

**Investigating cancer-like DNA  
methylation patterns and genome  
stability upon reprogramming to naïve  
human pluripotency**

---

Barts and the London School of Medicine and Dentistry  
Queen Mary University of London

Thesis submitted in partial fulfillment of the requirements  
of the Degree of Doctor of Philosophy

June 2019

**Author:**

**Hemalvi Patani**

## Statement of originality

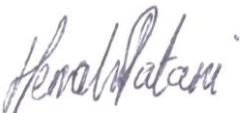
I, Hemalvi Patani, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged below and my contribution indicated. Previously published material is also acknowledged below.

I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material.

I accept that the College has the right to use plagiarism detection software to check the electronic version of the thesis.

I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university.

The copyright of this thesis rests with the author and no quotation from it or information derived from it may be published without the prior written consent of the author.

Signature: 

Date: 20th June 2019

Details of collaboration and publications:

Overlap analysis of sequencing data and CHIP-qPCR analysis was performed by Dr Michael D. Rushton, QMUL.

Analysis of whole genome sequencing was performed by Dr Gabriella Ficz, QMUL.

Analysis of DNA methylation array data from The Cancer Genome Atlas was performed by Dr Jonathan Higham and Dr Duncan Sproul at the MRC Institute of Genetics & Molecular Medicine, University of Edinburgh.

Processing of samples for protein mass spectrometry and raw data analysis was performed by Dr Saul Alvarez Teijeiro and Dr Pedro Cutillas, QMUL.

Mass spectrometry of nucleosides was performed by Dr David Oxley at the Mass Spectrometry facility at the Babraham Institute, Cambridge.

Microscopy imaging of markers of DNA damage and chromosomal instability was performed by Dr Alice Mazzagatti and Dr Sarah McClelland, QMUL.

Work presented in Chapters 3 & 4 of this thesis has been submitted for publication and is currently under review.

## Acknowledgements

Firstly, I thank my supervisor Gabriella Ficz for the opportunity to be the first PhD student in the lab and for providing all the training, tools and scientifically stimulating discussion that have allowed me to learn so much over the last four years. From our very first meeting, she always showed remarkable belief in me, and I could not be more grateful for the balance of freedom and support she has provided to enable me to become the scientist that I am today.

Next, I thank my colleagues in the Ficz lab for their constant day-to-day guidance, training and support both inside outside of the lab. I feel incredibly fortunate to have been part of such a fantastic group of people and scientists. A particular thank you to Emily Saunderson and Michael Rushton for answering all my silly questions, teaching me new techniques, and never saying no or losing patience with me – I admire you both for that. I thank Ateequllah Hayat and Lily Hoa for the endless entertaining discussions in the lab, both scientific and otherwise, which made my time in the lab so much fun. I thank you all for your friendship and I hope our tradition of social outings tied to some healthy competition continues.

It has been an absolute pleasure to work in the department of Haemato-oncology, and I thank all members, past and present, for creating such a pleasant and welcoming working atmosphere. I particularly thank the individuals with whom I shared an office, for all the scientific discussion and laughter. I'm grateful to have made such good friends with whom I have shared the joys and the hitches of the PhD process - I thank Mariette, Arran, Emma, Faith, Tahrima, Dave, Maru, Ryan, Henry, and Federico.

My thanks go to our collaborators Jonathan Higham and Duncan Sproul for their analysis of cancer methylation data, David Oxley for mass spectrometry of nucleosides, Saul Alvarez Teijeiro and Pedro Cutillas for generation and analysis of proteomics data, and Alice Mazzagatti and Sarah McClelland for microscopy of DNA damage markers. I am also grateful to the Medical Research Council and the Greg Wolf Fund for their funding and support.

Finally, I thank my parents for always ensuring I have had the best opportunities possible to enable me to be where I am today, and for a lifetime of love, support and belief in me. Thank you to my sister Harshnira for her lifelong friendship, for all our scientific discussions, and for paving the way for me at every stage in life. Last but by no means least, I thank Shahil Mehta for being my constant pillar of support; for the constant love and patience alongside all the motivation and encouragement. Thank you all for always being my number one fan.

## Abstract

Aberrant DNA hypermethylation of promoter CpG islands in the context of a hypomethylated genome is a hallmark of cancer. DNA hypomethylation is typically associated with genomic instability while CpG island hypermethylation is often linked to gene repression. The mechanisms underlying these changes and the role they play in cancer initiation and progression remain elusive as they are challenging to study once cancers have progressed, and no human experimental models exist to mechanistically investigate this phenomenon.

We observed that upon reprogramming of primed human embryonic stem cells to the naïve state, the acquisition of a globally hypomethylated genome is accompanied by hypermethylation of bivalent promoter CpG islands, thus resembling DNA methylation patterns characteristic of the human cancer methylome. We dissected the kinetics of these DNA methylation changes at high temporal resolution. We found that a subset of bivalent genes which are enriched in developmental pathways become hypermethylated, and showed that this is mirrored across multiple cancers, suggesting common underlying mechanisms of DNA hypermethylation.

To gain insight into the mechanism of hypermethylation, we investigated the dynamic expression of DNA methylation regulators upon reprogramming. We identified the *de novo* DNA methyltransferase DNMT3A as the enzyme primarily responsible for DNA hypermethylation, and characterised the consequences of its absence on the naïve pluripotent state. Additionally, we demonstrated a role of transcription factors and the pluripotency network in coordinating *de novo* methylation.

In parallel, we explored the impact of reprogramming on the genomic stability of naïve hESCs and investigated a potential relationship between reprogramming and DNA mutations. We observed evidence of chromosomal instability upon reprogramming, though the mutation frequency appears to remain unchanged.

The similarities between DNA methylation patterns acquired during reprogramming to naïve pluripotency and oncogenic transformation, as well as indications of genomic instability upon reprogramming suggest a wider utility for this reprogramming system in understanding cancer formation.

## Table of Contents

<b>Statement of originality</b> .....	2
<b>Acknowledgements</b> .....	4
<b>Abstract</b> .....	5
<b>List of Figures</b> .....	11
<b>List of Tables</b> .....	14
<b>Abbreviations</b> .....	15
<b>Introduction</b> .....	18
<b>1.1 Primed and Naïve Human Pluripotent States</b> .....	18
1.1.1 The origin of Embryonic Stem Cells .....	18
1.1.2 Human Embryonic Stem Cells in culture .....	18
1.1.3 Differences between primed and naïve human pluripotent states .....	21
<b>1.2 Epigenetic modifications of the DNA and histones</b> .....	24
1.2.1 What is epigenetics? .....	24
1.2.2 DNA methylation: an epigenetic modification .....	24
1.2.3 Biological functions of DNA methylation in mammalian development .....	26
1.2.4 DNA methylation occurs in different genomic contexts .....	27
1.2.5 Histone modifications: the histone code .....	28
<b>1.3 Mechanisms of DNA methylation</b> .....	32
1.3.1 The DNA methyltransferase family .....	32
1.3.2 Cross-talk between DNA methylation and histone modifications .....	36
1.3.3 Epigenetic mechanisms during development .....	40
<b>1.4 Mechanisms of DNA demethylation</b> .....	41
1.4.1 Active demethylation .....	41
1.4.2 Passive demethylation .....	43
<b>1.5 Epigenetic changes in cancer</b> .....	45
1.5.1 Aberrant DNA methylation patterns in cancer .....	45
1.5.2 DNA hypermethylation in cancer .....	45
1.5.3 Global hypomethylation and genomic instability .....	48
1.5.4 Disruption of epigenetic regulators in cancer .....	49
1.5.5 DNA methylation and mutation .....	50
<b>1.6 Epigenetic reprogramming</b> .....	52
1.6.1 Somatic cell reprogramming .....	52
1.6.2 Cellular reprogramming and the pluripotency network .....	54
1.6.3 Reprogramming and genomic stability .....	55

1.6.4 Reprogramming and cancer.....	55
1.7 Project Aims.....	58
<b>Chapter 2: Materials and Methods.....</b>	<b>59</b>
<b>2.1 Cell Culture .....</b>	<b>59</b>
2.1.1 Cell culture reagents .....	59
2.1.2 Cell lines.....	60
2.1.3 Isolation of Mouse Embryonic Fibroblasts .....	60
2.1.4 Human embryonic stem cell culture.....	60
2.1.5 HEK293T cell culture.....	61
2.1.6 Brightfield Microscopy of cells in culture .....	61
2.2 Embryoid body generation .....	61
2.3. Immunofluorescence .....	61
2.3.1 Immunostaining .....	61
2.3.2 Microscopy .....	63
2.4 Flow cytometry.....	63
2.5 Reverse Transcription Quantitative Real-time PCR (rt-qPCR) .....	63
2.6 Western Blotting .....	65
2.6.1 Protein Isolation .....	65
2.6.2 Protein Quantification .....	65
2.6.3 Gel Electrophoresis .....	65
2.6.4 Western Blot Transfer .....	66
2.6.5 Detection .....	66
2.7 Alkaline phosphatase assay .....	67
2.8 Stable shRNA knockdown cell line generation.....	67
2.8.1 Acquisition and preparation of shRNA plasmids .....	67
2.8.2 Generation of lentiviral particles.....	67
2.8.3 Transduction of hESCs and generation of stable knock down cell lines .....	68
2.9 Stable overexpression cell line generation .....	69
2.9.1 Acquisition of entry clones of target genes.....	69
2.9.2 Generation of expression vectors for target genes .....	69
2.9.3 Generation of lentiviral particles.....	69
2.9.4 Transduction of hESCs and generation of stable overexpression cell lines .....	70
2.10 Chromatin Immunoprecipitation and qPCR (ChIP-qPCR) .....	70
2.10.1 Chromatin extraction .....	70
2.10.2 Chromatin Immunoprecipitation.....	70
2.11 Targeted Bisulphite Sequencing .....	72

2.11.1 Primer design.....	72
2.11.2 DNA isolation and bisulphite conversion .....	73
2.11.3 PCR amplification of target regions.....	73
2.11.4 Fluidigm targeted amplicon sequencing .....	73
2.11.5 Amplicon library Sequencing .....	74
2.11.6 Data Analysis .....	74
2.12 Mass Spectrometry based Proteomics .....	77
2.12.1 Cell lysis and trypsin digestion.....	77
2.12.2 Desalting .....	78
2.12.3 Mass spectrometry data acquisition .....	78
2.12.4 Peptide identification and quantification .....	79
2.13 Mass Spectrometry of Nucleosides .....	79
2.13.1 Isolation and digestion of genomic DNA.....	79
2.13.2 Mass Spectrometry quantification of nucleosides .....	79
2.14 Infinium MethylationEPIC BeadChip assay .....	80
2.15 RNA sequencing.....	81
2.15.1 Library Preparation .....	81
2.15.2 Library Quantification and Quality Control.....	81
2.15.3 Library Sequencing .....	81
2.15.4 Data analysis.....	82
2.16 Targeted sequencing for mutational analysis .....	82
2.16.1 DNA isolation.....	82
2.16.2 PCR amplification of target regions.....	83
2.16.3 Library quantification.....	83
2.16.4 Amplicon library Sequencing .....	83
2.16.5 Data Analysis .....	83
2.17 Whole genome sequencing.....	84
2.17.1 Library Preparation .....	84
2.17.2 Library Quantification and Quality Control.....	85
2.17.3 Library Sequencing .....	85
2.17.4 Data Analysis .....	85
2.18 Bioinformatics Analysis .....	86
2.18.1 Bisulfite Sequencing Analysis .....	86
2.18.2 Overlap analysis .....	86
2.18.3 TCGA Analysis.....	86
2.19 Statistical Analysis.....	87



<b>Chapter 3. Results 1: The DNA methylation landscape during primed to naïve hESC reprogramming.....</b>	<b>88</b>
<b>3.1 Introduction.....</b>	<b>88</b>
<b>3.2 Naïve hESCs are morphologically and transcriptionally distinct from primed hESCs....</b>	<b>89</b>
<b>3.3 Naïve hESCs are characterised by global DNA hypomethylation and regional hypermethylation.....</b>	<b>92</b>
<b>3.4 Global demethylation occurs gradually during the primed to naïve transition .....</b>	<b>94</b>
<b>3.5 Reprogramming-associated hypermethylation is non-random and reproducible .....</b>	<b>97</b>
<b>3.6 Hypermethylation is a feature of the naïve human pluripotent state.....</b>	<b>99</b>
<b>3.7 Reprogramming-associated hypermethylation occurs primarily at bivalent CGIs .....</b>	<b>100</b>
<b>3.8 Reprogramming-associated hypermethylation is associated with developmental genes .....</b>	<b>103</b>
<b>3.9 Genes from developmental pathways are hypermethylated and downregulated during reprogramming.....</b>	<b>104</b>
<b>3.10 A subset of functionally distinct bivalent loci are hypermethylated upon reprogramming.....</b>	<b>107</b>
<b>3.11 Reprogramming-associated hypermethylation mirrors cancer hypermethylation... </b>	<b>108</b>
<b>3.12 Discussion .....</b>	<b>113</b>
<b>Chapter 4. Results 2: The mechanism of DNA hypermethylation during primed to naïve reprogramming.....</b>	<b>117</b>
<b>4.1 Introduction.....</b>	<b>117</b>
<b>4.2 Downregulation of maintenance methylation machinery coincides with global demethylation upon reprogramming.....</b>	<b>119</b>
<b>4.3 DNA hypermethylation is carried out by DNMT3A during early hESC reprogramming .....</b>	<b>121</b>
<b>4.4 DNMT3A and DNMT3B contribute to hypermethylation during the late transition of reprogramming.....</b>	<b>126</b>
<b>4.5 Hypermethylation has a functional role in naïve pluripotency.....</b>	<b>127</b>
<b>4.6 Hypermethylation of bivalent loci correlates with loss of H3K4me3.....</b>	<b>130</b>
<b>4.7 DNA hypermethylation is independent of TET1 loss .....</b>	<b>132</b>
<b>4.8 DNA hypermethylation is coordinated by the transcription factor and core pluripotency networks.....</b>	<b>134</b>
<b>4.9 Discussion .....</b>	<b>140</b>
<b>Chapter 5. Results 3: Investigating genome stability during primed to naïve reprogramming .....</b>	<b>145</b>
<b>5.1 Introduction.....</b>	<b>145</b>
<b>5.2 Levels of DNA damage markers remain constant across the transition from primed to naïve pluripotency.....</b>	<b>148</b>

5.3 hESC reprogramming is associated with increased chromosome errors during mitotic anaphase and increased ultrafine bridges .....	150
5.4 hESC reprogramming is associated with centrosome amplification and increased multipolar spindles.....	155
5.5 Long-term naïve hESCs exhibit increased aneuploidy .....	157
5.6 p53 protein expression is stabilised in naïve hESCs.....	161
5.7 Loss of DNMT3B does not enhance genomic instability upon reprogramming.....	162
5.8 The genomic mutation frequency is unaffected by global demethylation upon reprogramming.....	165
5.9 Hypermethylation of CGIs during reprogramming is not associated with an increased mutation frequency .....	166
5.10 Discussion .....	169
<b>Chapter 6. Conclusions and Future Directions .....</b>	<b>173</b>
6.1 Discussion & Future Directions .....	173
6.1.1 Dynamics and functions of DNA methylation in hESCs and cancer.....	173
6.1.2 Mechanisms of aberrant DNA methylation .....	176
6.1.3 Reprogramming of hESCs as a model system for cancer development .....	182
6.2 Concluding Remarks.....	186
<b>Bibliography.....</b>	<b>187</b>

## List of Figures

- 1.1. Summary of key changes that occur in the transition from primed to naïve pluripotency
- 1.2. DNA methylation reprogramming during development
- 1.3. An overview of DNA methylation and histone modifications
- 1.4. Common post-translational modifications associated with histone proteins
- 1.5. Resolution of bivalent chromatin domains
- 1.6. Structures of the mammalian DNMT family of proteins
- 1.7. Cross-talk between DNA methylation and histone modifications
- 1.8. The ten-eleven translocation methyltransferase dioxygenase (TET) family
- 1.9. Mechanisms of DNA demethylation
- 1.10. DNA methylation patterns in cancer
- 1.11. Cell state changes and Waddington's landscape
- 3.1. Naïve hESCs are morphologically and transcriptionally distinct from primed hESCs
- 3.2. Naïve hESCs can differentiate into all three germ layers
- 3.3. Naïve hESCs are characterised by global DNA hypomethylation and regional hypermethylation
- 3.4. Global demethylation occurs gradually during the primed to naïve transition
- 3.5. Global DNA demethylation can be detected at an individual CpG level during primed to naïve reprogramming
- 3.6. Reprogramming-associated hypermethylation is non-random and reproducible
- 3.7. Hypermethylation is a feature of the naïve human pluripotent state
- 3.8. Reprogramming-associated hypermethylation occurs primarily at bivalent CGIs
- 3.9. Reprogramming-associated hypermethylation is associated with developmental genes

- 3.10. Genes from developmental pathways are hypermethylated and downregulated during reprogramming
- 3.11. HOX genes are hypermethylated and upregulated
- 3.12. A subset of functionally distinct bivalent loci are hypermethylated upon reprogramming
- 3.13. Naïve bivalent hypermethylated CpGs exhibit increased methylation in tumour vs. normal tissues
- 3.14. Reprogramming-associated hypermethylation mirrors cancer hypermethylation
- 3.15. Common tumour suppressor genes are hypermethylated upon primed to naïve reprogramming
- 3.16. Cancer hypermethylation shows the strongest overlap with late transition hypermethylation
- 4.1. Possible mechanisms of hypermethylation in cancer
- 4.2. Downregulation of maintenance methylation machinery and upregulation of the TETs coincides with global demethylation upon reprogramming
- 4.3. *De novo* DNMTs are dynamically expressed during primed to naïve hESC reprogramming
- 4.4. Knockdown of DNMT3A abolishes hypermethylation in the early transition of hESC reprogramming
- 4.5. Hypermethylation is carried out by the dominantly expressed DNMT3A2 isoform
- 4.6. Both DNMT3A and DNMT3B contribute to hypermethylation during the late transition of reprogramming
- 4.7. Knockdown of DNMT3A and DNMT3B impacts naïve pluripotency
- 4.8. A subset of bivalent loci lose H3K4me3 and gain DNA methylation
- 4.9. Hypermethylation occurs independently of MECP2 loss
- 4.10. Hypermethylation occurs independently of TET1 loss
- 4.11. Transcription factors that bind hypermethylated loci are upregulated upon reprogramming

- 4.12. Knockdown of transcription factors reduces hypermethylation at target loci
- 4.13. Hypermethylation occurs independently of the 2i inhibitors
- 5.1. Levels of  $\gamma$ -H2AX remain unchanged upon primed to naïve reprogramming
- 5.2. Levels of 53BP1 show a slight increase in naïve compared to primed hESCs
- 5.3. There is a transient increase in the detection of lagging chromosomes during primed to naïve hESC reprogramming
- 5.4. There is a transient increase in the number of ultrafine bridges detected during primed to naïve hESC reprogramming
- 5.5. Spindle abnormalities can be detected in naïve hESCs
- 5.6. Aneuploidy can be detected in naïve hESCs through analysis of cellular DNA content
- 5.7. Abnormal chromosome numbers are more prevalent in naïve compared to primed hESCs
- 5.8. p53 protein expression is stabilised in naïve hESCs
- 5.9. Downregulation of DNMT3B does not enhance genomic instability during early reprogramming
- 5.10. Global demethylation does not affect the C to T mutation frequency
- 5.11. AID is transiently upregulated during the early transition of reprogramming
- 5.12. DNA hypermethylation at CGIs does not impact the frequency of C to T mutations
- 6.1 Model of de novo DNA hypermethylation upon primed to naïve hESC reprogramming
- 6.2 Reprogramming and cancer

## List of Tables

- 1.1. Culture conditions for the maintenance of naïve hESCs
- 1.2. Writers and erasers of histone lysine methylation
- 1.3. Specificities, roles and knockout phenotypes of DNMTs
- 2.1. Cell culture reagents
- 2.2. Primary antibodies used for immunostaining
- 2.3. Primer sequences used for rt-qPCR
- 2.4. Primary antibodies used for Western Blotting
- 2.5. shRNA sequences used to generate knock down cell lines
- 2.6. Entry clones used for overexpression of genes
- 2.7. Primers used for ChIP-qPCR
- 2.8. Bisulphite primer sequences for targeted bisulphite sequencing
- 2.9. Illumina P5 and P7 indexed adapters used for targeted bisulphite sequencing
- 2.10. Primer sequences for Fluidigm targeted bisulphite sequencing
- 2.11. Illumina P5 and P7 indexed adapters used for RNA-seq and WGS
- 2.12. Primer sequences used for targeted sequencing for mutation analysis

## Abbreviations

2iL	2 small inhibitors of MEK and GSK3 $\beta$ and hLIF
53BP1	p53 binding protein 1
5caC	5-carboxylcytosine
5fC	5-formylcytosine
5hmC	5-hydroxymethylcytosine
5mC	5-methylcytosine
AF	Alexa Fluor
AID/AICDA	Activation induced cytidine deaminase
APC	Allophycocyanin
B-CLL	B-cell chronic lymphocytic leukaemia
BER	Base excision repair
bFGF	Basic fibroblast growth factor
cDNA	Complementary DNA
CENP	Centromeric protein
CGI	CpG Islands
ChIP	Chromatin immunoprecipitation
DAPI	4',6-diamidino-2-phenylindole
DDR	DNA damage response
Dox	Doxycycline
DMP	Differentially methylated probe
DNMT	DNA methyltransferase
DSB	Double stranded break
ENCODE	Encyclopaedia of DNA elements

EpiSC	Epiblast stem cell
ESC	Embryonic stem cell
EZH2	Enhancer of Zeste homolog 2
FOXC1	Forkhead box C1
Gö	Protein kinase C inhibitor
GO	Gene ontology
GSK3 $\beta$	Glycogen synthase kinase 3 beta
hESC	Human embryonic stem cells
hLIF	Human recombinant leukaemia inhibitory factor
HR	Homologous recombination
ICM	Inner cell mass
iMEF	Irradiated mouse embryonic fibroblasts
iPS	Induced pluripotent stem cells
KD	Knock down
KLF2	Kruppel like factor 2
KSR	Knockout serum replacement
LIF	Leukaemia inhibitory factor
MBD	Methyl CpG binding domain protein
MECP2	Methyl CpG binding protein 2
MEK	Mitogen activated protein kinase kinase
mESC	Mouse embryonic stem cells
NFKB1	Nuclear factor kappa B subunit 1
NHEJ	Non-homologous end joining
OE	Overexpression
OSKM	OCT3/4, SOX2, KLF4, c-MYC



PCNA	Proliferating cell nuclear antigen
PE	Phycoerythrin
PKC	Protein kinase C
PRC2	Polycomb repressive complex 2
qPCR	Quantitative polymerase chain reaction
ROCKi	Rho-associated kinase inhibitor
RPA	Replication protein A
shRNA	Short hairpin RNA
SOX15	SRY-box 15
SUSD2	Sushi domain containing 2
TCGA	The Cancer Genome Atlas
TDG	Thymine DNA glycosylase
TET	Ten-eleven translocation methylcytosine dioxygenase
TSS	Transcriptional start site
UFB	Ultrafine Bridge
UHRF1	Ubiquitin like with PHD and ring finger domains 1
UNG	Uracil-N-glycosylase
VAF	Variant allele frequency
WGBS	Whole genome bisulfite sequencing
WGS	Whole genome sequencing
WT	Wild type
ZFX3	Zinc finger homeobox 3
$\gamma$ H2AX	Phosphorylated histone variant H2AX

# Introduction

## 1.1 Primed and Naïve Human Pluripotent States

### 1.1.1 The origin of Embryonic Stem Cells

A pluripotent stem cell is a single cell that is able to divide to give rise to all embryonic lineages that go on to make up an adult organism, as well as divide to produce daughter cells with comparable proliferative and developmental potential to the parent cell (Nichols and Smith 2009). Such cells can be found *in vivo* within early mammalian embryos (Nichols and Smith 2012). Pluripotency can be translated into an *in vitro* property through the derivation of stem cell lines, which can be propagated in culture, and enable maintenance of the pluripotent phenotype (Nichols and Smith 2012). The precise developmental stages of the embryo from which the stem cells are isolated determine the characteristics of the cells in culture. Typically, embryonic stem cells (ESCs) are derived from the inner cell mass (ICM) of the pre-implantation blastocyst, while epiblast stem cells (EpiSCs) are derived from the post-implantation epiblast (Brons et al. 2007; Tesar et al. 2007).

### 1.1.2 Human Embryonic Stem Cells in culture

Human embryonic stem cells (hESCs) are derived from the ICM of pre-implantation human blastocysts (Thomson et al. 1998). Their morphology, and their transcriptional and epigenetic profiles, however, exhibit more resemblance to murine EpiSCs rather than the more naïve mouse embryonic stem cells (mESCs) (Nichols and Smith 2009). It has become increasingly common to view pluripotency as a state with fluctuating cellular phenotypes. *In vitro*, two such states, with distinct culture conditions have been defined: naïve pluripotency that represents the pluripotent state of cells in the pre-implantation epiblast, and primed pluripotency that corresponds to cells primed for differentiation in the post-implantation epiblast (Nichols and Smith 2009). More recently, a third stage of formative pluripotency has been proposed as an intermediate period between naïve and primed pluripotency, but remains to be tested *in vitro* (Smith 2017). Until recently, the majority of hESC lines have been maintained in the primed state, in culture conditions containing fibroblast growth factor (FGF). In these conditions, hESCs exist as a population of cells heterogeneous for pluripotency markers, and are primed for differentiation (Hackett and Surani 2014). Manipulation of the culture conditions by a

variety of methods enables the reprogramming of primed hESCs to a more naïve state, as well as subsequent maintenance and survival of naïve hESCs, as shown in Table 1.1 (Hanna et al. 2010; Gafni et al. 2013; Chan et al. 2013; Ware et al. 2014; Theunissen et al. 2014; Takashima et al. 2014; Chen et al. 2015; Duggal et al. 2015; Qin et al. 2016; Guo et al. 2017). In parallel, human preimplantation epiblast cells have been comprehensively characterised at a molecular level, providing a guideline for assessing naïve human pluripotency and enabling direct comparisons between the *in vitro* and *in vivo* naïve pluripotent states (Guo et al. 2014; Smith et al. 2014; Yan et al. 2013; Blakeley et al. 2015; Stirparo et al. 2018).

Each of the naïve state culture conditions involves the use of two small inhibitors (2i) of mitogen-activated protein kinase kinase 1/2 (MEK1/2) and glycogen synthase kinase 3 beta (GSK3 $\beta$ ) in the presence of leukaemia inhibitory factor (LIF). This is based on the knowledge that culturing mESCs in 2i+LIF results in a naïve state of pluripotency that better represents the cells of the preimplantation epiblast (Ying et al. 2008). Inhibition of the FGF signalling pathway that occurs through the use of the MEK1/2 and GSK3 $\beta$  inhibitors has been shown to drive genome-wide DNA methylation of mESCs, resulting in a globally hypomethylated state representative of the ICM (Ficz et al. 2013; Habibi et al. 2013). Inhibition of GSK3 $\beta$  additionally upregulates Wnt signalling which is also activated through other mechanisms and helps to sustain the naïve state (Xu et al. 2016).

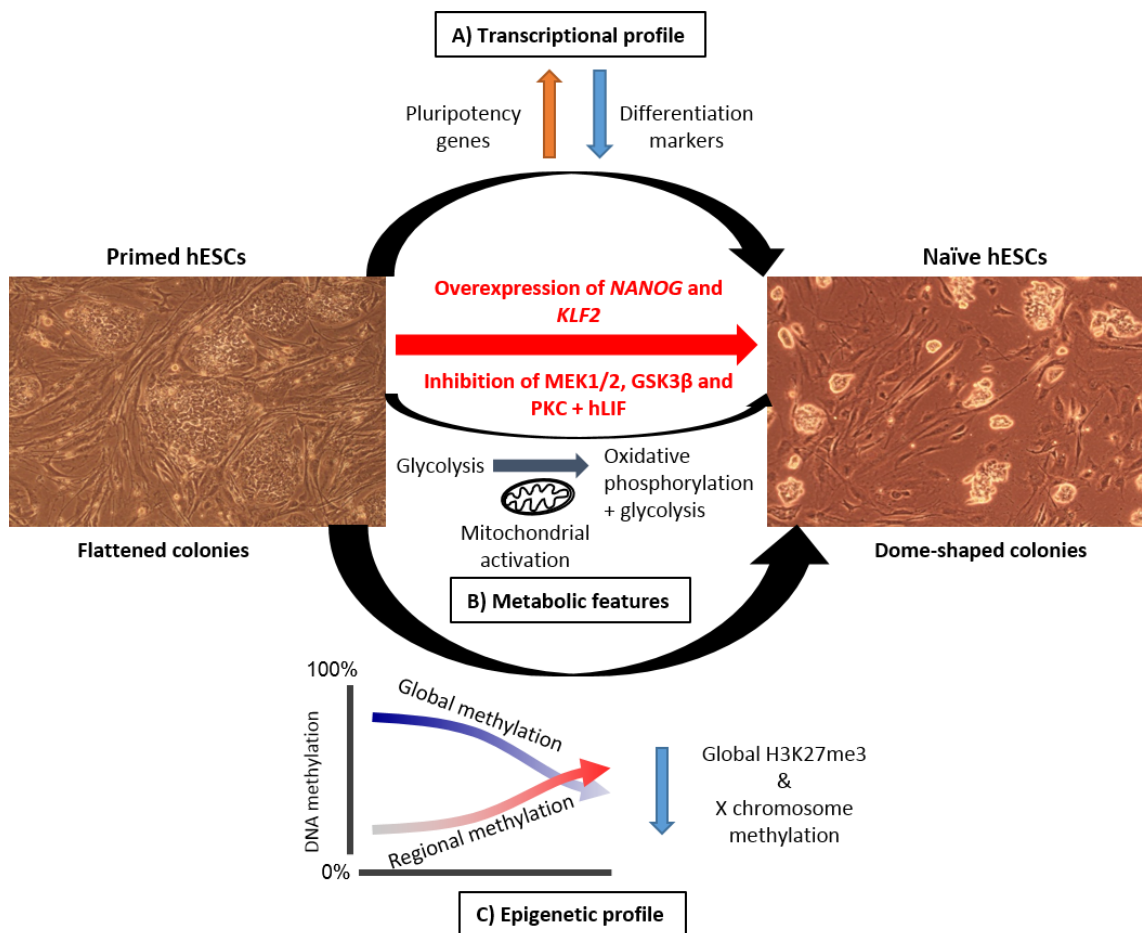
	Hanna (2010)	Gafni: NHSM (2013)	Chan: 3iL (2013)	Ware (2014)	Theunissen: 5iLAF (2014)	Takashima: t2iL+PKCi (2014)	Chen: TL2i (2015)	Duggal (2015)	Qin (2016)	Guo: t2iL+PKCi, cR (2017)
Inhibitors	MEKi	MEKi	MEKi	MEKi	MEKi	MEKi	MEKi	MEKi	MEKi	MEKi
	GSK3i	GSK3i	GSK3i	GSK3i	GSK3i	GSK3i	GSK3i	GSK3i	GSK3i	GSK3i
		JNKi								
		P38i								
		PKCi				PKCi				PKCi
		ROCKi			ROCKi					ROCKi
			BMPi							
					BRAFi					
					SRCi					
				HDACi						HDACi
								PKAi	PKAi	
								Ascorbic acid		
	Growth factors		bFGF	bFGF	bFGF				bFGF	
		TGFβ	TGFβ							
					Activin					
Cytokines	hLIF	hLIF	hLIF	hLIF	hLIF	hLIF	hLIF	hLIF	hLIF	
Base	N2B27	N2B27	mTeSR1	DMEM, 20% KSR	N2B27	N2B27	N2B27	mTeSR1	N2B27	N2B27
O <sub>2</sub> level	20%	20% or 5%	-	5%	5%	5%	5%	5%	-	5%
Transgenes	OCT4, KLF2, KLF4					NANOG, KLF2	STAT3			

**Table 1.1 Culture conditions for the maintenance of naïve hESCs.** This table shows the combinations of small molecule inhibitors, cytokines and growth factors with varying basal media and oxygen levels, used by different groups to reprogram primed hESCs and sustain the survival of naïve hESCs in culture. Some methods additionally use transgenes to establish the naïve state. Each approach gives rise to naïve cells that display varying molecular profiles, perhaps representing subtly different stages in the transition from primed to naïve pluripotency.

In this thesis, we use the *NANOG*/Kruppel-like factor 2 (*KLF2*) overexpression + 2iLGö reprogramming method derived by Takashima *et al* (Takashima et al. 2014), which along with the 5iLAF method is considered to give rise to naïve hESCs that most faithfully represent the transcriptional profile of the human preimplantation blastocyst (Huang, Maruyama, and Fan 2014). In this method, primed hESCs are reprogrammed to a naïve state more analogous to cells of the ICM (Ying et al. 2008), through transient induced overexpression of *NANOG* and *KLF2* transgenes, and inhibition of FGF signalling through the use of two small inhibitors of MEK1/2 and GSK3β in the presence of human recombinant leukaemia inhibitory factor (hLIF; collectively referred to as 2iL). For long-term propagation of the naïve hESCs, after removal of doxycycline (which is used to induce *NANOG* and *KLF2*), a pan-protein kinase C (PKC) inhibitor, Gö, is used to stabilise the cell state.

### 1.1.3 Differences between primed and naïve human pluripotent states

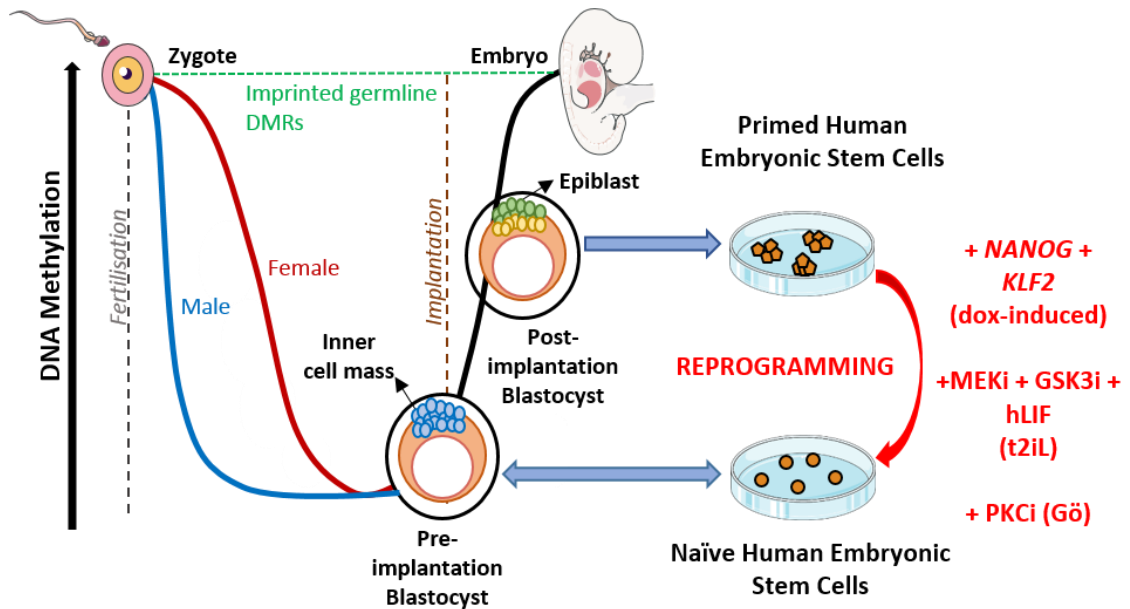
Naïve hESCs differ from their primed counterparts morphologically, transcriptionally, metabolically and epigenetically, as summarised in Figure 1.1. Primed hESCs grow as large, flat colonies resembling tightly packed cobblestones, while naïve hESCs grow in smaller, dome-shaped colonies. The gene expression profiles of primed and naïve hESCs have been extensively characterised, demonstrating a cohort of pluripotency genes that is enriched in naïve hESCs and human preimplantation epiblast cells, though the expression of these genes varies to a large degree between various naïve hESC lines derived under different conditions (Huang, Maruyama, and Fan 2014) (Figure 1.1a). The naïve gene regulatory network includes genes such as *KLF4*, *KLF17*, *TFCP2L1*, and *DPPA5*, while the primed gene regulatory network includes genes such as *ZIC2*, *ZIC3*, *OTX2* and *DUSP6*, and both primed and naïve hESCs express the core pluripotency genes *OCT4*, *SOX2* and *NANOG* (Stirparo et al. 2018; Pastor et al. 2016; Takashima et al. 2014; Theunissen et al. 2014). In addition to differences in gene expression, primed and naïve hESCs can be distinguished based on specific cell surface markers that are unique to each cell state (Collier et al. 2017). Primed hESCs, which are primarily glycolytic, undergo metabolic realignment upon the transition to the naïve state, through mitochondrial activation that enables the use of oxidative phosphorylation (Takashima et al. 2014) (Figure 1.1b).



**Figure 1.1 Summary of key changes that occur in the transition from primed to naïve pluripotency.** This Figure summarizes some of the known key differences between the primed hESCs which grow in large, flattened colonies, and the naïve hESCs which grow in smaller, dome-shaped colonies. a) Upon reprogramming, the cells undergo transcriptional upregulation of pluripotency factors and a downregulation of genes involved in differentiation pathways. b) Primed hESCs are primarily glycolytic, but upon reprogramming, undergo profound mitochondrial activation and begin to utilize oxidative phosphorylation in addition to glycolysis (Takashima et al. 2014). c) Upon reprogramming, the cells undergo global demethylation alongside some regional hypermethylation (Guo et al. 2017). The X chromosome also undergoes widespread demethylation and cells display greatly reduced levels of repressive histone 3 lysine 27 trimethylation (H3K27me3) (Takashima et al. 2014).

The epigenome undergoes global remodelling as cells transition from the primed to the naïve state of pluripotency (Figure 1.1c). Global hypomethylation of the DNA is a hallmark of both human and mouse preimplantation embryos, constituting the second wave of global epigenetic reprogramming during early mammalian development (Smith et al. 2012; Smith et al. 2014; Guo et al. 2014), with the first wave of epigenetic erasure occurring in primordial germ cells (Seisenberger et al. 2012) (Figure 1.2). The cells of the ICM from which conventional hESCs are derived would therefore be expected to be hypomethylated. Paradoxically, primed hESCs in culture display global hypermethylation with levels resembling that of somatic cells (Hackett and Surani 2014). Upon reprogramming of primed hESCs to the naïve state, the cells

undergo global DNA demethylation across a range of genomic contexts, including demethylation of imprints (Takashima et al. 2014; Pastor et al. 2016). Furthermore, female naïve hESCs display greatly reduced levels of DNA methylation on the X-chromosome, resulting in two active X-chromosomes (Figure 1.1c) (Takashima et al. 2014).



**Figure 1.2. DNA methylation reprogramming during development.** During early development, there are two waves of global epigenetic reprogramming. The first wave of erasure takes place upon in primordial germ cells as they migrate towards the genital ridge. Following this, the DNA methylation landscapes are established in male and female cells at varying rates. The second wave of epigenetic erasure occurs post-fertilisation, once again at distinct rates for male and female cells, with the lowest levels of methylation coinciding with the pre-implantation blastocyst. hESCs derived from the inner cell mass would be expected to have low levels of methylation, however primed hESCs are actually hypermethylated and can be reprogrammed *in vitro* to a naïve, hypomethylated state. Figure adapted from: Smallwood *et al* (Smallwood and Kelsey 2012).

In this thesis, we will focus primarily on the epigenetic changes occurring between the two states, with a specific emphasis on DNA methylation and its interplay with histone modifications. In parallel, we will explore the genomic stability of the cells as their epigenomes are remodelled during the transition from the primed to the naïve hESC state.

## 1.2 Epigenetic modifications of the DNA and histones

### 1.2.1 What is epigenetics?

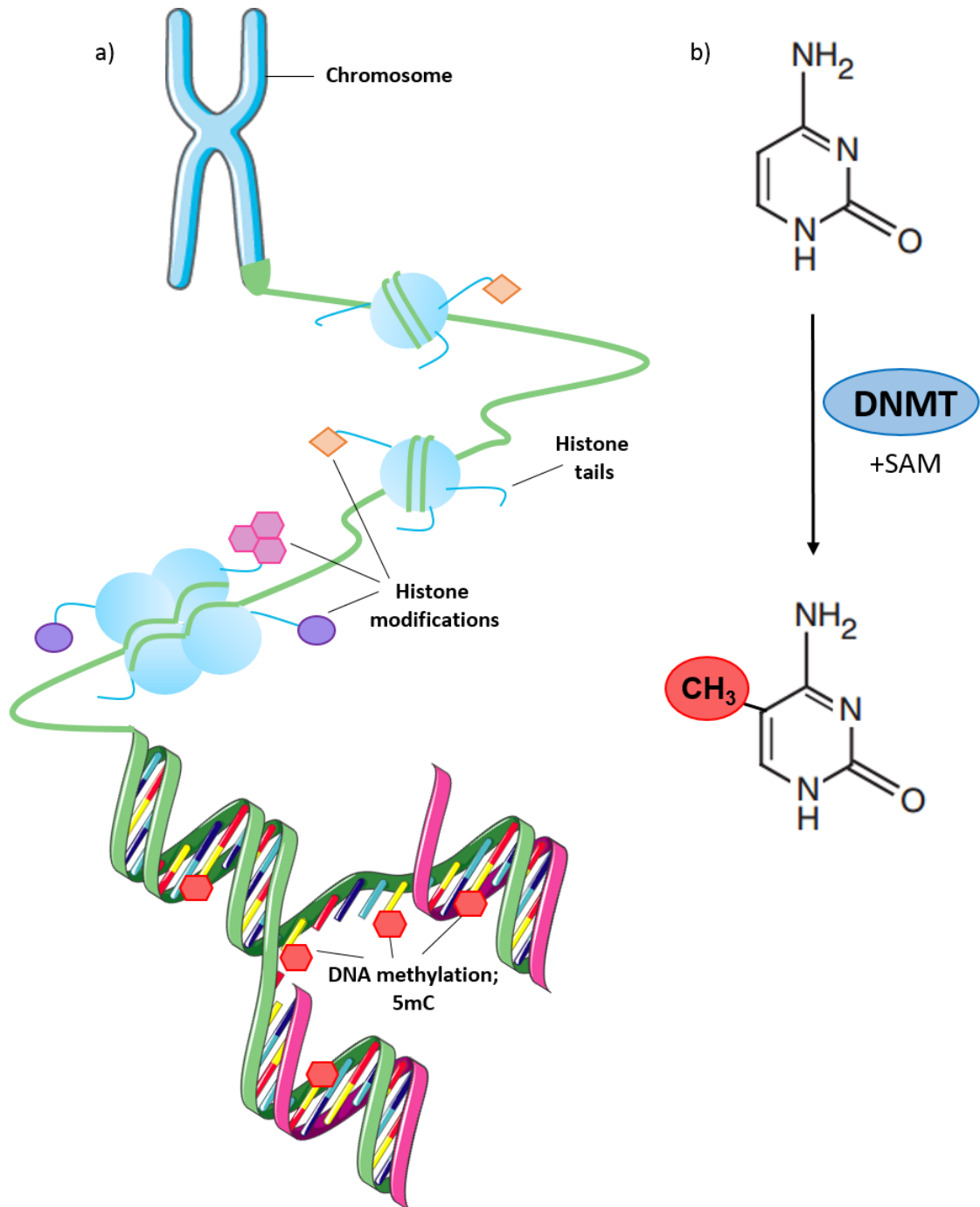
The term epigenetics was first coined in the 1940s by Conrad Waddington, derived from the Greek word “epigenesis”. Epigenetics was defined as “a branch of biology which studies the causal interactions between genes and their products which bring the phenotype into being” (Waddington 1942, 1968), incorporating all molecular pathways capable of modulating genotype and phenotype. Over the years, the definition of the word has narrowed, and the current accepted definition is “the study of changes in gene function that are mitotically and/or meiotically heritable and that do not entail a change in DNA sequence” (Wu and Morris 2001). DNA methylation and covalent modifications to the tails of histone proteins are two of the most broadly studied classes of epigenetic modifications.

### 1.2.2 DNA methylation: an epigenetic modification

DNA methylation, outlined in Figure 1.3a, is the most well characterised epigenetic modification that occurs in a cell (Wu and Zhang 2010). Modifications to tails of histone proteins, around which the DNA is wrapped, are the other major class of epigenetic modifications (Cedar and Bergman 2009) (Figure 1.3a). DNA methylation is carried out by a group of proteins called DNA methyltransferases (DNMTs), which catalyse the addition of a methyl group to the 5-position of a cytosine (C) residue, using S-adenosyl methionine (SAM) as a methyl donor, resulting in 5-methylcytosine (5mC) (Goll and Bestor 2005) (Figure 1.3b). Although DNA methylation is a relatively stable epigenetic mark and is maintained across cell division and replication, it can be lost from the DNA by both active and passive mechanisms (Wu and Zhang 2014). In mammals, DNA methylation typically occurs at a CpG dinucleotide, where a cytosine nucleotide is followed by a guanine nucleotide in the 5'-3' direction. CpG sites are underrepresented globally across the genome due to the high mutation rate of methylated cytosines (Bird 1980), but 70-80% of those that are present are methylated in somatic cells (Ehrlich et al. 1982). Among the globally sparse CpG landscape, short regions of high CpG density, called CpG islands (CGIs) are interspersed in the genome, and are commonly unmethylated in germ cells, early developmental cell types and somatic cells (Deaton and Bird 2011; Bird et al. 1985). CGIs are defined as regions with an observed/expected CpG ratio > 0.6 (Gardiner-Garden and Frommer 1987). Approximately 70% of gene promoters in the human genome are associated with a CGI including all housekeeping genes and many developmental regulators and tissue-specific genes (Saxonov, Berg, and Brutlag 2006). Tissue specific



methylation patterns have been observed at a subset of CGIs associated with developmental genes (Illingworth et al. 2008).



**Figure 1.3. An overview of DNA methylation and histone modifications.** a) DNA in the nucleus is wrapped tightly around octamers of histone proteins to form chromatin, which is further tightly packed to make up chromosomes. Two of the main epigenetic changes that can occur are post-translational modifications to histone tails or direct methylation of the cytosine base (depicted in yellow) on the DNA. b) DNA methylation is the addition of a methyl group to the 5' position of a cytosine base by a DNA methyltransferase (DNMT) enzyme, using S-adenosyl methionine (SAM) as a methyl donor.

### 1.2.3 Biological functions of DNA methylation in mammalian development

Large-scale studies of DNA methylation have observed limited heterogeneity in DNA methylation patterns between individuals, while the patterns across cell and tissue types exhibit greater variation, indicative of a relationship between DNA methylation and differentiation (Eckhardt et al. 2006). DNA methylation typically occurs in a tissue and cell type specific pattern. Correct patterns of DNA methylation are crucial for early development and cell differentiation. At the level of individual genes, the role of DNA methylation in influencing gene expression remains a subject of debate. Early studies on DNA methylation dating back to 1979 provided indirect indications for a role of DNA methylation in silencing gene expression (McGhee and Ginder 1979). This was supported by more direct studies showing that the perturbation of DNA methylation in mESCs using the cytidine analog 5-azacytidine inhibited cell differentiation that was dependent on gene expression changes (Jones and Taylor 1980), and that 5-azacytidine could restore the expression of previously silenced genes (Mohandas, Sparkes, and Shapiro 1981). Two models evolved to explain the mechanisms by which 5mC may operate once added to the DNA. The modified base can inhibit the DNA-binding ability of factors that are required for transcription if the methylated cytosine falls within their respective DNA sequence recognition motifs (Watt and MoUoy 1988; Zhu, Wang, and Qian 2016). Alternatively, 5mC can be recognised by methyl-CpG binding domain proteins (MBDs), and this evidence initially came from the identification of methyl-CpG binding proteins 1 and 2 (MECP1 and MECP2) (Meehan et al. 1989). These MBD-containing proteins interact with other co-repressor proteins that are able to influence transcriptional regulation and modify surrounding chromatin (Jones et al. 1998; Nan et al. 1998), as well as influencing DNA repair (Meng et al. 2015). Several studies in later years added to the pool of evidence supporting a causal relationship between DNA methylation at gene promoters and transcriptional repression, particularly as actively transcribed gene promoters were found to feature very low levels of DNA methylation (Siegfried et al. 1999; Boyes and Bird 1991; Naveh-Manly and Cedar 1981). Moreover, studies into the developmental processes of X-chromosome inactivation and genomic imprinting have provided evidence for DNA methylation in mediating gene silencing. In both of these contexts, DNA methylation has been shown to silence the underlying DNA, enabling cells to retain a single active X-chromosome or expression of imprinted genes from a single gene copy (Wolf 1984; Li, Beard, and Jaenisch 1993). Interestingly, despite the association between *de novo* methylation and gene silencing, upon global demethylation that occurs in primordial germ cells and during conversion of mESCs to the naïve state, DNA

methylation and transcription appear to be uncoupled, as the hypomethylated state does not result in re-expression of genes (Ficz et al. 2013; Seisenberger et al. 2012).

On a more global scale, DNA methylation also plays a role, as approximately 70% of the genome is methylated in somatic cells, and modern methylation mapping technologies have indicated that these highly methylated sequences constitute repetitive elements and transposons, satellite DNA, intergenic DNA and gene exons (Li and Zhang 2014). Global DNA methylation is key to maintenance of genomic integrity through silencing of retrotransposons (Bestor 2004) and it is known that DNA methylation is essential for developmental viability as targeted disruption of the DNMTs in mice results in embryonic lethality (Okano 1999; Li, Bestor, and Jaenisch 1992).

#### 1.2.4 DNA methylation occurs in different genomic contexts

Although typically considered a repressive mark associated with transcriptional silencing at CGI gene promoters, it has become increasingly evident that DNA methylation can have varying outcomes, dependent on the underlying genomic context (Jones 2012). CGIs located at the transcriptional start sites (TSS) of genes are commonly unmethylated, and their methylation is associated with long-term gene silencing. DNA methylation is typically considered to provide a stable lockdown on chromatin that has been silenced by other mechanisms, though there is evidence that DNA methylation can also have an instructive role in initiating silencing (Jones 2012). Genes without CGIs at their TSS can also display substantial variation in the levels of promoter methylation, but the functional relationship of this methylation with gene expression is less clear (Jones 2012). CGIs also exist within the bodies of genes, and outside of genes. Their functions here remain unknown, though it has been postulated that these regions may represent orphan or alternative promoters (Illingworth et al. 2010), however DNA methylation within gene bodies is not associated with transcriptional repression. Methylation of alternative promoters within gene bodies can, however, contribute to the control of alternative promoter usage (Maunakea et al. 2010). Variable levels of DNA methylation are also detected within gene enhancers and may have an impact upon enhancer activity (Schmidl et al. 2009).

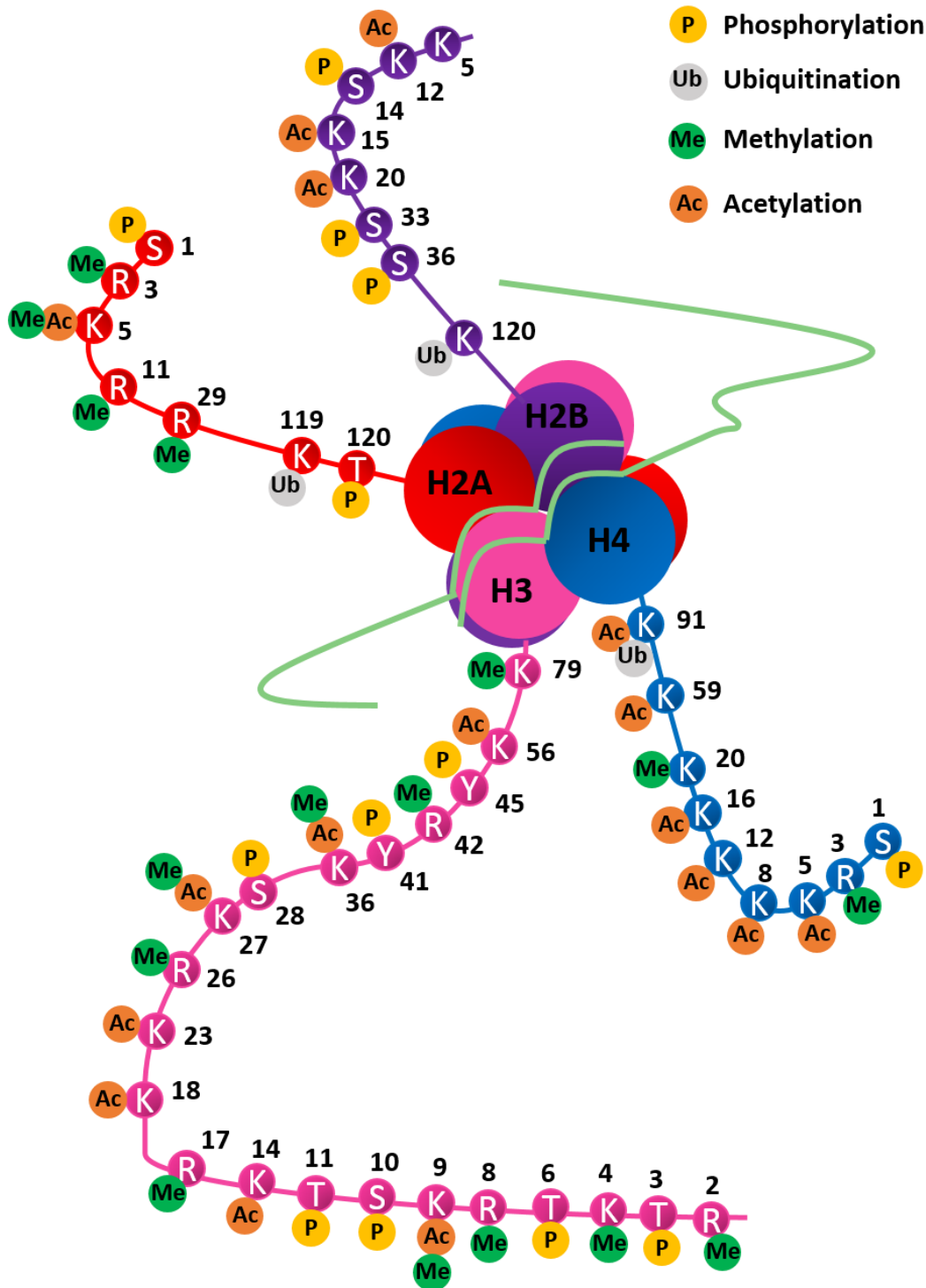
Aside from the genomic context of DNA methylation, studies have shown that the DNA sequences up to 2 kilobases (kb) away from the CGIs, termed CGI shores, are prone to large variation in DNA methylation levels across tissue types and in disease contexts (Irizarry et al.

2009). Moreover, these CGI shores have been shown to be susceptible to changes in methylation upon cellular reprogramming (Doi et al. 2009).

### 1.2.5 Histone modifications: the histone code

The other well characterised class of epigenetic modifications are post-translational modifications to the N-terminal tails of histone proteins, which were first discovered in 1964 (Allfrey, Faulkner, and Mirksy 1964). The histone tails can be modified in a variety of ways, each with a different impact upon chromatin structure, and the various combinations of histone modifications can give rise to numerous chromatin states that can influence transcription, DNA replication, DNA repair and recombination, by modifying chromatin structure or providing a binding platform for transcription factors (Kouzarides 2007). The most well characterised post-translational modifications to histone tails include methylation, acetylation, phosphorylation and ubiquitination, which can occur on lysine (K), arginine (R), serine (S), threonine (T) or tyrosine (Y) residues on any of the variants of histone proteins that constitute a histone octamer (Kouzarides 2007). Figure 1.4 depicts the majority of possible histone modifications that have been detected to date, though additional modifications are continually discovered as technologies advance.

For this thesis, we will focus primarily on the methylation of lysine residues, specifically those occurring on histone H3, as these are particularly involved in the interplay with DNA methylation (Cedar and Bergman 2009). Lysine residues can be either mono-, di- or trimethylated, which adds further complexity to the histone code and the associated functional response (Kouzarides 2007). A combination of the methylation state and the position of the lysine residue on the histone tail determines whether the lysine methylation confers an active or repressed transcriptional or chromatin state (Hyun et al. 2017). Each individual type of histone lysine methylation on histone H3 is written, read and erased by distinct proteins (Hyun et al. 2017), which are summarised in Table 1.2. Typically, H3K4, H3K36 and H3K79 modifications are thought to mark active transcription along with histone acetylation, H3K9 and H3K27 are associated with a repressive chromatin state, and H3K4me1 is enriched at gene enhancers (Kouzarides 2007).

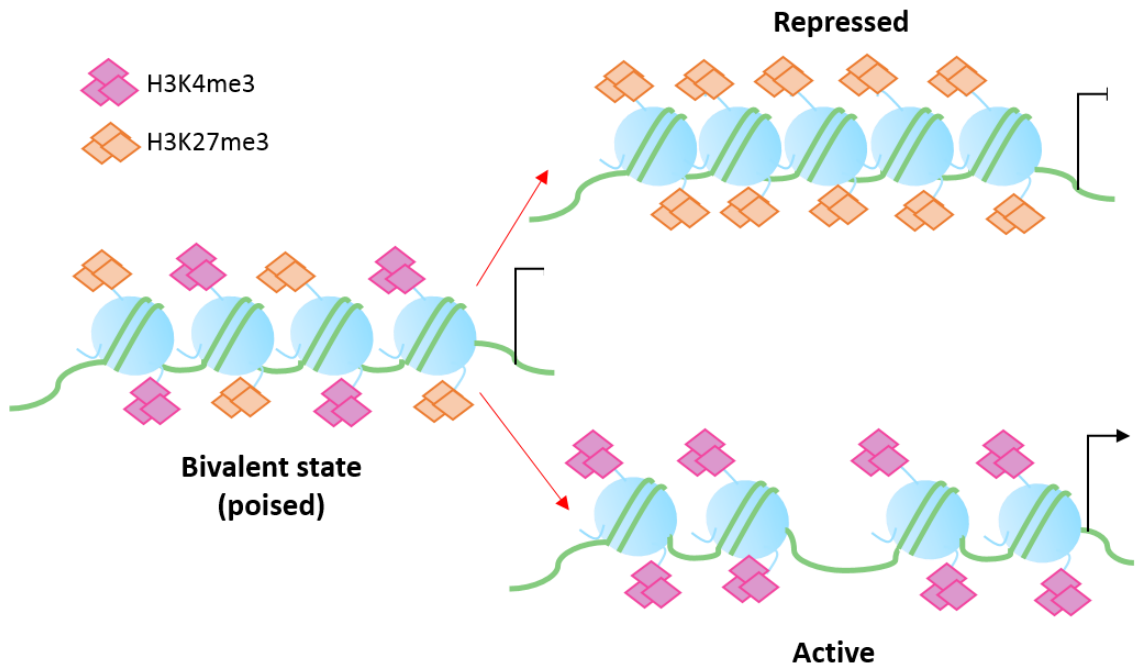


**Figure 1.4. Common post-translational modifications associated with histone proteins.** Histone octamers consisting of H2A, H2B, H3 and H4 make up nucleosomes. DNA (shown in green) is wrapped around nucleosomes. This figure depicts the positions of post-translational modifications (phosphorylation, ubiquitination, methylation and acetylation) located on each of the histone tails, with the location of the amino acid on the histone tail indicated with a number and the type of amino acid indicated with the single letter amino acid code.

Histone modification	Writers (methyltransferases)	Methylation state specificity	Erasers (demethylases)	Methylation state specificity
<b>H3K4</b>	SET1A/B, MLL1/2, PRDM9	me1/2/3	LSD1/2	me1/2
	MLL3/4	me1/2	NO66	me1/3
	SMYD1/2, SET7/9	me1	JARID1A/C/D	me2/3
	PRMD9	me1/2/3	JARID1B	me1/2/3
<b>H3K9</b>	SUV39H1/2	me2/3	JHDM2 family, PHF8	me1/2
	G9a, GLP, SETDB1	me1/2	JMHD3 family	me2/3
	PRDM family	-		
<b>H3K27</b>	EZH1, EZH2	me2/3	UTX, UTY, JMJD3	me2/3
			PHF8, KIAA1718	me2
<b>H3K36</b>	SETD2	me3	JHDM1/3 family	me1/2
	SETD3, NSD1-3	me1/2	JHMD3 family	me2/3
	SMYD2, ASH1L, SETMAR	me2		
<b>H3K79</b>	DOT1L	me1/2/3	-	-

**Table 1.2. Writers and erasers of histone lysine methylation.** This table summarises the known human modifiers (methyltransferases and demethylases) of histone lysine methylation at histone H3, and their methylation state specificities. Information compiled from (Hyun et al. 2017).

In ESCs, regions of co-enrichment of H3K4me3 and H3K27me3 mark regions of DNA termed bivalent regions which regulates changes in transcriptional activity from poised to active or silenced states upon lineage specification and cell differentiation (Mikkelsen et al. 2007; Bernstein et al. 2006; Azuara et al. 2006). Upon differentiation, the resolution of bivalent chromatin occurs through removal of either H3K4me3 or H3K27me3, along with changes to a number of other histone modifications and DNA methylation, resulting in repressed or active chromatin respectively (Figure 1.5). Although bivalency has primarily been studied in ESCs, bivalent chromatin is not unique to pluripotent cells as adult stem and progenitor cells such as neural progenitors also exhibit chromatin domains marked by H3K4me3 and H3K27me3 (Mohn et al. 2008; Oguro et al. 2010).

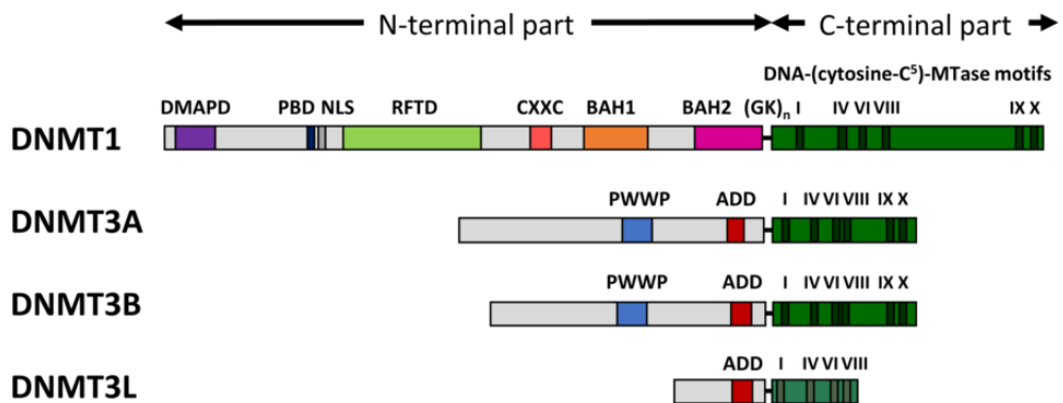


**Figure 1.5 Resolution of bivalent chromatin domains.** Bivalent chromatin at poised promoters in ESCs, marked by H3K4me and H3K27me3, is resolved into either a repressed state marked by H3K27me3 or an actively transcribing state marked by H3K4me3.

## 1.3 Mechanisms of DNA methylation

### 1.3.1 The DNA methyltransferase family

The DNMT family of proteins (Figure 1.6) are responsible for establishing and maintaining DNA methylation marks on the DNA.



**Figure 1.6. Structures of the mammalian DNMT family of proteins.** The DNMT family has a number of highly conserved C-terminal motifs in the catalytic region, shown as thick black lines (indicated as I–X). DNMT3L does not have catalytic activity. Functional domains required for enzymatic regulation are present in the N-terminal regions of DNMTs. Sub-domains for DNMT1 include a proliferating cell nuclear antigen binding site (PBD), nuclear localisation signal region (NLS), a replication fork targeting domain (RFTD), a cysteine rich CXXC domain (binds DNA containing CpG dinucleotides), and two bromo-adjacent homology domains (BAH1/2). For the DNMT3s, domains include an ADD (ATRX DNMT3 DNMT3L) domain, and for DNMT3A/B, an additional PWWP domain (highly conserved proline–tryptophan–tryptophan–proline motif that is involved in protein–protein interactions). Figure taken from: Jeltsch and Jurkowska, 2016 (Jeltsch and Jurkowska 2016).

There are four main members of the mammalian DNMT family, of which three are enzymatically active. DNMT1 shows a stronger preference for methylating hemimethylated DNA *in vitro*, while DNMT3A and DNMT3B do not show a preference for hemimethylated DNA compared to unmethylated DNA (Okano 1998; Okano 1999). The third of the DNMT3 family, DNMT3L, is catalytically inactive (Figure 1.6). An additional DNMT, DNMT2 shows activity *in vitro* but does not play a role in CpG methylation in mammalian cells (Okano, Xie, and Li 1998; Hermann, Schmitt, and Jeltsch 2003). Much of what is known about the functions of these enzymes is determined by gene knockout studies in embryos or ESCs, typically in murine cells and more recently in human ESCs (Table 1.3). DNMT3A and DNMT3B are thought to primarily be responsible for *de novo* methylation of DNA and are required for establishing DNA methylation patterns in the early developing embryo and setting up genomic imprints during



germ cell development (Okano 1999). The two proteins have distinct functions throughout embryonic development, both temporally and spatially, with DNMT3A being primarily responsible for maternal imprinting and DNMT3B for methylation of pericentromeric repeats and X-chromosome inactivation (Meng et al. 2015). This suggests that each enzyme has regional specificity determined by regulation in their N-terminal domains. DNMT3A and DNMT3B are expressed at high levels in the early developing embryo, but their expression decreases as cells differentiate. While DNMT3L does not exhibit catalytic activity, it is highly expressed in germ cells and in early embryo development, where it is essential for the establishment of maternal imprints in the developing oocyte, and as a mandatory cofactor for the *de novo* methyltransferases (Goll and Bestor 2005; Liao et al. 2012). DNMT1 is predominantly a maintenance methyltransferase that preserves established patterns of DNA methylation on the daughter DNA strand as cells undergo DNA replication upon dividing (Goll and Bestor 2005). DNMT1 is highly expressed in dividing cells and is ordinarily localised to replication foci (Leonhardt 1992). DNMT1 is considered to exhibit imperfect fidelity of maintenance methylation activity in the absence of *de novo* methylation activity of DNMT3A and DNMT3B (Liao et al. 2015). Additionally, *de novo* methylation activity of DNMT1 has also been reported in various cellular settings (Jair et al. 2006; Fatemi et al. 2002; Vertino et al. 1996; Bestor 2000).

The catalytic domain is present in the C-terminal of the DNMT enzymes, with the exception of DNMT3L (Figure 1.6). The N-terminal regions of the DNMT proteins differ between DNMT1 and the DNMT3 enzymes (Jurkowska, Jurkowski, and Jeltsch 2011). The primary role of the N-terminal is in regulation and targeting of the enzymes by mediating their interaction with chromatin and with other proteins. The N-terminus of DNMT1 contains a number of domains that mediate its interaction with replicating DNA (Jurkowska, Jurkowski, and Jeltsch 2011) (Figure 1.6). The N-termini of the DNMT3 enzymes contain domains that mediate interactions of the enzymes with chromatin. DNMT3A/B/L each possess an ADD domain capable of mediating allosteric activation of the enzymatic activity upon interaction with unmethylated H3K4 (Otani et al. 2009; Ooi et al. 2007; Guo et al. 2015; Zhang et al. 2010), and both DNMT3A and DNMT3B also possess a PWWP domain capable of interacting with H3K36me3 and targeting the enzymes to pericentromeric chromatin and gene bodies (Ge et al. 2004; Dhayalan et al. 2010; Baubec et al. 2015; Chen, Tsujimoto, and Li 2004).

Each of the human DNMT genes makes use of splicing and alternative promoter usage, generating isoforms of the proteins whose expression can vary with different stages of development and in disease. DNMT3A has two main isoforms in human cells: a full length

isoform DNMT3A1 that has an extended N-terminal region and is typically expressed in somatic cells, and a shorter isoform DNMT3A2, whose expression is typically restricted to early developmental cell types (Chen et al. 2002). DNMT3B has multiple isoforms identified in various cell types (Duymich et al. 2016; Ostler et al. 2007). The main isoform expressed in developmental cells is DNMT3B1. Levels of DNMT3B1 decrease sharply during differentiation, as cells begin to express an alternative, inactive isoform of the protein (Gifford et al. 2013; Gordon, Hartono, and Chedin 2013).

Gene	DNA Specificity	Major Function	Mouse knock-out phenotype	Human ESC knockout phenotype
DNMT1	Hemimethylated DNA	Maintenance methyltransferase	Genome-wide loss of DNA methylation, embryonic lethality at embryonic day 9.5 (E9.5), abnormal imprinted gene expression, activation of silent retrotransposons (Li, Bestor, and Jaenisch 1992; Howell et al. 2001)	Immediate lethality. Homozygous mutants rescued with ectopic DNMT1 show that acute DNMT1 loss results in rapid global demethylation in 1 week (Liao et al. 2015)
DNMT3A	Equal preference for hemimethylated and unmethylated DNA	<i>De novo</i> DNA methyltransferase	Postnatal lethality at 4–8 weeks, failure to establish methylation imprints in both male and female germ cells (Okano 1999)	Viable. Gradual loss of methylation at a subset of CpGs, retain up to 65% of original global methylation levels (Liao et al. 2015)
DNMT3B	Equal preference for hemimethylated and unmethylated DNA	<i>De novo</i> DNA methyltransferase	Demethylation of minor satellite DNA, embryonic lethality around E14.5 days (Okano 1999)	Viable. Gradual loss of methylation at a subset of CpGs, retain up to 65% of original global methylation levels (Liao et al. 2015)
DNMT3L	N/A	Cofactor required for <i>de novo</i> methyltransferase activity in ESCs	Lack of appropriate methylation of maternal allele and lack of sperm in homozygous males (Webster et al. 2005).	Not Reported

**Table 1.3 Specificities, roles and knockout phenotypes of DNMTs.** This table summarises the properties of DNMT proteins deduced from a number of knockout studies in mouse and human cells.

### 1.3.2 Cross-talk between DNA methylation and histone modifications

To carry out DNA methylation, the DNMTs must be targeted to the correct regions of the DNA. Mammalian DNA methyltransferases themselves have little or no innate sequence specificity beyond the CpG dinucleotide (Goll and Bestor 2005), although it has been shown that both DNMT3A and DNMT3B are sensitive to the sequences flanking their target CpG sites (Handa and Jeltsch 2005; Lin et al. 2002). Additionally, mechanisms must be in place to prevent CGIs that do not require methylation from becoming *de novo* methylated. Deposition of DNA methylation within the genome is generally not considered to occur randomly, and several studies have highlighted a relationship between DNA methylation and modifications to histone tails. This relationship may work in both directions, with histone methylation directing DNA methylation patterns in certain contexts, and DNA methylation guiding histone modifications in other contexts such as upon DNA replication. Methylation of histone H3, particularly at lysine 4 (H3K4), 9 (H3K9), 27 (H3K27) and 36 (H3K36) is particularly involved in the interplay with DNA methylation, and histone acetylation negatively correlates with DNA methylation (Cedar and Bergman 2009). There is evidence to suggest that some of this cross-talk may be brought about through direct interactions between the proteins responsible for carrying out DNA methylation and histone lysine methylation, rather than through the modifications themselves (Cedar and Bergman 2009).

DNMT1 is recruited to the DNA by ubiquitin-like with PHD and RING finger domain-containing protein 1 (UHRF1), a member of the Ring finger-type E3-ubiquitin ligase family, which also regulates the degradation and stability of DNMT1 (Sharif et al. 2007; Bostick et al. 2007; Qin, Leonhardt, and Spada 2011) (Figure 1.7a). UHRF1 associates directly with dimethylated or trimethylated H3K9 (H3K9me<sub>2/3</sub>), present in regions of heterochromatin, as well as binding to hemimethylated CpGs during the S phase of DNA replication (Liu et al. 2013; Rothbart et al. 2012). It thereby brings DNMT1 into proximity of the hemimethylated DNA, enabling DNMT1 to copy methylation onto the daughter DNA strand. This is further assisted by the interaction of DNMT1 with proliferating cell nuclear antigen (PCNA) through its N-terminal PCNA binding domain (Figure 1.7a). PCNA is an auxiliary factor involved in DNA replication and repair, so this interaction enhances methylation maintenance at DNA replication forks (Chuang et al. 1997). In vivo, studies have also shown that the DNMT1-UHRF1 complex interacts with ubiquitin carboxyl-terminal hydrolase 7 (USP7), which stimulates the activity of DNMT1 and regulates the stability of UHRF1 (Felle et al. 2011). Other studies have shown that the H3K9me<sub>2</sub>-specific methyltransferase G9a/GLP interacts directly with DNMT1 and may be required for DNA methylation at some loci (Esteve et al. 2006).

The lysine methyltransferase G9a/GLP has been reported to interact with DNMT3A, mediated by the M-phase phosphoprotein 8 (MPP8) (Chang et al. 2011). The H3K9 methyltransferase SETDB1 has also been shown to interact with DNMT3A, and SUV39H1 has been reported to interact with both DNMT3A and DNMT3B (Li et al. 2006; Fuks, Hurd, Deplus, et al. 2003). Collectively, this points to coordination of H3K9me2 and DNA methylation in regions of heterochromatin (Figure 1.7b). Once DNA methylation has been established in regions of heterochromatin, MBDs such as MECP2 can bind to methylated DNA and recruit histone deacetylases (HDACs) (Jones et al. 1998) to repress transcription, as well as interacting with H3K9 methyltransferases to further maintain the repressive chromatin state (Fuks, Hurd, Wolf, et al. 2003).

