Evolutionary usage and developmental roles of vertebrate non-methylated DNA

Hannah Katherine Long

Exeter College

UNIVERSITY OF OXFORD

Trinity Term 2013

Thesis submitted for the degree of Doctor of Philosophy.

Research Declaration

The work presented in this thesis is my own except where indicated. Any figures or results which are not my own work are indicated and the source acknowledged.

Chapter 3

The work in this chapter was performed in collaboration with Dr Neil Blackledge. All experiments are my own, apart from the Bio-CAP human amplicon spiking experiments which were performed by Dr Neil Blackledge (Department of Biochemistry, University of Oxford). MeDIP was performed by Dr Skirmantas Kriaucionis (Ludwig Institute for Cancer Research, Oxford University). *In vitro* biotinylation of the KDM2B-4 construct was performed by Dr Rob Klose.

Chapters 4 and 5

The bioinformatics analysis in Chapters 4 and 5 was performed in collaborated with Dr David Sims (CGAT, MRC Functional Genomics Unit, Department of Physiology, Anatomy and Genetics, University of Oxford).

Hannah Long

11th October 2013

Publications arising from this thesis

Long HK *, Blackledge NP* and Klose RJ. **ZF-CxxC domain-containing proteins, CpG islands, and the chromatin connection.** ***Biochemical Society transactions*** 41, 727-740 (Jun 1, 2013).

Long HK *, Sims D *, Heger A, Blackledge NP, Kutter C, Wright M, Grützner F, Odom DT, Patient R, Ponting CP, and Klose RJ. **Epigenetic conservation at gene regulatory elements revealed by non-methylated DNA profiling in seven vertebrates.** ***eLife*** 2, e00348 (2013).

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

(* co-first author)

Acknowledgements

Firstly I would like to thank my supervisors Dr Rob Klose and Professor Roger Patient for their support and encouragement throughout my DPhil. Your guidance and scientific insights were invaluable.

Thank you to all members of the Patient and Klose labs for help with experiments, for scientific discussion and for creating such a wonderful environment to work in. Thank you to Dr Philip Pinheiro for showing me the ropes of zebrafish developmental biology, and to Dr Filipa Simoes and Dr Wenchao Gu for their help with early developmental phenotypes. For early help with protein biochemistry, thank you to Dr Jin Zhou, and thank you to Dr Neil Blackledge for all of our collaborative projects.

I am grateful to Dr David Sims, Dr Andreas Heger and Professor Chris Ponting for working closely with Rob and I analysing non-methylated DNA profiles for seven vertebrates. This was a thoroughly enjoyable collaboration. I owe thanks also to the many people who provided samples for this project, and to Dr Duncan Odom and Dr Claudia Kutter for providing Tc1 mouse tissues. For bioinformatics support, I am also grateful to the Computational Biology Research Group at the Weatherall Institute of Molecular Medicine including Dr Simon McGowan, Dr Nicki Gray and Mr Stephen Taylor.

To friends in the department, from Exeter College and beyond, thank you for making my Oxford years thoroughly enjoyable and memorable.

And finally, thank you to my parents for being there throughout.

Table of contents

List of figures

List of tables

List of abbreviations

5hmC	5-hydroxymethylcytosine
5mC	5-methylcytosine
aa	amino acids
ATP	adenosine triphosphate
bp	base pair
C-terminus	carboxyl terminus
cDNA	complementary DNA
CFP1	CXXC finger protein 1
CGI	CpG island
CpG obs/exp	CpG observed over expected ratio
Da	Dalton
DMR	differentially methylated region
dNTP	deoxyribonucleotide triphosphate
dpc	days post coitum (mouse development)
dpf	days post fertilisation (zebrafish development)
DMSO	dimethyl sulfoxide
DNA	deoxyribonucleic acid
DNMT	DNA methyl transferase
Dr	*Danio rerio* (zebrafish)
E	embryonic day (mouse development)
EMSA	electrophoretic mobility shift assay
ES	embryonic stem
EZH2	enhancer of zeste homolog 2
gDNA	genomic DNA
HAT	histone acetyltransferase
HDAC	histone deacetylase
Hs	*Homo sapiens* (human)
ICR	imprinting control region
KDM	lysine demethylase
KMT	histone lysine methyltransferase
kb	kilobase
MBD	methyl-CpG binding domain
MBT	mid-blastula transition
Mm	*Mus musculus* (mouse)
N-terminus	amino terminus
NMI	non-methylated island
OD	optical density
PAGE	polyacrylamide gel electrophoresis
PCGF1	Polycomb group RING finger protein 1
PCR	polymerase chain reaction
PHD	plant homeo domain
PRC	Polycomb repressive complex
RNF2 / RING1B	ring finger protein 2
RNA	ribonucleic acid

RNAPII	RNA Polymerase II
RT	room temperature
SAH	S-adenosyl-L-homocysteine
SAM	S-adenosyl-L-methionine
SDS	sodium dodecyl sulphate
seq	sequencing
SF	short form
Tc1	Transchromosomic mouse model (containing human chromosome 21)
TE	transposable element
TET	ten-eleven translocation
ZF-CxxC	zinc finger CxxC domain
ZGA	zygotic genome activation

List of measures

SI units

M	mega	10^{6}
k	kilo	10^{3}
m	milli	10^{-3}
µ	micro	10^{-6}
n	nano	10^{-9}
p	pico	10^{-12}

Time

s	seconds
min	minutes
hr	hours

Volume and mass

g	gramme
L	litre (dm^3)
m	metre

Centrifuge

g	relative centrifugal force
rpm	revolutions per minute

Other

mol	mole
M	molar
A	ampere
V	volt
°C	degree centigrade
cpm	counts per minute
w/v	% weight per volume

Evolutionary usage and developmental roles of vertebrate non-methylated DNA

Hannah Long
Exeter College UNIVERSITY OF OXFORD
Submitted for the degree of Doctor of Philosophy
Trinity 2013

Abstract

Vertebrate genomes exhibit global methylation of cytosine residues where they occur in a cytosine-guanine dinucleotide (CpG) context and this epigenetic mark is generally thought to be repressive to transcription. Punctuating this pervasive DNA methylation landscape are short, contiguous regions of non-methylated DNA which are found at two thirds of mammalian gene promoters. These non-methylated regions exhibit CpG content close to expected levels as they escape the depletion of CpGs observed across the methylated fraction of the genome. The unique nucleotide properties of these CpG island (CGI) regions enable their identification by computational prediction in mammalian genomes. Owing to a lack of high-resolution genome-wide DNA methylation profiles in non-mammalian species, these CGI predictions have often been used as a proxy for non-methylated DNA in these organisms. In contrast to mammals, CGI predictions in cold-blooded vertebrates rarely coincide with gene promoters, leading to the belief that CGIs are significantly divergent between vertebrate species, and that unique promoter-associated features may have been acquired during warm-blooded vertebrate evolution.

This thesis is primarily concerned with the location, establishment and biological function of non-methylated islands of DNA in vertebrate genomes. To experimentally determine genome-wide profiles of non-methylated DNA, a novel biochemical technique was established called biotinylated ZF-CxxC affinity purification (Bio-CAP), and development of this method is discussed in Chapter 3. Experimental analysis of non-methylated DNA profiles in this thesis initially addresses two main questions: (1) 'How does the non-methylated DNA landscape compare genome-wide for seven vertebrates considering distinct tissue types and developmental stages?' (2) 'How are vertebrate non-methylated regions of DNA defined and interpreted in the nuclear environment?'

To address the first question, non-methylated DNA was profiled by Bio-CAP sequencing across the genomes of seven diverse vertebrate species, representing all major branch points of vertebrate evolution, and the results are discussed in Chapters 4 and 5. Contrary to previously held dogma, experimentally determined non-methylated islands of DNA (NMIs) constitute an ancient epigenetic feature of vertebrate gene regulatory elements. However, despite having numerous high-resolution maps of vertebrate non-methylated DNA, the means by which NMIs are identified and maintained in the nuclear environment remains poorly understood. To address the second question and identify features which determine the methylation state of DNA, exogenous DNA sequences were introduced into mouse embryonic stem (ES) cells. Non-methylated DNA was profiled by Bio-CAP sequencing to investigate how different features, such as sequence-specific binding motifs, chromatin architecture and nucleotide composition of a given DNA sequence impact local DNA methylation patterns. Interestingly, the majority of exogenous promoters were appropriately non-methylated in mouse ES cells, germline and somatic cells suggesting that gene promoters have retained strong signals for the non-methylated state across millions of years of evolution (discussed in Chapter 6).

During mouse embryogenesis, genome-scale DNA demethylation and remethylation events occur to remodel the epigenetic landscape and loss of DNA methylation during this time leads to embryonic lethality. To investigate the biological function of non-methylated DNA, the third question addressed in this thesis is (3) 'What is the developmental importance of non-methylated islands of DNA during vertebrate embryogenesis?' To investigate this, members of the ZF-CxxC domain-containing family of chromatin modifiers were ablated in zebrafish embryos to perturb the chromatin landscape at NMIs, and therefore interfere with their function during early development (Chapter 7). Early embryonic development and patterning was disrupted in knock-down embryos, suggesting that interpretation of non-methylated DNA and placement of chromatin modifications at NMIs is essential for normal zebrafish embryogenesis. Together this work sheds light on the evolutionary origins of NMIs, the mechanisms involved in the recognition and establishment of non-methylated loci and provides an insight into the function of non-methylated DNA during early embryonic development.

Chapter 1 Introduction

1.1 **Epigenetics and chromatin**

In multicellular organisms, all cells have essentially the same genomic sequence to that of the totipotent zygote from which they were derived. Despite this, distinct terminally-differentiated cell-types in an organism exhibit extremely diverse morphology, gene expression profiles and function. The mechanisms by which one cell at fertilisation can give rise to numerous diverse cell-types must therefore rely on a non-genetic component, which can integrate both developmental signals and inputs from the environment. The term 'epigenetics', first coined by the embryologist Conrad Waddington in 1942 (Waddington, 2012), has been used to describe this phenomena or "the study of mitotically and/or meiotically heritable changes in gene function that cannot be explained by changes in DNA sequence" (Russo et al., 1996). Since the first use of the term there have been many diverse definitions of epigenetics, although in general an epigenetic mechanism is thought to be heritable, self-propagating and reversible (Riddihough and Zahn, 2010).

The term 'chromatin' is used to describe the complex of DNA in a eukaryotic nucleus with its associated proteins. DNA therefore does not exist free in the eukaryotic nucleus, but is condensed through the coordination with histone proteins to form a chromatin fibre. Chromatin generally exists in two main states in the cell, first observed by differential condensation of chromosomal regions in moss nuclei (Heitz, 1928). Decondensed euchromatin is generally associated with gene-rich regions of the genome which are transcriptionally active. Conversely, heterochromatic regions of the genome appear cytologically more condensed and are associated with non-coding or repetitive portions of the genome which tend to be transcriptionally inert. A more recent definition of epigenetic events therefore incorporates an appreciation of the relationship between chromatin architecture and gene expression status, such that epigenetic changes involve: "the structural adaptation of chromosomal regions so as to register, signal or perpetuate altered activity states" (Bird, 2007). Overall, the term epigenetics has now come to refer to many non-genetic regulatory

systems, including non-coding RNAs, chromatin architecture and the modification of DNA and histones.

1.2 **The modification of DNA and histones**

An octamer of histone proteins forms the core of the nucleosome structure which coordinates and compacts DNA into the eukaryotic nucleus. Both DNA and histones are subject to chemical modifications, which can impact the functionality and accessibility of the underlying genetic material.

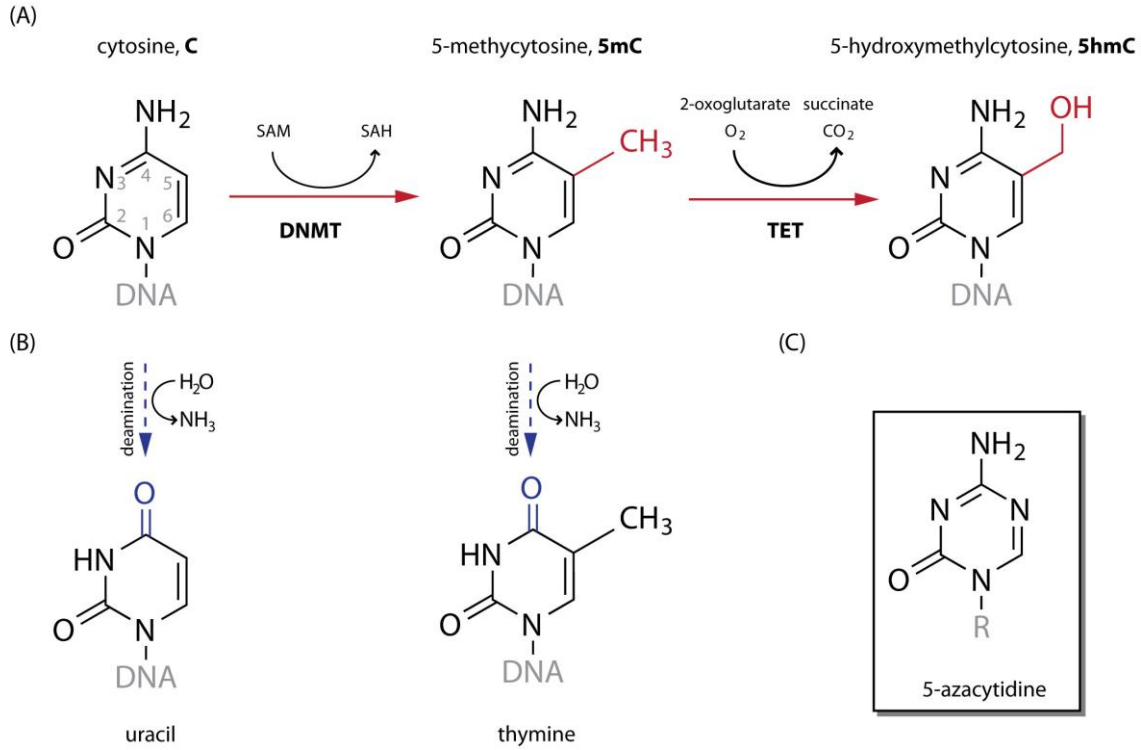


Figure 1.1 Methylation and deamination reactions of the cytosine base

(A) Cytosine residues can be methylated on the C5-position of the pyrimidine ring by DNA methyltransferase enzymes (DNMTs) with the methyl group provides by the methyl donor S-adenosylmethionine (SAM). The resulting 5-methylcytosine can be oxidised by TET enzymes to yield 5-hydroxymethylcytosine). (B) Both cytosine and 5-methylcytosine are subject to mutagenic deamination. Deamination converts cytosine to uracil and 5mC to thymine. If the resultant base pair mismatch is not recognised and repaired prior to DNA replication, this can lead to a C/G to T/A point mutation. (C) Chemical structure of 5-aza analogue of cytosine, 5-azacytidine. 5-azacytidine can be incorporated into DNA where it inhibits DNA methylation. R=ribose.

1.2.2 Histones and histone modifications

The histone octamer consists of two dimers of H2A:H2B and H3:H4 and wraps approximately 147 bp of DNA twice around its core (Peterson and Laniel, 2004). Each histone protein extends an

unstructured N-terminal tail from the globular nucleosomal structure, which can be post-translationally modified by chromatin-modifying enzymes. These modifications include methylation, acetylation, phosphorylation, ubiquitylation, sumoylation, ADP-ribosylation (Strahl and Allis, 2000) and biotinylation (Camporeale et al., 2004) which can affect the compaction of nucleosomes directly or lead to the recruitment of non-histone proteins to chromatin (Kouzarides, 2007). These modifications, in addition to variant histone proteins, play important roles in many cellular processes including gene regulation, the DNA damage response, chromosome segregation and transcriptional termination (Talbert and Henikoff, 2010).

1.2.3 Modification of DNA

DNA, the genetic material of inheritance, is composed of a double-stranded helix of two complementary phosphate-deoxyribose backbones bridged by the four bases adenine, cytosine, guanine and thymine. The fifth base uracil, normally only incorporated into RNA, is only detected in DNA as a deamination product of cytosine (Figure 1.1B). Despite the presence of only 5 nucleotides, chemical modification of DNA is known to add another layer of complexity to the DNA sequence. Across all kingdoms of life, the most frequent modification of DNA is the methylation of cytosine on the C5 position of the pyrimidine ring (5mC) (Figure 1.1A), which was first identified by biochemical analysis of DNA (Hotchkiss, 1948; Johnson and Coghill, 1925). The majority of DNA methylation in eukaryotes occurs at CpG dinucleotides and the symmetric nature of this site allows DNA methylation to be propagated following semi-conservative DNA replication which lead to the hypothesis in 1975 that DNA methylation could act as an epigenetic mark in vertebrates (Holliday and Pugh, 1975; Riggs, 1975). DNA methylation in eukaryotes has been implicated in many important cellular processes including differentiation, gene regulation, DNA repair and embryonic development (Bird, 2002; Challen et al., 2012; Cuozzo et al., 2007; Jones, 2012; Reik et al., 2001). In prokaryotes, DNA methylation can also occur on the N4 position of cytosine and N6 position of adenine and here DNA methylation has been implicated in DNA repair and defence against foreign DNA via a host-restriction mechanism (Wilson and Murray, 1991). More sensitive biochemical

analyses have recently identified other DNA modifications including oxidised derivatives of 5mC, such as 5-hydroxymethylation, 5-formylcytosine and 5-carboxylcytosine (5hmC, 5fC and 5caC; Figure 1.1A) (He et al., 2011b; Ito et al., 2011; Tahiliani et al., 2009).

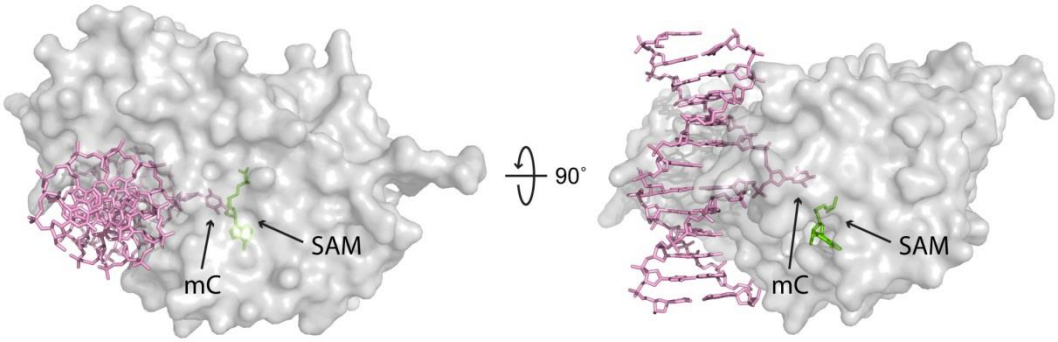


Figure 1.2 Bacterial HhaI methylates cytosine via a base-flipping mechanism

A crystal structure of a chemically-trapped covalent reaction intermediate between the bacterial DNA methyltransferase M. HhaI, a 13-mer DNA oligonucleotide which contains a methylated 5-fluorocytosine at its target position and the cofactor S-adenosyl-L-homocysteine (Klimasauskas et al., 1994). The target cytosine is flipped out from the DNA helix (pink) and wedged into a binding pocket in close proximity to the S-adenosyl-L-homocysteine (SAM, green). Images were generated in PyMol (Protein Data Bank ID: 1MHT). Two views are shown, looking down the axis of the DNA double helix (left) and a view perpendicular to this (right).

1.3 **The DNA methylation machinery**

DNA methylation is catalysed by the enzymatic activity of two classes of DNA methyltransferase (DNMT) enzymes with the methyl group provided by the methyl donor S-adenosylmethionine (SAM) (Figure 1.1, C to 5mC). The first eukaryotic methyltransferase to be identified was the maintenance DNMT1 enzyme from homology to known bacterial enzymes (Bestor et al., 1988). DNMT1 is able to recognise hemi-methylated DNA following DNA replication (Chen and Li, 2006; Goll and Bestor, 2005) or during DNA repair (Mortusewicz et al., 2005) and faithfully copy the methyl mark onto the second DNA strand. A second class of *de novo* DNMT3 methyltransferases were later identified which are able to place DNA methylation on unmodified DNA.

1.3.1 The mechanism of DNA methylation

The first insights into the catalytic mechanism of DNA methylation came from structural studies of the bacterial M.HhaI cytosine methyltransferase (Klimasauskas et al., 1994). The target cytosine base was observed to be extruded out from the DNA double helix, disrupting base-pairing interactions via

a base-flipping mechanism (Figure 1.2). This eversion of the target cytosine base facilitates the formation of a transient covalent protein-nucleotide complex, similar to that observed for the thymidylate synthetase enzyme (Wu and Santi, 1987), and transfer of a methyl group from the methyl donor SAM onto the target cytosine. The substrate cytosine sits deep in a hydrophobic pocket in the large domain of M.HhaI which forms the active site and the orphan guanine is stabilised by interaction with glutamine residues from the small domain of the enzyme (Klimasauskas et al., 1994).

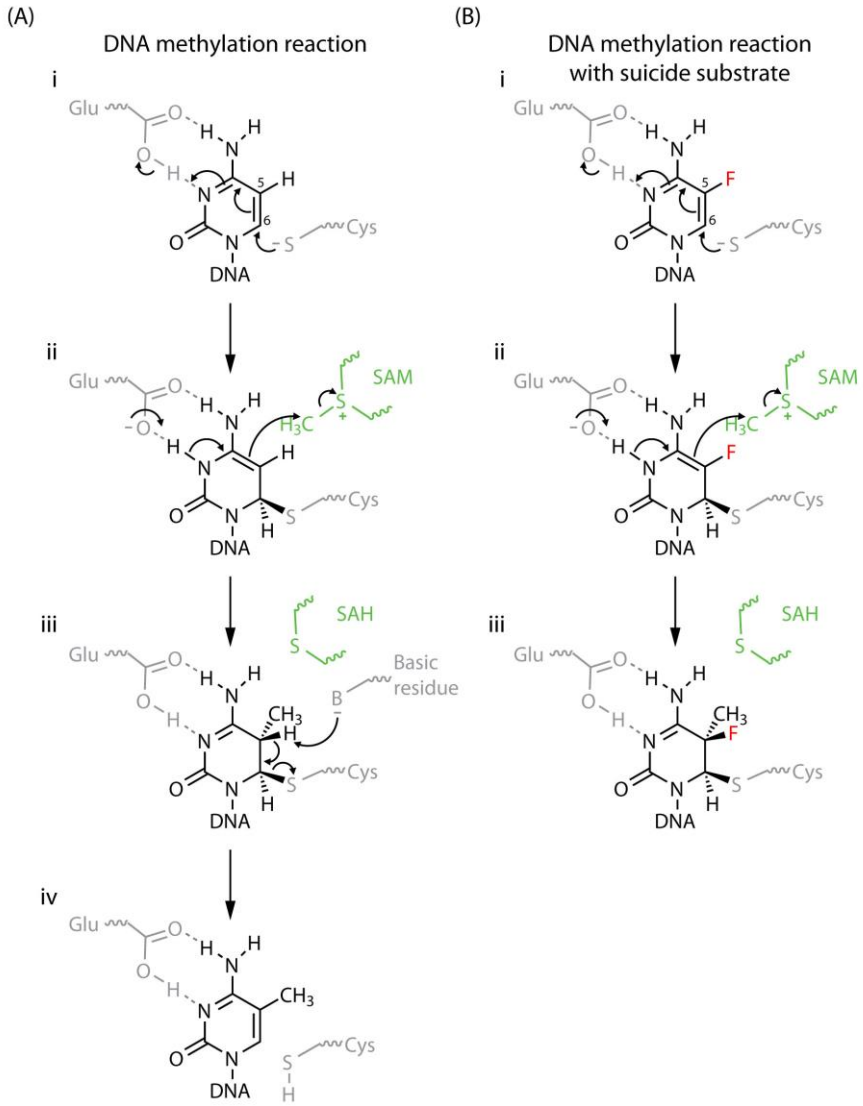


Figure 1.3 Mechanism of DNA methylation

The mechanism proposed for the DNA methylation reaction for (A) a cytosine base and (B) a so-called suicide substrate 5-fluoro-2′-deoxycytidine with fluorine at the C5 position of the target cytosine. (A i) Nucleophilic attack of the C6 position of cytosine by a reactive cysteine thiol group creates a transient covalent cysteine-cytosine covalent bond in the active site. (A ii) A methyl group is then transferred from SAM by electrophilic attack at the C5 position of cytosine. (A iii-iv) A proton is then lost from cytosine C5

and the covalent intermediate released. (B) The substrate 5-fluoro-2'-deoxycytidine cannot release F+ therefore the reaction intermediate is trapped in the active site along with the demethylated SAH. Figure was adapted from a number of sources (Goll and Bestor, 2005; Ryazanova et al., 2012; Zangi et al., 2010) and https://www.google.com/patents/US5856090.

Nucleophilic attack of cytosine position C6 by a reactive cysteine thiol group in the enzymatic active site creates a cysteine-cytosine covalent bond (Figure 1.3Ai). This then permits electrophilic attack by cytosine C5 on the methyl group from the methyl-donor SAM (Figure 1.3Aii), followed by release of the C5 proton, production of SAH (S-adenosyl-L-homocysteine, Figure 1.3Aiii) and elimination of the covalent intermediate of a reactive cysteine (Figure 1.3Aiv) (Cheng, 1995; Wu and Santi, 1987). Importantly, active site cysteine mutants lead to loss of function for a number of DNA methyltransferase enzymes supporting this proposed mechanism (Chen et al., 1993; Mi and Roberts, 1993; Wyszynski et al., 1992). Furthermore, the active site cysteine residue has been identified covalently attached to DNA with the suicide substrate 5-fluoro-2'-deoxycytidine (Chen et al., 1991; Friedman and Ansari, 1992; Hanck et al., 1993; Smith et al., 1992). In this trapped state, the methyl group has transferred to the C5 cytosine (Figure 1.3Bii) but the elimination step cannot take place as F^+ cannot be released (Figure 1.3Biii). Use of similar suicide DNA substrates has enabled determination of crystal structures of trapped DNA methyltransferase bound to DNA. For example, the structure of the bacterial C5-cytosine DNA methyltransferase M.HhaI was co-crystallised covalently linked to a 13-mer oligonucleotide containing a 5-fluoro-2'-deoxycytidine at the target base (Figure 1.2) (Klimasauskas et al., 1994).

Base-flipping has also very recently been observed for a trapped mouse DNMT1 enzyme covalently linked to a hemi-methylated DNA substrate via a reactive cysteine in the enzyme active site (Song et al., 2012b). Similar to the bacterial enzyme, the cytosine on the target strand was flipped out and anchored within the catalytic pocket of the enzyme. Both the major and minor grooves of the DNA fragment were interrogated by catalytic and recognition loops from DNMT1 with a methionine side chain from the catalytic loop occupying the space left by the extruded cytosine (Song et al., 2012b). The methyl group from the non-substrate 5mC was positioned in a hydrophobic concave surface of

DNMT1, and together these interactions may be important for determining the substrate specificity of the enzyme.

1.3.2 Inhibition of the DNMT enzymes using 5-azacytosine

Compounds were identified in the 1960s which were able to inhibit DNA methylation and lead to aberrant differentiation of cells in culture (Taylor and Jones, 1979). The ability to cause DNA hypomethylation was limited to cytosine analogues with modifications at the 5-position of the pyrimidine ring and their ability to incorporate into DNA and target DNMTs for degradation (Jones and Taylor, 1980; Yoo and Jones, 2006). Incorporation of cytosine analogues, such as 5-azacytosine (Figure 1.1C), into DNA causes passive loss of DNA methylation in a DNA replication-dependent manner as the base cannot be methylated (Glover and Leyland-Jones, 1987). Resultant hypomethylation can lead to reactivation of genes (Issa and Kantarjian, 2009), and therefore cytosine analogues are being investigated as potential therapeutics to reactivate genes silenced during tumorigenesis (Beumer et al., 2008).

1.3.3 The maintenance DNA methylation machinery

The mammalian maintenance DNA methyltransferase DNMT1 is a large protein which encodes many N-terminal regulatory domains including a replication foci-targeting sequence (RFTS), a ZF-CxxC DNA-binding domain and a pair of bromo-adjacent homology domains (BAH1 and BAH2), with a catalytic domain situated at the C-terminus of the protein (Figure 1.4). All eukaryotes which perform DNA methylation, with the exception of a lineage of ascomycete fungi, possess a DNMT1 enzyme (Figure 1.6) (Zemach and Zilberman, 2010). Following semiconservative DNA replication, hemimethylated DNA is inherited by both daughter strands and DNMT1 conducts the majority of maintenance methylation to faithfully restore symmetrical methylation at CpG dinucleotides (Iida et al., 2002). In keeping with its role in maintenance methylation, DNMT1 exhibits a 30 to 50-fold greater efficiency at catalysing the addition of a methyl group to hemi-methylated as opposed to unmodified CpG dinucleotides (Bestor and Ingram, 1983; Chuang et al., 1997). This specificity does

not seem to be absolute however, as DNMT1 has been reported to direct *de novo* methylation at both CpG and non-CpG sites *in vitro* and *in vivo* (Fatemi et al., 2002; Gowher et al., 2005b; Grandjean et al., 2007; Song et al., 2011). DNMT1 is observed to relocate to replication foci as cells enter S phase (Leonhardt et al., 1992; Margot et al., 2001), and has been proposed to associate with PCNA (proliferating cell nuclear antigen) at replication forks via its RFTS domain (Chuang et al., 1997). However, disruption of this interaction only led to a small reduction in DNA methylation levels suggesting that DNMT1 is targeted to replication foci by additional mechanisms (Schermelleh et al., 2007; Spada et al., 2007). Recently, another interaction partner UHRF1 (ubiquitin-like plant homeodomain and RING finger domain 1) was shown to be required for targeting DNMT1 to chromatin, with UHRF1 depletion causing severe reduction in DNA methylation (Bostick et al., 2007; Sharif et al., 2007).

Despite a lack of sequence or structural similarity with DNMT enzymes, UHRF1 domain appears to utilise a similar base-flipping mechanism to bind hemi-methylated DNA via a SRA (SET and RING finger associated) domain (Arita et al., 2008; Avvakumov et al., 2008; Bostick et al., 2007; Hashimoto et al., 2008; Liu et al., 2013; Qian et al., 2008; Sharif et al., 2007). Two loops from UHRF1 interrogate both the major and minor grooves, with the flipped base binding in a concave pocket of the protein (Arita et al., 2008; Hashimoto et al., 2008). This mode of binding may act to clearly distinguish the parental methylated strand from the unmethylated daughter and present the latter to DNMT1. However, crystal structures reveal that DNMT1 should be unable to catalyse DNA methylation on UHRF1-bound DNA (Arita et al., 2012; Song et al., 2011), suggesting that the unmethylated cytosine may be transferred to DNMT1 for maintenance methylation. A recent study has suggested that UHRF1 may indirectly recruit DNMT1 to newly replicated DNA by catalysing ubiquitylation of the histone H3 tail (H3K23) which is subsequently recognised by the DNMT1 RFTS domain (Nishiyama et al., 2013).

Further structural studies of isolated DNMT1 fragments have provided evidence that the enzyme possesses mechanisms for auto-regulation. When bound to non-methylated CpG DNA, the ZF-CxxC

domain of DNMT1 has been reported to occlude access of the active site to DNA via insertion of an 'auto-inhibitory linker' (Song et al., 2011). However, structures of larger DNMT1 fragments suggest that in fact the RFTS domain can insert into the DNMT1 DNA-binding pocket and play an inhibitory role that prevails over the 'auto-inhibitory linker' implicated from the previous study (Syeda et al., 2011; Takeshita et al., 2011). Therefore, binding of the DNMT1 RFTS domain either directly to UHRF1 (Achour et al., 2007; Leonhardt et al., 1992) or to its chromatin mark H3K23ub (Nishiyama et al., 2013) may alleviate the auto-repression of DNMT1 and stimulate maintenance DNA methylation activity at newly replicated DNA.

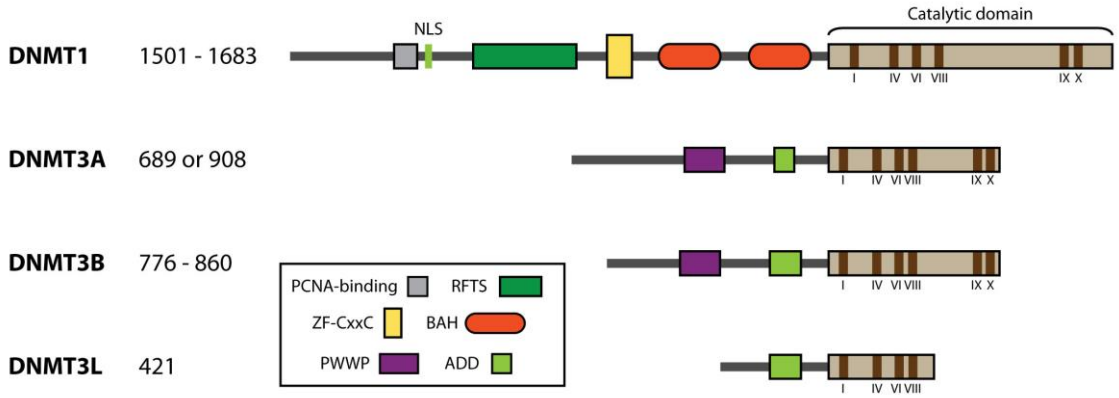


Figure 1.4 Domain architecture of mammalian DNA methyltransferase enzymes

An illustration of the domain architecture of the mammalian maintenance (DNMT1) and *de novo* (DNMT3A and DNMT3B) DNA methyltransferases and the helper DNMT3-like (DNMT3L) which lacks catalytic activity. Proteins are drawn to scale, and are adapted from (Ryazanova et al., 2012), with the number of amino acids indicated on the left for the mouse proteins, with both or a range of lengths indicated if two or more isoforms have been identified (number of amino acids taken from NCBI). RFTS, replication focus targeting sequence; BAH, bromo-adjacent homology; PWWP, domain name is derived from a conserved Pro-Trp-Trp-Pro motif; ADD, ATRX-DNMT3-DNMT3L. The proteins are shown with the N-terminus on the left.

1.3.4 The *de novo* DNA methylation machinery

In mammals, the *de novo* methylation machinery consists of two active enzymes DNMT3A and DNMT3B and a regulatory factor DNMT3-like (DNMT3L) which together are responsible for the establishment of genomic patterns of DNA methylation (Bird and Wolffe, 1999; Okano et al., 1999; Xu et al., 1999). DNMT3A and DNMT3B exhibit similar domain architecture, with an N-terminal variable region, followed by a PWWP domain, a Cys-rich 3-Zn-binding domain called ADD (ATRX-DNMT3-DNMT3L) and a C-terminal catalytic domain containing six highly conserved C5-DNA methyltransferase motifs (Figure 1.4). The DNMT3L amino acid sequence is very similar to DNMT3A

and DNMT3B in the Cys-rich 3-Zn-binding domain, but lacks conservation in the C-terminal domain required for catalytic activity (Figure 1.4). DNMT3A and DNMT3B have been shown to methylate CpG sites with no preference for hemi-methylated DNA (Okano et al., 1998a) and have some detectable activity towards non-CpG sites (Gowher and Jeltsch, 2001). Cells lacking DNMT3A and DNMT3B slowly lose genomic methylation perhaps due to lack of fidelity of the maintenance machinery (Chen et al., 2003). Restoration of DNA methylation in these cells by replacement of the DNMT3 enzymes verifies the role of these enzymes as *de novo* DNA methyltransferases (Chen et al., 2003).

The third DNMT3 family member, DNMT3L, is unable to bind SAM and lacks enzymatic activity (Figure 1.4) (Gowher et al., 2005a; Kareta et al., 2006). DNMT3L has been co-immunoprecipitated with the C-terminal portion of both DNMT3A and DNMT3B (Hata et al., 2002) via its own C-terminal domain (Chen et al., 2005). Interaction with DNMT3L is known to increase the enzymatic activity of both DNMT3A and DNMT3B by around 15 fold (Chedin et al., 2002; Chen et al., 2005; Gowher et al., 2005a) suggesting that DNMT3L may stimulate DNA methyltransferase activity as part of its regulatory action.

No crystal structures are yet available for the mammalian *de novo* class of enzyme bound to DNA, however a structure of a DNMT3A and DNMT3L tetramer has shed light into the mechanism by which DNMT3A is stimulated by DNMT3L and suggests a possible mechanism of action for the *de novo* methyltransferases. In this structure, C-terminal DNMT3L fragments were observed to bind the DNMT3A catalytic domain. DNMT3A and DNMT3B have the smallest active sites of the known DNA methyltransferases, and dimerisation of DNMT3A observed in the crystal structure doubles the available surface area by bringing two active sites together (Jia et al., 2007). By superimposing this structure onto that for M.HhaI (Klimasauskas et al., 1994), DNA modelled into the DNMT3A dimer suggested that the two active sites are localised in the major groove of DNA approximately one helical turn apart. Therefore, two CpGs separated by 8-10 base pairs could potentially be methylated simultaneously. The coordination of DNA methylation in this manner has been proposed to play a

role in recognition and methylation of differentially methylated regions (DMRs) at imprinted genes (Jia et al., 2007). Other reported interactions of DNMT3L and DNMT3A/B with chromatin (discussed in more detail in section 1.8), may act to target DNA methylation to appropriate regions of the genome (Chen et al., 2004; Dhayalan et al., 2010; Ooi et al., 2007).

1.3.5 DNMT2

DNMT2 exhibits high sequence and structural similarity to DNA methyltransferases (Dong et al., 2001; Okano et al., 1998a; Yoder and Bestor, 1998) and homologues have been identified in plants, animals and fungal species (Zemach and Zilberman, 2010). DNMT2 exhibits a weak cytosine methyltransferase activity *in vitro* (Hermann et al., 2003), however no detectable effect on global DNA methylation was observed in ES cells with a targeted deletion of the *Dnmt2* gene (Okano et al., 1998b). DNMT2 has since been found to methylate cytosine 38 in the anticodon loop of $tRNA^{Asp}$ via a similar mechanism to that proposed for DNA methyltransferases (Goll et al., 2006; Rai et al., 2007). It has been suggested that DNMT2 evolved from an ancient DNA methyltransferase which has switched its substrate preference to RNA and now has little involvement in global *de novo* or maintenance DNA methylation (Iyer et al., 2011; Jeltsch et al., 2006; Jurkowski and Jeltsch, 2011). For a number of species, including *Drosophila melanogaster* and *Caenorhabditis elegans*, DNMT2 represents the only putative DNMT present in the genome and whole genome bisulfite sequencing has determined that no defined methylation patterns can be detected in DNMT2-only genomes (Raddatz et al., 2013). Despite this finding, it has been suggested that the *Drosophila* DNMT2 orthologue is able to methylate DNA during a small window in early development as a tiny fraction of methylated cytosines (0.3%) were detected in the *Drosophila* embryo. However, these observations remain contentious (Kunert et al., 2003; Lyko et al., 2000).

1.4 **DNA methylation profiles**

DNA methylation is present in some, but not all, eukaryotic species where it plays important roles in silencing transposable elements for genome-defence and regulating gene expression via the

methylation of gene promoters and bodies. DNA methylation profiles, and total levels of DNA methylation, are not strictly conserved across eukaryotes as invertebrate species appear to have lost the requirement for DNA methylation in silencing repeats while gene body methylation is not detectable in fungi. The DNA methylation landscapes of the major eukaryotic clades will be discussed below.

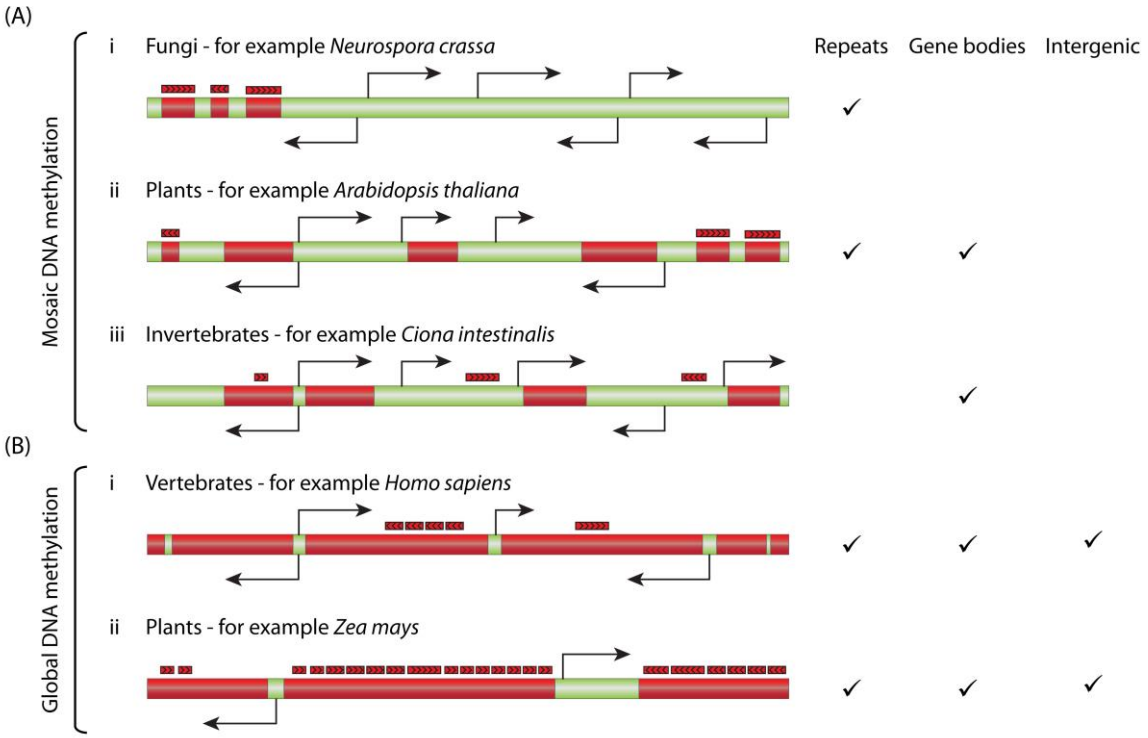


Figure 1.5 DNA methylation patterns in fungi, plants and animals

(A) (i) Mosaic DNA methylation is observed in the filamentous fungi *Neurospora crassa* where non-methylated regions (green) are interspersed with methylated sequences (red) with DNA methylation targeted to transposable elements (red boxes). (ii) The plant *Arabidopsis thaliana*, which has a small genome, also exhibits mosaic methylation with DNA methylation targeted to gene bodies and repetitive sequences. (iii) Mosaic methylation is also characteristic of invertebrates such as *Ciona intestinalis*, in which around 50 % of gene bodies are methylated, although the majority of transposable elements remain unmethylated. (B) (i) Global methylation is a feature of all vertebrate genomes, with non-methylated DNA found at CpG islands. Transposable elements, gene bodies and intergenic DNA are methylated. (ii) Plants with large genomes also appear to exhibit global DNA methylation, with genes separated by long tracts of transposable elements and their relics (SanMiguel et al., 1996) and gene bodies also tend to be methylated. Adapted from (Suzuki and Bird, 2008).

1.4.1 Fungi

The two most well-studied fungi, *Saccharomyces cerevisiae* (budding yeast) and *Schizosaccharomyces pombe* (fission yeast) exhibit no DNA methylation (Figure 1.6) (Becker et al., 2012; Proffitt et al., 1984), and the sole DNMT homologue, DNMT2, has been characterised as a

tRNA methyltransferase in *S. pombe* (Becker et al., 2012). Other fungi have been shown to exhibit low levels of DNA methylation which appears to be patterned across the genome (Figure 1.5 and Figure 1.6) (Antequera et al., 1984; Tang et al., 2012; Zemach et al., 2010). The common ancestor of all fungi appears to have lost the DNMT3 homologue as no fungal genomes have been identified to date with a DNMT3 orthologue (Zemach et al., 2010; Zemach and Zilberman, 2010). *P. blakesleeanus* for example encodes a DNMT1-like and a dim-2-like DNA methyltransferase which methylates cytosine in any sequence in context (Kouzminova and Selker, 2001; Selker et al., 2003), and has substantial amounts of DNA methylation at CG sites and to a lesser extent in non-CG contexts at CHG and CHH (where H is A, C or T). DNA methylation is targeted to transposable elements (TEs) and other repeats for silencing while genes remain unmethylated (Zemach et al., 2010).

A lineage of ascomycetes, or filamentous fungi, including *Uncinocarpus reessii* and *Neurospora crassa* have lost the DNMT1-like methyltransferase (Figure 1.6) and therefore appear to solely rely on dim-2 directed *de novo* methyltransferase activity for DNA methylation levels. The genome of *N. crassa* is almost completely devoid of DNA methylation except at very densely methylated loci which correspond to relics of RIP (repeat-induced point mutation) (Singer et al., 1995) at transposon relics (Selker et al., 2003). RIP detects duplicated DNA sequences prior to meiosis as a means of genome defence, and induces numerous C/G to T/A mutations dependent on the RID (RIP defective) protein, a non-catalytic DNMT1 homologue (Singer et al., 1995; Zemach and Zilberman, 2010). These regions are also targeted for DNA methylation by dim-2, the establishment of which is dependent on the H3K9 histone methyltransferase dim-5 (Tamaru and Selker, 2001; Tamaru et al., 2003) and the Hp1-like protein that recognises this histone modification (Freitag 2004). Therefore, with the exception of a small amount of gene body methylation detected for *U. reessii*, all species of fungi analysed so far appear to preferentially methylate transposable elements (TEs) as a means of genome defence to prevent the reactivation of parasitic genomic regions (Lewis et al., 2009).

1.4.2 Plants

Similar to fungi, plants have the capacity to methylate DNA at CpG and non-CpG sites, including symmetric CHG and asymmetric CHH positions (where H is A, G or T). *De novo* methylation is catalysed by DRM2 (domains rearranged methyltransferase 2), a homologue of the DNMT3 methyltransferases and is maintained by three different enzymatic activities depending on the methylation site. CG methylation is maintained by the plant DNMT1 homologue MET1 (or DMT1, DNA methyltransferase 1), CHG methylation is maintained by CMT3 (chromomethylase 3), a plant-specific methyltransferase proposed to be from the same family as the fungal dim-2 (Zemach and Zilberman, 2010), and CHH methylation is maintained by continued *de novo* placement by DRM1/2 (Law and Jacobsen, 2010). *De novo* methylation is targeted by short RNA molecules to homologous DNA sequences in retrotransposons and endogenous repeats via a process called RNA-directed DNA methylation (RdDM). RdDM requires the RNAi (RNA interference) machinery, two plant-specific RNA polymerases (PolIV and PolV) DRM1 and two chromatin remodellers (Matzke et al., 2009).

DNA methylation patterns have been determined at single nucleotide resolution for *Arabidopsis thaliana* (Cokus et al., 2008) and at lower resolution, by shotgun genomic bisulfite sequencing, in a number of other plant species (Feng et al., 2010; Zemach et al., 2010). Cytosine methylation was detected in all three sequences contexts, with non-CpG context DNA methylation detected at a much lower frequency in the genome than CpG methylation. In *Arabidopsis*, cytosine methylation levels are around 24, 6.7 and 1.7 % for CG, CHG and CHH contexts respectively (Cokus et al., 2008). Cytosine methylation, in all sequence contexts, is found predominantly at transposons and repetitive elements in the three flowering plants *Arabidopsis*, rice and poplar (Feng et al., 2010; Zemach et al., 2010). One-third of genes exhibit CpG (but not non-CpG) DNA methylation in their coding regions (Cokus et al., 2008; Feng et al., 2010; Lister et al., 2008; Zemach et al., 2010; Zhang et al., 2006; Zilberman et al., 2007) (Figure 1.5). The relationship between gene body methylation and transcription is unclear. It has been proposed that DNA methylation may silence antisense cryptic promoters (Tran et al., 2005; Zilberman et al., 2007), although increased antisense transcripts were

rare in *met1* knockout plants (*met1* maintains gene body methylation in wildtype plants) (Zhang et al., 2006). Two early diverging land plants, *Selaginella moellendorffii* and *Physcomitrella patens*, exhibit CG, CHG and CHH methylation of TEs, but exhibit little gene body or promoter methylation reminiscent of fungal methylomes (Figure 1.5Ai) (Zemach et al., 2010). Plants therefore appear to exhibit variable usage of genomic DNA methylation. Plants with large genomes, such as maize, are an extreme example of this as large tracts of DNA containing TEs dominate intergenic regions and are targeted for DNA methylation (Figure 1.5Bii) (Palmer et al., 2003; SanMiguel et al., 1996).

1.4.3 Invertebrates

Animals species appear to have lost the ancestral CMT1/DIM-2 DNA methyltransferase enzyme, therefore the vast majority of DNA methylation is found at CpG dinucleotides in these species (Figure 1.6) (Zemach and Zilberman, 2010). Similar to plants, invertebrate species which have DNA methylation exhibit fractional or mosaic patterns of DNA methylation, with relatively long tracts (tens of kb) of alternating methylated and non-methylated DNA (Okamura et al., 2010). The sea squirt, an ascidian, for example contains roughly the same proportion of methylated and non-methylated regions of DNA (Figure 1.5A) (Simmen et al., 1999). Generally, DNA methylation in invertebrates is found in gene bodies. It has been suggested that DNA methylation may be recruited to gene bodies during the process of splicing and interestingly, DNA methylation is enriched at exons in the honey bee and sea squirt genomes (Feng et al., 2010; Zemach and Zilberman, 2010). Transposable elements are generally hypomethylated in invertebrates, therefore there is little evidence that DNA methylation in invertebrates inhibits transcription or silences TEs (Zemach et al., 2010). However, the origin of gene body methylation appears to be primordial as it is shared between plants and animals, despite being lost in fungi. Interestingly, a number of invertebrate species, including the well-studied model organisms *D. melanogaster* and *C. elegans*, have lost their DNMT1 and 3 gene orthologues (Zemach and Zilberman, 2010) and generally lack detectable DNA methylation patterns (Raddatz et al., 2013) (see 1.3.5).

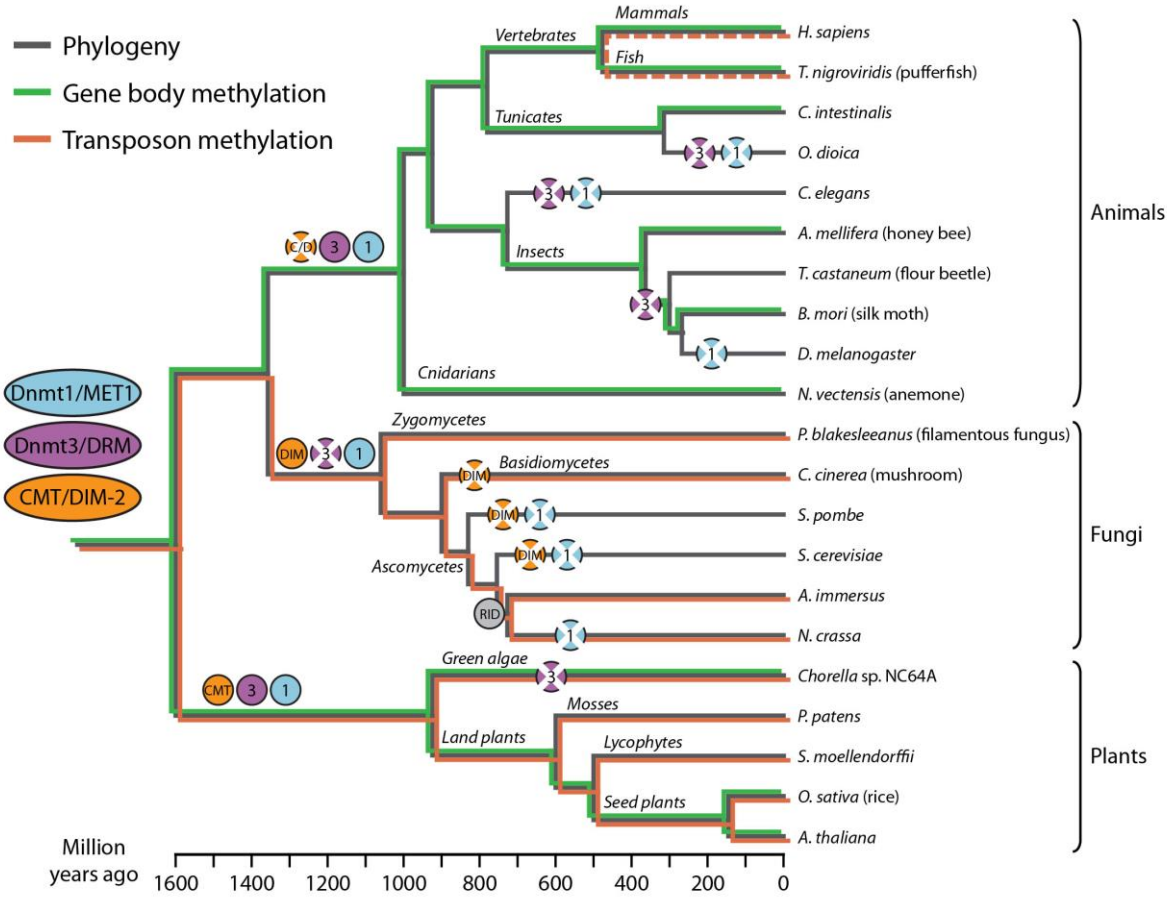


Figure 1.6 The evolution of eukaryotic DNA methylation

A phylogenetic tree illustrating the evolution of eukaryotic DNA methyltransferases and the associated profiles of DNA methylation. Representative species are shown from the animal, fungi and plant kingdoms. Circles with 1, 3, CMT, DIM and C/D represent DNMT1, DNMT3, CMT, Dim-2 and CMT/Dim-2 respectively. A white cross represents loss of the gene family in a given lineage. Green and red lines represent the presence of gene body and transposable element methylation in these species respectively. A scale bar indicates the when branch-points are estimate to have occurred during evolution. Adapted from (Zemach and Zilberman, 2010).