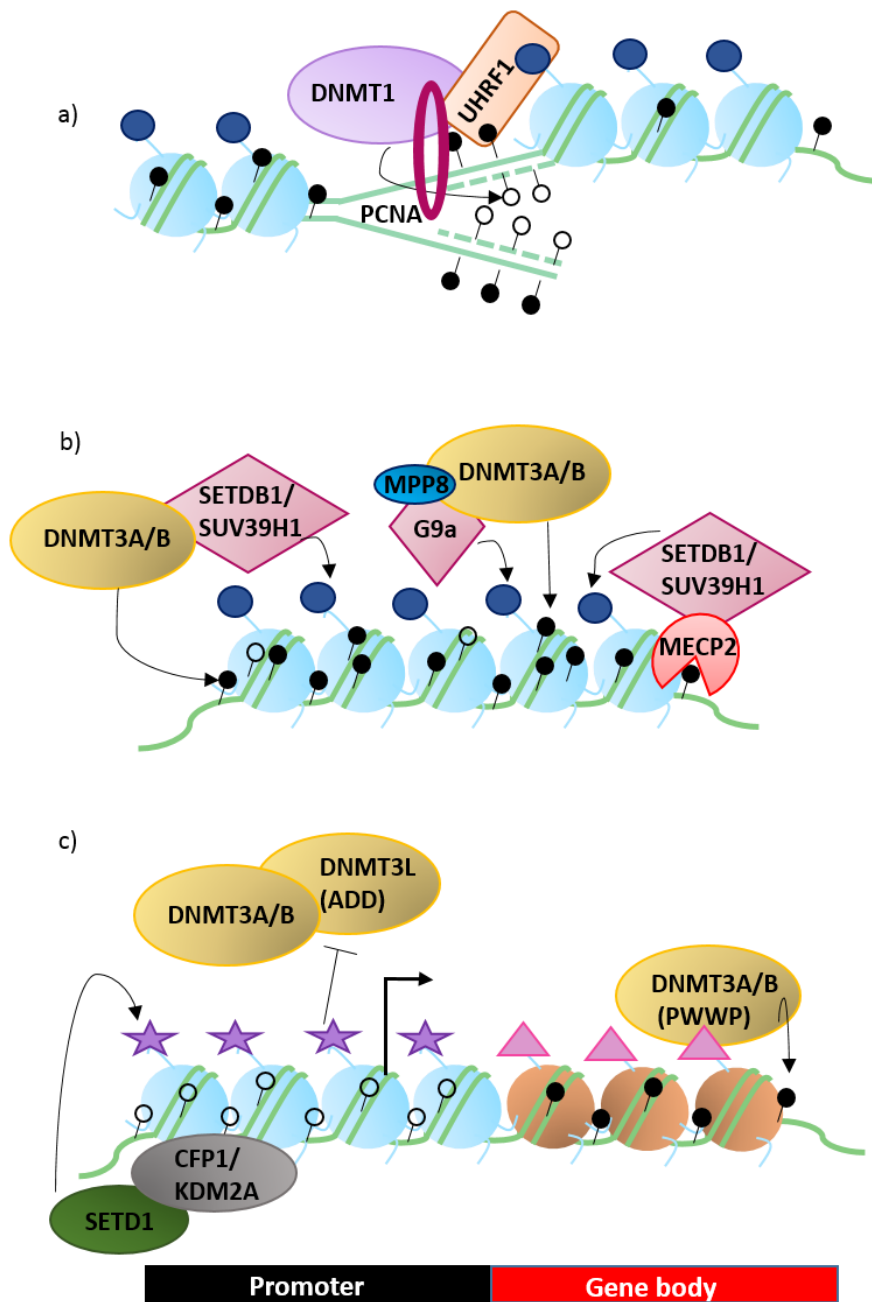
Trimethylated H3K27 (H3K27me3), like DNA methylation is associated with repressed DNA. The relationship between the two marks, however, is complex, as one or the other can precede (Mohn et al. 2008), indicating that DNA methylation can act as either a leader or a sealant in the silencing of chromatin (Ficz 2015). During development, however, H3K27me3 and DNA methylation are considered to be mutually exclusive at CpG rich regions (Brinkman et al. 2012). However, many of these regions undergo silencing and gain DNA methylation upon differentiation, suggesting a cell context dependent cross-talk between H3K27me3 and DNA methylation (Mohn et al. 2008; Meissner et al. 2008). It has been demonstrated that the polycomb repressive complex 2 (PRC2), which includes the H3K27me3 methyltransferase Enhancer of Zeste homolog 2 (EZH2), can directly interact with the DNMTs to direct DNA methylation (Vire et al. 2006). However, recent studies suggest that while PRC2 can recruit DNMT3A to the DNA, it is not sufficient to trigger *de novo* DNA methylation, indicating that additional factors are required (Rush et al. 2014). Additionally, deletion of PRC2 components does not lead to changes in DNA methylation (Hagarman et al. 2013; Boulard, Edwards, and Bestor 2015).

CGIs enriched with trimethylated H3K4 (H3K4me3), a mark of active gene promoters (Guenther et al. 2007), are typically protected from DNA methylation. This is supported by the finding of a strong anti-correlation between DNA methylation and H3K4me3 in many cellular contexts (Weber et al. 2007). The proteins Cfp1 and KDM2A are both able to bind specifically to unmethylated CpGs via their CXXC domains (Blackledge et al. 2010), and subsequently recruit the H3K4 methyltransferase SETD1 to the chromatin (Thomson et al. 2010) (Figure 1.7c). Additionally, it has also been shown that ten-eleven translocase 1 (TET1) can contribute to the protection of CGIs from DNA methylation (Williams et al. 2011; Wu et al. 2011; Verma et al. 2018), as can the lysine demethylase FBXL10 (also known as KDM2B or JHDM1B) (Boulard,

Edwards, and Bestor 2015). In certain developmental contexts, DNMT3A and DNMT3B function in complex with DNMT3L, which facilitates their enzymatic activity (Ooi et al. 2007) (Figure 1.7c). The ADD domain of DNMT3L selectively recognises unmethylated H3K4, and enhances access of either DNMT3A or DNMT3B to CGIs that require methylation (Ooi et al. 2007). The ADD domain of DNMT3A and DNMT3B is also able to directly interact with unmethylated H3K4, resulting in DNA methylation of chromatin with this modification state (Zhang et al. 2010).

Trimethylation of H3K36 (H3K36me3) is associated with the gene bodies of genes that are actively transcribed, and correlates with an enrichment of DNA methylation (Wagner and Carpenter 2012). The PWWP domains of DNMT3A and DNMT3B preferentially bind to H3K36me3, hence directing DNA methylation to gene bodies (Dhayalan et al. 2010; Baubec et al. 2015), as well as to pericentromeric heterochromatin (Figure 1.7c) (Chen, Tsujimoto, and Li 2004; Ge et al. 2004).

While the above mechanisms give us a general overview of how the DNMTs gain access to the DNA during early development, they do not shed light on how the enzymes are targeted more specifically both spatially and temporally to give rise to cell-specific epigenetic patterns. Various ideas have been reported, such as the interaction of DNMTs with transcription factors (Hervouet, Vallette, and Cartron 2010; Hervouet, Vallette, and Cartron 2014), the local influence of DNA-binding factors (Stadler et al. 2011), and the influence of micro RNAs (Denis, Ndlovu, and Fuks 2011), but a unified targeting mechanism applicable to both development and disease is yet to be described.



**Figure 1.7. Cross-talk between DNA methylation and histone modifications.** a) DNMT1 is recruited to replicating DNA by UHRF1, which recognises both H3K9me2/3 and hemimethylated DNA and interacts with PCNA. b) At regions of heterochromatin, DNMT3A and DNMT3B interact with the H3K9 methyltransferases SETDB1, SUV39H1 and G9a to coordinate DNA methylation and H3K9me3. MECP2 binds to methylated DNA and recruits more H3K9 methyltransferases to maintain the repressive state. c) At gene promoters marked by H3K4me3, the binding between the ADD domain of DNMT3s and unmethylated H3K4 is abrogated. CFP1 and KDM2A bind unmethylated CpGs through their CXXC domains and recruit SETD1 to methylated H3K4. Gene bodies are marked by H3K36me3, which is recognised by the PWWP domains of DNMT3A and DNMT3B to direct DNA methylation.

### 1.3.3 Epigenetic mechanisms during development

The global methylation of CpGs across the genome, with unmethylated CGIs results in a bimodal distribution of DNA methylation. During early mammalian development, cells undergo two waves of global erasure of DNA methylation. The first wave occurs in primordial germ cells and the second wave occurs in the early blastocyst upon fertilisation (Figure 1.2) (Seisenberger et al. 2012; Smith et al. 2012; Smith et al. 2014; Guo et al. 2014). DNA methylation patterns are subsequently re-established following implantation of the embryo by a wave of *de novo* methylation by the *de novo* methyltransferases DNMT3A and DNMT3B, in complex with DNMT3L (Okano 1999). It is possible that H3K4me3 patterns are established first and protect CGIs from this wave of re-methylation, allowing them to remain unmethylated (Ooi et al. 2007). Additionally cis-acting DNA-binding sequences play a role in protecting CGIs from *de novo* methylation, as has been demonstrated for specificity protein 1 (Sp1) elements (Brandeis et al. 1994; Macleod et al. 1994). The wave of *de novo* methylation re-instates the bimodal distribution of DNA methylation in the developing embryo.

Later in development, the genome undergoes further alterations to DNA methylation patterns. In particular, upon the formation of the germ layers during gastrulation, genes responsible for the maintenance of pluripotency are no longer protected from DNA methylation and instead are targeted for *de novo* methylation and repression (Gidekel and Bergman 2002). Unlike the global wave of *de novo* methylation upon implantation, *de novo* methylation of the pluripotency genes requires specific targeting of the DNMTs to pluripotency gene promoters. A multi-step mechanism has been outlined for the pluripotency gene Oct3/4, and is relevant for the control of other pluripotency genes (Epsztejn-Litman et al. 2008). The direct interaction of repressor molecules with the gene promoter can silence transcription, and this is followed by recruitment of the H3K9 methyltransferase G9a and HDACs by transcription factors. Following deacetylation of lysine residues and methylation of H3K9, the heterochromatin protein 1 (HP1) binds and facilitates formation of heterochromatin, and DNMT3A and DNMT3B are recruited as part of the G9a complex to carry out *de novo* methylation of the Oct3/4 promoter for stable gene repression (Fuhrmann et al. 2001; Feldman et al. 2006; Epsztejn-Litman et al. 2008).



## 1.4 Mechanisms of DNA demethylation

### 1.4.1 Active demethylation

Although patterns of DNA methylation are relatively stable in terminally differentiated cells, the genome can lose DNA methylation by both active and passive mechanisms of DNA methylation. These processes are at play during developmental cellular reprogramming (Wu and Zhang 2010). Unlike the machinery involved in regulating histone post-translational modifications, a dedicated DNA demethylase has not been identified. Instead the mechanisms that regulate removal of 5mC involve further modification to the base, or passive dilution of the base across cell divisions.

Like DNA methylation, the targeting of the active mechanisms of demethylation can be global or loci-specific. There are three members of the ten-eleven translocation methylcytosine dioxygenase (TET) family of proteins (Figure 1.8), each with cell-specific expression patterns and specific, non-overlapping target genes or genomic regions (Wu and Zhang 2011).

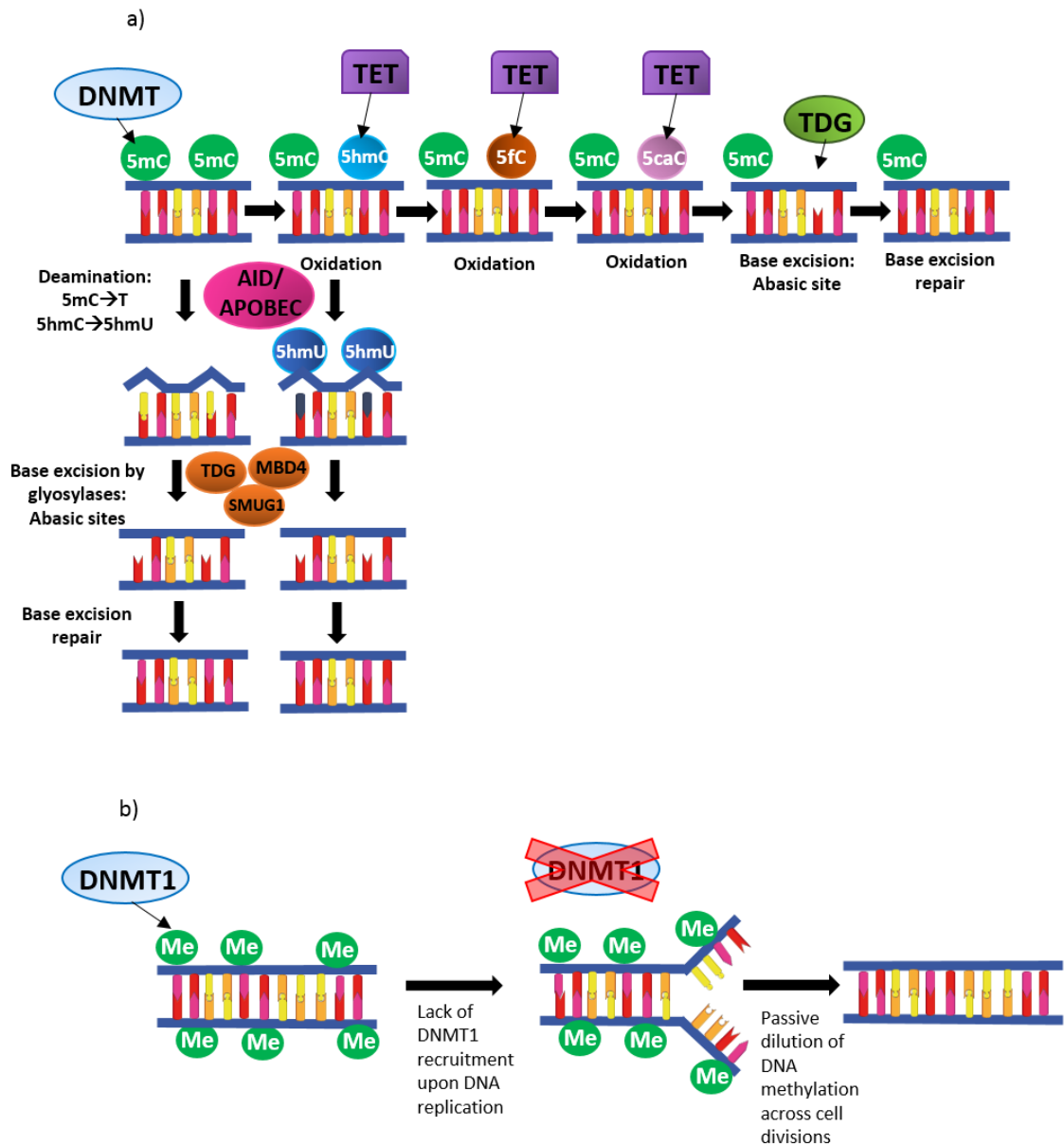


**Figure 1.8. The ten-eleven translocation methyltransferase dioxygenase (TET) family.** There are three members of the TET family. All three TETs have a double-stranded helix fold (DSBH) and a cysteine rich domain, and TET1 and TET3 have an additional CXXC domain which TET2 lacks.  
Figure taken from: Meng *et al.* (Meng *et al.* 2015)

TET enzymes can oxidise 5mC to 5-hydroxymethylcytosine (5hmC), which can subsequently be further oxidised to 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) (Tahiliani *et al.* 2009; Kriaucionis and Heintz 2009; Ito *et al.* 2011), as shown in Figure 1.9a. Once converted, these marks are not recognised by DNMT1 during DNA replication (Valinluck and Sowers 2007), leading to loss of DNA methylation as cells divide. Additionally, the 5mC oxidation products may however be cleaved by the thymine-DNA glycosylase (TDG) and activate the base excision repair (BER) pathway (Maiti and Drohat 2011) (Figure 1.9a). Alternatively, 5mC or 5hmC can be deaminated by the activation induced cytidine deaminase (AID/AICDA) or related APOBEC proteins to form thymine and 5-hydroxymethyluracil (5hmU), both of which can subsequently

also undergo excision by TDG and other glycosylases, and activate the BER pathway (Figure 1.9a) (Bhutani et al. 2010; Guo et al. 2011; Popp et al. 2010). Though the role of AID in active demethylation is controversial, an *in vivo* study in mice reported that AID-deficient primordial germ cells were up to three times more methylated than their wildtype counterparts (Popp et al. 2010), suggesting that AID may play a role in active DNA demethylation through deamination. Additionally, overexpression of AID in B-cell lymphomas leads to increased epigenetic heterogeneity driven by AID-driven demethylation rather than mutation (Teater et al. 2018).

Although 5hmC, 5fC and 5caC are intermediates in active demethylation, each of the modified bases can be identified in tissues, implying that they are stable bases and may have functions beyond acting as intermediates in the process of demethylation (Bachman et al. 2014; Bachman et al. 2015). Levels of 5hmC vary across tissue types and are most abundant in brain tissue and in ESCs, where they can be found at levels approximately 10-fold lower than 5mC (Nestor et al. 2012; Munzel, Globisch, and Carell 2011). Both 5fC and 5caC can be detected at levels approximately 10-100 fold lower than that of 5hmC (Bachman et al. 2015; Pfaffeneder et al. 2011). In ESCs, 5hmC is prevalent in euchromatin, particularly at transcriptional start sites, promoters, exons, gene bodies and active enhancers, and is typically found at genomic regions with intermediate CpG density (Shen and Zhang 2013; Ficz et al. 2011). Additionally, 5hmC is primarily enriched at developmental genes marked by bivalent histone modifications or polycomb target genes (Pastor et al. 2011). This correlates with the genome-wide pattern of TET1 enrichment which is proposed to protect bivalent CGIs from *de novo* DNA methylation (Williams et al. 2011; Wu et al. 2011). 5fC exhibits a similar pattern of enrichment to 5hmC in ESCs, but its enrichment at gene promoters correlates with active gene expression (Neri et al. 2015). Evidence from recent years indicates that each of these modified bases may exert their own regulatory functions on the DNA through binding to or recruitment of other factors (Song and He 2013). 5hmC is known to inhibit the binding of MBDs and has also been shown to recruit chromatin remodelling complexes to 5hmC-marked genes (Yildirim et al. 2011; Valinluck et al. 2004), and 5fC has also been shown to interact with transcription and chromatin regulating factors (Iurlaro et al. 2013; Spruijt et al. 2013).



**Figure 1.9. Mechanisms of DNA demethylation.** a) DNA demethylation can occur actively through oxidation of 5-methylcytosine (5mC) by the ten-eleven translocation methylcytosine dioxygenase (TET) enzymes to 5-hydroxymethylcytosine (5hmC), and further on to 5-formylcytosine and 5-carboxylcytosine (5caC). The modified bases may be removed by thymine-DNA glycosylase (TDG), resulting in an abasic site and activation of the base excision repair (BER) pathway and replacement of an unmodified cytosine base. Alternatively, deamination of 5mC of 5hmC by activation induced cytidine deaminase (AID) and related proteins results in a thymine or 5-hydroxyuridine (5hmU) base that causes a mismatch and activates the base excision repair pathway. b) Demethylation can also occur passively through dilution of DNA methylation upon DNA replication across cell divisions in the absence of faithful methylation maintenance by DNMT1.

### 1.4.2 Passive demethylation

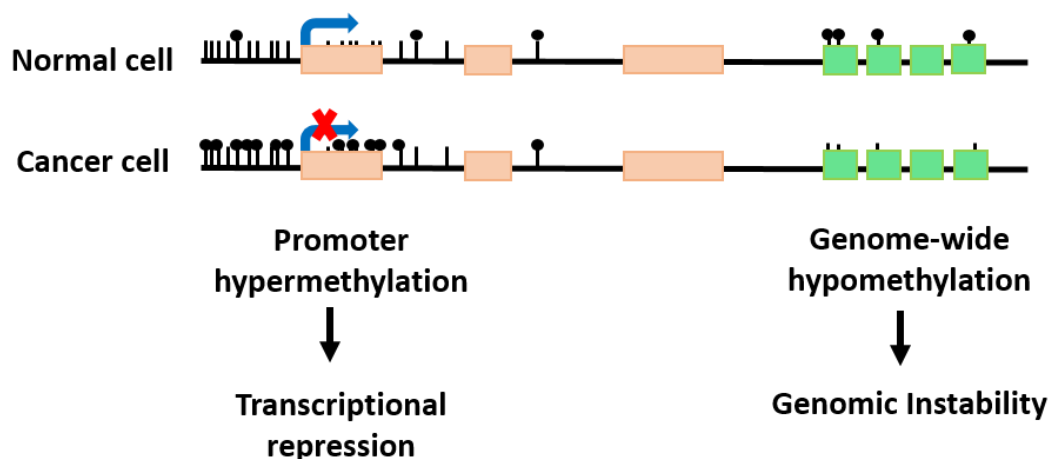
As well as active removal of 5mC, DNA demethylation can occur passively by dilution of methylation over cell divisions if it is not faithfully maintained by the maintenance

methyltransferase DNMT1 upon DNA replication (Wu and Zhang 2010) (Figure 1.9b). This could be as a result of absence or downregulation of the protein, or inactivity of its enzymatic function. In mESCs grown in 2i conditions, impairment of DNA methylation maintenance as a result of both downregulation of UHRF1 and globally reduced H3K9me2 has been identified as the primary mechanism of global demethylation (von Meyenn et al. 2016). Unlike active demethylation mechanisms, which are likely targeted to specific regions of the DNA in response to cellular stimuli, passive demethylation is likely more unspecific and would result in genome-wide loss of methylation.

## 1.5 Epigenetic changes in cancer

### 1.5.1 Aberrant DNA methylation patterns in cancer

DNA methylation patterns are commonly altered in a number of human diseases. Disruption of canonical DNA methylation patterns is a hallmark of human cancers, typically characterised by loss of global genomic DNA methylation accompanied by site-specific hypermethylation (Figure 1.10) (Jones and Baylin 2007; Baylin and Jones 2011; Esteller et al. 2001; Feinberg, Ohlsson, and Henikoff 2006; Ehrlich 2002). DNA hypomethylation is typically associated with genomic instability (Eden et al. 2003; Gaudet et al. 2003), while site-specific DNA hypermethylation occurs at promoter CpG islands and is often associated with repression of tumour suppressor genes in cancer cells (Jones and Baylin 2002). Whether DNA methylation changes are a cause or consequence of tumourigenesis remains a recurrent controversial question in the field, but emerging research is beginning to answer this question.



**Figure 1.10. DNA methylation patterns in cancer.** Typically, cancer cells exhibit genome-wide hypomethylation accompanied by hypermethylation of specific gene promoters compared to their normal counterparts. Genome-wide hypomethylation is associated with genomic instability, while promoter hypermethylation has been associated with transcriptional repression. Each black lollipop represents a methylated cytosine, while the black lines represent unmethylated cytosines.

### 1.5.2 DNA hypermethylation in cancer

Early experiments based on the study of individual genes indicated that it was primarily tumour-suppressor genes that underwent aberrant DNA hypermethylation in cancer (Jones and Baylin 2002). Multiple tumour suppressor genes, such as CDKN2A and BRCA1, that are

commonly mutated in cancers were found to be hypermethylated in various cancer types (Dobrovic and Simpfendorfer 1997; Merlo et al. 1995). Moreover, in some cases, such as with MLH1, hypermethylation was detected in the absence of any mutations to the gene, resulting in loss of function analogous to MLH1 gene mutations and a driving role in tumorigenesis (Kane et al. 1997; Herman et al. 1998). Hypermethylation of single tumour suppressor genes was thought to be propagated in cancer cells as a result of cellular selection due to the inhibition of tumour suppressor activity, as these genes are typically involved in pathways such as DNA repair and genomic stability, cell cycle and cell division, and metabolism (Llinas-Arias and Esteller 2017). Experiments carried out with more advanced DNA methylation mapping technologies showed that this phenomenon is more widespread and that a large number of CGIs are hypermethylated in cancer (Costello et al. 2000; Jones and Baylin 2007). While several studies have shown clear repressive roles of hypermethylation for individual tumour suppressor genes (Jones and Baylin 2002), many of the large number of loci hypermethylated in cancers are not tumour suppressor genes. It remains less well understood what the purpose of hypermethylation of a large number of loci might be. Many of the genes that are hypermethylated are shared across cancer types, while a number of them are tumour or tissue-type specific (Costello et al. 2000). Different cancer types display different levels of CGI hypermethylation, with a sub-group of cancers displaying a particularly high degree of hypermethylation at a specific subset of gene promoters, known as the CpG island methylator phenotype (CIMP) (Toyota et al. 1999). Moreover, many of the genes that undergo hypermethylation are already repressed in the normal tissue prior to transformation, and their expression is further attenuated upon hypermethylation (Keshet et al. 2006; Sproul et al. 2012; Easwaran et al. 2012). Developmental genes have been found to be frequently hypermethylated across cancer types, therefore it has been proposed that aberrant hypermethylation in cancer may act to block cellular differentiation and enable cancer cells to continue to propagate in their more primitive states (Widschwendter et al. 2007; Easwaran et al. 2012; Pfeifer 2018). Alternatively, it has been proposed that cancer hypermethylation may act to block cancer progression, by restricting the stochastic activation of genes that may enable adaptation of the cancers to environmental stimuli (Sproul and Meehan 2013).

There is some evidence that aberrant *de novo* methylation occurs early in tumorigenesis, as DNA hypermethylation can be detected in premalignant lesions (Chan et al. 2002; Hanley et al. 2017). Many of the genes that are hypermethylated in cancer are also hypermethylated during normal human ageing, which is perhaps unsurprising as cancer is a disease strongly associated with age, and this suggests that the two processes are driven by common epigenetic

mechanisms or that ageing predisposes to neoplasia (Rakyan et al. 2010; Teschendorff et al. 2010). Whether such DNA hypermethylation changes are a cause or consequence of tumorigenesis remains a recurrent controversial question in the field, but some recent studies have shed light on this. One recent study showed that targeting DNA hypermethylation to the CDKN2A gene promoter in normal cells initiates aberrant cellular processes (Saunders et al. 2017), and a second study demonstrated that ageing and cancer associated DNA hypermethylation accelerates cellular transformation in a *Braf*<sup>V600E</sup> mouse colon organoid system, through suppression of Wnt signalling regulators in a progressive manner (Tao et al. 2019). This study functionally links promoter CGI hypermethylation with oncogenic transformation, demonstrating a causal relationship.

The majority of the observations of aberrant DNA methylation have been made in cancer cell lines or primary cancer cells, but they are not fully representative of the processes occurring during the transition of normal cells into cancer cells. The underlying mechanisms that give rise to these opposing patterns of genomic DNA methylation in early stages of human cancer development remain elusive, as does the timing of such events in cancer initiation and progression. More specifically, while it has been postulated that recruitment of *de novo* DNMTs may be assisted by a chromatin- or DNA-binding factor, how *de novo* DNA methyltransferase activity is preferentially targeted to specific regions of the genome in the context of aberrant cancer methylation remains largely a mystery.

It has been proposed that *de novo* methylation of DNA in cancer operates through an instructive mechanism driven by cis-acting DNA sequences and trans-acting protein complexes that are capable of recruiting DNMTs (Keshet et al. 2006). In line with this, various studies have used computational models to show that sites prone to *de novo* DNA methylation in cancer can be predicted based on nearby DNA sequence motifs, and those that are protected from methylation are enriched for alternative motifs, such as zinc-finger protein binding sites (Feltus et al. 2003, 2006; Fan et al. 2007; Kim et al. 2008; McCabe, Lee, and Vertino 2009). However, there is limited consensus between the sequence motifs identified as being predictive for DNA methylation by each individual computational approach. A more recent hypothesis suggested that CGIs that become hypermethylated in cancer are typically associated with developmental genes and are regulated by tissue-specific transcription factors which are not ubiquitously expressed. CGIs that are resistant to *de novo* methylation in cancer typically reside within genes involved in basic cellular processes and are bound by combinations of ubiquitously expressed transcription factors which contribute to protection from DNA methylation (Gebhard et al. 2010).

A number of independent studies have observed the preferential susceptibility of loci marked by H3K27me3 or bound by polycomb to gain DNA methylation in cancer (Ohm et al. 2007; Schlesinger et al. 2007; Widschwendter et al. 2007; Nejman et al. 2014). Similarly, other studies have reported that hypermethylation in cancer occurs at the promoters of developmental genes and transcription factors which are marked by bivalent histone modifications in hESCs (Bernhart et al. 2016; Easwaran et al. 2012; Court and Arnaud 2017). Other studies have demonstrated site-specific recruitment of DNMTs to loci that undergo hypermethylation by DNA-binding factors, but thus far this has only been shown for individual genes or loci (Suzuki et al. 2006; Fuks et al. 2001; Di Croce et al. 2002; Brenner et al. 2005).

The process of inflammation, often as a result of bacterial or viral infection has also been linked to *de novo* methylation and carcinogenesis. It is thought that this association may arise either as a result of increased proliferation of inflammatory tissues or by means of epigenetic reprogramming induced by the infection (Niwa et al. 2010; Su et al. 2007; Ferrari et al. 2008). In line with this, associations have been made between DNA methylation and the processes of DNA damage and repair, which are known to play a role in carcinogenesis. Various studies have demonstrated *de novo* DNA methylation of CGIs at sites of DNA double-strand break repair or sites of oxidative damage. In each case, the DNA repair proteins responsible for repairing the damaged DNA recruit DNMTs to the DNA, thus linking DNA repair to *de novo* DNA methylation (Xia et al. 2017; O'Hagan, Mohammad, and Baylin 2008; Ding et al. 2015; Cuozzo et al. 2007).

### 1.5.3 Global hypomethylation and genomic instability

Cancer cells commonly exhibit reduced global levels of DNA methylation compared to normal somatic cells (Ehrlich 2002, 2009). More recently, it has also become evident that cancer cells also display reduced global levels of 5hmC (Ficz and Gribben 2014). Hypomethylation is likely an early event in cancer as it has been detected in pre-malignant tissue (Desrosiers et al. 1985). Global methylation in cancer is typically disrupted at repetitive DNA elements, and these can become overexpressed as a result of loss of silencing mechanisms (Ting et al. 2011). The presence of DNA methylation across the genome is beneficial in maintaining genomic and chromosomal stability (Chen et al. 1998). Global hypomethylation in cancer is often associated with genomic instability including chromosomal instability (Ehrlich 2009; Eden et al. 2003; Karpf and Matsui 2005; Chen et al. 1998; Sheaffer, Elliott, and Kaestner 2016).



Whether DNA hypomethylation is a cause or simply a consequence of carcinogenesis remains a question under investigation. Several studies have reported chromosomal breakages, fusions, aneuploidy, and re-activation of silenced DNA elements, in relation to DNA hypomethylation and downregulation or mutations of the DNMTs, and such instability may promote tumour initiation or progression (Eden et al. 2003; Gaudet et al. 2003; Sheaffer, Elliott, and Kaestner 2016; Dodge et al. 2005; Karpf and Matsui 2005). The absence of DNA methylation at certain genomic regions such as heterochromatic repetitive elements may lead to incorrect chromosome segregation that gives rise to chromosomal aberrations (Monier, Mouradian, and Sullivan 2006). Mice with a hypomorphic mutation of DNMT1 display greatly reduced levels of DNA methylation and develop aggressive T-cell lymphomas with a high frequency of chromosome 15 trisomy, suggesting that DNA methylation contributes to tumorigenesis by causing chromosomal instability that results in aneuploidy (Gaudet et al. 2003; Karpf and Matsui 2005). Inactivation of DNMT3B or mutations in DNMT3B which cause the immunodeficiency, centromere instability and facial anomalies (ICF) syndrome both result in DNA hypomethylation and chromosomal aberrations and aneuploidy (Ehrlich 2003; Dodge et al. 2005). Additionally, as DNA methylation is involved in the silencing of retrotransposons during development (Bestor 2004), hypomethylation may result in increased retrotransposition and re-expression of such elements, which may play an active role in tumorigenesis (Daskalos et al. 2009).

As well as global disruption of methylation at repetitive DNA sequences, hypomethylation can also result in the derepression of imprinted genes (Li, Beard, and Jaenisch 1993). Loss of imprinting is a commonly observed phenomenon across cancer types and is also thought to occur early on in tumorigenesis, with a putative driving role in predisposing cells to cancer development (Jelinic and Shaw 2007; Holm et al. 2005).

#### 1.5.4 Disruption of epigenetic regulators in cancer

Large-scale sequencing studies have enabled the identification of the landscape of mutations across cancer types. Among these landscapes, several mutations have been identified that affect epigenetic regulators (Plass et al. 2013), suggesting that epigenetic abnormalities may lie downstream of genetic events. However, while aberrant DNA methylation patterns including both genomic hypomethylation and CGI hypermethylation are a hallmark of cancer, mutations to the epigenetic modifiers of DNA methylation, the DNMTs, are less frequent. Nevertheless, mutations in DNMT3A are frequent in haematological malignancies (Yan et al. 2011; Ley et al.

2010), though the resulting hypermethylation is considered to be a consequence of the malignancy rather than a contributor to development of the malignancy (Spencer et al. 2017). Mutations in the TET genes, in particular TET2, are also an early event in haematological malignancies (Lorsbach et al. 2003; Langemeijer et al. 2009). Notably, both these types of mutations are present at considerable frequencies in the blood of healthy, ageing individuals (Desai et al. 2018; Abelson et al. 2018; Xie et al. 2014), and both types of mutations are predominantly found in haematological malignancies rather than solid tumours. Mutations in DNMT1 have been identified in colorectal cancers (Kanai et al. 2003). Aside from mutations to the DNMTs, many cancers exhibit aberrant overexpression of DNMT3B. Many studies have observed correlations between aberrant DNMT3B expression and CGI hypermethylation in cancer, but a direct causal role in establishing aberrant DNA methylation patterns has not been demonstrated (Roll et al. 2008; Linhart et al. 2007).

Though the writers and erasers of DNA methylation themselves may not be frequently mutated in cancer, mutation of readers and writers of histone modifications as well as chromatin remodelers is a common feature of many cancer types (Plass et al. 2013; Baylin and Jones 2011). Given the interplay between DNA methylation and the chromatin, such abnormalities may contribute to the aberrant DNA methylation patterns frequently observed in cancer. Alternatively, it is possible that aberrant DNA methylation patterns arise independently of genetic mutations to the epigenetic modifiers themselves. In line with this, it has been proposed that disruption of epigenetic 'modulators' or 'mediators', which can act either upstream or downstream of the epigenetic modifiers may also impact the epigenome in cancer (Feinberg, Koldobskiy, and Gondor 2016). Finally, given the commonality in aberrant DNA methylation patterns across different cancers that each have different driver mutations, it is also possible that the initial disruption of canonical DNA methylation patterns occurs independently of genetic mutations altogether.

### 1.5.5 DNA methylation and mutation

Cytosine (C) can deaminate spontaneously to give uracil (U), which causes a mismatch with guanine (G). This mismatch is recognised by uracil DNA glycosylases, and the uracil base is efficiently eliminated and replaced with cytosine. 5mC, however, is deaminated to give thymine (T) (Bird 1980). Compared to unmodified C, 5mC has increased susceptibility to spontaneously deaminate (Shen, Rideout, and Jones 1994). Aside from spontaneous deamination, AID is thought to mediate hydrolytic deamination of 5mC into T and C to U. The

U:G mismatch, if replicated over, will give rise to C to T mutations unless the U is excised by the uracil-N-glycosylase enzyme (UNG) (Perez-Duran et al. 2012). The T is repaired if AID activity is coupled to DNA repair, resulting in active demethylation. Unlike U, however, because T is a natural base, the mutant base can persist through DNA replication, and upon cell division, it can be passed on to daughter cells as a C to T transition mutation (Schmutte et al. 1995). Despite the existence of MBD4, a dedicated thymine DNA glycosylase that can selectively remove T from a T:G mismatch, these mismatches are repaired with reduced efficiency (Schmutte et al. 1995; Hendrich et al. 1999).

CpG sites are underrepresented globally across the genome due to the high mutation rate of methylated cytosines (Bird 1980). Moreover, methylated CpGs are considered to be hotspots for disease-related DNA mutations, supported by the fact that C to T mutations are the most frequent in human disease, constituting a third of all point mutations (Cooper and Youssoufian 1988). These mutations may occur in the coding regions of genes (Rideout et al. 1990), or on regulatory regions of DNA, hence altering binding sites of transcription factors and other proteins (Zemojtel et al. 2011). C to T mutations that occur in early-replicating DNA are likely to be repaired by BER machinery, but mutations that occur in late-replicating DNA where BER operates with reduced efficiency are more likely to result in a mutation that is propagated upon cell division (Blokzijl et al. 2016; Tomkova et al. 2018). Interestingly, while 5mC shows increased mutability compared to unmethylated C, 5hmC has in fact been shown to be protective against mutagenesis (Tomkova et al. 2016), and enrichment of 5hmC at sites of DNA damage has in fact been shown to promote genome stability by ensuring chromosome segregation occurs correctly (Kafer et al. 2016).

Some of the most frequently seen mutational signatures in human cancers feature high rates of C to T mutations, and are attributed to the deaminase hyperactivity of AID or APOBEC enzymes at methylated CpG dinucleotides, possibly coupled with inefficient BER pathway activity (Alexandrov et al. 2013). Recent reports suggest that tissue-specific mutations accumulate in normal adult stem cells during life. These patterns are similar between the normal stem cells of a given tissue and cancers that arise from those tissues, suggesting that intrinsic, non-random mutational processes such as that described above may be responsible for initiating tumourigenesis (Blokzijl et al. 2016).

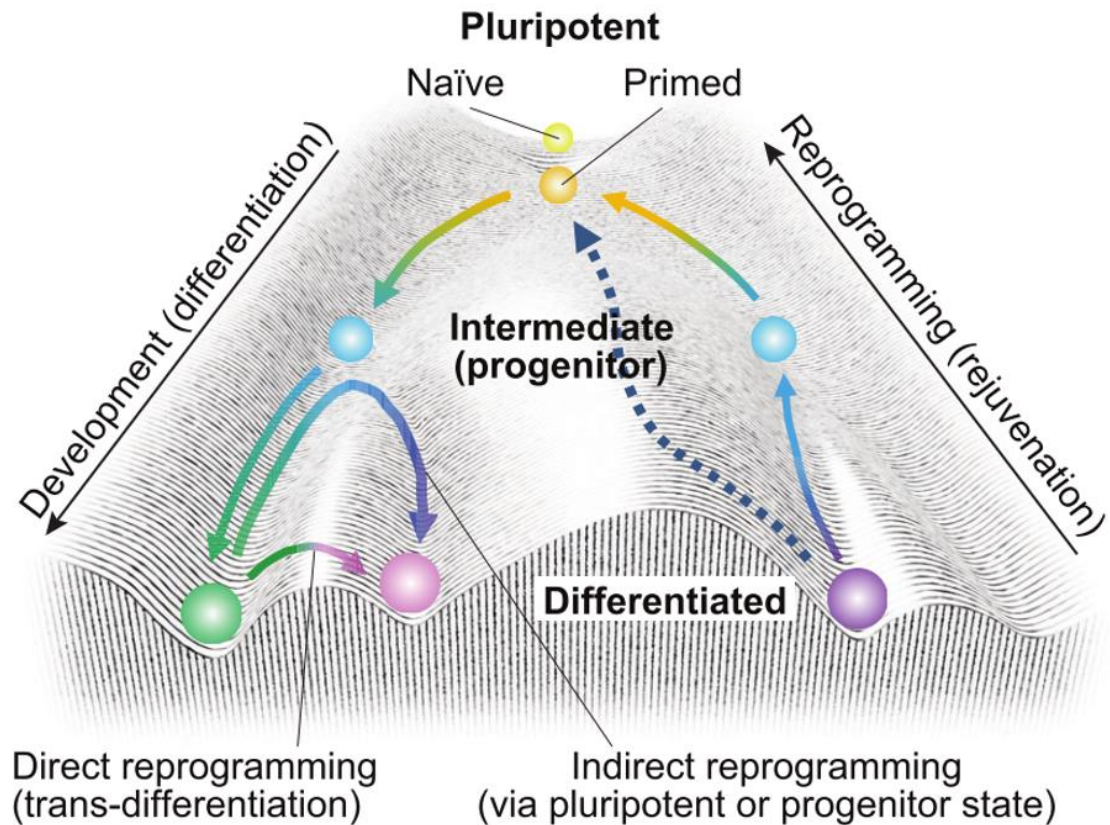
While 5mC is susceptible to DNA mutation, a global loss of methylation has been associated with an overall increased mutation frequency (Chen et al. 1998). This suggests that the interplay between CpG methylation and mutation is two-fold, in that global DNA methylation

maintains genomic stability, but methylated CpG dinucleotides individually have increased susceptibility to mutation compared to unmethylated CpGs.

## 1.6 Epigenetic reprogramming

### 1.6.1 Somatic cell reprogramming

Aside from the two waves of epigenetic reprogramming that take place *in vivo* during early mammalian development, it is now known that molecular reprogramming of human somatic cells can be induced by ectopically overexpressing a combination of four pluripotency transcription factors. The overexpression of OCT4, SOX2, KLF4 and c-MYC (collectively referred to as OSKM) in differentiated cells reprograms the cells, albeit with low efficiency, resulting in the generation of induced pluripotent stem (iPS) cells. (Takahashi and Yamanaka 2006; Okita, Ichisaka, and Yamanaka 2007; Takahashi et al. 2007). While iPS cells and ESCs share characteristics such as their pluripotency and differentiation potency, their molecular profiles are similar, yet distinct, and the molecular similarity between the two cell types remains a topic of controversy as iPS cells often exhibit molecular profiles attributable to somatic memory (Hochedlinger and Jaenisch 2015). Recently, it has been shown that the OSKM pluripotency transcription factors can be replaced with combinations of lineage-specific transcription factors to generate iPS cells, suggesting that cellular reprogramming may be influenced by the equilibrium between transcription factors controlling opposing cellular states rather than the direct specification of a pluripotent state by OSKM (Montserrat et al. 2013). Such reprogramming of mature cells into a pluripotent state challenges the unidirectional depiction of development proposed by Conrad Waddington, as does the further reprogramming of primed hESCs to the naïve state of pluripotency (Figure 1.11). Recently, it has also been demonstrated that naïve human iPS cells can be generated directly from somatic cells directly, without a primed intermediate, through OSKM induction and maintenance in 2iLGö media (Kilens et al. 2018).



**Figure 1.11. Cell state changes and Waddington's landscape.** This figure shows that during development, pluripotent cells (primed or naïve) commit to various somatic cell lineages via intermediate progenitor states. Using direct or indirect reprogramming, differentiated, somatic cells can be converted into alternative lineages or reprogrammed into a pluripotent cell state.  
 Figure from: Takahashi & Yamanaka, 2015 (Takahashi and Yamanaka 2015)

The nuclear reprogramming induced by the OSKM factors resets the epigenetic landscape of somatic cells into that of a pluripotent cell, and this is accompanied by global reprogramming of the transcriptome (Maherali et al. 2007). During reprogramming of somatic cells to iPS cells, certain components of the chromatin landscape are rapidly altered in a loci-specific manner to enable the first wave of transcriptional reprogramming that primarily involves silencing of somatic genes, and this is driven by transcription factors (Koche et al. 2011). Additionally, the early stages of iPS reprogramming involve the transient re-activation of developmental regulators (Cacchiarelli et al. 2015). The activation of pluripotency genes occurs at a later stage, as a final wave of transcriptional reprogramming (Polo et al. 2012). The majority of epigenetic modifications are not altered globally until the later stages of iPS reprogramming, at which stage global chromatin decompaction occurs and bivalency is established at developmental gene promoters (Mikkelsen et al. 2008). Global DNA demethylation occurs particularly late during reprogramming (Polo et al. 2012) and is a bottleneck for efficient

reprogramming, as it may trap cells in a partially reprogrammed state. In line with this, it has been shown that manipulation of DNMT1 levels or use of DNA demethylating agents increases the efficiency of cellular reprogramming. (Mikkelsen et al. 2008). Conversely, DNMT3A and DNMT3B are not required for efficient iPS reprogramming, indicating that *de novo* DNA methylation is dispensable for reprogramming (Pawlak and Jaenisch 2011).

Despite the dispensability of the *de novo* methyltransferases for efficient reprogramming of somatic cells into iPS cells, there have been several reports of aberrant DNA hypermethylation observed during reprogramming (Ruiz et al. 2012; Nishino et al. 2011). A number of these aberrantly hypermethylated sites have been shown to overlap with bivalent chromatin marks, as well as overlapping regions that are hypermethylated in cancer (Doi et al. 2009; Ohm et al. 2010). These observations insinuate that the process of reprogramming gives rise to epigenetic changes that are characteristic of neoplasia. Additionally, it has also been demonstrated that transient expression of OSKM in mice enables reprogramming to occur *in vivo*, resulting in the formation of teratomas in multiple organs (Abad et al. 2013). Moreover, partial or incomplete *in vivo* reprogramming using OSKM can result in a dysplastic phenotype resembling Wilms tumours when reprogramming is prematurely terminated, in the absence of any identifiable cancer-driving mutations (Ohnishi et al. 2014). This also implicates epigenetic reprogramming as a driver of carcinogenesis in the absence of cancer-causing mutations.

### 1.6.2 Cellular reprogramming and the pluripotency network

Of the four transcription factors that constitute the OSKM combination, OCT4, SOX2 and NANOG are considered to be part of the 'core' pluripotency network (Boyer et al. 2005; Wang et al. 2006). The core pluripotency network is known to interact with histone and chromatin modifying complexes, though the extent of the interactions and whether they are direct or indirect is less well characterised. OCT4, SOX2 and NANOG all associate with components of the Brg/Brahman-associated factors (BAF) complex, which is responsible for remodelling nucleosomes, and Oct4 and NANOG additionally associate with the nucleosome remodelling deacetylase (NuRD) complex (Ho et al. 2009; Orkin and Hochedlinger 2011). The core pluripotency factors also associate with HDAC complexes as well as chromatin scaffold proteins such as transcription intermediary factor-1b (TIF1b) (Seki et al. 2010; Wang et al. 2006). Moreover, the gene expression control of several histone modifying enzymes is under the control of pluripotency transcription factors. The H3K9 demethylases JMJD1A (KDM3A)

and JMJD1B (KDM3B), both of which are part of the JHDM2 family, are directly controlled by OCT4 (Loh et al. 2007).

### 1.6.3 Reprogramming and genomic stability

Although stem cells are constantly exposed to endogenous sources of DNA damage as a result of their rapid proliferation, stem cells are considered to be very sensitive to DNA damage, in that they activate robust DNA repair mechanisms against even low levels of DNA damage in order to maintain their genomic integrity (Fu et al. 2017; Seifert, Dejosez, and Zwaka 2017). However, the process of cellular reprogramming and generating iPS cells has been associated with genomic instability. Human iPS cells cultured for multiple passages display increased double-strand breaks and chromosomal aberrations and reduced DNA repair efficiency (Simara et al. 2017; Mayshar et al. 2010). Additionally, various studies have reported increased numbers of somatic mutations and copy number variation upon iPS reprogramming (Gore et al. 2011; Hussein et al. 2011; Ji et al. 2012). The OSKM reprogramming factors are thought to induce replication stress which ultimately leads to increased DNA damage and reduced genomic stability, and lowering the replication stress through either genetic or chemical manipulation during reprogramming can reduce the levels of genomic instability (Ruiz et al. 2015).

Moreover, naïve hESC lines generated through various methods of primed hESC reprogramming, as well as naïve hESCs generated directly through iPS reprogramming show an aberrant karyotype following multiple passages (Kilens et al. 2018; Theunissen et al. 2014; Pastor et al. 2016; Guo et al. 2017). It has been shown that this loss of genomic stability is attributable to the MEK1/2 inhibitor used as part of the 2i combination (Choi et al. 2017), and that reduction of MEK inhibition can reduce the genomic instability observed in naïve hESCs generated using the 5iLAF method of reprogramming (Di Stefano et al. 2018).

### 1.6.4 Reprogramming and cancer

The process of oncogenic transformation, whereby a normal becomes neoplastic, bears resemblance to the process of reprogramming, as both processes involve the acquisition of an altered developmental state and transcriptional programme along with enhanced self-renewal potential, with or without any changes to the DNA sequence (Suvà, Riggi, and Bernstein 2013; Ben-Porath et al. 2008). In particular, the loss of cellular identity, either through

dedifferentiation or transdifferentiation is considered to be an early event across many cancer types (Roy and Hebrok 2015). During both cellular reprogramming and malignant transformation, it has been observed that there is an inverse correlation between the differentiation state of the cell of origin and its reprogramming efficiency or transformation potential, with stem and progenitor cells being more receptive to pluripotent reprogramming and oncogene-induced transformation (Eminli et al. 2009; Silva et al. 2008; Morel et al. 2017; Cobaleda et al. 2000).

Each of the four factors that constitute the OSKM combination used in iPS reprogramming are also implicated in carcinogenesis. OCT3/4 has been shown to play a role in the generation of germ cell tumours (Gidekel et al. 2003), and SOX2 has been implicated as an oncogene in lung and oesophageal squamous cell carcinomas and small-cell lung carcinomas, as well as helping to support the cancer stem cell populations in Ewing sarcoma and a number of other cancer types (Bass et al. 2009; Rudin et al. 2012; Riggi et al. 2010). KLF4 has been shown to promote the development of various solid tumours such as breast and skin cancer (Rowland and Peeper 2006), and c-MYC is a well-known oncogene in a wide range of human cancers, functioning through its influence on gene expression programmes (Dang 2012; Lin et al. 2012).

Furthermore, NANOG, which is able to substitute some of the original OSKM factors and is used in alternative combinations of reprogramming factors (Jung et al. 2014) is frequently detected in the stem-like compartments of multiple human malignancies (Jeter et al. 2015). Moreover, it was recently shown that the induction of pluripotency regulators in cancers promotes their survival and resistance to therapy, conferring poor clinical outcomes (Hepburn et al. 2019). Finally, the finding that transient expression of OSKM in mice enables reprogramming to occur *in vivo*, resulting in the formation of teratomas in multiple organs (Abad et al. 2013) and that partial or incomplete *in vivo* reprogramming using OSKM can result in a dysplastic phenotype in the absence of any identifiable cancer-driving mutations (Ohnishi et al. 2014) implicates epigenetic reprogramming as a driver of carcinogenesis in the absence of cancer-causing mutations. The identification of pluripotency factors in cancers has driven hypotheses regarding the origins of cancer, such as the cancer stem-cell hypothesis (Beck and Blanpain 2013).

Aside from the pluripotency transcription factors themselves, reprogramming of the epigenome is a process common to both iPS reprogramming and cancer transformation which has led to the hypothesis that cancer initiation may involve a process of dedifferentiation that is analogous to cellular reprogramming. The epigenetic progenitor model of cancer was proposed as an alternative route of cancer initiation to the classical clonal genetic model of



cancer formation, which is based on a series of mutations and selection events (Feinberg, Ohlsson, and Henikoff 2006). The epigenetic progenitor model proposes that epigenetic reprogramming or alterations may first occur in stem or progenitor cells within tissues, resulting in the expansion of an epigenetically altered pool of progenitor cells that is refractory to differentiation. This altered progenitor pool can subsequently give rise to cancer by means of further mutations as well as epigenetic and genetic instability (Feinberg, Ohlsson, and Henikoff 2006). It has also been hypothesised that cancer cells follow an evolutionary trajectory towards a stem cell state, which allows for both self-renewal and differentiation that generate cellular hierarchies in cancer (Chen and He 2016). This convergent evolution could explain how cancers with largely varied mutational profiles still exhibit common phenotypic hallmarks (Hanahan and Weinberg 2011).

## 1.7 Project Aims

While it is well established that aberrant DNA methylation patterns are a hallmark of cancer, the underlying mechanisms that give rise to global hypomethylation and gene promoter hypermethylation in early stages of human cancer development remain elusive, as does the timing and biological function of such events in cancer initiation and progression. In particular, how *de novo* DNA methyltransferase activity is preferentially targeted to specific regions of the genome in the context of aberrant cancer methylation remains largely a mystery. Moreover, the relationship between DNA demethylation and genomic instability, alongside the increased mutability of methylated CpGs has not been extensively explored.

It has been hypothesised that cancer cells follow an evolutionary trajectory towards a stem cell state, which allows both self-renewal and differentiation, and several molecular parallels have been drawn between the acquisition of pluripotency during cellular reprogramming and molecular processes occurring during oncogenesis (Chen and He 2016; Suvà, Riggi, and Bernstein 2013). This suggests that common mechanisms may drive the two processes and that molecular dissection of epigenetic reprogramming may shed light on processes occurring during tumourigenesis. In line with this, for this thesis we use the recently developed NANOG/KLF2 overexpression + 2iLGö method of primed to naïve human embryonic stem cell reprogramming (Takashima et al. 2014) as a model system to explore mechanisms of DNA methylation and the relationship between changing DNA methylation patterns and genomic stability. The aims of the project were to:

- Temporally characterise the dynamics of DNA methylation during the transition from primed to naïve pluripotency in hESCs.
- Identify the epigenetic factors responsible for carrying out hypermethylation as well as identifying upstream mechanisms involved in targeting these factors to specific regions of the DNA.
- Determine whether the global demethylation observed upon resetting primed hESCs to the naïve state gives rise to genomic instability or an increased mutation frequency.

## Chapter 2: Materials and Methods

### 2.1 Cell Culture

#### 2.1.1 Cell culture reagents

Name	Source	Catalogue Number
DMEM/F-12 Ham	Sigma Aldrich	D6421
DMEM, high glucose	ThermoFisher Scientific	11965-092
DPBS	ThermoFisher Scientific	14190-094
KnockOut serum replacement (KSR)	ThermoFisher Scientific	10828-028
Fetal bovine serum (FBS)	ThermoFisher Scientific	11965-092
Neurobasal	ThermoFisher Scientific	21103-049
L-glutamine	ThermoFisher Scientific	25030-024
MEM Non-essential amino acids solution	ThermoFisher Scientific	11140-035
2-Mercaptoethanol (2-ME)	Sigma Aldrich	M3148
Doxycycline	Sigma Aldrich	D9891
N-2 Supplement	ThermoFisher Scientific	17502-048
B-27 Supplement	ThermoFisher Scientific	17504-044
Recombinant human FGF-basic (bFGF)	Peprtech	100-18B
Recombinant human LIF (hLIF)	Peprtech	300-05
PD0325901	Stem Cell Technologies	72182
CHIR99021	Stem Cell Technologies	72052
Y-27632 dihydrochloride (ROCKi)	Sigma Aldrich	Y0503
Gö6983 (PKCi)	Sigma Aldrich	G1918
Penicillin-Streptomycin	ThermoFisher Scientific	15140122
Gelatine solution	Sigma Aldrich	G1393
Accutase solution	Sigma Aldrich	A6964
Trypsin-EDTA	ThermoFisher Scientific	25200-072
Defined Trypsin Inhibitor (DTI)	ThermoFisher Scientific	R-007-100
Trypsin	ThermoFisher Scientific	15090-046
Collagenase, type IV	ThermoFisher Scientific	17104-019
FresR-S	Stem Cell Technologies	05859
Stem-Cell Banker	Takara Bio	11890

Table 2.1 Cell Culture Reagents

### 2.1.2 Cell lines

WA09/H9 NK2 primed human embryonic stem cells (hESCs): kindly provided by Austin Smith (Takashima et al. 2014) with permission from WiCell.

Primary mouse embryonic fibroblasts (MEF): Isolated from wild-type E12.5 mouse embryos.

### 2.1.3 Isolation of Mouse Embryonic Fibroblasts

Primary mouse embryonic fibroblasts (MEF) were isolated from wild-type mouse embryos at embryonic day 12.5 (E12.5). Primary MEFs were expanded for two passages, before being irradiated at 35Gy in a biological X-ray irradiator. Irradiated MEFs (iMEF) were frozen into stocks. Irradiated MEF (iMEF) were seeded at a density of  $1 \times 10^6$  cells on 0.1% gelatine-coated 6-well plates in DMEM supplemented with 10% foetal bovine serum (FBS) and 100U/ml Penicillin-Streptomycin. After 24 hours, media was aspirated, wells were washed with DPBS, and then hESCs were seeded on iMEF plated wells.

### 2.1.4 Human embryonic stem cell culture

Human H9-NK2 human embryonic stem cells (hESCs) containing doxycycline-inducible *KLF2* and *NANOG* transgenes coupled to Venus (H9-NK2) were maintained in conventional medium (KSR/FGF) comprised of DMEM/F-12 Ham with 20% KnockOut Serum Replacement (KSR) and 10ng/ml basic fibroblast growth factor (bFGF), supplemented with 2mM L-glutamine, 100uM 2-mercaptoethanol (2ME), 1% MEM non-essential amino acids, and 100U/ml Penicillin-Streptomycin. Cultures were passaged every 5-6 days as small clumps by dissociation with a buffer containing 1mg/ml Collagenase IV, 0.025% Trypsin, 1mM CaCl<sub>2</sub> and KSR at a final concentration of 20% in DPBS. Medium was changed daily. Cells were frozen in FresR-S freezing solution.

Resetting to the naïve state was carried out as previously described (Takashima et al. 2014). H9-NK2 cells were dissociated to single cells with trypsin-EDTA, trypsin inactivated using the Defined Trypsin Inhibitor (DTI), and re-plated in the presence of 10µM Rho-associated kinase inhibitor (ROCK inhibitor [Y-27632]). After 24 hours, media was changed to primed media with 1µM doxycycline. The following day, media was changed to 2iL+dox media composed of 50% DMEM/F12 and 50% Neurobasal supplemented with 2mM L-glutamine, 100µM 2ME, N2, B27, 1µM PD0325901 (MEK inhibitor), 1µM CHIR99021 (GSK3β inhibitor), 20ng/ml human recombinant LIF (hLIF), 100U/ml Penicillin-Streptomycin and 1µM doxycycline. Media was

changed daily. Cells were split every 4-5 days after dissociation to single cells using Accutase. After 2 weeks, doxycycline was withdrawn and PKC inhibitor Gö6983 was added at a concentration of 5µM. Cells in 2iL+Gö were split every 4-5 days after dissociation to single cells using Accutase. Cells were frozen in Stem-Cell Banker freezing solution. All hESCs were cultured on irradiated mouse embryonic fibroblasts (iMEF). All hESCs were maintained in 5% O<sub>2</sub>, 7% CO<sub>2</sub> at 37°C in a humidified incubator.

### **2.1.5 HEK293T cell culture**

Human embryonic kidney 293T (HEK293T) cells were maintained in DMEM supplemented with 10% FBS and 100U/ml Penicillin-Streptomycin. Cells were maintained in 5% CO<sub>2</sub> at 37°C in a humidified incubator.

### **2.1.6 Brightfield Microscopy of cells in culture**

Brightfield images of cells in culture were captured on the EVOS XL Core Cell Imaging System (ThermoFisher Scientific).

## **2.2 Embryoid body generation**

Embryoid body generation was performed as previously described (Takashima et al. 2014). Briefly, 10,000 naïve hESCs dissociated with Accutase were plated per well of a 96-well clear black round bottom ultra-low attachment spheroid microplate (Corning) in N2B27 with 10% KSR. Medium was changed every second day. RNA was isolated from cells from embryoid bodies harvested at days 0 and 5.

## **2.3. Immunofluorescence**

### **2.3.1 Immunostaining**

Cells grown on coverslips were washed with PBS and then fixed in 4% paraformaldehyde (PFA) for 15 minutes at room temperature. Fixed cells were blocked with 2% FBS/PBS + 0.1% BSA + 0.1% Triton (PBSBT) for 2 hours. Primary antibodies were diluted in PBSBT and incubated at overnight at 4°C. Cells were washed in excess PBSBT for 2 hours and incubated for 1 hour at room temperature in the dark in either goat anti-mouse or goat anti-rabbit Alexa Fluor (AF)

488 (A11029 and A11008; ThermoFisher Scientific) secondary antibodies at a 1:1000 dilution. Cells were washed in excess PBSBT for 2 hours before being counterstained with Vectashield Antifade Mounting Medium with 4',6-diamidino-2-phenylindole (DAPI) (VectorLabs) and mounted onto polysine® adhesion slides (VWR). Staining of 5-methylcytosine (5mC) was performed as previously described (Ficz et al. 2013). Briefly, fixed cells were permeabilised in PBS + 0.5% Triton for 1 hour at room temperature. Cells were incubated in 2N hydrochloric acid for 30 minutes before being blocked in PBSBT for 2 hours, and stained as above.

For staining of DNA damage markers, cells grown on coverslips were fixed with PTEMF (0.2% Triton X-100, 0.02 M PIPES [pH 6.8], 0.01 M EGTA, 1 mM MgCl<sub>2</sub>, and 4% formaldehyde) for 10 minutes at room temperature. After blocking with 3% BSA, cells were incubated with primary antibodies for 1 hour at room temperature. Secondary antibodies used were goat anti-mouse Alexa Fluor 488 (A11017; Invitrogen), goat anti-rabbit AF594 and AF488 (A11012 and A11008; Invitrogen), and goat anti-human AF647 (109-606-088-JIR [Stratech] or A21445 [Invitrogen]). Each secondary antibody was used at a 1:100 dilution for 30 minutes at room temperature, protected from light. DNA was stained with DAPI (Roche), and coverslips were mounted in Vectashield (Vector H-1000; Vector Laboratories) onto SuperFrost microscope slides (ThermoFisher Scientific).

Primary antibody details are listed in Table 2.2.

Antibody target	Species	Source	Catalogue number	Dilution
<b>KLF4</b>	Rabbit	Santa Cruz Biotechnology	sc-20691	1:200
<b>THY1</b>	Mouse	Merck Millipore	CBL415	1:200
<b>5mC</b>	Mouse	Active Motif	39649	1:200
<b>UHRF1</b>	Mouse	Merck Millipore	MABE308	1:1000
<b>DNMT1</b>	Mouse	Abcam	ab13537	1:1000
<b>Centrin 3</b>	Mouse	Abcam	ab54531	1:250
<b>CREST</b>	Human	Antibodies Inc	15-234-0001	1:200
<b>γ-H2AX</b>	Mouse	Merck Millipore	05-636	1:250
<b>53BP1</b>	Rabbit	Santa Cruz Biotechnology	sc-22760	1:250
<b>RPA70</b>	Rabbit	Abcam	ab79398	1:250
<b>Beta-tubulin</b>	Rabbit	Abcam	ab6046	1:250

Table 2.2 Primary antibodies used for immunostaining

### 2.3.2 Microscopy

For staining of epigenetic regulators and markers of pluripotency, slides were visualised and imaged on the Nikon Eclipse Ci fluorescence microscope.

For staining of DNA damage markers, images were acquired using an Olympus DeltaVision RT microscope (Applied Precision) equipped with a Coolsnap HQ camera. Three-dimensional image stacks were acquired in 0.2 mm steps, using Olympus 1003 (1.4 numerical aperture), 603, or 403 UPlanSApo oil immersion objectives. Deconvolution of image stacks and quantitative measurements was performed with SoftWorx Explorer (Applied Precision). Imaging analysis was performed using Softworx Explorer and ImageJ (for CREST spot counting only). Statistical analysis was performed using Prism v7.04 (GraphPad).

### 2.4 Flow cytometry

Cells were dissociated to single cells with Accutase and washed in PBS. Cells were fixed in 0.25% paraformaldehyde (PFA) at 37°C for 10 minutes, cooled at 4°C for 10 minutes and then permeabilised in ice-cold 88% methanol at -20°C for 10 minutes. Following fixation and permeabilisation, cells were washed in PBS/ 0.1% Tween-20/ 1% bovine serum albumin (PBST-BSA) and then blocked in 10% FBS in PBST-BSA for 30 minutes at 37°C. Cells were stained with an anti-feeder antibody conjugated to the fluorophore allophycocyanin (APC) for 10 minutes at 4°C in the dark, and with an anti-CD90 antibody conjugated to the fluorophore phycoerythrin (PE) for 30 minutes at room temperature in the dark. Alternatively, cells were stained with the anti-feeder APC antibody and then with DAPI for 15 minutes at 4°C in the dark. Samples were analysed on an LSR Fortessa cell analyser (BD Biosciences). Flow cytometry data analysis was carried out using FlowJo Version 10 software.

### 2.5 Reverse Transcription Quantitative Real-time PCR (rt-qPCR)

Cells were dissociated to single cells using Accutase and serially plated onto 6-well plates for 2 hours to eliminate excess iMEFs. Total RNA was isolated from pelleted hESCs using the Direct-zol RNA mini-prep kit (Zymo) as per manufacturer's instructions, and treated with the DNA-free™ DNA removal kit (ThermoFisher Scientific). Complementary DNA (cDNA) was made using the High-capacity RNA to cDNA™ kit (ThermoFisher Scientific). Real-time PCR was carried out using Sso advanced™ Universal SYBR® Green Supermix (Bio-Rad) on the CFX384 Touch™ Real Time PCR detection system (Bio-Rad). An endogenous control (GAPDH) was used to normalise

expression. Primers were designed using Primer-BLAST (which uses Primer3 for design and BLAST to ensure target specificity) and are listed in Table 2.3.

Gene	Primer Sequence (5' – 3')
GAPDH	GATTTGGTCGTATTGGGCGC
	TTCCCGTTCTCAGCCTTGAC
KLF4	TCTCCAATTCGCTGACCCAT
	CGGATCGGATAGGTGAAGCT
DNMT3A	AGTACGACGACGACGGCTA
	CACACTCCACGCAAAGCAC
DNMT3B	GAGTCTGCACGGGACCTATT
	GTCACGGGGAGGGATTTAGC
DNMT3L	CATAGCCTGGTGACCTCTGG
	CTGGTGGGTTCAAGTTCCA
DNMT3A1	ACTACATCAGCAAGCGCAAG
	CACAGCATTATTCTGCAA
DNMT3A2	GCTGCACCTGGCCTTATG
	CGTCTTTCAGGCTACGATCC
TET1	CGAGTTGGAAAGTTTGCCCG
	CACAAGGTTTTGGTCGCTGG
TET2	AATTTGCAAGCTCCTGGTGG
	GCAATTGTGATGGTGGTGGT
TET3	CAAGACACCTCGCAAGTTC
	CGTTGGTCACCTGGTTCTGAT
DNMT1	GACCACGGTTCCTCCTTCTG
	CGGCCTCGCATAACTCTCC
UHRF1	TCAGACAAGTCTCCACCCA
	TGTACAGCCCAATTCCGTC
FOXC1	GCTGTCAAATGGCCTTCCCT
	TCCTGCTTTGGGGTTGATT
ZFHX3	CTTCCAGAGGAGGACGAGGA
	AATGGCTTCTTCTGGGTCCG
SOX15	GGCTTTGGGTACAGACCCC
	GTTTGCAGTGGGAAGAGCCAT
NFKB1	GCTTAGGAGGGAGAGCCCA
	CTGCCATTCTGAAGCCGGG
MECP2	TGTTAGGGCTCAGGGAAGAAAAG
	AACTTGAGGGGTTTGTCTTGA
DPPA4	ATAGCAATCTTGGGGCAGGG
	TCTGTGGAGGTCCACTCCTT
AICDA	GGGAGGCAAGAAGACTCT
	GGTCCCAGTCCGAGATGTAG
NANOG	GCTACAAACAGGTGAAGACC
	CCTGCGTCACACCATTG
TFCP2L1	ATGAAGAGACGCTGACCTAC
	ACACGGATGATGCTCTTG

Table 2.3 Primer Sequences used for rt-qPCR



## 2.6 Western Blotting

### 2.6.1 Protein Isolation

Cells to be processed for protein analysis were dissociated from growth plates using Accutase, followed by centrifugation to obtain cellular pellets. Whole cell lysates were extracted in radioimmunoprecipitation assay buffer (RIPA buffer; Sigma Aldrich), with protease inhibitor cocktail (Sigma Aldrich). The suspension was centrifuged at 10,000g at 4°C for 15 minutes to pellet cell debris. The supernatant containing the protein was transferred into a new microcentrifuge tube and the pellet discarded. Protein lysates were subsequently stored at -80°C until required.

### 2.6.2 Protein Quantification

Proteins concentration was determined using the Pierce™ BCA Protein Assay kit (ThermoFisher Scientific) according to manufacturer's instructions. In a 96-well plate, 25µl of albumin protein standards (dilutions of 25-2000µg/ml prepared using serial dilution) and test protein samples were loaded in triplicate. 200µl of bicinchoninic acid (BCA) reagent was added to each well. This begins a two-step reaction, starting with the biuret reaction which involves chelation of copper with protein in an alkaline environment to form a light blue complex. The second step of the reaction involves the reaction of BCA with the reduced cuprous cation formed in step one, resulting in the formation of a water soluble complex that exhibits a strong linear absorbance at 562nm with increasing protein concentrations. The reaction is strongly influenced by the amino acid sequence of the protein. The absorbance of each well was quantified using the FLUOstar Omega Microplate reader (BMG Labtech) using an excitation of 595nm. The concentration of the BCA standards was plotted against their absorbance, and an equation for the standard curve was generated. Protein concentrations for the test samples were calculated based on the standard equation.

### 2.6.3 Gel Electrophoresis

25µg of protein was diluted in distilled water to a final volume of 13µl per sample. To this, 2µl of sample reducing agent (ThermoFisher Scientific) and 5µl of 4x NuPAGE LDS Sample loading buffer (ThermoFisher Scientific) were added, and samples were placed in a heat block at 99°C for 5 minutes. Denatured proteins were separated by electrophoresis on a 4-12% Bis-Tris gel in MOPS running buffer (ThermoFisher Scientific) at 200 Volts for 55 minutes, alongside the colour pre-stained broad range protein standard (NEB).

## 2.6.4 Western Blot Transfer

Resolved proteins were transferred to polyvinylidene difluoride (PVDF) membranes (Merck Millipore) using the Mini Trans-Blot® Electrophoretic transfer cell system (Bio-Rad). PVDF membranes were equilibrated in methanol for 1 minute. Cassettes were prepared in a plastic tray filled with 1L of Transfer Buffer. A sponge and 2 pieces of filter paper were placed on the negative side of the cassette, followed by the gel, the PVDF membrane, 2 further pieces of filter paper and another sponge. Throughout the procedure, each layer was rolled gently with a roller to ensure no air bubbles remained within the stack. The cassette was closed and placed in the transfer tank along with an ice block to prevent overheating during the transfer process. The transfer buffer was poured into the tank and transfer was carried out at 22 Volts overnight for 16 hours at 4°C.

## 2.6.5 Detection

Following transfer of proteins, PVDF membranes were removed from the cassettes and blocked with 5% skimmed milk for 45 min at room temperature. Membranes were then incubated overnight at 4°C with primary antibody in blocking buffer, with gentle rolling. The following day, membranes were washed three times in 0.1% Tween-20 in PBS (PBST) for 5 minutes each, before being incubated for 1 hour at room temperature with horseradish peroxidase-conjugated secondary antibodies: sheep-anti-mouse IgG or sheep-anti-rabbit IgG (1:5,000; GE Healthcare). Membranes were washed a further three times for 5 minutes each in PBST. Detection was performed with enhanced chemiluminescence (ECL; ThermoFisher Scientific), by adding 5ml of ECL reagent (a 1:1 mixture of peroxide solution and Luminol enhancer) to the membrane for 5 minutes. Visualisation of chemiluminescence was performed on the Amersham Imager 600 (GE Healthcare). Primary antibodies used are detailed in Table 2.4

Antibody target	Species	Source	Catalogue number	Dilution
<b>DNMT1</b>	Mouse	Abcam	ab13537	1:1000
<b>UHRF1</b>	Mouse	Merck Millipore	MABE308	1:5000
<b>GAPDH</b>	Rabbit	Cell Signalling Technology	2118S	1:2500
<b>Alpha-tubulin</b>	Mouse	Abcam	ab7291	1:5000
<b>TET1</b>	Rabbit	GeneTex	GTX124207	1:1000
<b>P53</b>	Mouse	Santa Cruz Biotechnology	sc-126	1:1000

Table 2.4 Primary antibodies used for Western Blotting

## 2.7 Alkaline phosphatase assay

To perform an alkaline phosphatase assay, the Amplite™ Colorimetric Alkaline Phosphatase Assay Kit (Strattech) was used according to manufacturer's instructions. Briefly, equal numbers of hESCs were seeded on wells of 0.1% gelatine and iMEF coated 96 well plates. After 24 hours, media was aspirated and wells were washed with DPBS. 50µl of each alkaline phosphatase standard solution (prepared by serial dilution) was plated on the same plate. P-nitrophenyl phosphate (pNPP) working solution was prepared and 50µl was added to each well of standards and 100µl to each test well containing cells. Plates were incubated at 37°C for 30 minutes in the dark, following which the absorbance increase was measured with the FLUOstar Omega Microplate reader (BMG Labtech) using the absorbance plate reader detection mode at 400nm.

## 2.8 Stable shRNA knockdown cell line generation

### 2.8.1 Acquisition and preparation of shRNA plasmids

Short hairpin RNA (shRNA) constructs were obtained from Dharmacon in The RNA Consortium (TRC) pLKO.1 HIV-based lentiviral vector. Constructs were obtained as glycerol stocks of plasmid DNA in DH5α E.Coli bacteria. Bacteria containing each construct was streaked onto an LB agar plate containing 100µg/ml Ampicillin and incubated overnight at 37°C. The following day, colonies were picked from the agar plate and 3ml of LB broth medium (containing 100µg/ml Ampicillin) was inoculated with each colony. Cultures were incubated at 37°C for 16 hours with vigorous shaking, following which plasmid DNA was purified from pelleted cultures using the Zyppy plasmid miniprep kit (Zymo Research), following manufacturer's instructions. shRNA sequences and identifiers are listed in table 2.5.

### 2.8.2 Generation of lentiviral particles

To generate lentiviral particles, HEK293T cells in a 6-well plate at 70% confluency were transfected with 3µg of the shRNA plasmid for a target gene, 2µg of the packaging construct pCMV Δ8.91, and 1µg of the vesicular stomatitis virus glycoprotein (VSV-G) containing envelope expressing plasmid pMD2.G, using jetPrime (Polyplus) at a ratio of 1:2. 48 hours later, media containing viral particles was collected and filtered through a 0.22µm filter.

### 2.8.3 Transduction of hESCs and generation of stable knock down cell lines

Primed hESCs were treated with 6µg/ml polybrene (Sigma Aldrich) for 15minutes, followed by transduction with filtered lentiviral particles. Stable hESC knock down cell lines were generated by puromycin selection (1µg/ml) of successful integrants for 48 hours.

Target Gene	Mature antisense sequence	Identifier	Catalogue Number
<b>DNMT3A</b>	AATAATCTCCTTGACCTGGG	TRCN0000035754	RHS3979-201764981
	TTATTAGCGAAGAACATCTGG	TRCN0000035757	RHS3979-201764984
<b>DNMT3B</b>	TATAGCAATTTGTCTTGAGGC	TRCN0000035684	RHS3979-201764911
	ATTTGAGAGATCGTTGCATGG	TRCN0000035686	RHS3979-201764913
<b>DNMT3L</b>	TTGTCCACGAACATCCAGAAG	TRCN0000019683	RHS3979-201750800
	TTCTTCTCCGAAACCAGAGC	TRCN0000019680	RHS3979-201750797
<b>DNMT3A isoform 1</b>	ATTGGGTAATAGCTCTGAGGC	TRCN0000035756	RHS3979-201764983
<b>FOXC1</b>	AGGGTGATCTTCTTGTCCGGG	TRCN0000013964	RHS3979-201745542
	TTCAGGTACCACGAGGTGAGG	TRCN0000013965	RHS3979-201745543
<b>ZFH3</b>	ATTCTCATAATTCTTCTGGC	TRCN0000013558	RHS3979-201745136
	TTAGCTGTGGAACTAAAGGG	TRCN0000013560	RHS3979-201745138
<b>SOX15</b>	TTCCAGTTTGCAGTGGGAAGA	TRCN0000019150	RHS3979-201750357
	TAGTGGGTATAGGTGGGCAGC	TRCN0000019151	RHS3979-201750358
<b>NFKB1</b>	ATCATCAGATGTAACTCTGG	TRCN0000006518	RHS3979-201738791
	TTCAGGATAGTAGAGGAAAGG	TRCN0000006520	RHS3979-201738793
<b>Non-targeting control</b>	N/A	N/A	RHS6848

Table 2.5 shRNA sequences used to generate knock down cell lines

## 2.9 Stable overexpression cell line generation

### 2.9.1 Acquisition of entry clones of target genes

Entry clones for overexpression of genes, listed in Table 2.6, were obtained from Harvard PlasmID Repository. These were made up of full length sequences for the target gene in the pENTR223 gateway cloning compatible vector.

### 2.9.2 Generation of expression vectors for target genes

To generate expression vectors, a recombination reaction was performed with 100ng of the entry clone for each target gene and 150ng of the pLenti CMV puro DEST (w118-1) destination vector with Gateway LR clonase II (ThermoFisher Scientific), following manufacturer's instructions. Following a 1 hour incubation with the LR clonase II enzyme, 1µl of each of the reactions was transformed into 50µl of NEB 5-alpha high efficiency competent E.coli (NEB) by heat-shock (incubating at 42°C for 30 seconds). After 1 hour of incubation with shaking at 37°C in SOC medium, 100µl of the transformation was plated onto LB agar plates containing 100µg/ml Ampicillin. The following day, colonies were picked from the agar plate and 3ml of LB broth medium (containing 100µg/ml Ampicillin) was inoculated. Cultures were incubated at 37°C for 16 hours with vigorous shaking, following which plasmid DNA was purified from pelleted cultures using the Zyppy plasmid miniprep kit (Zymo Research), following manufacturer's instructions.

### 2.9.3 Generation of lentiviral particles

To generate lentiviral particles, HEK293T cells at 70% confluency were transfected with 3µg of the expression vector for a target gene, 2µg of the packaging construct pCMV Δ8.91, and 1µg of the vesicular stomatitis virus glycoprotein (VSV-G) containing envelope expressing plasmid pMD2.G, using jetPrime (Polyplus) at a ratio of 1:2. 48 hours later, media containing viral particles was collected and filtered through a 0.22µm filter.