1.4.4 Vertebrates

At the evolutionary transition from invertebrates to vertebrates, DNA methylation profiles change drastically, with cytosine methylation becoming highly pervasive throughout the genome. This is proposed to have occurred early during vertebrate evolution as the early vertebrates lamprey and hagfish exhibit vertebrate-like DNA methylation characteristics (Tweedie et al., 1997). The emergence of pervasive DNA methylation may have been beneficial due to an enhanced innate immunity through the capacity to differentiate self from CpG-rich non-methylated bacterial pathogen DNA (Bird, 1995; Suzuki and Bird, 2008). In humans around 70-80 % of CpG cytosines are

methylated (Ehrlich et al., 1982) with both gene bodies and repetitive elements being subject to high levels of DNA methylation (Figure 1.5B).

1.5 **Non-methylated DNA in vertebrate genomes**

1.5.1 Experimental identification of short non-methylated regions of DNA

Counteracting the pervasive DNA methylation across vertebrate genomes are contiguous regions of non-methylated DNA. First identified as a fraction of the mouse genome susceptible to digestion by the methyl-CpG sensitive restriction enzyme HpaII, these regions were named HpaII tiny fragments (HTFs). DNA sequencing later revealed that, in addition to exhibiting an elevated GC content, these short released fragments contained clustered CpG sites (Bird et al., 1985; Bird, 1987; Cooper et al., 1983) and were therefore renamed CpG islands (CGIs) (Bird, 1987; Gardiner-Garden and Frommer, 1987). The tight association with gene promoters suggested that CGIs may be involved in regulating gene expression (Antequera and Bird, 1993).

1.5.2 Predicting CpG island location using computational algorithms

Due to the unique nucleotide properties of CGIs, algorithms were developed to identify genomic regions which exhibit CGI-like properties. The first CGI prediction algorithm was proposed by Gardiner-Garden and Frommer and uses a sliding window approach to identify contiguous regions of DNA (greater than 200 bp in length) which exhibit a CpG observed/expected ratio greater than 0.6 based on sequence composition, and a GC content above 50 % (Gardiner-Garden and Frommer, 1987). The original CGI prediction algorithm exhibited a relatively high false discovery rate (Illingworth et al., 2008), mainly in repetitive sequences, which was improved by increasing the length requirement of the prediction algorithm (Takai and Jones, 2002). Many computational methods are now available for the identification of CGIs (Chuang et al., 2012; Hackenberg et al., 2006; Wang and Leung, 2004). However, it is important to bear in mind that all prediction approaches decouple the two features of CGIs, their nucleotide properties and tendency to be non-methylated.

1.5.3 Depletion of methylated cytosines over evolutionary time

In vertebrates, CpG dinucleotides occur at only 20-25% of the expected frequency (Josse et al., 1961; Lander et al., 2001; Swartz et al., 1962). Depletion of CpGs has been attributed to the tendency of methylated cytosine to spontaneously deaminate to thymine (Coulondre et al., 1978) (Figure 1.1B). Deamination of cytosine to uracil is efficiently recognised and corrected by uracil glycosylases, as uracil is an unnatural base in DNA (Visnes et al., 2009). However, deamination of 5mC yields thymine (Shen et al., 1994) and the resultant thymine-guanine (T-G) mismatch base-pair is recognised and removed in an error-prone mechanism involving methyl-CpG binding protein 4 (MBD4) or a thymine DNA glycosylase (TDG) (Cortázar et al., 2007; Hendrich et al., 1999; Waters and Swann, 2000). An increased deamination rate of 5mC compared to unmodified cytosine has also been reported (Shen et al., 1994). Therefore, increased deamination of 5mC and inefficient repair of T-G mismatches both likely contribute to the hypermutability of methylated cytosines. Importantly, if a deamination event is not repaired in the germline it will be inherited by subsequent generations and contribute to genomic CpG depletion. The enhanced mutability of 5mC is apparent in mammalian genomes due to a positive correlation between the level of CpG methylation and the extent of TpG and CpA elevation (Bird, 1980; Sved and Bird, 1990). In somatic cells CpG to CpA/TpG mutations are also implicated in human disorders (Barker et al., 1984; Cooper et al., 2010).

Intriguingly, the rarity of CpG dinucleotides genome-wide appears to be a consequence of the fact that they are subject to methylation. This CpG-depletion is not uniform throughout vertebrate genome and 'protected' non-methylated CpGs cluster at CGIs presumably due to either a shared function or a shared protection at these regions. It is this genome-wide depletion of CpGs and the local protection at CGIs which allows these elements to be predicted computationally. Interestingly, it appears that the raised GC content observed at CGIs cannot be explained by CpG density alone. Contributions from biased DNA repair pathways and the association with replication origins are postulated to have led to the raised GC content at CGIs (Antequera and Bird, 1999).

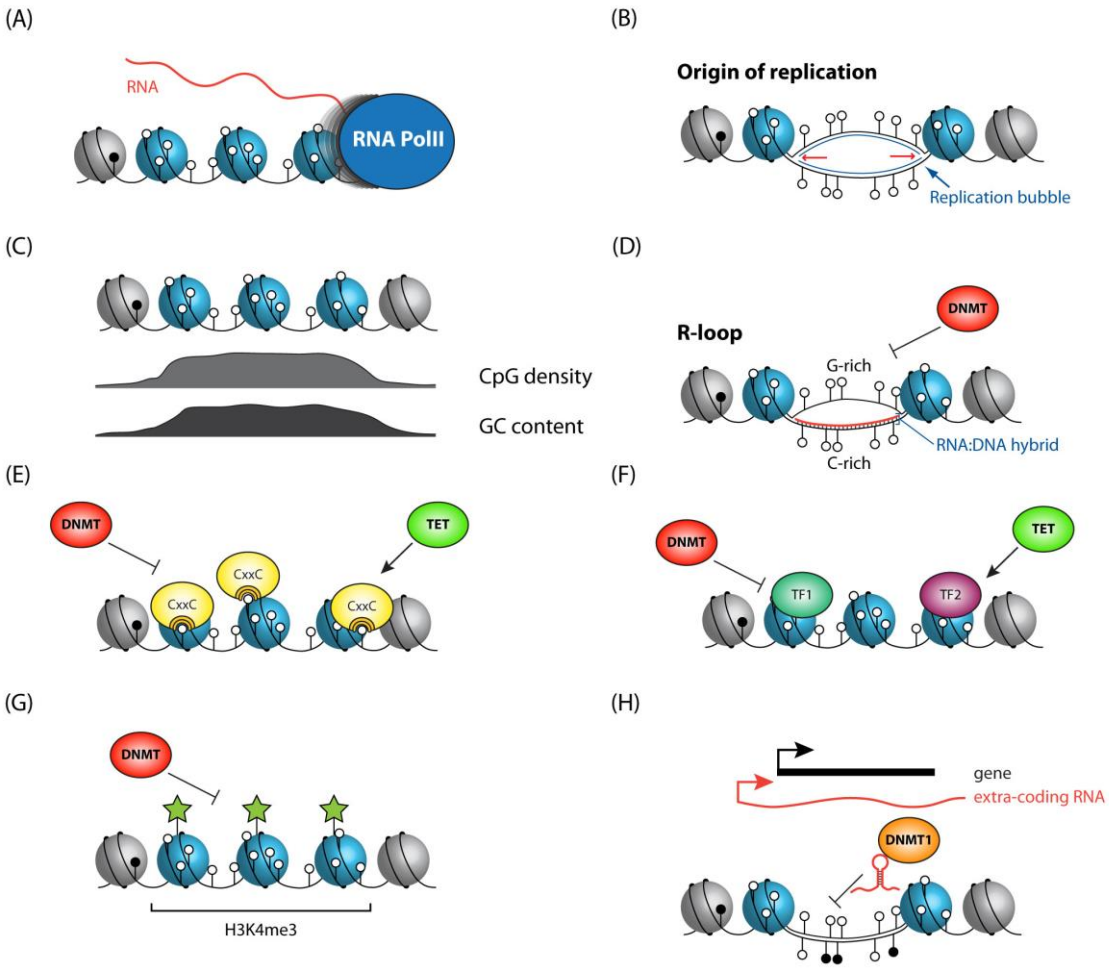


Figure 1.7 Mechanisms proposed to maintain CpG islands free of DNA methylation

Schematic diagrams of mechanisms proposed to specify or protect regions of non-methylated DNA.

1.5.3.2 **What maintains the high CpG density at CGIs?**

Given the tendency of deamination mediated mutagenesis to convert 5mC to T it would be expected that, given enough time, the genome would become completely depleted of CpGs. In fact, it was estimated that it would take 25 million years for CpG density to reduce to the current levels in vertebrates from an undepleted ancestor, a much shorter timescale than vertebrate evolution (around 416 million years) (Ponting, 2008) (Bird et al., 1987; Sved and Bird, 1990). Since complete CpG depletion has not occurred, it appears that a balance has been reached between the rapid rate of CpG loss and the rate of CpG creation by mutation (Antequera, 2003; Sved and Bird, 1990). Therefore CpG dinucleotides at CGIs must be maintained by some selective pressure or property of these regions. One possibility is that loss of CpGs may be opposed by natural selection at sites where a given cytosine residue is important for protein-coding requirements (for example, an arginine

codon CGX), or is found in a transcription factor binding site. Alternatively, CGIs may remain non-methylated in the germline, which protects these regions from the increased CpG depletion rates experienced by methylated cytosines, and prevents inheritance of CGI erosion by the next generation (Tykocinski and Max, 1984).

1.5.3.3 **Maintaining CpG islands free of DNA methylation**

A number of mechanisms have been proposed for the maintenance of non-methylated DNA at CGIs in the germline, and in somatic cells. For example, global transcription may occur at CGI genes during a critical time during gametogenesis when global methylation patterns are being established (Singh et al., 2013), and this may protect CGIs from DNA methylation (Figure 1.7A). Furthermore, CGIs have been proposed to act as origins of replication, which may also protect them from *de novo* DNA methylation (Antequera and Bird, 1999) (Figure 1.7B). The nucleotide content of CGIs may also play an important role in protection from DNA methylation (Figure 1.7C). For example, strand asymmetry of cytosine and guanine content at CGI promoters has been linked to the formation of R-loops, RNA:DNA hybrids formed from transcription through a region of GC skew (Ginno et al., 2012) (Figure 1.7D). These structures may form when CGI genes are transcribed and have been proposed to protect DNA from *de novo* methylation (Ginno et al., 2012) perhaps because RNA:DNA hybrids present a poor substrate for the DNMT enzymes (Ross et al., 2010). Furthermore, due to their CpG-rich nature, CGIs provide a large number of binding sites for non-methyl CpG binding proteins, and protein binding may act to sterically block access of DNMTs to DNA (Antequera, 2003; Cuadrado et al., 2001) (Figure 1.7E). Similarly, binding of transcription factors to DNA may also block access of DNMTs to CpG sites to protect DNA from DNA methylation (Hsieh, 1994) (Figure 1.7F). The proposed DNA demethylase enzymes TET1 and TET3 encode putative non-methylated DNA-binding domains and so may be directly recruited to remove aberrant methylation (Williams et al., 2011; Wu et al., 2011b; Xu et al., 2011b) or may be recruited indirectly by interaction with transcription factors (Figure 1.7E-F). Recent evidence also suggests that chromatin structure and histone modifications

(such as H3K4me3) may also act to prevent *de novo* methylation of CGIs (Figure 1.7G) and will be discussed in sections 1.8 and 1.10. A recent study revealed a novel mechanism for regulating genomic DNA methylation whereby an extra-coding RNA transcribed from the human *CEBPA* gene is proposed to interact with DNMT1 and block maintenance DNA methylation across the transcribed gene locus (Di Ruscio et al., 2013) Figure 1.7H). Together, there are a number of mechanisms proposed to counteract DNA methylation which may therefore act to specify regions of non-methylated DNA.

1.5.4 Segregation of vertebrate gene promoters based on CpG content

Vertebrate gene promoters can be distinguished based on CpG density, ranging from high to low CpG content. This measure is likely reflective of the time the promoter spends in the methylated state in the germline (Saxonov et al., 2006). Interestingly, analysis of all mammalian promoters revealed a bimodality of promoter CpG density, in keeping with the idea that a promoter either exists in a methylated (non-CGI) or non-methylated (CGI) state (Saxonov et al., 2006). More recently, a third class of promoters with intermediate CpG content has been described (Weber et al., 2007). In comparison to those with high or low CpG content, intermediate CpG promoters are subject to a greater degree of differential methylation in somatic cells, and acquisition of DNA methylation is associated with transcriptional silencing (Weber et al., 2007). Furthermore, deep bisulfite sequencing in mouse ES cells has detected regions of the genome which exhibit low levels of methylation (10-50%) and again tend to be CpG poor (Stadler et al., 2011). These regions are distal to gene promoters, appear to encompass transcription factor binding sites and are subject to differential methylation (Stadler et al., 2011).

Unsurprisingly, promoter bimodality is not detected in the invertebrate *Drosophila* (which has little or no DNA methylation), or the sea squirt (which has mosaic DNA methylation away from gene promoters) (Aerts et al., 2004; Elango and Yi, 2008; Sharif et al., 2010). Interestingly, promoter bimodality for fish and other cold-blooded vertebrates species is less clear than for mammals (Aerts

et al., 2004; Elango and Yi, 2008; Sharif et al., 2010), suggesting that cold-blooded vertebrate promoters may not segregate into high and low CpG classes in the same manner as warm-blooded species.

1.6 **Developmental systems for studying DNA methylation**

1.6.1 Mouse

The study of the DNA methylation system in mouse has provided considerable insight into the importance of DNA methylation during development and in disease. Mouse knockout models provide insight into the functionality of the DNA methylation machinery and readers of the DNA methylation state. Profiling genomic DNA methylation during early development has also revealed carefully orchestrated changes in DNA methylation patterning which are essential for reprogramming, establishment of imprinting and dosage compensation.

1.6.1.1 **DNA methylation is essential for embryonic and germ-line development**

Mouse models have been instrumental in addressing the biological importance of DNA methylation during development (Goll and Bestor, 2005). Knockout of DNMT1 causes a dramatic reduction of DNA methylation levels to around 5% of wild-type levels (Lei et al., 1996), with biallelic expression observed for normally imprinted genes (Li et al., 1993) and aberrant silencing of both X chromosomes in female cells due to derepression of both Xist alleles (Panning and Jaenisch, 1996). DNMT1 knockout mice die between 9.5-11 days post coitum (dpc) suggesting that DNA methylation is essential for embryonic development (Lei et al., 1996; Li et al., 1992).

DNMT3A and DNMT3B are highly expressed at around the blastocyst stage and during mouse germ cell development, coincident with the times when novel DNA methylation patterns are being deposited genome-wide (Li, 2002). The two enzymes appear to have separable functions during development, and DNMT3B has a more dominant role in *de novo* methylation during mouse embryonic development (Borgel et al., 2010; La Salle and Trasler, 2006; Lucifero et al., 2007) while DNMT3A is important for setting up DNA methylation patterns in germ cells (Kaneda et al., 2004).

This distinction may be reflected in the mutant mouse phenotypes as DNMT3B knockout mice die at 9.5 dpc with demethylation of minor satellite repeats while mice mutant for DNMT3A die at 4 weeks post birth (Okano et al., 1999). Conditional knockout of DNMT3A in germ cells caused embryonic lethality when inherited only from the mother, indicative of maternal effects and defective imprinting in the germline (Kaneda et al., 2004).

DNMT3L is expressed specifically in germ cells (Aapola et al., 2000) and is essential for establishing maternal imprints in oocytes and at dispersed repeated sequences in the developing sperm and consequently, despite being viable, both male and female DNMTL homozygous null mice are sterile (Bourc'his et al., 2001). Heterozygous offspring of DNMT3L homozygous females die at around 9 dpc due to imprinting defects (Bourc'his and Bestor, 2004; Bourc'his et al., 2001). The similarity between the DNMT3L and DNMT3A conditional phenotypes appear to reflect a shared role in the establishment of maternal imprints (Chedin et al., 2002; Kaneda et al., 2004). Together, genetic ablation of DNA methyltransferase enzymes has revealed that correct DNA methylation patterns are essential during development for establishment of monoallelic gene expression, germline development and viability.

1.6.1.2 **Early developmental changes in DNA methylation**

Despite the seemingly stable inheritance of DNA methylation between somatic cells, DNA methylation is highly dynamic at the genome-scale during early development (Bogdanovic and Veenstra, 2009; Reik et al., 2001). In mammals, two waves of demethylation followed by the reacquisition of DNA methylation patterns occur during development. The first wave occurs during gametogenesis when parent-specific imprinting marks are reset. The second wave initiates following fertilisation, when the paternal genome undergoes a selective, rapid loss of DNA methylation prior to the onset of replication (Figure 1.8A). The oocyte genome, which is hypomethylated compared to sperm (Shirane et al., 2013; Smallwood et al., 2011), avoids this early wave of demethylation (Mayer et al., 2000; Oswald et al., 2000; Santos et al., 2002). The early window of paternal-specific demethylation is followed by continued replication-dependent demethylation of both the maternal

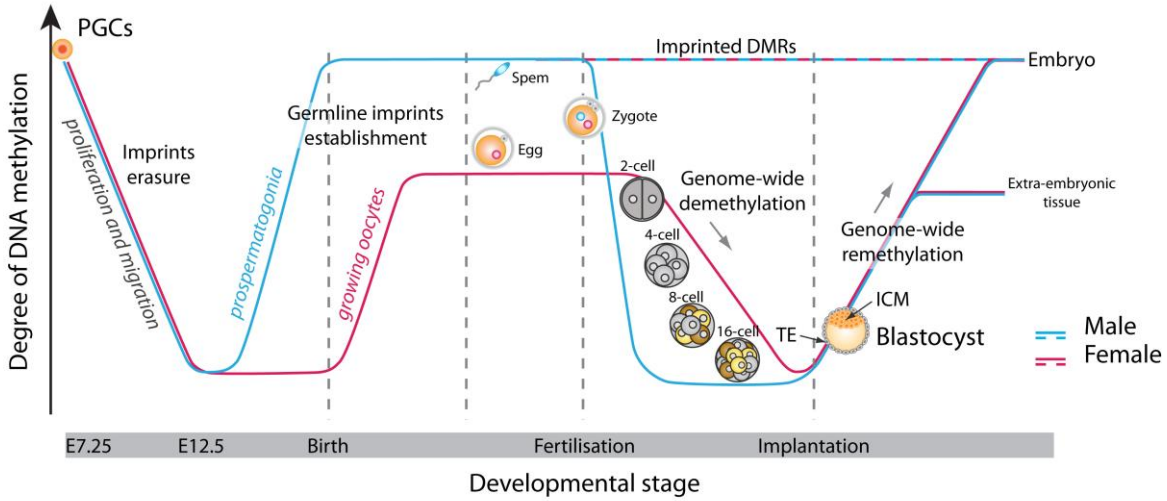
and paternal genomes, which reach a DNA methylation minimum by the blastocyst stage, 4 days after fertilisation (Santos et al., 2002). Interestingly, intracisternal A particle (IAP) repeats, other retroelements and imprinted loci are largely resistant to DNA demethylation (Howlett and Reik, 1991; Kim et al., 2004; Lane et al., 2003; Smith et al., 2012) (Figure 1.8A).

Concomitant with the loss of 5mC in the male pronucleus, 5hmC has been observed to accumulate, suggesting a mechanism for active loss of DNA methylation through enzymatic oxidation of 5mC to 5hmC by the TET enzymes (specifically TET3) (Inoue and Zhang, 2011; Wossidlo et al., 2011). This mechanism appears to be supported by the finding that DNA methylation in the maternal pronucleus is protected from enzymatic conversion by the nucleation of PGC7/Stella specifically on the maternal genome (Nakamura et al., 2012; Wossidlo et al., 2011). Furthermore, TET3-null mice exhibit maternal dominant lethality (Gu et al., 2011). 5hmC on the paternal chromosome is proposed to be diluted during early divisions (Inoue and Zhang, 2011), suggesting that the majority of demethylation occurs through replicative loss (Iqbal et al., 2011; Wossidlo et al., 2011), although base-excision repair has also been implicated (Hajkova et al., 2010; Wossidlo et al., 2010). During very early embryonic divisions, an oocyte-specific degradation-resistant DNMT1 (DNMT1o) is excluded from the nucleus (Doherty et al., 2002; Howell et al., 2001; Ratnam et al., 2002), perhaps accounting for the passive demethylation in the mouse embryo observed during early cell cleavages. Low levels of a somatic DNMT1 isoform (DNMT1s) are detected in the early embryo (Cirio et al., 2008; Kurihara et al., 2008), and may be involved in the maintenance methylation of imprinted loci (Borowczyk et al., 2009).

De novo DNA methylation occurs at around the time of blastocyst implantation and occurs in embryonic cells, while DNA methylation levels remain relatively low in extraembryonic tissues (Figure 1.8A) (Dean et al., 2001). Methylation of CGIs is rare, however a small fraction of CGIs are methylated, including those associated with pluripotency genes and the germline expression programme (Borgel et al., 2010; Farthing et al., 2008; Meissner et al., 2008; Mohn et al., 2008; Weber et al., 2007). DNA methylation is also observed at developmental genes, suggesting that

acquisition of DNA methylation is associated with loss of pluripotency and maintenance of cellular identity (Borgel et al., 2010).

(A)



(B)

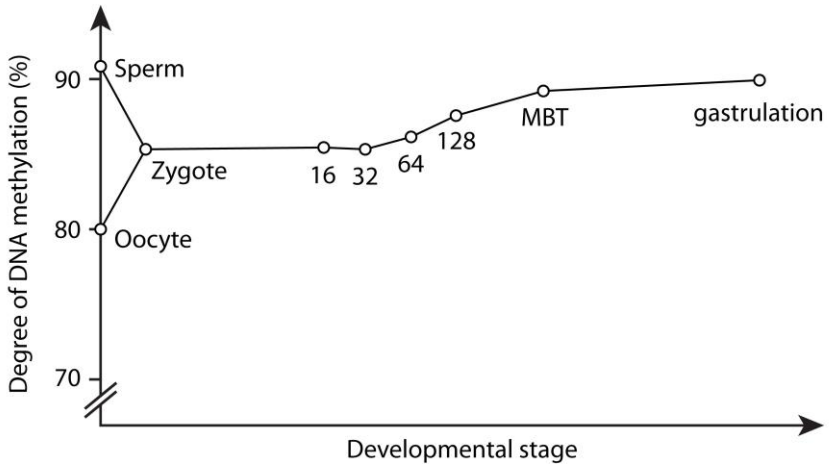


Figure 1.8 DNA methylation changes during mouse and zebrafish development

(A) As primordial germ cells (PGCs) proliferate and migrate towards the genital ridge in the mouse embryo from E7.5, global DNA methylation patterns are erased (imprints erasure). New methylation patterns are then established in germ-cell precursors, in prospermatogonia this occurs before meiosis and is completed before birth (blue line), whereas in the female germline primary oocytes enter meiosis and arrest in prophase I and DNA methylation patterns are only reestablished after birth during the growth phase. Oocytes are globally hypomethylated compared to sperm. Following fertilisation, a second wave of demethylation occurs, first methylation patterns are rapidly erased from the paternal genome by an active mechanism (blue line) while demethylation of the maternal genome occurs much more slowly and passively, dependent on DNA replication (pink line). Importantly, these demethylation events in the zygote do not include imprinted differentially methylated regions (DMRs) (blue and pink dashed lines), leading to parent-of-origin specific allelic methylation of these loci in early embryos and consequently leads to parental-specific expression of the associated imprinted gene. At blastocyst implantation, DNA methylation landscapes are re-established, concomitant with cell-lineage determination and cellular differentiation. Figure adapted from (Smallwood and Kelsey, 2012). (B) Average DNA methylation levels calculated for zebrafish gametes and different stages during development, including 16 cell, 32 cell, 64, cell 128 cell, 1000 cell (mid-blastula transition; MBT) and germ ring (gastrulation) stage. Zygotic methylation level is the mean of the methylation level of sperm and oocyte. DNA methylation levels are

shown as the percentage of methylated cytosine from all CpG contexts. The 16 cell embryo appears to have intermediate methylation status between the male and female gametes, DNA methylation levels then increase up to MBT when the maternal genome DNA methylation patterns are proposed to resemble that of the paternal genome in sperm (Jiang et al., 2013; Potok et al., 2013). Genome equivalence of DNA methylation patterns on the maternal and paternal chromosomes at MBT is proposed to create a totipotent chromatin state (Potok et al., 2013). DNA methylation changes continue to occur following MBT. Figure adapted from (Jiang et al., 2013).

1.6.1.3 **Imprinting**

Genomic imprinting is an epigenetic process by which a subset of genes is expressed in a parent-of-origin specific manner. Differential DNA methylation at imprinting control regions (ICRs), observed across a number of mammalian species, has been implicated in controlling the expression of clustered imprinted genes (Wood and Oakey, 2006). These ICRs generally fall into two categories, those which function as insulator elements and those which function as a promoter element for a regulatory non-coding RNA (Ideraabdullah et al., 2008). Genomic imprinting has therefore become a paradigm for studying the impact of DNA methylation on gene regulation (Ferguson-Smith, 2011).

Studies in the mouse have revealed a number of fundamental features of genomic imprinting. Imprints have been shown to be established during gametogenesis when DNA methylation patterns are initially completely erased and then replaced in a sex-specific manner at ICRs (Figure 1.8A). Many imprinted genes play important roles during fetal development and growth, and therefore accurate genomic imprinting is essential for normal mammalian development (Abramowitz and Bartolomei, 2012; Ferguson-Smith, 2011). Interestingly, a number of imprinted genes have also been implicated in post-natal processes, such as regulation of behaviour and metabolism (Charalambous et al., 2007; Wilkinson et al., 2007).

A well-studied example of genomic imprinting is the insulin-like growth factor gene (*Igf2*), which is expressed solely from the paternally inherited allele while the maternal allele is repressed (DeChiara et al., 1991; Ferguson-Smith et al., 1991). Conversely, the *H19* non-coding RNA gene, which lies 90 kb downstream of *Igf2*, is expressed only from the maternally-inherited chromosome (Bartolomei et al., 1991). The paternally methylated ICR at the *H19*-DMR (differentially methylated region) is required for normal imprinting of *H19* and the linked, but reciprocally imprinted, *Igf2* gene (Figure

1.9). The ICR element is bound by CTCF (CCCTC-binding factor) when non-methylated (maternal allele), and this binding prevents access of downstream enhancers to the *Igf2* gene rendering it silent. On the paternal chromosome, the enhancer region is free to preferentially interact with the *Igf2* promoter, while the *H19* promoter is repressed and gains DNA methylation. This paradigm appears widely applicable to other imprinted loci, as the establishment of germline-derived DMRs is consistently observed at imprinted loci, although how these lead to mono-allelic gene expression is different at each locus.