## 2.9.4 Transduction of hESCs and generation of stable overexpression cell lines

Primed hESCs were treated with 6µg/ml polybrene (Sigma Aldrich) for 15minutes, followed by transduction with filtered lentiviral particles. Stable hESC knock down cell lines were generated by puromycin selection (1µg/ml) of successful integrants for 48 hours.

Gene	Vector	Catalogue Number	Source
<b>TET1</b>	TET1 in pENTR223	HsCD00399189	Harvard PlasmID repository
<b>MECP2</b>	MECP2 in pENTR223	HsCD00373134	Harvard PlasmID repository

Table 2.6. Entry clones used for overexpression of genes

## 2.10 Chromatin Immunoprecipitation and qPCR (ChIP-qPCR)

### 2.10.1 Chromatin extraction

Cells were cross-linked with 1% formaldehyde for 10 minutes at room temperature with gentle rocking, after which the formaldehyde was quenched with 1.25M glycine. Chromatin was then extracted from the cross-linked cells using the chromatin extraction kit (Abcam), as per the manufacturer's instructions. The extracted chromatin was then fractionated by sonication at 4° (12 cycles of 15s on, 60s off; Diagenode Bioruptor® Plus). The size profile of the sonicated chromatin was checked by agarose gel electrophoresis.

### 2.10.2 Chromatin Immunoprecipitation

Chromatin immunoprecipitation was carried out using the ChIP – One Step kit (Abcam) with a starting total of 5µg of chromatin. Immunoprecipitation was carried out as per the manufacturer's instructions and the following quantities of antibody were used for immunoprecipitation: H3K4me3 0.5µg (Abcam; ab8580), H3K27me3 2µg (Abcam; ab195477). As a loading control for assessing immunoprecipitation, input DNA was isolated from 10µl of chromatin for each sample by incubation with 88µl of DNA release buffer and 2.5µl of proteinase K at 65°C for 15 minutes followed by 95°C for 10 minutes. The input DNA is representative of the starting quantity of chromatin prior to immunoprecipitation. Input and immunoprecipitated DNA were quantified by real-time PCR, and data was quantified as the %

enrichment of immunoprecipitated DNA relative to the input for each sample. Primers used for PCR reactions are listed in Table 2.7.

Target genomic region	Primer Sequence (5' – 3')
COL11A1	AGAGAACTGCACGTCCAACC
	TGCAACCAAGTGAGAAGCAG
DLX5	GGCAATCTGGGAGTTCCACA
	CTGAGCGGGGCTGTATCTTG
TRPC4	TATATGCACCCAGATGCCCC
	TTTAAAGCAGGGGAGAGGGC
PITX1	CATACACAGGGACGCTGTAAAC
	GGGAGGTCCATCTCAGAACA
SIM1	GCCTGGGGAGTAAGGAGACT
	AACTTCCTCCGCTGGTAGC
SLC17A9	GCTCCTATGGGGCAGCAG
	GGCTTGGGGTGCTCAGAC
LIN9	ATCTCAGGCACGTTGGTTTC
	CCAGTGACTCACCCAATCCT
WTIP	GGTTGGGACGAGGAAGGT
	AGTGTGCCCATGAACCTGAC
DPP6	TATTGGTAGCGGCCAAAAG
	GATCATGGCCTTTTCCTCA

Table 2.7 Primers used for ChIP-qPCR

## 2.11 Targeted Bisulphite Sequencing

### 2.11.1 Primer design

Bisulphite PCR primers were designed against an *in silico* bisulphite converted reference sequence using the Bisulphite Primer Seeker software (Zymo) or Methprimer (Urogene), and the following universal Illumina adapter sequences were added to the 5' end of each primer.

Forward primer: ACACTCTTCCCTACACGACGCTCTCCGATCT

Reverse primer: GTGACTGGAGTTCAGACGTGTGCTCTCCGATCT



### 2.11.2 DNA isolation and bisulphite conversion

Cells were dissociated to single cells using Accutase and serially plated for 2 hours to eliminate excess iMEFs. DNA was isolated from pelleted cells using the PureLink Genomic DNA mini kit (ThermoFisher Scientific) and quantified using the Nanodrop 2000 spectrophotometer (ThermoFisher Scientific). Bisulphite conversion of DNA was carried out using the Imprint<sup>®</sup> DNA Modification kit (Sigma Aldrich), following the manufacturer's instructions. This process involves a chemical reaction of the DNA with a sodium bisulphite reagent with which results in unmethylated cytosine bases being converted to uracils (which will be sequenced as thymine), whereas methylated cytosine bases are not converted.

### 2.11.3 PCR amplification of target regions

The modified DNA was amplified using the loci specific bisulphite PCR primers with HotStar Taq DNA Polymerase (Qiagen). The PCR conditions were as follows: 95 °C for 15 min; 94 °C for 30 seconds; 56 °C for 30 seconds; 72 °C for 1 min; Repeat steps 2-4 29X; 72 °C for 10 min; Hold 12°C. PCR products were purified using solid phase reverse immobilisation (SPRI) magnetic beads at a 1:1 ratio (Agencourt AMPure XP, Beckman Coulter). Purified PCR amplicons were PCR amplified with a further 8 cycles using a universal Illumina forward (P5) primer and an indexed reverse (P7) primer. PCR products were purified using SPRI beads at a 1:1 ratio. Purified amplicons were quantified using the Qubit™ dsDNA high sensitivity assay, measured on the Qubit 4 Fluorometer (ThermoFisher Scientific). Amplicons for each sample were pooled at equal concentrations, based on quantification on the Nanodrop 2000 spectrophotometer (ThermoFisher Scientific). Quantification of each sample was performed using the Kapa Library quantification kit for Illumina (Roche), and samples were pooled at equal concentrations to enable multiplexing. Bisulphite primer sequences for target loci are listed in table 2.8 and Illumina P5 and P7 index primers are listed in table 2.9.

### 2.11.4 Fluidigm targeted amplicon sequencing

For larger experiments, multiplex targeted bisulphite sequencing was performed on bisulphite converted DNA using the 48x48 layout on the Fluidigm C1 system (Fluidigm). This system allows amplification of 48 genomic regions (based on 48 primer pairs) for 48 samples of bisulphite converted genomic DNA. Bisulphite primer sequences for target loci analysed on the Fluidigm platform are listed in table 2.10.

### 2.11.5 Amplicon library Sequencing

Sequencing was performed on an Illumina MiSeq with 150bp paired-end reads, using v3 chemistry, at Barts and the London Genome Centre (London, UK).

### 2.11.6 Data Analysis

Sequencing reads were obtained in the form of FASTQ files and quality of reads was assessed using FastQC (Babraham Bioinformatics). Sequencing reads were trimmed to remove adapters and low-quality ends using Seqtk. The remaining sequences were mapped to a customised human genome composed of amplicon sequences or the human genome GRCh38 using Bowtie2-2.2.9 and Bismark v0.19.0 (Krueger and Andrews 2011) with default parameters, and methylation calls were extracted as a percentage for each individual CpG site using Bismark methylation extractor.

Gene/region	Primer Sequence (5' – 3')
<b>DLX5</b>	TGTTTAGTATTAGTTTAGTTTTATTTGGAGTGTGG
	AATCCAAACRCRAAAAACAAAAAATTAATACAC
<b>PITX1</b>	GGGGTTGTTYGTTTAGATAGAGGGTTATTTTTAG
	TAAAAAAACRATACCCCAACCCAAAATC
<b>FGF17</b>	GGTAYGAGGGTTGGTTTATGG
	AATAAAAAAACRACCCAAAAACTACTACCCCTAC
<b>SLC17A9</b>	TTGGTTTTGTTTTATGGGGTAGTAGGG
	TATATCCRAAACTACCTCCAACCCAACTAACC
<b>Col11A1</b>	GTTTTTYGTGTTTTAGTTTTTTTTTTTTTTTGAG
	ACCCACTAACRACATAAAACCTTAAAAACACAC
<b>HOXA3</b>	GAYGTTAGTATGTTTTGTTTTTTGATTTTTTTG
	CRAATAAAAATAACCAATCTACTAAACCTCACTAAC
<b>VAV3</b>	TTGAATTGTGATTTTTGAGTTGATTTAGGGTG
	AAAAACRAATCCAACCTCTCTCAACAAC
<b>TRPC4</b>	GAAGAGAGTYGGTTTTTAGGTAGTTTAGG
	TCAATCCTCRAATCCCATCACTTCAACC
<b>MEIS2</b>	AATAGTTAGTTTTTTTGTAGGTTGGAAATGG
	AAATCCCACRCAACTAAACAACCTCTCTC
<b>Hypomethylated_region_1</b>	GTTTGGGTYGAGTTTAGTGTATTTATTTATTTG
	ATATTTCAATTATCCAACTTCTCTCTCTCC
<b>Hypomethylated_region_2</b>	YGAAATGTTAGTGGGTTAGAATTGGGAG
	AACRAAAAAAATTAAAAACTTAACCCCATAC
<b>Hypomethylated_region_3</b>	GTTYGGAGGATTTAGGGTTTGGG
	ACRAAACCAAAAAAACAACAACACTACTACC
<b>Hypomethylated_region_4</b>	GGTYGATTTGTTTTTAGTTTTGAGGTTTTG
	ACRTTAACCACCAACCCAAAATATC

**Table 2.8. Bisulphite primer sequences for targeted bisulphite sequencing**

Primer	Primer Sequence (5' – 3') - unique index <u>underlined</u>
P5	AATGATACGGCGACCACCGAGATCTACTCTTTCCCTACACGACGCTCTTCCGATCT
P7	CAAGCAGAAGACGGCATAACGAGATT <u>ACAAGTT</u> GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
	CAAGCAGAAGACGGCATAACGAGAT <u>ATCACTGGG</u> TGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
	CAAGCAGAAGACGGCATAACGAGAT <u>CGCATCAAG</u> TGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
	CAAGCAGAAGACGGCATAACGAGAT <u>GCACGACC</u> GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
	CAAGCAGAAGACGGCATAACGAGATT <u>ACTCCG</u> TGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
	CAAGCAGAAGACGGCATAACGAGAT <u>CGGTCTAAG</u> TGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
	CAAGCAGAAGACGGCATAACGAGAT <u>ATGTTCCGG</u> TGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
	CAAGCAGAAGACGGCATAACGAGAT <u>CGTGGACC</u> GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
	CAAGCAGAAGACGGCATAACGAGAT <u>ATTGAGCC</u> GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
	CAAGCAGAAGACGGCATAACGAGATT <u>AGTTCGGG</u> TGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
	CAAGCAGAAGACGGCATAACGAGAT <u>CGGTGAGGG</u> TGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
	CAAGCAGAAGACGGCATAACGAGAT <u>CGTGAGTT</u> GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT

**Table 2.9. Illumina P5 and P7 indexed adapters used for targeted bisulphite sequencing**

Gene/region	Primer Sequence (5' – 3')
DLX5	TGTTTAGTATTAGTTTAGTTTTATTGGAGTGTGG
	AATCCAAACRCRAAAAACAAAAATTAATACAC
ZFHX3	GAGATGTTGATTTAGAGTTTTTTTT
	ACCACCTAAAATCCCTCTACTTCTT
FBXL13	GTAATTGGGGTTAGTTGGATGTTAG
	AAACAACACATAAACTAATTTTCTTTCTTA
RAB34	AGGTTTGGGAGGTGATTTATAGAGT
	ACAATAAACACCCATACCAAAAAA
ACHE	GAAGGAAGGGAAGGTTTAGTTTGA
	TTTAAAAAATCTCAAAACATCCTAAC
SIM1	TTTTTTGAGAGAGTGTAGGAGAGTTT
	ACTAATTACACCAATTTCCCTCTCTT
TBX4	GGGTTTTAGATATAGTTGGATTAG
	ACCCATAAAAATAAAATTAACAAAC
NFIX	TAGTAAATTGAAAGGATTAGTGAAT
	TCTAACCCCTACAAAAATAACACC
NR2F2	TTATTAATTGTGGAGTGTTTTTTTT
	ATACCATAATATTATTAACACTACATACAT
SHH	ATAGTAGGTTTGATAGAGATTTGGG
	ACTACAAATAACAACACTCACCTAAC
TBX5	AAAGTAAAGATTTTTAAGGTTGGTT
	TTCTATTCCCCAAAAA
FGF17	GGTAYGAGGGTTGGTTTATGG
	AATAAAAAACRACCCAAAAACTACTACCCCTAC
NKX6-1	TTGATTTGTGAGAATTAATAAATAA
	ACAATAAACTCCCTAACTATTTAAC
PITX1	GGGGTTGTTYGTTTAGATAGAGGGTTATTTTTAG
	TTAAAAAACRATACCCCAACCCAAAATC
ITGAM	TGGGGAATTTTTAGAAATTTAGAGT
	CCCCAATCACACAACCTAACAAAC
PAX8	TTAATTTTTGGGTGATATATTTGGT
	ATTTCTAACTCCTAAATCCACTCAAC
SLC17A9	TTGGTTYGTTTTATGGGGTAGTAGGG
	TATATCCRAAACTACCTCCAACCCAACTAAC

Table 2.10. Primer sequences for Fluidigm targeted bisulphite sequencing

## 2.12 Mass Spectrometry based Proteomics

### 2.12.1 Cell lysis and trypsin digestion

For proteomics experiments, 3 independent biological replicates per condition were used. Cells were washed twice with ice cold PBS supplemented with 1 mM Na<sub>3</sub>VO<sub>4</sub> and 1 mM NaF, lysed in urea buffer (8M urea in 20 mM in HEPES pH 8.0, 1 mM Na<sub>3</sub>VO<sub>4</sub>, 1 mM NaF, 1mM Na<sub>4</sub>P<sub>2</sub>O<sub>7</sub> and 1 mM sodium β-glycerophosphate) for 30 min and homogenized by sonication (15 cycles of

30s on 30s off; Diagenode Bioruptor<sup>®</sup> Plus, Liege, Belgium). Insoluble material was removed by centrifugation at 20,000 xg for 10min at 5C and protein levels in the cell extracts were quantified by bicinchoninic acid (BCA) analysis.

For trypsin digestion, 100 µg of protein were reduced and alkylated by sequential incubation with 10 mM DTT and 16.6 mM iodoacetamide for 1h and 30min, respectively. Urea concentration was diluted to 2 M with 20 mM HEPES (pH 8.0), 80 µL of preconditioned trypsin beads [(50% slurry of TLCK-trypsin (Thermo-Fisher Scientific; Cat. #20230)] were added and samples were incubated for 16h at 37C with agitation. Trypsin beads were preconditioned with 3 washes of 20 mM HEPES (pH 8.0)). Finally, samples were centrifuged for 2,000 xg for 5min at 5C and supernatants were transferred to fresh tubes.

### 2.12.2 Desalting

Peptide solutions were desalted using 10 mg OASIS-HLB cartridges (Waters, Manchester, UK). Briefly, OASIS cartridges were accommodated in a vacuum manifold (-5 mmHg), activated with 1 mL ACN and equilibrated with 1.5 mL washing solution (1% ACN, 0.1% TFA). Peptides were loaded into the cartridges, washed with 1 mL of washing solution, eluted with 500 µL of ACN solution (30% ACN, 0.1% TFA), dried in a speed vac (RVC 2-25, Martin Christ Gefriertrocknungsanlagen GmbH, Osterode am Harz, Germany) and stored at -80C.

### 2.12.3 Mass spectrometry data acquisition

Dried peptides were dissolved in 0.1% TFA and analysed by nanoflow ultimate 3,000 RSL nano instrument was coupled on-line to a Q Exactive plus mass spectrometer (Thermo Fisher Scientific). Gradient elution was from 3% to 35% solvent B in 120 min at a flow rate 300 nL/min with solvent A being used to balance the mobile phase (buffer A was 0.1% formic acid in water and B was 0.1% formic acid in acetonitrile) . The spray voltage was 1.95 kV and the capillary temperature was set to 255C. The Q-Exactive plus was operated in data dependent mode with one survey MS scan followed by 15 MS/MS scans. The full scans were acquired in the mass analyser at 375-1500 m/z with the resolution of 70,000 and the MS/MS scans were obtained with a resolution of 17,500. Overall duty cycle generated chromatographic peaks of approximately 30s at the base, which allowed the construction of extracted ion chromatograms (XICs) with at least 10 data points.

#### 2.12.4 Peptide identification and quantification

Mascot Daemon 2.5.0 was used to automate peptide identification from MS data. Peak list files (MGFs) from RAW data were generated with Mascot Distiller v2.5.1 and loaded into the Mascot search engine (v2.5) in order to match MS/MS data to peptides. Searches were performed against the SwissProt Database (release December 2015) with a FDR of ~1% and restricted to the human entries. Mass tolerance of  $\pm 10$  ppm for the MS scans and  $\pm 25$  mmu for the MS/MS scans, 2 trypsin missed cleavages, carbamidomethyl Cys as a fixed modification and PyroGlu on N-terminal Gln and oxidation of Met as variable modifications were allowed. The in-house developed Pescal software was used for label-free peptide quantification as described before (Alcolea et al. 2012). XICs for all the peptides identified across all samples were constructed with  $\pm 2$  min and  $\pm 7$  ppm retention time and mass windows, respectively. Peak areas from all XICs were calculated. The maximum intensity value for the 2 technical replicates was selected and used for further analysis. Intensity values for each peptide were normalized to total sample intensity. Statistical significance was calculated using two tail unpaired Student's t-test. Multiplicity correction was performed by applying the Benjamini-Hochberg method on the p-values, to control the false discovery rate (FDR). Differences were considered significant when  $FDR < 0.05$ . Proteins with a Mascot score  $> 40$  were used for analysis.

### 2.13 Mass Spectrometry of Nucleosides

#### 2.13.1 Isolation and digestion of genomic DNA

DNA was isolated from pelleted cells using the PureLink Genomic DNA mini kit (ThermoFisher Scientific). 500ng of genomic DNA (3 biological replicates per sample) was digested using DNA Degradase Plus (Zymo Research) by incubating genomic DNA with 1 $\mu$ l DNA Degradase Plus enzyme in a total volume of 25 $\mu$ l at 37° for 12 hours.

#### 2.13.2 Mass Spectrometry quantification of nucleosides

Nucleosides were analysed by LC-MS/MS on a Q-Exactive mass spectrometer (Thermo Scientific) fitted with a nanoelectrospray ion-source (Proxeon). All samples and standards had a heavy isotope-labelled nucleoside mix added prior to mass spectral analysis (2'-deoxycytidine- $^{13}\text{C}_1$ ,  $^{15}\text{N}_2$  (Santa Cruz), 5-(methyl- $^2\text{H}_3$ )-2'-deoxycytidine (Santa Cruz), 5-

(hydroxymethyl)-2'-deoxycytidine-<sup>2</sup>H<sub>3</sub> (Toronto Research Chemicals). MS2 data for 5hmC, 5fC, 5mC and C were acquired with both the endogenous and corresponding heavy-labelled nucleoside parent ions simultaneously selected for fragmentation using a 5 Th isolation window with a 1.5 Th offset. Parent ions were fragmented by Higher-energy Collisional Dissociation (HCD) with a relative collision energy of 10%, and a resolution setting of 70,000 for MS2 spectra. Peak areas from extracted ion chromatograms of the relevant fragment ions, relative to their corresponding heavy isotope labelled internal standards, were quantified against a six-point serial 2-fold dilution calibration curve, with triplicate runs for all samples and standards.

## 2.14 Infinium MethylationEPIC BeadChip assay

Genomic DNA was extracted using the PureLink Genomic DNA mini kit (ThermoFisher Scientific) and quantified using the Nanodrop 2000 spectrophotometer (ThermoFisher Scientific). Bisulphite conversion of DNA was carried out using the Imprint<sup>®</sup> DNA Modification kit (Sigma Aldrich), following the manufacturer's instructions. Infinium MethylationEPIC BeadChip assay (Illumina) was performed according to manufacturer instructions by Barts and the London Genome Centre (London, UK). The Illumina Infinium MethylationEPIC BeadChip assays DNA methylation levels at approximately 850,000 different CpG sites. For each CpG site, methylation levels are measured by probes attached to beads, one each for unmethylated and methylated DNA, followed by allele-specific base extension that includes a fluorescent label. Different labels are used for the T (unmethylated) or C (methylated) alleles. Methylation scores for each CpG site are classified as "Beta" values that range from 0 (unmethylated) to 1 (fully methylated) on a continuous scale, and are based on the ratio of methylated to unmethylated signal outputs. The Bioconductor package ChAMP (version 2.11.3: <https://bioconductor.org/packages/release/bioc/html/ChAMP.html>) (Tian et al. 2017) was used to process raw Infinium idat files using the GRCh37 human genome manifest file, and generate beta values. Data was normalised using the "SWAN" method of normalisation.



## 2.15 RNA sequencing

### 2.15.1 Library Preparation

Total RNA was extracted using Direct-zol RNA mini kit (Zymo) and DNase treated (ThermoFisher Scientific), before being quantified using the Nanodrop 2000 spectrophotometer (ThermoFisher Scientific). Messenger RNA (mRNA) was isolated from 500ng of total RNA using Dynabeads mRNA DIRECT purification kit (ThermoFisher Scientific) and fragmented with RNA fragmentation reagent (ThermoFisher Scientific). First strand cDNA synthesis was performed with SuperScript III First-Strand Synthesis System and 3  $\mu\text{g}$   $\mu\text{l}^{-1}$  random hexamers (ThermoFisher Scientific) followed by second strand synthesis with DNA polymerase I and RNase H. After purification using SPRI beads, the double stranded cDNA was ligated to in house designed adapters (based on TruSeq Indexed adapters (Illumina); listed in Table 2.11) using NEBNext Ultra II (NEB) followed by 15 cycles of amplification with universal P5 (AATGATACGGCGACCACCGAGATCTAC) and P7 (CAAGCAGAAGACGGCATACGAGAT) PCR primers and library purification.

### 2.15.2 Library Quantification and Quality Control

Library size distribution and molarity was assessed by the DNA 1000 assay on the 4200 Bioanalyzer (Agilent), and quantification was performed using the Qubit™ dsDNA high sensitivity assay, measured on the Qubit 4 Fluorometer (ThermoFisher Scientific). Libraries were also quantified using the Kapa Library quantification kit for Illumina (Roche), and final multiplexing of samples was based on concentrations calculated from the Kapa library quantification.

### 2.15.3 Library Sequencing

Sequencing was performed on an Illumina NextSeq500 with 75bp paired-end sequencing at Barts and the London Genome Centre (London, UK). Raw data was received in FASTQ format.

Primer	Primer Sequence (5' – 3') - unique index <u>underlined</u>
P5	AATGATACGGCGACCACCGAGATCTACA <u>CTCTTTCCCTACACGACGCTCTTCCGATC</u> *T
P7	GATCGGAAGAGCACACGTCTGAACTCCAGTCAC <u>CATCACGATCTCGTATGCCGTCTTCTGCTTG</u>
	GATCGGAAGAGCACACGTCTGAACTCCAGTCAC <u>CGATGT</u> ATCTCGTATGCCGTCTTCTGCTTG
	GATCGGAAGAGCACACGTCTGAACTCCAGTCAC <u>TTAGGC</u> ATCTCGTATGCCGTCTTCTGCTTG
	GATCGGAAGAGCACACGTCTGAACTCCAGTCAC <u>TGACCA</u> ATCTCGTATGCCGTCTTCTGCTTG
	GATCGGAAGAGCACACGTCTGAACTCCAGTCAC <u>ACAGTG</u> ATCTCGTATGCCGTCTTCTGCTTG
	GATCGGAAGAGCACACGTCTGAACTCCAGTCAC <u>GCCAAT</u> ATCTCGTATGCCGTCTTCTGCTTG
	GATCGGAAGAGCACACGTCTGAACTCCAGTCAC <u>CAGATC</u> ATCTCGTATGCCGTCTTCTGCTTG
	GATCGGAAGAGCACACGTCTGAACTCCAGTCAC <u>ACTTGA</u> ATCTCGTATGCCGTCTTCTGCTTG
	GATCGGAAGAGCACACGTCTGAACTCCAGTCAC <u>CTTGTA</u> ATCTCGTATGCCGTCTTCTGCTTG

**Table 2.11. Illumina P5 and P7 indexed adapters used for RNA-seq and WGS**

#### 2.15.4 Data analysis

Read quality of FASTQ files was determined using FASTQC. Genomic mapping of short reads was performed using hisat2 (v. 2.1.0) to the human genome (GRCh38). Mapped reads were counted for each sample using FeatureCounts (Subread, v. 1.6.3) (Liao, Smyth, and Shi 2014), with counting performed over all exons. Downstream RNA-sequencing analysis was performed using the R package EdgeR (v3.18.1) (Robinson, McCarthy, and Smyth 2010). Upregulated and downregulated genes were called as those with Benjamini-Hochberg corrected FDR < 0.05 and a log2 fold change > 1. Pathway enrichment analysis was performed using DAVID Bioinformatics Resources (Huang, Sherman, and Lempicki 2009; Huang, Sherman, and Lempicki 2008)

#### 2.16 Targeted sequencing for mutational analysis

##### 2.16.1 DNA isolation

Cells were dissociated to single cells using Accutase and serially plated for 2 hours to eliminate excess iMEFs. DNA was isolated from pelleted cells using the PureLink Genomic DNA mini kit (ThermoFisher Scientific) and quantified using the Nanodrop 2000 spectrophotometer (ThermoFisher Scientific).

### 2.16.2 PCR amplification of target regions

The modified DNA was amplified using loci specific PCR primers with HotStar Taq DNA Polymerase (Qiagen). The following universal Illumina adapter sequences were added to the 5' end of each primer.

Forward primer: ACACTCTTCCCTACACGACGCTCTCCGATCT

Reverse primer: GTGACTGGAGTTCAGACGTGTGCTCTCCGATCT

The PCR conditions were as follows: 95 °C for 15 min; 94 °C for 30 seconds; 56 °C for 30 seconds; 72 °C for 1 min; Repeat steps 2-4 29X; 72 °C for 10 min; Hold 12°C. PCR products were purified using SPRI magnetic beads at a 1:1 ratio (Agencourt AMPure XP, Beckman Coulter). Purified PCR amplicons were PCR amplified with a further 8 cycles using a universal Illumina forward (P5) primer and an indexed reverse (P7) primer (Table 2.9). PCR products were purified using SPRI beads at a 1:1 ratio. Primer sequences are listed in Table 2.12.

### 2.16.3 Library quantification

Purified amplicons were quantified using the Nanodrop 2000 and the Qubit™ dsDNA high sensitivity assay, measured on the Qubit 4 Fluorometer (ThermoFisher Scientific). Amplicons for each sample were pooled at equal concentrations, based on quantification on the Nanodrop 2000 spectrophotometer (ThermoFisher Scientific). Quantification of each sample was performed using the Kapa Library quantification kit for Illumina (Roche), and samples were pooled at equal concentrations to enable multiplexing.

### 2.16.4 Amplicon library Sequencing

Sequencing was performed on an Illumina MiSeq with 150bp paired-end reads, using v3 chemistry, at Barts and the London Genome Centre (London, UK).

### 2.16.5 Data Analysis

Data analysis was performed on the bioinformatics platform Galaxy (<https://usegalaxy.org/>). Sequencing reads were obtained in the form of FASTQ files and quality of reads was assessed using FastQC (Babraham Bioinformatics). Sequencing reads were aligned to the human GRCh38 genome using Bowtie2 with standard parameters. The function samtools mpileup

was used to generate a multi-way pileup file with a maximum read depth of 1,000,000 reads. Varscan (v2.4.2) was used for variant calling from the mpileup file, with a minimum variant allele frequency (VAF) threshold of 0.001 (0.1%). The resulting VCF file was used for downstream analysis which was performed using Microsoft excel. Variants with less than 10,000 reads were eliminated from the analysis.

DNA methylation status during reprogramming	Gene/region	Primer Sequence (5' – 3')
Hypermethylated	DLX5	TAGGGTTATCTGGCCTCCCC
		GCTGTGTTTTTCTGGTGCACT
Hypermethylated	PITX1	GGGGTGTCTGAGATGGACC
		GATTCCCGACCCCGTACAA
Hypermethylated	FGF17	TCCCTCCTCAGTCGTCCAAA
		CCTTGGTAGAGGCGCTTGAT
Non-hypermethylated	WTIP	GGTTGGGACGAGGAAGGT
		AGTGTGCCCATGAACCTGAC
Non-hypermethylated	CAMK2D	TAGGTCTCCTGCCTCCTTCC
		ACGTGTGCATCTTTGCGTTC

**Table 2.12. Primers sequences used for targeted sequencing for mutation analysis**

## 2.17 Whole genome sequencing

### 2.17.1 Library Preparation

DNA was isolated from pelleted cells using the PureLink Genomic DNA mini kit (ThermoFisher Scientific) and quantified using the Nanodrop 2000 spectrophotometer (ThermoFisher Scientific). 3 independent biological replicates per condition were used. 200ng of genomic DNA was fragmented using 2µl of NEBnext® dsDNA fragmentase® (NEB) for 15 minutes at 37°C in a total volume of 20µl. The reaction was stopped with 5µl of 0.5M EDTA pH8.0. Fragmented DNA was cleaned up using SPRI beads at a 1:1 ratio (Agencourt AMPure XP, Beckman Coulter). Sequencing libraries were prepared using the NEBNext Ultra II library preparation kit for Illumina (NEB). Fragmented DNA was subjected to end repair which results in 5' phosphorylation and dA tailing, followed by ligation with 1.5µM of TruSeq adaptors containing unique indexes for each sample (Table 2.11). Size selection of adaptor-ligated DNA was performed by using a 0.9X ratio of SPRI beads to sample, followed by 6 cycles of library amplification with Illumina universal P5 (AATGATACGGCGACCACCGAGATCTAC) and P7 (CAAGCAGAAGACGGCATAACGAGAT) PCR primers and a final library purification with 1:1 SPRI beads.

### 2.17.2 Library Quantification and Quality Control

Library size distribution and molarity was assessed by the DNA 1000 assay on the 4200 Bioanalyzer (Agilent), and quantification was performed using the Qubit™ dsDNA high sensitivity assay, measured on the Qubit 4 Fluorometer (ThermoFisher Scientific). Libraries were also quantified using the Kapa Library quantification kit for Illumina (Roche), and final multiplexing of samples was based on concentrations calculated from the Kapa library quantification.

### 2.17.3 Library Sequencing

Sequencing was performed on an Illumina HiSeq X-ten with 150bp paired-end sequencing at BGI Tech Solutions, Hong Kong. Raw data was received in FASTQ format.

### 2.17.4 Data Analysis

Sequencing reads were obtained in the form of FASTQ files and quality of reads was assessed using FastQC (Babraham Bioinformatics). Sequencing reads were aligned to the human GRCh38 genome using the BWA-MEM algorithm used in Burrows-Wheeler Aligner with standard parameters. The resulting SAM file was processed into a sorted and indexed BAM file using Samtools v1.3, and PCR duplicates were marked using Picard v2.6. BAM files were imported onto the Galaxy Bioinformatics platform and the FreeBayes tool was used to detect genetic variants and produce a Variant Call Format (VCF) file, with a minimum coverage of 1 to call a variant, due to low sample coverage (5x). BCF tools was used to filter out common variants between samples and retain only unique ones. The total number of single nucleotide variants for each sample was counted and normalised to the total read count, and the number of C to T mutations was also counted and normalised to the total read count for each sample. Additionally, the total number of C to T mutations and the number of A to G mutations was counted and a C to T/ G to A ratio was calculated.

## 2.18 Bioinformatics Analysis

### 2.18.1 Bisulfite Sequencing Analysis

Bismark coverage files downloaded from GEO were uploaded into SeqMonk (v1.41.0), where the genomes were binned into 300bp probe windows. Methylation quantitation was carried out using the 'Bisulphite methylation over features' pipeline in SeqMonk (v1.41.0), with a 300bp probe carried forward if it contained at least 5 CpGs each with at least 3 counts. Motif enrichment analysis was performed using the analysis of motif enrichment (AME) tool on the MEME suite (v5.0.4) (Bailey et al. 2009), searching against the human HOCOMOCO (v11 FULL) database. Sequences were scored using the average odds score and motif enrichment calculated using Fisher's exact test.

### 2.18.2 Overlap analysis

Overlap analysis was performed in R using the package regioneR (Version 3.8: <https://bioconductor.org/packages/release/bioc/html/regioneR.html>). Overlap was performed using the 'overlapPermTest' function with 1000 permutations. Random regions were generated for the hg19 genome using the 'circularRandomizeRegions' function. Random loci generation was restricted to loci present in the Illumina EPIC array (for overlaps performed with Illumina EPIC array probes) or to regions with a (G+C) fraction >0.55 and a CpG observed-to-expected ratio >0.6 (for overlaps performed with bisulfite sequencing data). ENCODE and ChromHMM data for the H1 hESC cell line were downloaded from the UCSC genome browser. For ENCODE data, StdPk files were downloaded for each histone modification and genomic coordinates extracted (as BED files) for use in the overlap analysis.

### 2.18.3 TCGA Analysis

Illumina 450K DNA Methylation data spanning 396965 CpGs and 9664 samples was downloaded from the Pan Cancer Atlas (<https://gdc.cancer.gov/about-data/publications/pancanatlas>). All samples from individuals without both a tumour and normal tissue sample were removed. Samples from tumour types with less than 30 individuals were removed. In order to assess only CpGs deemed "bivalent", CpGs outside of regions that showed a peak of H3K27me3 and H3K4me3 in ENCODE H1 hESCs were removed. For this analysis, raw infinium IDAT files from the hESC reprogramming experiment were processed using minfi and normalised via the singlesample Noob method. CpGs used for analysis were

filtered for those that are unmethylated in primed hESCs (mean beta < 0.3). Unmethylated probes were restricted to those CpGs with mean Beta < 0.3 during the primed to naïve transition. Hypermethylated probes were defined using ChAMP, and restricted to those CpGs with  $\Delta\text{Beta} > 0.1$  in the early transition, late transition or naïve state. For the creation of heatmaps, data was first ordered by sample based on mean methylation of all CpGs, and then by CpG based on mean methylation across all samples of every cancer type. Statistical significance was calculated using a paired Wilcoxon test.

## 2.19 Statistical Analysis

Significance testing was performed using Prism (v.7.04) and Student's T-test, one-way ANOVA or two-way ANOVA with Bonferroni post-hoc tests as specified in the Figure legends. Where applicable, data are plotted as mean  $\pm$  SEM. Representative data are shown where experiments were repeated at least twice with similar results.

## Chapter 3. Results 1: The DNA methylation landscape during primed to naïve hESC reprogramming

### 3.1 Introduction

Human embryonic stem cells are derived from the ICM of pre-implantation human blastocysts (Thomson et al. 1998). However paradoxically, their *in vitro* morphology and transcriptional and epigenetic profiles exhibit closer resemblance to murine EpiSCs rather than the more primitive mESCs (Nichols and Smith 2009). Notably, the genomes of conventional hESCs are globally hypermethylated at levels comparable to somatic cells (Hackett and Surani 2014). *In vitro*, two pluripotent states with distinct culture conditions have been defined: primed and naïve pluripotency (Nichols and Smith 2009). In conventional culture conditions, hESCs are dependent on FGF signalling (Xu et al. 2005) and exist as a population of cells heterogeneous for pluripotency markers that are primed for differentiation, with low clonogenicity (Hackett and Surani 2014). Manipulation of *in vitro* culture conditions by a variety of published methods (summarised in Chapter 1) enables the maintenance and survival of more naïve hESCs through reprogramming of primed hESCs. While each method uses a different combination of inhibitors, growth factors or transgenes to reprogram primed hESCs, a feature common to them all is the presence of hLIF and the use of two small molecule inhibitors (2i) which include a MEK inhibitor and a GSK3 $\beta$  inhibitor. Naïve hESCs are characterised by a transcriptional profile more representative of the human ICM and a globally hypomethylated genome, as in the ICM (Takashima et al. 2014; Pastor et al. 2016). Each variety of naïve hESCs has been thoroughly characterised at a molecular level, but the transition from the primed to naïve state, which is likely when molecular processes that mediate the changing cellular state occur, has not been studied. In particular, the dynamics of the large-scale remodelling of DNA methylation that occurs between primed and naïve pluripotent states has not been comprehensively characterised, and the biological context of the DNA methylation changes has not been investigated. Additionally, while there have been brief references to discrete loci in some naïve hESCs that exhibit higher levels of DNA methylation than their primed counterparts (Guo et al. 2017; Guo et al. 2016), this phenomenon has not been explored comprehensively.

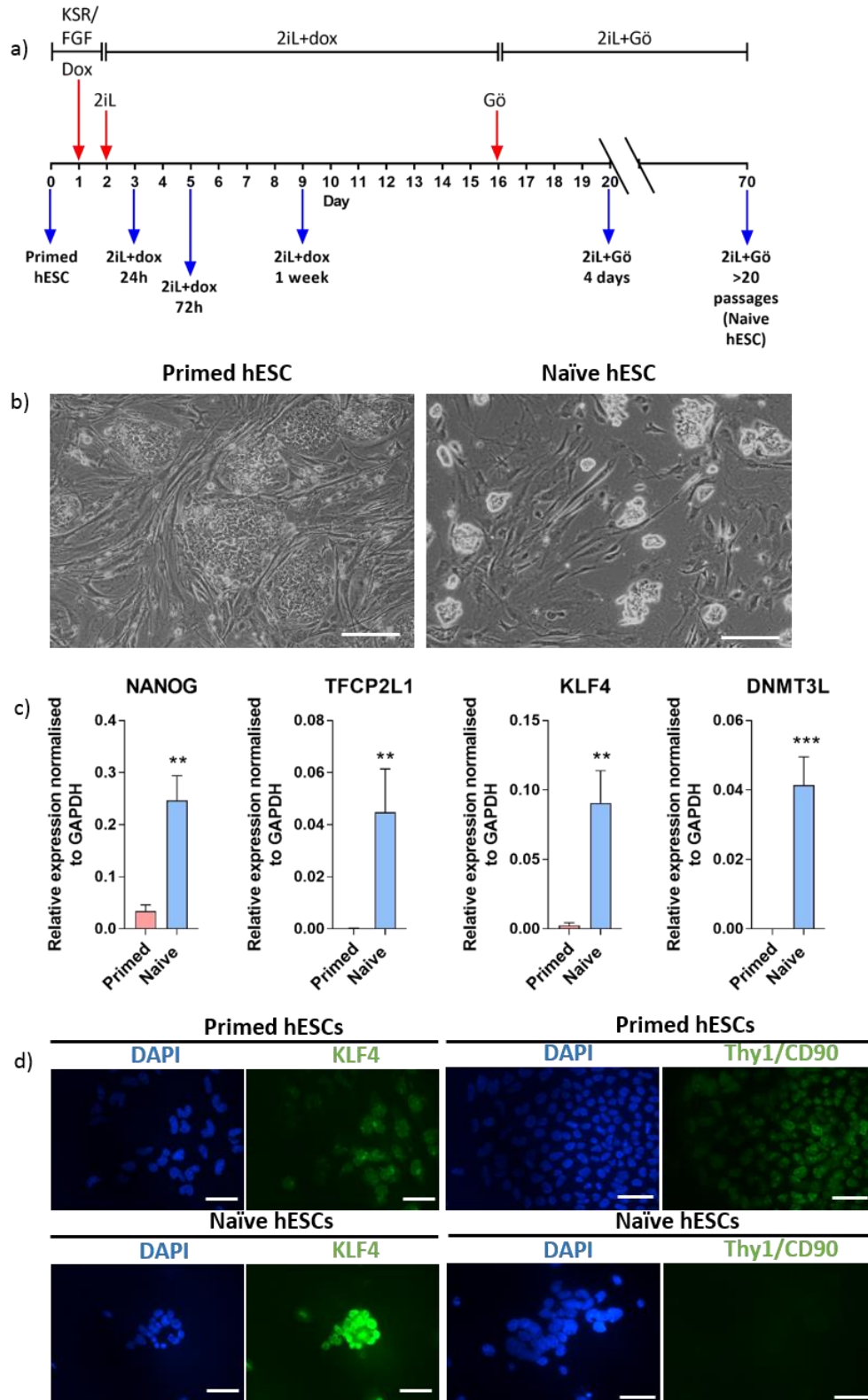
This chapter will explore the changing DNA methylation landscape during primed to naïve hESC reprogramming using the NANOG/KLF2 + 2iLGö method of reprogramming (Takashima et al. 2014). The timing and dynamics of these changes during the transition to naïve pluripotency will be discussed, as well as the genomic context of the methylation changes. Additionally, the



biological significance of the naïve hESC methylation patterns will be described. Finally, parallels between DNA methylation patterns upon reprogramming and methylation patterns frequently observed in cancer will be highlighted, and the possible utility of the primed to naïve reprogramming system as a mechanistic model for aberrant DNA methylation in cancer will be presented.

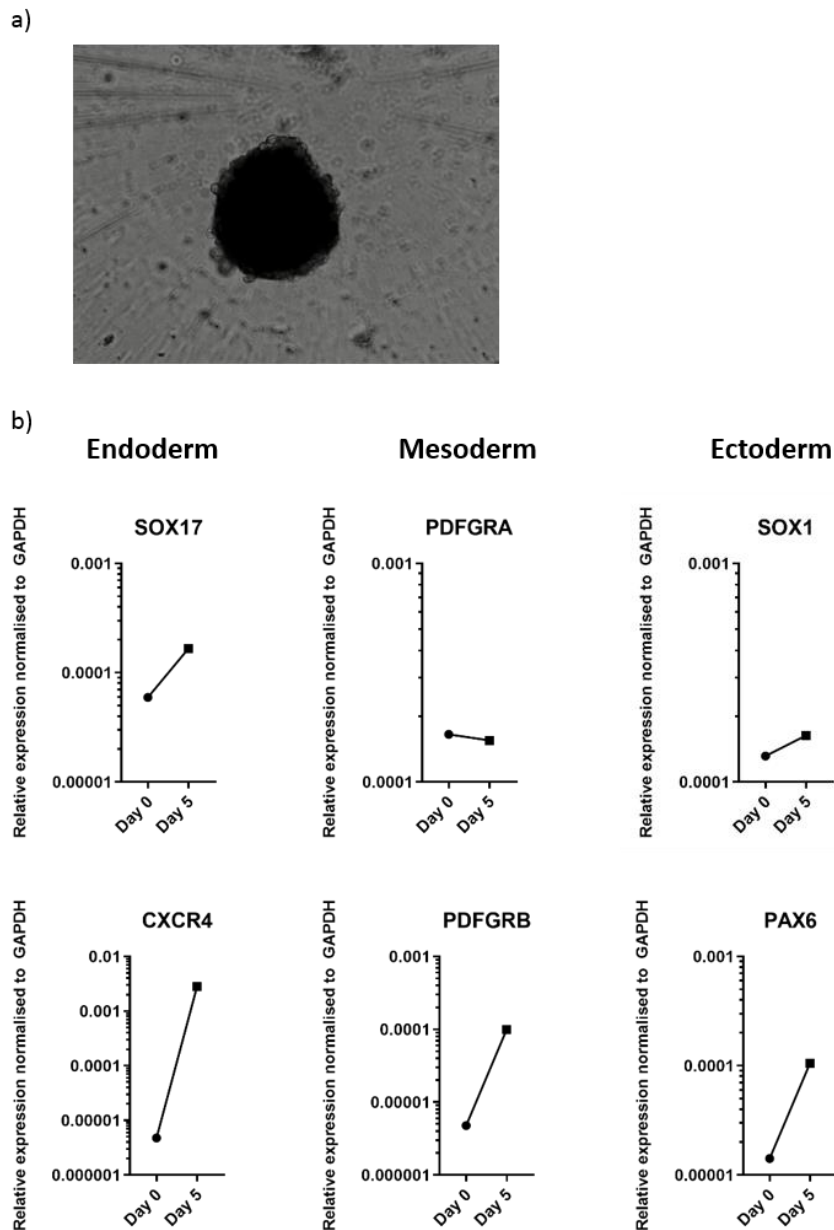
### 3.2 Naïve hESCs are morphologically and transcriptionally distinct from primed hESCs

To validate that reprogramming of primed H9-NK2 hESCs to the naïve state was successful in our hands, we began with primed hESCs cultured in KSR/FGF. We induced overexpression of the *NANOG* and *KLF2* transgenes using doxycycline, alongside the use of the 2i inhibitors in the presence of hLIF (collectively referred to as 2iL+dox). After two weeks, upon removal of doxycycline, we propagated the naïve hESCs in the presence of a PKC inhibitor, Gö, along with the 2i inhibitors (collectively referred to as 2iL+Gö; Figure 3.1a) (Takashima et al. 2014). The primed and naïve hESCs exhibit distinct morphologies as expected, with the primed hESCs growing in large, flattened colonies and the naïve hESCs growing in smaller dome-shaped colonies (Figure 3.1b). We measured the expression of a number of genes that are known to be upregulated upon primed to naïve hESC reprogramming based on previous studies (Takashima et al. 2014; Guo et al. 2017; Guo et al. 2016). The expression of *KLF4*, *NANOG*, *TFCP2L1* and *DNMT3L* was markedly higher in naïve hESCs compared to primed hESCs, characteristic of successful conversion to the naïve state (Figure 3.1c). Additionally, we could detect increased protein expression of *KLF4* in naïve hESCs by immunostaining, as well as reduced expression of the cell surface marker *Thy1/CD90* which is a marker of primed hESCs (Collier et al. 2017) (Figure 3.1d).



**Figure 3.1. Naïve hESCs are morphologically and transcriptionally distinct from primed hESCs.** a) The timeline of reprogramming from primed to naïve hESCs. b) Brightfield images of primed and naïve hESCs showing their different morphologies. Scale bars represent 125µm. c) qRT-PCR for naïve hESC markers in primed and naïve hESCs. Bars are representative of the mean of three technical replicates and three independent biological replicates. Error bars represent the standard error of the mean. Statistical difference between samples was analysed by a two-tailed Student's t-test. \*\* $P < 0.01$  and \*\*\* $P < 0.001$ . Human GAPDH was used to normalise expression. d) Immunostaining for KLF4, a naïve pluripotency marker, and THY1, a primed pluripotency marker. Scale bars represent 50µm.

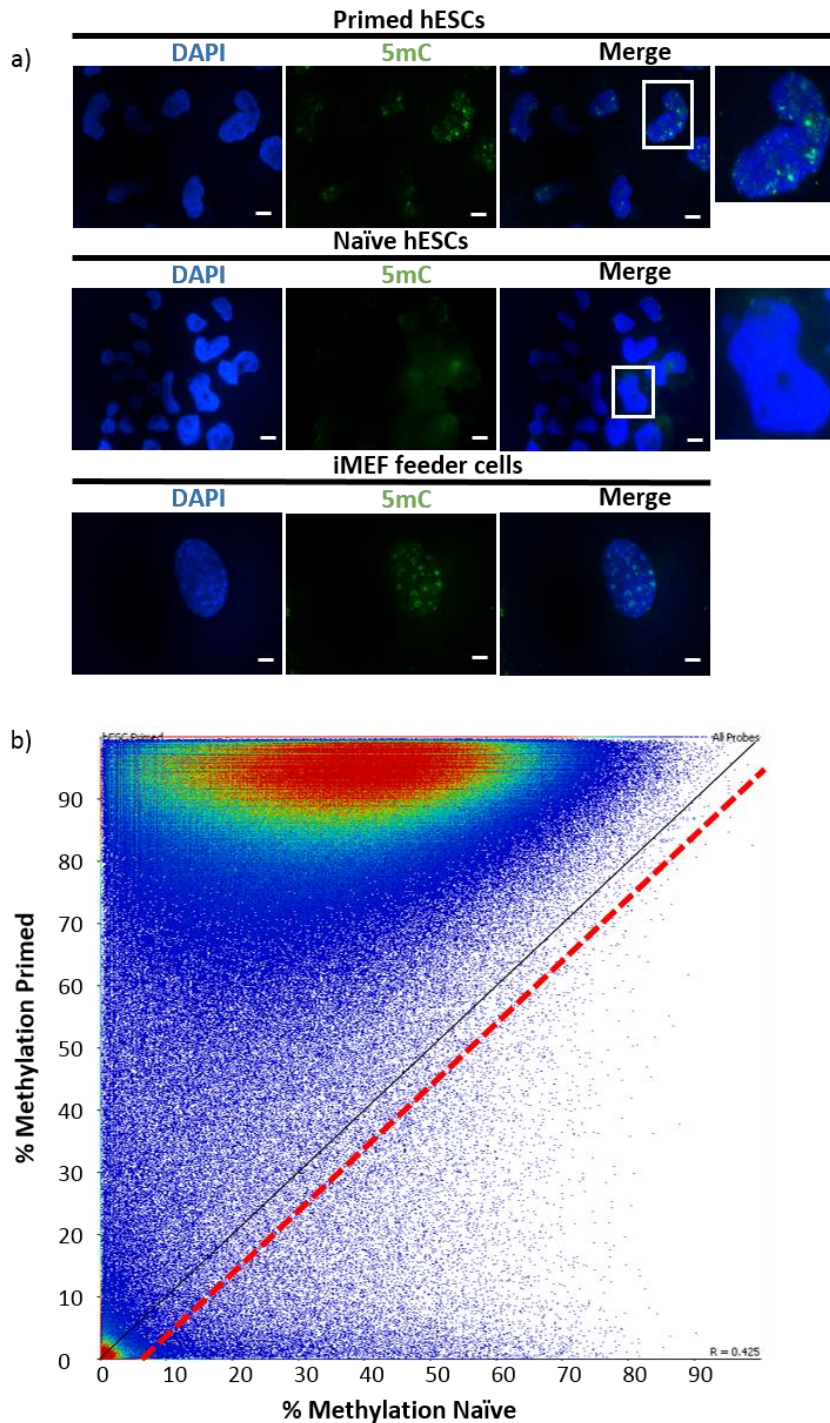
It has been previously demonstrated that naïve hESCs retain their differentiation potential (Takashima et al. 2014). To validate that this was also the case in our hands, we generated embryoid bodies directly from naïve hESCs. A representative embryoid body is displayed in Figure 3.2a. Embryoid bodies were harvested after 5 days and expression of transcripts representative of the three germ layers was measured. In line with published data, we saw upregulation of transcripts associated with the three germ layers (Figure 3.2b), indicative of retained differentiation capacity.



**Figure 3.2. Naïve hESCs can differentiate into all three germ layers.** a) Representative brightfield image of an embryoid body at day 5, taken at 40X magnification. b) qRT-PCR for markers of the endoderm, mesoderm and ectoderm in naïve hESCs (day 0) and embryoid bodies generated from naïve hESCs, harvested after 5 days. Data points are representative of the mean of three technical replicates (with RNA pooled from multiple embryoid bodies). Human GAPDH was used to normalise expression.

### 3.3 Naïve hESCs are characterised by global DNA hypomethylation and regional hypermethylation

It is well established that the primed to naïve hESC reprogramming is characterized by global demethylation of the genome (Takashima et al. 2014; Pastor et al. 2016), more closely resembling the hypomethylated ICM of human pre-implantation blastocysts (Guo et al. 2014; Smith et al. 2014), albeit with the additional loss of methylation at imprinted regions (Pastor et al. 2016). To confirm that this was also true in our hands, we performed immunostaining of 5mC in primed and naïve hESCs and observed a marked reduction in signal in naïve hESCs (Figure 3.3a). We also re-analysed published whole genome bisulfite-sequencing (WGBS) data for primed and naïve hESCs (Takashima et al. 2014). We divided the genome into 300 base pair (bp) regions and performed a comparison between the mean methylation of the three replicates of primed and naïve samples. We observed that while the majority of the genome is hypomethylated in naïve hESCs, this is accompanied by distinct regions that are hypermethylated in naïve hESCs compared to primed hESCs (Figure 3.3b). We identified 26,625 regions (300bp each) that are hypermethylated (>5% increase in methylation) in naïve hESCs. Interestingly, this pattern of global hypomethylation along with regional hypermethylation is characteristic of cancer cells, with respect to their normal counterparts (Ehrlich 2002).

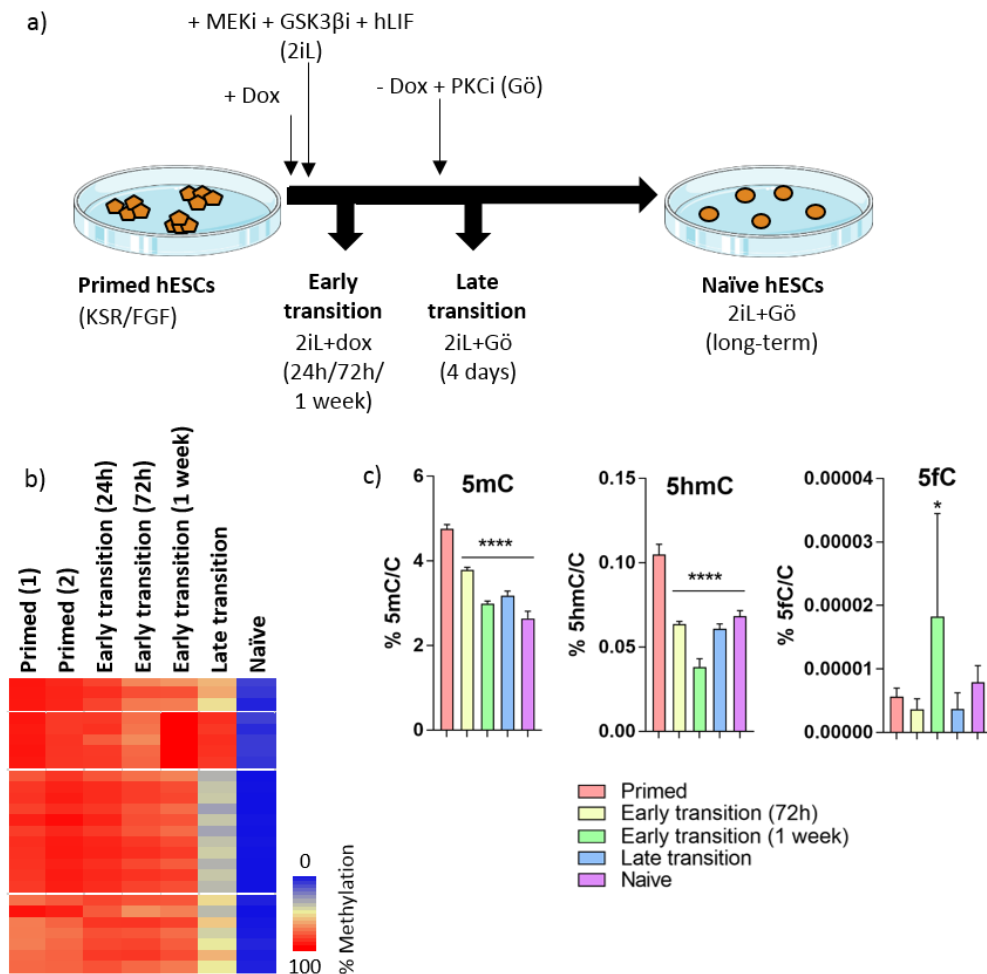


**Figure 3.3. Naïve hESCs are characterised by global DNA hypomethylation and regional hypermethylation.**

a) Immunostaining with an antibody against 5mC shows markedly reduced staining in the naïve cells compared to the primed cells. iMEF feeder cells are used as a positive control as they display a characteristic pattern of 5mC staining. Scale bars represent 10µm. b) Scatter plot showing re-analysed published whole genome bisulfite sequencing (WGBS) data (Takashima et al. 2014). Naïve hESCs display global hypomethylation compared to primed hESCs, along with regional hypermethylation. Each dot represents a 300 base pair genomic region and depicts the % methylation of the region in primed hESCs (y-axis) and naïve hESCs (x-axis). Red line represents the 5% threshold selected to identify hypermethylated regions in naïve hESCs.

### 3.4 Global demethylation occurs gradually during the primed to naïve transition

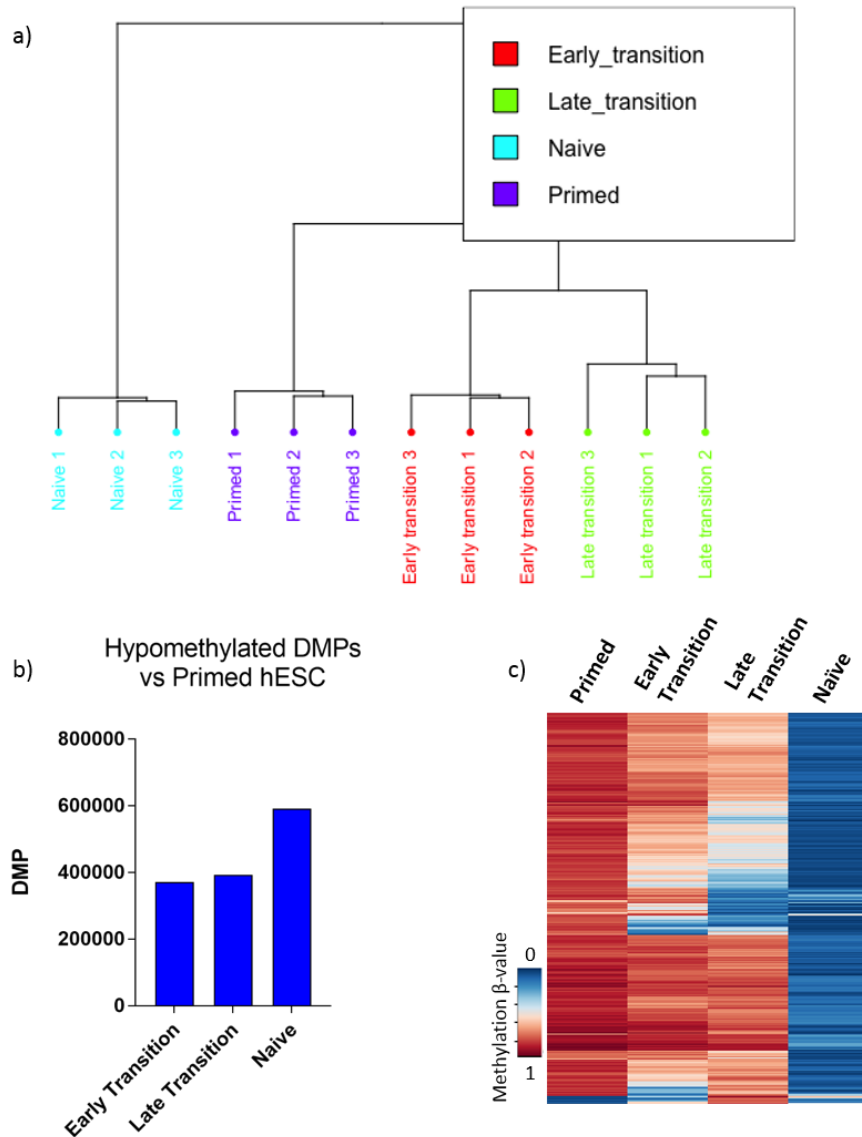
While primed and naïve hESCs have been studied thoroughly, the transitional period between the two states is not well characterised. As this intermediate period is likely when the changes in DNA methylation occur, it was in our interest to characterise the dynamics of these changes. We designed a time-course experiment to study this transition, capturing primed and naïve hESCs as well as the transition between the two states, termed 'early transition' and 'late transition' as the cells are in 2iL+dox or 2iL+Gö, respectively (Figure 3.4a). We employed a number of approaches to measure DNA methylation upon reprogramming of primed hESCs to the naïve state. We first used a targeted approach to measure DNA methylation at a number of regions that we had identified as being hypomethylated (>5% decrease in methylation) in naïve hESCs from the WGBS data (Figure 3.3b) and in line with previous studies, we were able to validate loss of methylation at these regions during primed to naïve reprogramming (Figure 3.4b). We next used mass spectrometric measurement of nucleosides as a more global measure of DNA methylation during the primed to naïve transition. We detected progressive global DNA demethylation of the genome in naïve cells, as measured by mass spectrometry of the 5mC nucleoside (Figure 3.4c). 5mC in the genome can undergo active demethylation; a process that begins with oxidation of 5mC by the TET proteins into 5hmC (Tahiliani et al. 2009; Kriaucionis and Heintz 2009), which can be further oxidised to 5fC (Ito et al. 2011). Alongside the reduction in 5mC, we detected globally reduced levels of 5hmC upon reprogramming, which is also a common phenomenon in the genomes of cancer cells (Ficz and Gribben 2014). Intriguingly, we were also able to detect a significant increase in the global level of 5fC during the early transition of reprogramming. This may be indicative of ongoing TET activity, and a lack of base excision by TDG DNA glycosylase, suggesting that the accumulated 5fC may play a functional role during reprogramming.



**Figure 3.4. Global demethylation occurs gradually during the primed to naïve transition.** a) Schematic detailing the model system and time points used in the study. hESCs; human embryonic stem cells, KSR/FGF; knockout serum replacement/fibroblast growth factor, dox; doxycycline. b) Targeted bisulfite-sequencing of four genomic regions. Each square represents the methylation % indicated by the colour key of a single CpG. c) Mass spectrometry analysis of 5-methylcytosine (5mC) and 5-hydroxymethylcytosine (5hmC) at the time points analysed, displayed as a % of total genomic cytosine (C) content. Data shown is representative of two biological replicates, each with three technical replicates. Statistical difference between samples was analysed by a one way ANOVA test with a Bonferroni post-hoc test of each sample compared to primed hESC. \* $P < 0.05$  and \*\*\*\* $P < 0.0001$ .

To further investigate global DNA methylation patterns during primed to naïve reprogramming at higher resolution, we carried out two independent reprogramming experiments and analysed DNA methylation using the Illumina Infinium MethylationEPIC BeadChip (each experiment with 2 or 3 replicates – independent populations of cells - per time point). The Illumina Infinium MethylationEPIC BeadChip allows determination of DNA methylation levels at single-CpG resolution of approximately 850,000 different CpG sites in the human genome, with the methylation score for each CpG classified as a  $\beta$ -value, which can range from 0 to 1, corresponding to 0 to 100% methylation per CpG. Our samples clustered together by time point, indicative of reproducible changes in DNA methylation upon hESC reprogramming

(Figure 3.5a). We used differential methylated probe (DMP) analysis to compare the methylation  $\beta$ -value of individual CpG probes at each stage of the transition and in naïve hESCs compared to primed hESCs. We identified CpGs that are hypomethylated at each time point compared to primed hESCs. Over 500,000 CpGs exhibit reduced levels of methylation upon reprogramming (Figure 3.5b), of which the top 20,000 variably methylation CpGs are displayed in Figure 3.5c. Collectively, these data show that global demethylation upon reprogramming occurs gradually across the transition to naïve pluripotency.

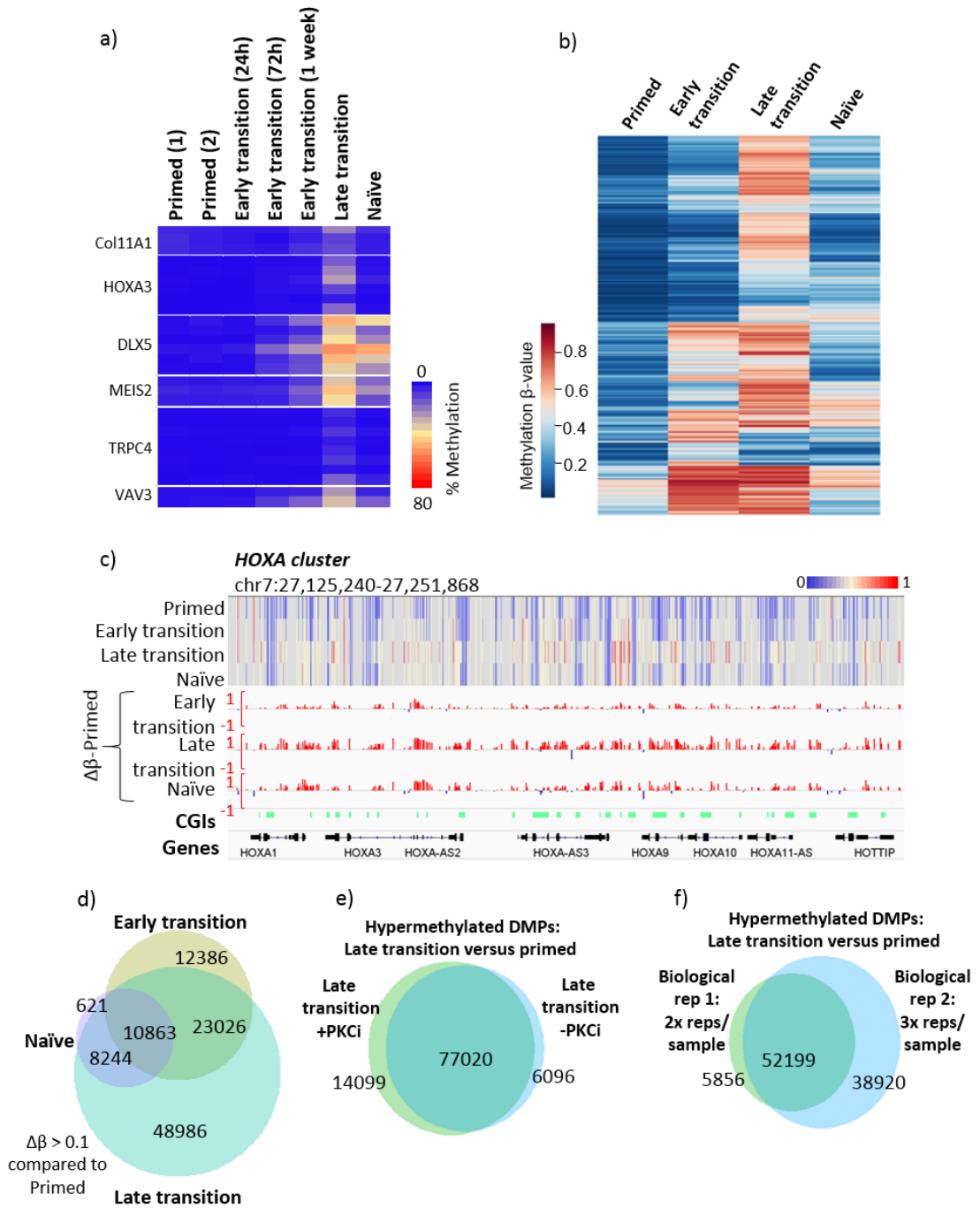


**Figure 3.5. Global DNA demethylation can be detected at an individual CpG level during primed to naïve reprogramming.** a) Dendrogram of Infinium MethylationEPIC BeadChip samples for pre-normalised data. b) Bar plot showing the number of CpGs hypomethylated ( $\Delta\beta$  vs Primed < -0.1, adjPval < 0.05) at each stage of reprogramming compared to primed hESCs c) Heatmap showing methylation levels of the top 20,000 variably methylated CpG probes across all samples. Methylation  $\beta$ -value is indicated by the colour key.



### 3.5 Reprogramming-associated hypermethylation is non-random and reproducible

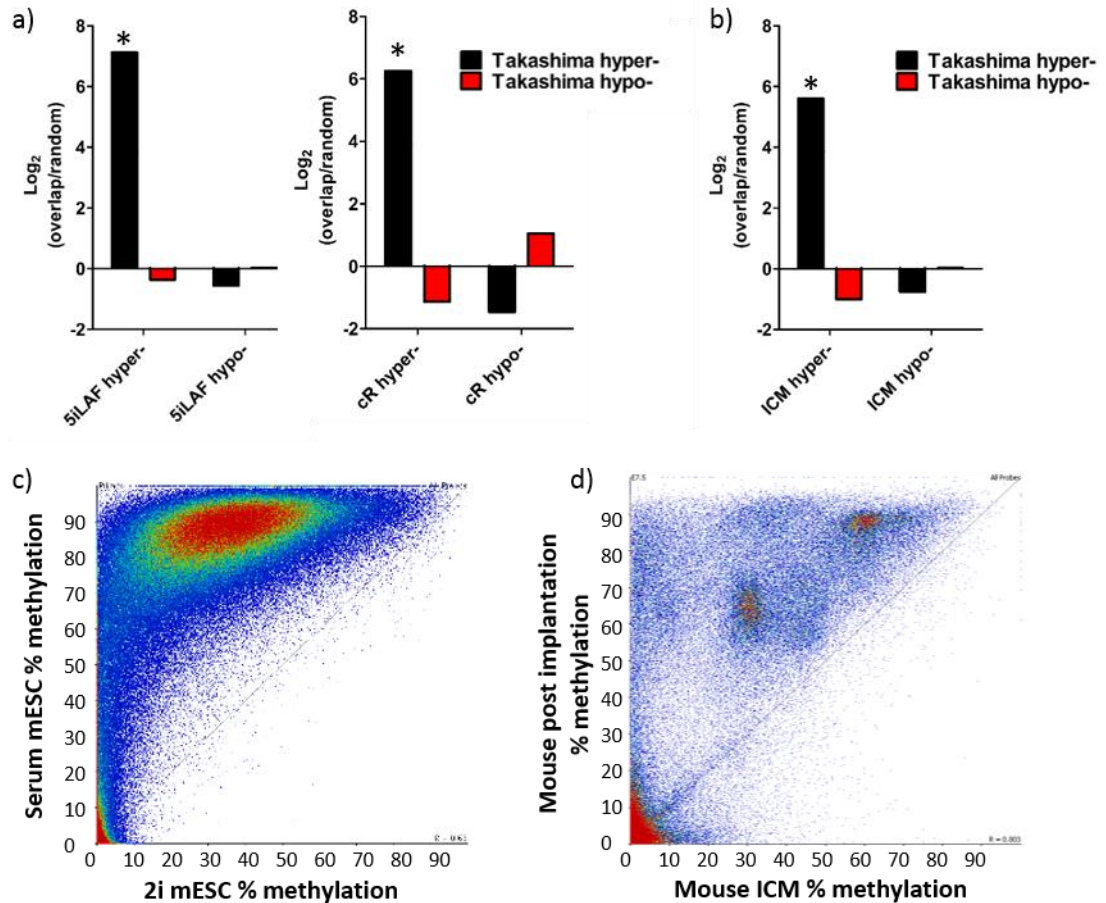
To investigate the dynamics of hypermethylation during primed to naïve reprogramming, we first used a targeted bisulfite-sequencing approach to measure DNA methylation at a number of regions that we had identified as being hypermethylated in naïve hESC from the WGBS data (Takashima et al. 2014) (Figure 3.3b). We measured DNA methylation at these regions across the transition to naïve pluripotency to determine at what stage hypermethylation occurs. We found that hypermethylation was detectable at some regions during the early transition, with a peak in the level and extent of hypermethylation during the late transition, following which hypermethylation was maintained at some regions and lost at others, indicative of a transitory nature of methylation at some genomic loci (Figure 3.6a). Using the Infinium MethylationEPIC BeadChip data and performing individual probe based analysis, we were able to confirm this pattern of hypermethylation at a genome-wide level and identify DMPs that were hypermethylated at each stage of reprogramming compared to primed hESCs. Of these, the top 10,000 hypermethylated DMPs at each time point are displayed in Figure 3.6b and a genome browser screen shot of the HOXA cluster is displayed in Figure 3.6c, exemplifying the phenomenon of hypermethylation upon reprogramming. A large proportion of the hypermethylated DMPs at each time point compared to primed hESC are shared (Figure 3.6d), though many are also uniquely hypermethylated at the late transition. This peak of hypermethylation that occurs during the late transition is measured after 4 days of removing dox and being in 2iL+Gö (18 days after the induction of reprogramming). This raised the possibility that the addition of the PKC inhibitor, Gö, may intensify or accelerate the process of hypermethylation. To verify whether this was the case, we cultured cells until the late transition of reprogramming in the presence and absence of the PKCi. We found that when compared to primed hESCs, hypermethylation of the same sites occurs in both conditions (Figure 3.6e), implying that the addition of the PKCi does not impact hypermethylation and that there is a time dependent accrual of DNA methylation instead. The reproducibility of the hypermethylation during the reprogramming process is apparent from the strong overlap between hypermethylated DMPs across biologically independent MethylationEPIC arrays (with 2 or 3 cell populations assayed within each array), suggesting that the site-specific gain in methylation is unlikely to be the result of a stochastic process (Figure 3.6f). This points towards a defined cellular mechanism controlling hypermethylation, which may have a biological function.



**Figure 3.6. Reprogramming-associated hypermethylation is non-random and reproducible.** a) Targeted bisulfite-sequencing of six genomic regions. Each square represents the methylation % indicated by the colour key of a single CpG. b) Heatmap showing methylation levels of the top 10,000 hypermethylated CpG probes at each time point compared to primed hESC. Methylation  $\beta$ -value is indicated by the colour key. c) Genome browser tracks for Infinium MethylationEPIC BeadChip data capturing a representative hypermethylated locus. The heatmap shows the raw methylation  $\beta$ -values per CpG for each sample, while the subsequent rows show the per-probe difference in methylation for each time point of reprogramming compared to primed hESCs. CpG islands (CGIs) are highlighted in green. d) Venn diagram showing the overlap of hypermethylated DMP at each stage of reprogramming compared to Primed hESC. e) Venn diagram showing the overlap of late transition hypermethylated DMP in the presence or absence of PKCi (Gö). f) Venn diagram showing the overlap of hypermethylated DMP between the late transition and primed hESCs between two biologically independent Infinium Methylation EPIC BeadChip arrays. DMP, differentially methylated probe. Hypermethylated CpG/DMP: ( $\Delta\beta$  vs Primed > 0.1, adjPval < 0.05). AdjPval is based on Benjamini-Hochberg adjustment.

### 3.6 Hypermethylation is a feature of the naïve human pluripotent state

As we were using an *in vitro* system of reprogramming, it was important to ensure that the hypermethylation we had observed was not simply an artefact of the NANOG/KLF2 + 2iLgö *in vitro* system. To address this, we compared our naïve hypermethylated and hypomethylated regions identified from the published WGBS data (Takashima et al. 2014) to hypermethylated and hypomethylated regions identified in naïve cells generated using alternative methods of primed to naïve hESC reprogramming. The first of these methods involves the generation of naïve cells by transferring primed hESCs to a media containing a cocktail of five inhibitors alongside LIF, Activin, and/or Fibroblast Growth Factor 2 (5iLAF) (Theunissen et al. 2014; Pastor et al. 2016). The second method involves the generation of naïve cells by chemical resetting (cR) (Guo et al. 2017), whereby primed hESCs are transiently exposed to histone deacetylase inhibitors before being propagated in 2iL+Gö. We saw a significant overlap between the hypermethylated regions in each of these data sets compared to the 2iLgö naïve cells (Figure 3.7a). The parallels between the three *in vitro* reprogramming systems suggest that hypermethylation is not simply an artefact of the NANOG/KLF2 + 2iLgö reprogramming method. We next compared our naïve hypermethylated regions to regions that are hypermethylated in the human ICM compared to the post-implantation embryo (Guo et al. 2014). Once again, we saw a significant overlap between the two data sets (Figure 3.7b) indicating that the hypermethylation we observed *in vitro* recapitulates the *in vivo* relationship between the ICM and the post-implantation embryo. Mouse embryonic stem cells (ESC) can also be cultured in naïve conditions with media containing the 2i inhibitors, but interestingly, when we performed a similar analysis of WGBS data for 2i ESCs and conventional ESCs cultured in serum (Ficz et al. 2013), we did not detect any hypermethylation in the mouse 2i ESCs (Figure 3.7c). Additionally, analysis of *in vivo* mouse developmental WGBS data also indicated that while there is a small amount of hypermethylation present in the mouse ICM compared to the mouse post-implantation epiblast (Figure 3.7d), it is far less extensive than the hypermethylation we see upon NANOG/KLF2 + 2iLgö hESC reprogramming (Figure 3.3b). Overall from these results, we can conclude that this hypermethylation pattern is a feature unique to human cells and is characteristic of *in vitro* naïve human pluripotency and the *in vivo* human ICM, suggesting that it may play a critical role in the biology of naïve human pluripotent stem cells.

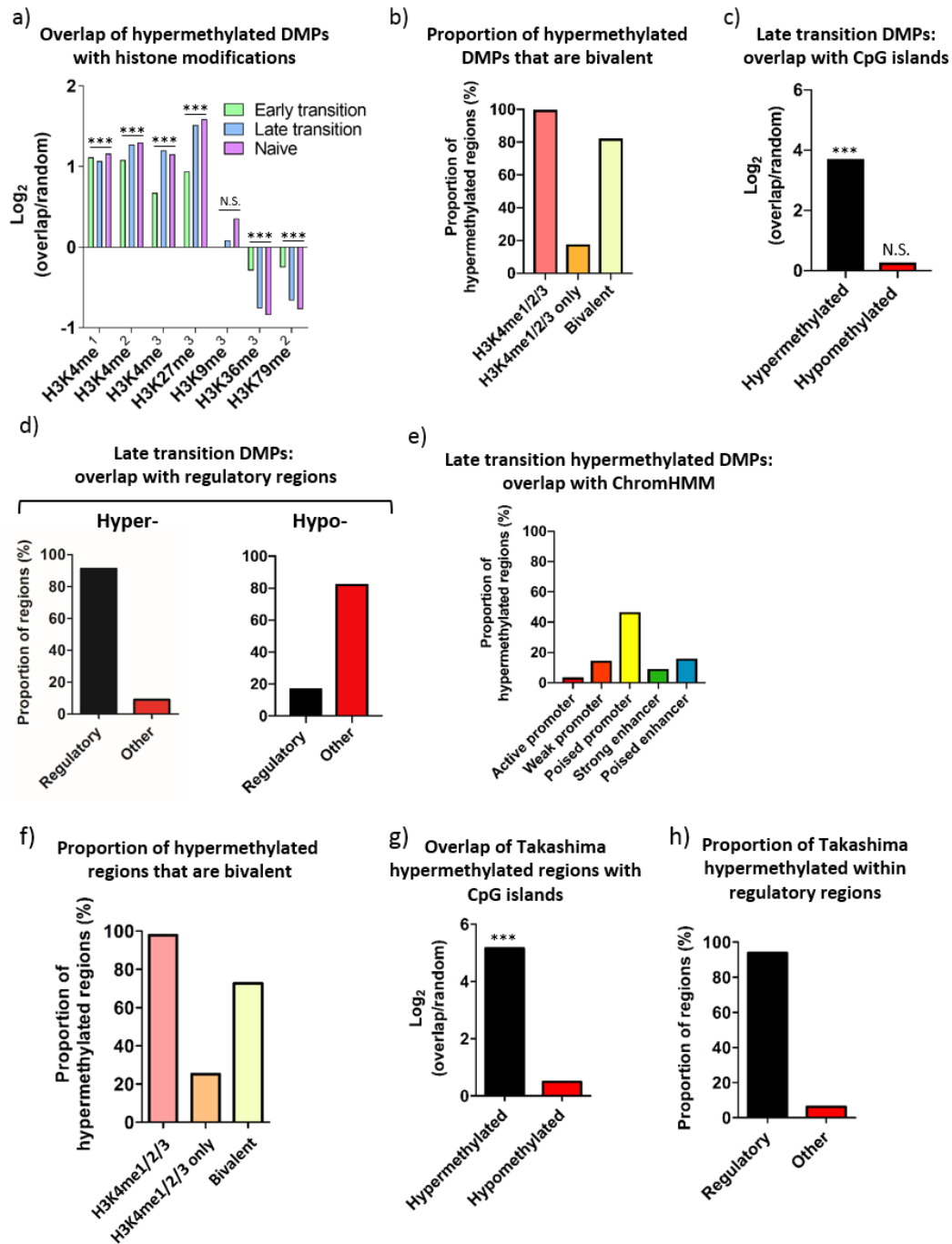


**Figure 3.7. Hypermethylation is a feature of the naïve human pluripotent state.** a) Overlap of hypomethylated and hypermethylated (26,625) 300bp regions identified from published WGBS data (Takashima et al. 2014) with naïve hypomethylated and hypermethylated (13,945) 300bp regions identified from published data for the 5iLAF resetting method (Pastor et al. 2016) and with naïve hypomethylated and hypermethylated (34,859) 300bp regions identified from published data for the chemical resetting method (Guo et al. 2017). b) Overlap of naïve hypomethylated and hypermethylated (26,625) 300bp regions from published WGBS data (Takashima et al. 2014) with hypomethylated and hypermethylated (13,400) 300bp regions identified in human ICM compared to post-implantation embryo from published data (Guo et al. 2014). Data is presented as the log<sub>2</sub> corrected fold increase in the observed overlap compared to the mean overlap of 1000 randomly generated regions. \* indicates  $p < 0.05$  c) Scatter plot showing methylation % for 300bp genomic regions in mouse ESCs cultured in serum or 2i (Ficz et al. 2013). d) Scatter plot showing methylation % for 300bp genomic regions in mouse ICM vs mouse post-implantation blastocyst (Smith et al. 2012).

### 3.7 Reprogramming-associated hypermethylation occurs primarily at bivalent CGIs

Using hypermethylated probes identified from the Infinium MethylationEPIC array, we sought to characterize the genomic features and locations of CpG loci that gain DNA methylation during primed to naïve hESC reprogramming. To gain a deeper insight into the underlying genomic and chromatin context of CpG hypermethylation, we performed a regional overlap analysis of reprogramming-associated hypermethylated DMPs with datasets from the Encyclopaedia of DNA elements (ENCODE), which contains CHIP-seq data for various histone

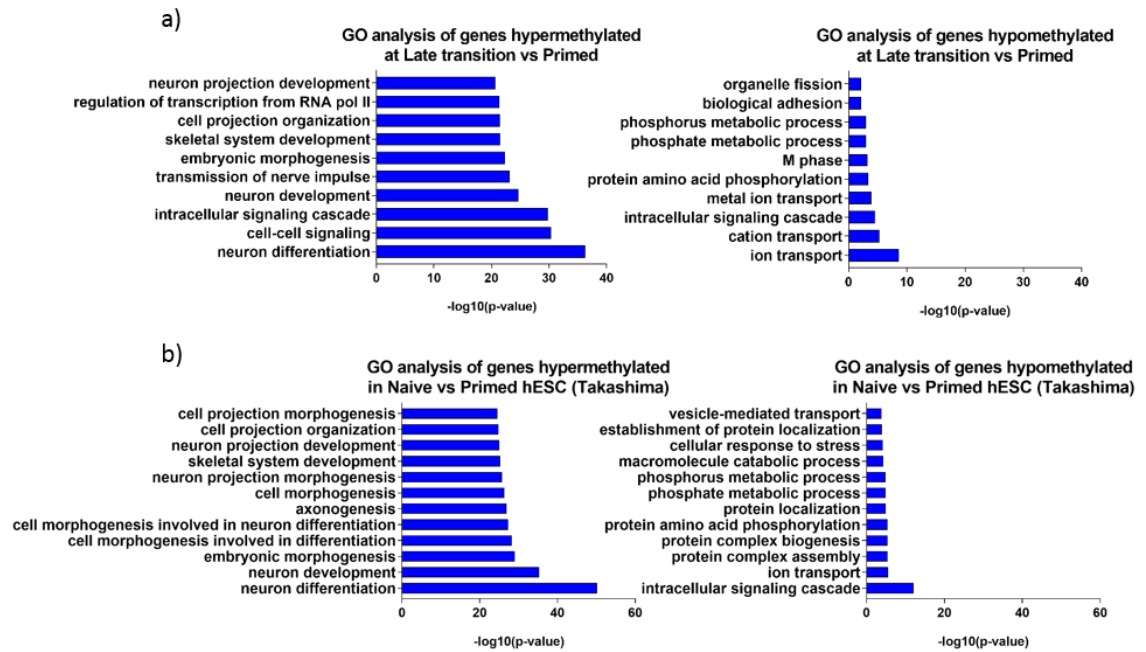
modifications in the primed H1 hESC cell line. We utilized pre-defined ChIP-seq peaks from ENCODE for each histone modification. We observed that hypermethylated probes at each stage of reprogramming are enriched within regions marked by H3K4me1/2/3 and H3K27me3 in primed hESCs, whereas no significant enrichment was observed for other histone modifications (Figure 3.8a). The majority of the hypermethylated probes fall within regions marked by bivalent histone modifications, defined by co-occurrence of H3K4me3 and H3K27me3 in the H1 hESC cell line (Figure 3.8b), and the remainder are marked by H3K4me3 alone. Bivalent regions typically mark regulatory regions (Bernstein et al. 2006) and to this end, we saw a striking overlap of hypermethylated DMPs with CGIs and regulatory regions of the genome, which encompass promoters and enhancers (Bernstein et al. 2006) (Figures 3.8c and 3.8d). We further compared our hypermethylated probes to the ChromHMM model, which uses genome-wide data to annotate the genome and enables functional interpretation of chromatin states (Ernst and Kellis 2017). This analysis showed that the majority of the hypermethylated probes reside within ChromHMM predicted poised promoters (Figure 3.8e) – again pointing to a relationship between hypermethylation with an existing bivalent histone signature; an association that has been observed previously in relation to hypermethylation in cancer (Ohm et al. 2007; Ohm et al. 2010; Easwaran et al. 2012). We next used the 300bp naïve hypermethylated regions identified from the WGBS data to validate our observations. Here too, we observed enrichment of the naïve hypermethylated regions within loci marked by bivalent histone modifications in primed hESCs (Figure 3.8f). Additionally, we observed that these hypermethylated regions were highly enriched within regulatory regions and CGIs (Figures 3.8g and 3.8h), as we saw with the Illumina MethylationEPIC array data, thus strongly reinforcing the highly reproducible nature of the DNA hypermethylation that occurs upon reprogramming to naïve human pluripotency.



**Figure 3.8. Reprogramming-associated hypermethylation occurs primarily at bivalent CGIs.** a) Overlap of hypermethylated probes at each stage of reprogramming with regions of histone modification enrichment (obtained from the ENCODE ChIP-seq data for hESC cell line H1). b) The proportion of hypermethylated probes that overlap H3K4me<sub>1/2/3</sub>, divided further into those that are enriched for H3K4me<sub>1/2/3</sub> alone or bivalent regions (marked by H3K4me<sub>3</sub> and H3K27me<sub>3</sub>). c) Overlap of late transition hypermethylated probes with CpG islands. d) The proportion of hypermethylated probes that overlap with ChromHMM regulatory regions (promoters or enhancers as defined in the hESC cell line H1). e) Overlap of late transition hypermethylated probes with ChromHMM predicted promoter and enhancer sub-categories. f) The proportion of hypermethylated regions (>5% methylation in naïve vs primed hESC) identified from published WGBS data (Takashima et al. 2014) that overlap H3K4me<sub>1/2/3</sub>, divided further into those that are enriched for H3K4me<sub>1/2/3</sub> alone or bivalent regions (marked by H3K4me<sub>3</sub> and H3K27me<sub>3</sub>). g) Overlap of Takashima hypermethylated regions with CpG islands. h) The proportion of Takashima hypermethylated regions that overlap with ChromHMM regulatory regions (promoters or enhancers as defined in the hESC cell line H1). DMP, differentially methylated probes. Where appropriate, data is presented as the log<sub>2</sub> corrected fold increase in the observed overlap compared to the mean overlap of 1000 randomly generated loci.

### 3.8 Reprogramming-associated hypermethylation is associated with developmental genes

In order to understand the genomic context of the hypermethylation observed upon reprogramming, we next used DAVID bioinformatics software to perform gene ontology (GO) analysis of the genes that exhibit reprogramming-associated hypermethylation (Huang, Sherman, and Lempicki 2008). We classified a gene as hypermethylated if it possessed a hypermethylated DMP within 1500bp upstream of its transcription start site (i.e. classical promoter). GO analysis revealed an extensive enrichment of hypermethylated genes in developmental pathways, particularly pathways involved in neuronal development, compared to hypomethylated genes which showed much weaker enrichment in pathways involved in ion transport and metabolism (Figure 3.9a). We validated this observation using the 300bp regions identified from the WGBS data. Here, we took the nearest overlapping gene to each hypermethylated or hypomethylated region and performed GO analysis. Once again, GO analysis revealed an extensive enrichment of hypermethylated genes in developmental pathways, compared to hypomethylated genes which show much weaker enrichment in pathways involved in ion transport and metabolism (Figure 3.9b). Interestingly, this observation draws parallels to the cancer methylome, where it has been observed that DNA hypermethylation is enriched in developmental genes, notably neuronal development (Easwaran et al. 2012).



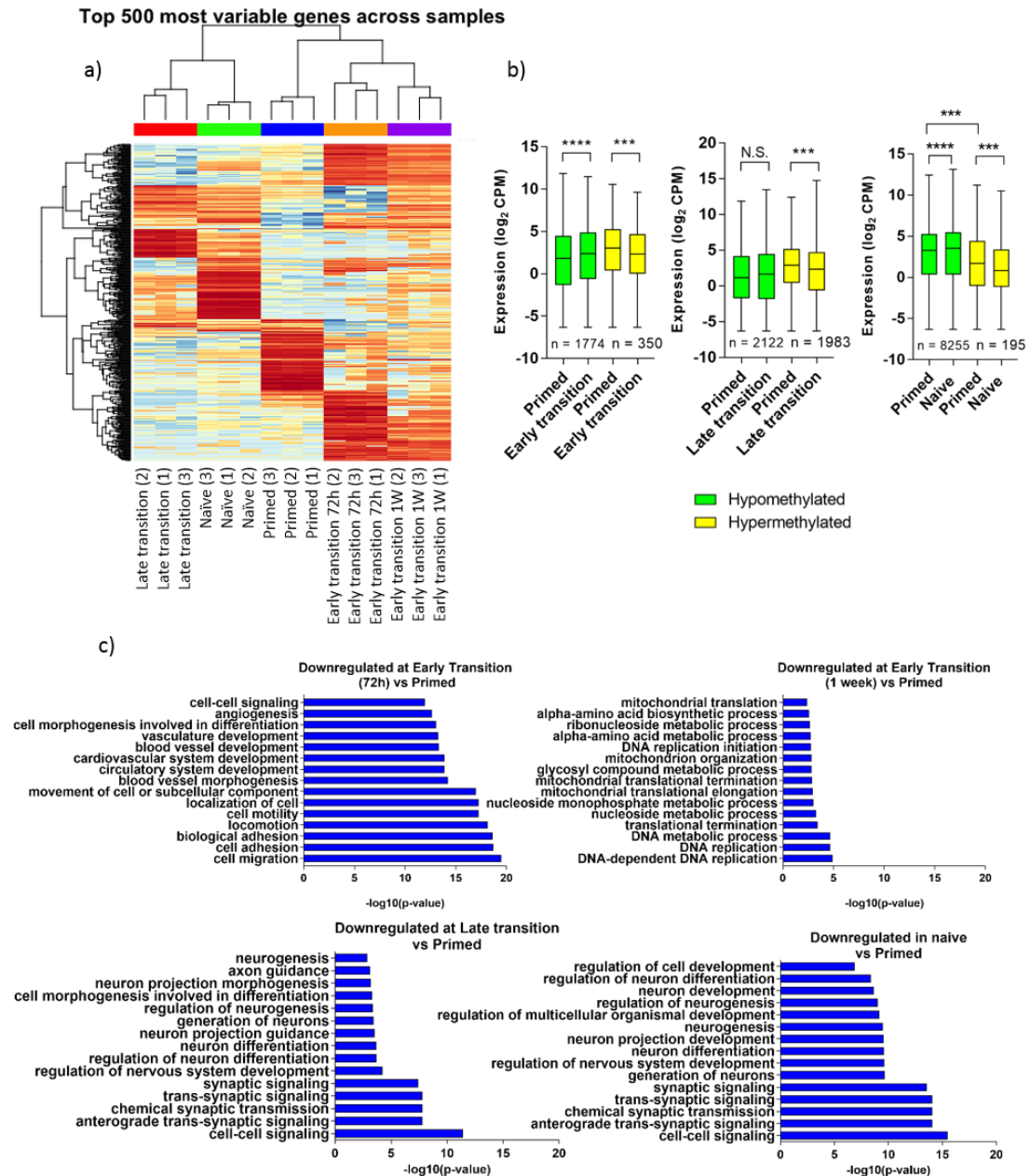
**Figure 3.9. Reprogramming-associated hypermethylation is associated with developmental genes.** a) GO term analysis of hypermethylated and hypomethylated genes at the late transition of reprogramming compared to primed hESCs. A gene was classified as hypermethylated if a hypermethylated probe was present within 1500bp upstream of the transcriptional start site. b) GO term analysis of hypermethylated and hypomethylated genes in naïve hESCs compared to primed hESCs. A gene was classified as hypermethylated or hypomethylated based on the closest overlapping gene to the 300bp regions identified as hypermethylated or hypomethylated from WGBS published data (Takashima et al. 2014). GO terms plotted are the most highly enriched biological processes with FDR < 0.05.

### 3.9 Genes from developmental pathways are hypermethylated and downregulated during reprogramming

Typically, DNA hypermethylation of gene promoters is associated with gene repression (Weber et al. 2007). To investigate whether this is true during reprogramming, we performed temporal transcriptome analysis of cells during the transition from primed to naïve pluripotency. Our RNA-seq samples clustered together by time point, once again indicative of reproducible processes occurring during hESC reprogramming (Figure 3.10a). We observed large numbers of genes differentially expressed at each stage of the transition compared to primed hESCs (Figure 3.10a). To determine if DNA methylation changes upon reprogramming had any impact upon gene expression, we used the Infinium MethylationEPIC data to compare the average promoter methylation (average  $\beta$ -value for probes within 1500bp upstream of the transcriptional start site (TSS)) for each gene to the gene expression. We observed a statistically significant reduction in the average expression of genes that are hypermethylated at each stage of reprogramming compared to primed hESCs, while hypomethylated genes



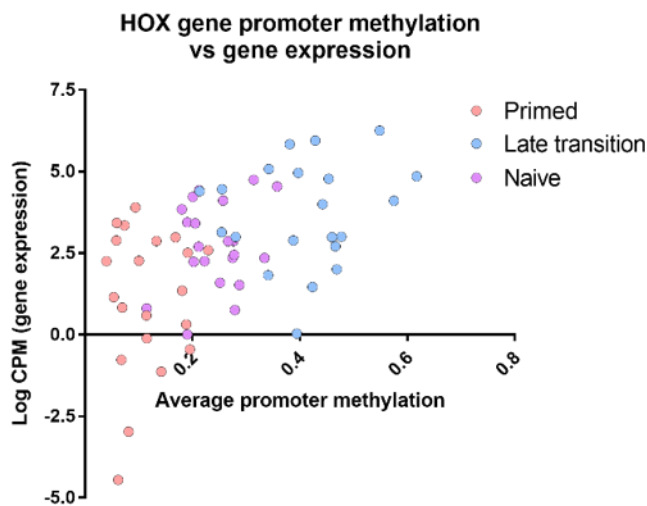
showed either no change or an increase in average expression compared to primed hESCs (Figure 3.10b). Moreover, we observed that when comparing primed to naïve hESCs, genes which undergo hypermethylation are characterised by low average expression in primed hESCs compared to hypomethylated genes. These lowly expressed genes are then further attenuated upon reprogramming, as is often observed in cancer (Figure 3.10b) (Sproul et al. 2012; Easwaran et al. 2012). GO analysis of genes downregulated during the early transition showed an enrichment of genes involved in cell adhesion and motility as well as DNA replication and metabolism. However, at both the late transition and in naïve hESCs compared to primed hESCs, GO analysis showed an enrichment of genes in developmental pathways (Figure 3.10c). As hypermethylation is also enriched in developmental genes, this finding suggests that the hypermethylation may play a functional role in these cells by contributing to downregulation of developmental and differentiation-related pathways, perhaps enhancing the pluripotent state by providing a more stable gene repression mechanism and associated differentiation block.



**Figure 3.10** Genes from developmental pathways are hypermethylated and downregulated during reprogramming. a) MDS plot showing the top 500 most variably expressed genes from RNA-seq data. Replicates for each time point cluster together. b) Average gene expression displayed as log<sub>2</sub> CPM for genes with promoter hypermethylation (average  $\beta$ -value of CpG probes within 1500bp of TSS > 0.1) or hypomethylation (average  $\beta$ -value of CpG probes within 1500bp of TSS < 0.1) during reprogramming. CPM, Counts per Million. Statistical significance between time points determined via paired Wilcoxon test. \*\*\*\*p<0.0001 and \*\*\*p<0.001. Statistical difference of average gene expression of hypermethylated and hypomethylated genes in primed hESCs (which go on to change in naive hESCs) determined via two-tailed Mann-Whitney U test. c) GO analysis of genes that are downregulated at each stage of reprogramming compared to primed hESCs. GO terms plotted are the most highly enriched biological processes with FDR < 0.05.

Intriguingly, whilst we observe the expected anti-correlation between hypermethylation and gene expression at a pathway level, there are subsets of genes, notably the HOX gene family, that become hypermethylated and are upregulated upon hESC reprogramming (Figure 3.11).

Such a phenomenon has been previously described in the context of cancer, where HOX genes are often found to be dysregulated (Bernhart et al. 2016; Su et al. 2018). Additionally, hypermethylation and upregulation of HOX genes has been observed in the context of a mutant form of DNMT3A with a point mutation in the PWWP domain (Sendzikaite et al. 2019).

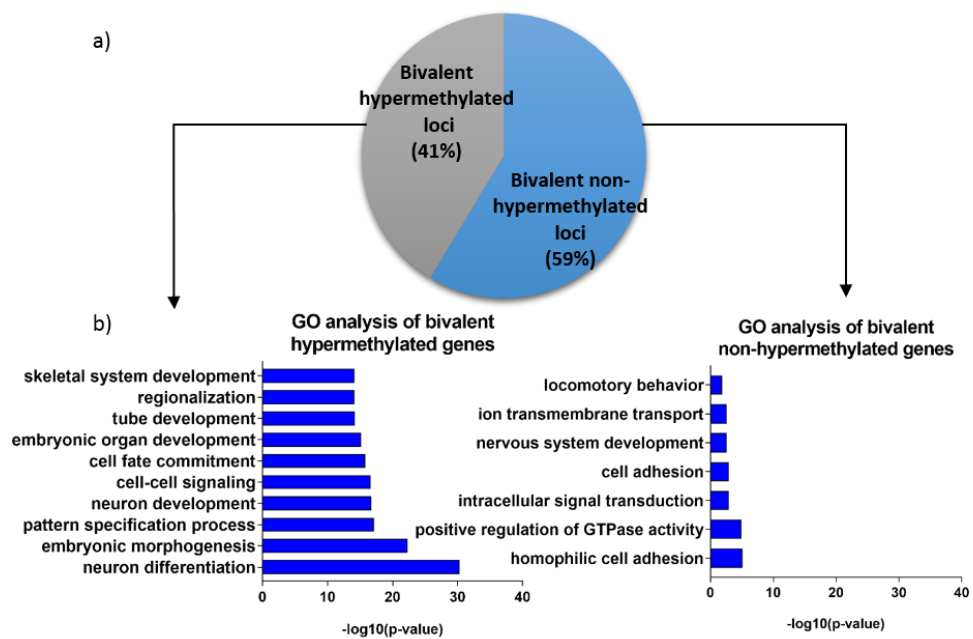


**Figure 3.11 HOX genes are hypermethylated and upregulated.** Scatter plot showing the average promoter methylation of 21 HOX genes (average  $\beta$ -value of CpG probes within 1500bp of TSS) versus the log<sub>2</sub> CPM (counts per million) for each gene from RNA-seq data. Data for each individual time point is indicated by the colour key.

### 3.10 A subset of functionally distinct bivalent loci are hypermethylated upon reprogramming

The significant overlap of hypermethylated regions with loci marked by bivalent histone modifications in primed hESCs led us to hypothesise that the hypermethylation that we observe upon reprogramming may simply be a consequence of bivalency, or that a pre-existing bivalent chromatin state is predictive of hypermethylation upon reprogramming. Utilising the ENCODE ChIP-seq data-set, we classified bivalent regions as those possessing significant peaks of both H3K4me<sub>3</sub> and H3K27me<sub>3</sub>. Having defined bivalent regions in primed hESCs, we observed that only 41% of these loci gain DNA methylation upon reprogramming (Figure 3.12a), indicating that a bivalent chromatin state is not an adequate prerequisite for a genomic region to undergo DNA hypermethylation. We therefore divided all bivalent regions in primed hESCs into two groups: those that become hypermethylated during reprogramming (bivalent hypermethylated) and those that do not become hypermethylated (bivalent non-hypermethylated) (Figure 3.12a). Taking the nearest gene to each region (within 1500bp

upstream of the transcriptional start site), we performed GO analysis of the two groups and observed a striking differential enrichment of pathways. The bivalent hypermethylated group showed a strong enrichment for developmental pathways, as was seen for hypermethylated regions overall, while the bivalent non-hypermethylated group showed lower enrichment of other biological processes (Figure 3.12b). This functional separation of bivalent hypermethylated and non-hypermethylated regions suggests that a shared mechanism coordinates hypermethylation of bivalent CGIs associated with developmental genes.

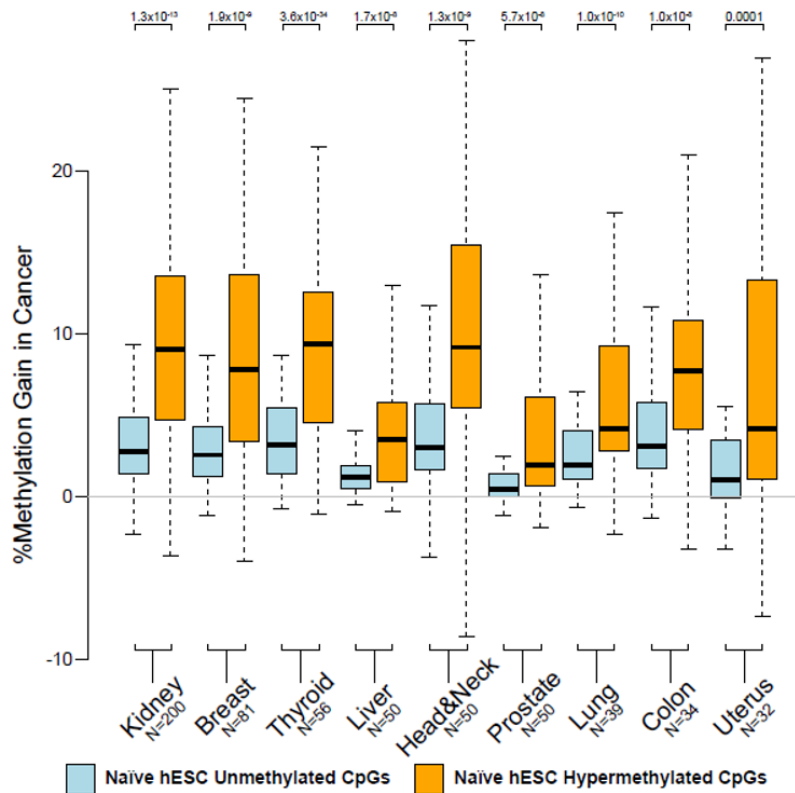


**Figure 3.12. A subset of functionally distinct bivalent loci are hypermethylated upon reprogramming.** a) Pie chart showing the proportion of loci marked by bivalent histone modifications in primed cells that do and do not gain DNA methylation upon reprogramming. b) Gene ontology analysis of bivalent hypermethylated and non-hypermethylated genes respectively (based on the nearest gene to each bivalent region, within 1500bp of the gene transcriptional start site). GO terms plotted are the most highly enriched biological processes with FDR < 0.05.

### 3.11 Reprogramming-associated hypermethylation mirrors cancer hypermethylation

Our data indicate that reprogramming of hESCs result in *de novo* methylation of DNA at loci marked by bivalent chromatin regions, associated with developmental genes. *De novo* DNA methylation of bivalent chromatin in the context of a hypomethylated genome has been reported in cancer cell lines and primary tumours (Bernhart et al. 2016; Easwaran et al. 2012). We set out to investigate a potential link between the hypermethylation patterns associated with hESC reprogramming and the re-emergence of such patterns in cancer. Specifically, we

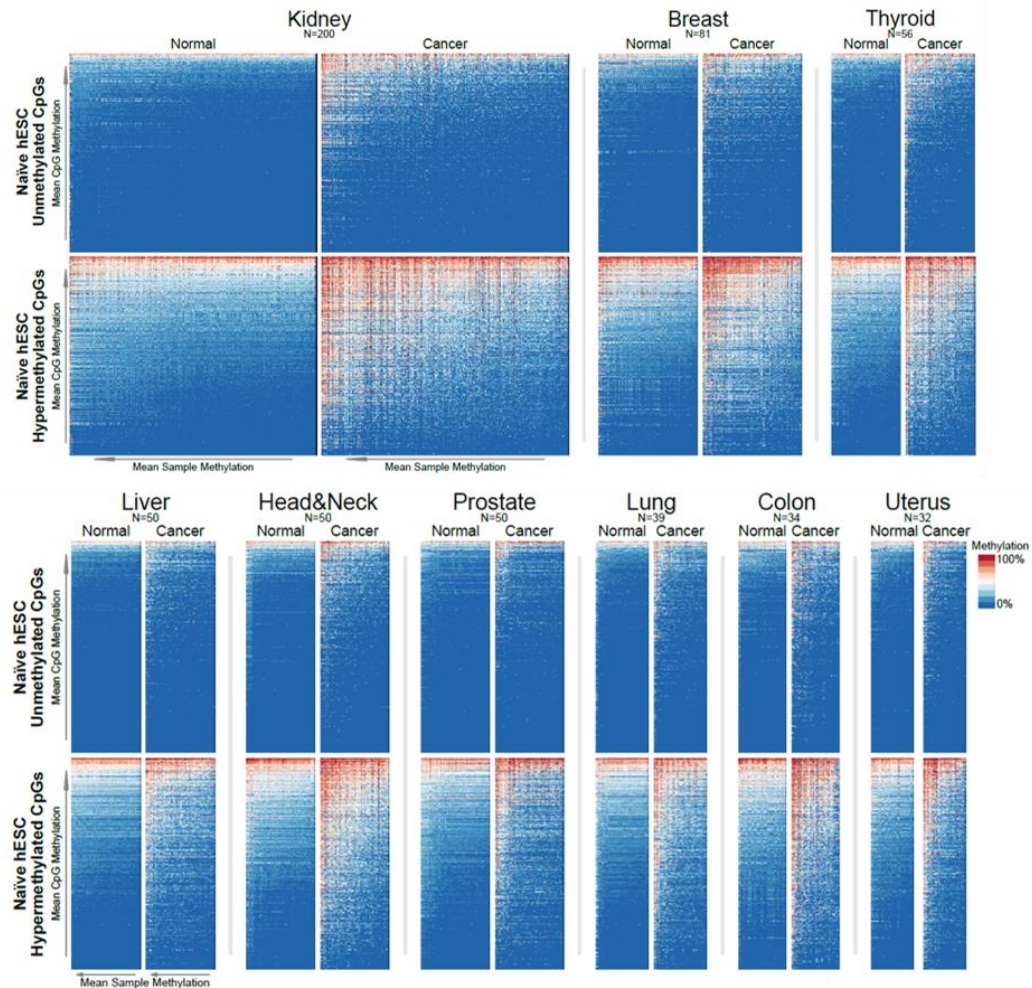
were interested to see whether the specific subset of bivalent regions that are hypermethylated upon reprogramming were also hypermethylated in cancer. We compared our reprogramming-associated bivalent hypermethylated and bivalent non-hypermethylated CpGs with data from the cancer genome atlas (TCGA) pan-cancer atlas.



**Figure 3.13. Naïve bivalent hypermethylated CpGs exhibit increased methylation in tumour vs. normal tissues.** Differences in mean methylation level between normal tissue and tumour samples (from TCGA pan-cancer atlas) of bivalent CpGs identified as hypermethylated or not hypermethylated during the transition to the naïve state in hESCs. Data is presented for 592 individuals, separated by tumour location. P-values determined via paired Wilcoxon test. CpGs used for analysis were filtered for those that are unmethylated in primed hESCs ( $\beta < 0.3$ ).

We saw a significantly higher gain in methylation between normal and cancer tissue for bivalent CpGs identified as hypermethylated during the reprogramming process compared to those that remain unmethylated (Figure 3.13). Strikingly, this was consistent across all cancer types analysed and indicates that reprogramming-associated hypermethylation parallels pan-cancer hypermethylation. Notably, normal control tissues are also generally more susceptible to methylation within these regions (Figure 3.14), suggesting some pre-existing heterogeneity at these sites, perhaps as a result of the ageing process (Rakyan et al. 2010), given that many of the datasets on TCGA are from elderly individuals. Additionally, as the normal controls are

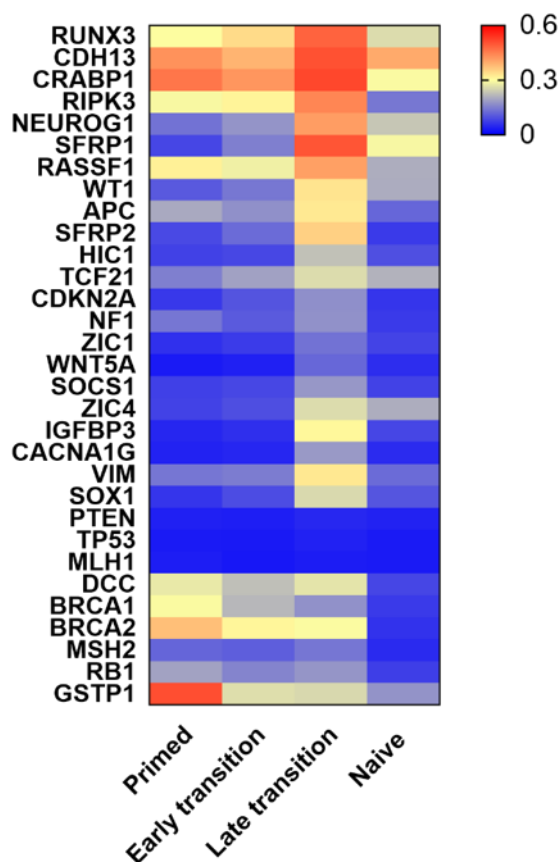
typically tumour-adjacent tissue samples derived from diseased patients, it is plausible that these tissues are already different to tissues obtained from a completely healthy individual.



**Figure 3.14. Reprogramming-associated hypermethylation mirrors cancer hypermethylation.** Heatmaps showing the mean methylation level of CpGs identified as hypermethylated or not hypermethylated in hESCs (as in Fig 3.12) during the transition from primed to naïve state, in tumour and corresponding normal tissue samples (from TCGA pan-cancer atlas) for a variety of cancer types. Data is ordered by mean methylation level, for each cancer type. Data was restricted to those cancer types that had at least 30 matched normal and cancer tissue datasets available.

In addition to the overall cancer-like hypermethylation pattern during reprogramming, we were also specifically able to detect hypermethylation of a number of tumour suppressor genes (using a list manually curated from the literature) (Llinas-Arias and Esteller 2017) that are commonly hypermethylated and inactivated in various cancer types (Figure 3.15), reinforcing the similarity between hypermethylation during reprogramming and in cancer.

### Average promoter methylation for common tumour suppressor genes

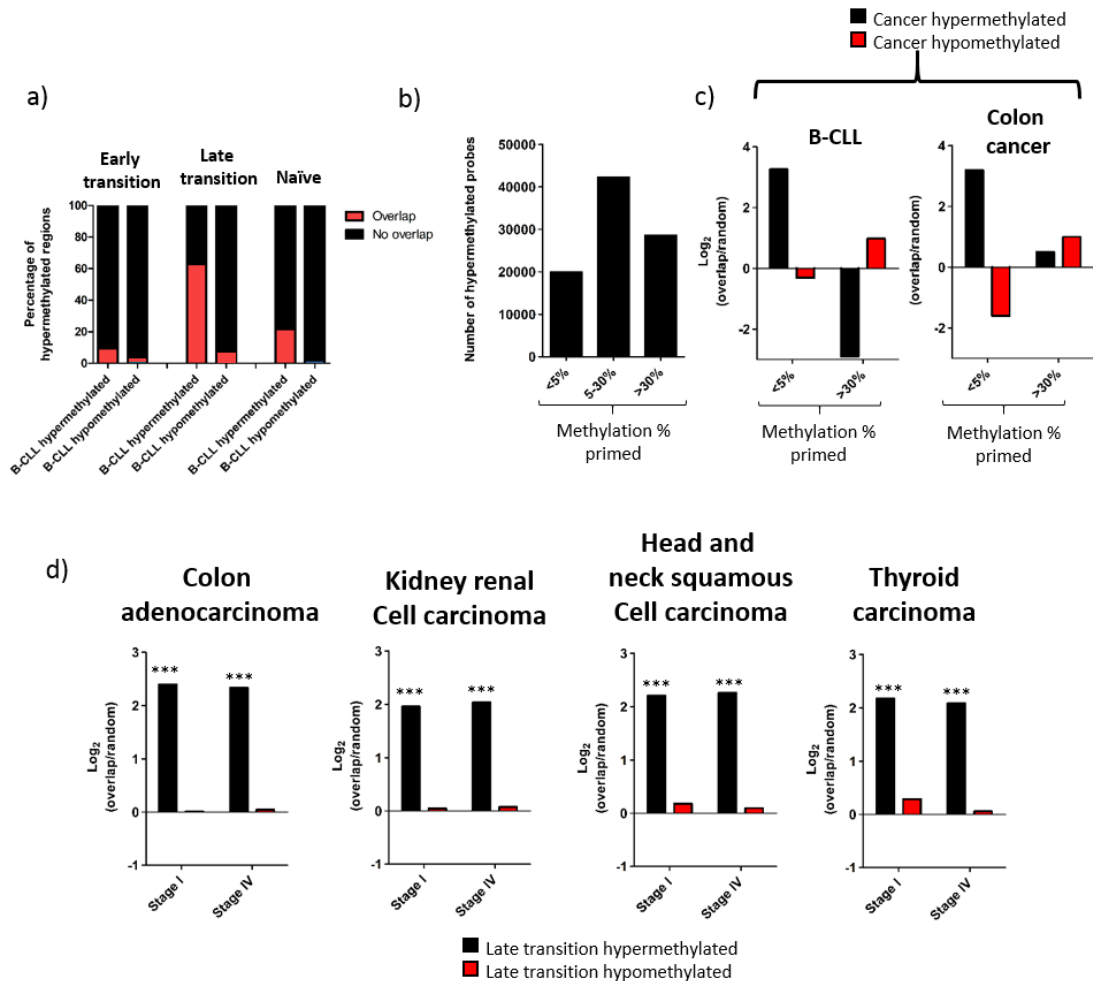


**Figure 3.15. Common tumour suppressor genes are hypermethylated upon primed to naïve reprogramming.** Heatmap showing the average promoter methylation level (average  $\beta$ -value of CpG probes within 1500bp of the transcriptional start site) of common tumour suppressor genes (manually curated list from the literature) at each time point. Methylation  $\beta$ -value is indicated by the colour key.

We further compared hypermethylated CpGs at each stage of reprogramming with regions previously identified as hypermethylated in B-cell chronic lymphocytic leukaemia (B-CLL) (Kushwaha et al. 2016). The most substantial overlap is observed between hypermethylated CpGs associated with the late transition of reprogramming and B-CLL hypermethylated regions, which corresponds to the peak of hypermethylation we observe during reprogramming (Figure 3.16a). Interestingly, we also see a more significant overlap between hypermethylated regions identified in B-CLL and colon cancer (Hansen et al. 2011) and reprogramming-associated hypermethylated CpGs with a low basal methylation level (<5%) in primed hESCs (Figures 3.16b, 3.16c). We see no further enrichment of reprogramming-associated hypermethylated CpGs when we overlap this dataset with more advanced stages of cancer (Figure 3.16d), suggesting that DNA hypermethylation occurs earlier in cancer



development and that this pattern might be maintained as the cancer progresses and evolves. These data support the idea that common underlying mechanisms of hypermethylation may be at play during hESC reprogramming and tumourigenesis and that further exploitation of this model system may shed further light on the cellular networks regulating this.



**Figure 3.16. Cancer hypermethylation shows the strongest overlap with late transition hypermethylation.** a) Proportion of hypermethylated and hypomethylated regions in B-CLL (from published data) that overlap hypermethylated probes identified during each stage of hESC reprogramming. Data shows an enrichment for late transition hypermethylated probes within B-CLL hypermethylated regions (and to a lesser extent naive hypermethylated probes). b) Bar plot showing the basal level of methylation in primed hESCs of probes that are hypermethylated during the late transition of reprogramming. c) Overlap of late transition hypermethylated probes (with <5% or >30% starting methylation in primed hESCs) with B-CLL and colon cancer hypermethylated regions from published data. d) Overlap of late transition DMPs with cancer hypermethylated probes obtained from TCGA. Overlaps were performed for both hyper- and hypomethylated probes compared to either stage I or stage IV hypermethylated probes (compared to normal controls) for each cancer type. For overlap analysis, data is presented as the log<sub>2</sub> corrected fold increase in the observed overlap compared to the mean overlap of 1000 randomly generated loci, where random loci generation was restricted to loci present in the Illumina EPIC array. \*\*\*P<0.001.



### 3.12 Discussion

The data presented in this chapter demonstrate for the first time the temporal dynamics of DNA methylation upon hESC reprogramming to the naïve state. Using both global and targeted approaches to measure DNA methylation, we see a high degree of similarity between the results using different techniques, highlighting the reproducible nature of reprogramming-associated DNA methylation changes. We observed that upon the transition from primed to naïve pluripotency, the genome is globally demethylated in a time-dependent manner. Whilst we do not focus on the genomic context of DNA demethylation in this chapter, other studies have shown that the DNA methylation landscape in naïve hESCs is comparable to that of the human ICM (Takashima et al. 2014; Pastor et al. 2016). Notably, however, DNA methylation is also lost from imprinted regions in naïve hESCs (Pastor et al. 2016). The loss of stable imprints is not reflective of the human ICM, but is often implicated in cancer (Cui et al. 2003; Holm et al. 2005; Jelinic and Shaw 2007).

In parallel with global DNA demethylation, we also observe a decrease in the level of 5hmC in the genome. This has been reported previously in naïve hESCs (Takashima et al. 2014), however our temporal analysis shows that the kinetic profile of 5hmC loss is distinct to the loss of 5mC. It appears that the level of 5hmC drops to its lowest level during the early transition of reprogramming, and subsequently recovers to some degree, though still at a significantly lower level than that in primed hESCs. This may indicate that the global loss of DNA methylation in the early stages of reprogramming can be attributed to a passive mechanism of demethylation, but that this may be accompanied by active oxidation of 5mC into 5hmC in the later stages of reprogramming. Interestingly, one week following induction of the naïve state, when 5hmC levels are at their lowest, we detected a significant spike in the level of 5fC in the genome. Such an increase in 5fC, which is usually present at very low levels in the genome, suggests that the modified base may be playing a role in the reprogramming process. 5fC has been shown to be bound by a number of transcriptional and chromatin regulatory proteins (Spruijt et al. 2013; Iurlaro et al. 2013) and it has been suggested that it may exhibit regulatory functions in cells (Song and He 2013; Song et al. 2013). However, additional replicates are required to draw any firm conclusions from this data as the two biological replicates measured display a high degree of variability.

We observed that upon hESC reprogramming, DNA methylation is acquired at specific loci, primarily CGI promoters, which are typically unmethylated in mammalian cells (Bird et al. 1985). Notably, we observe that the gain in DNA methylation is gradual, with a peak of hypermethylation during the late transition of reprogramming, after which hypermethylation

is maintained at some genomic regions but lost at others. The transient nature of hypermethylation at some genomic regions may be a result of improper maintenance of DNA methylation, particularly as the majority of the genome is undergoing demethylation during the reprogramming process. It may also be an indication that cells in the late transition of reprogramming have a unique cellular identity, which is not maintained upon complete generation of naïve hESCs. It is plausible that there may be some *in vitro* selective pressure on the cells as they transition to the naïve state, particularly during the later stages of reprogramming, which may select for hESCs with a specific DNA methylation signature, however further experimentation is required to test this hypothesis.

For Infinium MethylationEPIC BeadChip data, we classify hypermethylation as a >10% increase in methylation per CpG at any time point of reprogramming compared to primed hESCs. For re-analysis of published bisulphite sequencing data, we classify a hypermethylated region as a 300bp region containing 5 or more CpGs, with >5% increase in methylation in naïve compared to primed hESCs. These cut-off values were selected based on other publications, however it is difficult to determine the change in DNA methylation required to have a biological impact and there is no unified consensus on a biologically relevant cut-off for calling methylation changes in the literature. Additionally, when using data from a methylation array and performing DMP analysis, each CpG probe is treated individually. This makes it difficult to determine whether multiple CpGs on the same DNA, from the same cell are hypermethylated or whether the true pattern of hypermethylation in the population of cells is more stochastic. However, the reproducibility of results between multiple methylationEPIC arrays and also across multiple methods of measuring DNA methylation suggests that the site-specific gain in methylation we observe is the result of a non-random process.

While the hypermethylated sites are enriched in regulatory regions, there are hypermethylated CGIs present in promoters, enhancers and gene bodies, with individual hypermethylated CpGs residing within CGIs but also within their shores, shelves and outside of CGIs altogether. It is becoming increasingly evident that DNA methylation in each of these contexts can have a unique biological consequence (Jones 2012), implying that the long-standing anti-correlation between gene promoter hypermethylation and gene expression is not the only possible outcome of DNA hypermethylation. Nevertheless, with a focus on promoter methylation, we observed a reduction in the average expression of genes that are hypermethylated at each stage of reprogramming compared to primed hESCs. The exception to this was a subset of genes including several of the HOX genes, which appear to be simultaneously hypermethylated and upregulated. Such a phenomenon has been previously

described in the context of cancer, where HOX genes are often found to be dysregulated (Bernhart et al. 2016; Su et al. 2018) and more recently in the context of a DNMT3A PWWP domain mutation, where HOXA7 and HOXD8 were found to be hypermethylated and upregulated (Sendzikaite et al. 2019). However, we cannot currently exclude the possibility of 5-hydroxymethylcytosine being present at the promoters of these genes, or population heterogeneity whereby HOX gene overexpression bias is caused by a subpopulation of cells which do not undergo HOX promoter hypermethylation.

The hypermethylation pattern observed upon hESC reprogramming mirrors the frequently observed aberrant hypermethylation in human cancers, in both cases occurring in the context of a globally hypomethylated genome. Notably, along with the overall similar pattern of bivalent CGI hypermethylation of developmental genes, we also see hypermethylation of a number of common tumour suppressor genes, emphasising the commonality between the two phenomena. Such parallels with cancer hypermethylation have been drawn previously in other mammalian species and developmental contexts (Smith et al. 2017), however the data we present here demonstrates a hypermethylation phenomenon conserved across *in vitro* and *in vivo* human pluripotency (Guo et al. 2014), strengthened by its reproducibility across multiple *in vitro* reprogramming methods (Theunissen et al. 2014; Guo et al. 2017). Moreover, it is notable that we do not observe comparable hypermethylation in the mouse ICM or in *in vitro* mouse ESCs cultured in the presence of 2i inhibitors, though this may reflect the fact that hESCs and mESCs *in vitro* represent different pluripotent states (Ficz et al. 2013; Smith et al. 2012; Davidson, Mason, and Pera 2015). This observation has potential implications for making inferences with regards to epigenetic processes between species, both in development and in the study of cancer, as has been noted previously (Diede et al. 2013).

We show that the reprogramming-associated hypermethylated loci overlap with loci marked by H3K4me1/2/3 and H3K27me3 in primed hESCs. For this analysis, we use ENCODE ChIP-seq data from the H1 hESC line as this cell line, despite the experimental cell line in our study being H9. This is because the H1 cell line has been extensively characterised on ENCODE, and where data is available for both H1 and H9, the data sets are highly similar. Our data indicate that loci marked with the bivalent histone modification H3K4me3 and H3K27me3 are almost exclusively susceptible to DNA hypermethylation, but that the presence of bivalent chromatin is not a sufficient prerequisite for acquiring *de novo* DNA methylation upon reprogramming. This phenomenon parallels what is known about hypermethylation in cancer, where numerous studies have reported preferential susceptibility of H3K27me3-marked loci to gain DNA methylation (Ohm et al. 2007; Schlesinger et al. 2007; Widschwendter et al. 2007) or shown

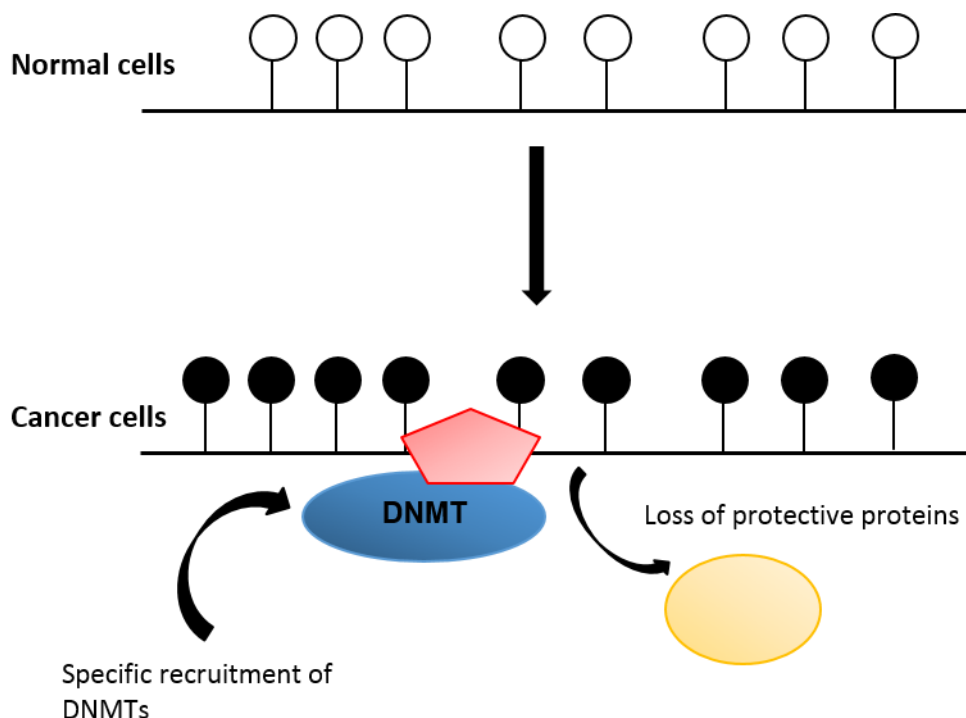
that bivalent regions in primed embryonic stem cells are frequently hypermethylated in cancer (Ohm et al. 2007; Easwaran et al. 2012). It is particularly noteworthy that the bivalent loci that undergo hypermethylation both upon reprogramming and across multiple cancer data sets are distinguishable from those that do not gain methylation despite having a comparable starting chromatin configuration, which adds a further dimension to the current understanding of the relationship between bivalent chromatin and hypermethylation in cancer. Aside from the difference in GO enrichment between the bivalent hypermethylated and bivalent non-hypermethylated subsets, it is likely that additional features such as the underlying DNA sequence separate these two groups and contribute to the mechanism of hypermethylation.

The data shows that hypermethylation during primed to naïve reprogramming affects genes belonging to developmental pathways. In the context of cancer, the function of hypermethylation remains a topic of debate. While several studies have shown clear functional roles of hypermethylation and gene repression for individual tumour suppressor genes (Jones and Baylin 2002; Saunderson et al. 2017), it remains less well understood what the purpose of hypermethylation of a large number of loci, many of which are common between cancer types, might be. It has been proposed that aberrant hypermethylation in cancer may act to block cellular differentiation, thus enabling cancer cells to continue to propagate in their more primitive states (Widschwendter et al. 2007; Easwaran et al. 2012; Pfeifer 2018), and this has been demonstrated experimentally in a recent study (Tao et al. 2019). It has also been proposed that hypermethylation may act to block cancer progression, with CGI hypermethylation restricting the epigenetic adaptability of cells during the process of metastasis or upon cancer treatment (Sproul and Meehan 2013). The commonality in methylation patterns across cancer types, each harbouring different driver mutations, suggests that these methylation changes occur early in tumourigenesis, and this has been demonstrated previously (Hanley et al. 2017) though models of early cancer development are limited. In line with this, the notion that cancer cells follow an evolutionary trajectory towards a stem cell state (Chen and He 2016; Ben-Porath et al. 2008) makes the transition from primed to naïve pluripotency an interesting model to study biological processes such as DNA methylation that likely occur early during cellular transformation or cancer initiation and may be analogous to dedifferentiation. In particular, the following chapter will detail the use of the primed to naïve reprogramming system as a model for the mechanism of DNA hypermethylation.

## Chapter 4. Results 2: The mechanism of DNA hypermethylation during primed to naïve reprogramming

### 4.1 Introduction

The acquisition of site-specific DNA hypermethylation in the context of a globally hypomethylated genome is a hallmark of cancer (Jones and Baylin 2007; Baylin and Jones 2011; Esteller et al. 2001; Feinberg, Ohlsson, and Henikoff 2006). Despite the frequency with which these observations have been made in cancer cell lines or primary cancer cells, how *de novo* DNA methyltransferase activity is preferentially targeted to specific regions of the genome in the context of aberrant cancer methylation remains largely a mystery. Two main hypotheses prevail in the current literature with regards to the mechanism of hypermethylation in cancer (Figure 4.1): site-specific recruitment of DNMTs to target loci either as a result of increased DNMT protein expression or loci-specific targeting by transcription factors, or loss of passive, protective mechanisms that typically act to maintain loci in a hypomethylated state (Sproul et al. 2012).



**Figure 4.1. Possible mechanisms of hypermethylation in cancer.** Possible mechanisms that result in the hypermethylation of CGI promoters in cancer. CGI promoter hypermethylation could result from either the loss of a protective mechanism that typically maintains CGIs in a hypomethylated state or a gain of *de novo* methyltransferase activity at the CGI (either targeted by transcription factors or through an increase in levels of the DNMTs in the cell).

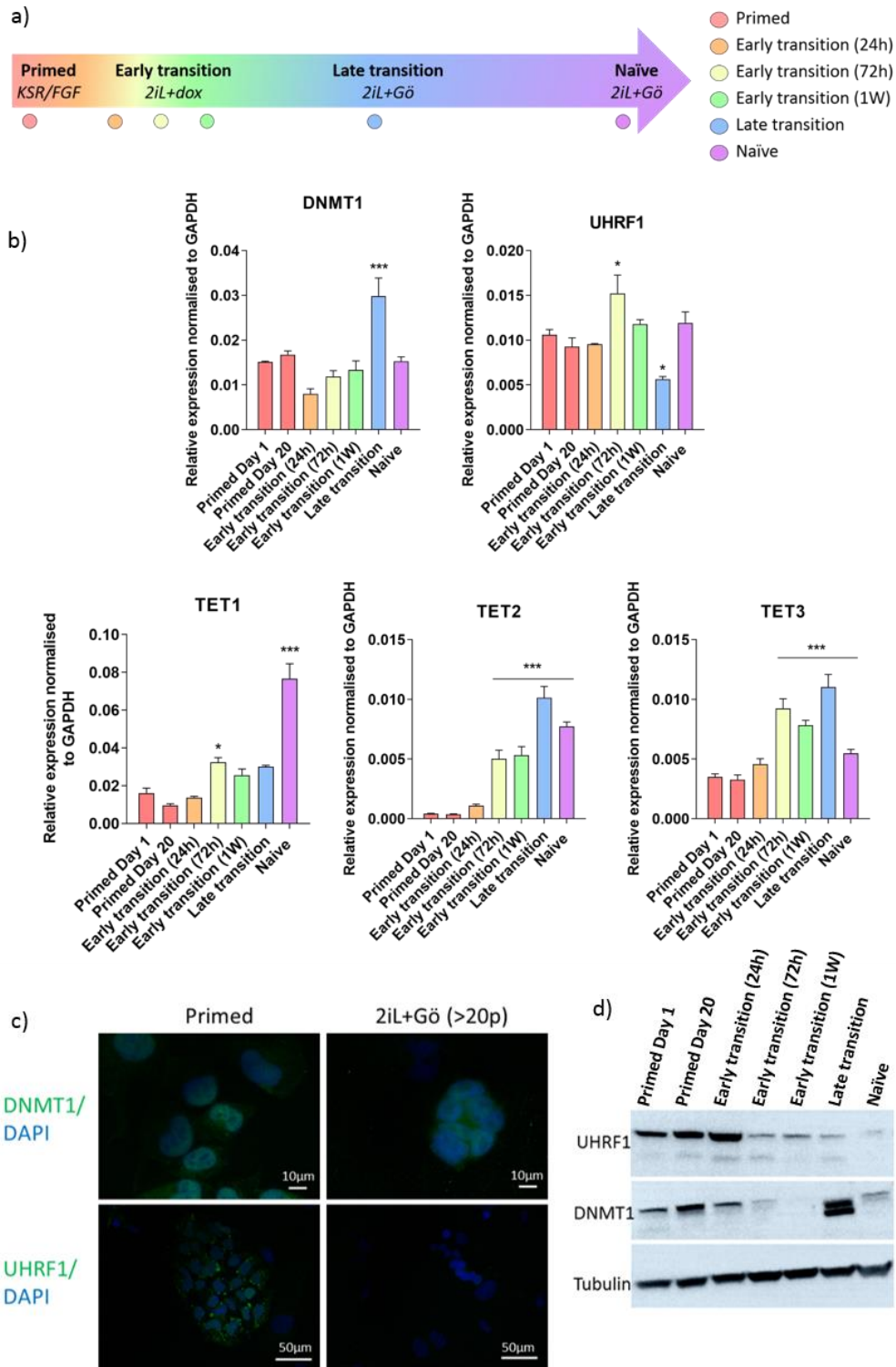
Despite considerable knowledge on the domain composition of the DNMTs, the understanding of how these enzymes are targeted and regulated to give rise to specific DNA methylation patterns remains limited. The recognition of H3K4 or H3K36me3 by the ADD and PWWP domains, respectively, have been proposed as general mechanisms of recruitment for the DNMT3 enzymes, implying that the interaction with chromatin is important for the generation of DNA methylation patterns (Otani et al. 2009; Ooi et al. 2007; Zhang et al. 2010; Ge et al. 2004; Dhayalan et al. 2010; Chen, Tsujimoto, and Li 2004). Additionally, the DNMT3 enzymes have been shown to be able to directly interact with numerous transcription factors *in vitro* (Hervouet, Vallette, and Cartron 2014), and it has been proposed that such interactions may influence localisation and targeting of the DNMTs. It is well established that CGIs in the genome are generally unmethylated (Bird et al. 1985; Weber et al. 2007). CpG-rich regions are thus thought to have mechanisms at play that protect them from DNA methylation, such as the binding of zinc finger CXXC domain-containing proteins which specifically bind to unmethylated CpG dinucleotides and attract chromatin environments refractory to DNA methylation (Voo et al. 2000; Long, Blackledge, and Klose 2013; Thomson et al. 2010). Both active targeting of DNMTs or loss of the inherent protection of CGIs may underlie CGI hypermethylation in cancer, or indeed a combination of such mechanisms may be at play.

It has been hypothesised that cancer cells follow an evolutionary trajectory towards a stem cell state, which allows both self-renewal and differentiation (Chen and He 2016). This, along with the parallels drawn with the cancer methylome (discussed in chapter 3), makes the transition from primed to naïve pluripotency an interesting model to study DNA methylation mechanisms that may be analogous to processes occurring during cancer formation. This chapter will use knowledge obtained from chapter 3 concerning the targeted hypermethylation at bivalent CGI promoters associated with developmental genes, and the parallels drawn with pan-cancer hypermethylation. The role of known components of the DNA methylation machinery in mediating hypermethylation upon hESC reprogramming will be described. Specifically, the DNMT responsible for hypermethylation will be highlighted and upstream mechanisms that may coordinate hypermethylation will be discussed. This chapter will explore the possibility of both active targeting of DNMTs or loss of protective mechanisms, through identification of candidate factors that may mediate each mechanism in the context of reprogramming-associated hypermethylation. Additionally, the possible functional relevance of hypermethylation in hESCs will be explored. Finally, the relevance of all these findings will be discussed in the context of cancer biology.

## 4.2 Downregulation of maintenance methylation machinery coincides with global demethylation upon reprogramming

To investigate the role of epigenetic machinery in regulating the DNA methylation changes associated with hESC reprogramming, we measured gene expression changes of known mediators of DNA methylation and demethylation by qRT-PCR in a temporal manner. The time points used coincide those used to measure DNA methylation changes, with additional increased resolution (Figure 4.2a), thus enabling inference of which of the epigenetic regulators may be responsible for the changing DNA methylation patterns. We found little change in expression levels of the maintenance methylation machinery genes, *DNMT1* and *UHRF1*, except for a peak in expression of *DNMT1* at the late transition (4 days after addition of Gö), coinciding with the peak of hypermethylation alluded to in chapter 3 (Figure 4.2b). We also measured the expression of the *TET* enzymes, which are able to demethylate the DNA by oxidation of 5mC (Tahiliani et al. 2009; Ito et al. 2011). We found that all three enzymes are upregulated, though the extent and timing of the upregulation varies subtly between the three (Figure 4.2b).

In parallel with transcriptional changes, we measured the level of protein expression of the two key proteins typically involved in maintenance of DNA methylation. We initially used immunofluorescence to detect expression of DNMT1 and UHRF1 primed hESCs and naïve hESCs. We saw little change in the expression of DNMT1 between primed and naïve hESCs (Figure 4.2c), while UHRF1 was detectable at high levels in the primed hESCs and markedly downregulated in naïve hESCs. We next used Western Blotting to detect the expression of the two proteins across the transition from primed to naïve pluripotency. While *UHRF1* gene expression levels were stable across the transition, UHRF1 protein is downregulated upon induction of the naïve state (Figure 4.2d). DNMT1 protein also appears to be downregulated during the early transition, but shows a peak in protein expression at the late transition (4 days after the addition of Gö), which coincides with the peak observed in expression of the *DNMT1* gene, as well as the peak in hypermethylation (Figure 4.2d). Interestingly, upon multiple experimental repeats, two bands can be detected at this stage on the Western blot, suggestive of an alternative isoform of DNMT1 protein, a degradation product, or a post-translationally modified version of the protein (Figure 4.2d).



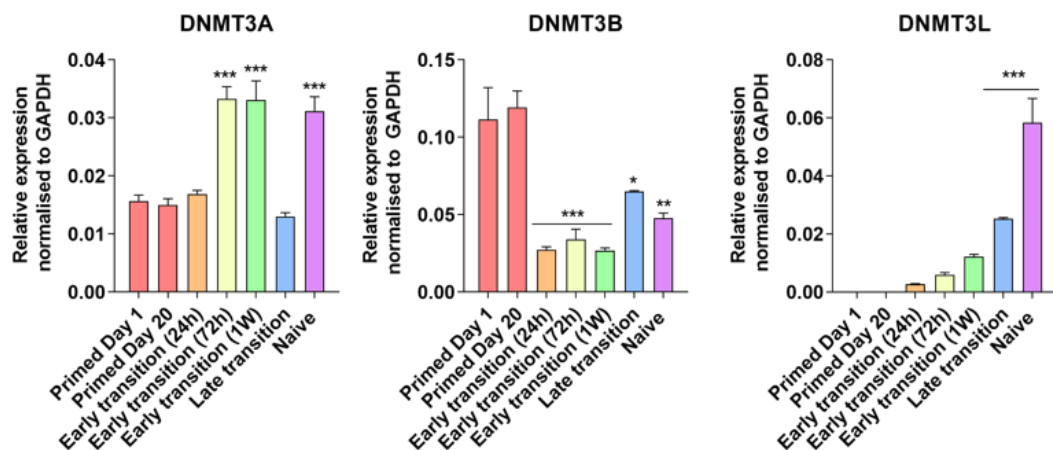
**Figure 4.2. Downregulation of maintenance methylation machinery and upregulation of the TET mRNA coincides with global demethylation upon reprogramming.** a) Schematic detailing the time points used for expression analysis of DNA methylation regulators. b) qRT-PCR for maintenance methylation genes and TET genes across the transition of hESC reprogramming. Bars are representative of the mean of three technical replicates and three independent biological replicates. Error bars represent the standard error of the mean. Statistical difference between samples was analysed by a two-tailed Student's t-test, with a Bonferroni post-hoc test comparing each time point to primed hESCs. \* $P < 0.05$ , \*\* $P < 0.01$  and \*\*\* $P < 0.001$ . Human GAPDH was used to normalise expression. c) Immunofluorescence of cells stained with DAPI and DNMT1 or UHRF1 in primed and naïve hESCs. Scale bars represent 10 $\mu$ m or 50 $\mu$ m as indicated. d) Detection of UHRF1 and DNMT1 protein expression by Western Blotting across the transition, with  $\alpha$ -tubulin as a loading control. Blots displayed are representative of 3 replicates.



Both downregulation of DNA methylation maintenance machinery and upregulation of the TET enzymes at the mRNA level coincide temporally with global demethylation of the genome upon hESC reprogramming. They may both contribute to genome-wide loss of methylation through passive or active demethylation, respectively, though the observation that we detected a global loss in the levels of 5hmC (detailed in chapter 3) suggest that TET driven active oxidation of 5mC is not the primary mechanism of demethylation.

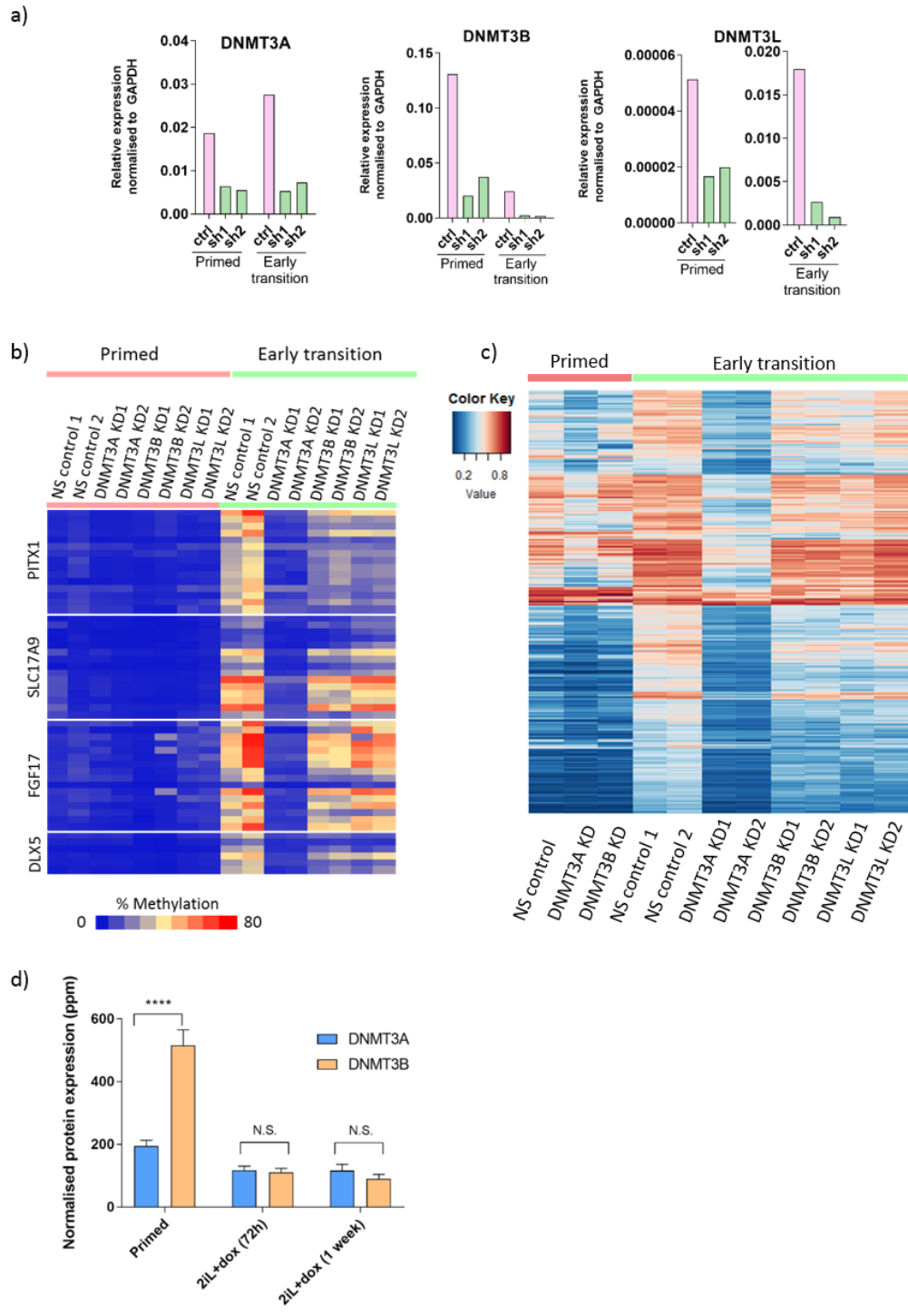
### 4.3 DNA hypermethylation is carried out by DNMT3A during early hESC reprogramming

To investigate the role of epigenetic machinery in driving hypermethylation, we sought to identify the DNMTs that are responsible for the deposition of *de novo* DNA methylation. We measured temporal gene expression changes of the *de novo* DNMTs, encompassing the DNMT3 family. Of the *de novo* DNMTs, *DNMT3A* undergoes a transient upregulation during the early transition of reprogramming, *DNMT3B* levels are downregulated across the transition, and the catalytically inactive *DNMT3L* is highly upregulated (Figure 4.3).



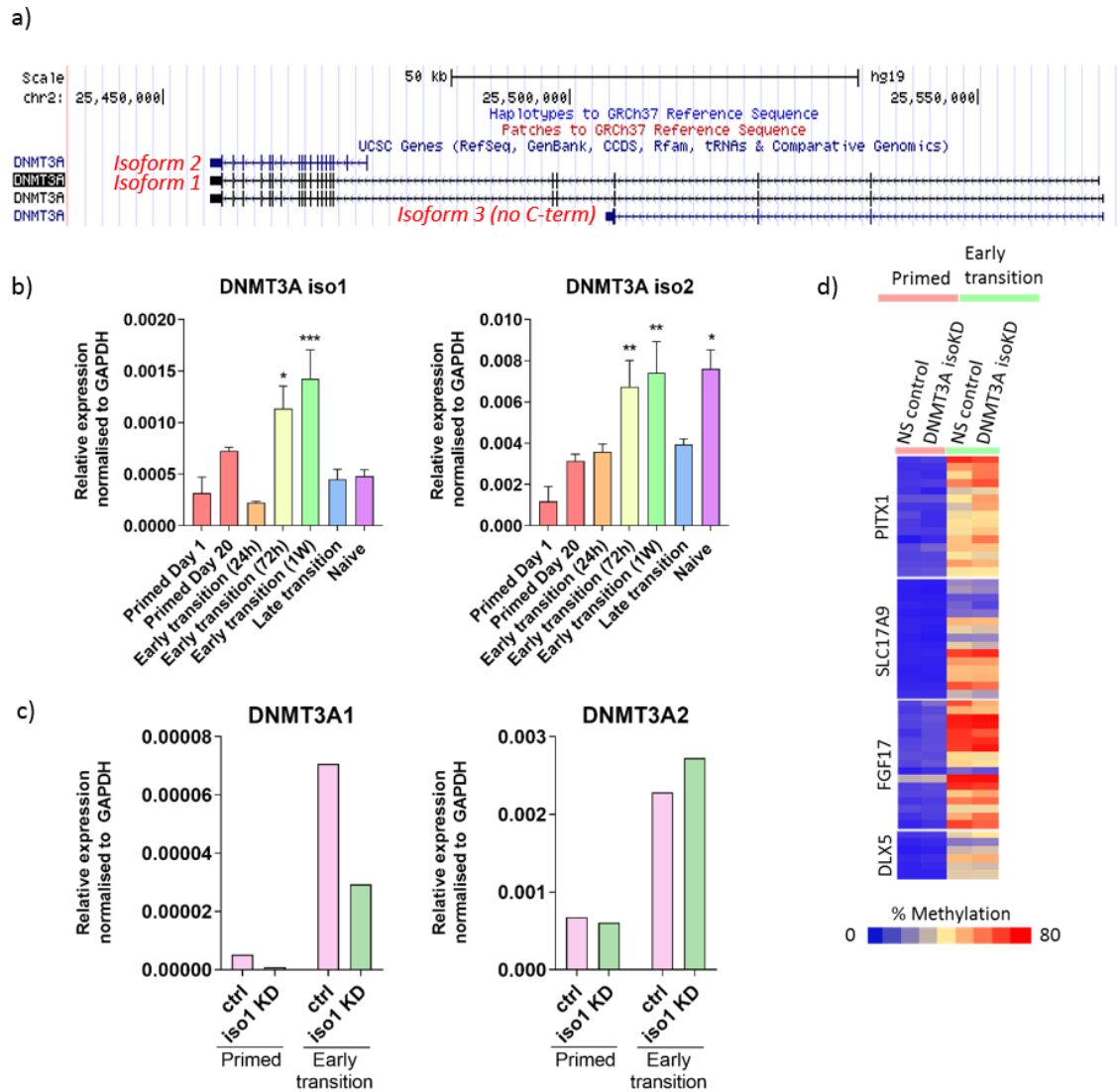
**Figure 4.3. De novo DNMTs are dynamically expressed during primed to naive hESC reprogramming.** qRT-PCR for DNMT3 genes across the transition of hESC reprogramming. Bars are representative of the mean of three technical replicates and three independent biological replicates. Error bars represent the standard error of the mean. Statistical difference between samples was analysed by a two-tailed Student's t-test, with a Bonferroni post-hoc test comparing each time point to primed hESCs. \* $P < 0.05$ , \*\* $P < 0.01$  and \*\*\* $P < 0.001$ . Human GAPDH was used to normalise expression.

To test which of the *de novo* methyltransferases drives hypermethylation, we generated constitutive knockdown primed hESC cell lines using two short hairpin RNAs (shRNAs) targeting each of the three DNMT genes. We subjected each of the cell lines to reprogramming until the early transition, at which stage reprogramming-associated hypermethylation is already detectable. We validated the efficiency of the shRNA knockdowns by measuring expression of the target genes in both primed and early transition hESCs (Figure 4.4a). We first used a targeted approach to measure DNA methylation at four genomic loci that are hypermethylated upon hESC reprogramming. In primed hESCs, we observed no difference in the levels of methylation in any of the knockdown cells compared to control cells, indicating that *de novo* DNMTs are dispensable at these loci in primed hESCs. In the early transition, knockdown of *DNMT3B* and *DNMT3L* had little impact on the level of methylation, as the cells exhibited comparable levels of hypermethylation to control cells (Figure 4.4b). Knockdown of *DNMT3A*, however, was able to abolish hypermethylation, as these cells retain the levels of methylation present in primed hESCs (Figure 4.4b). To ensure that the effect of the *DNMT3A* knockdown was not restricted to this selection of genomic regions, we performed an Infinium MethylationEPIC BeadChip array to observe the impact of loss of *DNMT3A* across the genome. As the numbers of replicates used for this experiment were insufficient to perform DMP analysis between the knockdowns and controls in primed and early transition hESCs, we extracted methylation beta-values for the probes identified as hypermethylated ( $\Delta\beta > 0.1$ ,  $p < 0.05$ ) in wild-type early transition hESCs compared to primed hESCs (of which the top 10,000 were presented in Figure 3.5 of Chapter 3), and plotted these as a heatmap. These data confirmed that in *DNMT3A* knockdown early transition hESCs, hypermethylation was not detected at any of the regions that gain methylation in control cells, while *DNMT3B* and *DNMT3L* early transition knockdown cells showed hypermethylation at comparable levels to the control (Figure 4.4c). This is despite comparable protein expression levels of *DNMT3A* and *DNMT3B* during the early transition of reprogramming, as measured by protein mass spectrometry (Figure 4.4d). Interestingly, with increased resolution across the genome compared to the targeted approach, it is evident that a knockdown of *DNMT3A* influences the level of methylation in primed hESCs too, as primed *DNMT3A* knockdown show a slight reduction in methylation compared to the control across all CpGs plotted. Collectively, this data indicates that hypermethylation during the early transition of reprogramming is carried out by *DNMT3A* and is independent of *DNMT3L*.