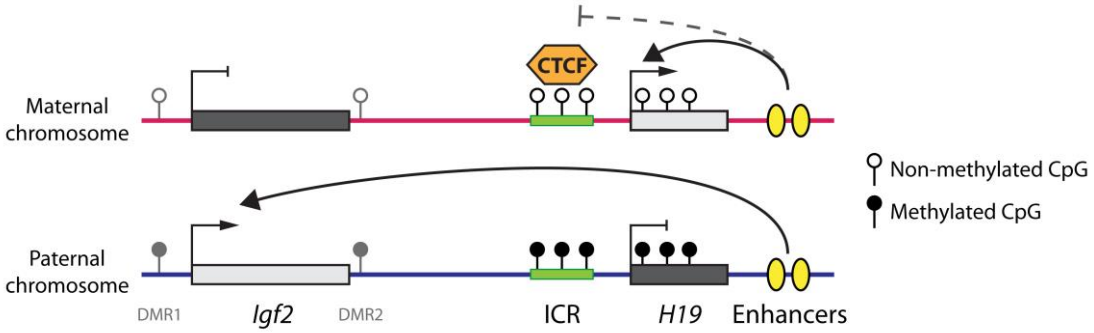


Figure 1.9 Imprinting at the *Igf2/H19* locus

Example of an imprinted loci where the *H19* gene is expressed only from the maternal chromosome and the *Igf2* gene is expressed only from the paternal chromosome (Reik and Murrell, 2000). The two genes share an enhancer which is downstream of the *H19* gene and the interaction of this enhancer with the *Igf2* and *H19* promoters is controlled by an imprinting-control region (ICR) upstream of *H19*. The ICR acts as a boundary element, and is bound by CTCF when non-methylated (open circles) on the maternal allele, which prevents interaction with and activation of the Igf2 promoter but allows activation of the *H19* gene. The ICR is methylated (filled circles) on the paternal chromosome, which prevents binding of CTCF and allows the enhancer to interact with the *Igf2* promoter, while the *H19* promoter becomes methylated and silenced. Differentially methylated regions (DMRs) are also found at the *Igf2* locus (Abramowitz and Bartolomei, 2012).

1.6.1.4 **X inactivation**

X chromosome inactivation (XCI) provides a second paradigm for understanding the role of DNA methylation in the epigenetic regulation of monoallelic gene expression. In mammalian species, males contain only one X chromosome, while females contain two, therefore in order to balance the expression of genes from the X chromosome between males and females one of the two female X chromosomes is randomly silenced during early development. Dosage compensation in mammals is mediated via the process of X inactivation in female mammals whereby one X chromosomes is inactivated by successive silencing mechanisms to form a compacted and silent X chromosome.

Most of the current knowledge about X chromosome inactivation has come from work in mice which initially undergo imprinted XCI during early embryonic divisions (of the paternal X chromosome). This imprinted silencing is reversed in the inner cell mass of the mouse embryo following which an X chromosome is silenced at random (Okamoto et al., 2004). Both forms of silencing appear to involve the monoallelic expression of the non-coding RNA Xist from the future inactive X chromosome (Xi), which regulates the onset of gene silencing, chromosome remodelling and recruitment of chromatin modifying complexes (Chow and Heard, 2009). DNA methylation occurs on the randomly inactivated X chromosome but not on the paternal imprinted Xi, which may potentially explain why the inactive state of the imprinted paternal Xi can be easily reversed in the inner cell mass prior to random X inactivation (Okamoto et al., 2004). Acquisition of DNA methylation occurs relatively late in random XCI, with CGIs on the inactive X becoming hypermethylated while gene-poor regions in fact exhibit lower levels of DNA methylation compared to the active X chromosome (Weber et al., 2005). *De novo* DNA methylation appears dispensable for XCI (Sado et al., 2004), suggesting that DNA methylation may act to maintain silencing of genes which have already been repressed by other chromatin-based processes (Jones, 2012; Lock et al., 1987). Skewing of XCI is rare, but is observed frequently in females with X-linked mutations (Muers et al., 2007), for example in ATR-X syndrome discussed in section 1.10.1.2.

1.6.2 Zebrafish

The zebrafish is an emerging model for the study of DNA methylation and encodes clear functional orthologues to mammalian DNMTs (Campos et al., 2012; Hendrich and Tweedie, 2003; Rai et al., 2008; Rai et al., 2010; Rai et al., 2006; Smith et al., 2011). (For a discussion of early zebrafish development, see Appendix 1.1 - Appendix 1.5). Unlike mammals, the zebrafish genome does not appear to undergo global demethylation following fertilisation (Macleod et al., 1999), suggesting this is not an essential process for vertebrate development. Zebrafish may not require global DNA methylation erasure as two processes associated with DNA methylation in mouse, imprinting and X chromosome inactivation, have not been reported in zebrafish. Similar to zebrafish, DNA

methylation levels appear to remain high and unchanging during *Xenopus* development (Veenstra and Wolffe, 2001), although some locus-specific changes have been reported at developmentally regulated promoters (Stancheva et al., 2002).

Although less dramatic than in the mouse, careful profiling of DNA methylation during zebrafish embryogenesis has revealed that DNA methylation changes occur during early development (Jiang et al., 2013; Potok et al., 2013). The zebrafish oocyte genome was observed to be hypomethylated in comparison to sperm, although this was not as striking as hypomethylation levels reported for mouse oocytes (Jiang et al., 2013; Shirane et al., 2013; Smith et al., 2012). Upon fertilisation, global methylation levels in the zebrafish zygote were observed shift to an intermediate methylation level and increase gradually to the mid-blastula transition (MBT, Appendix 1.4). Comparison of DNA methylation at 22 differentially methylated regions (DMRs) adjacent to SNPs revealed that in all cases the gametic methylation state was retained in the 16 cell stage embryo. However, by MBT the methylation state of the maternal locus matched that of the paternal genome (Jiang et al., 2013). It has been suggested that global DNA methylation changes may be required in zebrafish to attain a totipotent chromatin state at the time of zygotic genome activation (Appendix 1.4), and that maternal and paternal methylome equivalence may be required to induce the onset of zygotic transcription at MBT (Potok et al., 2013).

1.7 **Interpretation of DNA methylation**

1.7.1 Targeting DNA methylation

DNA methylation is thought to be targeted to regions of the genome by three main mechanisms. Firstly, *de novo* methyltransferases may be recruited by specific transcription factors (Figure 1.10A), secondly the DNMTs themselves may directly recognise DNA or chromatin modifications (Figure 1.10B) and finally RNA-based systems might target DNA methylation to target sequences (Figure 1.10C). For example, small RNAs known as Piwi-interacting RNAs (piRNAs) have been shown to target DNA methylation to transposable elements in animal germ cells (Luteijn and Ketting, 2013),

while RdDM in plants directs DNA methylation to transposons in an RNA-dependent manner (Matzke et al., 2009). The maintenance DNMT, DNMT1, can be recruited to replication forks in a UHRF1-dependent manner (Bostick et al., 2007; Nishiyama et al., 2013; Sharif et al., 2007) (see 1.3.3).

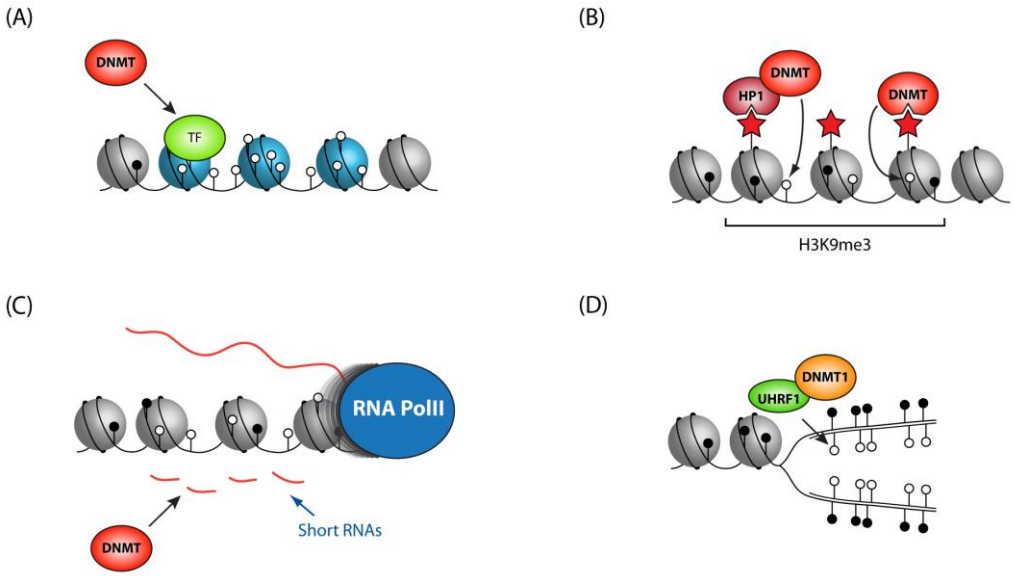


Figure 1.10 Mechanisms by which DNA methyltransferases are targeted to DNA

Schematic diagrams illustrating targeting of DNA methylation. (A) DNA methyltransferase (DNMT) enzymes can be targeted to DNA by interaction with transcription factors (TFs). (B) DNMTs can be recruited to chromatin through direct binding of histone tails (right, H3K9me3 or H3K4me0) or interaction with HP1 which binds H3K9me3 (left). (C) Small RNA-based pathways can also recruit DNMTs to DNA. (D) The maintenance DNMT, DNMT1, can be recruited to DNA replication forks by UHRF1.

Acquisition of DNA methylation can impact genome function by two main mechanisms (Klose and Bird, 2006). Firstly, DNA methylation can direct transcriptional repression by blocking the binding of transcription factors which can inhibit gene activation (Comb and Goodman, 1990; Kim et al., 2003; Mancini et al., 1999; Watt and Molloy, 1988). Secondly, proteins which recognise methyl-CpGs can recruit co-repressor complexes which modify and remodel the local chromatin landscape to silence transcription (Boyes and Bird, 1991; Hendrich and Bird, 1998; Jones et al., 1998; Nan et al., 1998). This second mechanism will be discussed in more detail below.

1.7.2 Interpretation of DNA methylation by a family of MBD-domain containing proteins

Methyl-CpG binding proteins are able to bind to methylated DNA and therefore interpret and determine the readout of DNA methylation, which is an otherwise inert modification (Antequera et al., 1989; Hendrich and Bird, 1998; Meehan et al., 1989). MeCP2 (methyl CpG binding protein 2) was the first methyl binding protein to be identified (Lewis et al., 1992; Meehan et al., 1989) and homology searches using the methyl-CpG binding domain (MBD) from MeCP2 (Nan et al., 1993) identified four gene homologues MBD1-4 (Hendrich and Bird, 1998) (Figure 1.11). The MBD domain has since been shown to interact with methylated CpG DNA via insertion of a loop from the MBD domain into the major groove of DNA. Specific recognition of the cytosine methyl group is dependent on a hydrophobic patch of the MBD domain (Ho et al., 2008; Ohki et al., 2001), and two arginine residues which form hydrogen bonds with the guanine bases and stack with the methyl-cytosines in a stair motif conformation (Zou et al., 2012). The MBD domains of MBD1, MBD2, MBD4 and MeCP2 have recently been shown to confer 5mC-binding *in vivo* (Baubec et al., 2013). However, human and mouse (Hendrich and Bird, 1998), but not *Xenopus* (Wade et al., 1999), MBD3 proteins have lost the capacity to bind methylated DNA due to a point mutation in the MBD domain (Figure 1.11).

Based on protein interaction studies, the MBD-containing proteins are proposed to recruit chromatin remodellers and modifying enzymes to methylated DNA to mediate transcriptional repression (Nan et al., 1998; Ng and Adrian, 1999). MeCP2 acts as a transcriptional repressor to silence methylated promoters via a number of interaction partners (Jones et al., 1998; Nan et al., 1997). Association with TFIIB interrupts the initiation stages of transcription (Kaludov and Wolffe, 2000) while incorporation into the SIN3A co-repressor complex mediates gene repression in a histone deacetylase (HDAC) dependent manner (Jones et al., 1998; Nan et al., 1998; Silverstein and Ekwall, 2005). Recently the pathological mutation in MeCP2 that leads to the human disease Rett syndrome was shown to disrupt the interaction of MeCP2 with the NCoR corepressor complex (see

1.10.2.3) (Lyst et al., 2013). MBD1 associates with the SUV39H1–HP1 chromatin modifying complex to mediate transcriptional repression at methylated regions of the genome (Fujita et al., 2003). MBD2 and 3 are gene paralogues which both associate with the Mi-2/NuRD complex, which contains both nucleosome remodelling and HDAC activities (Zhang et al., 1999), to confer DNA methylation-mediated transcriptional repression. MBD4 can also silence transcription through interaction with SIN3A and HDAC1 (Kondo et al., 2005) and has been implicated in targeting repair of sites of cytosine deamination (Hendrich et al., 1999) and active demethylation in zebrafish (Rai et al., 2008). The MBD proteins appear to have partial redundancy in function (Martin Caballero et al., 2009) which may be reflected in the mild developmental defects observed for MBD1, 2, 4 and MeCP2 knockdown (Table 1.1).

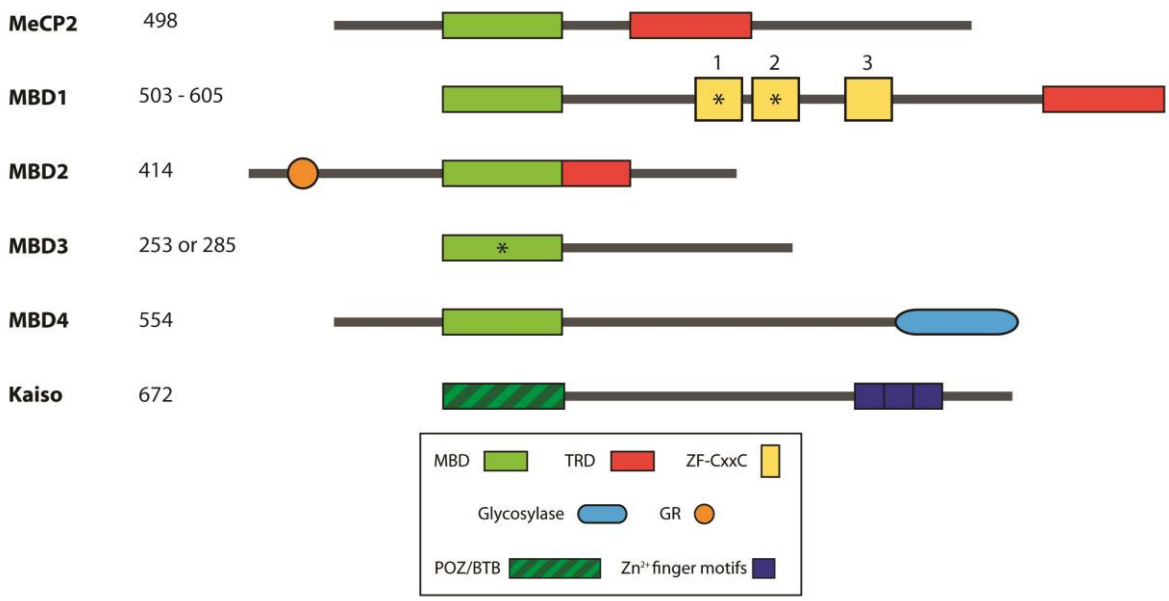


Figure 1.11 Methyl-CpG binding proteins

A schematic of methyl-CpG binding proteins adapted from (Bogdanovic and Veenstra, 2009). Number of amino acid residues are indicated (left), taken from (Fatemi and Wade, 2006) and NCBI. Where multiple isoforms are known, a range or two lengths are indicated. A legend for the illustrated domains is shown below the schematics, and an asterisk (*) indicates that the domain is non-functional. Methyl-CpG binding domain, MBD; transcriptional repression domain, TRD; glycine and arginine rich region, GR. The proteins are shown with the N-terminus on the left and are centred on the methyl-CpG binding domain.

Table 1.1 *Phenotypes of methyl-CpG binding protein knockout mice*

Gene	Methyl-CpG binding domain	Phenotype	Reference
MeCP2	MBD	At 3-8 weeks of age, mice develop an uncoordinated stiff gait and a loss of spontaneous movement. Death normally occurred by 54 days after birth.	(Chen et al., 2001; Guy et al., 2001)
Mbd1	MBD	Viable, healthy and fertile but exhibits a mild neurological phenotype. Adult neuronal stem cells display a reduced capacity to differentiate.	(Zhao et al., 2003)
Mbd2	MBD	Viable, healthy and fertile but exhibit defects in maternal behaviour.	(Hendrich et al., 2001)
Mbd3	MBD – lost binding specificity	Embryonic lethal, likely due to disruption of the NuRD co-repressor complex.	(Hendrich et al., 2001)
Mbd4	MBD	Viable, healthy and fertile but show a three-fold increase in CpG to TpG mutations.	(Millar et al., 2002; Wong et al., 2002)
Kaiso	Zinc finger	No apparent abnormalities.	(Prokhortchouk et al., 2006)

1.7.3 A second family of methyl-CpG binding proteins

A second family of proteins, including KAISO, are able to bind methylated DNA via an MBD-independent mechanism involving a zinc finger domain which has a binding site preference of CGCG (Prokhortchouk et al., 2001). KAISO also exhibits a transcriptional repressor activity which is mediated via the N-terminal POZ/BTB domain and interaction with the NCoR complex (Prokhortchouk et al., 2001; Yoon et al., 2003). Knockdown of Kaiso in *Xenopus* leads to a premature activation of a subset of genes prior to MBT, a phenocopy of Dnmt1 depletion, suggesting that Kaiso may act to repress methylated genes during *Xenopus* early development (Ruzov et al., 2004; Stancheva and Meehan, 2000). Kaiso has two close paralogues in mammals, ZBTB4 and ZBTB38 (Sasai et al., 2010). Similar to Kaiso, these two paralogues can bind to methylated DNA, but can also bind to non-methylated CpG DNA within a specific sequence consensus (Filion et al., 2006; Sasai et al., 2010).

Recently another zinc finger transcription factor, ZFP57, was also shown to bind methylated DNA in a sequence-specific manner (Quenneville et al., 2011). ZFP57 encodes five putative Cys2His2 zinc fingers and the adjacent zinc fingers ZF2 and ZF3 are able to recognise methylated CpG DNA in the sequence context TGCCGC (Quenneville et al., 2011), with reduced binding affinity for other oxidised

derivatives of 5mC or unmethylated cytosine (Liu et al., 2012). The ZF2 and ZF3 zinc fingers in Zfp57 interrogate the first (TGC) and second half (CGC) of the motif respectively (Liu et al., 2012). The 5' 5mC in the central CpG dinucleotide is coordinated by a layer of ordered water molecules while the 3' 5mC makes contacts with Arg178 from ZFP57. Zfp57 plays an important role in maintaining both maternal and paternal genomic imprinting and loss of maternal and zygotic function of Zfp57 leads to embryonic lethality (Li et al., 2008).

1.8 **Links between DNA methylation, histone modifications, nucleosome remodelling and their impact on gene expression**

The interplay between histone modifications and DNA methylation is now becoming apparent. It is clear that the DNA methylation state can impact the modification of nearby histone tails, while histone modifications can also influence the placement of DNA methylation.

1.8.1 Modification of histone lysine residues

1.8.1.1 **Histone lysine acetylation**

Acetylation of histone lysine residues (Figure 1.12B-C) can directly influence chromatin structure by reducing the affinity of histones for DNA via neutralisation of the basic charge of lysine residues (Krajewski, 2000; Tse et al., 1998). Furthermore, lysine acetylation can recruit bromodomain-containing coactivators to chromatin (Yang, 2004) and in some cases can antagonise placement of repressive methylation marks, such as those placed at lysine-9 (K9) and K27 on the histone H3 tail (Figure 1.12C) (Pasini et al., 2010; Yang, 2004). Histone acetylation is therefore associated with active, transcribed chromatin (Kouzarides, 2007). Heterochromatic regions exhibit low levels of histone acetylation, mediated in part by methyl-CpG binding proteins recruiting histone deacetylase activity (HDAC) to regions of DNA which are methylated (Jones et al., 1998; Nan et al., 1998) (see section 1.7.2). During gene repression, histone deacetylation can also be followed by acquisition of DNA methylation, revealing a further potential mechanistic link between histone acetylation and DNA methylation (Cedar and Bergman, 2009; Epsztejn-Litman et al., 2008; Feldman et al., 2006b).

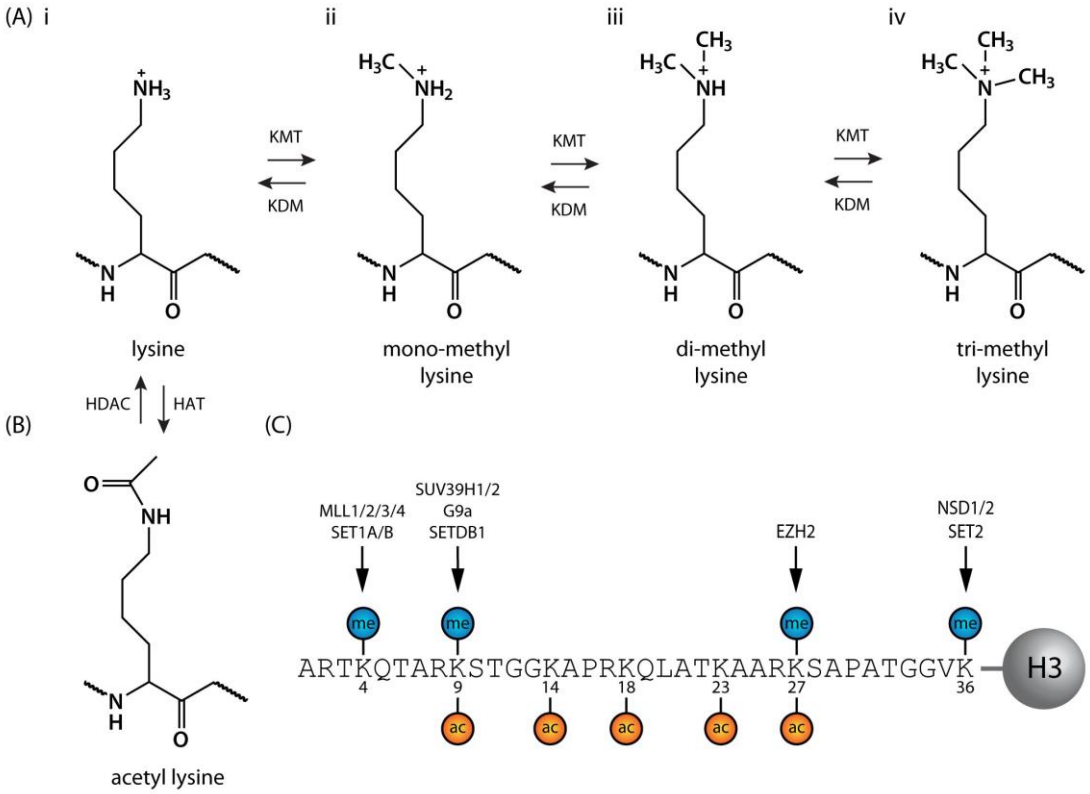


Figure 1.12 Histone lysine methylation and acetylation

Histones are subject to post-translational modification, including methylation and acetylation. (A) Lysine residues can be (i) mono-, (i) di- or (iii) tri-methylated, with lysine methyltransferase enzymes (KMTs) placing the methylation marks and lysine demethylase enzymes (KDMs) removing them. (B) Lysine residues can also be acetylated, catalysed by histone acetyltransferases (HATS) and removed by histone deacetylases (HDACs). (C) The histone H3 tail is subject to a large number of post-translational modifications, lysine methylation (me, blue) and lysine acetylation (ac, orange) are shown. Residues 1-36 of histone H3 are depicted. The KMT enzymes responsible for lysine methylation of mammalian histone H3 are listed above the target lysine. Part (C) was adapted from (Margueron et al., 2005).

1.8.1.2 **Histone lysine methylation**

Many histone tail lysine residues are subject to methylation and, depending on the position on the histone tail, these modifications can have a positive or negative impact on local transcription levels. Most lysine methylation is placed by SET-domain containing histone lysine methyltransferase enzymes (Martin and Zhang, 2005) and can occur in three modification states mono-, di- and tri-methylation (Figure 1.12A). Only recently was the opposing histone lysine demethylase activity identified, firstly for LSD1 (Shi et al., 2004) and a later for a distinct JmjC-domain containing class of demethylases (Klose et al., 2006). Unlike acetylation, histone methylation does not directly impact chromatin structure. However, many reader proteins have been identified which recognise specific lysine methylation marks and impact chromatin structure, histone modification, DNA methylation

and gene function (Cedar and Bergman, 2009; Kouzarides, 2007). A number of histone tail arginine residues are also methylated, with three possible modes of methylation: monomethylated, symmetrically dimethylated or asymmetrically dimethylated (Margueron et al., 2005). Histone arginine methylation can impact transcription by blocking modification of adjacent residues on histone tails, and through the recruitment of effector proteins (Di Lorenzo and Bedford, 2011), but will not be discussed in further detail here.

1.8.2 DNA methylation and histone lysine methylation

1.8.2.1 **DNA methylation and methylation of H3K9**

G9a and GLP are two related methyltransferases which in mammals form a heteromeric complex which is the primary enzyme for methylation of H3K9me1/2 in euchromatic regions of the genome (Tachibana et al., 2002; Tachibana et al., 2005). The formation of long tracts of the H3K9me1/2 modification by G9a/GLP is linked with the silencing of associated genes and the formation of facultative heterochromatin within regions of euchromatin (Tachibana et al., 2002; Tachibana et al., 2005). Loss of G9a in mouse ES cells is associated with DNA hypomethylation (Ikegami et al., 2007), however G9a catalytic mutants, which cannot rescue H3K9 methylation, do not lead to global loss of DNA methylation, suggesting that the G9a/GLP complex regulates H3K9 methylation and DNA methylation independently (Tachibana et al., 2008). A number of further studies have linked G9a-mediated H3K9 methylation with gene regulation and DNA methylation of imprinted genes (Nagano et al., 2008; Wagschal et al., 2008; Xin et al., 2003), in *de novo* methylation and silencing of the OCT-3/4 gene (Feldman et al., 2006a) and in coordinating the placement of H3K9 methylation and DNA methylation during DNA replication (Estève et al., 2006).

Tri-methylation of H3K9, placed by Suv39h1/2 (Rea et al., 2000) and Setdb1 (Schultz et al., 2002), is found associated with repetitive DNA elements and regions of centromeric constitutive heterochromatin (Martens et al., 2005; Rosenfeld et al., 2009). DNA methylation can facilitate the post-translational modification of histone 3 lysine 9 (H3K9) through recruitment of methyl binding

domain containing proteins (the MBD family, discussed in 1.7) which can recruit H3K9 methyltransferases to chromatin (Bogdanovic and Veenstra, 2009). For example, the association of MBD1 with SETDB1 is necessary for the accurate maintenance of heterochromatic H3K9me3 methylation during replication (Fujita et al., 2003; Sarraf and Stancheva, 2004). Interestingly, H3K9me3 has been reported at silenced genes in ES cells, suggesting that the tri-methylated state may also play a role in regulating gene expression at facultative heterochromatic regions (Bilodeau et al., 2009).