**Figure 4.4. Knockdown of DNMT3A abolishes hypermethylation in the early transition of hESC reprogramming.** a) qRT-PCR for the transcripts of the DNMT3 family in control and knock-down cell lines, in primed and early transition hESCs. Bars represent the mean of three technical replicates. Human GAPDH was used to normalise expression. b) Targeted bisulfite-sequencing of 4 regions of DNA. Each square represents the methylation % indicated by the colour key of a single CpG. Data is shown for 2 knockdown samples of DNMT3A/B/L and two non-silencing (NS) controls in primed and early transition hESCs. c) Heatmap showing methylation levels for primed and early transition control and DNMT3A/B/L knock down samples. Heatmap is composed of the top 20,000 CpG probes that are differentially methylated ( $\Delta\beta > 0.1$ ,  $p < 0.05$ ) in the wild-type early transition compared to primed hESCs. Methylation  $\beta$ -value is indicated by the colour key. d) Raw protein intensity values for DNMT3A and DNMT3B extracted from proteomics data, plotted in parts per million (ppm). Bars represent an average of 2 intensity values for each of the 3 replicates per sample, with error bars representing the SEM. Statistical difference between samples was analysed by a two-tailed student's t-test at each time point. \*\*\*\* $P < 0.0001$ .

DNMT3A is encoded by two main isoforms in human cells: a full length isoform DNMT3A1 that has an extended N-terminal region and is typically expressed in somatic cells, and a shorter isoform DNMT3A2, whose expression is typically restricted to early developmental cell types (Chen et al. 2002) (Figure 4.5a). A third transcript also exists but lacks the C-terminal that encodes the catalytic domain. A recent study demonstrated that the N-terminal of DNMT3A1 can bind DNA and localizes specifically to the shores of bivalent CpG islands in mouse ESCs, where it can carry out *de novo* methylation (Manzo et al. 2017). We measured the transcript expression of the individual isoforms of *DNMT3A* and found that the two isoforms are differentially expressed during hESC reprogramming (Figure 4.5b). Both isoforms are transiently upregulated during the early transition, and *DNMT3A2* is additionally upregulated in the naïve state, though the overall expression of *DNMT3A2* across the period of reprogramming is much higher than that of *DNMT3A1* (Figure 4.5b). To test whether *DNMT3A1* is responsible for the hypermethylation of bivalent CpG islands upon hESC reprogramming, we generated a specific shRNA knockdown cell line of *DNMT3A1* in primed hESCs by targeting its unique N-terminal domain, without affecting the expression of *DNMT3A2* (Figure 4.5c). We subjected the *DNMT3A1* knockdown cell line and a control cell line to reprogramming until the early transition and used a targeted approach to measure DNA methylation at three regions known to be hypermethylated upon reprogramming. In primed hESCs, there was no impact of the *DNMT3A1* knockdown on DNA methylation levels. Contrary to the study in mouse ESCs however (Manzo et al. 2017), *DNMT3A1* knockdown hESCs also exhibited comparable levels of hypermethylation to the control at the target regions analysed during the early transition of reprogramming (Figure 4.5d), hence suggesting that the more dominantly expressed *DNMT3A2* carries out *de novo* methylation upon hESC reprogramming.

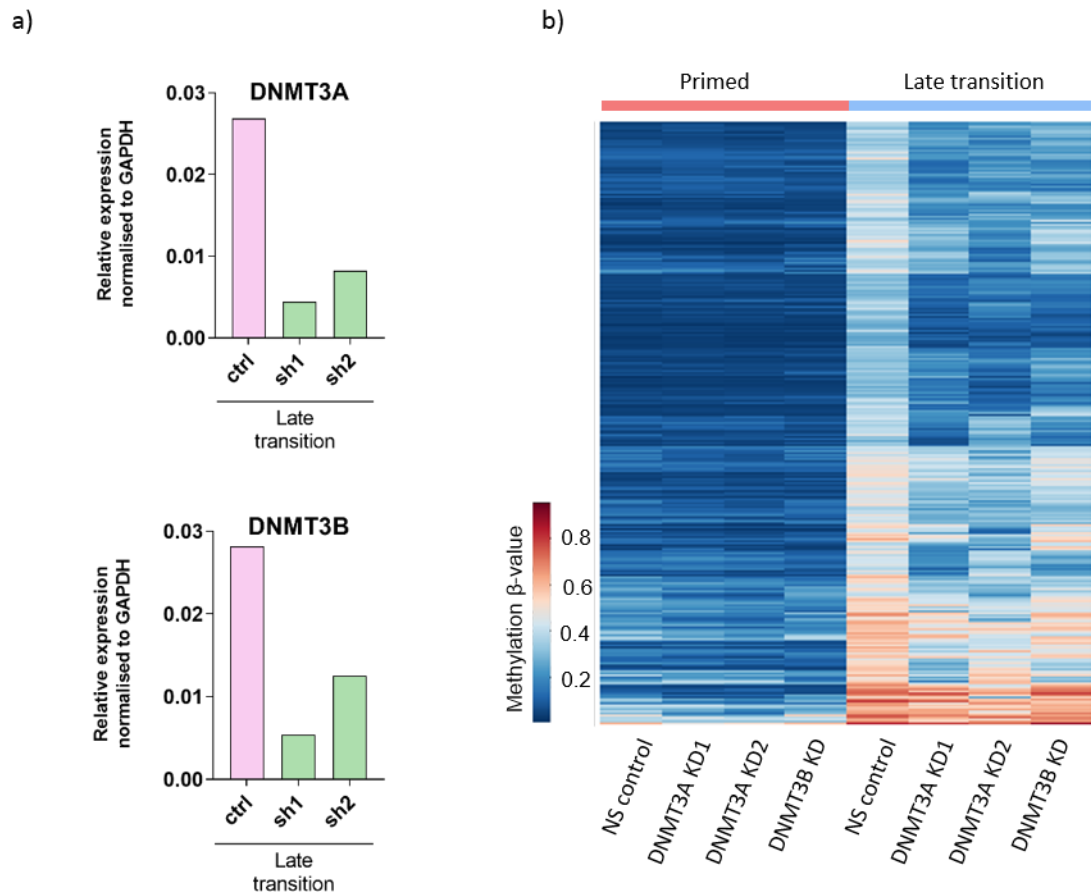


**Figure 4.5. Hypermethylation is carried out by the dominantly expressed DNMT3A2 isoform.** a) Genome browser screenshot depicting the transcripts encoding the main isoforms of DNMT3A. The longer isoform, DNMT3A1 has an extended N-terminal domain compared to the shorter isoform, DNMT3A2. b) qRT-PCR for DNMT3A gene isoforms across the transition of hESC reprogramming. Bars are representative of the mean of three technical replicates and three independent biological replicates. Error bars represent the standard error of the mean. Statistical difference between samples was analysed by a two-tailed Student's t-test, with a Bonferroni post-hoc test comparing each time point to primed hESCs. \* $P < 0.05$ , \*\* $P < 0.01$  and \*\*\* $P < 0.001$ . Human GAPDH was used to normalise expression. c) qRT-PCR for the transcripts of the DNMT3A isoforms in primed and early transition hESCs for a control and a DNMT3A1 knockdown cell line. Bars represent the mean of three technical replicates. Human GAPDH was used to normalise expression. d) Targeted bisulfite-sequencing of 4 regions of DNA. Each square represents the methylation % indicated by the colour key of a single CpG. Data is shown for a non-silencing (NS) control and knockdown of DNMT3A1 in primed and early transition hESCs.

#### 4.4 DNMT3A and DNMT3B contribute to hypermethylation during the late transition of reprogramming

As our analysis thus far had been focused on the early transition of reprogramming, we next subjected the *DNMT3A* and *DNMT3B* knockdown cell lines to reprogramming until the late transition of reprogramming and measured the expression of the two genes to validate that the genes were still knocked down (Figure 4.6a). We then performed an Infinium MethylationEPIC BeadChip array to observe the impact of a knockdown of *DNMT3A* and *DNMT3B* on hypermethylation. Once again, as the numbers of replicates used for this experiment were insufficient to perform DMP analysis between the knockdowns and controls in primed and late transition hESCs, we extracted methylation beta-values for the probes identified as hypermethylated ( $\Delta\beta > 0.1$ ,  $p < 0.05$ ) in wild-type early transition hESCs compared to primed hESCs (of which the top 10,000 were presented in Figure 3.5 of Chapter 3), and plotted these as a heatmap. Interestingly, at this stage of reprogramming, while *DNMT3A* knockdown cells show reduced levels of hypermethylation compared to the control (Figure 4.6b), *DNMT3A* is not solely accountable for all the hypermethylation present. *DNMT3B*, which by this stage is transcriptionally expressed at a higher level than during the early transition (Figure 4.3), also contributes to hypermethylation, as *DNMT3B* knockdown cells also fail to hypermethylate to the same extent as control cells (Figure 4.6b).

Collectively, these data show that *DNMT3A*, and more specifically the shorter isoform *DNMT3A2*, is responsible for de novo methylation during the early transition of hESC reprogramming. However, by the late transition of reprogramming at which point we detect the peak of hypermethylation and the strongest overlap with cancer hypermethylation (shown in chapter 3), both *DNMT3A* and *DNMT3B* seem to contribute to hypermethylation.



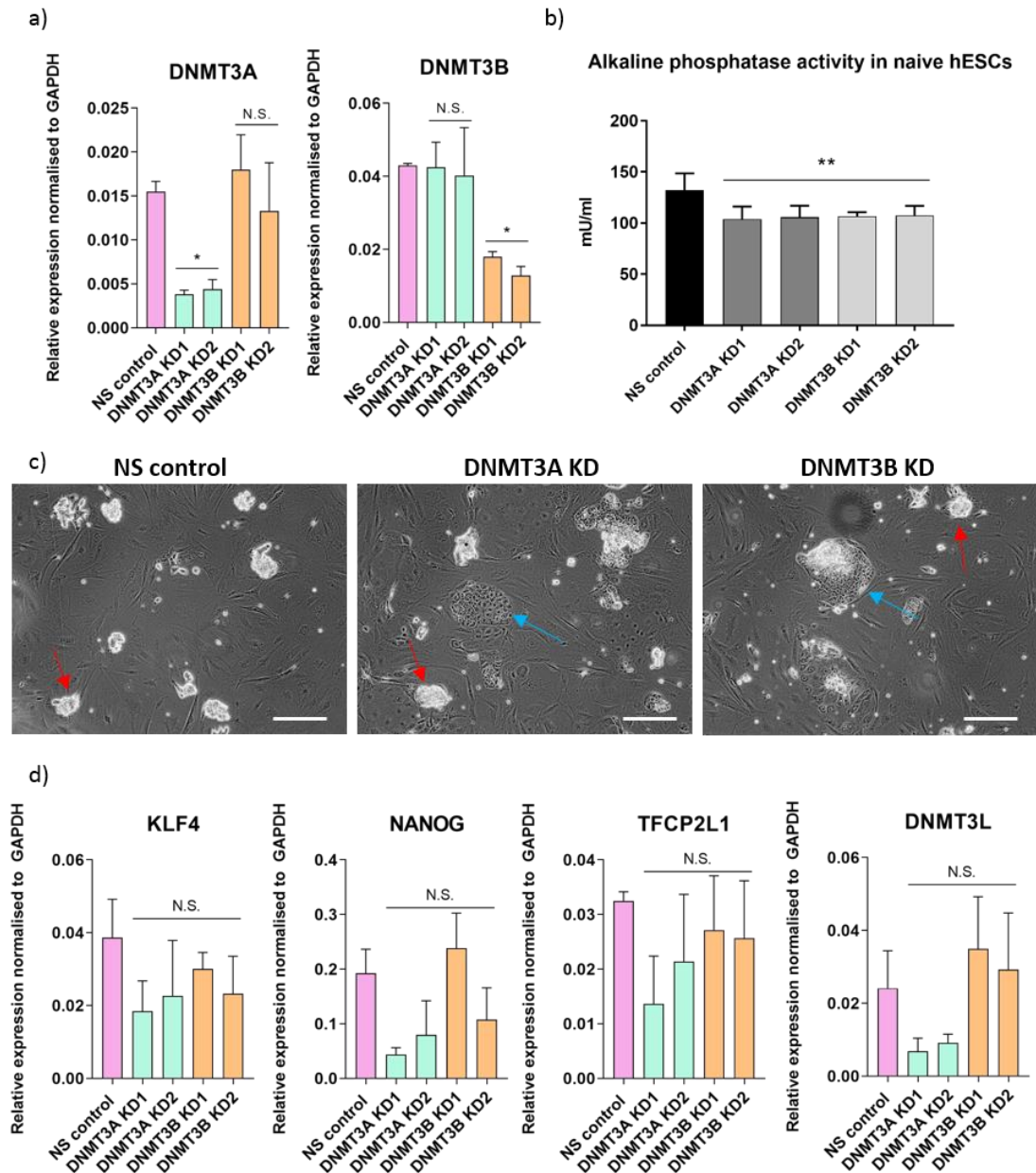
**Figure 4.6. Both DNMT3A and DNMT3B contribute to hypermethylation during the late transition of reprogramming.** a) qRT-PCR for the transcripts of DNMT3A and DNMT3B in primed and early transition hESCs for a control knockdown cell lines. Bars represent the mean of three technical replicates. Human GAPDH was used to normalise expression. b) Heatmap showing methylation levels for primed and late transition control and DNMT3A/B knock down hESCs. Heatmap is composed of the top 20,000 CpG probes that are differentially methylated ( $\Delta\beta > 0.1$ ,  $p < 0.05$ ) in the wild-type late transition compared to primed hESCs. Methylation  $\beta$ -value is indicated by the colour key.

#### 4.5 Hypermethylation has a functional role in naïve pluripotency

Given that both DNMT3A and DNMT3B knockdown cells display reduced hypermethylation compared to control cells by the late transition of reprogramming, we reasoned that we could use these cells to investigate whether hypermethylation plays a functional role during reprogramming or in naïve hESCs. We further reprogrammed the DNMT3A and DNMT3B knockdown cells to the stable naïve state, and measured the expression of the two genes to validate that the genes were still knocked down (Figure 4.7a). We performed an alkaline phosphatase assay on the naïve DNMT3A and DNMT3B knockdown cells, alongside the control hESCs. High alkaline phosphatase activity is a marker for an undifferentiated pluripotent stem cell state (O'Connor et al. 2008). We saw a reduction in the alkaline phosphatase activity of

both the DNMT3A and DNMT3B knockdown cells compared to control cells (Figure 4.7b), suggesting that pluripotency is impaired in the knockdown naïve cells. Additionally, while control cells form dome-shaped colonies as is typical of naïve hESCs, the DNMT3A and DNMT3B knockdown naïve hESCs grow in a mixture of both dome-shaped and more flattened colonies (Figure 4.7c). This may indicate either failure of the knockdown cells to successfully transition to the naïve state, or that the lack of hypermethylation in the absence of DNMT3A or DNMT3B impacts the stability of the naïve hESCs. We further measured the gene expression of a panel of genes that are markers of naïve pluripotency. While we did not see a consistent significant change in the expression of pluripotency markers in DNMT3A and DNMT3B knockdown hESCs compared to control cells, there seems to be a general trend of reduced expression of pluripotency markers, particularly in the DNMT3A knockdown cells, which may be more apparent with further biological replicates (Figure 4.7d). Collectively, these data indicate a putative role of *de novo* methylation by DNMT3A and DNMT3B either in the conversion to the naïve state or in stabilization of naïve pluripotency.

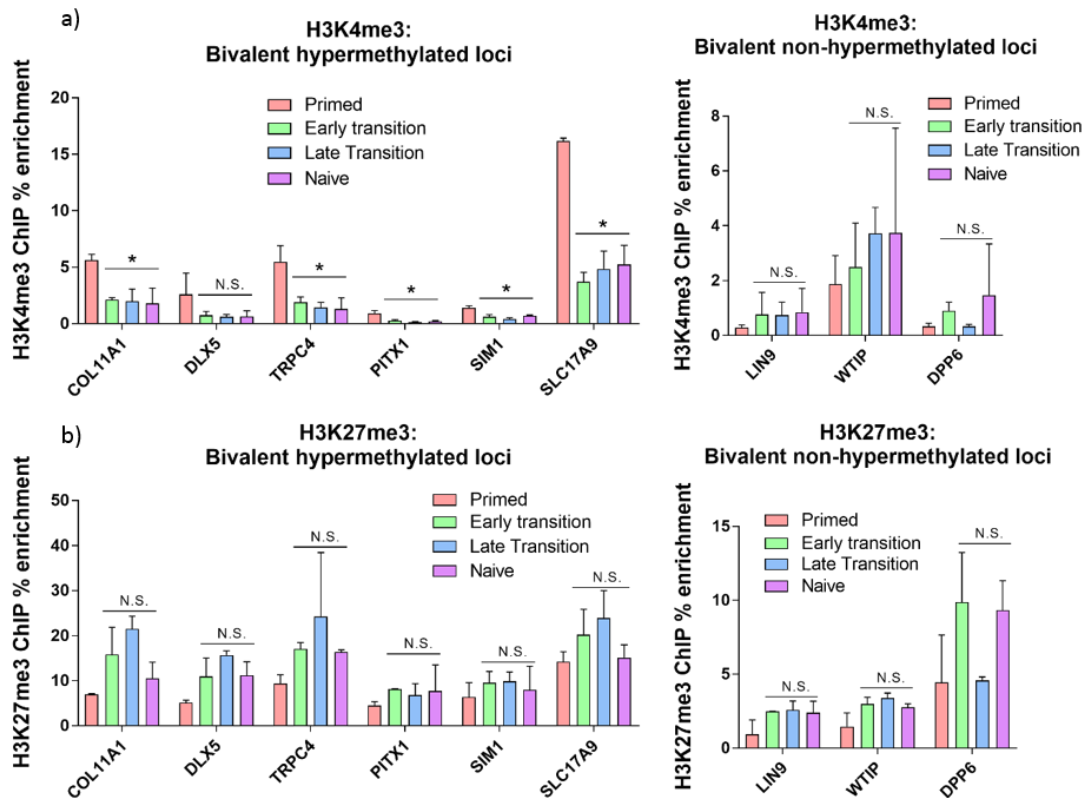




**Figure 4.7. Knockdown of DNMT3A and DNMT3B impacts naïve pluripotency.** a) qRT-PCR for DNMT3A and DNMT3B in control and knockdown naïve hESCs. Bars are representative of the mean of three technical replicates and two independent biological replicates. Error bars represent the standard error of the mean. Statistical difference between samples was analysed by one way ANOVA, with a Bonferroni post-hoc test comparing each time point to the non-silencing (NS) control cells. \* $P < 0.05$ , N.S. denotes not significant. Human GAPDH was used to normalise expression. b) Alkaline phosphatase activity measured in DNMT3A and DNMT3B knock down and control naïve hESCs. Data shown are the mean of 2 biological replicates, each with 5 technical replicates. Error bars represent SEM. Statistical difference between samples was analysed by a one way ANOVA with a Bonferroni post-hoc test compared to the control. \*\* $P < 0.01$ . c) Brightfield images of non-silencing (NS) control, DNMT3A and DNMT3B knockdown naïve hESCs showing their different morphologies. Red arrows indicate naïve dome-shaped colonies and blue arrows indicate flattened colonies. Scale bars represent  $125\mu\text{m}$ . d) qRT-PCR for naïve pluripotency markers in control and knockdown naïve hESCs. Bars are representative of the mean of three technical replicates and two independent biological replicates. Error bars represent the standard error of the mean. Statistical difference between samples was analysed by one way ANOVA, with a Bonferroni post-hoc test comparing each time point to the non-silencing (NS) control cells. N.S. denotes not significant. Human GAPDH was used to normalise expression.

#### 4.6 Hypermethylation of bivalent loci correlates with loss of H3K4me3

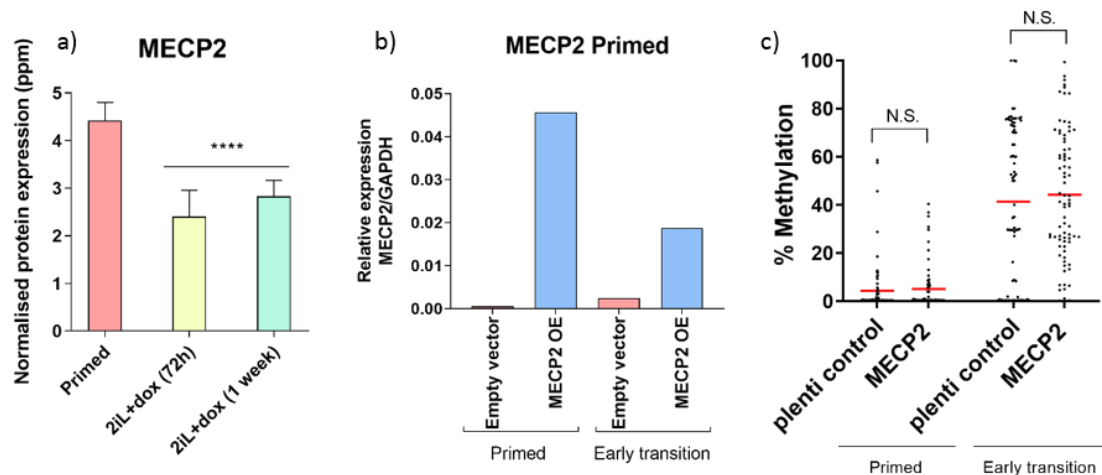
The interplay between bivalent chromatin and DNA methylation has been referred to many times, particularly in the context of cancer. Moreover, our data thus far had highlighted an enrichment of hypermethylated regions in genomic loci marked by H3K4me3 and H3K27me3 in primed hESCs. To investigate any relationship between bivalent chromatin and DNA methylation upon hESC reprogramming, we used the grouping of bivalent regions generated in chapter 3 to address the intrinsic differences between the bivalent hypermethylated and bivalent non-hypermethylated groups, which both begin with the same chromatin states. As DNA methylation and H3K4me3 are known to be mutually exclusive (Ooi et al. 2007), we used a targeted approach to measure the enrichment of the modified histone at bivalent DNA regions. We performed CHIP-qPCR of H3K4me3, across the transition from primed to naïve pluripotency. We observed a loss of H3K4me3 at bivalent regions that become hypermethylated, whilst bivalent non-hypermethylated regions retain their levels of H3K4me3 (Figure 4.8a). The loss of H3K4me3 is already apparent by the early transition of reprogramming. It is known that loss of H3K4me3 is permissive to the gain of DNA methylation, but this on its own cannot explain the specific gain of methylation at these regions, as both DNMT3A and DNMT3B possess an ADD domain capable of mediating the interaction of the enzymes with unmethylated H3K4 (Otani et al. 2009; Ooi et al. 2007; Guo et al. 2015; Zhang et al. 2010). Despite comparable absolute protein levels of DNMT3A and DNMT3B, as measured by mass spectrometry (Figure 4.4d), only DNMT3A deposits DNA methylation during the early transition of hESC reprogramming. In contrast to the reduction in H3K4me3, the levels of H3K27me3 exhibit little change at bivalent regions, despite the presence of DNA methylation (Figure 4.8b). While H3K27me3 and DNA methylation are considered to be mutually exclusive at CpG rich regions during development (Brinkman et al. 2012), co-existence of the two modifications has previously been reported in the context of cancer (Gao et al. 2014).



**Figure 4.8. A subset of bivalent loci lose H3K4me3 and gain DNA methylation.** a) ChIP-qPCR enrichment of H3K4me3 and b) H3K27me3 are shown for candidate bivalent regions (possessing both H3K4me3 and H3K27me3 histone modifications in primed cells) that do (6 regions) and do not (3 regions) become hypermethylated during reprogramming. Data is shown as the signal enrichment relative to the input sample and error bars represent the minimum and maximum values from 2 independent experiments. Statistical difference between samples was analysed by a one way ANOVA test, with Bonferroni post-hoc test of each time point compared to primed hESC. \* indicates  $p < 0.05$ , N.S. denotes not significant.

A recent study demonstrated that the binding of the DNA methylation reader protein, MECP2, to the ADD domain of DNMT3A inhibits the activity of the DNA methyltransferase by stabilising its auto-inhibitory confirmation (Rajavelu et al. 2018). At genomic sites with unmodified H3K4, the binding of H3 to the ADD domain of DNMT3A disrupts the interaction between DNMT3A and MECP2, thereby allosterically activating DNMT3A and enabling *de novo* DNA methylation (Rajavelu et al. 2018). Mass spectrometry based protein quantification showed that MECP2 is expressed in primed hESCs and is downregulated upon reprogramming, coinciding with the reduction of H3K4me3 at bivalent DNA loci (Figure 4.9a). To test whether the concomitant loss of MECP2 and H3K4me3 enable *de novo* DNA methylation upon hESC reprogramming, we generated a stable MECP2-overexpressing cell line in primed hESCs and subjected it to reprogramming until the early transition. We validated overexpression by measuring transcriptional expression of MECP2 in primed and early transition hESCs (Figure 4.9b). We

then used a targeted-bisulfite sequencing approach to measure DNA methylation at selected loci. We observed no impact of MECP2 overexpression on DNA methylation levels in primed or early transition hESCs (Figure 4.9c). These data indicate that hypermethylation of bivalent loci upon reprogramming occurs independently of MECP2 loss.

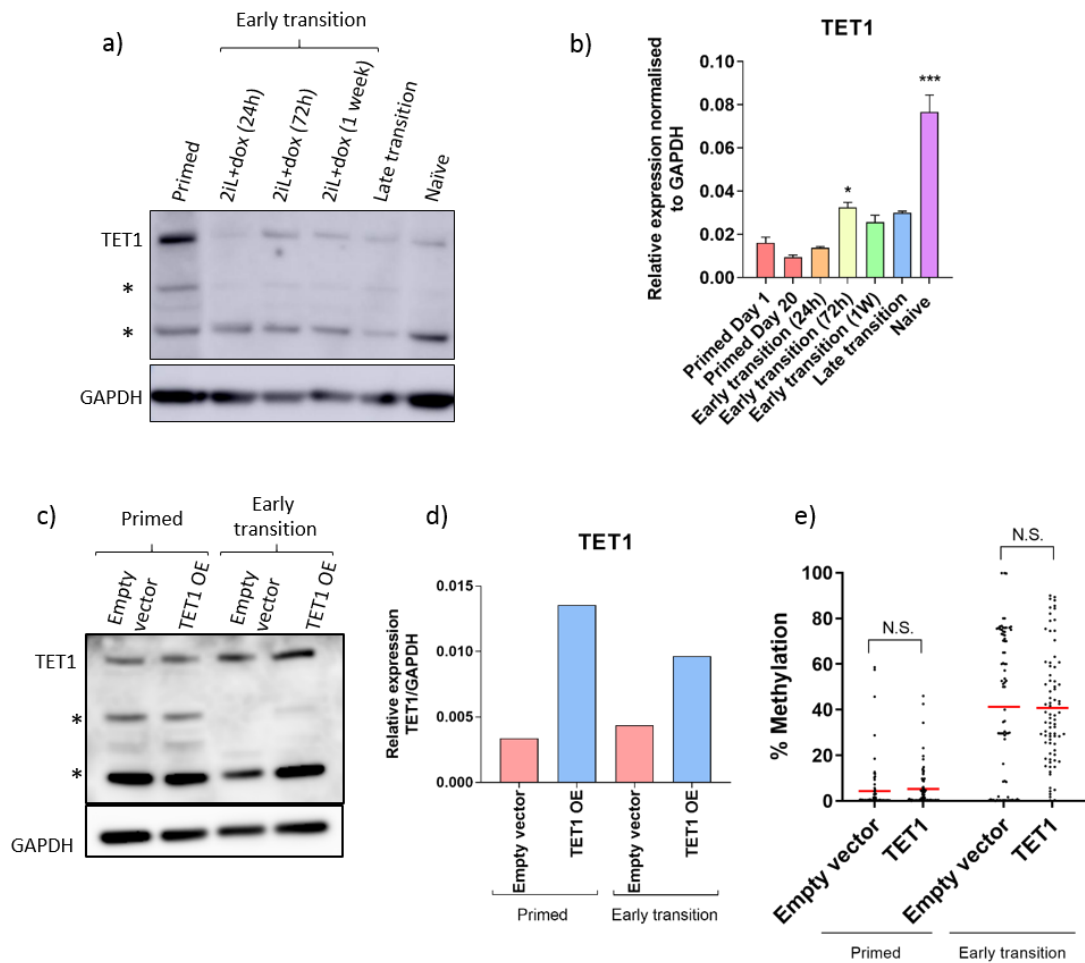


**Figure 4.9. Hypermethylation occurs independently of MECP2 loss.** a) Raw protein intensity values for MECP2 extracted from proteomics data, plotted in parts per million (ppm). Bars represent an average of 2 intensity values for each of the 3 replicates per sample, with error bars representing the SEM. Statistical difference between samples was analysed by a one-way ANOVA with a Bonferroni post-hoc test comparing each early transition time point to primed hESCs. \*\*\*\* $P < 0.0001$ . b) qRT-PCR for TET1 in an empty vector control and overexpression cell line, in primed and early transition cells. Bars represent the mean of three technical replicates. Human GAPDH was used to normalise expression. c) Plot showing the % methylation in the early transition and primed state, for a MECP2 overexpressing hESC line and empty vector control. Each dot represents the methylation % of single CpGs from 6 genomic regions and the red bars represent the mean methylation level for each sample. Statistical difference between samples was analysed by a two way ANOVA test, with Bonferroni post-hoc test of the MECP2 overexpressing sample compared to the control (empty vector). N.S. denotes not significant ( $p > 0.05$ ).

#### 4.7 DNA hypermethylation is independent of TET1 loss

We next sought to investigate the hypothesis that the loss of protective mechanisms that typically act to maintain CGIs in a hypomethylated state may result in hypermethylation. Bivalent loci in mice have been shown to be protected from DNA methylation through binding of ten-eleven translocation (TET) enzymes via their CXXC domains (Xu et al. 2011), and a recent study demonstrated that TET triple knock out in primed hESCs results in hypermethylation of bivalent loci (Verma et al. 2018). TET2 does not have a CXXC domain, and TET3 is expressed at very low levels in hESCs. TET1, however, which has a functional CXXC domain is expressed in primed hESCs and is subsequently downregulated at the protein level as hESCs progress through the early transition of reprogramming (Figure 4.10a). This protein expression pattern

appears to be uncoupled from the transcriptional expression of TET1 (Figure 4.10b), suggesting that the expression of TET1 protein is controlled by a post-transcriptional mechanism. To test whether loss of TET1-mediated antagonism of DNA methylation is responsible for hypermethylation of bivalent loci upon reprogramming, we generated a stable TET1-overexpressing cell line in primed hESCs and subjected it to reprogramming until the early transition. We validated overexpression by measuring both transcriptional and protein expression of TET1 in primed and early transition hESCs (Figures 4.10c and 4.10d). We then used a targeted-bisulfite sequencing approach to measure DNA methylation at selected loci. We observed no impact of TET1 overexpression on DNA methylation levels in primed or early transition hESCs (Figure 4.10e). These data indicate that hypermethylation of bivalent loci upon reprogramming is independent of TET1 loss.

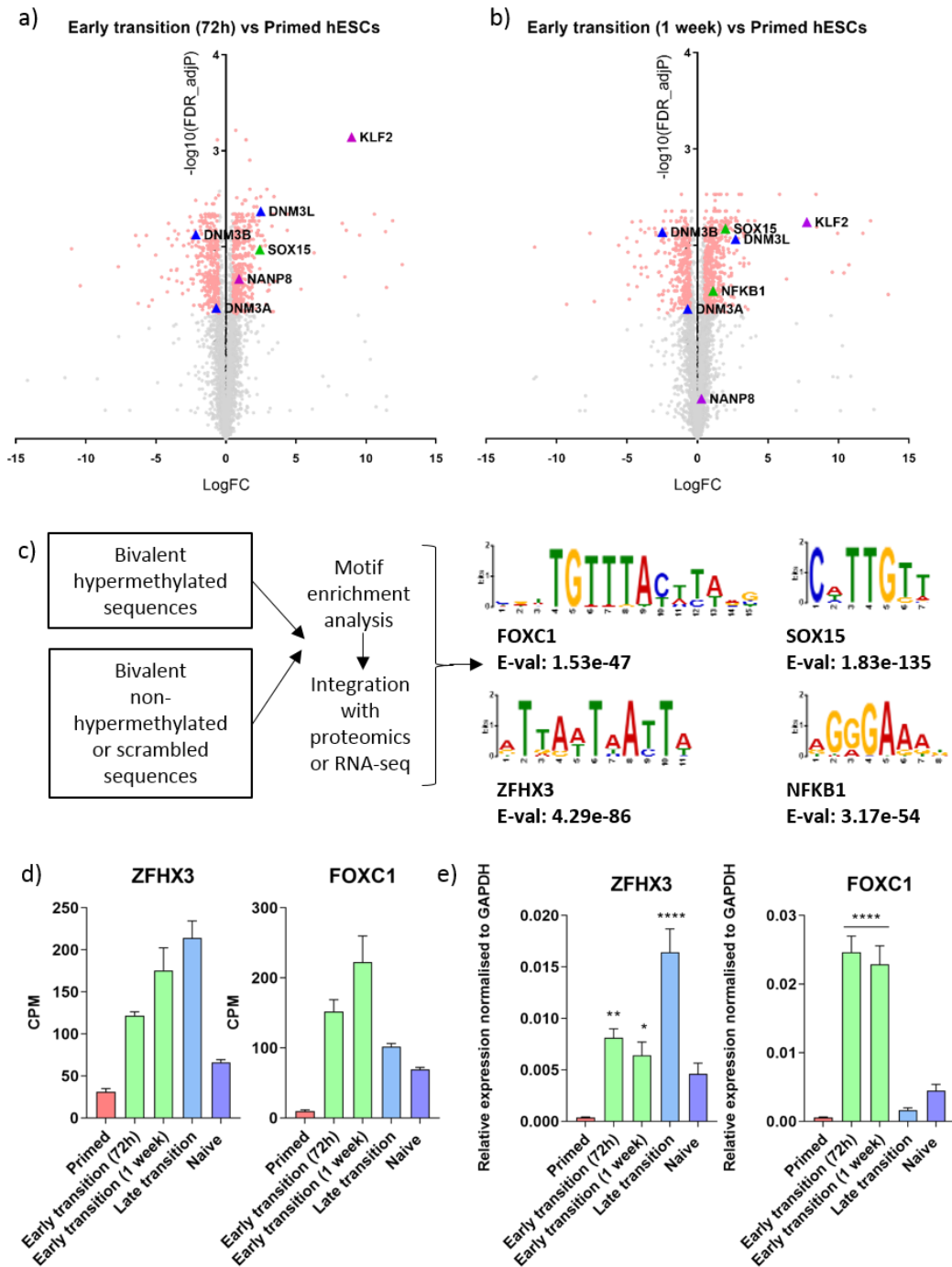


**Figure 4.10. Hypermethylation occurs independently of TET1 loss.** a) Western blot analysis of TET1 across the transition from primed to naïve pluripotency. \* denotes degraded fragments of TET1 protein. GAPDH is used as a loading control. b) qRT-PCR for TET1 across the period of resetting. Bars are representative of the mean of two technical replicates and three biological replicates. Error bars represent the standard error of the mean. Statistical difference between samples was analysed by a one way ANOVA test. \*P<0.05 and \*\*\*P<0.0001. Human GAPDH was used to normalise expression. c) Western blot analysis of TET1 in empty vector and TET1 overexpressing primed and early transition hESCs. \* denotes degraded fragments of TET1 protein. GAPDH is used as a loading control. d) qRT-PCR for TET1 in an empty vector control and overexpression cell line, in primed and early transition cells. Bars represent the mean of three technical replicates. Human GAPDH was used to normalise expression. e) Plot showing the % methylation in the early transition and primed state, for a TET1 overexpressing hESC line and empty vector control. Each dot represents the methylation % of single CpGs from 6 genomic regions and the red bars represent the mean methylation level for each sample. Statistical difference between samples was analysed by a two way ANOVA test, with Bonferroni post-hoc test of the TET1 overexpressing sample compared to the control (empty vector). N.S. denotes not significant (p > 0.05).

#### 4.8 DNA hypermethylation is coordinated by the transcription factor and core pluripotency networks.

Having seen no evidence thus far for loss of protective mechanisms in facilitating DNA hypermethylation upon reprogramming, we hypothesized that an additional player, likely a DNA-binding factor, is involved in targeting DNMT3A-mediated hypermethylation in the early

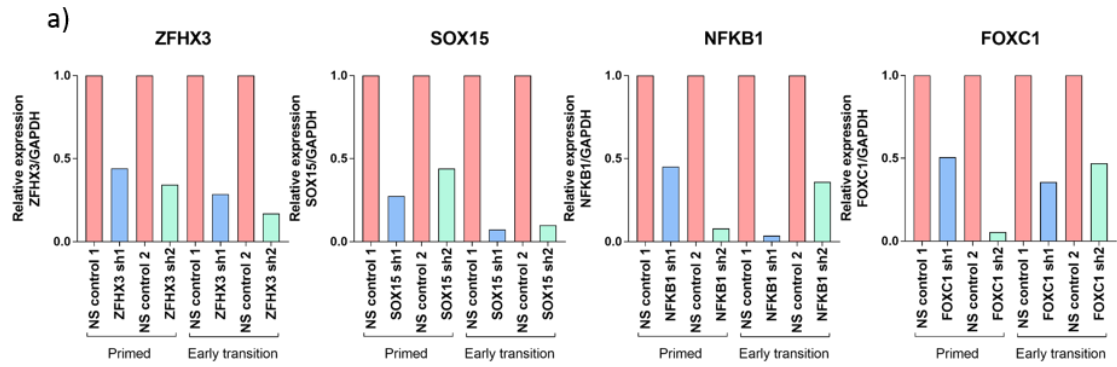
stages of hESC reprogramming. We reasoned that a more thorough understanding of the unique properties of the bivalent hypermethylated group compared to the bivalent non-hypermethylated group may provide insight into such a mechanism. We thus performed differential transcription factor-binding motif analysis of bivalent hypermethylated DNA regions, with the bivalent non-hypermethylated regions as a control set. This analysis performs a search of known DNA-binding motifs in the genomic sequences input, by searching against a database of human transcription factor motifs. We identified a large number of DNA-binding transcription factors with motifs enriched at regions that undergo hypermethylation. In order to play a role, these proteins must be expressed in the cells, particularly at the time at which hypermethylation occurs. We therefore analysed the total proteomics of primed hESCs and hESCs at two time points during the early transition of reprogramming. In total, we identified 406 proteins that were significantly differentially expressed ( $\text{Log}_2 \text{FC} > 1$  or  $< -1$ ,  $\text{FDR} < 0.05$ ) 72 hours into the early transition compared to primed hESCs, and 424 proteins that were significantly differentially expressed ( $\text{Log}_2 \text{FC} > 1$  or  $< -1$ ,  $\text{FDR} < 0.05$ ) 1 week into the early transition compared to primed hESCs (Figures 4.11a and 4.11b). Additionally, we were able to validate that the experiment had been successful technically, as we could detect overexpression of NANOG and KLF2 in early transition cells, as well as a significant upregulation of DNMT3L. We were also able to detect downregulation of both DNMT3A and DNMT3B, to varying degrees (Figures 4.11a and 4.11b). In order to identify proteins with a putative role in hypermethylation, we filtered the list of proteins identified to specifically look at those that are upregulated upon reprogramming. We next performed a cross-comparison of results of the motif analysis with the upregulated proteins from the proteomic analysis. We short-listed two candidate transcription factors, SRY-box 15 (SOX15) and Nuclear factor kappa B subunit 1 (NFKB1), which are upregulated during early reprogramming and show an enrichment of binding sites at hypermethylated regions (Figures 4.11a-c). We performed a similar cross-comparison of the results from the motif analysis with the RNA-seq data described in chapter 3. Through this analysis, we identified an additional two candidate transcription factors, Forkhead box C1 (FOXC1) and Zinc finger homeobox 3 (ZFHX3; Figure 4.11c), which were transcriptionally upregulated during the early transition based on RNA-seq data (Figure 4.11d) but not detected in any samples by proteomic analysis, likely due to technical limitations of the method in detecting nuclear transcription factors (Simicevic and Deplancke 2017). We measured the expression of the transcripts of ZFHX3 and FOXC1 by qRT-PCR to confirm their upregulation during the early transition of reprogramming (Figure 4.11e).



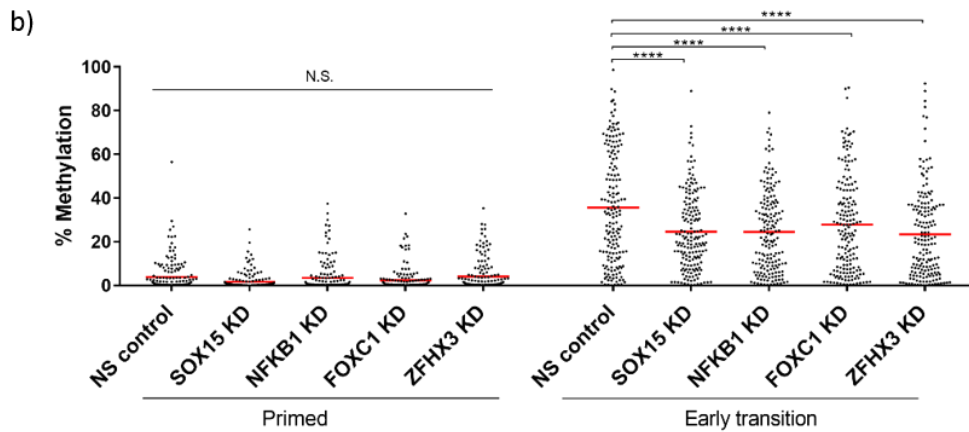
**Figure 4.11. Transcription factors that bind hypermethylated loci are upregulated upon reprogramming.** a&b) Volcano plots showing the difference in protein expression in early transition (72h and 1W) hESCs compared to primed hESCs (Supplementary Table 6). Each dot represents the log2 fold change based on three biological replicates. Statistical difference between samples was analysed by a student's t-test, corrected for multiple testing. Red dots indicate statistically significant changes ( $\text{adjP} < 0.05$ ). Proteins of interest are highlighted with coloured and labelled symbols. c) A selection of the transcription factors with motifs enriched in bivalent hypermethylated regions, with bivalent non-hypermethylated regions used as a background control. Motif analysis was performed using the analysis of motif enrichment (AME) tool on the MEME suite. d) Expression levels of FOXC1 and ZFH3 extracted from RNA-seq data, displayed in normalised counts per million (CPM). e) qRT-PCR for FOXC1 and ZFH3 across the period of reprogramming. Bars are representative of the mean of three technical replicates and three biological replicates. Error bars represent the standard error of the mean. Statistical difference between samples was analysed by a one way ANOVA test. \* $P < 0.05$ , \*\* $P < 0.01$ , \*\*\* $P < 0.001$  and \*\*\*\* $P < 0.0001$ . Human GAPDH was used to normalise expression.



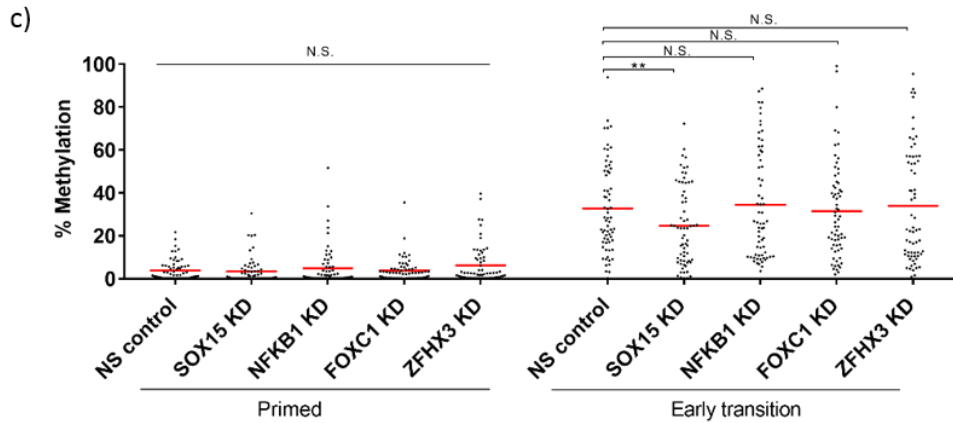
In order to determine whether our candidate transcription factors play a role in facilitating hypermethylation, we generated constitutive knockdown cell lines using two short hairpin RNAs (shRNAs) targeting each of the four candidate genes, and subjected each of the cell lines to reprogramming until the early transition. We validated the efficiency of the shRNA knockdowns by measuring expression of the target genes in both primed and early transition hESCs (Figure 4.12a). We then used a targeted-bisulfite sequencing approach to measure DNA methylation at selected loci, assaying both loci where the candidate transcription factors are predicted to bind and loci with no predicted binding sites. There was no effect on the level of methylation in primed hESCs in any of the knockdown cell lines compared to the control. Strikingly, however, upon reprogramming, knockdown of each of the transcription factors was able to reduce the level of hypermethylation at target loci analysed, suggesting that the early reprogramming transcription factor network coordinates bivalent promoter CGI hypermethylation (Figure 4.12b). The impact seems to be higher in regions where the highly expressed transcription factors are also predicted to bind, indicating a network synergy in preferentially mediating *de novo* methyltransferase recruitment to these sites (Figure 4.12c). These data support the hypothesis that a DNA-binding event may be the first step required in *de novo* DNA methylation, to actively target DNMT activity to specific genomic sites.



Loci with TF binding sites

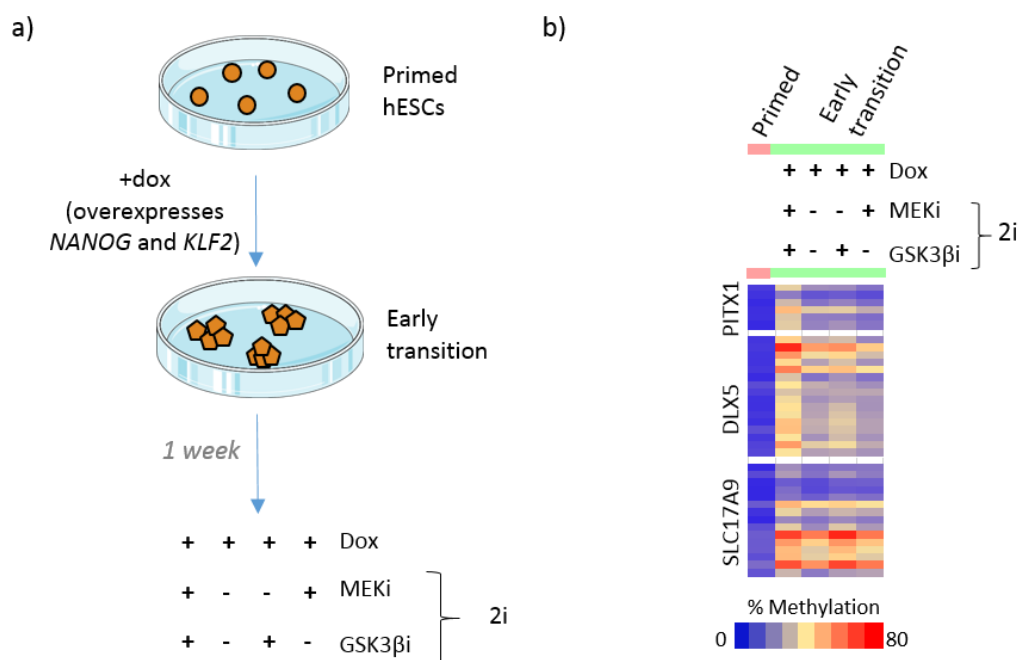


Loci without TF binding sites



**Figure 4.12. Knockdown of transcription factors reduces hypermethylation at target loci.** a) qRT-PCR for the transcripts of 4 transcription factors in control and knock-down cell lines, in primed and early transition cells. Bars represent the mean of three technical replicates. Human GAPDH was used to normalise expression. b&c) Plots showing the % methylation in the early transition and primed state, for 4 different transcription factor knock downs and a non-silencing control. Data for each sample are an average of 2 independent shRNA knock downs. Each dot represents the methylation % of single CpGs from 12 genomic regions (with TF binding sites) or 4 genomic regions (without TF binding sites) and the red bars represent the mean methylation level for each sample. Statistical difference between samples was analysed by a two way ANOVA test, with Bonferroni post-hoc test of each TF knock down compared to the control. \* $P < 0.05$ , \*\* $P < 0.01$ , \*\*\* $P < 0.001$ , \*\*\*\* $P < 0.0001$ . N.S. denotes not significant.

As reprogramming hESCs to the naïve state is a multifaceted process involving a number of inhibitors and growth factors that affect the signalling network of hESCs, we sought to test whether signalling changes associated with factors required for induction of the naïve state induction could influence hypermethylation. We conducted hESC reprogramming until the early transition, each time removing one of the factors required for reprogramming (Figure 4.13a). We then used a targeted approach to measure DNA methylation at loci that become hypermethylated upon reprogramming. Reprogramming in the absence of the ERK pathway inhibitor or GSK3 $\beta$  inhibitor (also a WNT pathway agonist) or concomitant removal of both inhibitors still resulted in hypermethylation at target loci analysed (Figure 4.13b). As the factor common to all the conditions tested is doxycycline, this suggests that hypermethylation may be coordinated by the overexpressed *NANOG* and *KLF2* or the associated pluripotency network. Collectively, these data indicate that upon reprogramming hESCs to the naïve state, hypermethylation is driven by the transcription factor network that becomes active upon reprogramming, and that this may be synchronised by the core pluripotency network.



**Figure 4.13. Hypermethylation occurs independently of the 2i inhibitors.** A) Schematic showing the experimental setup to test the impact of the 2i inhibitors on hypermethylation during the early transition of hESC reprogramming. b) Targeted bisulfite-sequencing of 3 genomic regions. Each square represents the methylation % indicated by the colour key of a single CpG. The first column represents data from primed hESCs, and the subsequent columns represent data from early transition hESCs cultured in a variety of culture conditions indicated by the +/- symbols above.

## 4.9 Discussion

The data presented in this chapter detail the DNA methylation machinery involved in mediating the changing patterns of DNA methylation upon primed to naive hESC reprogramming. We observed downregulation of the maintenance methylation proteins DNMT1 and UHRF1, and upregulation of the TET enzymes, both of which temporally coincided with the global demethylation that occurs upon reprogramming. To this end, it has been demonstrated in mESCs that *TET1*<sup>-/-</sup> and *TET2*<sup>-/-</sup> ESCs are still able to demethylate their genomes, (Ficz et al. 2013) and that *TET1/2/3* triple knock out only affects a limited specific number of loci (von Meyenn et al. 2016). This suggests that *TET* enzymes are dispensable for global demethylation. In mESCs, the primary cause of global demethylation has been attributed to impaired methylation maintenance as a result of downregulated UHRF1 protein and a global loss of the H3K9me2 mark, which is responsible for the recruitment of DNMT1 to replicating DNA (von Meyenn et al. 2016). We hypothesize that impaired methylation maintenance due to downregulated UHRF1 protein and the resulting alteration in targeting and regulation of DNMT1 activity and stability is primarily responsible for the global demethylation induced upon hESC reprogramming. This is supported by the observation that a disruption of the interactions in the DNMT1/UHRF1/PCNA complex can result in tumorigenesis associated with global demethylation (Hervouet et al. 2010; Pacaud et al. 2014).

We identified DNMT3A as the *de novo* DNA methyltransferase responsible for the deposition of DNA methylation early during hESC reprogramming and we demonstrate loss of H3K4me3 specifically from bivalent regions that become hypermethylated. Intriguingly, the levels of H3K27me3 exhibit little change at bivalent regions, despite the presence of DNA methylation. H3K27me3 and DNA methylation are considered to be mutually exclusive at CpG rich regions during development (Brinkman et al. 2012), however co-existence of the two modifications has previously been reported in the context of cancer (Gao et al. 2014). It is known that loss of H3K4me3 is permissive to the gain of DNA methylation, but this on its own cannot explain the specific gain of methylation at these regions, as both DNMT3A and DNMT3B possess an ADD domain capable of mediating the interaction of the enzymes with unmethylated H3K4 (Otani et al. 2009; Ooi et al. 2007; Guo et al. 2015; Zhang et al. 2010). Interestingly, during the early transition of reprogramming, our proteomics data shows that both DNMT3A and DNMT3B are downregulated to varying degrees. Despite comparable absolute protein levels of DNMT3A and DNMT3B, as measured by mass spectrometry, only DNMT3A deposits DNA methylation during the early transition of hESC reprogramming. Generating a stable DNMT3B overexpression primed hESC cell line and reprogramming it to the early transition would

provide verification that protein expression levels of DNMT3B are not the limiting factor preventing the enzyme from contributing to DNA hypermethylation during the early transition of reprogramming.

We deduced that the shorter isoform of DNMT3A, DNMT3A2, is specifically responsible for *de novo* methylation upon reprogramming, as we observed no impact on levels of hypermethylation when we knocked down DNMT3A1. Due to the lack of unique exons in DNMT3A2, we were unable to knock down this isoform and show a direct impact on DNA hypermethylation. An alternative to confirm the unique role of DNMT3A2 would be to generate a DNMT3A knockout cell line and re-introduce specifically the expression of DNMT3A2 and measure DNA methylation at bivalent CGIs upon reprogramming.

The data indicate that DNMT3B also contributes to hypermethylation by the late transition of reprogramming, as is often observed in the context of cancer cells (Roll et al. 2008). In line with this, naïve hESCs generated from either DNMT3A or DNMT3B knockdown primed hESCs exhibit reduced stability of naïve pluripotency. It is difficult to distinguish whether this is reflective of an inability to successfully transition to naïve pluripotency in the absence of the DNMT3s, or whether there is a direct impact on the stability of the naïve pluripotent state. Additionally, it is unclear whether the impact is dependent on the hypermethylation associated with the catalytic activity of the two enzymes, or is a result of a non-catalytic role of DNMT3A and DNMT3B. This could be further tested by using an independent method to erase DNA hypermethylation at bivalent CGIs, such as treatment of naïve hESCs with the DNA demethylating agent 5-azacytidine, and then evaluating of the impact of lack of hypermethylation on the stability of the naïve pluripotent state. Moreover, it would be useful to validate the findings of both DNMT3A and DNMT3B knockdowns by generating double knockout cell lines of the two enzymes, in order to be sure that any residual protein in the knockdown cell lines is not influencing the cellular phenotype, or that the two proteins are not compensating for each other.

We investigated the potential role of loss of protective mechanisms in facilitating DNA hypermethylation upon reprogramming. Specifically, we looked at the impact of TET1 overexpression on hypermethylation during the early transition of reprogramming. To validate overexpression of TET1, we measured both gene and protein expression of the enzyme in TET1-overexpressing and control cells. The gene expression data showed only a modest 2-fold increase in TET1 transcript levels at the early transition, compared to a 10-fold increase observed in MECP2 transcript levels in an MECP2-overexpression cell line generated using the

same overexpression system. Western blot analysis of the protein also showed a modest overexpression of TET1 in the overexpressing cells compared to the control, particularly when taking into account all the bands on the blot, which likely correspond to degradation products or isoforms of the protein. As the gene and protein expression of TET1 are uncoupled during hESC reprogramming, it is plausible that despite trying to overexpress TET1 mRNA, post-transcriptional or post-translational mechanisms retain overarching control over TET1 protein expression. This may explain the lack of impact seen on levels of DNA hypermethylation upon reprogramming in cells containing a TET1-overexpression construct. Further verification of TET1 protein expression by western blotting, or measurement of global levels of 5hmC as a readout of TET1 overexpression would enable more firm conclusions to be drawn with regards to the impact of TET1 loss on DNA hypermethylation.

We explored the hypothesis that hypermethylation of bivalent CGIs occurs as a result of active targeting of DNMT activity by chromatin or DNA-interacting proteins. It has been demonstrated that PRC2, which is responsible for deposition of H3K27me<sub>3</sub>, can directly regulate DNA methylation (Vire et al. 2006), however recent studies suggest that while PRC2 can recruit DNMT3A to the DNA, it is not sufficient to trigger *de novo* DNA methylation, indicating that additional factors are required (Rush et al. 2014). Our data points towards the transcription factor network established upon reprogramming playing a role in the targeting or recruitment of DNMT3A to loci that gain methylation. We identified four candidate transcription factors that were upregulated during the early transition of reprogramming and had motifs enriched at bivalent regions that become hypermethylated, of which SOX15 and NFKB1 were found to be upregulated as measured by mass spectrometry, while FOXC1 and ZFHX3 were not detected in any samples by mass spectrometry but were transcriptionally upregulated based on RNA-seq data. It is likely that as we performed total proteomics on whole cell lysates, technical limitations of the method in detecting nuclear transcription factors (Simicevic and Deplancke 2017) can explain why we did not detect FOXC1 and ZFHX3. Subcellular fractionation and subsequent proteomics of the nuclear fraction of proteins may provide a more accurate measure of the transcription factor network active upon reprogramming. Additionally, the expression of each of the candidate proteins can be validated by western blotting.

Whilst we cannot currently differentiate between a direct interaction of DNMT3A with transcription factors or an indirect transcription factor network-driven effect on targeting of the enzyme, loci-specific recruitment of DNMT3A via transcription factors has been previously demonstrated (Brenner et al. 2005) and *in vitro* data supports the ability of DNMT3A to

interact directly with numerous transcription factors (Hervouet, Vallette, and Cartron 2014). Moreover, this data is in support of the notion that underlying DNA sequence is important in the establishment of DNA methylation patterns (Long et al. 2016; Krebs et al. 2014; Lienert et al. 2011). Targeted ChIP of the transcription factors at bivalent DNA regions, and a co-immunoprecipitation of the transcription factors bound to DNMT3A would be useful to confirm whether the transcription factors actually bind their predicted binding sites during reprogramming, and whether they exhibit a direct interaction with DNMT3A.

Our approach for shortlisting candidate transcription factors with a putative role in the mechanism of hypermethylation relied on identifying transcription factors that were upregulated either at the level of mRNA or protein upon reprogramming. This approach is based on the assumption that an increase in the expression of a transcription factor is correlated with an increase in its activity or function. However, there is much evidence in the literature that transcription factors can undergo post-translational modifications that may alter their activity, localisation or ability to interact with other co-factors (Tootle and Rebay 2005; Filtz, Vogel, and Leid 2014). It may therefore be of additional value to investigate the impact on reprogramming-associated hypermethylation of transcription factors whose expression levels are not significantly altered upon hESC reprogramming.

Our data are also indicative of the overexpressed NANOG and KLF2 coordinating *de novo* methylation, however studies have shown that KLF2 is not expressed *in vivo* in the human inner cell mass (Yan et al. 2013; Blakeley et al. 2015), where we also observed hypermethylation (detailed in chapter 3). Additionally, we observed comparable hypermethylation in naïve hESCs generated using two transgene-independent methods of reprogramming (detailed in chapter 3). This collectively suggests that the core pluripotency network, to which NANOG belongs, is likely responsible for coordinating the transcriptional changes that drive DNA hypermethylation. There is growing evidence in the literature regarding the acquisition of stem-like properties and expression of pluripotency genes in cancers (Ben-Porath et al. 2008; Chen and He 2016). This makes it intriguing to speculate that a transcriptional programme associated with the pluripotency network could drive a shared mechanism of hypermethylation during reprogramming to naïve pluripotency and in cancer development, either preceding or in conjunction with genetic mutations. Additional molecular features of the primed to naïve state transition appear analogous to cancer hallmarks (Hanahan and Weinberg 2011), such as altered metabolism (Takashima et al. 2014), loss of imprints (Pastor et al. 2016), loss of DNA hydroxymethylation (Ficz and Gribben 2014) and genomic instability (Pastor et al. 2016; Liu et al. 2017). Whether they are related to the

changing epigenetic landscape remains unexplored, but further use of this model system may shed light on the emergence of these characteristics during cellular transformation. We propose that this may provide a good model system to understand whether the molecular processes associated with cellular reprogramming play a role in tumourigenesis. To this end, the next chapter will explore the genomic stability of hESCs upon reprogramming, as well as any potential relationship between the changing DNA methylation patterns and the genomic stability of the cells.



## Chapter 5. Results 3: Investigating genome stability during primed to naïve reprogramming

### 5.1 Introduction

CpG sites are underrepresented globally across the genome due to the high mutation rate of methylated cytosines. As CpG residues within CGIs are commonly unmethylated, CpG islands remain protected from the depletion of CpGs (Bird 1980; Bird et al. 1985). Cytosine can deaminate spontaneously to give uracil, and the resulting U:G mismatch is recognised by uracil DNA glycosylases and the uracil base is efficiently eliminated and replaced with cytosine. 5mC, however, is deaminated to give thymine (Bird 1980), and the resulting T:G mismatch is repaired with reduced efficiency (Schmutte et al. 1995). Compared to unmodified C, 5mC has increased susceptibility to spontaneously deaminate (Shen, Rideout, and Jones 1994). Aside from spontaneous deamination, the activation-induced cytidine deaminase (AID) is thought to mediate hydrolytic deamination of 5mC into T and C to U. The T is repaired if AID activity is coupled to DNA repair, resulting in active demethylation. The U:G mismatch, if replicated over will give rise to C to T mutations unless it is excised by UNG (Perez-Duran et al. 2012). Unlike U, however, because T is a natural base, the mutant base can persist through DNA replication, and upon cell division, it can be passed on to daughter cells as a C to T transition mutation (Schmutte et al. 1995). Despite the existence of MBD4, a dedicated thymine DNA glycosylase that can selectively remove T from a T:G mismatch, these mismatches are repaired with reduced efficiency (Schmutte et al. 1995; Hendrich et al. 1999).

Methylated CpGs are considered to be hotspots for DNA mutation, supported by the fact that C to T mutations are the most frequent in human disease, constituting a third of all point mutations (Cooper and Youssoufian 1988; Pfeifer 2017). In particular, it has been shown for the p53 gene, which is frequently mutated across many cancers, that the mutations frequently occur within the methylated CGIs within the gene promoter (Denissenko et al. 1997; Greenblatt et al. 1994). C to T mutations may occur in the coding regions of genes (Rideout et al. 1990), or on regulatory regions of DNA, hence altering binding sites of transcription factors and other proteins (Zemojtel et al. 2011). When the mutations occur in early-replicating DNA, they are typically repaired by BER machinery, but mutations can also occur in late-replicating DNA where BER operates with reduced efficiency, thereby resulting in mutation (Blokzijl et al. 2016; Tomkova et al. 2018).

While 5mC shows increased mutability compared to unmethylated C, its oxidised derivative, 5hmC (Tahiliani et al. 2009), has in fact been shown to be protective against mutagenesis

(Tomkova et al. 2016), and enrichment of 5hmC at sites of DNA damage has been shown to promote genome stability by ensuring chromosome segregation occurs correctly (Kafer et al. 2016).

Some of the most frequently seen mutational signatures in human cancers feature high rates of C to T mutations, and are attributed to the deaminase hyperactivity of AID or APOBEC enzymes at methylated CpG dinucleotides, possibly coupled with inefficient BER pathway activity (Alexandrov et al. 2013). Recent reports suggest that tissue-specific mutations accumulate in normal adult stem cells during life. These patterns are similar between the normal stem cells of a given tissue and cancers that arise from those tissues, suggesting that intrinsic, non-random mutational processes such as that described above may be responsible for initiating tumorigenesis (Blokzijl et al. 2016).

While DNA methylation may promote mutation due to the mutability of 5mC, its presence across the genome is beneficial in maintaining genomic and chromosomal stability (Chen et al. 1998). Recently, it has been reported that hypomethylation of late-replicating regions of the DNA begins during foetal development and is linked to cell division, such that cancer hypomethylation reflects the mitotic history of the cells (Zhou et al. 2018). Several studies have reported chromosomal breakages, fusions, and aneuploidy in relation to DNA hypomethylation and downregulation or mutations of the DNMTs, and such instability may promote tumour initiation or progression (Eden et al. 2003; Gaudet et al. 2003; Sheaffer, Elliott, and Kaestner 2016; Dodge et al. 2005; Karpf and Matsui 2005). As DNA methylation is involved in the silencing of retrotransposons during development (Bestor 2004), hypomethylation may also result in increased retrotransposition and re-expression of such elements, which may play an active role in tumorigenesis (Daskalos et al. 2009). The absence of DNA methylation at certain genomic regions such as heterochromatic repetitive elements may lead to incorrect chromosome segregation that gives rise to chromosomal aberrations and resulting aneuploidy, which is found across many tumour types (Monier, Mouradian, and Sullivan 2006; Duijf, Schultz, and Benezra 2013). Such chromosomal instability has been detected early on in the process of carcinogenesis and is a common feature of cancers (Shih et al. 2001; Kops, Weaver, and Cleveland 2005). A number of different types of chromosomal aberrations can be detected depending on the underlying molecular cause, which include various mitotic abnormalities that often arise as a result of cellular replication stress (Burrell et al. 2013). The aberrations can manifest as multipolar spindles arising from multiple centrioles, lagging chromosomes during anaphase, or ultrafine anaphase bridges, among others (Burrell et al. 2013; Kops, Weaver, and Cleveland 2005; Thompson, Bakhom, and Compton 2010).

Chromosomal aberrations such as those mentioned above can be detected in cells *in vitro* through a variety of methods including immunofluorescent staining of proteins that are associated with or bind to the various structural aberrations, as well as staining for centromeres. Additionally, as genomic instability can activate DNA damage pathways, detection of proteins associated with the DNA damage response are also used as a proxy for the level of DNA damage occurring in a cell. In particular, one of the most frequently used markers is the phosphorylation of the histone H2A variant H2AX at Ser 139 ( $\gamma$ -H2AX), as this phosphorylation is induced rapidly and abundantly upon the detection of a DNA double-stranded break (DSB) in a cell (Ji et al. 2017).

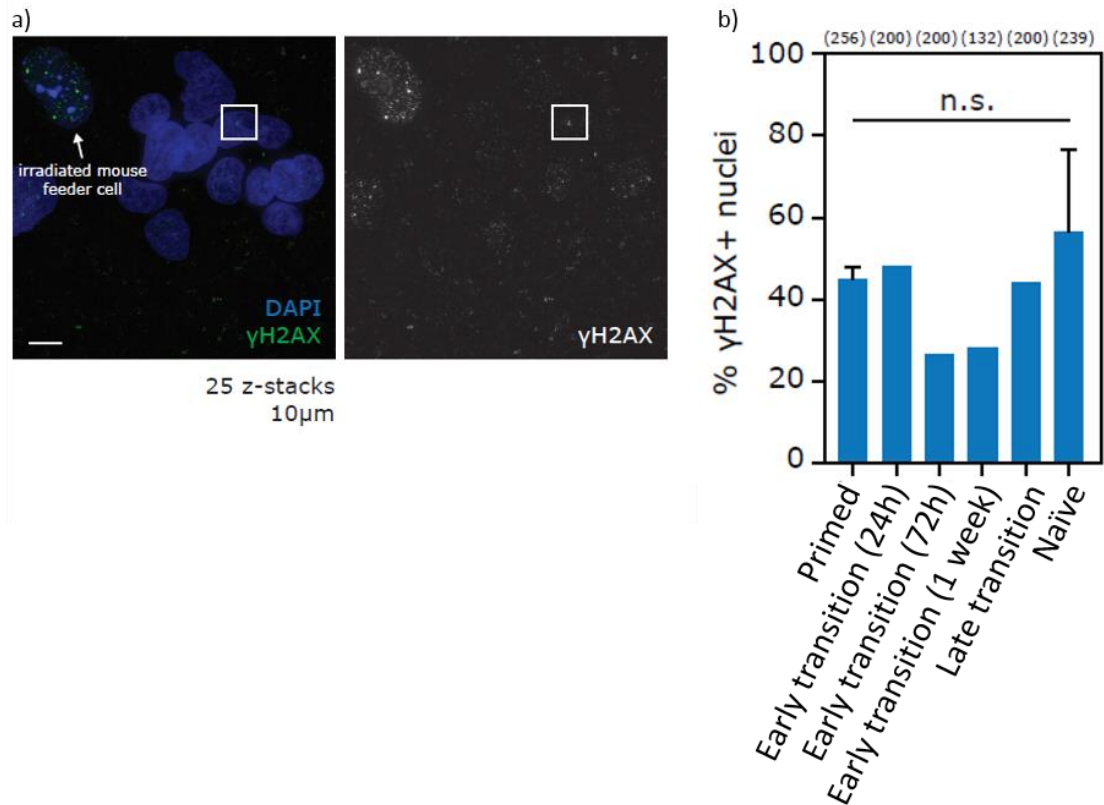
Primed to naïve hESC reprogramming is associated with large-scale changes in DNA methylation patterns, as detailed in chapter 3. Both global demethylation of the genome and regional hypermethylation may facilitate altered genomic stability of the cells. To this end, recent studies have reported reduced long-term genomic stability in naïve hESCs generated using alternative reprogramming methods, with naïve hESCs generated using the 5iLAF method particularly displaying abnormal karyotypes following multiple passages (Liu et al. 2017; Pastor et al. 2016; Guo et al. 2017). Such karyotypic instability has also been observed upon the generation of naïve human stem cells generated through iPS reprogramming (Kilens et al. 2018), in addition to increased chromosomal aberrations, somatic mutations and copy number variation (Hussein et al. 2011; Ji et al. 2012; Mayshar et al. 2010; Gore et al. 2011). Additionally, it has been previously reported that *de novo* point mutations occur during iPS reprogramming in a manner that indicates that they are attributable to oxidative stress-induced DNA damage (Yoshihara et al. 2017). However, it has also been demonstrated that limiting replication stress during cellular reprogramming either through genetic or chemical manipulation can reduce genomic instability (Ruiz et al. 2015).

This chapter will explore the genomic stability of hESCs during the reprogramming of primed to naïve hESCs, in the context of changing DNA methylation patterns. In particular, a potential relationship between global demethylation and structural genomic instability will be investigated with preliminary experiments, as well as a potential relationship between DNA demethylation and an increased mutation frequency. Additionally, a relationship between CGI hypermethylation and increased mutability of the CpG dinucleotide will be tested.

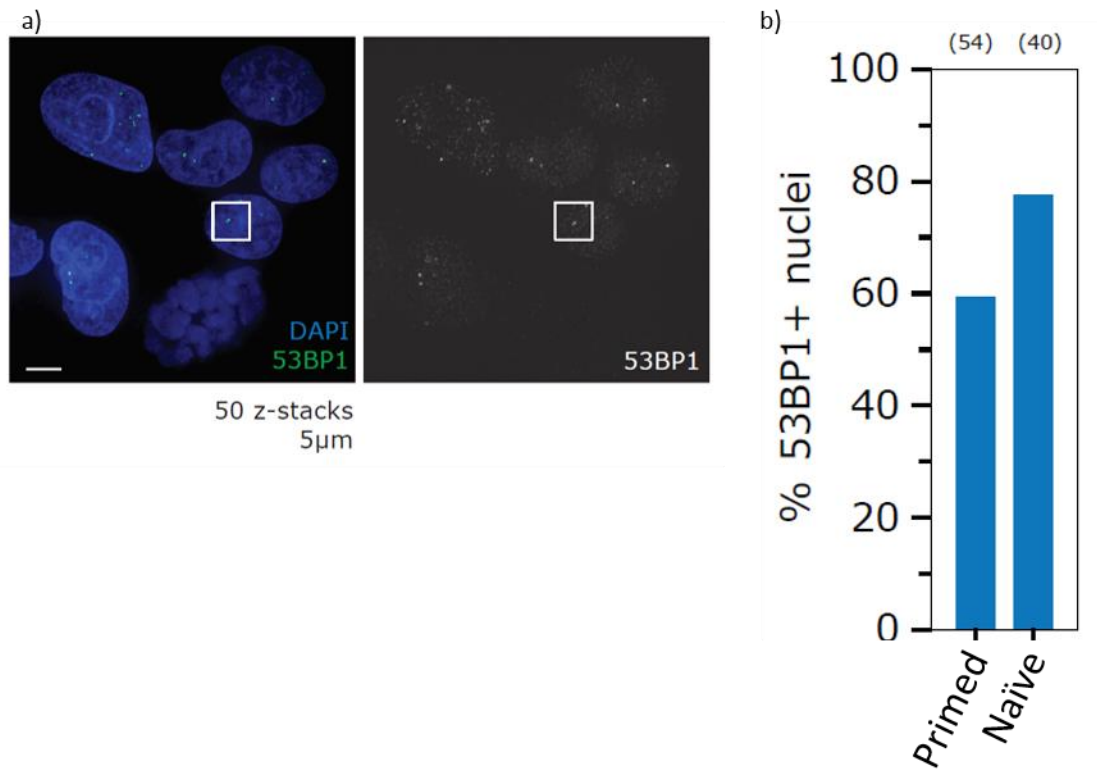
## 5.2 Levels of DNA damage markers remain constant across the transition from primed to naïve pluripotency

The DNA damage response (DDR) is typically activated in response to stress signals encountered during physiological processes. In particular, DSBs rapidly result in the accumulation of  $\gamma$ -H2AX, and the specific phosphorylation of serine 139 on H2AX is used as a marker for DNA damage (Ji et al. 2017). To understand whether hESCs undergo DNA damage upon reprogramming, we used immunofluorescence to measure the number of  $\gamma$ -H2AX foci present in hESCs across the transition from primed to naïve pluripotency. We observed no significant difference in the number of interphase nuclei positive for  $\gamma$ -H2AX at any point during the reprogramming transition compared to primed hESCs, though the basal level of  $\gamma$ -H2AX positive nuclei is high (Figure 5.1). This may be a reflection of human stem cells being highly sensitive to DNA damage and activating robust DNA repair responses against even low levels of DNA damage in order to maintain their genomic integrity (Fu et al. 2017; Seifert, Dejosez, and Zwaka 2017).

We next stained primed and naïve hESCs for p53 binding protein 1 (53BP1), which is also involved in DSB repair. Specifically, 53BP1 is known to regulate the DSB repair pathway choice, favouring non-homologous end joining (NHEJ) over homologous recombination (HR) (Panier and Boulton 2014). We observed a small difference from 59.26% to 77.5% in the percentage of 53BP1-positive interphase nuclei in naïve compared to primed hESCs (Figure 5.2). However, as the levels of DSBs remain unchanged between primed and naïve hESCs based on the  $\gamma$ -H2AX staining, it is possible that the increased 53BP1 is attributable to its role in responding to replication stress, which may occur during reprogramming (Lukas et al. 2011; Harrigan et al. 2011).



**Figure 5.1. Levels of  $\gamma$ -H2AX remain unchanged upon primed to naïve reprogramming.** a) A representative microscope image of hESCs grown on iMEFs stained with an antibody against  $\gamma$ -H2AX (phosphorylated S139; green) and DAPI (blue). Irradiated mouse feeder cells act as an internal positive control for staining as they contain several, large  $\gamma$ -H2AX foci. The white box highlights an example of a hESC nucleus with a  $\gamma$ -H2AX foci. Scale bar represents 10µm. b) Quantification of the percentage of nuclei that have at least one  $\gamma$ -H2AX foci during each stage of reprogramming. The number at the top of each bar represents the total number of nuclei analysed for each time point. Error bars represent the standard deviation of three technical replicates for primed and naïve hESCs. A semi-automated spot count approach was used to score  $\gamma$ -H2AX positive nuclei and to quantify the number of foci/cell. Statistical significance was determined using a one-way ANOVA comparing each time point to each other. N.S. denotes not significant.