In turn, the targeting of the *de novo* methyltransferases DNMT3A/B is in some contexts dependent on the presence of H3K9me3, first noted from studies in the filamentous fungi *Neurospora crassa* (Tamaru and Selker, 2001). At pericentric heterochromatin in mammals, the catalytic activity of SUV39H1/2 is required for the placement of H3K9me3 and the establishment of DNA methylation (Lehnertz et al., 2003). The H3K9me3 modification is recognised by the chromodomain of heterochromatin protein 1 (HP1) (Bannister et al., 2001; Jacobs et al., 2001; Lachner et al., 2001) which can then recruit DNMTs to place DNA methylation (Lehnertz et al., 2003) (Figure 1.10B). Furthermore, it appears that DNMT3A/B can interact directly, via their ADD domains, with the H3K9 histone methyltransferases SUV39H1 (Fuks et al., 2003) and SETDB1 (Li et al., 2006). It has also been reported that the DNMT3A and ATRX ADD domains are also able to recognise H3K9me3 directly (Dhayalan et al., 2011; Zhang et al., 2010b) (Figure 1.10C). H3K9me2/3 is also bound by UHRF1 via a Tandem Tudor Domain (TTD) (Arita et al., 2012; Nady et al., 2011; Rothbart et al., 2012) as part of a mechanism that targets maintenance DNA methylation to pericentric heterochromatin and telomeres (Ebert et al., 2006; Ernst et al., 2011; Grewal and Elgin, 2007; Hawkins et al., 2010). Together, H3K9me3 and DNA methylation appear to reinforce one another to maintain a repressive chromatin environment at heterochromatic regions of the genome.

1.8.2.2 **DNA methylation and methylation of H3K36**

Histone 3 lysine 36 trimethylation (H3K36me3) is placed co-transcriptionally throughout gene bodies by SETD2 (Edmunds et al., 2008) which associates with the phosphorylated C-terminal domain of RNAPII (Sun et al., 2005). The PWWP domain of DNMT3A/B is proposed to bind to H3K36me3 (Dhayalan et al., 2010) and indeed DNA methylation is reported to be higher at genes with intermediate expression levels compared to lowly expressed genes (Zemach et al., 2010) (see section 1.8.3.2). Deletion of the DNMT3A/B PWWP domain reduced their ability to bind to and methylate major repeats at pericentric heterochromatin (Chen et al., 2004; Ge et al., 2004), suggesting that H3K36me3 may have a role in heterochromatic DNA methylation independent of its co-transcriptional role (Chantalat et al., 2011).

1.8.2.3 **DNA methylation and methylation of H3K4**

The H3K4me3 modification and DNA methylation appear to be mutually exclusive across the genome. DNMT3L has been shown to bind to unmodified H3K4 through its ADD domain (Ooi et al., 2007). The presence of H3K4me3 at non-methylated regions of the genome may therefore help to repel *de novo* methyltransferase activity. Genome-wide analyses have also concluded that regions modified by H3K4me2 methylation are protected from CpG methylation (Okitsu and Hsieh, 2007; Weber et al., 2007).

1.8.3 The impact of DNA methylation on gene expression

Although DNA methylation is generally associated with gene silencing, it has become clear that the function of DNA methylation varies with genomic context and the relationship between DNA methylation and gene expression is not as straight forward as once thought. The impact of DNA methylation on gene promoters, gene bodies, regulatory elements and repetitive elements will be discussed briefly.

1.8.3.1 **Gene promoters and enhancers**

During development and differentiation, a small number of CGI promoters acquire DNA methylation and this event is strongly correlated with gene silencing (Hashimshony et al., 2003; Kass et al., 1997; Venolia and Gartler, 1983) likely due to inhibition of transcription factor binding and recruitment of transcriptionally repressive MBD-containing complexes (Klose and Bird, 2006). Methylation of enhancer elements has been shown to reduce their activity *in vivo* (Stadler et al., 2011) and in reporter assays (Schmidl et al., 2009), suggesting that DNA methylation may be able to regulate the activity of distal regulatory elements in a similar mechanism to gene promoters. In most cases DNA methylation is unlikely to be the initiating mechanism for promoter silencing, as DNA methylation is usually limited to genes which have already entered the repressed state (Jones, 2012). For example, methylation of the *Hprt* gene on the inactive X chromosome occurs after initiation of the silenced state, suggesting that DNA methylation acts to lock in a repressed state (Lock et al., 1987). This is supported by the observation that CGI promoters already marked by polycomb repression have a greater chance of being methylated in cancer cells suggesting that the silent state precedes DNA methylation (Gal-Yam et al., 2008; Schlesinger et al., 2007).

1.8.3.2 **Gene body methylation**

In general, gene bodies exhibit low CpG density and are broadly methylated. However, in contrast to DNA methylation at promoter CGIs, gene body methylation does not necessary block gene expression (Jones, 1999). 'Orphan' CGIs have been identified in gene bodies (Illingworth et al., 2010) and acquisition of DNA methylation at an intragenic CGI has been reported to cause a reduction in gene expression, most likely by reducing gene expression from an alternative promoter within the body of the gene (Deaton et al., 2011; Maunakea et al., 2010). In another study, methylation of an intergenic CGI lead to increased gene expression, however here the CGI methylation may silence an antisense transcript which may oppose the expression of the gene (Straussman et al., 2009). More generally, a correlation has recently been observed between gene transcription and levels of gene

body methylation with the highest levels of DNA methylation observed at genes with intermediate levels of expression (Ball et al., 2009; Cokus et al., 2008; Zemach et al., 2010). In summary, the precise role of gene body DNA methylation in transcriptional regulation remains unclear and appears to be highly context dependent.

1.8.3.3 **Silencing repetitive regions of the genome**

Repetitive regions of the genome are methylated in vertebrates, plants and fungi (Zemach, 2010). DNA methylation induces silencing through the recruitment of methyl-CpG binding proteins (see section 1.7.2) which recruit co-repressor complexes to remodel and compact chromatin and place repressive histone modifications including H3K9me3 (Kondo et al., 2004; Martens et al., 2005). However, it is not known whether DNA methylation is the default state or if it is actively targeted to these regions. Interestingly, H3K9me3 appears to be able to target DNA methylation to repetitive DNA (Tamaru et al., 2003), and is in fact essential for DNA methylation and generating silent heterochromatic regions in *N. crassa* (Honda et al., 2012; Tamaru and Selker, 2001). Furthermore, a hypomorphic DNMT1 mouse model, with globally reduced DNA methylation levels, exhibited derepression of retroviral elements (Howard et al., 2007). Together, DNA methylation appears to lock in transcriptional silencing at heterochromatic regions of the genome.

1.9 **Interpretation of non-methylated DNA**

1.9.1 A family of non-methylated cytosine binding proteins

Given that a family of methyl-CpG binding proteins exist, it seemed likely that non-methylated CpG dinucleotides may also act as a protein binding site. Through a phage-based ligand screen, Skalnik and colleagues identified the first non-methylated CpG binding protein (CGBP or CFP1) whose DNA binding activity relied on a cysteine-rich zinc finger CxxC (ZF-CxxC) domain (Voo et al., 2000). Bioinformatic analyses subsequently identified an extended family of ZF-CxxC domain-containing proteins (Figure 1.13 and Table 1.2). All family members contain two cysteine-rich clusters composed of $CxxCxxCx_{4/5}CGxCxxC$ and CxxRxC motifs (where x is any amino acid). Variations in the

linker region between the two clusters distinguish three classes of ZF-CxxC domain and define their capacity or selectivity for binding non-methylated DNA (Figure 1.14A).

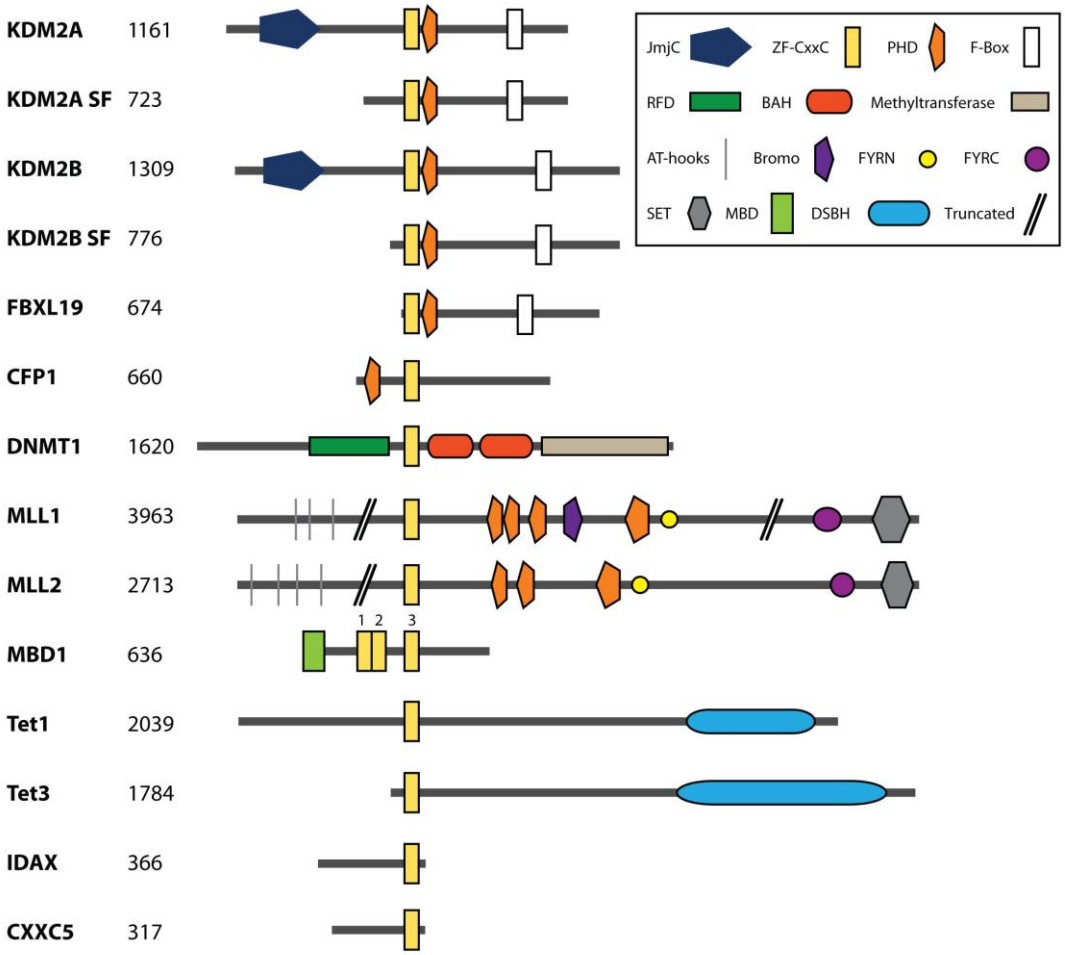


Figure 1.13 A family of ZF-CxxC domain containing proteins

A schematic of the domain architecture for all 12 mouse ZF-CxxC domain containing proteins. The proteins are drawn to scale with the number of amino acids in the protein indicated on the left. All proteins are centred at the ZF-CxxC domain with the N-terminus on the left. In the case of KDM2A and KDM2B, alternative downstream promoters give rise to short forms of each protein (SF). For MBD1, the three ZF-CxxC domains are numbered 1-3 and the protein was aligned using the third ZF-CxxC domain with non-methylated CpG DNA-binding activity. Domain annotation was performed using a sequence search from the Pfam database (http://pfam.sanger.ac.uk/). All sequences apart from for TET3 were taken from NCBI. NCBI reference sequence: KDM2A NP_001001984.2; KDM2A SF (*Homo sapiens*) NP_001243334.1; KDM2B NP_001003953.1; KDM2B SF NP_038938.1; FBXL19 NP_766336.2; CFP1 NP_083144.1; DNMT1 NP_001186360.2; MLL1 NP_001074518.1; MLL2 NP_083550.2; TET1 NP_001240786.1; TET3 NP_898961.2; IDAX NP_001004367.2. GenBank: MBD1 AAC68869.1; CXXC5 AAH89314.1. From publication: TET3 (Iyer et al., 2009).

1.9.2 How the ZF-CxxC domain binds to non-methylated CpG DNA

Recent crystal structures have revealed how the short primary sequence (35-42 amino acids) of the type-1 ZF-CxxC domain enables precise recognition of a CpG dinucleotide (Allen et al., 2006; Cierpicki

et al., 2010; Song et al., 2011; Xu et al., 2011a; Xu et al., 2012). The two cysteine-rich clusters coordinate two Zn^{2+} ions in a tetrahedral manner and stabilise the ZF-CxxC domain in an extended crescent-shaped structure (Figure 1.14B). Lying perpendicular to the DNA axis, the domain interrogates the major groove via a DNA-binding loop while regions flanking the ZF-CxxC domain reach around to interact with the minor groove (Figure 1.14B and Figure 1.15A). This requirement for access to both faces of DNA is reflected in the preferential binding of the ZF-CxxC domain to nucleosome-free linker DNA, presumably because the interaction between nucleosomal DNA and histone proteins precludes interaction with both the major and minor groove of DNA (Zhou et al., 2012).

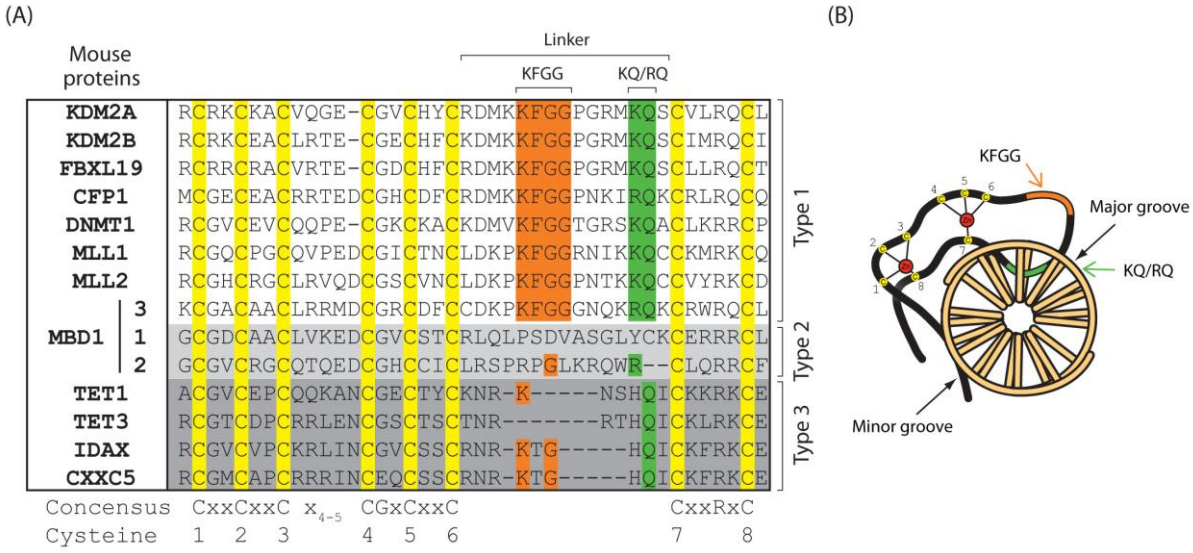


Figure 1.14 Primary sequence variation in ZF-CxxC domains

(A) A manually curated multiple sequence alignment of all ZF-CxxC domains from mouse. ZF-CxxC domains can be split into three types depending on their sequence similarity, labelled as type-1, -2 and -3 on the right of the alignment. Eight cysteine residues 'C' 1-8 are fully conserved across all of the ZF-CxxC domains (yellow). In the linker region between the two cysteine rich clusters (1-6 and 7-8), a lysine-phenylalanine-glycine-glycine 'KFGG' motif is conserved across the type-1 ZF-CxxC domains (orange) and a lysine-glutamine or arginine-glutamine 'KQ'/'RQ' DNA binding motif which binds specifically to the CpG dinucleotide is present in all of the type-1 ZF-CxxC domains (green), but is lost in the type-2 ZF-CxxC domains and substituted to histidine followed by glutamine 'HQ' in the type-3 ZF-CxxC domains. Notably, while the type-3 ZF-CxxC domains are truncated in the linker region between the cysteine-rich clusters, the linker length is retained in the type-2 ZF-CxxC domains, but the sequence similarity to type-1 ZF-CxxC domains is completely lost. (B) A cartoon of the ZF-CxxC domain highlighting the crescent structure and interaction with both the major and minor groove of DNA. Zinc ions (red) coordinate the eight cysteine residues (yellow, C 1-8). The KQ/RQ motif region (green) of the domain wedges into the major groove of the DNA while the N-terminal (NT) and C-terminal (CT) regions of the ZF-CxxC domain interrogate the minor groove. The KFGG region (orange) forms part of the linker between the cysteine-rich regions but does not from specific interactions with the CpG dinucleotide. DNA is viewed down the double helix with bases shown as rods.

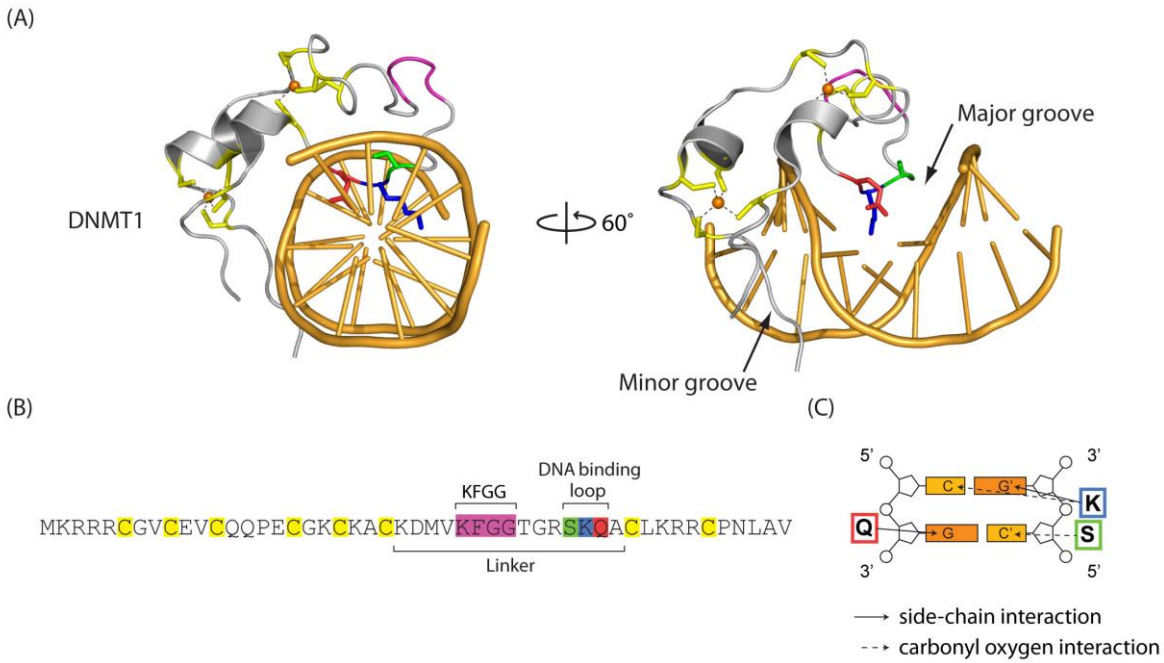


Figure 1.15 Structural insights into the DNA binding properties of ZF-CxxC proteins

(A) Crystal structures of the DNMT1 (*Homo sapiens*; PDB ID: 3PTA) ZF-CxxC domain in complex with DNA, viewed down the double helix axis (left) and rotated 60° to the right (right). The eight cysteine residues are highlighted in yellow and their interaction with the Zn^{2+} ions (represented as orange spheres) are shown by dashes. The KFGG motif is highlighted in pink and the DNA binding tripeptide is highlighted in green, blue and red for S, K and Q respectively. The DNA-binding tri-peptide loop interrogates the CpG dinucleotide via the major groove of DNA (right), while the N- and C-terminus of the ZF-CxxC domain interact with the minor groove. (B) The amino acid sequence of the DNMT1 ZF-CxxC domain; residues are highlighted as in A. (C) Schematic showing the base-specific hydrogen bond interactions between the DNA-binding tripeptide and a CpG dinucleotide. Side-chain interactions are shown by a solid arrow and carbonyl oxygen interactions are shown by a dashed arrow.

Between the two cysteine-rich clusters the variable linker region confers non-methyl CpG binding specificity for the type-1 ZF-CxxC domains (Figure 1.14A). A highly conserved 'KFGG' motif provides rigidity to the ZF-CxxC domain fold while a hydrophilic, positively charged DNA binding loop inserts into the DNA major groove and mediates base-specific interactions via an important tripeptide containing a KQ/RQ motif (Figure 1.15A-B) (Cierpicki et al., 2010; Xu et al., 2011a). The KQ/RQ dipeptide makes specific side-chain and peptide backbone interactions via hydrogen-bonding with two of the cytosines and one guanine from the double-stranded CpG dinucleotide (Figure 1.15C). The remaining guanine is interrogated by the amino acid immediately N-terminal to the KQ/RQ motif via the carbonyl oxygen of the peptide backbone (Figure 1.15C). Importantly, methylation of the C-5 position would cause a severe steric clash and loss of hydrogen bonding leading to loss of binding (Cierpicki et al., 2010; Song et al., 2011; Xu et al., 2011a). The two type 2 ZF-CxxC domains from

MBD1 (CxxC-1 and CxxC-2) lack a functional DNA binding loop (Jorgensen et al., 2004) and instead appear to mediate protein-protein interactions (Lyst et al., 2006; Sarraf and Stancheva, 2004) (Figure 1.14A). Despite loss of the KQ/RQ motif, type-3 ZF-CxxC domains appear able to bind non-methylated CpG via a XHQ tripeptide with a more flexible mode of binding (Xu et al., 2012) (Figure 1.14A, see 0).

1.9.3 *In vivo* functions of ZF-CxxC domain-containing proteins

In vitro binding analyses and structural studies have elucidated the mechanism by which the type-1 and -3 ZF-CxxC domains recognises DNA (Cierpicki et al., 2010; Song et al., 2011; Xu et al., 2011a; Xu et al., 2012). *In vivo* functions have also been elucidated for many of the ZF-CxxC domain-containing proteins and a number of these will be discussed below.

1.9.3.1 **KDM2A, KDM2B and FBLX19**

The lysine demethylase 2 (KDM2) family are JmjC-containing histone demethylases which catalyse the removal of mono- and di-methylation from lysine 36 on the histone H3 tail (H3K36me1/2) (Klose and Zhang, 2007; Tsukada et al., 2006). In addition to the JmjC domain, the KDM2 proteins, KDM2A and KDM2B, encode type-1 ZF-CxxC domains which are able to bind specifically to non-methylated DNA *in vitro* (Blackledge et al., 2010; Farcas et al., 2012).

Profiling KDM2A binding genome-wide in mouse ES cells revealed that KDM2A is enriched at more than 90% of CGIs, including promoter CGIs of both expressed and silenced genes (Blackledge et al., 2010). Therefore KDM2A nucleation appears to be dependent on the DNA binding specificity of the ZF-CxxC domain and independent of the transcriptional state of the associated gene. The substrate of KDM2A, H3K36me1/2, is found on 30-50% of total histone H3, making it one of the most abundant histone modifications in mammalian cells (Garcia et al., 2008; Peters et al., 2003; Robin et al., 2007). The function of this modification is still poorly understood, although work in *Saccharomyces cerevisiae* suggests that H3K26me2 is inhibitory to transcriptional initiation, perhaps through the binding of the EAF3 chromodomain-containing protein and recruitment of the RPD3S

HDAC-containing co-repressor complex (Carrozza et al., 2005; Li et al., 2009). Furthermore, it was recently suggested that H3K36 methylation can prevent histone exchange by inhibiting the interaction between histone chaperones and histone H3 thereby potentially suppressing the capacity of non-regulatory regions to support transcriptional initiation (Venkatesh et al., 2012). Importantly, H3K36me1/2 appears to be specifically depleted at KDM2A-bound CGIs and was observed to increase at CGIs upon depletion of KDM2A (Blackledge et al., 2010). Although the transcriptionally repressive roles of H3K36 methylation have yet to be verified in mammals, it is tempting to speculate that recruitment of KDM2A to CGIs via its ZF-CxxC acts to deplete the repressive H3K36me2 modification from regulatory regions and create a favourable chromatin landscape for the initiation of transcription.

KDM2B, the sister protein to KDM2A, encodes a highly similar domain architecture and is also able to catalyse the removal of H3K36me1/2 (He et al., 2008). KDM2B has been implicated in cellular immortalization, reprogramming and exhibits a transformative capacity in cancer (He et al., 2008; He et al., 2011a; Liang et al., 2012; Pfau et al., 2008; Wang et al., 2011). Furthermore a knockout mouse model and a naturally occurring bovine mutant exhibit developmental defects (Fukuda et al., 2011; Testoni et al., 2012), suggesting that KDM2B has important roles during development and cellular homeostasis. Genome-wide profiling of KDM2B revealed that KDM2B is also enriched at CGIs *in vivo* (Farcas et al., 2012; He et al., 2013; Wu et al., 2013). Intriguingly, KDM2B appeared to be enriched, and KDM2A depleted, at a subset of CGIs which were associated with a developmentally regulated subset of genes. This gene class is often bound and transcriptionally repressed by the polycomb group repressive complexes during development (Simon and Kingston, 2009), suggesting that KDM2B may contribute to polycomb mediated transcriptional repression (Farcas et al., 2012).

In mammals, two classes of polycomb repressive complexes (PRCs) called PRC1 and PRC2 mediate gene silencing via specific histone modifying activities (Mikkelsen et al., 2007; Simon and Kingston, 2009). PRC1 contains the RING1B E3 ubiquitin ligase which mono-ubiquitylates lysine 119 on the histone H2A tail (H2AK119ub), while the PRC2 subunit EZH2 methylates the histone H3 tail at the

lysine 27 position (H3K27me3). Polycomb complexes appear to nucleate at CGI elements (Ku et al., 2008; Lynch et al., 2012), although the mechanisms by which this occurs are incompletely understood. Interestingly, KDM2B has been identified in a variant PRC1 complex containing BCoR, PCGF1, RYBP, YAF2 and RING1B in cancer cells (Gao et al., 2012; Gearhart et al., 2006; Sanchez et al., 2007) and also in mouse ES cells (Farcas et al., 2012) suggesting that this complex may be relevant in a non-malignant context. Given the lack of an apparent sequence-specific DNA binding domain within components of the canonical PRC1 complex, it is possible that the KDM2B ZF-CxxC may act as a targeting module in the variant PRC1 complex. Indeed shRNA-based knockdown of KDM2B lead to a reduction in RING1B occupancy at polycomb-targets and a concomitant increase in expression of some polycomb-repressed genes (Farcas et al., 2012; Wu et al., 2013).