**Figure 5.2. Levels of 53BP1 show a slight increase in naïve compared to primed hESCs.** a) A representative microscope image of hESCs grown on iMEFs stained with an antibody against 53BP1 (green) and DAPI (blue). The white box highlights an example of a hESC nucleus with a 53BP1 foci. Scale bar represents 5µm. b) Quantification of the percentage of nuclei that have at least one 53BP1 foci during each stage of reprogramming. The number at the top of each bar represents the total number of nuclei analysed for each time point. A semi-automated spot count approach was used to score  $\gamma$ H2AX positive nuclei and to quantify the number of foci/cell.

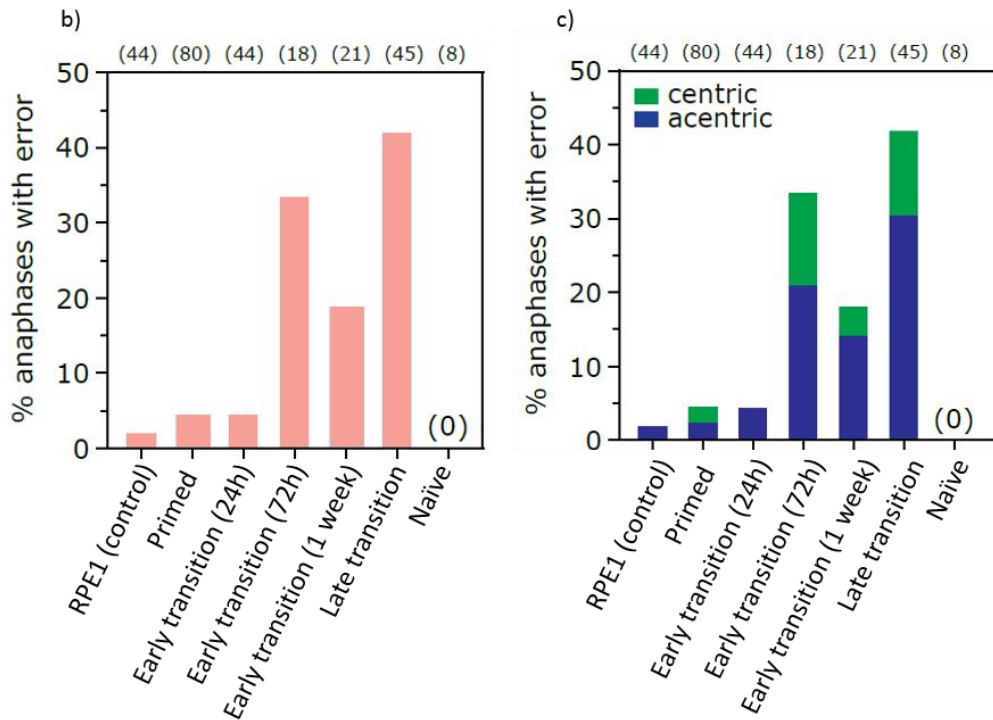
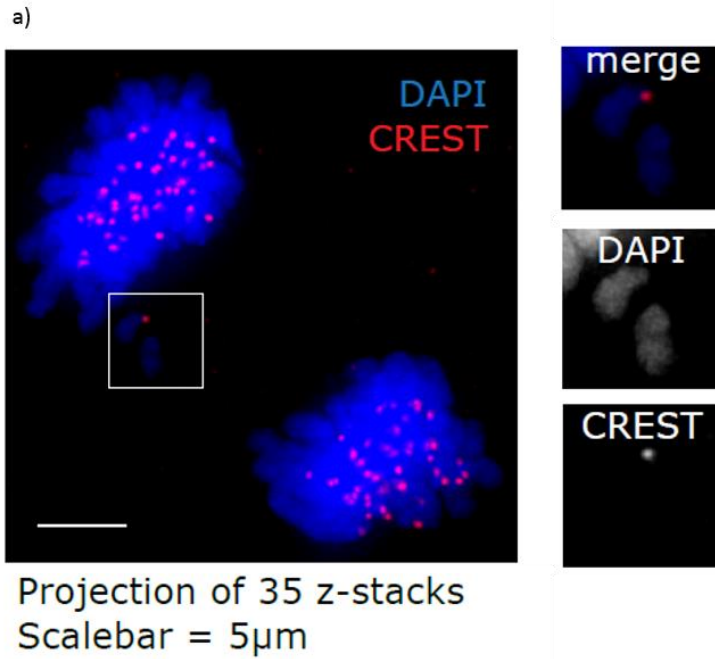
### 5.3 hESC reprogramming is associated with increased chromosome errors during mitotic anaphase and increased ultrafine bridges

In order to maintain genomic stability, cells must ensure that the DNA is correctly replicated and repaired prior to mitosis, and that chromosomes are properly segregated during cell division, the latter being the most critical for maintenance of genomic stability. Cells have mechanisms in place to ensure that these processes occur efficiently, however these mechanisms can be defective for a variety of reasons and can result in genomic instability. Defects during DNA replication as a result of endogenous or exogenous damage as well as inefficient DNA repair often result in the stalling of DNA replication forks (Lecona and Fernandez-Capetillo 2014). Such defects may arise as a result of replication stress, which can be a major contributor of structural abnormalities that arise during mitosis (Burrell et al. 2013). Replication stress can occur for a variety of reasons such as insufficient levels of

deoxynucleotides or DNA replication factors, or due to the misregulation of DNA repair pathways (Bester et al. 2011; Lecona and Fernandez-Capetillo 2014; Ruiz et al. 2015).

Many of the chromosomal abnormalities that arise as a result of DNA replication errors can be detected during anaphase if they bypass the mitotic checkpoint (Kops, Weaver, and Cleveland 2005). Moreover, defective chromosome segregation during mitosis can manifest as a number of structural abnormalities. One such abnormality is the presence of lagging chromosomes during anaphase, which can arise as a result of sister chromatid mis-segregation that can be attributed to pre-mitotic processes or defective chromosome attachments during mitosis (Bizard and Hickson 2018). Lagging chromosomes, or laggards, are fragments of chromosomes that are not physically connected to the rest of the segregating chromosomes and remain near the equatorial plate during anaphase (Bizard and Hickson 2018). Lagging chromosomes can be sub-divided further as either centric or acentric chromatin fragments. Acentric laggards lack a centromere and are therefore unable to interact with the mitotic spindle and segregate correctly during mitosis. Such fragments typically arise as a result of incorrect processing of DNA damage. Centric laggards, on the other hand, are typically full chromosomes containing a centromere, which arise as a result of defective microtubule attachment such as merotelic attachments during mitosis (Bizard and Hickson 2018; Cimini et al. 2001).

In order to determine whether chromosomes are correctly segregated through mitosis during hESC primed to naïve reprogramming, cells at various stages of reprogramming were stained with DAPI and anti-CREST, which stains all centromeres. The hTERT-RPE1 cell line, which is a karyotypically stable diploid human epithelial cell line, was used as a control. Primed hESCs exhibited very small number of lagging chromosomes, however we detected a transient increase in the percentage of anaphase nuclei with lagging chromosomes as cells progressed through the early transition of reprogramming, with 42% of anaphase nuclei containing laggards at the late transition of reprogramming. No lagging chromosomes were detected in the stable naïve cells, however this may be due to the very low numbers of anaphase nuclei present in the sample. Interestingly, the majority of lagging chromosomes detected were acentric, based on the absence of signal from anti-CREST staining. This is indicative of the majority of the instability being attributable to errors during DNA replication in the lead up to mitosis, rather than errors during mitosis.

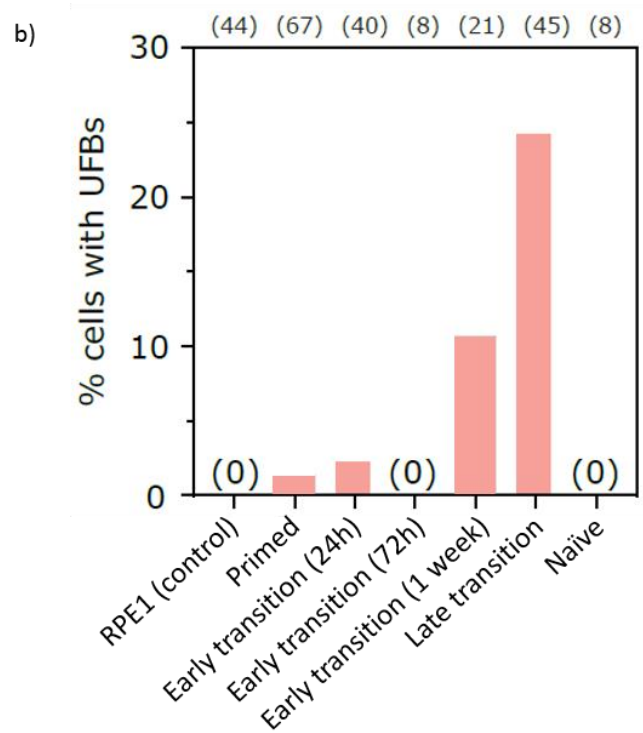
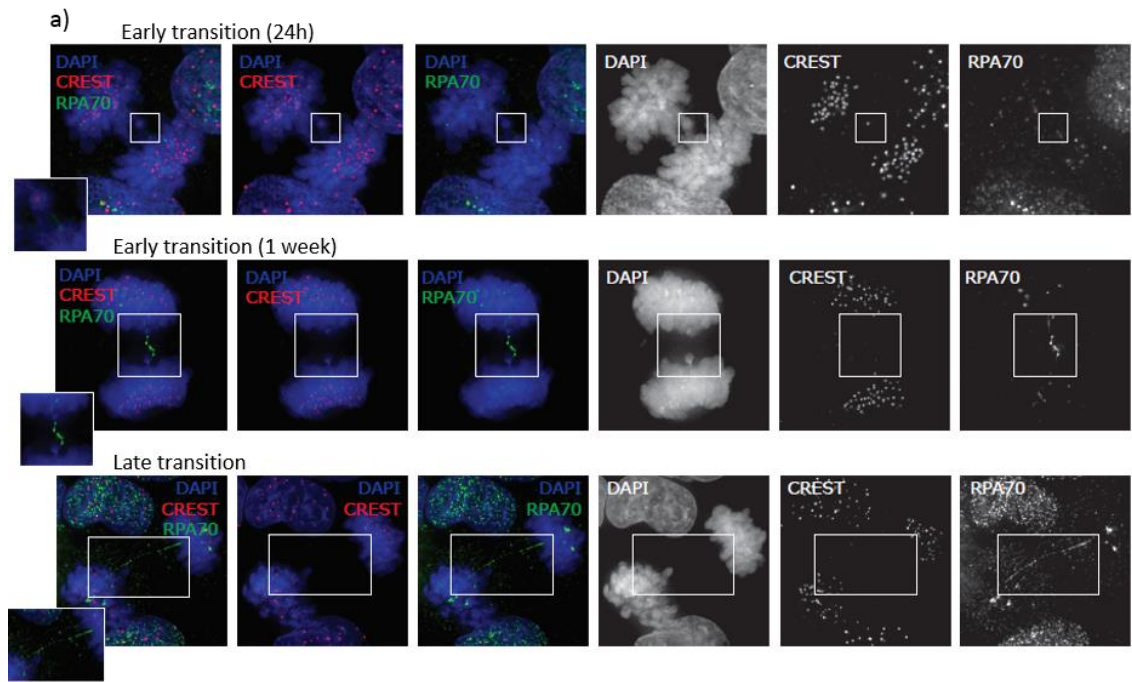


**Figure 5.3. There is a transient increase in the detection of lagging chromosomes during primed to naïve hESC reprogramming.** a) Representative image of hESCs stained with DAPI (blue) and anti-CREST which stains centromeres (red). Image shows an example of a cell in anaphase with lagging chromosomes. Scale bar represent 5 $\mu$ m. b) Quantification of the % of cells in anaphase with errors based on the detection of lagging chromosomes, at each stage of reprogramming. The RPE1 cell line was used as a control. The number at the top of each bar represents the total number of anaphase nuclei analysed for each time point. c) % of cells with lagging chromosomes, split into those with centric laggards (containing a centromere) or acentric laggards (not containing a centromere) based on anti-CREST staining.



Another class of abnormal structures that can arise during anaphase are ultrafine bridges (UFBs), which are distinct from chromatin bridges that also form during mitosis, and can be distinguished by the fact that they cannot be detected using conventional DNA dyes such as DAPI, due to the presence of single-stranded DNA (ssDNA) (Liu et al. 2014). Instead, UFBs consist of de-chromatinised single-stranded DNA filaments coated with UFB-associated proteins (Liu et al. 2014). These bridges are a potential source of genomic instability. UFBs can arise from interlocked DNA at centromeres, from late replication intermediates induced by replication stress, from telomeres, or from intertwined DNA generated by HR (Chan and West 2018). The common feature between the majority of UFBs is that they are typically associated with defects that occur during DNA replication that result in un-replicated DNA, or abnormal DNA structures which are carried over into mitosis. However, some UFBs may also arise from defects during mitosis such as improper sister chromatid cohesion (Liu et al. 2014). Once UFBs have formed, the replication protein A (RPA) is one of a number of proteins that is recruited to these bridges through its association with ssDNA, and plays a role in their resolution (Chan and Hickson 2009).

To detect the presence of ultrafine bridges, we stained hESCs at each stage of reprogramming with DAPI, anti-CREST to detect the centromeres, and an antibody against the replication protein A 70kDa DNA-binding subunit (RPA70). The number of anaphase nuclei with ultrafine-bridges was counted based on RPA70 staining. We did not detect the presence of any UFBs in the hTERT-RPE1 control cell line. Primed hESCs exhibited very small number of UFBs, however we detected a transient increase in the percentage of nuclei with UFBs as cells progressed through the early transition of reprogramming, with a peak of 24.4% of anaphase nuclei containing UFBs at the late transition of reprogramming. Interestingly, no UFBs were detected in the stable naïve cells, however this may be due to the very low numbers of anaphase nuclei present in the sample.



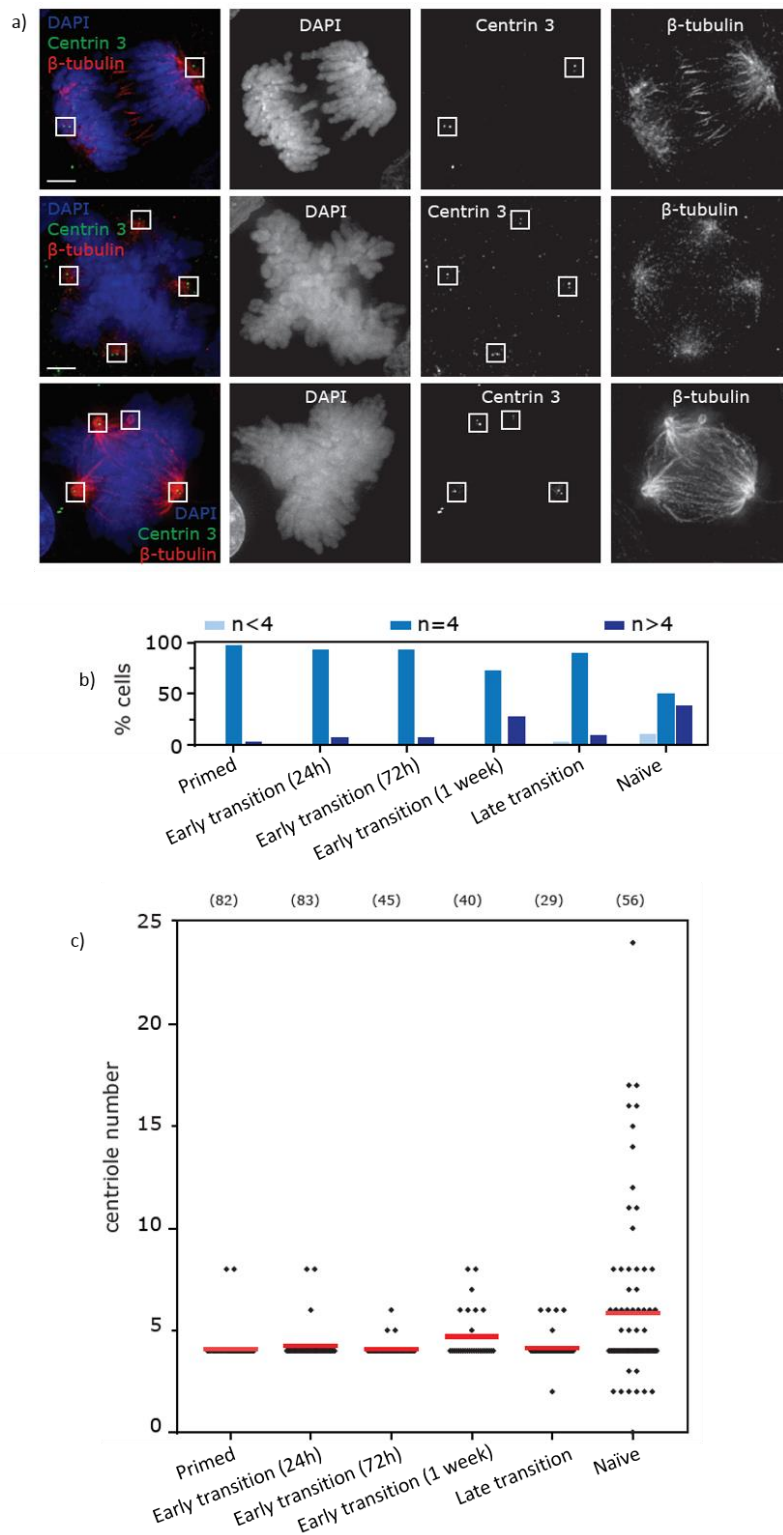
**Figure 5.4. There is a transient increase in the number of ultrafine bridges detected during primed to naïve hESC reprogramming.** a) Images of hESCs at various stages during reprogramming stained with DAPI (blue), anti-CREST (red) and anti-RPA70 (green). Images show representative examples of cells in anaphase with ultrafine bridges (UFBs) which stain negatively for DAPI and positively for RPA70 which associates with single-stranded DNA at UFBs. b) Quantification of the percentage of cells with UFBs at each stage of reprogramming. The RPE1 cell line was used as a control. The number at the top of each bar represents the total number of anaphase nuclei analysed for each time point.

The chromosomal aberrations detected during the transition from primed to naïve pluripotency include primarily acentric lagging chromosomes and ultrafine bridges during anaphase, both of which are typical of pre-mitotic defects (Gisselsson 2008). Only a small percentage of the laggards contained centromeres, indicating that mitotic errors involving incorrect attachment of chromosomes to the mitotic spindle cannot account for the instability observed upon reprogramming (Burrell et al. 2013). Collectively, this points to a putative role of replication stress induced by reprogramming in mediating the genomic instability detected during reprogramming.

#### 5.4 hESC reprogramming is associated with centrosome amplification and increased multipolar spindles

In addition to lagging chromosomes and UFBs, multipolar mitotic spindles are another type of mitotic error that can be detected by staining mitotic cells. Multipolar spindles are strongly associated with supernumerary centrosomes that may arise as a result of centrosome amplification. Multipolar spindles can perturb balanced separation of sister chromatids and can often result in random segregation to daughter cells, hence giving rise to aneuploid cells (Gisselsson 2008).

In order to determine whether the phenomenon of centrosome amplification and multipolar spindles was associated with primed to naïve hESC reprogramming, cells from various time points during reprogramming were stained with DAPI, anti-centrin3 which stains centrosomes, and anti- $\beta$ -tubulin which stains microtubules. Across the period from primed to naïve hESCs, a number of different types of multipolar spindles could be detected (Figure 5.5a). While nearly 100% of primed and early transition hESCs typically displayed 4 centrioles as expected (Figure 5.5b), there were an elevated number of naïve hESCs with more than 4 centrioles (figure 5.5c), corresponding to mitotic cells with multipolar spindles. This suggests that there may be an increased likelihood of an aneuploid population of cells in the naïve state as a result of random chromosome segregation associated with multipolar spindles.

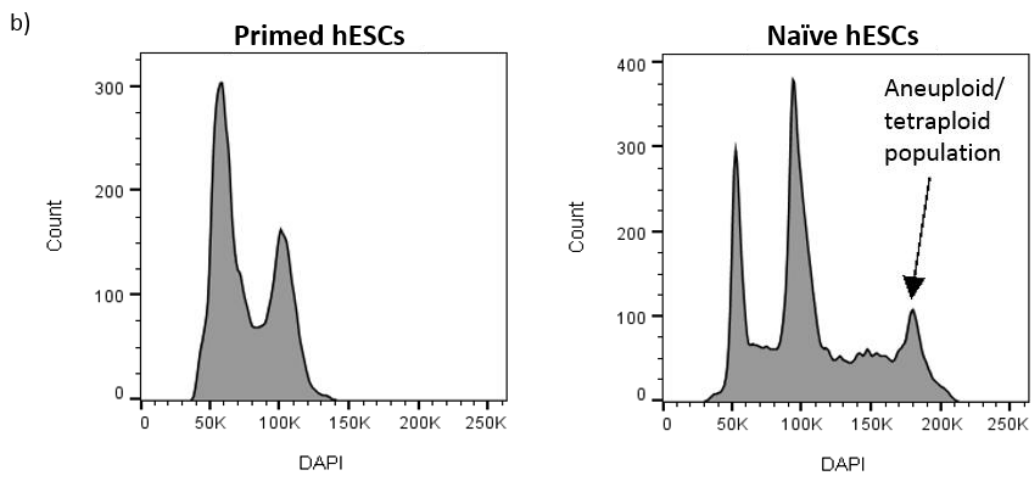
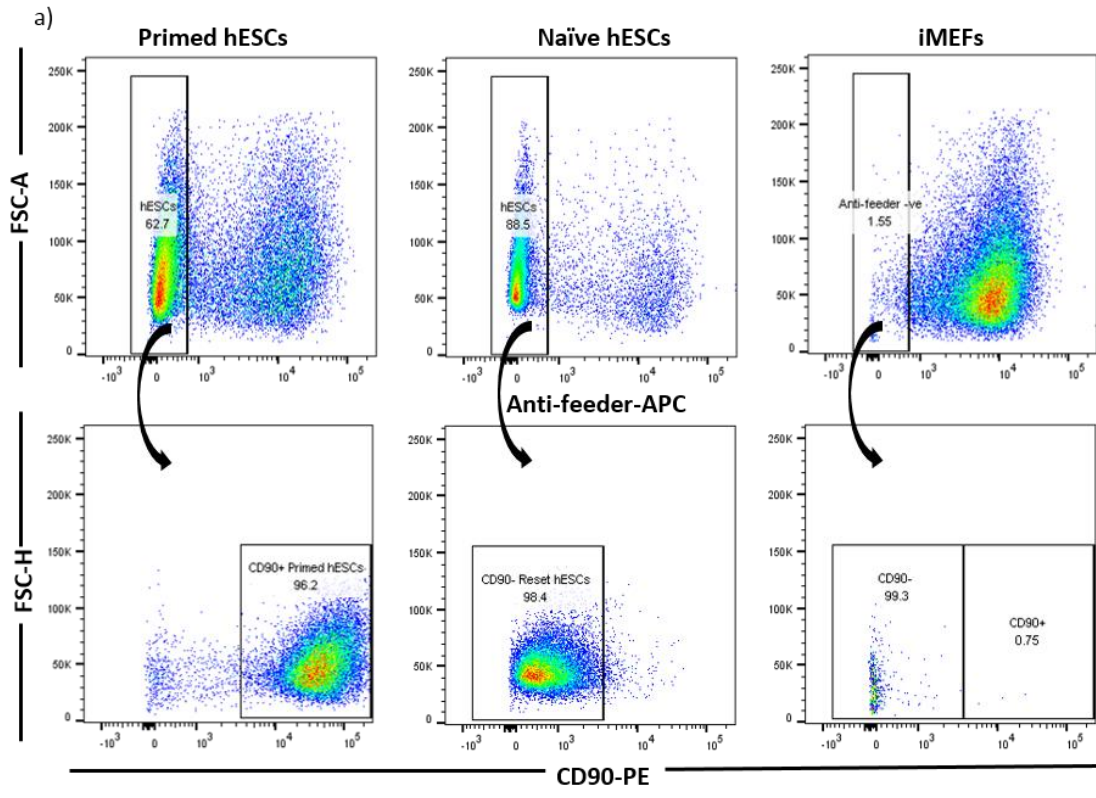


**Figure 5.5. Spindle abnormalities can be detected in naïve hESCs.** a) Images of cells stained with DAPI (blue), anti-centrin 3 (green) which stains centrioles, and  $\beta$ -tubulin (red), which allows the visualisation of spindle fibres. The first row represents an example of a normal spindle, with 4 centrioles and an organised spindle. The second two rows show representative examples of abnormal multi-polar spindles, each with more than 4 centrioles. Scale bars represent 5 $\mu$ m. b) Quantification of the percentage of cells with less than 4 centrioles, 4 centrioles, or more than 4 centrioles at each stage of reprogramming. c) Centriole number per cell (black dots) during reprogramming. Red lines represent the mean. The number at the top of each bar represents the total number of nuclei analysed for each time point.

## 5.5 Long-term naïve hESCs exhibit increased aneuploidy

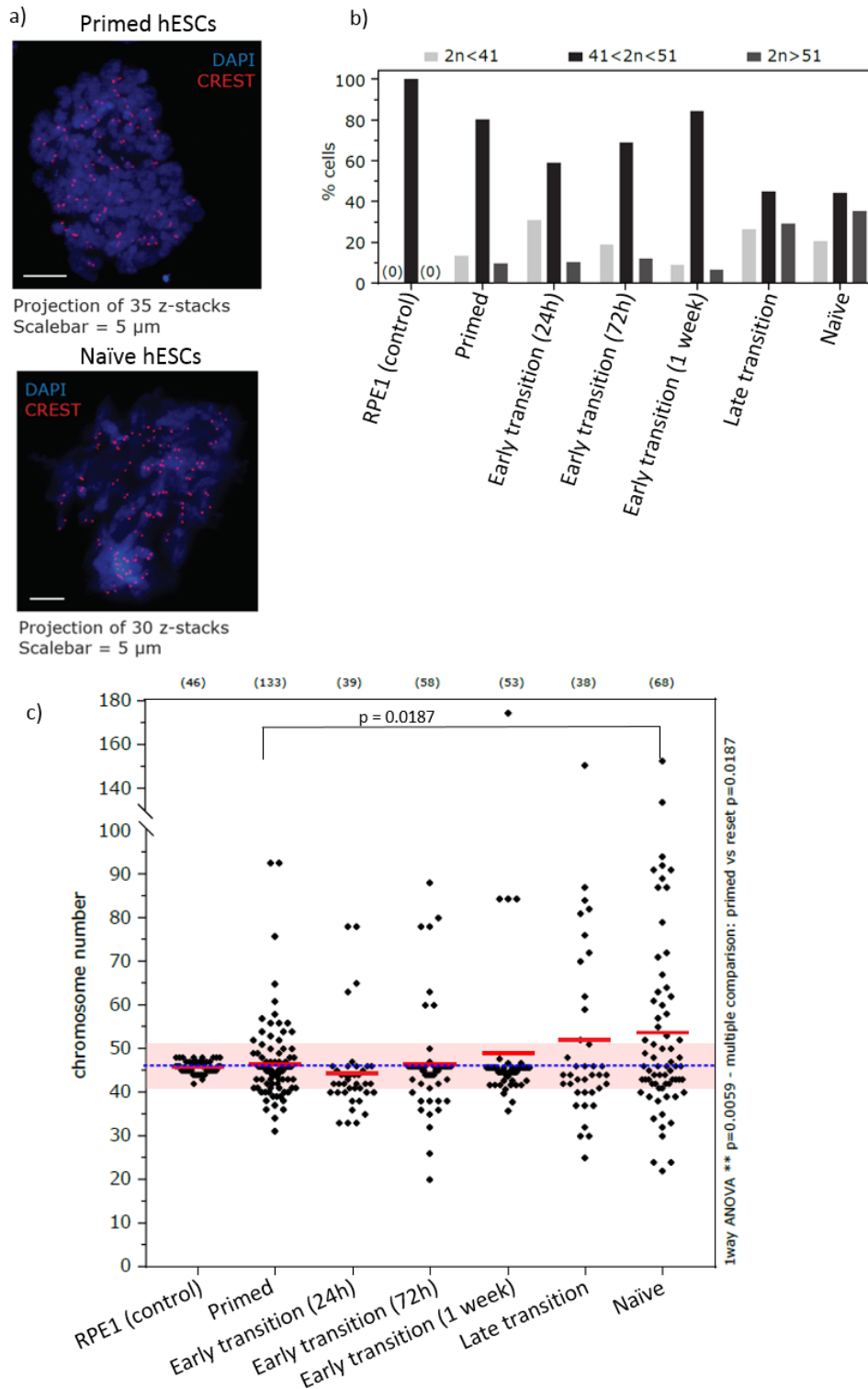
Having observed signs of chromosomal instability, we next investigated the potential outcomes of these aberrations by using various techniques to measure the chromosomal content of individual cells during reprogramming.

We first performed DAPI staining of fixed cells followed by flow cytometry simply to measure the DNA content of cells. As hESCs are grown on a layer of iMEF feeder cells, the flow cytometry staining protocol required a separation strategy to eliminate the feeder cells from the analysis. The most common published method for this involves dissociating the mixture of hESCs and iMEFs, and performing serial plating to allow separation of the two cell types on the basis that iMEFs adhere to plastic surfaces faster than hESC. In our hands, this was only able to eliminate 50% of feeder cells. We therefore used an anti-feeder antibody for positive discrimination of iMEFs and to test whether this could efficiently separate iMEFs from hESCs, we further stained the cells with an antibody for human Thy1/CD90 which is a cell surface marker that is highly expressed in primed hESCs but is absent in naïve hESCs (Chapter 3; Figure 3.1c). We performed flow cytometry to visualise the separation of primed and naïve hESCs from each other and from feeder cells and confirmed that the anti-feeder antibody was successfully able to separate iMEFs from hESCs (Figure 5.6a). We repeated the staining in the absence of the anti-CD90 antibody, instead staining cells with anti-feeder and DAPI and analysed the DNA content of primed and naïve hESCs by comparing levels of DAPI. Primed hESCs have a DAPI profile that is considered normal, whereby there is a peak of cells in the G1 phase of the cell cycle with a single copy of the genome, a smaller number of cells in S phase with intermediate levels of DNA content, and a second peak of cells in the G2/M phase of the cell cycle with duplicated DNA content (Figure 5.6b). The naïve hESCs also retain this profile, but in addition, there is an additional tail and peak on the histogram, depicting cells with an abnormal DNA content greater than that of cells undergoing mitosis. This is indicative of aneuploid and tetraploid cells in the naïve hESC population (Figure 5.6b).



**Figure 5.6. Aneuploidy can be detected in naïve hESCs through analysis of cellular DNA content.** a) Representative flow cytometry plots of anti-feeder/CD90 stained primed and naïve hESCs, with iMEFs as a control. Plots shown have already filtered out debris and doublets. Plots are shown for forward scatter area (FSC-A) against anti-feeder APC and for forward scatter height (FSC-H) against CD90-PE. Mouse feeder cells can be separated from a mixed population to give primed (CD90+) and naïve (CD90-) hESC populations with 96.2% and 98.4% purity respectively. b) DAPI-area histogram for primed and naïve hESCs following removal of debris, doublets and iMEFs from the analysis. The histogram shows a normal cell cycle profile for primed hESCs, with an expected distribution of cells in G1, S, or G2/M phase of the cell cycle. In the naïve hESCs, an abnormal aneuploidy/tetraploid population can be detected in addition to this profile.

We next performed immunofluorescence staining of cells across the transition from primed to naïve hESCs to determine how early on during reprogramming these abnormalities are apparent. Cells were stained with DAPI and the anti-CREST antibody as an indication of chromosome number. As a control, we used the hTERT-RPE1 cell line which is expected to have normal numbers of chromosomes. Figure 5.7a depicts examples of aneuploid cells detected with abnormal numbers of chromosomes. The number of centromeres stained by the anti-CREST antibody was used as a readout of the number of chromosomes present in each cell. We counted the number of chromosomes per cell and used a threshold of 41-51 to define a cell with a normal diploid chromosome number to allow for staining error. Cells with less than 41 or more than 51 chromosomes were considered to be aneuploid (Figure 5.7b). Compared to the hTERT-RPE1 cell line which has a very stable diploid genome with no signs of aneuploidy, the hESCs generally display higher levels of variability in the number of chromosomes per cell. However, the mean chromosome number per cell for primed hESCs and hESCs in the early transition of reprogramming remains within  $\pm 5$  of the expected 46 chromosomes in a diploid genome, and then begins to deviate from the expected mean during the late transition (Figure 5.7c). The stable naïve hESCs which have been passaged 20 times exhibit higher levels of aneuploidy, showing a significant increase in chromosome number per cell compared to primed hESCs.



**Figure 5.7. Abnormal chromosome numbers are more prevalent in naïve compared to primed hESCs.** a) A representative microscope image of a primed and naïve hESC stained with the CREST antibody which stains all centromeres (red) and DAPI (blue), in each case depicting a cell with more than the normal number of chromosomes. Scale bar represents 5 μm. b) Quantification of the percentage of cells with less than 41 chromosomes, between 41 and 51 chromosomes, or more than 51 chromosomes at each stage of reprogramming. The RPE1 cell line was used as a control. c) Chromosome number per cell (black dots) during reprogramming. Red lines represent the mean. The number at the top of each bar represents the total number of nuclei analysed for each time point. Statistical significance between the time points was analysed using a one-way ANOVA with multiple testing. Significant difference between primed and naïve hESCs is indicated with a p-value.

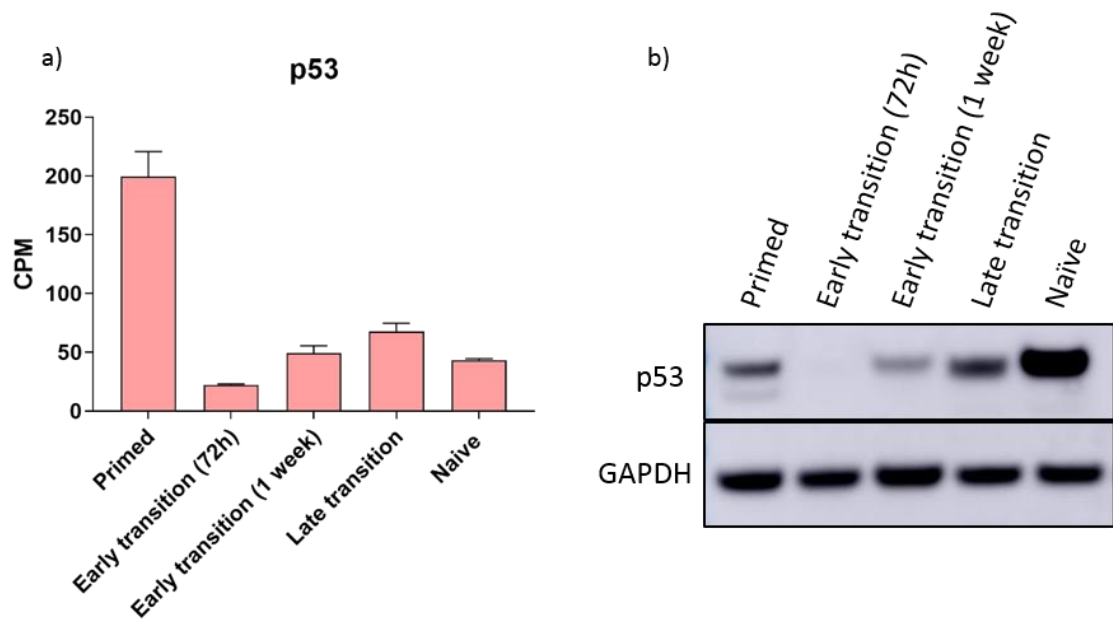


## 5.6 p53 protein expression is stabilised in naïve hESCs

The mitotic checkpoint is a mechanism employed by cells to ensure that chromosome segregation occurs accurately during mitosis, and is one of a number of cell cycle checkpoints that act to regulate cell cycle progression. The mitotic checkpoint should prevent the majority of structural aberrations that have been detected during primed to naïve hESC reprogramming. One of the key proteins associated with each of the cell cycle checkpoints is p53. P53 is implicated during each of the cell cycle checkpoints, and in the response to DNA damage, where it mediates cell cycle arrest or programmed cell death in response to various signals (Giono and Manfredi 2006). In normal cells, p53 is maintained at low levels by proteins that regulate its expression, but it is stabilised and activated in response to various cellular stresses and stimuli (Shieh et al. 1997; Kubbutat, Jones, and Vousden 1997). P53 plays a role in preventing the propagation of aneuploid cells that result from incorrect mitosis, thereby playing a role in maintaining genomic stability (Giono and Manfredi 2006).

P53 is one of the most frequently mutated tumour suppressor genes across multiple cancer types, and mutations in p53 are also positively selected for in some human pluripotent stem cells (Merkle et al. 2017). Moreover, it has recently been suggested that p53 may contribute to DNA methylation homeostasis in ESCs by regulating the expression of DNMT3A and DNMT3B (Tovy et al. 2017). Additionally, there is evidence to suggest that p53 acts as a barrier to efficient cellular reprogramming, and that mutations to p53 enhance cellular reprogramming and increase the malignant potential of the reprogrammed cells (Marion et al. 2009; Sarig et al. 2010).

We measured both gene and protein expression of p53 across the transition from primed to naïve pluripotency in hESCs. At the level of gene expression p53 is highly expressed in primed hESCs and is downregulated upon the early transition of reprogramming and maintained at reduced levels throughout reprogramming to naïve hESCs (Figure 5.8a). At the protein level, however, p53 is initially downregulated as cells go through the early stages of reprogramming, but the protein appears to be later stabilised as cells go through the late transition of reprogramming, and is particularly abundant in stable naïve hESCs (Figure 5.8b). It is worth noting that the antibody used for detection is able to detect both wild-type and mutant forms of the p53 protein.



**Figure 5.8. p53 protein expression is stabilised in naïve hESCs.** a) Counts per million (CPM) from RNA-seq data of p53 across the transition of hESC reprogramming. Bars are representative of the mean of three biological replicates. b) Western blot analysis of p53 across the transition from primed to naïve pluripotency. GAPDH is used as a loading control.

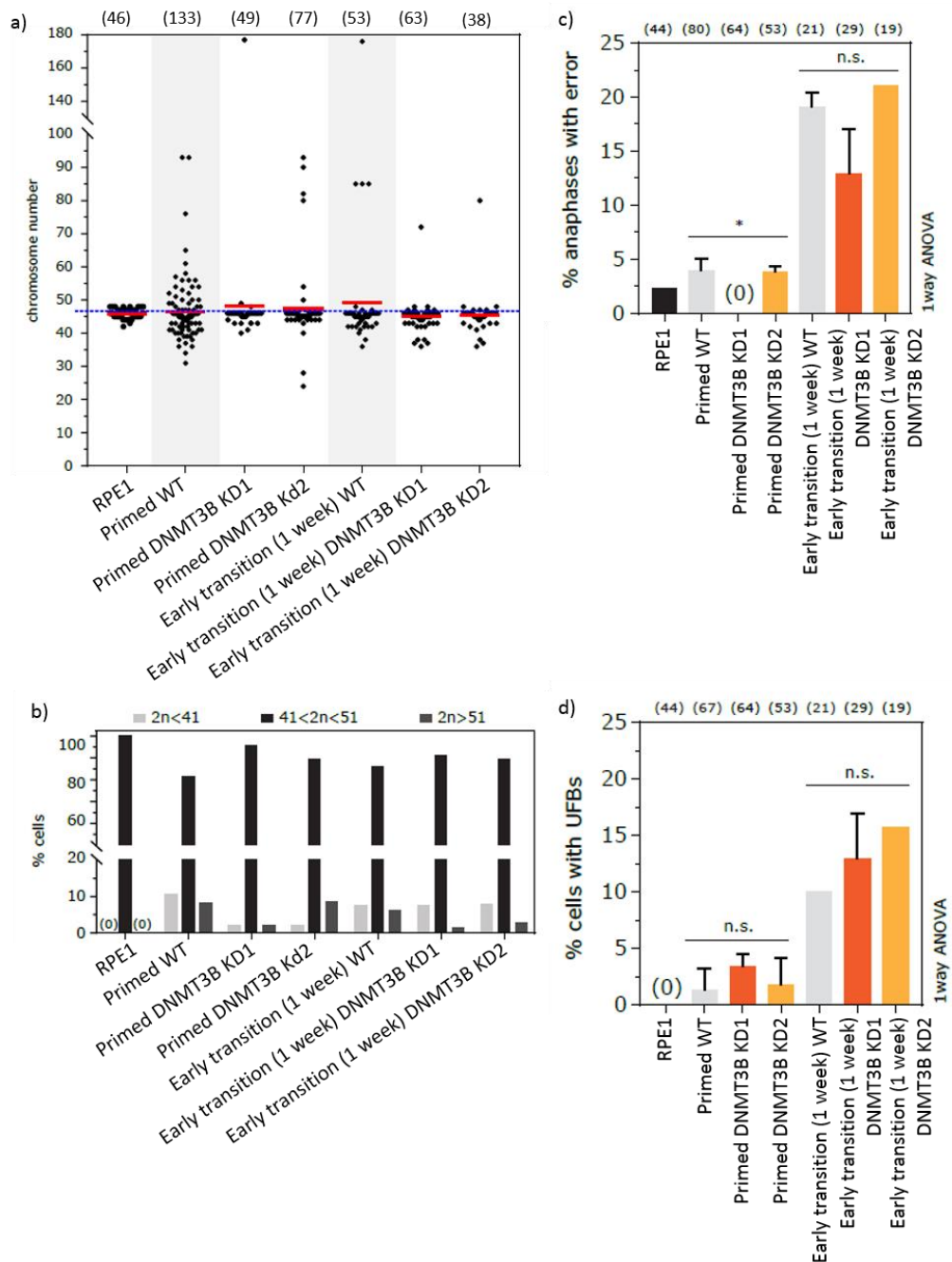
## 5.7 Loss of DNMT3B does not enhance genomic instability upon reprogramming

The centromeres that hold sister chromatids together and ensure proper chromosome segregation during mitosis are composed primarily of repetitive DNA satellites. The centromeric proteins B and C (CENP-B and CENP-C) localise to alpha satellite or minor satellite DNA at the centromeres (Guenatri et al. 2004; Muro et al. 1992; Politi et al. 2002). Centromeric regions also exhibit distinct epigenetic modifications, including dense DNA hypermethylation and histone lysine di- and trimethylation, particularly at H3K9 (Lehnertz et al. 2003).

A major link between DNA methylation and genomic stability comes from the study of patients with ICF syndrome which is caused by mutations in DNMT3B and is characterised by mitotic defects (Ehrlich 2003). One of the key defining features of the disease is the DNA demethylation of centromeric and pericentromeric repeats and loss of chromosome condensation during mitosis (Ehrlich 2003). It has been demonstrated that DNMT3B interacts with CENP-C and that CENP-C recruits DNMT3B and DNA methylation to centromeric and pericentromeric satellite repeats, which further regulates the characteristic histone code at these regions (Gopalakrishnan et al. 2009). In this context, loss of either CENP-C or DNMT3B

through experimental manipulation led to increased chromosomal segregation defects and genomic instability (Gopalakrishnan et al. 2009).

Using the DNMT3B knock down hESC lines generated in chapter 3, we sought to investigate a potential role of diminished DNMT3B expression on the chromosomal and genomic instability detected early during reprogramming. Our hypothesis was that in the absence of DNMT3B, we may see an increase of cells with abnormal chromosomal structures as a result of loss of DNMT3B-mediated epigenetic regulation of centromeres. Using cells stained with DAPI and anti-CREST, we once again used the number of CREST positive foci as a readout of chromosome number in primed and early transition wild-type (WT) and DNMT3B knock down cells (Figure 5.9a). We also counted the number of chromosomes per cell and used a threshold of 41-51 to define a cell with a normal diploid chromosome number to allow for staining error. Cells with less than 41 or more than 51 chromosomes were considered to be aneuploid (Figure 5.9b). The mean chromosome number per cell for WT and DNMT3B knock down primed and early transition hESCs remains within  $\pm 5$  of the expected 46 chromosomes in a diploid genome, indicating that reduced expression of DNMT3B does not influence chromosomal stability in the early stages of hESC reprogramming. Additionally, we quantified the percentage of anaphase cells with lagging chromosomes and UFBs in WT and DNMT3B knock down hESCs (Figure 5.9c and 5.9d). Once again, there was no significant difference between the percentages of mitotic cells with laggards or UFBs in the DNMT3B knock down cells compared to the WT in either primed or early transition hESCs. Collectively, these results indicate that a reduction in the expression of DNMT3B does not affect chromosomal instability in the early stages of primed to naïve hESC reprogramming.

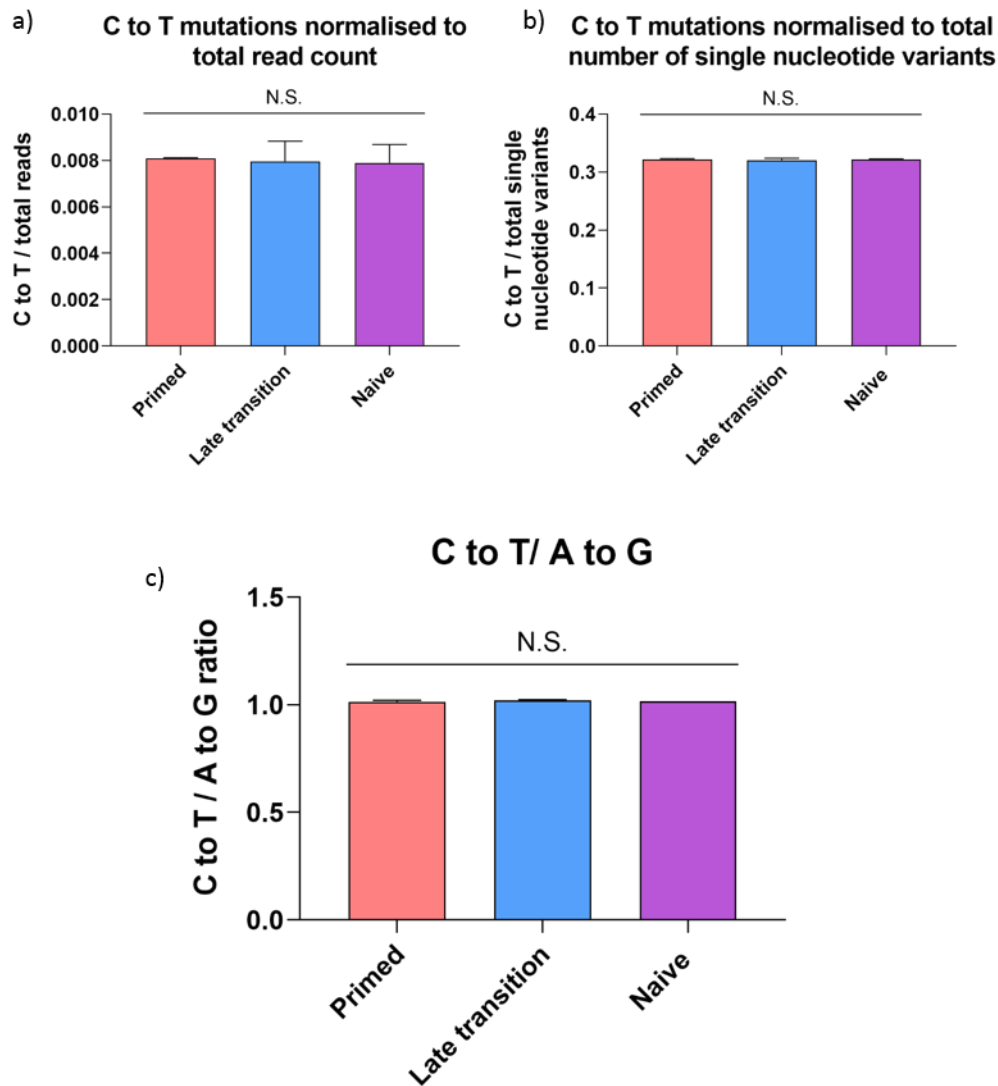


**Figure 5.9. Downregulation of DNMT3B does not enhance genomic instability during early reprogramming.** a) Chromosome number per cell (black dots) during early reprogramming in wild-type (WT) and DNMT3B knock down (KD) hESCs. Red lines represent the mean. The number at the top of each bar represents the total number of nuclei analysed for each time point. b) Quantification of the percentage of cells with less than 41 chromosomes, between 41 and 51 chromosomes, or more than 51 chromosomes at each stage of reprogramming. The RPE1 cell line was used as a control. c) Quantification of the % of cells in anaphase with errors based on the detection of lagging chromosomes during early reprogramming in WT and DNMT3B KD hESCs. The RPE1 cell line was used as a control. The number at the top of each bar represents the total number of anaphase nuclei analysed for each time point. Bars represent the mean of 2 biological replicates. Error bars represent SEM. Statistical significance was calculated using a one-way ANOVA with a post-hoc test comparing each KD sample to the WT for each time point. N.S. denotes not significant. d) Quantification of the percentage of cells with ultrafine bridges (UFBs) during early reprogramming in WT and DNMT3B KD hESCs. The RPE1 cell line was used as a control. The number at the top of each bar represents the total number of anaphase nuclei analysed for each time point. Bars represent the mean of 2 biological replicates. Error bars represent SEM. Statistical significance was calculated using a one-way ANOVA with a post-hoc test comparing each KD sample to the WT for each time point. N.S. denotes not significant.

## 5.8 The genomic mutation frequency is unaffected by global demethylation upon reprogramming

The assays detailed above indicated that reprogramming is associated with genomic instability. We were next interested in determining whether the demethylation that occurs during reprogramming affects the global mutation frequency in hESCs, as has been reported in other contexts (Chen et al. 1998). We used a whole-genome sequencing approach to measure the mutation frequency during hESC reprogramming. We opted to measure hESC in the late transition of reprogramming and in naïve hESCs grown for 20 passages following induction of the stable naïve state, with primed hESCs used as a basal measure of the ongoing mutation rate. The cells in the late transition of reprogramming were used as an intermediate time point to detect any potential increase in mutation frequency directly influenced by the ongoing genomic demethylation, while the naïve hESCs were used to try to detect any change in the mutation frequency that accrued over time in the population as a result of the cells having a global demethylated genome. For both time points, we were interested in detecting all mutations with a low variant allele frequency (VAF), irrespective of the genomic location of the mutation, as hESCs in culture are unlikely to face any specific selective pressures that would select for certain mutations.

We performed WGS for primed, late transition and naïve hESCs, and obtained whole genome sequencing reads at a coverage of 5x per sample. We counted the total number of C to T variants per sample and normalised this value to either the total read count per sample (Figure 5.10a) or the total number of single nucleotide variants detected (Figure 5.10b), and additionally calculated the ratio between C to T and A to G variants (Figure 5.10c). We observed no significant difference in the frequency of C to T mutations between primed, late transition and naïve hESCs, indicating that global demethylation upon reprogramming does not affect the mutation frequency in the cells. However, due to the low coverage of the data, the low level of variability between the different time points is comparable to the low variability between replicates. This suggests that the variability is attributable to the background error rate that arises as a result of PCR and sequencing errors. It is therefore difficult to distinguish true variation between samples from background noise, suggesting that a considerably higher coverage of sequencing reads would be required to detect true variation in the frequency of mutations.

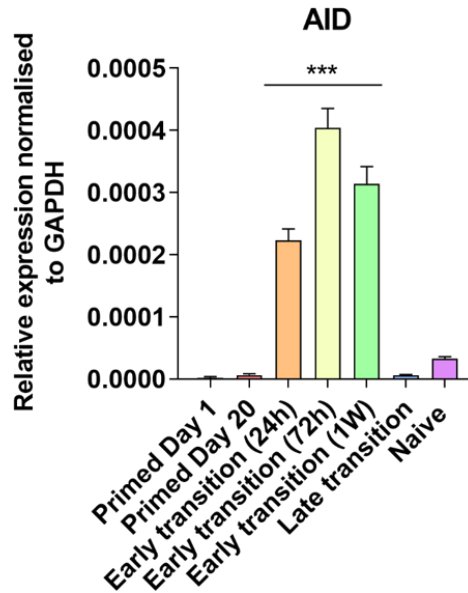


**Figure 5.10. Global demethylation does not affect the C to T mutation frequency.** a) The total number of C to T mutations per sample, normalised to the total read count per sample. Bars are representative of two biological replicates, and error bars represent SEM. Statistical significance between samples was calculated using a one-way ANOVA. N.S. denotes not significant. b) The total number of C to T mutations per sample, normalised to the total number of single nucleotide variants per sample. Bars are representative of two biological replicates, and error bars represent SEM. Statistical significance between samples was calculated using a one-way ANOVA. N.S. denotes not significant. c) The total number of C to T mutations per sample, divided by the total number of A to G variants, displayed as a C to T / A to G ratio. Bars are representative of two biological replicates, and error bars represent SEM. Statistical significance between samples was calculated using a one-way ANOVA. N.S. denotes not significant.

## 5.9 Hypermethylation of CGIs during reprogramming is not associated with an increased mutation frequency

Methylated CpGs are considered to be more prone to spontaneous deamination than unmethylated CpGs, and T:G mismatches generated through both spontaneous deamination or AID-induced deamination undergo less efficient repair (Shen, Rideout, and Jones 1994; Schmutte et al. 1995). We used rt-qPCR to measure the expression of AID/AICDA across the

time course of reprogramming and detected a transient upregulation of its expression during the early stages of reprogramming (Figure 5.11).

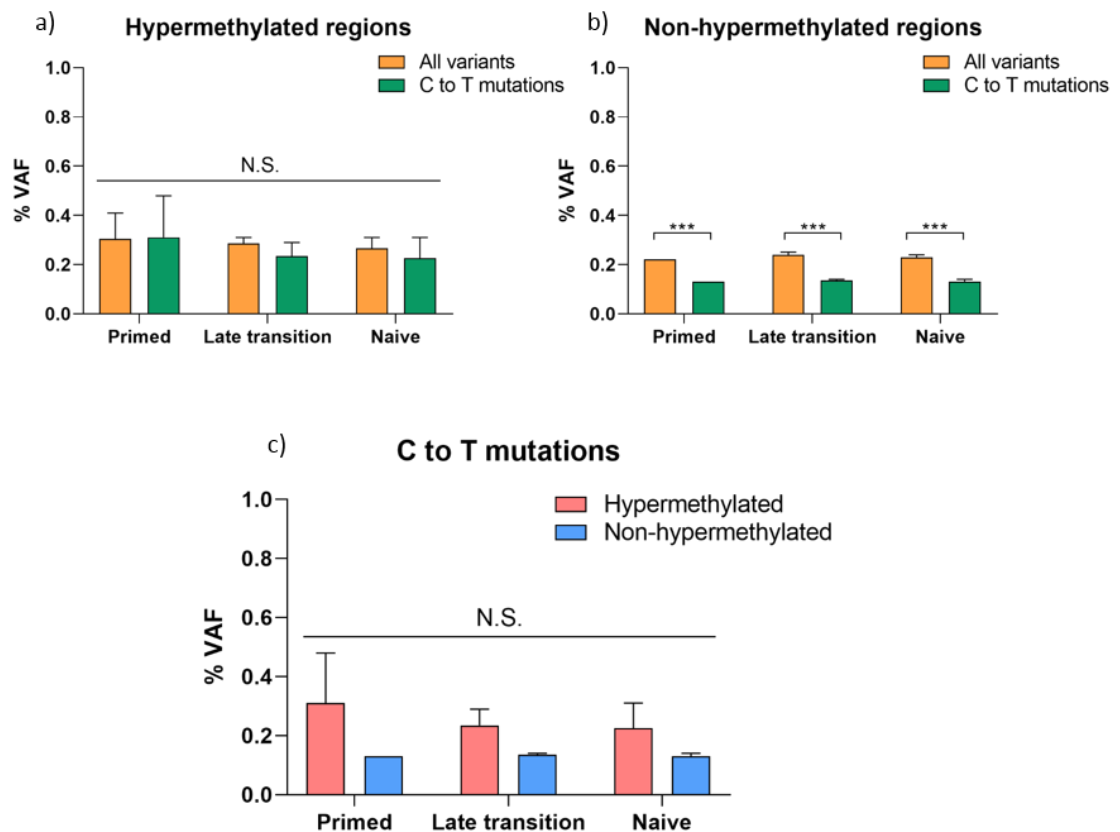


**Figure 5.11. AID is transiently upregulated during the early transition of reprogramming.** qRT-PCR for AID across the transition of hESC reprogramming. Bars are representative of the mean of three technical replicates and three independent biological replicates. Error bars represent the standard error of the mean. Statistical difference between samples was analysed by a two-tailed Student's t-test, with a Bonferroni post-hoc test comparing each time point to primed hESCs. \* $P < 0.05$ , \*\* $P < 0.01$  and \*\*\* $P < 0.001$ . Human GAPDH was used to normalise expression.

We subsequently set out to determine whether the CGI hypermethylation that accumulates during primed to naïve reprogramming increases the mutability of CpG residues at CGIs. To allow the cells enough time to undergo multiple divisions following *de novo* methylation, we opted to study the effect of CGI hypermethylation on DNA mutation in cells in the late transition of reprogramming and in naïve hESCs grown for 20 passages following induction of the stable naïve state. The cells in the late transition of reprogramming were used as an intermediate time point to detect any potential early mutations at CGIs which are not subsequently maintained in the population following multiple rounds of cell division, while the naïve hESCs would allow us to detect mutations that accrued over time in the population. Both time points were compared to primed hESCs, which contain primarily unmethylated CGIs.

We used our WGBS analysis (detailed in chapter 3) to identify two control regions within CGIs that are unmethylated in primed hESCs and remain unmethylated throughout the process of

reprogramming to the naïve state. We also selected three test CGI regions that are unmethylated in primed hESCs that become hypermethylated during reprogramming and remain hypermethylated in naïve hESCs. Each of regions selected contained at least 6 CpG sites. We amplified DNA from the five regions at each of the three time points using and then used a targeted sequencing approach to obtain high-depth sequencing of each of the regions for our analysis.



**Figure 5.12. DNA hypermethylation at CGIs does not impact the frequency of C to T mutations.** a) The percentage variant allele frequency (% VAF) for C to T mutations compared to all other single nucleotide variants in primed, naïve, and late transition hESCs, at regions that become hypermethylated during reprogramming and b) regions that remain unmethylated during reprogramming. Bars are representative of the mean of two independent biological replicates. Error bars represent the standard error of the mean. Statistical significance between samples was calculated using a two-way ANOVA, comparing the time points of reprogramming and the type of variant, with a post-hoc test comparing C to T mutations and all variants for each time point. N.S. denotes not significant, \*\*\* denotes  $p < 0.001$ . c) The %VAF of C to T mutations for primed, naïve and late transition hESCs for hypermethylated and non-hypermethylated regions. Bars are representative of the mean of two independent biological replicates. Error bars represent the standard error of the mean. Statistical significance between samples was calculated using a two-way ANOVA, comparing the time points of reprogramming and the methylation status of the region, with a post-hoc test comparing C to T mutations at hypermethylated and non-hypermethylated regions for each time point. N.S. denotes not significant.



Using an analysis pipeline that enabled the calculation of VAFs of all single nucleotide variants, we compared the average VAFs of C to T mutations compared to all variants at each time point. We used a minimum threshold of 0.001 (0.1%) to identify variants. This threshold is 10-fold lower than the 1% cut-off typically used to variant calling, however as we were interested in comparing rare variants between samples with two replicates per time point, and because the non-hypermethylated regions and the primed hESCs both act as controls, we reasoned that this cut-off was sufficient to test our hypothesis. Moreover, as with the WGS analysis, hESCs in culture are unlikely to face any particular selective pressures that would select for specific mutations and increase their VAFs. We observed no significant difference between the VAFs of C to T mutations compared to all variants at the regions that become hypermethylated upon reprogramming (Figure 5.12a). For the control regions that remain unmethylated during reprogramming, there was a statistically significant difference between the VAFs of C to T mutations compared to all variants, however this was observed at all time points (Figure 5.12b). Moreover, when we looked specifically at C to T mutations and compared the VAFs across the time points at both hypermethylated and non-hypermethylated regions, we found no significant differences between any of the average VAFs, indicating that there was no difference between regions that become hypermethylated compared to those that remain unmethylated throughout the reprogramming process (Figure 5.12c). Altogether, we observed that C to T mutations were not more frequent than other substitutions at any regions or at any time point, and that hypermethylation did not increase the VAFs of C to T mutations in a site-specific manner. This suggests that the CGI hypermethylation associated with reprogramming does not influence the intrinsic mutability of CpGs and that the upregulation of AID detected early during the reprogramming process does not have an impact of the rate of C to T mutations, or that the AID protein is not actually expressed despite the mRNA being present, though this remains to be tested.

## 5.10 Discussion

The data presented in this chapter demonstrate preliminary evidence of some genomic instability in cells as they undergo reprogramming from the primed to the naïve state of human pluripotency. The genomic instability observed appears to be related to the reprogramming process, but largely independent of the global DNA demethylation and regional CGI hypermethylation that occurs upon reprogramming, though this remains to be explored further due to their temporal overlap.

Our data showed that upon primed to naïve hESC reprogramming, the levels of DNA DSBs based on  $\gamma$ -H2AX positive foci remain constant, which may either be due to steady levels of DNA damage, or the lack of activation of the DDR. Interestingly, while we did not observe evidence of an increased frequency of DNA DSBs, we observed evidence of a number of different manifestations of chromosomal segregation errors, such as lagging chromosomes, UFBs in anaphase nuclei, and multipolar mitotic spindles with extra centrosomes. The correlation between the frequencies of these mitotic segregation errors is expected, as a mitotically unstable population will typically display different types of errors, likely in a similar fraction of cells (Gisselsson 2008), though this is difficult to infer from staining of fixed cells. Moreover, we observed a significant increase in the number of aneuploid or tetraploid cells in the naïve hESC population, which is likely the outcome of the chromosomal errors that bypass the various check points in the cell cycle. This increased karyotypic instability is in line with previous studies that have reported abnormal chromosomal content in certain naïve hESC lines generated either from primed hESCs or through iPS reprogramming of somatic cells (Pastor et al. 2016; Kilens et al. 2018).

The specific types of chromosomal aberrations observed such as UFBs and primarily acentric lagging chromosomes are indicative of pre-mitotic defects and suggest that replication stress may play a role in generating instability. Replication stress has previously been associated with similar chromosomal aberrations, and has been implicated as a potential cause of genomic instability in other types of cellular reprogramming (Burrell et al. 2013; Ruiz et al. 2015). To ascertain whether replication stress is indeed a contributing factor to chromosomal instability during reprogramming, one possible experimental approach would be to induce reprogramming of primed hESCs in the presence of additional exogenous nucleotide supplements in case exhaustion of the nucleotide pool results in DNA replication defects and the associated chromosomal instability (Bester et al. 2011).

Genome alterations such as those detected upon primed to naïve hESC reprogramming only become a threat to the cell once they pass anaphase. It has been shown in other contexts that if such chromosome structures are left unresolved, they can give rise to chromosomal aberrations such as those typically found in cancer. In line with this, the stabilisation of p53 protein in naïve hESCs is an interesting observation as it could reflect stabilisation of the wild-type p53 protein, indicating that damaged naïve hESCs may be undergoing apoptosis. The apoptosis of damaged naïve hESCs could be a possible explanation for the lower percentage of cells with lagging chromosomes and ultrafine bridges in the stable naïve hESC population compared to the cells in the early and late transition of reprogramming. Alternatively, as the

antibody used recognises both wild-type and mutant forms of the protein, accumulation of p53 in naïve hESCs could reflect stabilisation of a mutant form of p53, as has previously been observed in hESC lines and is a frequent observation in cancer where mutant p53 favours cancer cell survival (Merkle et al. 2017; Mantovani, Collavin, and Del Sal 2019).

Our data indicated that a reduction in the expression of DNMT3B did not affect the levels of abnormal mitotic structures observed early during hESC reprogramming. As we did not explicitly assess the DNA methylation status of centromeric DNA in DNMT3B knock down hESCs however, we cannot conclude whether the expected demethylation of centromeric and pericentromeric DNA repeats occurred upon knock down of DNMT3B. Moreover, we did not assess chromosomal instability in the DNMT3B knock down cells at later stages of reprogramming or in naïve hESCs lacking DNMT3B, which correspond to the time points where the most instability was evident in wild type hESCs. However, from the data we have, we can conclude that diminished expression of the DNMT3B protein does not acutely influence chromosomal instability during the early stages of primed to naïve hESC reprogramming.

We tested a putative relationship between DNA demethylation and the frequency of ongoing mutations in hESCs during the late transition of reprogramming and in naïve hESCs. Our data showed no detectable change in the mutation frequency between primed, late transition and naïve hESCs. The experimental approach used enabled us to sequence the genome at 5x coverage for each sample, with duplicate samples of primed, naïve and late transition hESCs, each of which represent a heterogeneous population of cells. The average mutation rate in normal human somatic cells is estimated to range between  $3 \times 10^{-7}$  and  $3 \times 10^{-8}$  mutations per nucleotide per generation of cells, equating to roughly 10 to 100 mutations per genome per generation for the human genome that is made up of  $3 \times 10^9$  base pairs (Xue et al. 2009; Genomes Project et al. 2010; Milholland et al. 2017) . With the 5x coverage we obtained, it is difficult to differentiate between true mutations and false positives, as the VAFs for each of the variants identified were below the estimated 1% error rate that is a limitation of the next-generation sequencing process. This 1% error rate constitutes polymerase errors, bias during PCR amplification as well as errors during cluster amplification and sequencing (Fox et al. 2014). It is therefore difficult to draw strong conclusions from this experiment. An improved approach for future studies would be to employ sequencing techniques that are designed to minimise the background error rate, such as duplex sequencing, which is reported to have an error rate of less than 1 artefactual mutation per billion nucleotides by operating on the premise that true mutations must be found on both strands of the DNA (Schmitt et al. 2012). Additionally, as hESCs in culture may not undergo any selective pressure that would facilitate

the accumulation of mutations, it may also be beneficial to grow up single cell clones of hESCs and perform sequencing on the resulting cell populations in order to be able to detect variants at a sufficient depth.

We also tested the differential susceptibility of C to T mutations arising from methylated and unmethylated cytosine nucleotides, through either spontaneous deamination or deamination by AID. We saw no significant difference in the VAFs of C to T mutations across the time course of reprogramming as CGIs become hypermethylated, or between regions that are hypermethylated compared to those that remain unmethylated. Despite using a targeted sequencing approach that enabled a greater depth of sequencing for each region analysed, the VAFs of the mutations identified were all once again below 1%, thereby making it difficult to differentiate them from background noise and to draw strong conclusions with regards to the relationship between DNA methylation and mutation susceptibility. Moreover, the primed to naïve hESC *in vitro* reprogramming system is unlikely to be subjected to specific selective pressures that would enable the clonal selection and outgrowth of cells with specific mutations as is the case for clonal mutations in cancer. However, given the high depth of sequencing using a targeted approach (a minimum of 10,000 reads per region), it is likely that the low VAFs of below 1% indicate that there is no increase in the frequency of C to T mutations as a result of hypermethylation during reprogramming. Once again, employing a sequencing technique such as duplex sequencing may be a better approach to ascertain any relationship between DNA methylation and mutation.

The preliminary characterisation of genomic instability and altered mutation frequency upon primed to naïve hESC reprogramming detailed in this chapter indicates that some degree of genomic instability is associated with reprogramming to the naïve state of pluripotency. The instability appears to be associated with replication stress and manifests as structural chromosomal aberrations that are detectable in mitotic cells during anaphase. On the contrary, the mutation rate of primed, naïve and late transition hESCs appears to be comparable, indicating that reprogramming and the associated DNA methylation changes do not affect the cellular mutation rate. Due to limitations of the experimental and sequencing approaches used, however, further experiments using more sensitive technologies are necessary to enable firm conclusions to be drawn with regards to reprogramming, DNA methylation and mutation frequency.

## Chapter 6. Conclusions and Future Directions

### 6.1 Discussion & Future Directions

#### 6.1.1 Dynamics and functions of DNA methylation in hESCs and cancer

In this thesis, we demonstrate that upon reprogramming of primed hESCs to the naïve state of pluripotency, cells acquire DNA methylation at a subset of CGI promoters associated with developmental genes, while globally, the genome undergoes demethylation. By performing a temporal analysis of DNA methylation levels during the transition between the two pluripotent states, we show that the acquisition of DNA hypermethylation is gradual, and that global erasure of DNA methylation also occurs at a late stage of reprogramming, despite several changes in gene and protein expression occurring rapidly upon induction of reprogramming. This is consistent with the profile of iPS reprogramming of somatic cells, where global demethylation occurs at a late stage, after several waves of transcriptional reprogramming and global chromatin remodelling (Polo et al. 2012; Mikkelsen et al. 2008). The level of hypermethylation accrued at each CGI varies, with some CGIs gaining 10% methylation, while others gain up to 80-90% methylation. This indicates that not all the cells in the population of reprogramming hESCs are gaining DNA methylation at each region, pointing to population heterogeneity in the epigenetic landscape of cells upon reprogramming. Moreover, only a subset of the hypermethylated sites retain the hypermethylation in stable naïve hESCs, suggesting that the reprogramming process may be selecting for cells with a particular DNA methylation profile. To verify whether primed to naïve hESC reprogramming involves active reprogramming of cells or the selection of a subset of cells, hESCs could be labelled with bromodeoxyuridine (BrdU) and the rate of loss of BrdU monitored upon induction of reprogramming to determine the rate of actively proliferating cells. Additionally, it would also be of interest to determine whether cells that become hypermethylated during the transition phase are those that successfully transition to the naïve state, by sorting cells during the early transition of reprogramming based on positive staining for the cell surface marker Sushi domain containing 2 (SUSD2), which has recently been reported to be a reliable marker for the purification of naïve hESCs (Bredenkamp et al. 2019).

The DNA methylation landscape that results upon primed to naïve hESC reprogramming mirrors the human cancer DNA methylome, where aberrant CGI hypermethylation is frequently observed amidst a globally hypomethylated genome. Such parallels to the cancer DNA methylome have been drawn previously in other mammalian species and developmental

contexts, such as in mouse trophoblast cells (Smith et al. 2017). However, the data we present here demonstrates a hypermethylation phenomenon conserved across *in vitro* and *in vivo* human pluripotency (Guo et al. 2014), strengthened by its reproducibility across multiple *in vitro* reprogramming methods (Theunissen et al. 2014; Guo et al. 2017). Moreover, it is notable that we do not observe comparable hypermethylation in the mouse ICM or in *in vitro* mouse ESCs cultured in the presence of 2i inhibitors. This observation may reflect the fact that hESCs and mESCs *in vitro* represent different pluripotent states, which is further highlighted by the differences in their gene expression (Davidson, Mason, and Pera 2015). Of particular relevance for this study, it is notable that naïve human and mouse ESCs exhibit varied expression of DNMTs, in particular DNMT3L which is highly upregulated in naïve hESCs as demonstrated in our data, but is downregulated in mESCs cultured in 2i (Ficz et al. 2013; von Meyenn et al. 2016). In recent years, the emergence of novel sequencing technologies that allow for more detailed characterisation of the molecular processes underlying mouse and human early development have exposed a number of differences between the two, likely with functional consequences (Hanna, Demond, and Kelsey 2018; Boroviak et al. 2018). Such differences have potential implications for making inferences with regards to epigenetic processes between species, both in development and in the study of cancer, as has been noted previously (Diede et al. 2013).

The presence of methylated CGIs both in naïve hESCs *in vitro* and in human ICM cells *in vivo* suggests that the hypermethylation may play a functional role in the maintenance of naïve pluripotency. This is supported by the finding that similar developmental pathways undergo hypermethylation and gene repression, perhaps further attenuating the expression of lowly expressed genes and maintaining cells in a naïve state. We observed that naïve hESCs generated from either DNMT3A or DNMT3B knockdown primed hESCs exhibited reduced stability in the naïve pluripotent state. It is currently difficult to distinguish whether this is reflective of an inability of the primed hESCs to successfully transition to the naïve state in the absence of the DNMT3s, or whether the lack of *de novo* DNA methylation directly impacts the stability of the naïve pluripotent state. Additionally, it is unclear whether the impact on naïve hESCs is dependent on the catalytic *de novo* methylation activity of the two enzymes, or is a result of a non-catalytic role of DNMT3A and DNMT3B. This could be further tested by using an independent method to erase DNA hypermethylation at bivalent CGIs, such as treatment of naïve hESCs with the DNA demethylating agent 5-azacytidine, followed by evaluation of the impact of lack of hypermethylation on the stability of the naïve pluripotent state. Moreover, it would be useful to validate the findings of both DNMT3A and DNMT3B knockdowns by

generating CRISPR knockout primed hESC lines of the two enzymes and reprogramming them to the naïve state, in order to be sure that any residual protein in the knockdown cell lines is not influencing the cellular phenotype. In parallel, it would also be of interest to generate CRISPR knockout cell lines directly in naïve hESCs, in order to clarify whether DNMT3A and DNMT3B and their *de novo* methylation activity are required for the process of reprogramming or for the stable naïve state. In addition to assessing the impact of DNMT3A and DNMT3B knockout on the naïve pluripotent state in knockout cell lines generated using both strategies, it would also be interesting to evaluate the differentiation potential of the knockout hESCs to ascertain whether the *de novo* methylation is functionally important during the process of differentiation. The SUSD2 cell-surface marker could also be used in this context to verify how efficiently the DNMT knockdown or knockout hESCs transition to naïve pluripotency (Bredenkamp et al. 2019). Current protocols for differentiating naïve hESCs require the cells to be transitioned through the primed state *in vitro* conditions before they differentiate into progenitors of various cell lineages (Takashima et al. 2014), suggesting that a molecular barrier which may involve DNA methylation restricts their ability to differentiate directly from the naïve state. To this end, the abnormal morphology detected in naïve hESCs generated from DNMT3A and DNMT3B knockdown hESCs may be indicative of spontaneous differentiation occurring in the absence of such a barrier and this could be further explored by staining for cellular markers of differentiation.

The relationship between a bivalent histone signature at developmental gene promoters and the acquisition of DNA methylation that we observe upon primed to naïve hESC reprogramming is consistent with reports of hypermethylated bivalent CGIs identified upon iPS reprogramming (Ohm et al. 2010; Doi et al. 2009). Moreover, it mirrors the wealth of literature that has detailed a similar predisposition of CGIs marked by polycomb or bivalent histone modifications in stem cells to exhibit DNA hypermethylation in cancer (Easwaran et al. 2012; Widschwendter et al. 2007; Ohm et al. 2007; Schlesinger et al. 2007; Bernhart et al. 2016). Additionally, our observation that bivalent regions that become hypermethylated upon reprogramming lose H3K4me3 but retain H3K27me3 is consistent with studies in cancer that have also reported the co-occurrence of H3K27me3 and DNA methylation (Gao et al. 2014). Conversely, H3K27me3 and DNA methylation are considered to be mutually exclusive during mammalian development, and genome-wide analysis of histone modifications in naïve hESCs generated using the 5iLAF reprogramming protocol showed a reduction of H3K27me3 at developmental gene promoters (Brinkman et al. 2012; Theunissen et al. 2014). A more comprehensive global analysis of histone modifications upon reprogramming to the naïve

state may improve our understanding of the relationship between bivalency and DNA methylation.

We observed a striking overlap between the CpGs that become hypermethylated upon reprogramming and those hypermethylated in a range of cancer types. The data also revealed that a number of the reprogramming-associated hypermethylated CpGs that exhibit high levels of DNA methylation in cancer also show moderate levels of methylation in the matched normal tissue. This may be because the normal tissue used on TCGA is typically healthy tissue neighbouring the site of the tumour, which may also be undergoing some aberrant molecular processes despite being considered non-malignant. Moreover, the majority of the patients from which the tissues are taken on TCGA are elderly individuals, and the CGI methylation observed may be attributable to ageing-associated DNA methylation which has also been shown to overlap with cancer hypermethylation (Rakyan et al. 2010). To ascertain more comprehensively the degree of overlap between reprogramming-associated hypermethylation and hypermethylation associated with both ageing and cancer, it would be of value in future analyses to compare the data to samples from the NIH Roadmap Epigenomics Project (Bernstein et al. 2010), which comprises DNA methylation data of various tissue types from both young and old healthy individuals.

### 6.1.2 Mechanisms of aberrant DNA methylation

The global demethylation that occurs upon primed to naïve hESC reprogramming is accompanied by a downregulation of the DNMT1 binding partner UHRF1, which is responsible for recruiting DNMT1 to replicating DNA. This downregulation occurs at the protein level but not at the mRNA level, indicating post-transcriptional or post-translational regulation of the protein. While we did not experimentally test whether this is the primary mechanism responsible for the global loss of DNA methylation upon hESC reprogramming, the finding is in line with data from mESCs converted from serum to 2i conditions, where downregulation of UHRF1 protein and the resulting impairment of maintenance methylation has been reported as the major mechanism resulting in a hypomethylated genome in 2i mouse ESCs (von Meyenn et al. 2016). Further exploration is required in future studies to verify whether impaired maintenance methylation is the primary cause of DNA demethylation upon hESC reprogramming.

The DNA loci that are hypermethylated upon reprogramming of primed hESCs to the naïve state exhibit a high degree of overlap with CpGs that are hypermethylated across a variety of



different cancer types. Despite some degree of tissue-specific DNA hypermethylation in cancers from distinct tissue types (Sproul et al. 2012), studies that have performed pan-cancer DNA methylation analysis have reported a considerable overlap in the regions that are hypermethylated across cancer types (Easwaran et al. 2012). Moreover, these studies have highlighted an enrichment of these pan-cancer hypermethylated loci in developmental pathways, often more specifically in pathways associated with neuronal development (Kim et al. 2012; Easwaran et al. 2012). Collectively, the findings of these studies suggest that aberrant DNA methylation does not occur through a stochastic process, as only a subset of genomic loci are affected by this phenomenon, with substantial similarity across cancer types, independent of their underlying genetic mutations. Instead, they point toward an instructive mechanism controlling DNA methylation, relying on the interaction of DNMTs with trans-acting protein complexes and specific DNA sequences (Keshet et al. 2006).

Our data indicates that DNMT3A, and particularly the shorter isoform DNMT3A2 is responsible for the deposition of *de novo* DNA methylation at bivalent promoter CGIs during the early stages of reprogramming primed hESCs to the naïve state of pluripotency. This finding is not in line with a study performed in mouse extraembryonic tissues, where DNMT3B has been identified as the DNMT responsible for depositing DNA methylation at bivalent CGI promoters associated with developmental genes (Smith et al. 2017). It is also not consistent with a study performed in differentiating mESCs, where the longer isoform of DNMT3, DNMT3A1, was found to control *de novo* DNA methylation at bivalent CGIs. Additionally, two independent studies investigating the function of a mutant form of DNMT3A with a point mutation in the PWWP domain have reported aberrant DNA hypermethylation of DNA regions marked by polycomb or bivalent histone modifications, but in each case, hypermethylation is observed upon differentiation when DNMT3A1 becomes more dominantly expressed (Heyn et al. 2018; Sendzikaite et al. 2019). It is likely that the precise roles of DNMT3A and DNMT3B are dependent on the cellular context and the specific isoforms expressed. We observed that during the late transition of reprogramming, as the gene expression of DNMT3B increases and we detect the peak of hypermethylation, both DNMT3A and DNMT3B appear to contribute to the hypermethylation. This finding is in support of correlations drawn in cancer settings between increased DNMT3B expression and gene hypermethylation (Roll et al. 2008; Linhart et al. 2007). To verify whether both enzymes contribute to hypermethylation in the later stages of reprogramming or to ascertain whether one is able to compensate for the other, a DNMT3A and DNMT3B double knockout cell line could be transitioned to the naïve state and

DNA methylation measured to assess whether the absence of both enzymes abrogates *de novo* methylation altogether.

The data presented in this thesis demonstrates that CGIs marked by bivalent histone modifications in primed hESCs are prone to deposition of *de novo* DNA methylation upon reprogramming to the naïve state. However, while the bivalent state appears to be predictive of regions that are hypermethylated during primed to naïve hESC reprogramming and of aberrant DNA hypermethylation in cancer, many bivalent regions do not become hypermethylated, indicating that bivalency is not sufficient to prime a region for hypermethylation. This suggests that other factors may also contribute to this phenomenon. To this end, it has been found that the proximity of a CGI to retrotransposon elements can influence their predisposition to DNA methylation independently of other chromatin features (Estecio et al. 2010). More recently, it has been shown that partially methylated domains which undergo demethylation in cancer acquire intermediate DNA methylation levels, regardless of the underlying functional genomic elements, such that CGIs become hypermethylated (Brinkman et al. 2019). Additionally, a number of studies have identified short DNA motifs that discriminate between DNA loci that are sensitive or resistant to DNA methylation (Keshet et al. 2006; Feltus et al. 2006; Lienert et al. 2011). It is possible that a simple unifying mechanism of aberrant DNA methylation does not exist, but that various determinants are reconciled to generate the DNA methylation patterns that are characteristic of cancer genomes.

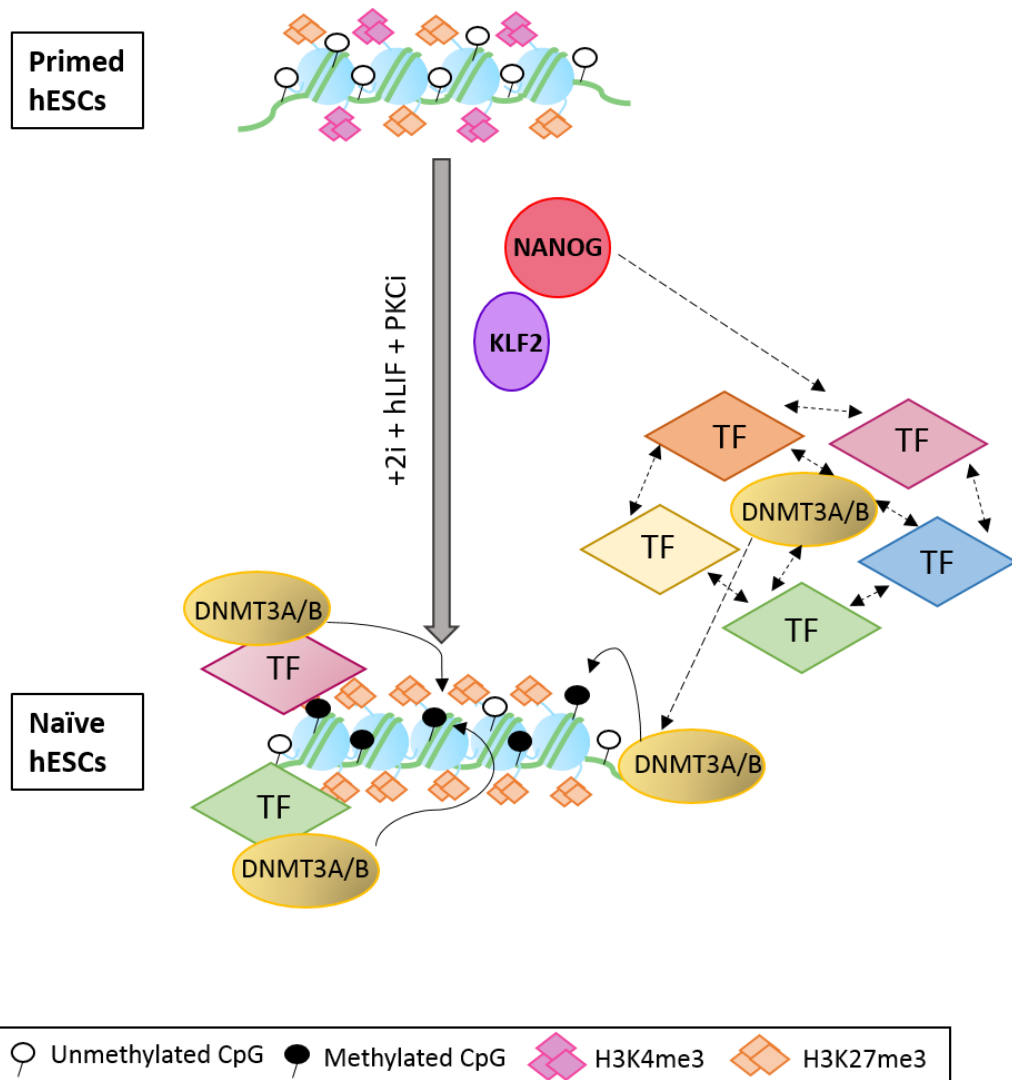
Our data points towards the transcription factor network established upon reprogramming playing a role in the targeting or recruitment of DNMT3A to loci that gain methylation. In particular, our validation of four transcription factors NFKB1, SOX15, ZFX3 and FOXC1 implies that they may each play a role in regulating *de novo* methylation. Based on the current literature, the four transcription factors are not involved in common pathways. However, SOX15 is associated with a stem cell phenotype and is also upregulated upon iPS reprogramming (Nishino et al. 2010; Maruyama et al. 2005). FOXC1 is upregulated in many cancer types and is considered to play a role in conferring stem-like properties in cancer cells (Yang et al. 2017). The NFKB1 gene encodes the 105 kDa protein p105, which undergoes further processing to produce the 50 kDa protein p50, which is a DNA-binding subunit of the NFKB protein complex. Specifically, p50 can form a p50-p50 homodimer or can form a p50-p65 heterodimer, and each of these combinations have been reported to act as transcriptional repressors or to drive tumour-promoting inflammation in a context-dependent manner (Concetti and Wilson 2018). Moreover, p50 has previously been shown to interact with the

H3K4me3 demethylase KDM5A and recruit it to gene promoters to maintain a repressive chromatin state (Zhao et al. 2016). While none of these studies directly report a relationship between the transcription factors and DNA methylation, their roles in stem cells, cancer, and in transcriptional repression suggest that they may also play a role in DNA methylation. The combinations of transcription factors active in different cell types or cancer types may result in the tissue-specific variation observed among the DNA methylomes of different cancer types (Sproul et al. 2012). It would be useful in future studies to evaluate the effect of the knock down of a transcription factor that is not overexpressed during reprogramming to determine whether the high expression of these transcription factors is related to their role in controlling *de novo* methylation. Moreover, although our data suggests that the effect of the transcription factor knockdown may be independent of their binding, the data is based on predicted binding sites and may not be reflective of the true binding profile of each transcription factor in hESCs. Therefore, if each transcription factor plays a role in controlling *de novo* methylation at a subset of sites, it is plausible that generating a multi-knockdown cell line and assessing the impact on DNA methylation may produce a similar effect to the knockdown of DNMT3A. Finally, it would also be of value to reprogram the transcription factor knock down cell lines beyond the early transition through to the stable naïve state to confirm whether or not they successfully reprogram, to distinguish whether the reduction in DNA methylation is due to reduced *de novo* methylation or inefficient reprogramming. Furthermore, whilst we cannot currently differentiate between a direct interaction of DNMT3A with transcription factors or an indirect effect on the targeting of the enzyme by the active transcription factor network, it has previously been demonstrated that DNMT3A can be recruited to the DNA in a loci-specific manner by transcription factors (Brenner et al. 2005) and *in vitro* data supports the ability of DNMT3A to interact directly with numerous transcription factors (Hervouet, Vallette, and Cartron 2014). Future studies could test the putative protein-protein interaction by performing co-immunoprecipitation of DNMT3A and the various transcription factors. Additionally, targeted or global CHIP analysis of the transcription factors would aid the interpretation of the data by clarifying which of the predicted binding sites is actually bound by each of the transcription factors in hESCs and how this is altered upon reprogramming.

Although we did not find evidence of a role for TET1 in protecting CGIs from DNA methylation upon primed to naïve hESC reprogramming, further verification is required in future experiments to confirm whether TET1 was successfully overexpressed at the protein level. Moreover, several other CXXC domain-containing proteins have been implicated in playing a protective role against DNA methylation in other cellular contexts (Long, Blackledge, and Klose

2013), and these proteins may be relevant for the mechanism of hypermethylation in naïve hESCs and in cancer. The selection criteria used to identify candidate proteins that may be involved either in site-specific targeting of DNMTs to the DNA, or in protection of CGIs from DNA methylation was the changing expression levels of the proteins. However, this does not take into account mechanisms such as post-translational modifications or protein localisation that may impact their activity and functions upon reprogramming. It remains possible that the protection of CGIs by CXXC domain-containing proteins acts alongside the transcription factor mediated recruitment of *de novo* DNMTs to the DNA to regulate DNA hypermethylation and this should be addressed further in future studies.

Our data are also indicative of the overexpressed pluripotency factors NANOG and KLF2 coordinating *de novo* methylation, however studies have shown that KLF2 is not expressed *in vivo* in the human ICM (Yan et al. 2013; Blakeley et al. 2015), where we also observe hypermethylation. Additionally, we observe comparable hypermethylation in naïve hESCs generated using two transgene-independent methods of reprogramming. This collectively suggests that the core pluripotency network, to which NANOG belongs, may be responsible for coordinating the transcriptional changes that drive DNA hypermethylation. Figure 6.1 depicts the model of *de novo* methylation during reprogramming that emerges through the intersection of various lines of evidence from the data in this thesis, including the *de novo* DNMTs, bivalent histone modifications, the transcription factor network and pluripotency factors. There is growing evidence in the literature regarding the acquisition of stem-like properties and expression of pluripotency genes in cancers, which contribute to intratumour heterogeneity and plasticity and may facilitate the formation of cancer stem cells (Ben-Porath et al. 2008; Chen and He 2016; Friedmann-Morvinski and Verma 2014). This makes it intriguing to speculate that a transcriptional programme associated with the pluripotency network could drive a shared mechanism of hypermethylation during reprogramming to naïve pluripotency and in cancer development, either preceding or in conjunction with genetic mutations.



**Figure 6.1. Model of de novo DNA hypermethylation upon primed to naïve hESC reprogramming.** The model that emerges from the data in this thesis is that upon reprogramming of primed hESCs to the naïve state, DNMT3A, and later DNMT3B carry out de novo methylation at bivalent CGI promoters, which concomitantly lose H3K4me3. The de novo DNMTs are targeted to these loci by DNA-binding transcription factors (TFs) which act either through a direct or indirect interaction with the DNMTs. This process is further regulated by NANOG and the naïve pluripotency network.

### 6.1.3 Reprogramming of hESCs as a model system for cancer development

Alongside the epigenetic, transcriptional, and metabolic changes occurring during the transition from primed to naïve human pluripotency, our data shows preliminary evidence of multiple manifestations of chromosomal instability upon reprogramming, resulting in the accumulation of aneuploid cells in the naïve hESC population. Similar to frequent aberrant DNA methylation patterns, aneuploidy is a common feature of human cancers (Gordon, Resio, and Pellman 2012). Whether or not the genomic instability is related to the changing epigenetic landscape cannot be inferred from the current data, but it is evident that both processes are induced as cells acquire a more primitive pluripotent state. These processes may be analogous to those that occur in differentiated somatic cells, as they acquire enhanced self-renewal and proliferative capacity during iPS reprogramming and in cancer development. In line with this, there is now early evidence of *in vivo* processes that resemble reprogramming, such as the recently described process of paligenosis (Willet et al. 2018). In this process, fully differentiated stomach cells regain plasticity and proliferative potential and re-enter the cell cycle in response to tissue injury, with the potential to result in dysplasia in the context of chronic injury or inflammation that enables the accumulation of mutations (Jin and Mills 2019). Similar processes involving the plasticity of differentiated cells and their ability to undergo dedifferentiation and reprogramming of their cellular identity in response to tissue injury or stress have been reported in several other tissue types (Puri, Folias, and Hebrok 2015; Wang et al. 2017; Boerboom et al. 2017; Yu et al. 2018; Desai, Brownfield, and Krasnow 2014; Miyajima, Tanaka, and Itoh 2014). Moreover, signalling pathways such as the MAPK pathway and Wnt signalling which are involved in primed to naïve hESC reprogramming and stem cell maintenance have also been implicated in the regulation of such processes (Boerboom et al. 2017), and these pathways are some of the most frequently misregulated signalling pathways in cancer (Zhan, Rindtorff, and Boutros 2017; Dhillon et al. 2007). While the epigenetic landscape during these *in vivo* dedifferentiation processes has not yet been assessed, the change in cellular state implies that epigenetic changes are involved as cellular identity is epigenetically regulated. As with partial reprogramming induced by transient expression of the OSKM factors *in vivo* which results in dysplasia, the improper regulation of *in vivo* dedifferentiation processes in response to tissue injury, particularly chronic conditions, may increase the likelihood of the generation of dysplastic cells, as increased cellular and epigenetic plasticity can facilitate the hallmarks of cancer (Ohnishi et al. 2014; Flavahan, Gaskell, and Bernstein 2017). Figure 6.2 summarises these ideas into a model depicting dedifferentiation

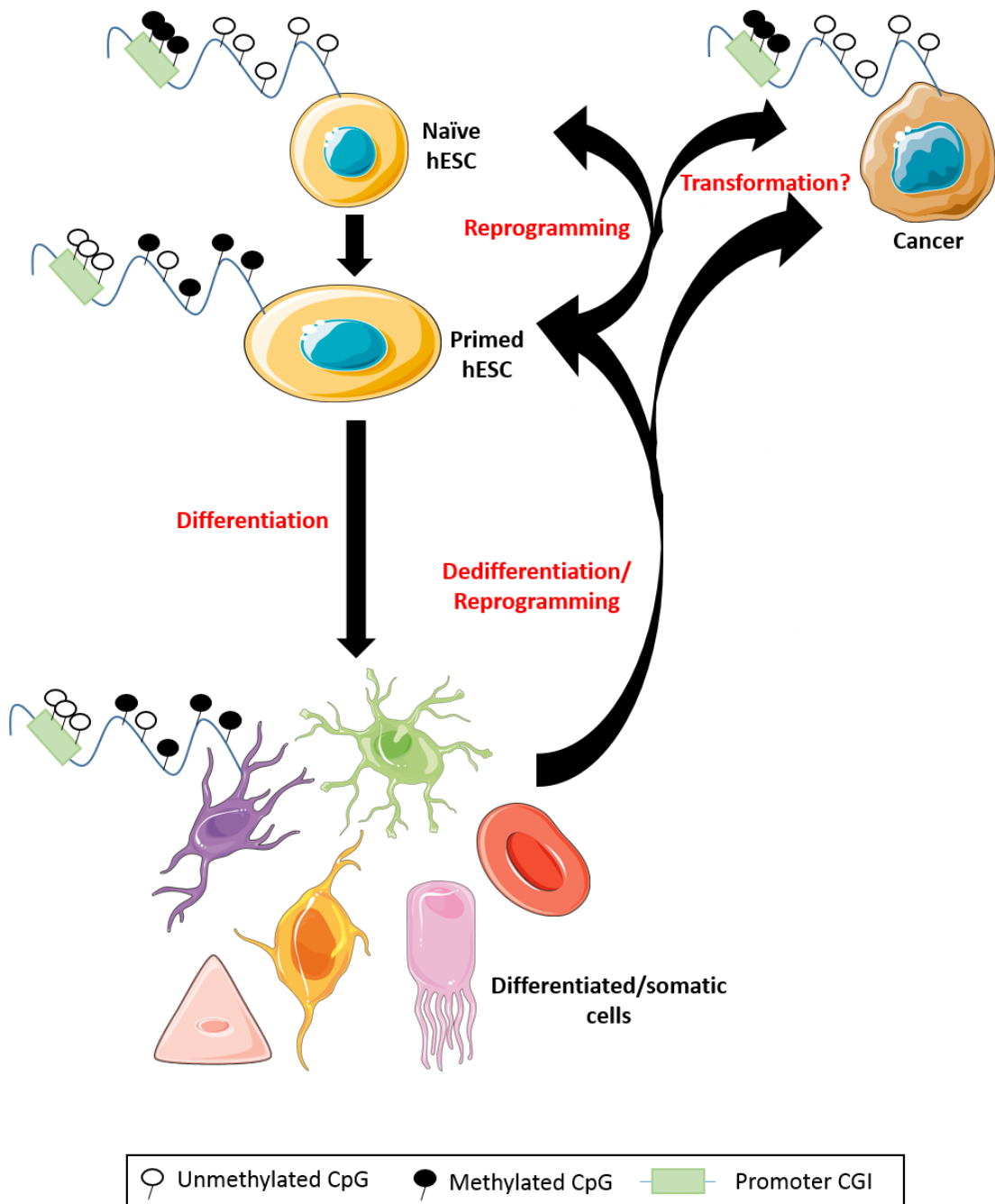
and reprogramming as potential routes to cancer formation through shared molecular processes, including DNA methylation.

The understanding of the development and early stages of cancer is limited by the scarcity of suitable model systems. Cancer has long been considered a genetic disease, with the clonal model of cancer having prevailed in the literature, but the initial processes that facilitate mutations at specific sites in specific tissues of origin and result in the transformation of a normal cell into a cancer cell remain ambiguous (Polak et al. 2015; Feinberg, Ohlsson, and Henikoff 2006). There is growing evidence that clonal expansion of cancer driver mutations exists at low frequencies in the healthy ageing population, indicating that additional processes are required for disease initiation (Xie et al. 2014; Young et al. 2016; Blokzijl et al. 2016). Whether epigenetic changes precede mutations and prime cells for genomic instability or whether aberrant epigenetic changes are the result of DNA mutations also remains an area that requires further study, as both possibilities have been proposed previously (Feinberg, Ohlsson, and Henikoff 2006; Youn et al. 2018). The majority of cancer studies focus on comparing cancer cells to normal tissue of the closest analogous tissue type. As a result, much of the data regarding aberrant epigenetic modifications in cancer is correlative, while causal relationships have been more difficult to demonstrate. Moreover, the molecular processes that underlie many of the aberrations that are detected in cancer remain poorly characterised as they may no longer be active in the cancer cells of patients at the time of study. A recent technological advance that has enabled the modification of the epigenome in a targeted manner *in vitro* using CRISPR technology has facilitated the demonstration of hypermethylation driving aberrant molecular processes in the absence of cancer driver mutations (Saunders et al. 2017). Moreover, using organoid models which are a closer representation of *in vivo* tissues compared to cell lines, it has been shown that hypermethylation can accelerate oncogene-driven transformation *in vitro* (Tao et al. 2019). While these studies represent major advances in our understanding of the role of DNA methylation in cancer by indicating that DNA hypermethylation can play a driving role in carcinogenesis, the underlying molecular mechanisms that bring about aberrant hypermethylation still have not been elucidated.

The primed to naïve hESC reprogramming system is unique in that it enables thorough temporal characterisation of the molecular processes that give rise to the abnormal epigenetic patterns observed in both naïve hESCs and across many cancer types. Despite hESCs being far removed from somatic cells which form the bulk of human tumours, the parallels between the epigenetic landscapes of naïve hESCs and cancer cells suggest that reprogramming and the

transformation of normal cells into cancer cells may share common epigenetic trajectories. Moreover, it is also plausible that the MEK1/2 and GSK3 $\beta$  pathway inhibition, along with the overexpression of NANOG and KLF2 act to phenocopy the signalling effects of cancer driver mutations, thus resulting in comparable DNA methylomes. Molecular mechanistic insight gained from studying primed to naïve reprogramming may be relevant for understanding cancer formation. Mechanisms deciphered in stem cells could be further tested in more physiological model systems such as organoids to enhance the understanding of the epigenetic and mutational trajectories that facilitate transformation.





**Figure 6.2. Reprogramming and cancer.** During normal human development, stem cells give rise to various lineages of differentiated/somatic cells through the process of differentiation, which involves epigenetic changes that modulate changes in cell identity. Differentiation can be reversed in vivo in response to tissue injury that results in dedifferentiation, or through induced pluripotent stem cell reprogramming, and these stem cells can be further reprogramming to a more primitive naïve state of pluripotency in vitro. The molecular processes underlying these changes in cellular state, particularly the epigenetic reprogramming, may parallel processes occurring during malignant transformation. Of particular relevance to this thesis, both naïve hESCs and cancer cells exhibit genome-wide DNA hypomethylation and promoter CpG islands (CGI) hypermethylation, while somatic cells and primed hESCs exhibit a globally hypermethylated genome with unmethylated CGIs.

## 6.2 Concluding Remarks

The commonality in DNA methylation patterns across cancer types, each harbouring different driver mutations, suggests that DNA methylation changes occur early in tumourigenesis, and this has been demonstrated previously (Hanley et al. 2017), though models of early cancer development are limited. In line with this, the notion that cancer cells follow an evolutionary trajectory towards a stem cell state (Chen and He 2016; Ben-Porath et al. 2008) makes the transition from primed to naïve pluripotency an interesting model to study biological processes such as DNA methylation and DNA damage that results in chromosomal and genomic instability, as these processes likely occur early during cellular transformation or cancer initiation, and may be analogous to dedifferentiation. Using this system, we have demonstrated that upon reprogramming primed hESCs to the naïve state, transcription factors and the pluripotency network facilitate *de novo* DNA methylation amidst global demethylation that likely results from impaired maintenance methylation. In parallel, chromosomal instability accumulates in the cells and results in aneuploidy. Whether or not additional molecular features of the primed to naïve state transition that appear analogous to cancer hallmarks (Hanahan and Weinberg 2011) - such as altered metabolism (Takashima et al. 2014), loss of imprints, and loss of DNA hydroxymethylation (Ficz and Gribben 2014) - are related to the changing epigenetic landscape remains unexplored, but further use of this model system may shed light on the emergence of these characteristics during cellular transformation. Primed to naïve hESC reprogramming may provide a good model system to understand whether cellular reprogramming and the molecular processes associated with it play a role in tumourigenesis.

## Bibliography

- Abad, M., L. Mosteiro, C. Pantoja, M. Canamero, T. Rayon, I. Ors, O. Grana, D. Megias, O. Dominguez, D. Martinez, M. Manzanares, S. Ortega, and M. Serrano. 2013. 'Reprogramming in vivo produces teratomas and iPS cells with totipotency features', *Nature*, 502: 340-5.
- Abelson, S., G. Collord, S. W. K. Ng, O. Weissbrod, N. Mendelson Cohen, E. Niemeyer, N. Barda, P. C. Zuzarte, L. Heisler, Y. Sundaravadanam, R. Luben, S. Hayat, T. T. Wang, Z. Zhao, I. Cirlan, T. J. Pugh, D. Soave, K. Ng, C. Latimer, C. Hardy, K. Raine, D. Jones, D. Hoults, A. Britten, J. D. McPherson, M. Johansson, F. Mbabaali, J. Eagles, J. K. Miller, D. Pasternack, L. Timms, P. Krzyzanowski, P. Awadalla, R. Costa, E. Segal, S. V. Bratman, P. Beer, S. Behjati, I. Martincorena, J. C. Y. Wang, K. M. Bowles, J. R. Quiros, A. Karakatsani, C. La Vecchia, A. Trichopoulou, E. Salamanca-Fernandez, J. M. Huerta, A. Barricarte, R. C. Travis, R. Tumino, G. Masala, H. Boeing, S. Panico, R. Kaaks, A. Kramer, S. Sieri, E. Riboli, P. Vineis, M. Foll, J. McKay, S. Polidoro, N. Sala, K. T. Khaw, R. Vermeulen, P. J. Campbell, E. Papaemmanuil, M. D. Minden, A. Tanay, R. D. Balicer, N. J. Wareham, M. Gerstung, J. E. Dick, P. Brennan, G. S. Vassiliou, and L. I. Shlush. 2018. 'Prediction of acute myeloid leukaemia risk in healthy individuals', *Nature*, 559: 400-04.
- Alcolea, M.P., P. Casado, J.C. Rodríguez-Prados, B. Vanhaesebroeck, and P.R. Cutillas. 2012. 'Phosphoproteomic Analysis of Leukemia Cells under Basal and Drug-treated Conditions Identifies Markers of Kinase Pathway Activation and Mechanisms of Resistance', *Molecular and Cellular Proteomics*, 11: 453-66.
- Alexandrov, L. B., S. Nik-Zainal, D. C. Wedge, S. A. Aparicio, S. Behjati, A. V. Biankin, G. R. Bignell, N. Bolli, A. Borg, A. L. Borresen-Dale, S. Boyault, B. Burkhardt, A. P. Butler, C. Caldas, H. R. Davies, C. Desmedt, R. Eils, J. E. Eyfjord, J. A. Foekens, M. Greaves, F. Hosoda, B. Hutter, T. Ilicic, S. Imbeaud, M. Imielinski, N. Jager, D. T. Jones, D. Jones, S. Knappskog, M. Kool, S. R. Lakhani, C. Lopez-Otin, S. Martin, N. C. Munshi, H. Nakamura, P. A. Northcott, M. Pajic, E. Papaemmanuil, A. Paradiso, J. V. Pearson, X. S. Puente, K. Raine, M. Ramakrishna, A. L. Richardson, J. Richter, P. Rosenstiel, M. Schlesner, T. N. Schumacher, P. N. Span, J. W. Teague, Y. Totoki, A. N. Tutt, R. Valdes-Mas, M. M. van Buuren, L. van 't Veer, A. Vincent-Salomon, N. Waddell, L. R. Yates, Initiative Australian Pancreatic Cancer Genome, IcgC Breast Cancer Consortium, IcgC Mmml- Seq Consortium, IcgC PedBrain, J. Zucman-Rossi, P. A. Futreal, U. McDermott, P. Lichter, M. Meyerson, S. M. Grimmond, R. Siebert, E. Campo, T. Shibata, S. M. Pfister, P. J. Campbell, and M. R. Stratton. 2013. 'Signatures of mutational processes in human cancer', *Nature*, 500: 415-21.
- Allfrey, V.G., R. Faulkner, and A.E. Mirksy. 1964. 'Acetylation and methylation of histones and their possible role in the regulation of RNA synthesis', *Biochemistry*, 51.
- Azuara, V., P. Perry, S. Sauer, M. Spivakov, H.F. Jørgensen, R.M. John, M. Gouti, M. Casanova, G. Warnes, M. Merckenschlager, and A.G. Fisher. 2006. 'Chromatin signatures of pluripotent cell lines', *Nature Cell Biology*, 8.
- Bachman, M., S. Uribe-Lewis, X. Yang, M. Williams, A. Murrell, and S. Balasubramanian. 2014. '5-Hydroxymethylcytosine is a predominantly stable DNA modification', *Nat Chem*, 6: 1049-55.
- Bachman, Martin, Santiago Uribe-Lewis, Xiaoping Yang, Heather E. Burgess, Mario Iurlaro, Wolf Reik, Adele Murrell, and Shankar Balasubramanian. 2015. '5-Formylcytosine can be a stable DNA modification in mammals', *Nature Chemical Biology*, 11: 555.

- Bailey, T. L., M. Boden, F. A. Buske, M. Frith, C. E. Grant, L. Clementi, J. Ren, W. W. Li, and W. S. Noble. 2009. 'MEME SUITE: tools for motif discovery and searching', *Nucleic Acids Res*, 37: W202-8.
- Bass, A. J., H. Watanabe, C. H. Mermel, S. Yu, S. Perner, R. G. Verhaak, S. Y. Kim, L. Wardwell, P. Tamayo, I. Gat-Viks, A. H. Ramos, M. S. Woo, B. A. Weir, G. Getz, R. Beroukhim, M. O'Kelly, A. Dutt, O. Rozenblatt-Rosen, P. Dziunycz, J. Komisarof, L. R. Chirieac, C. J. Lafargue, V. Scheble, T. Wilbertz, C. Ma, S. Rao, H. Nakagawa, D. B. Stairs, L. Lin, T. J. Giordano, P. Wagner, J. D. Minna, A. F. Gazdar, C. Q. Zhu, M. S. Brose, I. Cecconello, U. Ribeiro, Jr., S. K. Marie, O. Dahl, R. A. Shivdasani, M. S. Tsao, M. A. Rubin, K. K. Wong, A. Regev, W. C. Hahn, D. G. Beer, A. K. Rustgi, and M. Meyerson. 2009. 'SOX2 is an amplified lineage-survival oncogene in lung and esophageal squamous cell carcinomas', *Nat Genet*, 41: 1238-42.
- Baubec, T., D. F. Colombo, C. Wirbelauer, J. Schmidt, L. Burger, A. R. Krebs, A. Akalin, and D. Schubeler. 2015. 'Genomic profiling of DNA methyltransferases reveals a role for DNMT3B in genic methylation', *Nature*, 520: 243-7.
- Baylin, S. B., and P. A. Jones. 2011. 'A decade of exploring the cancer epigenome - biological and translational implications', *Nat Rev Cancer*, 11: 726-34.
- Beck, B., and C. Blanpain. 2013. 'Unravelling cancer stem cell potential', *Nat Rev Cancer*, 13: 727-38.
- Ben-Porath, I., M. W. Thomson, V. J. Carey, R. Ge, G. W. Bell, A. Regev, and R. A. Weinberg. 2008. 'An embryonic stem cell-like gene expression signature in poorly differentiated aggressive human tumors', *Nat Genet*, 40: 499-507.
- Bernhart, S. H., H. Kretzmer, L. M. Holdt, F. Juhling, O. Ammerpohl, A. K. Bergmann, B. H. Northoff, G. Doose, R. Siebert, P. F. Stadler, and S. Hoffmann. 2016. 'Changes of bivalent chromatin coincide with increased expression of developmental genes in cancer', *Sci Rep*, 6: 37393.
- Bernstein, B. E., T. S. Mikkelsen, X. Xie, M. Kamal, D. J. Huebert, J. Cuff, B. Fry, A. Meissner, M. Wernig, K. Plath, R. Jaenisch, A. Wagschal, R. Feil, S. L. Schreiber, and E. S. Lander. 2006. 'A bivalent chromatin structure marks key developmental genes in embryonic stem cells', *Cell*, 125: 315-26.
- Bernstein, B. E., J. A. Stamatoyannopoulos, J. F. Costello, B. Ren, A. Milosavljevic, A. Meissner, M. Kellis, M. A. Marra, A. L. Beaudet, J. R. Ecker, P. J. Farnham, M. Hirst, E. S. Lander, T. S. Mikkelsen, and J. A. Thomson. 2010. 'The NIH Roadmap Epigenomics Mapping Consortium', *Nat Biotechnol*, 28: 1045-8.
- Bester, A. C., M. Roniger, Y. S. Oren, M. M. Im, D. Sarni, M. Chaoat, A. Bensimon, G. Zamir, D. S. Shewach, and B. Kerem. 2011. 'Nucleotide deficiency promotes genomic instability in early stages of cancer development', *Cell*, 145: 435-46.
- Bestor, T. 2000. 'The DNAmethyltransferases of mammals', *Hum Mol Genet*, 9: 2395-402.
- Bestor, TH. Bourc'his, D. 2004. 'Transposon Silencing and Imprint Establishment in Mammalian Germ Cells', *COLD Spring Harb Symp Quant Biol*, 69: 381-87.
- Bhutani, N., J. J. Brady, M. Damian, A. Sacco, S. Y. Corbel, and H. M. Blau. 2010. 'Reprogramming towards pluripotency requires AID-dependent DNA demethylation', *Nature*, 463: 1042-7.
- Bird, A. 1980. 'DNA methylation and the frequency of CpG in animal DNA', *Nucleic Acids Res*, 8.
- Bird, A., M. Taggart, M. Frommer, O.J. Miller, and D. Macleod. 1985. 'A fraction of the mouse genome that is derived from islands of non-methylated, CpG-rich DNA', *Cell*, 40.
- Bizard, A. H., and I. D. Hickson. 2018. 'Anaphase: a fortune-teller of genomic instability', *Curr Opin Cell Biol*, 52: 112-19.
- Blackledge, N. P., J. C. Zhou, M. Y. Tolstorukov, A. M. Farcas, P. J. Park, and R. J. Klose. 2010. 'CpG islands recruit a histone H3 lysine 36 demethylase', *Mol Cell*, 38: 179-90.

- Blakeley, P., N. M. Fogarty, I. Del Valle, S. E. Wamaita, T. X. Hu, K. Elder, P. Snell, L. Christie, P. Robson, and K. K. Niakan. 2015. 'Defining the three cell lineages of the human blastocyst by single-cell RNA-seq', *Development*, 142: 3613.
- Blokzijl, F., J. de Ligt, M. Jager, V. Sasselli, S. Roerink, N. Sasaki, M. Huch, S. Boymans, E. Kuijk, P. Prins, I. J. Nijman, I. Martincorena, M. Mokry, C. L. Wiegerinck, S. Middendorp, T. Sato, G. Schwank, E. E. Nieuwenhuis, M. M. Verstegen, L. J. van der Laan, J. de Jonge, I. Jzermans JN, R. G. Vries, M. van de Wetering, M. R. Stratton, H. Clevers, E. Cuppen, and R. van Boxtel. 2016. 'Tissue-specific mutation accumulation in human adult stem cells during life', *Nature*, 538: 260-64.
- Boerboom, A., V. Dion, A. Chariot, and R. Franzen. 2017. 'Molecular Mechanisms Involved in Schwann Cell Plasticity', *Front Mol Neurosci*, 10: 38.
- Boroviak, T., G. G. Stirparo, S. Dietmann, I. Hernando-Herraez, H. Mohammed, W. Reik, A. Smith, E. Sasaki, J. Nichols, and P. Bertone. 2018. 'Single cell transcriptome analysis of human, marmoset and mouse embryos reveals common and divergent features of preimplantation development', *Development*, 145.
- Bostick, M., J.K. Kim, P.O. Estève, A. Clark, S. Pradhan, and S.E. Jacobsen. 2007. 'UHRF1 Plays a Role in Maintaining DNA Methylation in Mammalian Cells', *Science*, 317: 1760-64.
- Boulard, M., J. R. Edwards, and T. H. Bestor. 2015. 'FBXL10 protects Polycomb-bound genes from hypermethylation', *Nat Genet*, 47: 479-85.
- Boyer, L. A., T. I. Lee, M. F. Cole, S. E. Johnstone, S. S. Levine, J. P. Zucker, M. G. Guenther, R. M. Kumar, H. L. Murray, R. G. Jenner, D. K. Gifford, D. A. Melton, R. Jaenisch, and R. A. Young. 2005. 'Core transcriptional regulatory circuitry in human embryonic stem cells', *Cell*, 122: 947-56.
- Boyes, J., and A. Bird. 1991. 'DNA methylation inhibits transcription indirectly via a methyl-CpG binding protein', *Cell*, 64.
- Brandeis, M., D. Frank, I. Keshet, Z. Siegfried, M. Mendelsohn, A. Nemest, V. Temper, and A.R.B.H. Cedar. 1994. 'Sp1 elements protect a CpG island from de novo methylation', *Nature*, 371.
- Bredenkamp, N., G. G. Stirparo, J. Nichols, A. Smith, and G. Guo. 2019. 'The Cell-Surface Marker Sushi Containing Domain 2 Facilitates Establishment of Human Naive Pluripotent Stem Cells', *Stem Cell Reports*.
- Brenner, C., R. Deplus, C. Didelot, A. Lorient, E. Vire, C. De Smet, A. Gutierrez, D. Danovi, D. Bernard, T. Boon, P.G. Pelicci, B. Amati, T. Kouzarides, Y. de Launoit, L. Di Croce, and F. Fuks. 2005. 'Myc represses transcription through recruitment of DNA methyltransferase corepressor', *The EMBO Journal*, 24: 336-46.
- Brinkman, A. B., H. Gu, S. J. Bartels, Y. Zhang, F. Matarese, F. Simmer, H. Marks, C. Bock, A. Gnirke, A. Meissner, and H. G. Stunnenberg. 2012. 'Sequential ChIP-bisulfite sequencing enables direct genome-scale investigation of chromatin and DNA methylation cross-talk', *Genome Res*, 22: 1128-38.
- Brinkman, A. B., S. Nik-Zainal, F. Simmer, F. G. Rodriguez-Gonzalez, M. Smid, L. B. Alexandrov, A. Butler, S. Martin, H. Davies, D. Glodzik, X. Zou, M. Ramakrishna, J. Staaf, M. Ringner, A. Sieuwerts, A. Ferrari, S. Morganella, T. Fleischer, V. Kristensen, M. Gut, M. J. van de Vijver, A. L. Borresen-Dale, A. L. Richardson, G. Thomas, I. G. Gut, J. W. M. Martens, J. A. Foekens, M. R. Stratton, and H. G. Stunnenberg. 2019. 'Partially methylated domains are hypervariable in breast cancer and fuel widespread CpG island hypermethylation', *Nat Commun*, 10: 1749.
- Brons, I. G., L. E. Smithers, M. W. Trotter, P. Rugg-Gunn, B. Sun, S. M. Chuva de Sousa Lopes, S. K. Howlett, A. Clarkson, L. Ahrlund-Richter, R. A. Pedersen, and L. Vallier. 2007. 'Derivation of pluripotent epiblast stem cells from mammalian embryos', *Nature*, 448: 191-5.

- Burrell, R. A., S. E. McClelland, D. Endesfelder, P. Groth, M. C. Weller, N. Shaikh, E. Domingo, N. Kanu, S. M. Dewhurst, E. Gronroos, S. K. Chew, A. J. Rowan, A. Schenk, M. Sheffer, M. Howell, M. Kschischo, A. Behrens, T. Helleday, J. Bartek, I. P. Tomlinson, and C. Swanton. 2013. 'Replication stress links structural and numerical cancer chromosomal instability', *Nature*, 494: 492-96.
- Cacchiarelli, D., C. Trapnell, M. J. Ziller, M. Soumillon, M. Cesana, R. Karnik, J. Donaghey, Z. D. Smith, S. Ratanasirintrao, X. Zhang, S. J. Ho Sui, Z. Wu, V. Akopian, C. A. Gifford, J. Doench, J. L. Rinn, G. Q. Daley, A. Meissner, E. S. Lander, and T. S. Mikkelsen. 2015. 'Integrative Analyses of Human Reprogramming Reveal Dynamic Nature of Induced Pluripotency', *Cell*, 162: 412-24.
- Cedar, H., and Y. Bergman. 2009. 'Linking DNA methylation and histone modification: patterns and paradigms', *Nat Rev Genet*, 10: 295-304.
- Chan, Annie On-On, Russell R. Broaddus, Patrick S. Houlihan, Jean-Pierre J. Issa, Stanley R. Hamilton, and Asif Rashid. 2002. 'CpG Island Methylation in Aberrant Crypt Foci of the Colorectum', *The American Journal of Pathology*, 160: 1823-30.
- Chan, K. L., and I. D. Hickson. 2009. 'On the origins of ultra-fine anaphase bridges', *Cell Cycle*, 8: 3065-6.
- Chan, Y. S., J. Goke, J. H. Ng, X. Lu, K. A. Gonzales, C. P. Tan, W. Q. Tng, Z. Z. Hong, Y. S. Lim, and H. H. Ng. 2013. 'Induction of a human pluripotent state with distinct regulatory circuitry that resembles preimplantation epiblast', *Cell Stem Cell*, 13: 663-75.
- Chan, Y. W., and S. C. West. 2018. 'A new class of ultrafine anaphase bridges generated by homologous recombination', *Cell Cycle*, 17: 2101-09.
- Chang, Y., L. Sun, K. Kokura, J. R. Horton, M. Fukuda, A. Espejo, V. Izumi, J. M. Koomen, M. T. Bedford, X. Zhang, Y. Shinkai, J. Fang, and X. Cheng. 2011. 'MPP8 mediates the interactions between DNA methyltransferase Dnmt3a and H3K9 methyltransferase GLP/G9a', *Nat Commun*, 2: 533.
- Chen, H., I. Aksoy, F. Gonnot, P. Osteil, M. Aubry, C. Hamela, C. Rognard, A. Hochard, S. Voisin, E. Fontaine, M. Mure, M. Afanassieff, E. Cleroux, S. Guibert, J. Chen, C. Vallot, H. Aclouque, C. Genthon, C. Donnadiou, J. De Vos, D. Sanlaville, J. F. Guerin, M. Weber, L. W. Stanton, C. Rougeulle, B. Pain, P. Y. Bourillot, and P. Savatier. 2015. 'Reinforcement of STAT3 activity reprogrammes human embryonic stem cells to naive-like pluripotency', *Nat Commun*, 6: 7095.
- Chen, H., and X. He. 2016. 'The Convergent Cancer Evolution toward a Single Cellular Destination', *Mol Biol Evol*, 33: 4-12.
- Chen, R.Z., U. Pettersson, C. Beard, L. Jackson-Grusby, and R. Jaenisch. 1998. 'DNA hypomethylation leads to elevated mutation rates', *Nature*, 395.
- Chen, T., N. Tsujimoto, and E. Li. 2004. 'The PWWP domain of Dnmt3a and Dnmt3b is required for directing DNA methylation to the major satellite repeats at pericentric heterochromatin', *Mol Cell Biol*, 24: 9048-58.
- Chen, T., Y. Ueda, S. Xie, and E. Li. 2002. 'A novel Dnmt3a isoform produced from an alternative promoter localizes to euchromatin and its expression correlates with active de novo methylation', *J Biol Chem*, 277: 38746-54.
- Choi, J., A. J. Huebner, K. Clement, R. M. Walsh, A. Savol, K. Lin, H. Gu, B. Di Stefano, J. Brumbaugh, S. Y. Kim, J. Sharif, C. M. Rose, A. Mohammad, J. Odajima, J. Charron, T. Shioda, A. Gnirke, S. Gygi, H. Koseki, R. I. Sadreyev, A. Xiao, A. Meissner, and K. Hochedlinger. 2017. 'Prolonged Mek1/2 suppression impairs the developmental potential of embryonic stem cells', *Nature*.
- Chuang, L.S., H.I. Ian, T.W. Koh, H.H. Ng, G. Xu, and B.F. Li. 1997. 'Human DNA-(Cytosine-5) Methyltransferase-PCNA Complex as a Target for p21WAF1', *Science*, 277: 1996-2000.

- Cimini, D., B. Howell, P. Maddox, A. Khodjakov, F. Degrossi, and E.D. Salmon. 2001. 'Merotelic Kinetochore Orientation Is a Major Mechanism of Aneuploidy in Mitotic Mammalian Tissue Cells', *Journal of Cell Biology*, 153: 517-27.
- Cobaleda, C., N. Gutiérrez-Cianca, J. Pérez-Losada, T. Flores, R. García-Sanz, M. González, and I. Sánchez-García. 2000. 'A primitive hematopoietic cell is the target for the leukemic transformation in human Philadelphia-positive acute lymphoblastic leukemia', *Blood*, 95: 1007.
- Collier, A. J., S. P. Panula, J. P. Schell, P. Chovanec, A. Plaza Reyes, S. Petropoulos, A. E. Corcoran, R. Walker, I. Douagi, F. Lanner, and P. J. Rugg-Gunn. 2017. 'Comprehensive Cell Surface Protein Profiling Identifies Specific Markers of Human Naive and Primed Pluripotent States', *Cell Stem Cell*.
- Concetti, J., and C. L. Wilson. 2018. 'NFkB1 and Cancer: Friend or Foe?', *Cells*, 7.
- Cooper, David N., and Hagop Youssoufian. 1988. 'The CpG dinucleotide and human genetic disease', *Human Genetics*, 78: 151-55.
- Costello, J.F. , M.C. Frühwald, D.J. Smiraglia, L.J. Rush, G.P. Robertson, X. Gao, F.A. Wright, J.D. Feramisco, P. Peltomäki, J.C. Lang, D.E. Schuller, L. Yu, C.D. Bloomfield, M.A. Caligiuri, A. Yates, R. Nishikawa, H.J.S. Huang, N.J. Petrelli, X. Zhang, M.S. O'Dorisio, W.A. Held, W.K. Cavenee, and C. Plass. 2000. 'Aberrant CpG-island methylation has non-random and tumour-type-specific patterns', *Nature Genetics*, 25.
- Court, F., and P. Arnaud. 2017. 'An annotated list of bivalent chromatin regions in human ES cells: a new tool for cancer epigenetic research', *Oncotarget*, 8.
- Cui, H., M. Cruz-Correa, F.M. Giardiello, D.F. Hutcheon, D.R. Kafonek, S. Brandenburg, Y. Wu, X. He, N.R. Powe, and A.P. Feinberg. 2003. 'Loss of IGF2 Imprinting: A Potential Marker of Colorectal Cancer Risk', *Science*, 299.
- Cuozzo, C., A. Porcellini, T. Angrisano, M. Morano, B. Lee, A. Di Pardo, S. Messina, R. Iuliano, A. Fusco, M.R. Santillo, M.T. Muller, L. Chiariotti, M.E. Gottesman, and E.V. Avvedimento. 2007. 'DNA Damage, Homology-Directed Repair, and DNA Methylation', *PLoS Genet*, 3.
- Dang, C. V. 2012. 'MYC on the path to cancer', *Cell*, 149: 22-35.
- Daskalos, A., G. Nikolaidis, G. Xinarianos, P. Savvari, A. Cassidy, R. Zakopoulou, A. Kotsinas, V. Gorgoulis, J. K. Field, and T. Liloglou. 2009. 'Hypomethylation of retrotransposable elements correlates with genomic instability in non-small cell lung cancer', *Int J Cancer*, 124: 81-7.
- Davidson, K. C., E. A. Mason, and M. F. Pera. 2015. 'The pluripotent state in mouse and human', *Development*, 142: 3090-9.
- Deaton, A. M., and A. Bird. 2011. 'CpG islands and the regulation of transcription', *Genes Dev*, 25: 1010-22.
- Denis, H., M. N. Ndlovu, and F. Fuks. 2011. 'Regulation of mammalian DNA methyltransferases: a route to new mechanisms', *EMBO Rep*, 12: 647-56.
- Denissenko, Mikhail F., James X. Chen, Moon-shong Tang, and Gerd P. Pfeifer. 1997. 'Cytosine methylation determines hot spots of DNA damage in the human P53 gene', *Proceedings of the National Academy of Sciences*, 94: 3893.
- Desai, P., N. Mencia-Trinchant, O. Savenkov, M. S. Simon, G. Cheang, S. Lee, M. Samuel, E. K. Ritchie, M. L. Guzman, K. V. Ballman, G. J. Roboz, and D. C. Hassane. 2018. 'Somatic mutations precede acute myeloid leukemia years before diagnosis', *Nat Med*, 24: 1015-23.
- Desai, T. J., D. G. Brownfield, and M. A. Krasnow. 2014. 'Alveolar progenitor and stem cells in lung development, renewal and cancer', *Nature*, 507: 190-4.
- Desrosiers, R.C., A. Bakker, J. Kamine, L.A. Falk, R.D. Hunt, and N.W. King. 1985. 'Hypomethylation of DNA from Benign and Malignant Human Colon Neoplasms', *Science*.

- Dhayalan, A., A. Rajavelu, P. Rathert, R. Tamas, R. Z. Jurkowska, S. Ragozin, and A. Jeltsch. 2010. 'The Dnmt3a PWWP domain reads histone 3 lysine 36 trimethylation and guides DNA methylation', *J Biol Chem*, 285: 26114-20.
- Dhillon, A. S., S. Hagan, O. Rath, and W. Kolch. 2007. 'MAP kinase signalling pathways in cancer', *Oncogene*, 26: 3279-90.
- Di Croce, L., V.A. Raker, M. Corsaro, F. Fazi, M. Fanelli, M. Faretta, F. Fuks, F. Lo Coco, T. Kouzarides, C. Nervi, S. Minucci, and P.G. Pelicci. 2002. 'Methyltransferase Recruitment and DNA Hypermethylation of Target Promoters by an Oncogenic Transcription Factor', *Science*, 295.
- Di Stefano, B., M. Ueda, S. Sabri, J. Brumbaugh, A. J. Huebner, A. Sahakyan, K. Clement, K. J. Clowers, A. R. Erickson, K. Shioda, S. P. Gygi, H. Gu, T. Shioda, A. Meissner, Y. Takashima, K. Plath, and K. Hochedlinger. 2018. 'Reduced MEK inhibition preserves genomic stability in naive human embryonic stem cells', *Nat Methods*.
- Diede, S. J., Z. Yao, C. C. Keyes, A. E. Tyler, J. Dey, C. S. Hackett, K. Elsaesser, C. J. Kemp, P. E. Neiman, W. A. Weiss, J. M. Olson, and S. J. Tapscott. 2013. 'Fundamental differences in promoter CpG island DNA hypermethylation between human cancer and genetically engineered mouse models of cancer', *Epigenetics*, 8: 1254-60.
- Ding, N., E. M. Bonham, B. E. Hannon, T. R. Amick, S. B. Baylin, and H. M. O'Hagan. 2015. 'Mismatch repair proteins recruit DNA methyltransferase 1 to sites of oxidative DNA damage', *J Mol Cell Biol*.
- Dobrovic, A., and D. Simpfendorfer. 1997. 'Methylation of the BRCA1 Gene in Sporadic Breast Cancer', *Cancer Res*, 57: 3347-50.
- Dodge, J. E., M. Okano, F. Dick, N. Tsujimoto, T. Chen, S. Wang, Y. Ueda, N. Dyson, and E. Li. 2005. 'Inactivation of Dnmt3b in mouse embryonic fibroblasts results in DNA hypomethylation, chromosomal instability, and spontaneous immortalization', *J Biol Chem*, 280: 17986-91.
- Doi, A., I. H. Park, B. Wen, P. Murakami, M. J. Aryee, R. Irizarry, B. Herb, C. Ladd-Acosta, J. Rho, S. Loewer, J. Miller, T. Schlaeger, G. Q. Daley, and A. P. Feinberg. 2009. 'Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts', *Nat Genet*, 41: 1350-3.
- Duggal, G., S. Warriar, S. Ghimire, D. Broekaert, M. Van der Jeught, S. Lierman, T. Deroo, L. Peelman, A. Van Soom, R. Cornelissen, B. Menten, P. Mestdagh, J. Vandesompele, M. Roost, R. C. Sliker, B. T. Heijmans, D. Deforce, P. De Sutter, S. C. De Sousa Lopes, and B. Heindryckx. 2015. 'Alternative Routes to Induce Naive Pluripotency in Human Embryonic Stem Cells', *Stem Cells*, 33: 2686-98.
- Duijf, P. H., N. Schultz, and R. Benezra. 2013. 'Cancer cells preferentially lose small chromosomes', *Int J Cancer*, 132: 2316-26.
- Duymich, C. E., J. Charlet, X. Yang, P. A. Jones, and G. Liang. 2016. 'DNMT3B isoforms without catalytic activity stimulate gene body methylation as accessory proteins in somatic cells', *Nat Commun*, 7: 11453.
- Easwaran, H., S. E. Johnstone, L. Van Neste, J. Ohm, T. Mosbrugger, Q. Wang, M. J. Aryee, P. Joyce, N. Ahuja, D. Weisenberger, E. Collisson, J. Zhu, S. Yegnasubramanian, W. Matsui, and S. B. Baylin. 2012. 'A DNA hypermethylation module for the stem/progenitor cell signature of cancer', *Genome Res*, 22: 837-49.
- Eckhardt, F., J. Lewin, R. Cortese, V. K. Rakyan, J. Attwood, M. Burger, J. Burton, T. V. Cox, R. Davies, T. A. Down, C. Haefliger, R. Horton, K. Howe, D. K. Jackson, J. Kunde, C. Koenig, J. Liddle, D. Niblett, T. Otto, R. Pettett, S. Seemann, C. Thompson, T. West, J. Rogers, A. Olek, K. Berlin, and S. Beck. 2006. 'DNA methylation profiling of human chromosomes 6, 20 and 22', *Nat Genet*, 38: 1378-85.



- Eden, A., F. Gaudet, A. Waghmare, and R. Jaenisch. 2003. 'Chromosomal Instability and Tumors Promoted by DNA Hypomethylation', *Science*, 300.
- Ehrlich, M. 2002. 'DNA methylation in cancer: too much, but also too little', *Oncogene*, 21: 5400-13.
- . 2009. 'DNA hypomethylation in cancer cells', *Epigenomics*, 1: 239-59.
- Ehrlich, M., MA. Gama-Sosa, L-H. Huang, RM. Midgett, KC. Kuo, RA. McCune, and C. Gehrke. 1982. 'Amount and distribution of 5-methylcytosine in human DNA from different types of tissues or cells', *Nucleic Acids Res*, 10.
- Ehrlich, Melanie. 2003. 'The ICF syndrome, a DNA methyltransferase 3B deficiency and immunodeficiency disease', *Clinical Immunology*, 109: 17-28.
- Eminli, S., A. Foudi, M. Stadtfeld, N. Maherali, T. Ahfeldt, G. Mostoslavsky, H. Hock, and K. Hochedlinger. 2009. 'Differentiation stage determines potential of hematopoietic cells for reprogramming into induced pluripotent stem cells', *Nat Genet*, 41: 968-76.
- Epsztejn-Litman, S., N. Feldman, M. Abu-Remaileh, Y. Shufaro, A. Gerson, J. Ueda, R. Deplus, F. Fuks, Y. Shinkai, H. Cedar, and Y. Bergman. 2008. 'De novo DNA methylation promoted by G9a prevents reprogramming of embryonically silenced genes', *Nat Struct Mol Biol*, 15: 1176-83.
- Ernst, J., and M. Kellis. 2017. 'Chromatin-state discovery and genome annotation with ChromHMM', *Nat Protoc*, 12: 2478-92.
- Estecio, M. R., J. Gallegos, C. Vallot, R. J. Castoro, W. Chung, S. Maegawa, Y. Oki, Y. Kondo, J. Jelinek, L. Shen, H. Hartung, P. D. Aplan, B. A. Czerniak, S. Liang, and J. P. Issa. 2010. 'Genome architecture marked by retrotransposons modulates predisposition to DNA methylation in cancer', *Genome Res*, 20: 1369-82.
- Esteller, M., P.G. Corn, S.B. Baylin, and J.G. Herman. 2001. 'A Gene Hypermethylation Profile of Human Cancer', *Cancer Research*, 61: 3225-29.
- Esteve, P. O., H. G. Chin, A. Smallwood, G. R. Feehery, O. Gangisetty, A. R. Karpf, M. F. Carey, and S. Pradhan. 2006. 'Direct interaction between DNMT1 and G9a coordinates DNA and histone methylation during replication', *Genes Dev*, 20: 3089-103.
- Fan, S., F. Fang, X. Zhang, and M. Q. Zhang. 2007. 'Putative zinc finger protein binding sites are over-represented in the boundaries of methylation-resistant CpG islands in the human genome', *PLoS One*, 2: e1184.
- Fatemi, Mehrnaz, Andrea Hermann, Humaira Gowher, and Albert Jeltsch. 2002. 'Dnmt3a and Dnmt1 functionally cooperate during de novo methylation of DNA', *European Journal of Biochemistry*, 269: 4981-84.
- Feinberg, A. P., M. A. Koldobskiy, and A. Gondor. 2016. 'Epigenetic modulators, modifiers and mediators in cancer aetiology and progression', *Nat Rev Genet*, 17: 284-99.
- Feinberg, A. P., R. Ohlsson, and S. Henikoff. 2006. 'The epigenetic progenitor origin of human cancer', *Nat Rev Genet*, 7: 21-33.
- Feldman, N., A. Gerson, J. Fang, E. Li, Y. Zhang, Y. Shinkai, H. Cedar, and Y. Bergman. 2006. 'G9a-mediated irreversible epigenetic inactivation of Oct-3/4 during early embryogenesis', *Nat Cell Biol*, 8: 188-94.
- Felle, M., S. Joppien, A. Nemeth, S. Diermeier, V. Thalhammer, T. Dobner, E. Kremmer, R. Kappler, and G. Langst. 2011. 'The USP7/Dnmt1 complex stimulates the DNA methylation activity of Dnmt1 and regulates the stability of UHRF1', *Nucleic Acids Res*, 39: 8355-65.
- Feltus, F. A., E. K. Lee, J. F. Costello, C. Plass, and P. M. Vertino. 2003. 'Predicting aberrant CpG island methylation', *Proc Natl Acad Sci U S A*, 100: 12253-8.
- . 2006. 'DNA motifs associated with aberrant CpG island methylation', *Genomics*, 87: 572-9.
- Ferrari, R., M/ Pellegrini, G.A. Horowitz, W. Xie, A.J. Berk, and S.K. Kurdistani. 2008. 'Epigenetic Reprogramming by Adenovirus e1a', *Science*, 321.

- Ficz, G. 2015. 'New insights into mechanisms that regulate DNA methylation patterning', *J Exp Biol*, 218: 14-20.
- Ficz, G., M. R. Branco, S. Seisenberger, F. Santos, F. Krueger, T. A. Hore, C. J. Marques, S. Andrews, and W. Reik. 2011. 'Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation', *Nature*, 473: 398-402.
- Ficz, G., and J. G. Gribben. 2014. 'Loss of 5-hydroxymethylcytosine in cancer: cause or consequence?', *Genomics*, 104: 352-7.
- Ficz, G., T. A. Hore, F. Santos, H. J. Lee, W. Dean, J. Arand, F. Krueger, D. Oxley, Y. L. Paul, J. Walter, S. J. Cook, S. Andrews, M. R. Branco, and W. Reik. 2013. 'FGF signaling inhibition in ESCs drives rapid genome-wide demethylation to the epigenetic ground state of pluripotency', *Cell Stem Cell*, 13: 351-9.
- Filtz, T. M., W. K. Vogel, and M. Leid. 2014. 'Regulation of transcription factor activity by interconnected post-translational modifications', *Trends Pharmacol Sci*, 35: 76-85.
- Flavahan, W. A., E. Gaskell, and B. E. Bernstein. 2017. 'Epigenetic plasticity and the hallmarks of cancer', *Science*, 357.
- Fox, E. J., K. S. Reid-Bayliss, M. J. Emond, and L. A. Loeb. 2014. 'Accuracy of Next Generation Sequencing Platforms', *Next Gener Seq Appl*, 1.
- Friedmann-Morvinski, D., and I. M. Verma. 2014. 'Dedifferentiation and reprogramming: origins of cancer stem cells', *EMBO Rep*, 15: 244-53.
- Fu, X., K. Cui, Q. Yi, L. Yu, and Y. Xu. 2017. 'DNA repair mechanisms in embryonic stem cells', *Cell Mol Life Sci*, 74: 487-93.
- Fuhrmann, G., A.C.K. Chung, K.J. Jackson, G. Hummelke, A. Baniahmad, J. Sutter, I. Sylvester, H.R. Schöler, and A.J. Cooney. 2001. 'Mouse Germline Restriction of Oct4 Expression by Germ Cell Nuclear Factor', *Developmental Cell*, 1: 377-87.
- Fuks, F., W.A. Burgers, N. Godin, M. Kasai, and T. Kouzarides. 2001. 'DNMT3A binds deacetylases and is recruited by a sequence-specific repressor to silence transcription', *The EMBO Journal*, 20: 2536-44.
- Fuks, F., P. J. Hurd, D. Wolf, X. Nan, A. P. Bird, and T. Kouzarides. 2003. 'The methyl-CpG-binding protein MeCP2 links DNA methylation to histone methylation', *J Biol Chem*, 278: 4035-40.
- Fuks, F., P.J. Hurd, R. Deplus, and T. Kouzarides. 2003. 'The DNA methyltransferases associate with HP1 and the SUV39H1 histone methyltransferase', *Nucleic Acids Res*, 31: 2305-12.
- Gafni, O., L. Weinberger, A. A. Mansour, Y. S. Manor, E. Chomsky, D. Ben-Yosef, Y. Kalma, S. Viukov, I. Maza, A. Zviran, Y. Rais, Z. Shipony, Z. Mukamel, V. Krupalnik, M. Zerbib, S. Geula, I. Caspi, D. Schneir, T. Shwartz, S. Gilad, D. Amann-Zalcenstein, S. Benjamin, I. Amit, A. Tanay, R. Massarwa, N. Novershtern, and J. H. Hanna. 2013. 'Derivation of novel human ground state naive pluripotent stem cells', *Nature*, 504: 282-6.
- Gao, F., G. Ji, Z. Gao, X. Han, M. Ye, Z. Yuan, H. Luo, X. Huang, K. Natarajan, J. Wang, H. Yang, and X. Zhang. 2014. 'Direct ChIP-bisulfite sequencing reveals a role of H3K27me3 mediating aberrant hypermethylation of promoter CpG islands in cancer cells', *Genomics*, 103: 204-10.
- Gardiner-Garden, M., and M. Frommer. 1987. 'CpG Islands in vertebrate genomes', *Journal of Molecular Biology*, 196: 261-82.
- Gaudet, F., J. G. Hodgson, A. Eden, L. Jackson-Grusby, J. Dausman, J. W. Gray, H. Leonhardt, and R. Jaenisch. 2003. 'Induction of Tumors in Mice by Genomic Hypomethylation', *J Biol Chem*, 280: 17986-91.
- Ge, Y. Z., M. T. Pu, H. Gowher, H. P. Wu, J. P. Ding, A. Jeltsch, and G. L. Xu. 2004. 'Chromatin targeting of de novo DNA methyltransferases by the PWWP domain', *J Biol Chem*, 279: 25447-54.
- Gebhard, C., C. Benner, M. Ehrich, L. Schwarzfischer, E. Schilling, M. Klug, W. Dietmaier, C. Thiede, E. Holler, R. Andreesen, and M. Rehli. 2010. 'General transcription factor

- binding at CpG islands in normal cells correlates with resistance to de novo DNA methylation in cancer cells', *Cancer Res*, 70: 1398-407.
- Genomes Project, Consortium, G. R. Abecasis, D. Altshuler, A. Auton, L. D. Brooks, R. M. Durbin, R. A. Gibbs, M. E. Hurles, and G. A. McVean. 2010. 'A map of human genome variation from population-scale sequencing', *Nature*, 467: 1061-73.
- Gidekel, S., and Y. Bergman. 2002. 'A unique developmental pattern of Oct-3/4 DNA methylation is controlled by a cis-demodification element', *J Biol Chem*, 277: 34521-30.
- Gidekel, S., G. Pizov, Y. Bergman, and E. Pikarsky. 2003. 'Oct-3/4 is a dose-dependent oncogenic fate determinant', *Cancer Cell*, 4.
- Gifford, C. A., M. J. Ziller, H. Gu, C. Trapnell, J. Donaghey, A. Tsankov, A. K. Shalek, D. R. Kelley, A. A. Shishkin, R. Issner, X. Zhang, M. Coyne, J. L. Fostel, L. Holmes, J. Meldrim, M. Guttman, C. Epstein, H. Park, O. Kohlbacher, J. Rinn, A. Gnirke, E. S. Lander, B. E. Bernstein, and A. Meissner. 2013. 'Transcriptional and epigenetic dynamics during specification of human embryonic stem cells', *Cell*, 153: 1149-63.
- Giono, L. E., and J. J. Manfredi. 2006. 'The p53 tumor suppressor participates in multiple cell cycle checkpoints', *J Cell Physiol*, 209: 13-20.
- Gisselsson, D. 2008. 'Classification of chromosome segregation errors in cancer', *Chromosoma*, 117: 511-9.
- Goll, M. G., and T. H. Bestor. 2005. 'Eukaryotic cytosine methyltransferases', *Annu Rev Biochem*, 74: 481-514.
- Gopalakrishnan, S., B. A. Sullivan, S. Trazzi, G. Della Valle, and K. D. Robertson. 2009. 'DNMT3B interacts with constitutive centromere protein CENP-C to modulate DNA methylation and the histone code at centromeric regions', *Hum Mol Genet*, 18: 3178-93.
- Gordon, C.A., S.R. Hartono, and F. Chedin. 2013. 'Inactive DNMT3B splice variants modulate de novo DNA methylation', *PLoS One*, 8.
- Gordon, D. J., B. Resio, and D. Pellman. 2012. 'Causes and consequences of aneuploidy in cancer', *Nat Rev Genet*, 13: 189-203.
- Gore, A., Z. Li, H. L. Fung, J. E. Young, S. Agarwal, J. Antosiewicz-Bourget, I. Canto, A. Giorgetti, M. A. Israel, E. Kiskinis, J. H. Lee, Y. H. Loh, P. D. Manos, N. Montserrat, A. D. Panopoulos, S. Ruiz, M. L. Wilbert, J. Yu, E. F. Kirkness, J. C. Izpisua Belmonte, D. J. Rossi, J. A. Thomson, K. Eggan, G. Q. Daley, L. S. Goldstein, and K. Zhang. 2011. 'Somatic coding mutations in human induced pluripotent stem cells', *Nature*, 471: 63-7.
- Greenblatt, M.S., W.P. Bennett, M. Hollstein, and C.C. Harris. 1994. 'Mutations in the p53 Tumor Suppressor Gene: Clues to Cancer Etiology and Molecular Pathogenesis', *Cancer Res*, 54: 4855-78.
- Guenatri, M., D. Bailly, C. Maison, and G. Almouzni. 2004. 'Mouse centric and pericentric satellite repeats form distinct functional heterochromatin', *J Cell Biol*, 166: 493-505.
- Guenther, M. G., S. S. Levine, L. A. Boyer, R. Jaenisch, and R. A. Young. 2007. 'A chromatin landmark and transcription initiation at most promoters in human cells', *Cell*, 130: 77-88.
- Guo, G., F. von Meyenn, M. Rostovskaya, J. Clarke, S. Dietmann, D. Baker, A. Sahakyan, S. Myers, P. Bertone, W. Reik, K. Plath, and A. Smith. 2017. 'Epigenetic resetting of human pluripotency', *Development*, 144: 2748-63.
- Guo, G., F. von Meyenn, F. Santos, Y. Chen, W. Reik, P. Bertone, A. Smith, and J. Nichols. 2016. 'Naive Pluripotent Stem Cells Derived Directly from Isolated Cells of the Human Inner Cell Mass', *Stem Cell Reports*, 6: 437-46.
- Guo, H., P. Zhu, L. Yan, R. Li, B. Hu, Y. Lian, J. Yan, X. Ren, S. Lin, J. Li, X. Jin, X. Shi, P. Liu, X. Wang, W. Wang, Y. Wei, X. Li, F. Guo, X. Wu, X. Fan, J. Yong, L. Wen, S. X. Xie, F. Tang, and J. Qiao. 2014. 'The DNA methylation landscape of human early embryos', *Nature*, 511: 606-10.

- Guo, J. U., Y. Su, C. Zhong, G. L. Ming, and H. Song. 2011. 'Hydroxylation of 5-methylcytosine by TET1 promotes active DNA demethylation in the adult brain', *Cell*, 145: 423-34.
- Guo, X., L. Wang, J. Li, Z. Ding, J. Xiao, X. Yin, S. He, P. Shi, L. Dong, G. Li, C. Tian, J. Wang, Y. Cong, and Y. Xu. 2015. 'Structural insight into autoinhibition and histone H3-induced activation of DNMT3A', *Nature*, 517: 640-4.
- Habibi, E., A. B. Brinkman, J. Arand, L. I. Kroeze, H. H. Kerstens, F. Matarese, K. Lepikhov, M. Gut, I. Brun-Heath, N. C. Hubner, R. Benedetti, L. Altucci, J. H. Jansen, J. Walter, I. G. Gut, H. Marks, and H. G. Stunnenberg. 2013. 'Whole-genome bisulfite sequencing of two distinct interconvertible DNA methylomes of mouse embryonic stem cells', *Cell Stem Cell*, 13: 360-9.
- Hackett, J. A., and M. A. Surani. 2014. 'Regulatory principles of pluripotency: from the ground state up', *Cell Stem Cell*, 15: 416-30.
- Hagarman, J. A., M. P. Motley, K. Kristjansdottir, and P. D. Soloway. 2013. 'Coordinate regulation of DNA methylation and H3K27me3 in mouse embryonic stem cells', *PLoS One*, 8: e53880.
- Hanahan, D., and R. A. Weinberg. 2011. 'Hallmarks of cancer: the next generation', *Cell*, 144: 646-74.
- Handa, V., and A. Jeltsch. 2005. 'Profound flanking sequence preference of Dnmt3a and Dnmt3b mammalian DNA methyltransferases shape the human epigenome', *J Mol Biol*, 348: 1103-12.
- Hanley, M. P., M. A. Hahn, A. X. Li, X. Wu, J. Lin, J. Wang, A. H. Choi, Z. Ouyang, Y. Fong, G. P. Pfeifer, T. J. Devers, and D. W. Rosenberg. 2017. 'Genome-wide DNA methylation profiling reveals cancer-associated changes within early colonic neoplasia', *Oncogene*, 36: 5035-44.
- Hanna, C. W., H. Demond, and G. Kelsey. 2018. 'Epigenetic regulation in development: is the mouse a good model for the human?', *Hum Reprod Update*.
- Hanna, J., A. W. Cheng, K. Saha, J. Kim, C. J. Lengner, F. Soldner, J. P. Cassady, J. Muffat, B. W. Carey, and R. Jaenisch. 2010. 'Human embryonic stem cells with biological and epigenetic characteristics similar to those of mouse ESCs', *PNAS*, 107: 9222-27.
- Hansen, K. D., W. Timp, H. C. Bravo, S. Sabunciyan, B. Langmead, O. G. McDonald, B. Wen, H. Wu, Y. Liu, D. Diep, E. Briem, K. Zhang, R. A. Irizarry, and A. P. Feinberg. 2011. 'Increased methylation variation in epigenetic domains across cancer types', *Nat Genet*, 43: 768-75.
- Harrigan, J. A., R. Belotserkovskaya, J. Coates, D. S. Dimitrova, S. E. Polo, C. R. Bradshaw, P. Fraser, and S. P. Jackson. 2011. 'Replication stress induces 53BP1-containing OPT domains in G1 cells', *J Cell Biol*, 193: 97-108.
- Hendrich, B., U. Hardeland, H. Ng, J. Jiricny, and A. Bird. 1999. 'The thymine glycosylase MBD4 can bind to the product of deamination at methylated CpG sites', *Nature*, 401.
- Hepburn, A. C., R. E. Steele, R. Veeratterapillay, L. Wilson, E. E. Kounatidou, A. Barnard, P. Berry, J. R. Cassidy, M. Moad, A. El-Sherif, L. Gaughan, I. G. Mills, C. N. Robson, and R. Heer. 2019. 'The induction of core pluripotency master regulators in cancers defines poor clinical outcomes and treatment resistance', *Oncogene*.
- Herman, J. G., A. Umar, K. Polyak, J.R. Graff, N. Ahuja, J. P. Issa, S. D. Markowitz, J.K.V. Willson, S.R. Hamilton, K. W. Kinzler, M.F. Kane, R.D. Kolodner, B. Vogelstein, T.A. Kunkel, and S. B. Baylin. 1998. 'Incidence and functional consequences of hMLH1 promoter hypermethylation in colorectal carcinoma', *PNAS*, 95: 6870-75.
- Hermann, A., S. Schmitt, and A. Jeltsch. 2003. 'The human Dnmt2 has residual DNA-(cytosine-C5) methyltransferase activity', *J Biol Chem*, 278: 31717-21.
- Hervouet, E., L. Lalier, E. Debien, M. Cheray, A. Geairon, H. Rogniaux, D. Loussouarn, S. A. Martin, F. M. Vallette, and P. F. Cartron. 2010. 'Disruption of Dnmt1/PCNA/UHRF1

- interactions promotes tumorigenesis from human and mice glial cells', *PLoS One*, 5: e11333.
- Hervouet, E., F. M. Vallette, and P. F. Cartron. 2010. 'Dnmt1/Transcription factor interactions: an alternative mechanism of DNA methylation inheritance', *Genes Cancer*, 1: 434-43.
- Hervouet, Eric, François M. Vallette, and Pierre-François Cartron. 2014. 'Dnmt3/transcription factor interactions as crucial players in targeted DNA methylation', *Epigenetics*, 4: 487-99.
- Heyn, P., C. V. Logan, A. Fluteau, R. C. Challis, T. Auchynnikava, C. A. Martin, J. A. Marsh, F. Taglini, F. Kilanowski, D. A. Parry, V. Cormier-Daire, C. T. Fong, K. Gibson, V. Hwa, L. Ibanez, S. P. Robertson, G. Sebastiani, J. Rappsilber, R. C. Allshire, M. A. M. Reijns, A. Dauber, D. Sproul, and A. P. Jackson. 2018. 'Gain-of-function DNMT3A mutations cause microcephalic dwarfism and hypermethylation of Polycomb-regulated regions', *Nat Genet*.
- Ho, L., R. Jothi, J. L. Ronan, K. Cui, K. Zhao, and G. R. Crabtree. 2009. 'An embryonic stem cell chromatin remodeling complex, esBAF, is an essential component of the core pluripotency transcriptional network', *Proc Natl Acad Sci U S A*, 106: 5187-91.
- Hochedlinger, K., and R. Jaenisch. 2015. 'Induced Pluripotency and Epigenetic Reprogramming', *Cold Spring Harb Perspect Biol*, 7.
- Holm, T. M., L. Jackson-Grusby, T. Brambrink, Y. Yamada, W. M. Rideout, 3rd, and R. Jaenisch. 2005. 'Global loss of imprinting leads to widespread tumorigenesis in adult mice', *Cancer Cell*, 8: 275-85.
- Howell, CY., TH. Bestor, F. Ding, KE. Latham, C. Mertineit, JM. Trasler, and JR. Chaillet. 2001. 'Genomic Imprinting Disrupted by a Maternal Effect Mutation in the Dnmt1 Gene', *Cell*, 104: 829-38.
- Huang, D.W., B. T. Sherman, and R. A. Lempicki. 2009. 'Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists', *Nucleic Acids Res*, 37: 1-13.
- Huang, Da Wei, Brad T. Sherman, and Richard A. Lempicki. 2008. 'Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources', *Nature Protocols*, 4: 44.
- Huang, K., T. Maruyama, and G. Fan. 2014. 'The naive state of human pluripotent stem cells: a synthesis of stem cell and preimplantation embryo transcriptome analyses', *Cell Stem Cell*, 15: 410-15.
- Hussein, S. M., N. N. Batada, S. Vuoristo, R. W. Ching, R. Autio, E. Narva, S. Ng, M. Sourour, R. Hamalainen, C. Olsson, K. Lundin, M. Mikkola, R. Trokovic, M. Peitz, O. Brustle, D. P. Bazett-Jones, K. Alitalo, R. Lahesmaa, A. Nagy, and T. Otonkoski. 2011. 'Copy number variation and selection during reprogramming to pluripotency', *Nature*, 471: 58-62.
- Hyun, K., J. Jeon, K. Park, and J. Kim. 2017. 'Writing, erasing and reading histone lysine methylations', *Exp Mol Med*, 49: e324.
- Illingworth, R., A. Kerr, D. Desousa, H. Jorgensen, P. Ellis, J. Stalker, D. Jackson, C. Clee, R. Plumb, J. Rogers, S. Humphray, T. Cox, C. Langford, and A. Bird. 2008. 'A novel CpG island set identifies tissue-specific methylation at developmental gene loci', *PLoS Biol*, 6: e22.
- Illingworth, R. S., U. Gruenewald-Schneider, S. Webb, A. R. Kerr, K. D. James, D. J. Turner, C. Smith, D. J. Harrison, R. Andrews, and A. P. Bird. 2010. 'Orphan CpG islands identify numerous conserved promoters in the mammalian genome', *PLoS Genet*, 6: e1001134.
- Irizarry, R. A., C. Ladd-Acosta, B. Wen, Z. Wu, C. Montano, P. Onyango, H. Cui, K. Gabo, M. Rongione, M. Webster, H. Ji, J. Potash, S. Sabunciyan, and A. P. Feinberg. 2009. 'The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores', *Nat Genet*, 41: 178-86.