The type-1 ZF-CxxC domain-containing FBXL19 has a very similar domain architecture to KDM2A and KDM2B, lacking only the N-terminal catalytic JmjC domain (Figure 1.13). Interestingly, both KDM2A and KDM2B encode short forms which are transcribed from an alternative transcription start site downstream of the JmjC domain and therefore give rise to proteins which closely resemble FBXL19 (Figure 1.13). The role of the short form of KDM2A and KDM2B along with FBXL19 remain poorly understood, however due to the presence of a presumably functional ZF-CxxC domain it appears likely that they will localise to and affect CGI function.

1.9.3.2 **CFP1**

CFP1 is a component of the mammalian H3K4 methyltransferase SETD1 complex which includes SETD1A or 1B, ASH2L, RbBP5, WDR5 and WDR8, and DPY-30 (Ernst and Vakoc, 2012; Lee and Skalnik, 2005; Lee et al., 2007). In agreement with its capacity to bind non-methylated DNA *in vitro* (Voo et al., 2000), genome-wide profiling of CFP1 in mouse brain revealed that CFP1 associates with more than 80% of CGIs, the majority of which were enriched for H3K4me3 (Thomson et al., 2010). Although the contribution of H3K4me3 to gene expression remains poorly defined, this mark is generally thought to be permissive to active transcription (Bernstein et al., 2005) through the

recruitment of plant homeodomain (PHD) or tudor domain effector proteins (Bian et al., 2011; Blackledge and Klose, 2011; Lauberth et al., 2013; Vermeulen et al., 2007; Wysocka et al., 2006). Importantly, an exogenous CpG-rich DNA sequence lacking gene regulatory features was able to recruit CFP1 and nucleate H3K4me3, apparently in the absence of transcription factors and RNAPII (Thomson et al., 2010), suggesting that non-methylated DNA can recruit CFP1 and lead to modification of H3K4me3. Interestingly however, CFP1 appears to be recruited to a CGI within the alpha-globin locus only in erythroid cell-types in which the alpha-globin genes are active (Vernimmen et al., 2011) therefore suggesting that CFP1 may not responsible for the presence of the H3K4me3 modification at this locus in non-erythroid cells. CFP1 is essential for early mouse development (Carlone and Skalnik, 2001) and CFP1 null mouse ESCs fail to effectively differentiate *in vitro* (Carlone et al., 2005), losing H3K4me3 at up to half of CGIs (Clouaire et al., 2012). Therefore the CFP1 ZF-CxxC appears sufficient to nucleate activating chromatin modifying activity to CGIs, which appears to be essential for embryogenesis and differentiation.

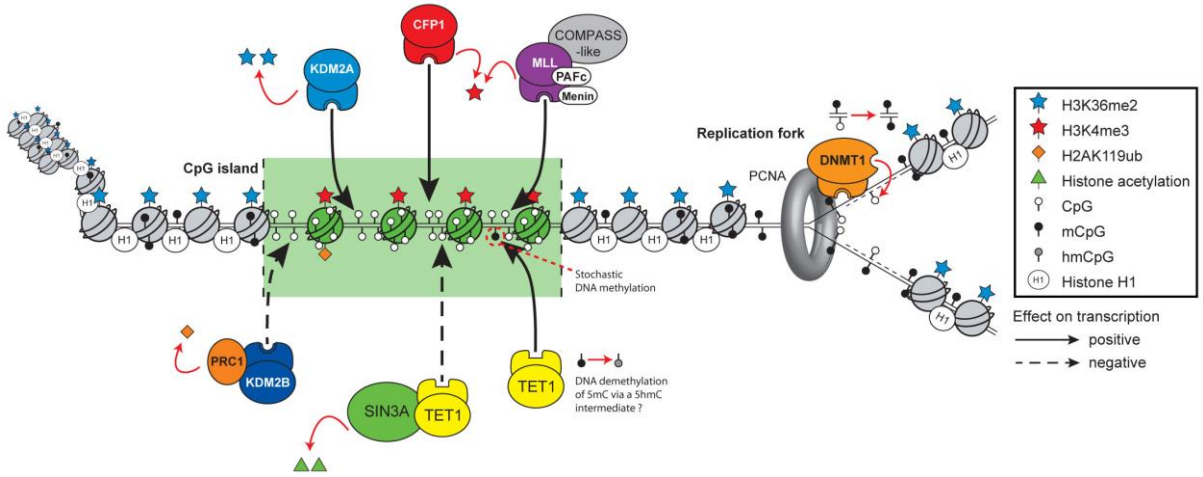


Figure 1.16 The role of ZF-CxxC proteins at CGIs and during DNA replication

A schematic illustrating the convergence of ZF-CxxC domain containing protein function at CGI elements and the role of the ZF-CxxC domain containing DNMT1 during DNA replication.

1.9.3.3 **MLL1 and MLL2**

In addition to CFP1, two other H3K4 methyltransferases, MLL1 and MLL2, members of the mixed-lineage leukaemia protein family contain ZF-CxxC domains able to bind to non-methylated DNA *in vitro* (Bach et al., 2009; Birke et al., 2002). MLL1 and MLL2 share many of the interaction partners of

CFP1 including ASH2L, WDR5, RbBP5 and DPY-30 (Ruthenburg et al., 2007) and similarly play an essential role in early mammalian development and also in definitive haematopoiesis (Ernst et al., 2004; Yu et al., 1995). However, unlike CFP1, MLL1 has been shown to localise to only to a small subset of gene promoters (Guenther et al., 2005). In human lymphoma cells approximately 5000 promoters were observed to be bound by MLL1, coincident with H3K4me3, RNAPII and active transcription (Guenther et al., 2005). Therefore it appears that MLL1 is recruited to CGIs via its ZF-CxxC domain, but is restricted to a subset of CGI elements which are actively transcribed. This reduced CGI binding preference of MLL1 is perhaps mediated via its N-terminal partner, MENIN, which frequently coincides with MLL1 binding sites (Scacheri et al., 2006; Yokoyama et al., 2005), interaction with the PAF1C complex members (Figure 1.16) (Milne et al., 2010; Muntean et al., 2010) or via the action of other chromatin-interacting domains such as the AT-hook (Zeleznik-Le et al., 1994) and PHD domain (PHD3) (Chang et al., 2010; Milne et al., 2010; Park et al., 2010; Wang et al., 2010). Numerous MLL1 protein fusions have been implicated in driving aggressive adult and childhood leukaemias (Meyer et al., 2009). Together with the embryonic lethality observed for MLL1 null mice, this strongly indicates that the correct targeting of MLL1 is essential to drive appropriate gene expression from MLL1 target CGI genes.

1.9.3.4 **DNMT1**

As mentioned in section 1.3.3, DNMT1 contains a functional type-1 ZF-CxxC domain which initially seemed counterintuitive given that CGIs are for the most-part free of DNA methylation and DNMT1 appears to favour a hemi-methylated DNA substrate (Bestor and Ingram, 1983; Chuang et al., 1997). However, structural studies suggest that the DNMT1 ZF-CxxC domain may act as an inbuilt mechanism to help limit the catalytic activity of DNMT1 on unmodified CpG sites (Song et al., 2011) (see section 1.3.3).

1.9.3.5 **MBD1**

The CxxC-3 domain from MBD1 (Figure 1.14A) is able to bind non-methylated DNA *in vitro* (Clouaire et al., 2010; Jorgensen et al., 2004), but does not lead to MBD1 localisation to CGIs (Baubec et al., 2013). MBD1 instead appears to nucleate at sites with methylated CpGs via its methyl-CpG binding domain (Clouaire et al., 2010; Cross et al., 1997; Jorgensen et al., 2004). It is possible that during times of genome-wide hypomethylation, such as in preimplantation embryos, that the CxxC-3 domain may act as a relevant targeting module (Clouaire et al., 2010).

Table 1.2 *ZF-CxxC domain-containing protein nomenclature*

Gene name	CXXC nomenclature	Other names	NCBI Gene ID
Kdm2a	Cxxc8	Fbxl11, Jhdm1a, Ndy2	225876
Kdm2b	Cxxc2	Fbxl10, Jhdm1b, Ndy1	30841
Fbxl19	-		233902
Cfp1	Cxxc1	Cgbp, Phf18	74322
Dnmt1	Cxxc9	Met1	13433
Mll1	Cxxc7	All1, Htrx1, Kmt2a	214162
Mll2	-	Wbp7, Kmt2b	75410
Mbd1	Cxxc3	Pcm1	17190
Tet1	Cxxc6	Lcx	52463
Tet3	Cxxc10		194388
Idax	Cxxc4		319478
Cxxc5	Cxxc5		67393

1.9.3.6 **TET1 and TET3**

The TET protein family has been suggested to form part of the mammalian DNA demethylation system as they are able to convert 5mC to 5hmC (Inoue and Zhang, 2011; Kriaucionis and Heintz, 2009; Tahiliani et al., 2009). TET1 and TET3 encode type-3 ZF-CxxC domains which have lost the KFGG motif and KQ/RQ motif central to specific non-methylated CpG binding (Figure 1.14A). There are conflicting reports whether the TET enzymes bind non-methylated cytosine (with a slight preference for CpG dinucleotides) (Xu et al., 2012), bind CpG dinucleotides irrespective of methylation status (Xu et al., 2011b; Zhang et al., 2010a) or exhibit no sequence preference (Frauer et al., 2011). *In vivo* evidence suggests that in some cases type-3 ZF-CxxC domains may act as a CGI-

targeting module as a number of studies have profiled TET1 genome-wide and generally concluded that TET1 is enriched at gene promoters with a positive correlation for CpG density (Williams et al., 2011; Wu et al., 2011b; Xu et al., 2011b).

1.9.4 Non-methylated DNA, histone modifications and the impact on gene expression

It is becoming clear that non-methylated DNA can recruit both activatory and repressive chromatin modifying enzymes to CGIs via nucleation of ZF-CxxC domain-containing proteins (Blackledge et al., 2010; Farcas et al., 2012; Guenther et al., 2005; Thomson et al., 2010). H3K4me3 is placed at CGIs by the ZF-CxxC domain-containing CFP1 and MLL1/2 complexes. Although the contribution of H3K4me3 to gene expression remains poorly defined, this mark is generally thought to be permissive to active transcription through the recruitment of plant homeodomain (PHD) or tudor domain effector proteins (Bian et al., 2011; Blackledge and Klose, 2011; Lauberth et al., 2013; Vermeulen et al., 2007; Wysocka et al., 2006). Furthermore, KDM2A nucleation at non-methylated CGIs catalyses the removal of the transcriptionally repressive H3K36me2 mark (Blackledge et al., 2010) while TET 5-methylcytosine dioxygenases have been proposed to bind to CGIs and remove spurious DNA methylation (Long et al., 2013; Williams et al., 2011; Wu et al., 2011b; Xu et al., 2011b) (Figure 1.16). Together, a number of ZF-CxxC targeted chromatin modifiers appear to establish and reinforce a transcriptionally permissive chromatin environment at CGIs (Blackledge and Klose, 2011).

Interestingly, opposing repressive chromatin modifiers have also recently been reported to be targeted to CGIs. For example, polycomb-repressive activities appear to be recruited to CGIs (Lynch et al., 2012). In fact, nearly all H3K27me3-marked polycomb targets are CGIs (Mikkelsen et al., 2007) and in ES cells genome-wide H3K27me3 appears to anti-correlate with DNA methylation (Hagarman et al., 2013). In vertebrates, the mechanisms leading to the placement of H3K27me3 by polycomb repressive complex 2 (PRC2) are still unclear, especially given that none of the components (Ezh2, Eed, Suz12, Rbap46/48, Jarid2, Aebp2 or PHC1-3) contain a known non-methylated DNA binding capacity, although Jarid2 has some preference for binding to GC-rich sequences (Mendenhall et al.,

2010). PRC1 was previously thought to be recruited to DNA in a hierarchical manner through recognition of the PRC2 H3K27me3 modification (Min et al., 2003; Wang et al., 2004). However, this mechanism has recently been challenged as PRC1 can be recruited to targets independently of PRC2 (Puschendorf et al., 2008; Schoeftner et al., 2006; Tavares et al., 2012). Instead, a variant PRC1 complex containing the ZF-CxxC domain-containing KDM2B protein has been identified which appears to direct PRC1 binding to CGIs (Farcas et al., 2012; He et al., 2013; Wu et al., 2013) (Figure 1.16). Should PRC1 then be able to recruit PRC2, this could constitute a mechanism for establishing full polycomb-based silencing at a specific CGI. Furthermore, an indirect mechanism for polycomb recruitment at CGIs has been suggested whereby the lack of H3K36me2/3 at CGIs, which would normally inhibit PRC2 activity, is permissive for the establishment of polycomb repression at CGIs (Klose et al., 2013; Schmitges et al., 2011).

By whichever means H3K27me3 is directed to CGIs, this modification is associated with repression of transcription (Simon and Kingston, 2013), perhaps through directing local chromatin compaction (Grau et al., 2011) or disassembly of the RNAPII pre-initiation complex by interfering with Mediator (Lehmann et al., 2012). Therefore, the presence of H3K27me3 at a CGI will lead to gene silencing. Furthermore, the TET enzymes have been shown to play a repressive role at CpG-rich promoters, through the direct recruitment of the SIN3A co-repressor complex (Williams et al., 2011). Together, regions of non-methylated DNA appear to be subject to both activating and repressive chromatin modifications. It has therefore been suggested that CGIs may in fact facilitate a bistable chromatin state, dominated either by transcriptionally positive or negative chromatin marks, which is ultimately responsive to levels of transcription factors or the expression state of the associated gene (Klose et al., 2013; Lynch et al., 2012).

1.9.5 Non-methylated DNA and nucleosome occupancy

It has been reported that CGIs may create a nucleosome depleted region, perhaps due to their elevated GC content, and this has been proposed to support poised transcription or ease of gene

activation (Fenouil et al., 2012; Ramirez-Carrozzi et al., 2009). However, other studies have revealed that active promoters and enhancers are in fact marked by nucleosomes containing the histone variants H2AZ and H3.3 which are highly unstable and are therefore lost during preparative methods used in other studies (Jin and Felsenfeld, 2007; Jin et al., 2009). Nucleosomes containing both H2AZ and H3.3 appear to rapidly exchange at gene regulatory regions (Jin et al., 2009). This suggests that CGI promoters may be associated with unstable double variant nucleosomes which may facilitate DNA binding by transcription factors and chromatin modifiers and therefore enable switching between an active and repressed chromatin state (Klose et al., 2013).

1.10 **DNA methylation, remodelling the nucleosome and links to human disease**

Along with the 'nucleosome-free regions' reported at CGIs, nucleosome remodelling has also been implicated in modulating DNA methylation, and disruption of this process has been linked with human disease. These diseases, and other human conditions, have provided insights into the importance of appropriate regulation of DNA methylation in normal cellular development and homeostasis.

1.10.1 Chromatin remodelers, DNA methylation and disease

1.10.1.1 **DDM1 and LSH**

It has been known for some time that chromatin remodelling is intimately linked with DNA methylation (Meehan et al., 2001). The first evidence for this came from a screen in *A. thaliana* aiming to identify genes required for the maintenance of DNA methylation. This revealed a gene named DDM1 (decrease in DNA methylation) which reduced global DNA methylation by 70 % (Vongs et al., 1993), causing progressive reduction of DNA methylation over a number of generations at a number of loci. This suggested that the *Ddm1* mutation impairs the ability to propagate DNA methylation after DNA replication (Kakutani et al., 1996). A highly homologous protein LSH was identified in mouse which is expressed ubiquitously during fetal development (Raabe et al., 2001). *Lsh*$^{-/-}$ mice exhibit up to 70% loss of global DNA methylation in fetal and new born pups (Dennis et

al., 2001; Geiman et al., 2001), including loss from major and minor satellite repeats (Dennis et al., 2001; Muegge, 2005), and the mice exhibit postnatal lethality (Geiman et al., 2001). DDM1 and LSH are closely related to the ISWI family of chromatin remodelers (Dennis et al., 2001; Geiman and Muegge, 2000; Meehan et al., 2001) which are able to disrupt the interaction between histones and DNA (Jeddeloh et al., 1999). Therefore, as cellular DNA methyltransferase activity appears unaffected in *ddm1* or *Lsh* mutants (Dennis et al., 2001; Kakutani et al., 1995), it was proposed that LSH remodels chromatin, to regulate the accessibility of DNA to the *de novo* methyltransferases (Brzeski and Jerzmanowski, 2003; Meehan et al., 2001) . Furthermore, LSH may also function to recruit DNMT3A, DNMT3B and G9a/GLP directly to establish a transcriptionally repressive chromatin domain (Myant and Stancheva, 2008; Myant et al., 2011; Zhu et al., 2006). LSH has been detected to be over-expressed in some cancers, and may be essential for bypassing cellular senescence (Keyes et al., 2011).

1.10.1.2 **ATRX**

ATRX is another member of the SNF2-like family of chromatin remodellers which is recruited to heterochromatic repeats, including centromeres, telomeres and ribosomal DNA (Law et al., 2010; McDowell et al., 1999). Targeting of ATRX involves interactions with MeCP2 (Nan et al., 2007), HP1 (Lechner et al., 2005) and interrogation of H3K9me3 and H3K4me0 histone modifications by the ATRX ADD domain (ATRX-DNMT3-DNMT3L domain) (Dhayalan et al., 2011; Eustermann et al., 2011; Iwase et al., 2011). ATRX has also been observed to bind to G-rich regions of the genome, which have a tendency to form a non-B form DNA structure called a G-quadruplex (G4) (Lipps and Rhodes, 2009). ATRX is proposed to resolve these G4 structures in tandem repetitive DNA in partnership with DAXX through deposition of the histone variant H3.3 (Clynes et al., 2013; Drane et al., 2010; Elsasser et al., 2012; Gibbons and Higgs, 2010; Lewis et al., 2010; Shay et al., 2012).

Mutations in ATRX, gives rise to alpha-thalassemia mental retardation (ATR-X) syndrome, a very rare X-linked disease. Disease-related mutations are clustered in the two functional domains, the N-

terminal ADD domain and a C-terminal helicase domain, which may impair nuclear localisation, protein-protein interactions or chromatin remodelling functions (Cardoso et al., 2000; Gibbons et al., 1997; Mitson et al., 2011; Tang et al., 2004). Interestingly, diverse changes in DNA methylation are commonly observed in ATRX patients including hypermethylation at the DYZ2 Y-chromosome repeat and hypomethylation at ribosomal DNA repeats (Gibbons et al., 2000). However, how loss of function of ATRX causes these changes is incompletely understood. Mutations in ATRX have also been observed in a number of malignancies (Gibbons et al., 2003; Steensma et al., 2004), although whether DNA methylation is affected in these tumours remains to be seen (Clynes et al., 2013).

1.10.2 DNA methylation and other human diseases

1.10.2.1 **ICF syndrome**

Mutations in the C-terminal catalytic or PWWP domain of DNMT3B have been associated with the extremely rare autosomal recessive ICF (immunodeficiency, centromeric region instability, and facial anomalies) syndrome mutants in humans. ICF is the only known genetic disease in humans known to involve a DNA methyltransferase enzyme, and mutations are proposed to reduce, but not abolish, catalytic activity (Gowher and Jeltsch, 2002; Ueda et al., 2006). Patients with ICF exhibit aberrant hypomethylation, chromatin decondensation and distinctive chromosomal rearrangements at pericentromeric heterochromatin of chromosome 1 and 16 and hypomethylation of satellite repeat II (Ehrlich et al., 2008). These abnormalities are almost exclusively observed in stimulated B cells (Ehrlich, 2003; Xu et al., 1999), therefore ICF syndrome can provide insights into the impact of DNA methylation on the maturation and activation of lymphocytes (Ehrlich, 2003).

1.10.2.2 **Imprinting control related diseases**

DNA methylation plays an important role in the mammalian germline in ensuring that for a subset of genes only one allele is expressed, a phenomenon known as imprinting (Kaneda et al., 2004). Failure to accurately methylate the correct parental allele at imprinted genes has been causally linked to many human diseases including Angelman's syndrome (AS) and Prader-Willi syndrome (PWS) (Lim

and Maher, 2009) which result from maternal or paternal deficiency of the same chromosomal fragment (Horsthemke and Wagstaff, 2008; White et al., 2006). Deficiencies in imprinting such as those occurring for AS and PWS illustrate that abnormal patterns of DNA methylation at specific loci are sufficient to cause human disease.

1.10.2.3 **Rett syndrome**

Rett syndrome, the most common form of mental retardation in females (Amir et al., 1999), is caused in 96 % of cases by spontaneous mutation of the methyl DNA binding protein MeCP2 (Moretti and Zoghbi, 2006) (see section 1.7.2). Mutations cluster in the MBD and transcriptional repressor domain (TRD, Figure 1.11) (Kriaucionis and Bird, 2003) which impairs binding to DNA methylation and other chromatin modifiers including the NCoR co-repressor complex (Lyst et al., 2013), resulting in the inability to appropriately silence genes in a subset of neurons during development (Amir et al., 1999; Guy et al., 2001). This therefore suggests that the ability to recognise and interpret patterns of DNA methylation is crucial for normal cell function, and that MeCP2 may be essential in neuronal lineages for synaptic function and neuronal plasticity (Moretti and Zoghbi, 2006).

1.10.2.4 **Cancer**

Deregulation of DNA methylation is a very common trait of cancer. In heterochromatic regions, aberrant hypomethylation causes derepression of repetitive regions of the genome and reactivation of mobile elements. An increase in mutation rate and incidence of chromosomal translocations can lead to the subsequent silencing of tumour repressors or activation of oncogenes in mice (Eden et al., 2003; Gaudet et al., 2003). Acquisition of DNA methylation at tumour suppressor genes, such as *CDKN2A*, *DAPK1* and *CDH1* can also lead to tumourigenic features including increased proliferation, invasiveness and escape from apoptosis (Kim et al., 2001; Merlo et al., 1995; Nakata et al., 2006). Conversely, hypomethylation of CGIs can also lead to abnormal expression of oncogenes, such as *NAT1* in breast cancer (Kim et al., 2008). The mechanisms which underlie DNA methylation

alterations in cancer are still under investigation. However, recent studies indicate that DNA methylation changes may occur very early in cancer development and in some cases contribute to cancer initiation (Baylin and Jones, 2011). The study of human cancer therefore provides insight into the importance of appropriate gene expression control, and the role of DNA methylation in this process both at the genome-scale and at individual loci.

1.11 **Experimental methods for identifying the methylation status of DNA**

Many methods have been developed for the analysis of DNA methylation state, a number of which will be summarized here, distinguishing between techniques to detect methylated and non-methylated DNA.

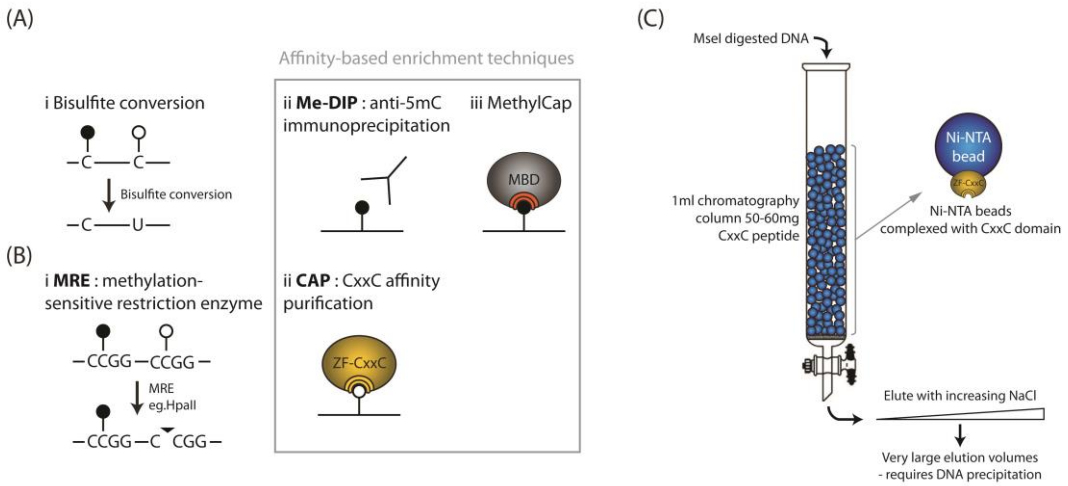


Figure 1.17 Methods for detection of DNA methylation status

(A) Techniques for the detection of methylated DNA. (i) Bisulphite conversion converts non-methylated cytosines, but not methylated cytosines, to uracil. (ii) Me-DIP and (iii) MethylCap (Methyl-CpG capture) rely on affinity for 5-methylcytosine to purify methylated DNA using an anti-5mC antibody or the MBD domain from MeCP2 respectively. (B) Techniques to detect non-methylated DNA include (i) MRE (methylation-sensitive restriction enzyme) which utilises the fact that methylation-sensitive restriction enzymes can only digest DNA when their recognition sequence is non-methylated and (ii) CAP (CxxC-affinity purification) utilises the ZF-CxxC domain to specifically purify non-methylated DNA. (C) The original CAP protocol utilised the MBD1 ZF-CxxC domain complexed to Ni-NTA beads via a His-tag to purify non-methylated DNA. The CAP method was performed on a chromatography column using MseI digested DNA. Elution of non-methylated DNA was performed using increasing concentration of NaCl, with non-methylated DNA eluting at higher salt concentrations. Methylated cytosine is depicted as a black filled circle and non-methylated cytosine as an open white circle.

1.11.1 Profiling methylated DNA

1.11.1.1 **Detection of total genomic DNA methylation levels**

Following total acid hydrolysis of DNA, base composition can be precisely quantified by high-resolution separation using a number of highly sensitive techniques. High-performance liquid chromatography (HPLC) allows quantitative determination of the relative levels of cytosine and 5-methylcytosine (Armstrong et al., 2011; Oakeley, 1999). This and other methods including thin layer chromatography (TLC) (Dahl and Guldberg, 2003) have proven extremely useful for analysis of the methylation status of entire genomes, for example in cancer cell lines or during early development (Lyko et al., 2000; Rhee et al., 2000), however they do not yield locus-specific information.

1.11.1.2 **Affinity-based purification of methylated DNA**

There are a number of DNA methylation profiling techniques based on direct affinity purification of methylated DNA, including the use of anti-5-methylcytosine antibodies (methylated DNA immunoprecipitation (MeDIP)) or recombinant methyl binding domain proteins (methylated DNA capture by affinity purification (MethylCap), or MBD-isolated Genome Sequencing (MiGS)) (Figure 1.17Aii-iii) (Bock et al., 2010; Bogdanovic et al., 2011; Brinkman et al., 2010; Cross et al., 1994; Down et al., 2008; Kangaspeska et al., 2008; Serre et al., 2010). These methodologies, coupled to massively parallel sequencing, directly profile DNA methylation genome-wide (Down et al., 2008; Serre et al., 2010). As the majority of CpGs are methylated in vertebrate genomes, methyl-CpG affinity approaches require a relatively high sequencing depth, and it can be challenging to assess methylation status in regions sparse in CpGs (Weber et al., 2007). Algorithms which compensate for CpG density are therefore used to detect genomic regions with low levels of DNA methylation (Down et al., 2008; Pelizzola et al., 2008).