- Ito, S., L. Shen, Q. Dai, S.C. Wu, L.B. Collins, J.A. Swenberg, C. He, and Y. Zhang. 2011. 'Tet Proteins Can Convert 5-Methylcytosine to 5-Formylcytosine and 5-Carboxylcytosine', *Science*, 333.
- Iurlaro, M., G. Ficuz, D. Oxley, E. Raiber, M. Bachman, M.J. Booth, S. Andrews, S. Balasubramanian, and W. Reik. 2013. 'A screen for hydroxymethylcytosine and formylcytosine binding proteins suggests functions in transcription and chromatin regulation', *Genome Biology*, 14.
- Jair, K. W., K. E. Bachman, H. Suzuki, A. H. Ting, I. Rhee, R. W. Yen, S. B. Baylin, and K. E. Schuebel. 2006. 'De novo CpG island methylation in human cancer cells', *Cancer Res*, 66: 682-92.
- Jelnic, P., and P. Shaw. 2007. 'Loss of imprinting and cancer', *J Pathol*, 211: 261-8.
- Jeltsch, A., and R. Z. Jurkowska. 2016. 'Allosteric control of mammalian DNA methyltransferases - a new regulatory paradigm', *Nucleic Acids Res*, 44: 8556-75.
- Jeter, C. R., T. Yang, J. Wang, H. P. Chao, and D. G. Tang. 2015. 'Concise Review: NANOG in Cancer Stem Cells and Tumor Development: An Update and Outstanding Questions', *Stem Cells*, 33: 2381-90.
- Ji, J., S. H. Ng, V. Sharma, D. Neculai, S. Hussein, M. Sam, Q. Trinh, G. M. Church, J. D. McPherson, A. Nagy, and N. N. Batada. 2012. 'Elevated coding mutation rate during the reprogramming of human somatic cells into induced pluripotent stem cells', *Stem Cells*, 30: 435-40.
- Ji, J., Y. Zhang, C. E. Redon, W. C. Reinhold, A. P. Chen, L. K. Fogli, S. L. Holbeck, R. E. Parchment, M. Hollingshead, J. E. Tomaszewski, Q. Dudon, Y. Pommier, J. H. Doroshow, and W. M. Bonner. 2017. 'Phosphorylated fraction of H2AX as a measurement for DNA damage in cancer cells and potential applications of a novel assay', *PLoS One*, 12: e0171582.
- Jin, R. U., and J. C. Mills. 2019. 'The cyclical hit model: how paligenosis might establish the mutational landscape in Barrett's esophagus and esophageal adenocarcinoma', *Curr Opin Gastroenterol*.
- Jones, P. A. 2012. 'Functions of DNA methylation: islands, start sites, gene bodies and beyond', *Nat Rev Genet*, 13: 484-92.
- Jones, P. A., and S. B. Baylin. 2002. 'The fundamental role of epigenetic events in cancer', *Nat Rev Genet*, 3: 415-28.
- . 2007. 'The epigenomics of cancer', *Cell*, 128: 683-92.
- Jones, P.A., and S.M. Taylor. 1980. 'Cellular Differentiation, Cytidine Analogs and DNA Methylation', *Cell*, 20.
- Jones, P.L. , G.J.C. Veenstra, P.A. Wade, D. Vermaak, S.U. Kass, N. Landsberger, J. Strouboulis, and A.P. Wolffe. 1998. 'Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription', *Nature Genetics*, 19.
- Jung, L., P. Tropel, Y. Moal, M. Teletin, E. Jeandidier, R. Gayon, C. Himmelspach, F. Bello, C. Andre, A. Tosch, A. Mansouri, C. Bruant-Rodier, P. Bouille, and S. Viville. 2014. 'ONSL and OSKM cocktails act synergistically in reprogramming human somatic cells into induced pluripotent stem cells', *Mol Hum Reprod*, 20: 538-49.
- Jurkowska, R. Z., T. P. Jurkowski, and A. Jeltsch. 2011. 'Structure and function of mammalian DNA methyltransferases', *Chembiochem*, 12: 206-22.
- Kafer, G. R., X. Li, T. Horii, I. Suetake, S. Tajima, I. Hatada, and P. M. Carlton. 2016. '5-Hydroxymethylcytosine Marks Sites of DNA Damage and Promotes Genome Stability', *Cell Rep*, 14: 1283-92.
- Kanai, Yae, Saori Ushijima, Yukihiro Nakanishi, Michiie Sakamoto, and Setsuo Hirohashi. 2003. 'Mutation of the DNA methyltransferase (DNMT) 1 gene in human colorectal cancers', *Cancer Letters*, 192: 75-82.

- Kane, M.F., M. Loda, G.M. Gaida, J. Lipman, R. K. Mishra, H. Goldman, J.M. Jessup, and R. Kolodner. 1997. 'Methylation of the hMLH1 Promoter Correlates with Lack of Expression of hMLH1 in Sporadic Colon Tumors and Mismatch Repair-defective Human Tumor Cell Lines', *Cancer Res*, 57: 808-11.
- Karpf, A. R., and S. Matsui. 2005. 'Genetic disruption of cytosine DNA methyltransferase enzymes induces chromosomal instability in human cancer cells', *Cancer Res*, 65: 8635-9.
- Keshet, I., Y. Schlesinger, S. Farkash, E. Rand, M. Hecht, E. Segal, E. Pikarski, R. A. Young, A. Niveleau, H. Cedar, and I. Simon. 2006. 'Evidence for an instructive mechanism of de novo methylation in cancer cells', *Nat Genet*, 38: 149-53.
- Kilens, S., D. Meistermann, D. Moreno, C. Chariou, A. Gaignerie, A. Reignier, Y. Lelievre, M. Casanova, C. Vallot, S. Nedellec, L. Flippe, J. Firmin, J. Song, E. Charpentier, J. Lammers, A. Donnart, N. Marec, W. Deb, A. Bihouee, C. Le Caignec, C. Pecqueur, R. Redon, P. Barriere, J. Bourdon, V. Pasque, M. Soumillon, T. S. Mikkelsen, C. Rougeulle, T. Freour, L. David, and Consortium Milieu Interieur. 2018. 'Parallel derivation of isogenic human primed and naive induced pluripotent stem cells', *Nat Commun*, 9: 360.
- Kim, J.H., A. Karnovsky, V. Mahavisno, T. Weymouth, M. Pande, D.C. Dolinoy, L.S. Rozek, and M.A. Sartor. 2012. 'LRpath analysis reveals common pathways dysregulated via DNA methylation across cancer types', *BMC Genomics*, 13.
- Kim, S., M. Li, H. Paik, K. Nephew, H. Shi, R. Kramer, D. Xu, and T. H. Huang. 2008. 'Predicting DNA methylation susceptibility using CpG flanking sequences', *Pac Symp Biocomput*: 315-26.
- Koche, R. P., Z. D. Smith, M. Adli, H. Gu, M. Ku, A. Gnirke, B. E. Bernstein, and A. Meissner. 2011. 'Reprogramming factor expression initiates widespread targeted chromatin remodeling', *Cell Stem Cell*, 8: 96-105.
- Kops, G. J., B. A. Weaver, and D. W. Cleveland. 2005. 'On the road to cancer: aneuploidy and the mitotic checkpoint', *Nat Rev Cancer*, 5: 773-85.
- Kouzarides, T. 2007. 'Chromatin modifications and their function', *Cell*, 128: 693-705.
- Krebs, A. R., S. Dessus-Babus, L. Burger, and D. Schubeler. 2014. 'High-throughput engineering of a mammalian genome reveals building principles of methylation states at CG rich regions', *Elife*, 3: e04094.
- Kriaucionis, S. , and N. Heintz. 2009. 'The Nuclear DNA Base 5-Hydroxymethylcytosine Is Present in Purkinje Neurons and the Brain', *Science*, 324.
- Krueger, F., and S. R. Andrews. 2011. 'Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications', *Bioinformatics*, 27: 1571-2.
- Kubbutat, M.H.G., S.N. Jones, and K.H. Vousden. 1997. 'Regulation of p53 stability by Mdm2', *Nature*, 387.
- Kushwaha, G., M. Dozmorov, J. D. Wren, J. Qiu, H. Shi, and D. Xu. 2016. 'Hypomethylation coordinates antagonistically with hypermethylation in cancer development: a case study of leukemia', *Hum Genomics*, 10 Suppl 2: 18.
- Langemeijer, S. M., R. P. Kuiper, M. Berends, R. Knops, M. G. Aslanyan, M. Massop, E. Stevens-Linders, P. van Hoogen, A. G. van Kessel, R. A. Raymakers, E. J. Kamping, G. E. Verhoef, E. Verburgh, A. Hagemeijer, P. Vandenbergh, T. de Witte, B. A. van der Reijden, and J. H. Jansen. 2009. 'Acquired mutations in TET2 are common in myelodysplastic syndromes', *Nat Genet*, 41: 838-42.
- Lecona, E., and O. Fernandez-Capetillo. 2014. 'Replication stress and cancer: it takes two to tango', *Exp Cell Res*, 329: 26-34.
- Lehnertz, Bernhard, Yoshihide Ueda, Alwin A. H. A. Derijck, Ulrich Braunschweig, Laura Perez-Burgos, Stefan Kubicek, Taiping Chen, En Li, Thomas Jenuwein, and Antoine H. F. M. Peters. 2003. 'Suv39h-Mediated Histone H3 Lysine 9 Methylation Directs DNA

- Methylation to Major Satellite Repeats at Pericentric Heterochromatin', *Current Biology*, 13: 1192-200.
- 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: 865-73.
- Ley, Timothy J., Li Ding, Matthew J. Walter, Michael D. McLellan, Tamara Lamprecht, David E. Larson, Cyriac Kandoth, Jacqueline E. Payton, Jack Baty, John Welch, Christopher C. Harris, Cheryl F. Lichti, R. Reid Townsend, Robert S. Fulton, David J. Dooling, Daniel C. Koboldt, Heather Schmidt, Qunyan Zhang, John R. Osborne, Ling Lin, Michelle O'Laughlin, Joshua F. McMichael, Kim D. Delehaunty, Sean D. McGrath, Lucinda A. Fulton, Vincent J. Magrini, Tammi L. Vickery, Jasreet Hundal, Lisa L. Cook, Joshua J. Conyers, Gary W. Swift, Jerry P. Reed, Patricia A. Alldredge, Todd Wylie, Jason Walker, Joelle Kalicki, Mark A. Watson, Sharon Heath, William D. Shannon, Nobish Varghese, Rakesh Nagarajan, Peter Westervelt, Michael H. Tomasson, Daniel C. Link, Timothy A. Graubert, John F. DiPersio, Elaine R. Mardis, and Richard K. Wilson. 2010. 'DNMT3A Mutations in Acute Myeloid Leukemia', *New England Journal of Medicine*, 363: 2424-33.
- Li, E., and Y. Zhang. 2014. 'DNA methylation in mammals', *Cold Spring Harb Perspect Biol*, 6: a019133.
- Li, En., C. Beard, and R. Jaenisch. 1993. 'Role for DNA methylation in genomic imprinting', *Nature*, 366: 362-265.
- Li, En., TH. Bestor, and R. Jaenisch. 1992. 'Targeted Mutation of the DNA Methyltransferase Gene Results in Embryonic Lethality', *Cell*, 69: 915-26.
- Li, H., T. Rauch, Z. X. Chen, P. E. Szabo, A. D. Riggs, and G. P. Pfeifer. 2006. 'The histone methyltransferase SETDB1 and the DNA methyltransferase DNMT3A interact directly and localize to promoters silenced in cancer cells', *J Biol Chem*, 281: 19489-500.
- Liao, H. F., K. Y. Tai, W. S. Chen, L. C. Cheng, H. N. Ho, and S. P. Lin. 2012. 'Functions of DNA methyltransferase 3-like in germ cells and beyond', *Biol Cell*, 104: 571-87.
- Liao, J., R. Karnik, H. Gu, M. J. Ziller, K. Clement, A. M. Tsankov, V. Akopian, C. A. Gifford, J. Donaghey, C. Galonska, R. Pop, D. Reyon, S. Q. Tsai, W. Mallard, J. K. Joung, J. L. Rinn, A. Gnirke, and A. Meissner. 2015. 'Targeted disruption of DNMT1, DNMT3A and DNMT3B in human embryonic stem cells', *Nat Genet*, 47: 469-78.
- Liao, Y., G. K. Smyth, and W. Shi. 2014. 'featureCounts: an efficient general purpose program for assigning sequence reads to genomic features', *Bioinformatics*, 30: 923-30.
- Lienert, F., C. Wirbelauer, I. Som, A. Dean, F. Mohn, and D. Schubeler. 2011. 'Identification of genetic elements that autonomously determine DNA methylation states', *Nat Genet*, 43: 1091-7.
- Lin, C. Y., J. Loven, P. B. Rahl, R. M. Paranal, C. B. Burge, J. E. Bradner, T. I. Lee, and R. A. Young. 2012. 'Transcriptional amplification in tumor cells with elevated c-Myc', *Cell*, 151: 56-67.
- Lin, I. G., L. Han, A. Taghva, L. E. O'Brien, and C. L. Hsieh. 2002. 'Murine De Novo Methyltransferase Dnmt3a Demonstrates Strand Asymmetry and Site Preference in the Methylation of DNA In Vitro', *Molecular and Cellular Biology*, 22: 704-23.
- Linhart, H. G., H. Lin, Y. Yamada, E. Moran, E. J. Steine, S. Gokhale, G. Lo, E. Cantu, M. Ehrich, T. He, A. Meissner, and R. Jaenisch. 2007. 'Dnmt3b promotes tumorigenesis in vivo by gene-specific de novo methylation and transcriptional silencing', *Genes Dev*, 21: 3110-22.
- Liu, X., Q. Gao, P. Li, Q. Zhao, J. Zhang, J. Li, H. Koseki, and J. Wong. 2013. 'UHRF1 targets DNMT1 for DNA methylation through cooperative binding of hemi-methylated DNA and methylated H3K9', *Nat Commun*, 4: 1563.
- Liu, X., C. M. Nefzger, F. J. Rossello, J. Chen, A. S. Knaupp, J. Firas, E. Ford, J. Pflueger, J. M. Paynter, H. S. Chy, C. M. O'Brien, C. Huang, K. Mishra, M. Hodgson-Garms, N. Jansz, S.



- M. Williams, M. E. Blewitt, S. K. Nilsson, R. B. Schittenhelm, A. L. Laslett, R. Lister, and J. M. Polo. 2017. 'Comprehensive characterization of distinct states of human naive pluripotency generated by reprogramming', *Nat Methods*.
- Liu, Y., C. F. Nielsen, Q. Yao, and I. D. Hickson. 2014. 'The origins and processing of ultra fine anaphase DNA bridges', *Curr Opin Genet Dev*, 26: 1-5.
- Llinas-Arias, P., and M. Esteller. 2017. 'Epigenetic inactivation of tumour suppressor coding and non-coding genes in human cancer: an update', *Open Biol*, 7.
- Loh, Y. H., W. Zhang, X. Chen, J. George, and H. H. Ng. 2007. 'Jmjd1a and Jmjd2c histone H3 Lys 9 demethylases regulate self-renewal in embryonic stem cells', *Genes Dev*, 21: 2545-57.
- Long, H. K., N. P. Blackledge, and R. J. Klose. 2013. 'ZF-CxxC domain-containing proteins, CpG islands and the chromatin connection', *Biochem Soc Trans*, 41: 727-40.
- Long, H. K., H. W. King, R. K. Patient, D. T. Odom, and R. J. Klose. 2016. 'Protection of CpG islands from DNA methylation is DNA-encoded and evolutionarily conserved', *Nucleic Acids Res*, 44: 6693-706.
- Lorsbach, R. B., J. Moore, S. Mathew, S. C. Raimondi, S. T. Mukatira, and J. R. Downing. 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: 637-41.
- Lukas, C., V. Savic, S. Bekker-Jensen, C. Doil, B. Neumann, R. S. Pedersen, M. Grofte, K. L. Chan, I. D. Hickson, J. Bartek, and J. Lukas. 2011. '53BP1 nuclear bodies form around DNA lesions generated by mitotic transmission of chromosomes under replication stress', *Nat Cell Biol*, 13: 243-53.
- Macleod, D., J. Charlton, J. Mullins, and A.P. Bird. 1994. 'Spl sites in the mouse aprt gene promoter are required to prevent methylation of the CpG island', *Genes Dev*, 8: 2282-92.
- Maherali, N., R. Sridharan, W. Xie, J. Utikal, S. Eminli, K. Arnold, M. Stadtfeld, R. Yachechko, J. Tchieu, R. Jaenisch, K. Plath, and K. Hochedlinger. 2007. 'Directly reprogrammed fibroblasts show global epigenetic remodeling and widespread tissue contribution', *Cell Stem Cell*, 1: 55-70.
- Maiti, A., and A. C. Drohat. 2011. 'Thymine DNA glycosylase can rapidly excise 5-formylcytosine and 5-carboxylcytosine: potential implications for active demethylation of CpG sites', *J Biol Chem*, 286: 35334-8.
- Mantovani, F., L. Collavin, and G. Del Sal. 2019. 'Mutant p53 as a guardian of the cancer cell', *Cell Death Differ*, 26: 199-212.
- Manzo, M., J. Wirz, C. Ambrosi, R. Villasenor, B. Roschitzki, and T. Baubec. 2017. 'Isoform-specific localization of DNMT3A regulates DNA methylation fidelity at bivalent CpG islands', *EMBO J*, 36: 3421-34.
- Marion, R. M., K. Strati, H. Li, M. Murga, R. Blanco, S. Ortega, O. Fernandez-Capetillo, M. Serrano, and M. A. Blasco. 2009. 'A p53-mediated DNA damage response limits reprogramming to ensure iPS cell genomic integrity', *Nature*, 460: 1149-53.
- Maruyama, M., T. Ichisaka, M. Nakagawa, and S. Yamanaka. 2005. 'Differential roles for Sox15 and Sox2 in transcriptional control in mouse embryonic stem cells', *J Biol Chem*, 280: 24371-9.
- Maunakea, A. K., R. P. Nagarajan, M. Bilenky, T. J. Ballinger, C. D'Souza, S. D. Fouse, B. E. Johnson, C. Hong, C. Nielsen, Y. Zhao, G. Turecki, A. Delaney, R. Varhol, N. Thiessen, K. Shchors, V. M. Heine, D. H. Rowitch, X. Xing, C. Fiore, M. Schillebeeckx, S. J. Jones, D. Haussler, M. A. Marra, M. Hirst, T. Wang, and J. F. Costello. 2010. 'Conserved role of intragenic DNA methylation in regulating alternative promoters', *Nature*, 466: 253-7.
- Mayshar, Y., U. Ben-David, N. Lavon, J. C. Biancotti, B. Yakir, A. T. Clark, K. Plath, W. E. Lowry, and N. Benvenisty. 2010. 'Identification and classification of chromosomal aberrations in human induced pluripotent stem cells', *Cell Stem Cell*, 7: 521-31.

- McCabe, M. T., E. K. Lee, and P. M. Vertino. 2009. 'A multifactorial signature of DNA sequence and polycomb binding predicts aberrant CpG island methylation', *Cancer Res*, 69: 282-91.
- McGhee, J.D., and G.D. Ginder. 1979. 'Specific DNA methylation sites in the vicinity of the chicken B-globin genes', *Nature*, 280.
- Meehan, R.R., J.D. Lewis, S. McKay, E.L. Kleiner, and A.P. Bird. 1989. 'Identification of a Mammalian Protein That Specifically Binds to DNA Containing Methylated CpGs', *Cell*, 58.
- Meissner, A., T. S. Mikkelsen, H. Gu, M. Wernig, J. Hanna, A. Sivachenko, X. Zhang, B. E. Bernstein, C. Nusbaum, D. B. Jaffe, A. Gnirke, R. Jaenisch, and E. S. Lander. 2008. 'Genome-scale DNA methylation maps of pluripotent and differentiated cells', *Nature*, 454: 766-70.
- Meng, H., Y. Cao, J. Qin, X. Song, Q. Zhang, Y. Shi, and L. Cao. 2015. 'DNA methylation, its mediators and genome integrity', *Int J Biol Sci*, 11: 604-17.
- Merkle, F. T., S. Ghosh, N. Kamitaki, J. Mitchell, Y. Avior, C. Mello, S. Kashin, S. Mekhoubad, D. Ilic, M. Charlton, G. Saphier, R. E. Handsaker, G. Genovese, S. Bar, N. Benvenisty, S. A. McCarroll, and K. Eggen. 2017. 'Human pluripotent stem cells recurrently acquire and expand dominant negative P53 mutations', *Nature*, 545: 229-33.
- Merlo, A., J. G. Herman, L. Mao, D.J. Lee, E. Gabrielson, P.C. Burger, S. B. Baylin, and D. Sidransky. 1995. '5' CpG island methylation is associated with transcriptional silencing of the tumour suppressor p16/CDKN2/MTS1 in human cancers', *Nature Medicine*, 1.
- Mikkelsen, T. S., J. Hanna, X. Zhang, M. Ku, M. Wernig, P. Schorderet, B. E. Bernstein, R. Jaenisch, E. S. Lander, and A. Meissner. 2008. 'Dissecting direct reprogramming through integrative genomic analysis', *Nature*, 454: 49-55.
- Mikkelsen, T. S., M. Ku, D. B. Jaffe, B. Issac, E. Lieberman, G. Giannoukos, P. Alvarez, W. Brockman, T. K. Kim, R. P. Koche, W. Lee, E. Mendenhall, A. O'Donovan, A. Presser, C. Russ, X. Xie, A. Meissner, M. Wernig, R. Jaenisch, C. Nusbaum, E. S. Lander, and B. E. Bernstein. 2007. 'Genome-wide maps of chromatin state in pluripotent and lineage-committed cells', *Nature*, 448: 553-60.
- Milholland, B., X. Dong, L. Zhang, X. Hao, Y. Suh, and J. Vijg. 2017. 'Differences between germline and somatic mutation rates in humans and mice', *Nat Commun*, 8: 15183.
- Miyajima, A., M. Tanaka, and T. Itoh. 2014. 'Stem/progenitor cells in liver development, homeostasis, regeneration, and reprogramming', *Cell Stem Cell*, 14: 561-74.
- Mohandas, T., R.S. Sparkes, and L.J. Shapiro. 1981. 'Reactivation of an inactive human X chromosome: evidence for X inactivation by DNA methylation', *Science*, 211.
- Mohn, F., M. Weber, M. Rebhan, T. C. Roloff, J. Richter, M. B. Stadler, M. Bibel, and D. Schubeler. 2008. 'Lineage-specific polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors', *Mol Cell*, 30: 755-66.
- Monier, K., S. Mouradian, and K.F. Sullivan. 2006. 'DNA methylation promotes Aurora-B-driven phosphorylation of histone H3 in chromosomal subdomains', *Journal of Cell Science*, 120.
- Montserrat, N., E. Nivet, I. Sancho-Martinez, T. Hishida, S. Kumar, L. Miquel, C. Cortina, Y. Hishida, Y. Xia, C. R. Esteban, and J. C. Izpisua Belmonte. 2013. 'Reprogramming of human fibroblasts to pluripotency with lineage specifiers', *Cell Stem Cell*, 13: 341-50.
- Morel, A. P., C. Ginestier, R. M. Pommier, O. Cabaud, E. Ruiz, J. Wicinski, M. Devouassoux-Shisheboran, V. Combaret, P. Finetti, C. Chassot, C. Pinatel, F. Fauvet, P. Saintigny, E. Thomas, C. Moyret-Lalle, J. Lachuer, E. Despras, J. L. Jauffret, F. Bertucci, J. Guitton, A. Wierinckx, Q. Wang, N. Radosevic-Robin, F. Penault-Llorca, D. G. Cox, F. Hollande, S. Ansieau, J. Caramel, D. Birnbaum, A. M. Vigneron, A. Tissier, E. Charafe-Jauffret, and A. Puisieux. 2017. 'A stemness-related ZEB1-MSRB3 axis governs cellular pliancy and breast cancer genome stability', *Nat Med*, 23: 568-78.

- Munzel, M., D. Globisch, and T. Carell. 2011. '5-Hydroxymethylcytosine, the sixth base of the genome', *Angew Chem Int Ed Engl*, 50: 6460-8.
- Muro, Y., H. Masumoto, K. Yoda, N. Nozaki, M. Ohashi, and T. Okazaki. 1992. 'Centromere protein B assembles human centromeric alpha-satellite DNA at the 17-bp sequence, CENP-B box', *J Cell Biol*, 116: 585-96.
- Nan, X., H. Ng, C.A. Johnson, C.D. Laherty, B.M. Turner, R.N. Eisenman, and A. Bird. 1998. 'Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex', *Nature*, 393.
- Naveh-Many, T., and H. Cedar. 1981. 'Active gene sequences are undermethylated', *PNAS*, 78.
- Nejman, D., R. Straussman, I. Steinfeld, M. Ruvolo, D. Roberts, Z. Yakhini, and H. Cedar. 2014. 'Molecular rules governing de novo methylation in cancer', *Cancer Res*, 74: 1475-83.
- Neri, F., D. Incarnato, A. Krepelova, S. Rapelli, F. Anselmi, C. Parlato, C. Medana, F. Dal Bello, and S. Oliviero. 2015. 'Single-Base Resolution Analysis of 5-Formyl and 5-Carboxyl Cytosine Reveals Promoter DNA Methylation Dynamics', *Cell Rep*, 10: 674-83.
- Nestor, C. E., R. Ottaviano, J. Reddington, D. Sproul, D. Reinhardt, D. Dunican, E. Katz, J. M. Dixon, D. J. Harrison, and R. R. Meehan. 2012. 'Tissue type is a major modifier of the 5-hydroxymethylcytosine content of human genes', *Genome Res*, 22: 467-77.
- Nichols, J., and A. Smith. 2009. 'Naive and primed pluripotent states', *Cell Stem Cell*, 4: 487-92.
- . 2012. 'Pluripotency in the embryo and in culture', *Cold Spring Harb Perspect Biol*, 4: a008128.
- Nishino, K., M. Toyoda, M. Yamazaki-Inoue, Y. Fukawatase, E. Chikazawa, H. Sakaguchi, H. Akutsu, and A. Umezawa. 2011. 'DNA methylation dynamics in human induced pluripotent stem cells over time', *PLoS Genet*, 7: e1002085.
- Nishino, K., M. Toyoda, M. Yamazaki-Inoue, H. Makino, Y. Fukawatase, E. Chikazawa, Y. Takahashi, Y. Miyagawa, H. Okita, N. Kiyokawa, H. Akutsu, and A. Umezawa. 2010. 'Defining hypo-methylated regions of stem cell-specific promoters in human iPS cells derived from extra-embryonic amnions and lung fibroblasts', *PLoS One*, 5: e13017.
- Niwa, T., T. Tsukamoto, T. Toyoda, A. Mori, H. Tanaka, T. Maekita, M. Ichinose, M. Tatematsu, and T. Ushijima. 2010. 'Inflammatory processes triggered by Helicobacter pylori infection cause aberrant DNA methylation in gastric epithelial cells', *Cancer Res*, 70: 1430-40.
- O'Connor, M. D., M. D. Kardel, I. Iosfina, D. Youssef, M. Lu, M. M. Li, S. Vercauteren, A. Nagy, and C. J. Eaves. 2008. 'Alkaline phosphatase-positive colony formation is a sensitive, specific, and quantitative indicator of undifferentiated human embryonic stem cells', *Stem Cells*, 26: 1109-16.
- O'Hagan, H. M., H. P. Mohammad, and S. B. Baylin. 2008. 'Double strand breaks can initiate gene silencing and SIRT1-dependent onset of DNA methylation in an exogenous promoter CpG island', *PLoS Genet*, 4: e1000155.
- Oguro, H., J. Yuan, H. Ichikawa, T. Ikawa, S. Yamazaki, H. Kawamoto, H. Nakauchi, and A. Iwama. 2010. 'Poised Lineage Specification in Multipotential Hematopoietic Stem and Progenitor Cells by the Polycomb Protein Bmi1', *Cell Stem Cell*, 6.
- Ohm, J. E., P. Mali, L. Van Neste, D. M. Berman, L. Liang, K. Pandiyan, K. J. Briggs, W. Zhang, P. Argani, B. Simons, W. Yu, W. Matsui, W. Van Criekinge, F. V. Rassool, E. Zambidis, K. E. Schuebel, L. Cope, J. Yen, H. P. Mohammad, L. Cheng, and S. B. Baylin. 2010. 'Cancer-related epigenome changes associated with reprogramming to induced pluripotent stem cells', *Cancer Res*, 70: 7662-73.
- Ohm, J. E., K. M. McGarvey, X. Yu, L. Cheng, K. E. Schuebel, L. Cope, H. P. Mohammad, W. Chen, V. C. Daniel, W. Yu, D. M. Berman, T. Jenuwein, K. Pruitt, S. J. Sharkis, D. N. Watkins, J. G. Herman, and S. B. Baylin. 2007. 'A stem cell-like chromatin pattern may predispose tumor suppressor genes to DNA hypermethylation and heritable silencing', *Nat Genet*, 39: 237-42.

- Ohnishi, K., K. Semi, T. Yamamoto, M. Shimizu, A. Tanaka, K. Mitsunaga, K. Okita, K. Osafune, Y. Arioka, T. Maeda, H. Soejima, H. Moriwaki, S. Yamanaka, K. Woltjen, and Y. Yamada. 2014. 'Premature termination of reprogramming in vivo leads to cancer development through altered epigenetic regulation', *Cell*, 156: 663-77.
- Okano, M., S. Xie, and E. Li. 1998. 'Dnmt2 is not required for de novo and maintenance methylation of viral DNA in embryonic stem cells', *Nucleic Acids Res*, 26.
- 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: 247-57.
- Okano, M. Xie, S. Li, E. 1998. 'Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases', *Nat Genet*, 19: 219-20.
- Okita, K., T. Ichisaka, and S. Yamanaka. 2007. 'Generation of germline-competent induced pluripotent stem cells', *Nature*, 448: 313-7.
- Ooi, S. K., C. Qiu, E. Bernstein, K. Li, D. Jia, Z. Yang, H. Erdjument-Bromage, P. Tempst, S. P. Lin, C. D. Allis, X. Cheng, and T. H. Bestor. 2007. 'DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA', *Nature*, 448: 714-7.
- Orkin, S. H., and K. Hochedlinger. 2011. 'Chromatin connections to pluripotency and cellular reprogramming', *Cell*, 145: 835-50.
- Ostler, K. R., E. M. Davis, S. L. Payne, B. B. Gosalia, J. Exposito-Céspedes, M. M. Le Beau, and L. A. Godley. 2007. 'Cancer cells express aberrant DNMT3B transcripts encoding truncated proteins', *Oncogene*, 26: 5553-63.
- Otani, J., T. Nankumo, K. Arita, S. Inamoto, M. Ariyoshi, and M. Shirakawa. 2009. 'Structural basis for recognition of H3K4 methylation status by the DNA methyltransferase 3A ATRX-DNMT3-DNMT3L domain', *EMBO Rep*, 10: 1235-41.
- Pacaud, R., E. Brocard, L. Lalier, E. Hervouet, F. M. Vallette, and P. F. Cartron. 2014. 'The DNMT1/PCNA/UHRF1 disruption induces tumorigenesis characterized by similar genetic and epigenetic signatures', *Sci Rep*, 4: 4230.
- Panier, S., and S. J. Boulton. 2014. 'Double-strand break repair: 53BP1 comes into focus', *Nat Rev Mol Cell Biol*, 15: 7-18.
- Pastor, W. A., D. Chen, W. Liu, R. Kim, A. Sahakyan, A. Lukianchikov, K. Plath, S. E. Jacobsen, and A. T. Clark. 2016. 'Naive Human Pluripotent Cells Feature a Methylation Landscape Devoid of Blastocyst or Germline Memory', *Cell Stem Cell*, 18: 323-9.
- Pastor, W. A., U. J. Pape, Y. Huang, H. R. Henderson, R. Lister, M. Ko, E. M. McLoughlin, Y. Brudno, S. Mahapatra, P. Kapranov, M. Tahilian, G. Q. Daley, X. S. Liu, J. R. Ecker, P. M. Milos, S. Agarwal, and A. Rao. 2011. 'Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells', *Nature*, 473: 394-7.
- Pawlak, M., and R. Jaenisch. 2011. 'De novo DNA methylation by Dnmt3a and Dnmt3b is dispensable for nuclear reprogramming of somatic cells to a pluripotent state', *Genes Dev*, 25: 1035-40.
- Perez-Duran, P., L. Belver, V. G. de Yébenes, P. Delgado, D. G. Pisano, and A. R. Ramiro. 2012. 'UNG shapes the specificity of AID-induced somatic hypermutation', *J Exp Med*, 209: 1379-89.
- Pfaffeneder, T., B. Hackner, M. Truss, M. Munzel, M. Müller, C. A. Deiml, C. Hagemeyer, and T. Carell. 2011. 'The discovery of 5-formylcytosine in embryonic stem cell DNA', *Angew Chem Int Ed Engl*, 50: 7008-12.
- Pfeifer, G. P. 2018. 'Defining Driver DNA Methylation Changes in Human Cancer', *Int J Mol Sci*, 19.
- Pfeifer, Gerd P. 2017. 'DNA Methylation and Mutation', *eLS*.
- Plass, C., S. M. Pfister, A. M. Lindroth, O. Bogatyrova, R. Claus, and P. Lichter. 2013. 'Mutations in regulators of the epigenome and their connections to global chromatin patterns in cancer', *Nat Rev Genet*, 14: 765-80.

- Polak, P., R. Karlic, A. Koren, R. Thurman, R. Sandstrom, M. Lawrence, A. Reynolds, E. Rynes, K. Vlahovicek, J. A. Stamatoyannopoulos, and S. R. Sunyaev. 2015. 'Cell-of-origin chromatin organization shapes the mutational landscape of cancer', *Nature*, 518: 360-64.
- Politi, Valeria, Giovanni Perini, Stefania Trazzi, Artem Pliss, Ivan Raska, William C. Earnshaw, and Giuliano Della Valle. 2002. 'CENP-C binds the alpha-satellite DNA in vivo at specific centromere domains', *Journal of Cell Science*, 115: 2317.
- Polo, J. M., E. Anderssen, R. M. Walsh, B. A. Schwarz, C. M. Nefzger, S. M. Lim, M. Borkent, E. Apostolou, S. Alaei, J. Cloutier, O. Bar-Nur, S. Cheloufi, M. Stadtfeld, M. E. Figueroa, D. Robinton, S. Natesan, A. Melnick, J. Zhu, S. Ramaswamy, and K. Hochedlinger. 2012. 'A molecular roadmap of reprogramming somatic cells into iPS cells', *Cell*, 151: 1617-32.
- Popp, C., W. Dean, S. Feng, S. J. Cokus, S. Andrews, M. Pellegrini, S. E. Jacobsen, and W. Reik. 2010. 'Genome-wide erasure of DNA methylation in mouse primordial germ cells is affected by AID deficiency', *Nature*, 463: 1101-5.
- Puri, S., A. E. Foliás, and M. Hebrok. 2015. 'Plasticity and dedifferentiation within the pancreas: development, homeostasis, and disease', *Cell Stem Cell*, 16: 18-31.
- Qin, H., M. Hejna, Y. Liu, M. Percharde, M. Wossidlo, L. Blouin, J. Durruthy-Durruthy, P. Wong, Z. Qi, J. Yu, L. S. Qi, V. Sebastiano, J. S. Song, and M. Ramalho-Santos. 2016. 'YAP Induces Human Naive Pluripotency', *Cell Rep*, 14: 2301-12.
- Qin, W., H. Leonhardt, and F. Spada. 2011. 'Usp7 and Uhrf1 control ubiquitination and stability of the maintenance DNA methyltransferase Dnmt1', *J Cell Biochem*, 112: 439-44.
- Rajavelu, A., C. Lungu, M. Emperle, M. Dukatz, A. Brohm, J. Broche, I. Hanelt, E. Parsa, S. Schiffers, R. Karnik, A. Meissner, T. Carell, P. Rathert, R. Z. Jurkowska, and A. Jeltsch. 2018. 'Chromatin-dependent allosteric regulation of DNMT3A activity by MeCP2', *Nucleic Acids Res.*
- Rakyan, V. K., T. A. Down, S. Maslau, T. Andrew, T. P. Yang, H. Beyan, P. Whittaker, O. T. McCann, S. Finer, A. M. Valdes, R. D. Leslie, P. Deloukas, and T. D. Spector. 2010. 'Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains', *Genome Res*, 20: 434-9.
- Rideout, W. M. 3rd, G. A. Coetzee, A. F. Olumi, and P. A. Jones. 1990. '5-Methylcytosine as an endogenous mutagen in the human LDL receptor and p53 genes.', *Science*, 249: 1288-90.
- Riggi, N., M. L. Suva, C. De Vito, P. Provero, J. C. Stehle, K. Baumer, L. Cironi, M. Janiszewska, T. Petricevic, D. Suva, S. Tercier, J. M. Joseph, L. Guillou, and I. Stamenkovic. 2010. 'EWS-FLI-1 modulates miRNA145 and SOX2 expression to initiate mesenchymal stem cell reprogramming toward Ewing sarcoma cancer stem cells', *Genes Dev*, 24: 916-32.
- Robinson, M. D., D. J. McCarthy, and G. K. Smyth. 2010. 'edgeR: a Bioconductor package for differential expression analysis of digital gene expression data', *Bioinformatics*, 26: 139-40.
- Roll, J. D., A. G. Rivenbark, W. D. Jones, and W. B. Coleman. 2008. 'DNMT3b overexpression contributes to a hypermethylator phenotype in human breast cancer cell lines', *Mol Cancer*, 7: 15.
- Rothbart, S. B., K. Krajewski, N. Nady, W. Tempel, S. Xue, A. I. Badeaux, D. Barsyte-Lovejoy, J. Y. Martinez, M. T. Bedford, S. M. Fuchs, C. H. Arrowsmith, and B. D. Strahl. 2012. 'Association of UHRF1 with methylated H3K9 directs the maintenance of DNA methylation', *Nat Struct Mol Biol*, 19: 1155-60.
- Rowland, B. D., and D. S. Peeper. 2006. 'KLF4, p21 and context-dependent opposing forces in cancer', *Nat Rev Cancer*, 6: 11-23.
- Roy, N., and M. Hebrok. 2015. 'Regulation of Cellular Identity in Cancer', *Dev Cell*, 35: 674-84.
- Rudin, C. M., S. Durinck, E. W. Stawiski, J. T. Poirier, Z. Modrusan, D. S. Shames, E. A. Bergbower, Y. Guan, J. Shin, J. Guillory, C. S. Rivers, C. K. Foo, D. Bhatt, J. Stinson, F.

- Gnad, P. M. Haverty, R. Gentleman, S. Chaudhuri, V. Janakiraman, B. S. Jaiswal, C. Parikh, W. Yuan, Z. Zhang, H. Koeppen, T. D. Wu, H. M. Stern, R. L. Yauch, K. E. Huffman, D. D. Paskulin, P. B. Illei, M. Varella-Garcia, A. F. Gazdar, F. J. de Sauvage, R. Bourgon, J. D. Minna, M. V. Brock, and S. Seshagiri. 2012. 'Comprehensive genomic analysis identifies SOX2 as a frequently amplified gene in small-cell lung cancer', *Nat Genet*, 44: 1111-6.
- Ruiz, S., D. Diep, A. Gore, A.D. Panopoulos, N. Montserrat, N. Plongthongkum, S. Kumar, H.L. Fung, A. Giorgetti, J. Bilic, E.M. Batchelder, H. Zaehres, N.G. Kan, H.R. Schöler, M. Mercola, K. Zhang, and J.C.I. Belmonte. 2012. 'Identification of a specific reprogramming-associated epigenetic signature in human induced pluripotent stem cells', *PNAS*, 109: 16196-201.
- Ruiz, S., A. J. Lopez-Contreras, M. Gabut, R. M. Marion, P. Gutierrez-Martinez, S. Bua, O. Ramirez, I. Olalde, S. Rodrigo-Perez, H. Li, T. Marques-Bonet, M. Serrano, M. A. Blasco, N. N. Batada, and O. Fernandez-Capetillo. 2015. 'Limiting replication stress during somatic cell reprogramming reduces genomic instability in induced pluripotent stem cells', *Nat Commun*, 6: 8036.
- Rush, Margaret, Ruth Appanah, Sandra Lee, Lucia L. Lam, Preeti Goyal, and Matthew C. Lorincz. 2014. 'Targeting of EZH2 to a defined genomic site is sufficient for recruitment of Dnmt3a but not de novo DNA methylation', *Epigenetics*, 4: 404-14.
- Sarig, R., N. Rivlin, R. Brosh, C. Bornstein, I. Kamer, O. Ezra, A. Molchadsky, N. Goldfinger, O. Brenner, and V. Rotter. 2010. 'Mutant p53 facilitates somatic cell reprogramming and augments the malignant potential of reprogrammed cells', *J Exp Med*, 207: 2127-40.
- Saunderson, E. A., P. Stepper, J. J. Gomm, L. Hoa, A. Morgan, M. D. Allen, J. L. Jones, J. G. Gribben, T. P. Jurkowski, and G. Ficiz. 2017. 'Hit-and-run epigenetic editing prevents senescence entry in primary breast cells from healthy donors', *Nat Commun*, 8: 1450.
- Saxonov, S., P. Berg, and D. L. Brutlag. 2006. 'A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters', *Proc Natl Acad Sci U S A*, 103: 1412-7.
- Schlesinger, Y., R. Straussman, I. Keshet, S. Farkash, M. Hecht, J. Zimmerman, E. Eden, Z. Yakhini, E. Ben-Shushan, B. E. Reubinoff, Y. Bergman, I. Simon, and H. Cedar. 2007. 'Polycomb-mediated methylation on Lys27 of histone H3 pre-marks genes for de novo methylation in cancer', *Nat Genet*, 39: 232-6.
- Schmidl, C., M. Klug, T. J. Boeld, R. Andreesen, P. Hoffmann, M. Edinger, and M. Rehli. 2009. 'Lineage-specific DNA methylation in T cells correlates with histone methylation and enhancer activity', *Genome Res*, 19: 1165-74.
- Schmitt, Michael W., Scott R. Kennedy, Jesse J. Salk, Edward J. Fox, Joseph B. Hiatt, and Lawrence A. Loeb. 2012. 'Detection of ultra-rare mutations by next-generation sequencing', *Proceedings of the National Academy of Sciences*, 109: 14508.
- Schmutte, C., A. S. Yang, R. W. Beart, and P. A. Jones. 1995. 'Base Excision Repair of U:G Mismatches at a Mutational Hotspot in the p53 Gene Is More Efficient Than Base Excision Repair of T:G Mismatches in Extracts of Human Colon Tumors', *Cancer Res*, 55: 3742-46.
- Seifert, B. A., M. Dejosez, and T. P. Zwaka. 2017. 'Ronin influences the DNA damage response in pluripotent stem cells', *Stem Cell Res*, 23: 98-104.
- Seisenberger, S., S. Andrews, F. Krueger, J. Arand, J. Walter, F. Santos, C. Popp, B. Thienpont, W. Dean, and W. Reik. 2012. 'The dynamics of genome-wide DNA methylation reprogramming in mouse primordial germ cells', *Mol Cell*, 48: 849-62.
- Seki, Yasuhiro, Akira Kurisaki, Kanako Watanabe-Susaki, Yoshiro Nakajima, Mio Nakanishi, Yoshikazu Arai, Kunio Shiota, Hiromu Sugino, and Makoto Asashima. 2010. 'TIF1 $\beta$  regulates the pluripotency of embryonic stem cells in a phosphorylation-dependent manner', *Proceedings of the National Academy of Sciences*, 107: 10926.

- Sendzikaite, G., C. W. Hanna, K. R. Stewart-Morgan, E. Ivanova, and G. Kelsey. 2019. 'A DNMT3A PWWP mutation leads to methylation of bivalent chromatin and growth retardation in mice', *Nat Commun*, 10: 1884.
- Sharif, J., M. Muto, S. Takebayashi, I. Suetake, A. Iwamatsu, T. A. Endo, J. Shinga, Y. Mizutani-Koseki, T. Toyoda, K. Okamura, S. Tajima, K. Mitsuya, M. Okano, and H. Koseki. 2007. 'The SRA protein Np95 mediates epigenetic inheritance by recruiting Dnmt1 to methylated DNA', *Nature*, 450: 908-12.
- Sheaffer, K. L., E. N. Elliott, and K. H. Kaestner. 2016. 'DNA Hypomethylation Contributes to Genomic Instability and Intestinal Cancer Initiation', *Cancer Prev Res (Phila)*, 9: 534-46.
- Shen, J.C., W.M. Rideout, and P.A. Jones. 1994. 'The rate of hydrolytic deamination of 5-methylcytosine in double-stranded DNA', *Nucleic Acids Res*, 22.
- Shen, L., and Y. Zhang. 2013. '5-Hydroxymethylcytosine: generation, fate, and genomic distribution', *Curr Opin Cell Biol*, 25: 289-96.
- Shieh, S., M. Ikeda, Y. Taya, and C. Prives. 1997. 'DNA Damage-Induced Phosphorylation of p53 Alleviates Inhibition by MDM2', *Cell*, 91: 325-34.
- Shih, Ie-Ming, Wei Zhou, Steven N. Goodman, Christoph Lengauer, Kenneth W. Kinzler, and Bert Vogelstein. 2001. 'Evidence That Genetic Instability Occurs at an Early Stage of Colorectal Tumorigenesis', *Cancer Research*, 61: 818.
- Siegfried, Z., S. Eden, M. Mendelsohn, X. Feng, B.Z. Tsuberi, and H. Cedar. 1999. 'DNA methylation represses transcription in vivo', *Nature Genetics*, 22.
- Silva, J., O. Barrandon, J. Nichols, J. Kawaguchi, T. W. Theunissen, and A. Smith. 2008. 'Promotion of reprogramming to ground state pluripotency by signal inhibition', *PLoS Biol*, 6: e253.
- Simara, P., L. Tesarova, D. Rehakova, P. Matula, S. Stejskal, A. Hampl, and I. Koutna. 2017. 'DNA double-strand breaks in human induced pluripotent stem cell reprogramming and long-term in vitro culturing', *Stem Cell Res Ther*, 8: 73.
- Simicevic, J., and B. Deplancke. 2017. 'Transcription factor proteomics-Tools, applications, and challenges', *Proteomics*, 17.
- Smallwood, S. A., and G. Kelsey. 2012. 'De novo DNA methylation: a germ cell perspective', *Trends Genet*, 28: 33-42.
- Smith, A. 2017. 'Formative pluripotency: the executive phase in a developmental continuum', *Development*, 144: 365-73.
- Smith, Z. D., M. M. Chan, K. C. Humm, R. Karnik, S. Mekhoubad, A. Regev, K. Eggan, and A. Meissner. 2014. 'DNA methylation dynamics of the human preimplantation embryo', *Nature*, 511: 611-5.
- Smith, Z. D., M. M. Chan, T. S. Mikkelsen, H. Gu, A. Gnirke, A. Regev, and A. Meissner. 2012. 'A unique regulatory phase of DNA methylation in the early mammalian embryo', *Nature*, 484: 339-44.
- Smith, Z. D., J. Shi, H. Gu, J. Donaghey, K. Clement, D. Cacchiarelli, A. Gnirke, F. Michor, and A. Meissner. 2017. 'Epigenetic restriction of extraembryonic lineages mirrors the somatic transition to cancer', *Nature*, 549: 543-47.
- Song, C. X., and C. He. 2013. 'Potential functional roles of DNA demethylation intermediates', *Trends Biochem Sci*, 38: 480-4.
- Song, C. X., K. E. Szulwach, Q. Dai, Y. Fu, S. Q. Mao, L. Lin, C. Street, Y. Li, M. Poidevin, H. Wu, J. Gao, P. Liu, L. Li, G. L. Xu, P. Jin, and C. He. 2013. 'Genome-wide profiling of 5-formylcytosine reveals its roles in epigenetic priming', *Cell*, 153: 678-91.
- Spencer, D. H., D. A. Russler-Germain, S. Ketkar, N. M. Helton, T. L. Lamprecht, R. S. Fulton, C. C. Fronick, M. O'Laughlin, S. E. Heath, M. Shinawi, P. Westervelt, J. E. Payton, L. D. Wartman, J. S. Welch, R. K. Wilson, M. J. Walter, D. C. Link, J. F. DiPersio, and T. J. Ley. 2017. 'CpG Island Hypermethylation Mediated by DNMT3A Is a Consequence of AML Progression', *Cell*, 168: 801-16 e13.

- Sproul, D., R.K. Kitchen, C.E. Nestor, J.M. Dixon, A.H. Sims, D.J. Harrison, B.H. Ramsahoye, and R.R. Meehan. 2012. 'Tissue of origin determines cancer-associated CpG island promoter hypermethylation patterns', *Genome Biol*, 13.
- Sproul, D., and R. R. Meehan. 2013. 'Genomic insights into cancer-associated aberrant CpG island hypermethylation', *Brief Funct Genomics*, 12: 174-90.
- Spruijt, C. G., F. Gnerlich, A. H. Smits, T. Pfaffeneder, P. W. Jansen, C. Bauer, M. Munzel, M. Wagner, M. Muller, F. Khan, H. C. Eberl, A. Mensinga, A. B. Brinkman, K. Lephikov, U. Muller, J. Walter, R. Boelens, H. van Ingen, H. Leonhardt, T. Carell, and M. Vermeulen. 2013. 'Dynamic readers for 5-(hydroxy)methylcytosine and its oxidized derivatives', *Cell*, 152: 1146-59.
- Stadler, M. B., R. Murr, L. Burger, R. Ivanek, F. Lienert, A. Scholer, E. van Nimwegen, C. Wirbelauer, E. J. Oakeley, D. Gaidatzis, V. K. Tiwari, and D. Schubeler. 2011. 'DNA-binding factors shape the mouse methylome at distal regulatory regions', *Nature*, 480: 490-5.
- Stirparo, G. G., T. Boroviak, G. Guo, J. Nichols, A. Smith, and P. Bertone. 2018. 'Integrated analysis of single-cell embryo data yields a unified transcriptome signature for the human pre-implantation epiblast', *Development*, 145.
- Su, J., Y. H. Huang, X. Cui, X. Wang, X. Zhang, Y. Lei, J. Xu, X. Lin, K. Chen, J. Lv, M. A. Goodell, and W. Li. 2018. 'Homeobox oncogene activation by pan-cancer DNA hypermethylation', *Genome Biol*, 19: 108.
- Su, P. F., T. C. Lee, P. J. Lin, P. H. Lee, Y. M. Jeng, C. H. Chen, J. D. Liang, L. L. Chiou, G. T. Huang, and H. S. Lee. 2007. 'Differential DNA methylation associated with hepatitis B virus infection in hepatocellular carcinoma', *Int J Cancer*, 121: 1257-64.
- Suvà, M.L. , N. Riggi, and B.E. Bernstein. 2013. 'Epigenetic Reprogramming in Cancer', *Science*, 339.
- Suzuki, M., T. Yamada, F. Kihara-Negishi, T. Sakurai, E. Hara, D. G. Tenen, N. Hozumi, and T. Oikawa. 2006. 'Site-specific DNA methylation by a complex of PU.1 and Dnmt3a/b', *Oncogene*, 25: 2477-88.
- Tahiliani, M., K. Peng Koh, Y. Shen, W.A. Pastor, H. Bandukwala, Y. Brudno, S. Agarwal, L.M. Iyer, D.R. Liu, L. Aravind, and A. Rao. 2009. 'Conversion of 5-Methylcytosine to 5-Hydroxymethylcytosine in Mammalian DNA by MLL Partner TET1', *Science*, 324.
- Takahashi, K., K. Tanabe, M. Ohnuki, M. Narita, T. Ichisaka, K. Tomoda, and S. Yamanaka. 2007. 'Induction of pluripotent stem cells from adult human fibroblasts by defined factors', *Cell*, 131: 861-72.
- Takahashi, K., and S. Yamanaka. 2006. 'Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors', *Cell*, 126: 663-76.
- . 2015. 'A developmental framework for induced pluripotency', *Development*, 142: 3274-85.
- Takashima, Y., G. Guo, R. Loos, J. Nichols, G. Ficz, F. Krueger, D. Oxley, F. Santos, J. Clarke, W. Mansfield, W. Reik, P. Bertone, and A. Smith. 2014. 'Resetting transcription factor control circuitry toward ground-state pluripotency in human', *Cell*, 158: 1254-69.
- Tao, Y., B. Kang, D. A. Petkovich, Y. R. Bhandari, J. In, G. Stein-O'Brien, X. Kong, W. Xie, N. Zachos, S. Maegawa, H. Vaidya, S. Brown, R. W. Chiu Yen, X. Shao, J. Thakor, Z. Lu, Y. Cai, Y. Zhang, I. Mallona, M. A. Peinado, C. A. Zahnow, N. Ahuja, E. Fertig, J. P. Issa, S. B. Baylin, and H. Easwaran. 2019. 'Aging-like Spontaneous Epigenetic Silencing Facilitates Wnt Activation, Stemness, and Braf(V600E)-Induced Tumorigenesis', *Cancer Cell*, 35: 315-28 e6.
- Teater, M., P. M. Dominguez, D. Redmond, Z. Chen, D. Ennishi, D. W. Scott, L. Cimmino, P. Ghione, J. Chaudhuri, R. D. Gascoyne, I. Aifantis, G. Inghirami, O. Elemento, A. Melnick, and R. Shaknovich. 2018. 'AICDA drives epigenetic heterogeneity and accelerates germinal center-derived lymphomagenesis', *Nat Commun*, 9: 222.



- Tesar, P. J., J. G. Chenoweth, F. A. Brook, T. J. Davies, E. P. Evans, D. L. Mack, R. L. Gardner, and R. D. McKay. 2007. 'New cell lines from mouse epiblast share defining features with human embryonic stem cells', *Nature*, 448: 196-9.
- Teschendorff, A. E., U. Menon, A. Gentry-Maharaj, S. J. Ramus, D. J. Weisenberger, H. Shen, M. Campan, H. Noushmehr, C. G. Bell, A. P. Maxwell, D. A. Savage, E. Mueller-Holzner, C. Marth, G. Kocjan, S. A. Gayther, A. Jones, S. Beck, W. Wagner, P. W. Laird, I. J. Jacobs, and M. Widschwendter. 2010. 'Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer', *Genome Res*, 20: 440-6.
- Theunissen, T. W., B. E. Powell, H. Wang, M. Mitalipova, D. A. Faddah, J. Reddy, Z. P. Fan, D. Maetzel, K. Ganz, L. Shi, T. Lungjangwa, S. Imsoonthornruksa, Y. Stelzer, S. Rangarajan, A. D'Alessio, J. Zhang, Q. Gao, M. M. Dawlaty, R. A. Young, N. S. Gray, and R. Jaenisch. 2014. 'Systematic identification of culture conditions for induction and maintenance of naive human pluripotency', *Cell Stem Cell*, 15: 471-87.
- Thompson, S. L., S. F. Bakhoun, and D. A. Compton. 2010. 'Mechanisms of chromosomal instability', *Curr Biol*, 20: R285-95.
- Thomson, J. P., P. J. Skene, J. Selfridge, T. Clouaire, J. Guy, S. Webb, A. R. Kerr, A. Deaton, R. Andrews, K. D. James, D. J. Turner, R. Illingworth, and A. Bird. 2010. 'CpG islands influence chromatin structure via the CpG-binding protein Cfp1', *Nature*, 464: 1082-6.
- Thomson, James A., Joseph Itskovitz-Eldor, Sander S. Shapiro, Michelle A. Waknitz, Jennifer J. Swiergiel, Vivienne S. Marshall, and Jeffrey M. Jones. 1998. 'Embryonic Stem Cell Lines Derived from Human Blastocysts', *Science*, 282: 1145.
- Tian, Y., T. J. Morris, A. P. Webster, Z. Yang, S. Beck, A. Feber, and A. E. Teschendorff. 2017. 'ChAMP: updated methylation analysis pipeline for Illumina BeadChips', *Bioinformatics*, 33: 3982-84.
- Ting, D. T., D. Lipson, S. Paul, B. W. Brannigan, S. Akhavanfard, E. J. Coffman, G. Contino, V. Deshpande, A. J. Iafrate, S. Letovsky, M. N. Rivera, N. Bardeesy, S. Maheswaran, and D. A. Haber. 2011. 'Aberrant overexpression of satellite repeats in pancreatic and other epithelial cancers', *Science*, 331: 593-6.
- Tomkova, M., M. McClellan, S. Kriaucionis, and B. Schuster-Bockler. 2018. 'DNA Replication and associated repair pathways are involved in the mutagenesis of methylated cytosine', *DNA Repair (Amst)*, 62: 1-7.
- Tomkova, M., M. McClellan, S. Kriaucionis, and B. Schuster-Boeckler. 2016. '5-hydroxymethylcytosine marks regions with reduced mutation frequency in human DNA', *Elife*, 5.
- Tootle, T. L., and I. Rebay. 2005. 'Post-translational modifications influence transcription factor activity: a view from the ETS superfamily', *Bioessays*, 27: 285-98.
- Tovy, A., A. Spiro, R. McCarthy, Z. Shipony, Y. Aylon, K. Alton, E. Ainbinder, N. Furth, A. Tanay, M. Barton, and M. Oren. 2017. 'p53 is essential for DNA methylation homeostasis in naive embryonic stem cells, and its loss promotes clonal heterogeneity', *Genes Dev*.
- Toyota, M., N. Ahuja, M. Ohe-Toyota, J. G. Herman, S. B. Baylin, and J. P. Issa. 1999. 'CpG island methylator phenotype in colorectal cancer', *PNAS*, 96.
- Valinluck, V., and L. C. Sowers. 2007. 'Endogenous cytosine damage products alter the site selectivity of human DNA maintenance methyltransferase DNMT1', *Cancer Res*, 67: 946-50.
- Valinluck, V., H. H. Tsai, D. K. Rogstad, A. Burdzy, A. Bird, and L. C. Sowers. 2004. 'Oxidative damage to methyl-CpG sequences inhibits the binding of the methyl-CpG binding domain (MBD) of methyl-CpG binding protein 2 (MeCP2)', *Nucleic Acids Res*, 32: 4100-8.
- Verma, N., H. Pan, L. C. Dore, A. Shukla, Q. V. Li, B. Pelham-Webb, V. Teijeiro, F. Gonzalez, A. Krivtsov, C. J. Chang, E. P. Papapetrou, C. He, O. Elemento, and D. Huangfu. 2018. 'TET

- proteins safeguard bivalent promoters from de novo methylation in human embryonic stem cells', *Nat Genet*, 50: 83-95.
- Vertino, P. M., R. W. Yen, J. Gao, and S. B. Baylin. 1996. 'De Novo Methylation of CpG Island Sequences in Human Fibroblasts Overexpressing DNA (Cytosine-5)-Methyltransferase', *Mol Cell Biol*, 16: 4555-65.
- Vire, E., C. Brenner, R. Deplus, L. Blanchon, M. Fraga, C. Didelot, L. Morey, A. Van Eynde, D. Bernard, J. M. Vanderwinden, M. Bollen, M. Esteller, L. Di Croce, Y. de Launoit, and F. Fuks. 2006. 'The Polycomb group protein EZH2 directly controls DNA methylation', *Nature*, 439: 871-4.
- von Meyenn, F., M. Iurlaro, E. Habibi, N. Q. Liu, A. Salehzadeh-Yazdi, F. Santos, E. Petrini, I. Milagre, M. Yu, Z. Xie, L. I. Kroeze, T. B. Nesterova, J. H. Jansen, H. Xie, C. He, W. Reik, and H. G. Stunnenberg. 2016. 'Impairment of DNA Methylation Maintenance Is the Main Cause of Global Demethylation in Naive Embryonic Stem Cells', *Mol Cell*.
- Voo, K.S., D.L. Carlone, B. Jacobsen, A. Flodin, and D.G. Skalnik. 2000. 'Cloning of a Mammalian Transcriptional Activator That Binds Unmethylated CpG Motifs and Shares a CXXC Domain with DNA Methyltransferase, Human Trithorax, and Methyl-CpG Binding Domain Protein 1', *Molecular and Cellular Biology*, 20.
- Waddington, C. H. 1942. 'The epigenotype', *Endeavour*, 1: 18-20.
- . 1968. 'Towards a Theoretical Biology', *Nature*, 218: 525-27.
- Wagner, E. J., and P. B. Carpenter. 2012. 'Understanding the language of Lys36 methylation at histone H3', *Nat Rev Mol Cell Biol*, 13: 115-26.
- Wang, Jianlong, Sridhar Rao, Jianlin Chu, Xiaohua Shen, Dana N. Levasseur, Thorold W. Theunissen, and Stuart H. Orkin. 2006. 'A protein interaction network for pluripotency of embryonic stem cells', *Nature*, 444: 364-68.
- Wang, W. E., L. Li, X. Xia, W. Fu, Q. Liao, C. Lan, D. Yang, H. Chen, R. Yue, C. Zeng, L. Zhou, B. Zhou, D. D. Duan, X. Chen, S. R. Houser, and C. Zeng. 2017. 'Dedifferentiation, Proliferation, and Redifferentiation of Adult Mammalian Cardiomyocytes After Ischemic Injury', *Circulation*, 136: 834-48.
- Ware, C. B., A. M. Nelson, B. Mecham, J. Hesson, W. Zhou, E. C. Jonlin, A. J. Jimenez-Caliani, X. Deng, C. Cavanaugh, S. Cook, P. J. Tesar, J. Okada, L. Margaretha, H. Sperber, M. Choi, C. A. Blau, P. M. Treuting, R. D. Hawkins, V. Cirulli, and H. Ruohola-Baker. 2014. 'Derivation of naïve human embryonic stem cells', *PNAS*, 111: 4484-89.
- Watt, F., and P.L. MoUoy. 1988. 'Cytosine methylation prevents binding to DNA of a HeLa cell transcription factor required for optimal expression of the adenovirus major late promoter', *Genes Dev*, 2.
- Weber, M., I. Hellmann, M. B. Stadler, L. Ramos, S. Paabo, M. Rebhan, and D. Schubeler. 2007. 'Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome', *Nat Genet*, 39: 457-66.
- Webster, K. E., M. K. O'Bryan, S. Fletcher, P. E. Crewther, U. Aapola, J. Craig, D. K. Harrison, H. Aung, N. Phutikanit, R. Lyle, S. J. Meachem, S. E. Antonarakis, D. M. de Kretser, M. P. Hedger, P. Peterson, B. J. Carroll, and H. S. Scott. 2005. 'Meiotic and epigenetic defects in Dnmt3L-knockout mouse spermatogenesis', *Proceedings of the National Academy of Sciences*, 102: 4068-73.
- Widschwendter, M., H. Fiegl, D. Egle, E. Mueller-Holzner, G. Spizzo, C. Marth, D. J. Weisenberger, M. Campan, J. Young, I. Jacobs, and P. W. Laird. 2007. 'Epigenetic stem cell signature in cancer', *Nat Genet*, 39: 157-8.
- Willet, S. G., M. A. Lewis, Z. F. Miao, D. Liu, M. D. Radyk, R. L. Cunningham, J. Burclaff, G. Sibbel, H. G. Lo, V. Blanc, N. O. Davidson, Z. N. Wang, and J. C. Mills. 2018. 'Regenerative proliferation of differentiated cells by mTORC1-dependent paligenesis', *EMBO J*, 37.

- Williams, K., J. Christensen, M. T. Pedersen, J. V. Johansen, P. A. Cloos, J. Rappsilber, and K. Helin. 2011. 'TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity', *Nature*, 473: 343-8.
- Wolf, S.F. Jolly, D.J. Lunnen, K.D. Friedmann, T. Migeon, B.R. 1984. 'Methylation of the hypoxanthine phosphoribosyltransferase locus on the human X chromosome: Implications for X-chromosome inactivation', *Proc Natl Acad Sci U S A*, 81: 2806-10.
- Wu, C.T., and J.R. Morris. 2001. 'Genes, Genetics, and Epigenetics: A Correspondence', *Science*, 293.
- Wu, H., A. C. D'Alessio, S. Ito, K. Xia, Z. Wang, K. Cui, K. Zhao, Y. E. Sun, and Y. Zhang. 2011. 'Dual functions of Tet1 in transcriptional regulation in mouse embryonic stem cells', *Nature*, 473: 389-93.
- Wu, H., and Y. Zhang. 2011. 'Mechanisms and functions of Tet protein-mediated 5-methylcytosine oxidation', *Genes Dev*, 25: 2436-52.
- . 2014. 'Reversing DNA methylation: mechanisms, genomics, and biological functions', *Cell*, 156: 45-68.
- Wu, S. C., and Y. Zhang. 2010. 'Active DNA demethylation: many roads lead to Rome', *Nat Rev Mol Cell Biol*, 11: 607-20.
- Xia, L., W. Huang, M. Bellani, M. M. Seidman, K. Wu, D. Fan, Y. Nie, Y. Cai, Y. W. Zhang, L. R. Yu, H. Li, C. A. Zahnow, W. Xie, R. W. Chiu Yen, F. V. Rassool, and S. B. Baylin. 2017. 'CHD4 Has Oncogenic Functions in Initiating and Maintaining Epigenetic Suppression of Multiple Tumor Suppressor Genes', *Cancer Cell*, 31: 653-68 e7.
- Xie, M., C. Lu, J. Wang, M. D. McLellan, K. J. Johnson, M. C. Wendl, J. F. McMichael, H. K. Schmidt, V. Yellapantula, C. A. Miller, B. A. Ozenberger, J. S. Welch, D. C. Link, M. J. Walter, E. R. Mardis, J. F. Dpersio, F. Chen, R. K. Wilson, T. J. Ley, and L. Ding. 2014. 'Age-related mutations associated with clonal hematopoietic expansion and malignancies', *Nat Med*, 20: 1472-8.
- Xu, C., E. Rosler, J. Jiang, J. S. Lebkowski, J. D. Gold, C. O'Sullivan, K. Delavan-Boorsma, M. Mok, A. Bronstein, and M. K. Carpenter. 2005. 'Basic fibroblast growth factor supports undifferentiated human embryonic stem cell growth without conditioned medium', *Stem Cells*, 23: 315-23.
- Xu, Y., F. Wu, L. Tan, L. Kong, L. Xiong, J. Deng, A. J. Barbera, L. Zheng, H. Zhang, S. Huang, J. Min, T. Nicholson, T. Chen, G. Xu, Y. Shi, K. Zhang, and Y. G. Shi. 2011. 'Genome-wide regulation of 5hmC, 5mC, and gene expression by Tet1 hydroxylase in mouse embryonic stem cells', *Mol Cell*, 42: 451-64.
- Xu, Z., A.M. Robitaille, J.D. Berndt, K.C. Davidson, K.A. Fischer, J. Mathieu, J.C. Potter, H. Ruohola-Baker, and R.T. Moon. 2016. 'Wnt/ $\beta$ -catenin signaling promotes self-renewal and inhibits the primed state transition in naïve human embryonic stem cells', *PNAS*.
- Xue, Y., Q. Wang, Q. Long, B. L. Ng, H. Swerdlow, J. Burton, C. Skuce, R. Taylor, Z. Abdellah, Y. Zhao, Asan, D. G. MacArthur, M. A. Quail, N. P. Carter, H. Yang, and C. Tyler-Smith. 2009. 'Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree', *Curr Biol*, 19: 1453-7.
- Yan, L., M. Yang, H. Guo, L. Yang, J. Wu, R. Li, P. Liu, Y. Lian, X. Zheng, J. Yan, J. Huang, M. Li, X. Wu, L. Wen, K. Lao, R. Li, J. Qiao, and F. Tang. 2013. 'Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells', *Nat Struct Mol Biol*, 20: 1131-9.
- Yan, X. J., J. Xu, Z. H. Gu, C. M. Pan, G. Lu, Y. Shen, J. Y. Shi, Y. M. Zhu, L. Tang, X. W. Zhang, W. X. Liang, J. Q. Mi, H. D. Song, K. Q. Li, Z. Chen, and S. J. Chen. 2011. 'Exome sequencing identifies somatic mutations of DNA methyltransferase gene DNMT3A in acute monocytic leukemia', *Nat Genet*, 43: 309-15.

- Yang, Z., S. Jiang, Y. Cheng, T. Li, W. Hu, Z. Ma, F. Chen, and Y. Yang. 2017. 'FOXC1 in cancer development and therapy: deciphering its emerging and divergent roles', *Ther Adv Med Oncol*, 9: 797-816.
- Yildirim, O., R. Li, J. H. Hung, P. B. Chen, X. Dong, L. S. Ee, Z. Weng, O. J. Rando, and T. G. Fazio. 2011. 'Mbd3/NURD complex regulates expression of 5-hydroxymethylcytosine marked genes in embryonic stem cells', *Cell*, 147: 1498-510.
- Ying, Q. L., J. Wray, J. Nichols, L. Batlle-Morera, B. Doble, J. Woodgett, P. Cohen, and A. Smith. 2008. 'The ground state of embryonic stem cell self-renewal', *Nature*, 453: 519-23.
- Yoshihara, M., R. Araki, Y. Kasama, M. Sunayama, M. Abe, K. Nishida, H. Kawaji, Y. Hayashizaki, and Y. Murakawa. 2017. 'Hotspots of De Novo Point Mutations in Induced Pluripotent Stem Cells', *Cell Rep*, 21: 308-15.
- Youn, A., K. I. Kim, R. Rabadan, B. Tycko, Y. Shen, and S. Wang. 2018. 'A pan-cancer analysis of driver gene mutations, DNA methylation and gene expressions reveals that chromatin remodeling is a major mechanism inducing global changes in cancer epigenomes', *BMC Med Genomics*, 11: 98.
- Young, A. L., G. A. Challen, B. M. Birmann, and T. E. Druley. 2016. 'Clonal haematopoiesis harbouring AML-associated mutations is ubiquitous in healthy adults', *Nat Commun*, 7: 12484.
- Yu, S., K. Tong, Y. Zhao, I. Balasubramanian, G. S. Yap, R. P. Ferraris, E. M. Bonder, M. P. Verzi, and N. Gao. 2018. 'Paneth Cell Multipotency Induced by Notch Activation following Injury', *Cell Stem Cell*, 23: 46-59 e5.
- Zemojtel, T., S. M. Kielbasa, P. F. Arndt, S. Behrens, G. Bourque, and M. Vingron. 2011. 'CpG deamination creates transcription factor-binding sites with high efficiency', *Genome Biol Evol*, 3: 1304-11.
- Zhan, T., N. Rindtorff, and M. Boutros. 2017. 'Wnt signaling in cancer', *Oncogene*, 36: 1461-73.
- Zhang, Y., R. Jurkowska, S. Soeroes, A. Rajavelu, A. Dhayalan, I. Bock, P. Rathert, O. Brandt, R. Reinhardt, W. Fischle, and A. Jeltsch. 2010. 'Chromatin methylation activity of Dnmt3a and Dnmt3a/3L is guided by interaction of the ADD domain with the histone H3 tail', *Nucleic Acids Res*, 38: 4246-53.
- Zhao, D., Q. Zhang, Y. Liu, X. Li, K. Zhao, Y. Ding, Z. Li, Q. Shen, C. Wang, N. Li, and X. Cao. 2016. 'H3K4me3 Demethylase Kdm5a Is Required for NK Cell Activation by Associating with p50 to Suppress SOCS1', *Cell Rep*, 15: 288-99.
- Zhou, W., H. Q. Dinh, Z. Ramjan, D. J. Weisenberger, C. M. Nicolet, H. Shen, P. W. Laird, and B. P. Berman. 2018. 'DNA methylation loss in late-replicating domains is linked to mitotic cell division', *Nat Genet*, 50: 591-602.
- Zhu, H., G. Wang, and J. Qian. 2016. 'Transcription factors as readers and effectors of DNA methylation', *Nat Rev Genet*, 17: 551-65.