1.11.1.3 **Bisulfite conversion**

A leap forward in the resolution of DNA methylation detection was the development of bisulfite-induced modification of genomic DNA (Frommer et al., 1992). Treatment of genomic DNA with sodium bisulfite causes deamination of non-methylated cytosines to uracil, while methylated cytosines are unchanged (Figure 1.17Ai). Sequencing of bisulfite-converted DNA followed by comparison with an unconverted reference sequence therefore allows mapping of 5-methylcytosine at base-pair resolution (Zilberman and Henikoff, 2007). It is now relatively common for bisulfite sequencing to be coupled to massively parallel sequencing to generate genome-wide DNA methylation maps (Cokus et al., 2008; Lister et al., 2009; Meissner et al., 2008; Popp et al., 2010), however the depth of sequencing required renders bisulfite sequencing prohibitively expensive for the majority of applications. Furthermore, it can be challenging to align bisulfite converted reads, and methylation 'aware' mapping methods can lead to an increase in discarded reads and a potential bias in alignment (Krueger et al., 2012). Reduced representation bisulfite sequencing (RRBS) offers a method to reduce the sequencing requirements of bisulfite sequencing by incorporation of an initial purification step (Meissner et al., 2008). However this method is biased towards CpG-rich regions of the genome and does not report the methylation state of CpGs in CpG-poor regions of the genome.

1.11.2 Profiling non-methylated DNA

The non-methylated fraction of the genome represents only around 1-2 % of total genomic DNA in vertebrates and the clear association of non-methylated DNA with gene regulatory elements has gained much interest over recent years. Techniques have therefore been developed to directly interrogate non-methylated DNA genome-wide as this can reduce sequencing depth requirements and simplify downstream analysis.

1.11.2.1 **Methylation-sensitive restriction enzyme digestion**

A number of restriction endonucleases are inhibited by the presence of cytosine methylation within their recognition site and these enzymes have been exploited for detection of DNA methylation at both the genome scale and at specific loci. Methylation-sensitive restriction enzymes (MSREs) can only digest non-methylated sites (Laird, 2010), while methylation-insensitive isoschizomers are able to digest the same sites irrespective of the DNA methylation state (Fouse et al., 2010). DNA digested with an MSRE can be analysed by gel electrophoresis to determine genome-wide levels of DNA methylation, or probed by Southern blot for locus-specific analysis (Bird et al., 1985; Cooper et al., 1983). More recently, methylation-sensitive restriction enzyme digestion of DNA has been coupled to whole-genome sequencing, whereby small double-cut DNA fragments are released from the bulk genome and sequenced (methylation sensitive restriction enzyme sequencing, MRE-seq, Figure 1.17Bi) (Harris et al., 2010; Maunakea et al., 2010). MRE-seq has an inherent bias towards regions of the genome containing restriction sites, although this can be partially overcome by digestion with multiple MSREs and combining the samples prior to sequencing (Harris et al., 2010).

1.11.2.2 **Affinity-based purification of non-methylated DNA**

In order to directly profile the non-methylated fraction of the genome, Adrian Bird and colleagues recently developed a novel approach to profile non-methylated DNA called CxxC-affinity purification (CAP, Figure 1.17Bii and C) (Illingworth et al., 2008). This method exploits a zinc finger CxxC (ZF-CxxC) domain from methyl binding domain 1 (MBD1) which specifically recognises non-methylated CpG dinucleotides (Clouaire et al., 2010; Jorgensen et al., 2004). Therefore, CAP specifically purifies a much smaller fraction of the genome compared to the previously described affinity purification methods for methylated DNA. Therefore, CAP coupled to massively parallel sequencing requires a relatively low sequencing depth to yield sufficient coverage to accurately profile non-methylated regions of DNA genome-wide (Illingworth et al., 2010).

1.12 **Aims of this thesis**

DNA methylation is the best studied of the known epigenetic modifications; however it is becoming increasingly apparent that unmodified CpG dinucleotides have an equally important role in modulating genome function. In all vertebrate species that have been examined, DNA methylation appears to be pervasive throughout the genome (Tweedie et al., 1997), however the specific genomic location of non-methylated DNA has only been carefully defined in mammalian species (Illingworth et al., 2010; Laurent et al., 2010; Stadler et al., 2011). In non-mammalian vertebrate species, CGI predictions have often been used as a proxy for non-methylated DNA and the location of these elements appears divergent between warm and cold-blooded vertebrates (Aïssani and Bernardi, 1991; Sharif et al., 2010). Therefore the overarching aim of this thesis is to better understand the function and usage of non-methylated DNA across vertebrate species.

To experimentally profile non-methylated DNA genome-wide, a modified CAP technique was developed to affinity purify the non-methylated fraction of the genome called biotinylated ZF-CxxC affinity purification (Bio-CAP, described in detail in Chapter 3). To address whether usage of non-methylated DNA between vertebrate species is similar or divergent, non-methylated DNA maps were generated for seven diverse vertebrates and for distinct tissue types. The features of vertebrate non-methylated DNA are described in Chapters 4 and 5. The mechanism by which non-methylated regions of DNA are recognised and established in vertebrate genomes remains unresolved. To address this problem, the establishment of non-methylated DNA profiles was examined by introducing heterologous DNA sequences from one organism into another. The aim was to investigate how DNA sequences, which are non-methylated in the host species, are interpreted in a foreign nuclear environment (Chapter 6). The final aim of this thesis was to dissect the functions of non-methylated DNA during early development. DNA methylation is known to play important roles during early mammalian development; however, due to the limitations of studying embryogenesis in mouse, the zebrafish was explored as an alternative model system.

Together, this work reveals important and conserved features of vertebrate non-methylated DNA, indicates potential mechanisms for the establishment of these genomic elements and reveals that the accurate regulation of non-methylated DNA is essential for early zebrafish development. Future work will further investigate the molecular mechanisms behind the proposed functions of non-methylated DNA in the modulation of the chromatin landscape, regulation of gene expression and establishment of totipotency.

Chapter 2 -- Materials and Methods

2.1 **Zebrafish maintenance and manipulation**

2.1.1 Zebrafish husbandry

Zebrafish (*Danio rerio*) were reared and maintained in the BMS facility at the University of Oxford in flowing system water at 28.5°C with conductance in the range of 450-550 µS and pH 7.1-7.2. The light was cycled with 14 hr periods of light and 10 hr of dark. Zebrafish were fed three times a day (by BMS staff) with flake food, high protein pellets and enriched brine shrimp.

2.1.2 Collecting and staging zebrafish embryos

Zebrafish embryos were collected by placing transparent boxes filled with marbles into tanks in the evening to induce natural mating the following morning. Tanks in light boxes enabled matings to occur at different times during the day by shifting the timing of the light cycle. Pair matings were set-up in mating tanks in the evening; the next day, mating was induced by removing a divider between the male and female zebrafish. Embryos were collected in a small sieve and rinsed before being placed in a petri dish with system water and methylene blue to prevent fungal growth. Zebrafish embryos were grown at 22 – 32°C in system water or E3 medium. Embryos were examined and staged under a dissecting microscope according to (Kimmel et al., 1995).

2.1.3 Microinjection of morpholino or RNA into zebrafish embryos

Following natural mating, embryos were sorted under a dissecting microscope. Morpholino or RNA was diluted in milliQ water and injected into the yolk beneath the cell(s) of embryos at the 1-4 cell or 1-2 cell stage respectively.

For zebrafish embryo injection, needles were pulled from borosilicate glass capillary tubes (Harvard Apparatus, 1mm outer diameter, 0.58 inner diameter) using a P-97 Flaming/Brown Micropipette puller (Intracel Sutter Instrument Co.) and back-filled with the solution to be injected. Injection bubble size was calibrated in oil at 11.25 x magnification using a graticule. Staged embryos were

lined up along a glass slide placed inside a petri dish lid. Embryos were injected typically with 0.5-1 nL of RNA or morpholino. For titrations, embryos were injected up to three times.

2.1.4 Dechorionation of embryos

Embryos collected for *in situ* hybridisation were first fixed in 4% paraformaldehyde for at least 3 hr at room temperature before being washed into PBSTw (PBS Tween). The embryos were then manually dechorionated in PBSTw in a petri dish using forceps. Embryos collected for western blot or chromatin immunoprecipitation were manually dechorionated on an agarose-coated petri dish (1% agarose in E3 medium) using forceps.

2.2 **Cell culture**

2.2.1 V6.5 mouse embryonic stem cells for Bio-CAP experiments

Murine V6.5 embryonic stem (ES) cells (C57BL/6 (F) x 129/sv (M)) were cultured on inactivated MEFs in DMEM supplemented with 15% fetal calf serum (FCS), leukaemia-inhibiting factor, penicillin/streptomycin, L-glutamine and non-essential amino acids. Prior to genomic DNA extraction for Bio-CAP experiments, V6.5 cells were cultured for two passages under feeder-free conditions on 0.1% gelatin. Culturing and collection of V6.5 embryonic stem cells was performed by Dr Neil Blackledge.

2.2.2 E14 mouse embryonic stem cells for zebrafish BAC integration

Feeder-independent murine E14 mouse ES cells were grown on 0.1% gelatine in DMEM media supplemented with 15% fetal calf serum (FCS), leukaemia-inhibiting factor (LIF), penicillin/streptomycin, L-glutamine and non-essential amino acids.

2.2.3 293T cells for zKdm2b antigen overexpression

293T cells were grown in DMEM supplemented with 10% fetal calf serum (FCS) and penicillin/streptomycin. Transient transfection of 293T cells was performed using lipofectamine. 5 µg of plasmid DNA was transfected into 293T cells using Lipofectamine 2000 (Invitrogen, Carlsbad, CA).

Cells were grown for 12 hr without penicillin/streptomycin and for a further 12 hr with penicillin/streptomycin before harvesting.

2.3 **Obtaining tissue samples**

Dissected tissue samples were immediately frozen on dry ice and stored at -80°C until genomic DNA extraction. Genomic DNA was stored at -20°C.

2.3.1 *Anolis carolinensis* (green anole lizard)

Anole lizard testis tissue and liver genomic DNA was provided by Jonathan Losos and Shane Campbell-Staton from the Museum of Comparative Zoology, Department of Organismic and Evolutionary Biology, Harvard University, USA.

2.3.2 *Danio rerio* (zebrafish)

Zebrafish tissues were obtained from a mixed background of KCL WT, AB and Tubingen. Zebrafish testis and liver were dissected and immediately frozen on dry-ice. Zebrafish 24 hpf embryos were collected after natural mating and staged according to (Kimmel et al., 1995).

2.3.3 *Gallus gallus* (chicken)

Chicken testis and liver were dissected from an 18 month old male chicken by Mike Colley from FAI Farms and immediately frozen on dry-ice.

2.3.4 *Homo sapiens* (human)

Genomic DNA from human adult male normal tissue was obtained from Stratech (US Biological). Samples were from the following lot numbers T5595-1681 (Tissue, Genomic DNA, Human Adult Normal, Testis) and T5595-1663 (Tissue, Genomic DNA, Human Adult Normal, Liver).

2.3.5 *Mus musculus* (house mouse)

Mouse tissue was obtained from a cross between a female C57Bl/6 and male CBA/ca mouse (B6CF1). Mouse liver and testis tissue were dissected by Jonathan Godwin. Sperm was isolated from mouse testis by collection of live sperm able to swim into PBS at 37°C. Tissues were either

immediately processed for genomic DNA extraction or frozen on dry-ice. Mouse V6.5 ES cells were provided by Dr Neil Blackledge.

2.3.6 *Ornithorhynchus anatinus* (platypus)

Platypus testis and liver genomic DNA was provided by Frank Grützner and Megan Wright from The Robinson Institute, University of Adelaide, Australia.

2.3.7 *Xenopus tropicalis* (African clawed frog)

Frog testis and liver tissue was provided by Anna Noble from the School of Biological Sciences, University of Portsmouth. Stage 11–12 frog embryos were collected at the Developmental Biology Division, MRC National Institute for Medical Research with the help of Anita Abu-Daya. All samples were frozen on dry ice prior to genomic DNA extraction.

2.4 **Preparation of genomic DNA**

Fresh or frozen tissue was subjected to manual homogenisation and DNA was purified either by phenol chloroform extraction or the QIAGEN 100/G genomic tip kit.

2.4.1 Genomic DNA extraction from somatic tissue

25 mg of tissue was homogenised in a 1 mL dounce with 950 µL of genomic DNA (gDNA) extraction buffer (without SDS, 2.19.2). 0.5 % SDS and 200 µg/mL Proteinase K were added and the sample incubated at 50°C for 3 hr. 1/100 volumes of RNaseA/T1 was added and the sample was incubated at 37°C for 1 hr. A further 1 mL of extraction buffer was added to the sample (without SDS or Proteinase K) and the genomic DNA was purified by phenol chloroform extraction. Three extractions were performed using equal volumes of phenol:chloroform:isoamylalcohol (25:24:1 saturated with 10mM Tris pH8.0 and 1mM EDTA, Sigma Aldrich), followed by a final extraction with an equal volume of chloroform. Samples were vortexed briefly, separated by centrifugation for 5 min at 2,000 rpm and the aqueous upper fraction moved into a fresh tube. Genomic DNA was precipitated with 2.5 volumes of ethanol and 0.1 volume of 3 M sodium acetate (pH 5.2), washed with 70 % ethanol and air-dried. The genomic DNA was resuspended in milliQ water and the concentration was

determined using a Nanodrop 1000 spectrophotometer (Thermo Scientific) at 260 nm and stored at -20°C.

2.4.2 Genomic DNA extraction from sperm

Sperm heads are resilient to extraction by proteinase K and SDS alone as their membrane is enriched for disulphide bonds (Weyrich, 2001). Therefore for genomic DNA extraction of sperm or testis tissue, the strong reducing agent dithiothreitol (DTT) was added to the gDNA extraction buffer to reduce these disulphide bonds. The extraction protocol follows 2.4.1 using sperm extraction buffer instead of gDNA extraction buffer (DTT is added along with SDS and Proteinase K following homogenisation).

2.4.3 Genomic DNA extraction using the QIAGEN Genomic-tip 100/G kit

Up to 100 mg of embryos or testis tissue or 80 mg of liver was used to prepare genomic DNA with the Genomic-tip 100/G kit (QIAGEN). The manufacturer's instructions were followed for tissue samples. Briefly, tissue was homogenised in a 1 mL dounce using Buffer G2 with RNaseA (final concentration 200 µg/mL). The sample was incubated at 50°C for 2hr with Proteinase K (1 mg/mL final concentration). Following a 10 s vortex and centrifugation, samples were applied to a 100/G Genomic-tip column pre-equilibrated with Buffer QBT. The column was washed twice with Buffer QC and eluted with 5 mL of Buffer QF (warmed to 50°C). Genomic DNA was precipitated by addition of 0.7 volumes of room temperature isopropanol followed by centrifugation at 4°C. The DNA pellet was washed with 4°C 70 % ethanol, air-dried and resuspended in milliQ water.

2.5 **Extraction, preparation and manipulation of RNA**

2.5.1 Extraction of total RNA

RNA extraction was performed using the RNeasy Mini Kit (QIAGEN) for freshly collected or frozen (-80°C) zebrafish embryos. Up to 40 early stage zebrafish embryos (less than 30 mg) were homogenised by passing 10 times through a 21G needle with a 1 mL syringe in 600 µL RLT buffer. Lysates were centrifuged at 14,000 rpm for 3 min at 4°C and transferred to a fresh tube. 1 volume of

70 % ethanol was added to the lysate, mixed by pipetting, transferred to an RNeasy spin column and centrifuged for 15 s at 8,000 rpm. The columns were washed with 700 µL Buffer RW1, twice with 500 µL Buffer RPE and eluted in 30 µL of RNase-free (DEPC-treated) water. Diethylpyrocarbonate (DEPC) inactivates RNase enzymes, and is heat inactivated prior to use. RNA was quantified using a Nanodrop 1000 spectrophotometer (Thermo Scientific) at 260 nm and stored at -80°C.

2.5.2 Reverse transcription of total RNA to cDNA

RNA (200 ng – 1 µg) was brought up to a volume of 4 µL, 1 µL (0.5 µg) random primers were added and the mixture incubated for 5 min at 70°C followed by 5 min on ice. To this 1X ImProm-II Reaction buffer, 6 mM $MgCl_2$, 0.5 mM dNTPs, 20 u Recombinant RNasin Ribonuclease Inhibitor, and 1 µL ImProm-II Reverse Transcriptase were added. The samples were thermal cycled on a BioRad DNA Engine at 25°C for 5 min, 42°C for 1 hr and 70°C for 15 min. The samples were diluted, typically with 150 – 300 µL milliQ water. cDNA was stored at -20°C. 5 µL diluted cDNA was used for quantitative PCR analysis (see 2.8.1.6).

2.5.3 *In vitro* synthesis of sense mRNA for over-expression or morpholino rescue

Capped mRNA was generated by *in vitro* transcription using the mMessage mMachine kit (Ambion). Expression vectors containing the gene of interest (in a 5’ to 3’ orientation) downstream of an RNA polymerase promoter site (T7, T3 or SP6) were linearised 3’ of the gene and purified using phenol chloroform extraction. Templates were also generated by polymerase chain reaction (PCR) with an RNA polymerase promoter site in the forward primer.

Capped mRNA was *in vitro* transcribed from 1 µg linear plasmid DNA (or 0.2 µg PCR product DNA) using the appropriate T7, T3 or SP6 RNA polymerase enzyme mix, 1X reaction buffer and 1X NTP/CAP analogue in a 20 µL reaction volume for 1 – 3 hr at 37°C. The DNA template was digested using 1 µL TURBO DNase (Ambion) for 15 min at 37°C and the RNA was purified using the RNeasy mini kit (QIAGEN). Purity of the *in vitro* transcribed mRNA was determined by agarose gel

electrophoresis and the concentration was quantified using a Nanodrop 1000 spectrophotometer (Thermo Scientific) at 260 nm. mRNA was stored at -80°C.

2.5.4 *In vitro* synthesis of antisense mRNA probes for *in situ* hybridisation

Expression vectors containing the gene of interest (in a 3' to 5' orientation) downstream of an RNA polymerase promoter site (T7, T3 or SP6) were linearised 5' of the gene and purified. Templates were also generated from polymerase chain reaction (PCR) fragments with an RNA polymerase promoter site in the forward primer (at the 3' end of the gene).

RNA probes were *in vitro* transcribed from 1 µg of linear plasmid DNA (or 0.2 µg PCR product DNA) using the appropriate T7, T3 or SP6 RNA polymerase enzyme mix, 1X reaction buffer, 2 µL DIG-dUTP NTP mix (Roche) and 0.5 µL RNaseOUT (Invitrogen) in a 20 µL reaction volume for 2 – 6 hr at 37°C (or at room temperature overnight). The DNA template was digested using 1 µL TURBO DNase (Ambion) (in 1X TURBO DNase buffer) for 15 min at 37°C. The RNA probe was precipitated following dilution with 80 µL RNase-free water by the addition 0.33 volumes 10 M ammonium acetate and 2.5 volumes ethanol. The RNA probe was washed with 70 % ethanol, air-dried and resuspended in RNase-free water. Purity of the RNA probe was determined by agarose gel electrophoresis before and after DNase treatment. RNA probes were stored at -20°C and reused up to 10 times. Before use the RNA probe was diluted typically to 1:100 – 1:200 in Hybe +/+ for use in whole mount *in situ* hybridisation.

2.6 **Whole mount *in situ* hybridisation (WISH)**

2.6.1 Collecting and fixation of zebrafish embryos

Embryos were staged (according to (Kimmel et al., 1995)), collected and fixed in 4 % paraformaldehyde (PFA) for at least 3 hr at room temperature (or overnight at 4°C). Embryos were sequentially washed with 25 %, 50 % 75 % and 100 % ethanol (ethanol was diluted with PBSTw) and stored in 100 % ethanol at -20°C. Embryos collected before the 18-somite stage were fixed with intact chorions using the same conditions as above, washed with PBSTw for dechorionation and

stored as above. Embryos collected after the 18-somite stage were dechorionated prior to fixing with PFA to prevent fixation of tail curvature.

2.6.2 Whole mount in situ hybridisation

Fixed embryos, stored in 100 % ethanol, were rehydrated with PBSTw by 75 %, 50 % and 25 % ethanol washes (diluted with PBSTw) followed by four PBSTw washes at room temperature. Embryos between 18 somites and 4 days post fertilisation (dpf) were treated with 10 µg/ml proteinase K (Sigma) for 3 to 20 min as required. Embryos collected after 4 dpf were treated with 20 µg/ml proteinase K for 20 to 25 min. Embryos treated with proteinase K were washed twice with PBSTw, fixed once more with 4 % PFA for 20 min at room temperature and washed five times with PBSTw. Embryos collected before the 18 somite stage were not treated with proteinase K. Embryos were transferred from PBSTw into 50 % Hybe +/+ (diluted in PBSTw, containing tRNA and Heparin) for 5 min and then into 100 % Hybe +/+ for at least 1 hr at 65°C. Hybe +/+ was replaced with antisense probes diluted 1:200 in Hybe +/+ and embryos were incubated overnight at 65°C. The eppendorf tubes were positioned horizontally to allow maximal penetration of the embryos with the RNA probe and to prevent the embryos from sticking together.

The following day, RNA probes were removed and kept for future use at -20°C. Embryos were transferred into a mesh-bottomed multi-well plate compatible with a Biolane HT-1 *in situ* machine (Intavis) in Hybe -/- (without tRNA and Heparin). The Biolane HT-1 allowed for automation of temperature control, replacement of solutions and all steps were performed with gentle rocking. Embryos were washed for 10 min each at 65°C in 100 % Hybe -/-, 75 % Hybe -/- 25 % 2 x SSC, 50 % Hybe -/- 50 % 2 x SSC, 25 % Hybe -/- 75 % 2 x SSC and 2 x SSC, and then four times for 15 min each in 0.2 x SSC. Embryos were washed sequentially at room temperature for 5 min each in 75 % 0.2 x SSC 25 % MABTw, 50 % 0.2 x SSC 50 % MABTw, 25 % 0.2 x SSC 75 % MABTw and 100 % MABTw. To block potential non-specific binding sites for the antibody, embryos were first blocked with MAB block (containing Boehringers Blocking Reagent) for 1 hr at room temperature before incubation with

1:5000 anti-DIG antibody (Roche) for 10 hr at 4°C. Embryos were then washed eight times for 15 min in MABTw followed by three 5 min washes in AP buffer. The *in situ* signal was then developed with a 1:1 mixture of AP buffer and BM Purple (Roche) in the temperature range 4 - 37°C depending on the desired time for developing. When an appropriate level of signal was obtained signal development was stopped by addition of 4 % PFA for at least 20 min followed by three 5 min washes in PBSTw.

Embryos at stages later than 26 hpf will have developed pigmented cells which can obscure visualisation of the in situ signal. Where appropriate, embryos were placed in bleaching solution on a light box for 10 – 35 min as required to remove the pigment. Bleached or unbleached embryos were transferred into 50 % glycerol (in PBSTw) for storage and 90 % glycerol for imaging.

2.6.3 Imaging WISH zebrafish embryos

Embryo phenotypes were scored and imaged using a Nikon DXM1200 camera attached to a Nikon SMZ-U 1500 dissecting microscope. Exposure and white balance was set automatically using Q-Capture Pro 7 software and embryos were illuminated from above on a white depression tile.

2.7 **Bacterial methods**

2.7.1 Common media

Table 2.1 *Common media for bacterial work*

Media	Composition
Luria-Bertani (LB) media	Dissolved in distilled H2O: 10 g/L Tryptone 5 g/L Yeast extract 10 g/L NaCl Adjusted to pH 7.0 and autoclaved
2XTY media	Dissolved in distilled H2O: 16 g/L Tryptone 10 g/L Yeast extract 5 g/L NaCl Adjusted to pH 7.0 and autoclaved

SOC media	Dissolved in distilled H2O: 20 g/L Tryptone 5 g/L Yeast extract 0.59 g/L NaCl 0.186 g/L KCl 0.203 g/L MgCl2 1.2 g/L MgSO4 3.6 % Glucose Autoclave before addition of Glucose

The media kitchen in the Department of Biochemistry, University of Oxford prepared the majority of media used in this thesis.

2.7.2 Antibiotics

Table 2.2 *Antibiotic working concentrations*

Antibiotic	Stock concentration	Dissolved in	Working concentration
Ampicillin (Amp)	100 mg/ml	Distilled H_2O	100 µg/ml
Chloramphenicol (Cm)	34 mg/ml	Ethanol	34 µg/ml
Kanamycin (Kan)	50 mg/ml	Distilled H_2O	50 µg/ml
Tetracycline (Tet)	10 mg/ml	Ethanol	10 µg/ml

Ampicillin and Kanamycin were filter sterilised through a 0.2 µm filter. All antibiotics were stored at -20°C.

2.7.3 Culturing *E. coli* on plates

Prior to use, LB agar plates were warmed to 37°C. For culturing bacteria on plates, a bacterial culture in LB was spread onto agar plates using a sterile disposable bent rod, or a sterile disposable plastic loop was used to streak a single colony or a frozen glycerol stock onto an agar plate.

2.7.4 Culturing *E. coli* in liquid culture

For small cultures, 5 mL LB plus appropriate antibiotics was added to a 14 mL round-bottomed tube. A single colony was picked from an LB agar plate using a 200 µL pipette tip and added to the media. For larger cultures, small overnight cultures were grown to stationary phase and added to a larger volume of LB plus appropriate antibiotics at a 1:500 – 1:1000 dilution. Cultures were grown shaking at 200 rpm typically at 37°C.

2.7.5 Making chemically competent *E. coli*

Table 2.3 *Media for chemically competent cells*

Media	Composition
SOB media	Distilled H_2O 2 % w/v Bacto Tryptone 0.5 % w/v Bacto Yeast Extract 10 mM NaCl 2.5mM KCl Add 1/100 volume of 2M Mg solution before use. Store at 4°C.
2M Mg solution	1 M $MgSO4.7H_2O$ 1 M $MgCl.6H_2O$ Autoclave. Store at 4°C.
Transformation buffer	Distilled H_2O 10 mM PIPES 15 mM $CaCl.2H_2O$ 250 mM KCl 55 mM $MnCl_2.4H_2O$ Adjust pH 6.8. 0.22 µm filter. Store at 4°C.

XL-10 gold *E. coli* was streaked from a previous aliquot of competent cells onto an LB agar plate without antibiotic selection and incubated at 37°C overnight. A single colony was inoculated in 2 mL SOB media (or 1/10 volume of final culture volume) and grown with shaking at 200 rpm for 12 hr at 37°C. The culture was then added to 200 mL SOB media and incubated shaking at 200 rpm for 36 hr at 18°C or until the OD600 reached 0.4-0.8. The bacterial culture was then incubated on ice for 10 min and collected by centrifugation at 3,000 rpm for 15 min at 4°C. The pellet was resuspended in 67 mL of transformation buffer (or 1/3 volumes of the final culture volume) by gentle pipetting on ice. The bacterial suspension was incubated on ice for 10 min and collected by centrifugation at 3,000 rpm for 15 min at 4°C. The bacterial pellet was resuspended in 16 mL of transformation buffer (or 1/12.5 volumes of the final culture volume) by gentle pipetting on ice. DMSO was added to a final concentration of 7 %. The bacterial suspension was incubated on ice for 10 min before being snap-frozen in liquid nitrogen in 100 µL aliquots and stored at -80°C.

2.7.6 Transformation

Competent cells were defrosted on ice and 1-4 µL DNA (typically 1 – 100ng) was added with swirling. The transformation mix was incubated on ice for 20 min followed by heat-shock in a water bath at

42°C for 1 min 30 s followed immediately by 2 min on ice. 1 mL LB was added to the transformation mix followed by 1 hr recovery at 37°C with shaking. 100 µL of the transformation mix was spread onto an LB agar plate with appropriate selection. If the transformed DNA was the result of a cloning step (i.e. a ligation or LIC reaction), the remaining bacteria were pelleted by centrifugation (10,000 rpm for 30 s), resuspended in a smaller volume of LB and spread on a second LB agar plate with appropriate selection.

2.7.7 Isolation of plasmid DNA

2.7.7.1 **Small-scale plasmid preparation**

For a small-scale preparation of DNA, a single *E. coli* colony was grown overnight in 5 mL LB media containing antibiotic at 37°C, shaking at 200 rpm. The cells were collected by centrifugation at 4,200 rpm for 10 min at room temperature. DNA was extracted from the bacterial pellet by alkaline lysis using the QIAGEN Spin Miniprep Kit according to the manufacturer's instructions. DNA was eluted using distilled H_2O and stored at -20°C.

2.7.7.2 **Large-scale plasmid preparation**

For large-scale preparation of DNA, a single *E. coli* colony was grown overnight in 150 mL LB media with antibiotic selection at 37°C with shaking at 200 rpm. The cells were collected by centrifugation at 4,200 rpm, for 10 min at room temperature. DNA was extracted from the bacterial pellet by alkaline lysis using the GenElute HP Maxiprep kit (Sigma-Aldrich) according to the manufacturer's instructions. The eluted DNA was concentrated following the manufacturer's instructions, diluted to 1 mg/mL and stored at -20°C.

2.8 **Manipulation of DNA and cloning**

2.8.1 Polymerase chain reaction

2.8.1.1 **Oligonucleotides**

Custom DNA oligonucleotides for use as primers were obtained from Invitrogen, 100 µM stocks were

diluted in distilled H_2O and stored at -20°C. Working dilutions of oligonucleotides for PCR were diluted to 10 µM and to 3.2 µM for Sanger sequencing.

2.8.1.2 **Analytical PCR**

For TA cloning, colony PCR or analytical PCRs, recombinant Taq DNA polymerase (purified by Dr Jin Zhou) was used in a 25 µL reaction volume. 200 ng of purified recombinant Taq DNA polymerase, 12.5 pmol of the forward and reverse primer, 5 nmol of dNTP mix and 10 ng of plasmid DNA were incubated with 1X PCR buffer (10X Thermo Pol Reaction Buffer: 20 mM Tris HCl pH 8.8, 10 mM $(NH_4)_2SO_4$, 10 mM KCl, 2 mM $MgSO_4$, 0.1% Triton X 100).

Table 2.4 *Set-up of a typical analytical PCR reaction*

Component	Volume (1x)
Taq DNA polymerase (0.2 µg/µl)	0.25 µL
10 x Taq buffer	2.5 µL
dNTP (2mM)	2.5 µL
Forward primer (10 µM)	1.25 µL
Reverse primer (10 µM)	1.25 µL
milliQ water	16.25
Template (10 ng/µL)	1 µL

Table 2.5 *Thermal cycling of a typical analytical PCR*

Cycling step		Temperature	Time
Initial denaturation		95°C	2 min
35 cycles	Denaturation	95°C	30 s
	Annealing	45-72°C	30 s
	Extension	72°C	45 s – 1 min per kb
Final extension		72°C	10 min
Hold		10°C	∞

Thermal cycling was performed on a BioRad DNA Engine. Primer annealing temperature varied based on primer properties and extension time varied based on amplicon length and complexity.

2.8.1.3 **Colony PCR**

Colony PCR was performed as in 2.8.1.2 with a bacterial colony dipped into the PCR reaction mix instead of addition of plasmid DNA. Bacterial colonies were spotted onto a fresh LB-agar plate with

antibiotic selection prior to addition to the PCR reaction. One PCR reaction was left without a colony as a negative control.

2.8.1.4 **High fidelity PCR using Phusion polymerase**

Phusion High-Fidelity DNA Polymerase (New England Biolabs) was used for PCR reactions which required high fidelity amplification, for example during cloning or amplification of recombination cassettes. All reactions were performed in a 50 µL volume. A reaction without DNA was used as a negative control. For long primers, 1.5 µL of DMSO was added if the PCR yield was low.

Table 2.6 *Set-up of a typical high-fidelity PCR reaction*

Component	Volume (1x)
Phusion DNA polymerase (2 U/µL)	0.5 µL
5 x Phusion high fidelity buffer	10 µL
dNTP (2mM)	5 µL
Forward primer (10 µM)	2.5 µL
Reverse primer (10 µM)	2.5 µL
milliQ water	28.5 µL
Template (10 ng/µL)	1 µL

Thermal cycling was performed on a BioRad DNA Engine. Primer annealing temperature varied based on primer properties and extension time varied based on amplicon length and complexity.

Table 2.7 *Thermal cycling of a typical high-fidelity PCR reaction*

Cycling step		Temperature	Time
Initial denaturation		98°C	30 s
35 cycles	Denaturation	98°C	10 s
	Annealing	45-72°C	30 s
	Extension	72°C	15-30 s per kb
Final extension		72°C	10 min
Hold		10°C	∞

2.8.1.5 **Quantitative PCR (qPCR)**

Quantitative PCR reactions were performed using SYBR Green (2x Sensimix SYBR Hi-ROX, BIOLINE), 3.75 pmol of forward and reverse primers and 5 µL DNA sample in a final reaction volume of 15 µL. All qPCR reactions were performed in duplicate including a no template control.

Table 2.8 *Set-up of a quantitative PCR reaction*

Component	Volume (1x)
SYBR Green	7.5 µL
Forward primer (10 µM)	0.375 µL
Reverse primer (10 µM)	0.375 µL
milliQ water	1.75
DNA sample	5 µL

Table 2.9 *Thermal cycling of a quantitative PCR reaction*

Cycling step		Temperature	Time
Initial denaturation		95°C	10 min
40-50 cycles	Denaturation	95°C	10 s
	Annealing	60°C	15 s
	Extension	72°C	15 s
Melt		Ramp 65-95°C 1°C rise each step	90 s on first step 5s on each next step

Thermal cycling was performed on a Rotor-Gene 6000 (Corbett) for 40-50 cycles followed by melt curve analysis. Primers were designed to amplify an 80-120 bp region of the locus of interest with a Tm of approximately 60°C.

2.8.1.6 **Quantitative reverse transcription PCR (qRT-PCR)**

For analysis of RNA samples, total RNA was extracted (see 2.5.1), reverse transcribed to cDNA (see 2.5.1) and analysed by quantitative PCR. 5 µL diluted cDNA was used per quantitative PCR reaction in duplicate including a no template control.

2.8.1.7 **Analysis of qPCR data**

Quantitative PCR data was analysed using the comparative CT method (or 2-ΔΔCT method) which assumes that the primers used have a 100 % linear amplification (Schmittgen and Livak, 2008). Primer sets were initially analysed using a 10-fold dilution series of DNA or cDNA to assess the efficiency of the PCR (which should be close to 1). All reactions were performed in duplicate to control for pipetting errors and a diluted input control was included for normalisation. C_T values were measured for each sample (including input controls) using the Rotor-Gene 6000 Series Software on the Slope Correct setting. The C_T threshold was defined as an arbitrary position within

the exponential phase of amplification. The C_T value was determined as the cycle at which the amplification curve crossed the C_T threshold. The C_T value is therefore inversely related to the quantity of amplicon in the sample due to the lower number of cycles required to reach the C_T threshold. ΔCt values were then calculated for each sample using the equation shown in (1). Input ΔCt values were used to calculate the 'percentage of input' value for each of the samples. This allows for comparison between different qPCR experiments (2). An average was taken from the two technical duplicate reactions. For replicate qPCR experiments, the average and standard error of the mean (SEM) were calculated and plotted using error bars.

(1) $\Delta Ct = 2^{(-C_T)}$

(2) Normalisation factor = $\frac{(100 \times (\text{Dilution factor input}))}{((\text{Dilution factor sample}) \times (\text{Average input } \Delta Ct))}$

2.8.2 Adding 'A-tail' to blunt PCR products

PCR amplification with a proofreading polymerase creates a blunt PCR product. For TA cloning, A-overhangs must be added prior to cloning. The A-overhang reaction (Table 2.10) was assembled and left to proceed for 1 hr at 72°C. The reaction was then purified using the QIAquick PCR Purification Kit (QIAGEN) and used directly for TA cloning.

Table 2.10 *Addition of A-overhangs to a PCR product*

Component	Volume
PCR product (from PCR purification or gel extraction)	50 µL
ThermoPol Taq buffer (10x)	6 µL
dATP (10 µM)	2 µL
milliQ water	1 µL
Taq DNA polymerase (0.2 µg/µl)	1 µL

2.8.3 Restriction endonuclease digestion

Restriction digests were performed in a 30 µL volume for conventional analytical purposes or 50 µL for digestion of a cloning vector. The digestion buffer and addition of BSA (100 ng/µL) were chosen in line with the manufacturer's recommendation (New England Biolabs). Typically, digests were left for at least 2 hr at the suggested digestion temperature. Analytical digestions were resolved by agarose gel electrophoresis on a 0.8 – 1.5 % agarose gel (dependent on the size of the expected

fragment) in 1X TAE and visualised with ethidium bromide. The size of DNA fragments was determined by comparison to 1 Kb Plus DNA Ladder (Invitrogen). Digested vector DNA for cloning was resolved by gel electrophoresis and the appropriate band was cut from the gel. DNA was purified from the gel using the QIAquick Gel Extraction kit (QIAGEN) and eluted in 30 µL distilled H_2O.

Amplification of DNA flanked by restriction enzymes was performed using oligonucleotide primers with 5' overhangs containing a GC clamp and the restriction site upstream of the region complementary to the template. A fraction of the PCR product was analysed by gel electrophoresis to check that the desired product was present. The remaining PCR volume was purified if necessary using the QIAquick PCR Purification Kit (QIAGEN) and digested using appropriate restriction enzymes. The digested PCR product was then resolved by gel electrophoresis, cut from the gel and purified using the QIAquick Gel Extraction Kit (QIAGEN).

2.8.4 Dephosphorylation of DNA fragments

The 5' phosphate of vector DNA was enzymatically removed to prevent self-ligation during ligation reactions. 1-5 µg of digested DNA was supplemented with 1X Antarctic Phosphatase Buffer (New England BioLabs) and 1 µL of Antarctic Phosphatase. The reaction mixture was mixed and incubated for 15 min at 37°C for 5´ extensions or blunt-ends and 60 min for 3´ extensions. The enzyme was heat-inactivated for 5 min at 65°C. Subsequent ligations were performed without further purification of the DNA.

2.8.5 DNA ligation

Ligation reactions were performed using multiple vector-to-insert ratios typically 3:1, 1:1 and 1:3 by molar ratio. A vector-only control was used to monitor self-ligation. Ligations were performed using the Rapid DNA Ligation Kit (Roche). The vector and insert DNA were diluted in 1X DNA Dilution Buffer to a final volume of 10 µL. 10 µL of 1X DNA Ligation Buffer and 1 µL T4 DNA Ligase were

added to the reaction and mixed well. The reaction was allowed to proceed at room temperature for typically 30 min. 2 µL of the ligation mixture was transformed into competent cells (see 2.7.5).

2.8.6 Ligation independent cloning (LIC)

5 µg of a LIC-adapted vector containing the *sacB* gene, which allows negative selection on 5% sucrose, (i.e. pNIC28 or pCAGIpuro_LIC) was digested with BsaI in a 50 µl reaction volume for at least 1 hr. The vector was resolved by agarose gel electrophoresis in 1X TAE, visualised by ethidium bromide staining and excised from the gel. Vector DNA was purified using the QIAquick Gel Extraction kit (QIAGEN) and eluted in 30 µL distilled H_2O. The DNA to be inserted via LIC-cloning was amplified by PCR using primers containing LIC cloning hybridisation sequences at their 5' end. The PCR product was purified using the QIAquick PCR Purification Kit (QIAGEN) and eluted in 30 µL milliQ H_2O.

The digested vector and PCR-amplified insert were incubated separately with 10 units of T4 DNA polymerase (Fermentas) in 1X Reaction Buffer (supplied with the enzyme) in the presence of 2 µL of 2 mM dGTP and dCTP respectively (total reaction 50 µL). The reaction mix was incubated for 30 min at 22°C followed by heat inactivation at 75°C for 20 min. DNA was purified using the QIAquick PCR Purification Kit (QIAGEN) and eluted in 30 µL milliQ H_2O. The purified vector and insert DNA were mixed together in three different vector-to-insert molar ratios, 5:1, 1:1 and 1:2. The vector and insert were incubated for 30 min at room temperature to allow hybridisation to occur. 1 µL of each LIC hybridisation reaction was transformed into competent XL-10 Gold *E. coli*, plated onto LB plates with appropriate antibiotic selection and grown at 37°C overnight.

2.8.7 Screening colonies for successful cloning

Bacterial colonies were screened for successful conventional or ligation-independent cloning. A number of bacterial colonies were grown overnight in a 5 mL culture and DNA extracted by alkaline lysis using the QIAprep Spin Miniprep kit (QIAGEN) and eluted in 50 µL distilled H_2O. DNA from each clone was analysed by restriction endonuclease digestion or PCR to verify the presence of the insert.

2.8.8 TA cloning

Regions of gDNA or cDNA for which sequence information was required were amplified by PCR using Taq polymerase (2.8.1.2) or high fidelity Phusion Taq polymerase (2.8.1.4) followed by A-overhang addition (2.8.2) and TA cloned into the pGEM-T Easy Vector System (Promega). Bisulfite converted genomic DNA was amplified by PCR (using specially designed bisulfite primers) and cloned into the pGEM-T Easy vector. Ligations were set up with 1X Rapid Ligation Buffer, 1 µL pGEM-T Easy vector, 1 µL T4 DNA Ligase and 3 µL of the PCR product. The ligation was mixed well and incubated for at least 1 hr at room temperature. 2 µL of the ligation was transformed into competent XL-10 Gold *E. Coli*. The transformed bacteria were plated onto LB plates with Ampicillin, IPTG and X-gal and grown overnight at 37°C. White colonies were picked, cultured overnight and plasmid DNA purified using the QIAprep Spin Miniprep kit (QIAGEN). Clones were screened by digestion of 5 µL plasmid DNA with the EcoRI restriction enzyme. Positive clones were sequenced using T7 or SP6 primers.

2.8.9 Sanger sequencing

Sanger sequencing was performed by Source BioScience yielding, on average, 500-1000 bp of sequence. The quality of the sequencing dropped off more quickly for bisulfite sequencing reactions.

2.8.10 Preparation of DNA constructs

Preparation of DNA constructs used in this thesis is described in Appendix 2.1.

2.9 **Bacterial artificial chromosome (BAC) recombineering**

BAC recombineering was performed following the Gene Bridges protocol (http://www.genebridges.com/gb/pdf/K001_Q+E_BAC_Modification_Kit.pdf) and a protocol recommended by Stefan Schulte-Merker (http://www.hubrecht.eu/research/schulte-merker/documents/BACrecombineering_stepbystep.pdf).

The pRedET plasmid contains a tetracycline resistance gene along with the λ phage redγβα operon under the control of an arabinose-inducible pBAD promoter (Guzman et al., 1995). In bacteria containing this plasmid, addition of L-arabinose to the culture media, and incubation at 37°C, leads

to induction of the redα, β and γ genes. Recombination is induced between regions of homology (provided by the recombination cassette). The Redαβ pair mediates double strand break repair with the 5' - 3' exonuclease Redα digesting one strand of DNA from the double strand break leaving a 3' overhang which is then coated by the Redβ. The single stranded overhang then invades the recombination cassette homology arms to form a primer for DNA synthesis. Recombination is assisted by expression of the Redγ Gam protein, which inhibits RecBCD exonuclease activity in *E. coli*. The pRedET plasmid is dependent on oriR101, which requires the temperature sensitive RepA protein for replication. Therefore, bacteria are incubated at 30°C to maintain the pRedET plasmid, except during induction of pRedET which the temperature is raised to 37°C to ensure maximal expression and activity of the recombination proteins.

2.9.1 Ordering BACs

Two zebrafish bacterial artificial chromosomes (BACs; DKEY-194K4 and DKEY-5K13) and two mouse BACs (RP23-189D14 and RP23-477B14) were ordered from Source BioScience. All BACs confer chloramphenicol resistance. The BACs were delivered in the DH10B *E. coli* strain and were streaked out onto Cm plates (12.5 μg/mL).

2.9.2 Glycerol stocks

Glycerol stocks were made from the parent BACs and following each recombineering step by adding 500 μL of overnight culture to 500 μL 1:1 glycerol:LB. Glycerol stocks were stored at -80°C.

2.9.3 Designing recombination cassette

Primers were designed to amplify the recombination cassette with normal 18-24 bp primers with additional 50bp of homology at the 5' end of the primers to facilitate recombination.

For amplifying recombination cassettes for zebrafish BACs, primers from Stefan Schulte-Merker were used to amplify inverted Tol2 (iTol2) surrounding the ampicillin resistance gene with homology arms to the pIndigoBAC-536 backbone. For amplification of the Kan/Neo cassette (to confer antibiotic resistance to mouse embryonic stem cells during production of stable cell lines), a second

pair of homology arms were designed to another region in the pIndigoBAC-536 backbone. All cassettes were amplified using Phusion High-Fidelity DNA Polymerase (see 2.8.1.1). (New England Biolabs, M0530S).

Table 2.11 *Set-up for high fidelity PCR for amplifying zebrafish BAC recombineering cassettes*

Component	Volume
Phusion polymerase	0.5 µL
5 x Phusion high fidelity buffer	10 µL
dNTP	5 µL
Forward primer (10 mM)	2.5 µL
Reverse primer (10 mM)	2.5 µL
milliQ water	28.5
Template (10 ng/µL)	1 µL

2.9.4 BAC antibiotics and L-arabinose

Table 2.12 *Antibiotic and L-arabinose working concentrations for BAC recombineering*

Antibiotic	Stock concentration	Dissolved in	Working concentration
Chloramphenicol (Cm)	34 mg/ml	Ethanol	12.5 µg/ml
Tetracycline (Tet) *	10 mg/ml	Ethanol	3 µg/ml
Ampicillin (Amp)	100 mg/ml	Water	30 µg/ml
Kanamycin (Kan)	50 mg/ml	Water	15 µg/ml
L-arabinose	10 % (w/v)	Water	0.33 % (w/v)

* Tetracycline is light sensitive.

2.9.5 Electroporation of pRedET plasmid

2 mL LB containing 12.5 µg/ml Cm was inoculated with a BAC colony from the plate streaked in 2.9.1 and incubated overnight at 37°C. 30 µL of the overnight culture was added to 1.4 mL fresh LB with Cm and cultured for a further 2 – 3 hr at 37°C. The bacterial culture was centrifuged for 30 s at 11,000 rpm at 4°C and the bacterial pellet was resuspended in 1 mL of ice-cold milliQ water by gentle pipetting. This wash step was repeated twice. The supernatant was tipped out of the tube leaving around 30 µL water to which 1 µL of pRedET plasmid (around 20 ng/µL) was added. Bacteria were kept on ice throughout. 30 µL of the bacteria plus DNA was added to a chilled 1 mm

electroporation cuvette (BioRad, MicroPulser Cuvettes, 1652089) and electroporated at 1800 V, 25 µF and 200 Ohms using a GenePulser Xcell (Bio-Rad).

Following electroporation, the bacteria were resuspended in 1 mL SOC media and incubated without selection for 1 hr at 30°C. The bacteria were then spread onto LB agar plates containing Cm (12.5 µg/ml) and Tet (3 µg/ml) and incubated at 30°C overnight to retain the pRedET plasmid.

2.9.6 Electroporation of recombination cassette

A 2 mL overnight culture with chloramphenicol (12.5 µg/ml) and tetracycline (3 µg/ml) plus any other appropriate antibiotic selection was inoculated with a bacterial colony containing the BAC of interest and pRedET and incubated shaking overnight at 30°C (covered in foil as tetracycline is light-sensitive). A fresh 30 mL culture containing appropriate antibiotics was inoculated with 643 µL of the overnight culture and incubated for a further 2.5 – 3 hr at 30°C until the OD_{600} reached 0.3. Two 4 mL aliquots of this culture were prepared in 14 mL tube, one of which was inoculated with 143 µL of 10 % L-arabinose. Both tubes were incubated at 37°C for 45 min – 1 hr. 2 mL of the culture was centrifuged for 30 s at 11,000 rpm at 4°C and made electrocompetent by three washes with 1 mL ice-cold water (as described in 2.9.5). 1-4 µL of the recombination cassette (around 100 ng) was added to the bacteria in 30 µL ice-cold milliQ water. The bacteria plus DNA was added to a chilled 1 mm electroporation cuvette (BioRad, MicroPulser Cuvettes, 1652089) and electroporated at 1800 V, 25 µF and 200 Ohms using a GenePulser Xcell (Bio-Rad).

Following electroporation, the bacteria were resuspended in 1 mL SOC media and incubated without selection for 1 hr at 37°C. The bacteria were then spread onto LB agar plates containing Cm (12.5 µg/ml) and Tet (3 µg/ml) and other appropriate antibiotic selection. The plates were incubated at 30°C to retain the pRedET plasmid.

Following the final recombination step, plates were incubated at 37°C without Tet selection to induce loss of the pRedET plasmid.

2.9.7 Screening for colonies

A successful BAC recombineering should yield colonies on the induced (+ L-arabinose) plates but not on the uninduced (– L-arabinose) plates. Colonies obtained on both plates were screened by colony PCR with primer sets designed across both the left and right homology arms. Positive colonies were further screened by restriction digestion and compared to the parent BAC to verify that the BAC remained intact during recombineering (2.9.8).

2.9.8 Small-scale BAC preparation

Purification of BAC DNA was performed to determine whether the pRedET plasmid had been maintained in the bacteria following electroporation of pRedET or to check BAC integrity following recombination can be checked by restriction enzyme digestion. To check integrity of BACs, they were digested with a restriction enzyme that gave a distinct digestion pattern when separated on a 0.8 % agarose gel. This was done to ensure the BAC had undergone appropriate recombination and not aberrant rearrangement.

For small-scale alkaline lysis BAC preparation, 5 mL LB containing the appropriate antibiotic selection was inoculated with a single BAC colony and incubated overnight at either 37°C or 30°C as appropriate. Growth at 30°C with tetracycline would allow detection of the pRedET plasmid (9270 bp, runs between 5-6kb on a 1% agarose gel), while growth at 37°C without tetracycline will cause loss of pRedET and allow restriction digest analysis. The bacterial culture was pelleted by centrifugation at 4500rpm for 10 min, resuspended in 250 µL of buffer P1 (2.19.2) containing RNaseA (0.1 mg/ml) and transferred to a 1.5 mL tube. 250 µL buffer P2 lysis buffer (2.19.2) was then added and mixed thoroughly by inversion followed by incubation at room temperature for less than 5 min to allow lysis to occur fully. 250 µL of buffer P3 (2.19.2) was then added, the reaction was mixed by inversion, and incubated on ice for 5 min. The precipitate was pelleted by centrifugation at 13,000 rpm for 15 min at 4°C and the supernatant transferred to a fresh tube and pelleted by centrifugation once more. The supernatant was then precipitated by adding 750 µL of isopropanol,

mixing and incubation on ice for 10 min. The DNA was pelleted at 13,000 rpm for 10 min at 4°C, washed with 1 mL 70 % ethanol and pelleted again. The DNA was air dried for 10 min at room temperature and resuspended in 40 µL milliQ water. 20-40 µL of the BAC preparation was used for analytical restriction enzyme digestion.

2.9.9 Large-scale BAC preparation

Large-scale BAC preparations were performed using the Nucleobond BAC100 kit (740579). 500 µL from a 5 mL overnight culture (grown at 37°C in appropriate antibiotic selection from a single BAC colony) was inoculated into a 300 mL overnight culture and grown overnight at 37°C in the appropriate antibiotic selection. The bacterial culture was pelleted at 4,200 rpm for 15 min and the BAC DNA extracted according to the manufacturer's instructions.

2.9.10 Generation of stable cell lines

Stable E14 cell lines (see 2.2.2 for E14 cell culture conditions) were generated by integration of a zebrafish BAC containing the Kan/Neo selection cassette into the mouse genome. 1 µg of BAC DNA was transfected along with either 5 µg transposase plasmid DNA (plasmid 446) or 5 µg of pBSII-LIC (392, as a control for Tol2-mediated integration) into E14 cells using Lipofectamine 2000 (Invitrogen, Carlsbad, CA), and stable integrants were selected using 400 µg/mL G418 for 10 days. Colonies were picked into wells of a 96-well plate and screened by PCR for stable integration of the BAC.

2.10 **Protein expression and purification**

2.10.1 Protein expression and purification of recombinant His-tagged proteins

The zKdm2b antigen, zKdm2 ZF-CxxC-PHD constructs and human hKDM2B ZF-CxxC-Avi construct were expressed by freshly transforming a BL21 expression strain carrying the pRARE2 plasmid with the appropriate expression vector. Cultures were grown in 2XTY and grown to an OD_{600} of 0.6 at 37°C. When expressing proteins containing a ZF-CxxC domain, the media was supplemented with 0.25mM $ZnCl_2$. The culture was then cooled to 30°C and expression was induced with 1 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) for 3 hr or 18°C overnight. The cultures were then

pelleted and the bacteria lysed in 20mM Tris-HCl (pH 8.0), 500mM NaCl and 0.1% Nonidet P-40 with cOmplete EDTA-free protease inhibitors (Roche) by sonicating at 60 % amplitude 30s on and 30s off for 6 cycles using a Vibra-Cell (SONICS) sonicator. The lysate was centrifuged at 4°C for 20 min at 12,500 rpm, and Ni-NTA beads (Ni-charged IMAC Sepharose 6 Fast Flow, GE Healthcare) pre-equilibrated with lysis buffer were added to the supernatant. The recombinant 6-His protein was allowed to bind the beads by rotating for 1 h at 4°C, the lysate and beads were then applied to a 10 or 25mL disposable column. The column was washed with 20 column volumes of wash buffer (50mM NaH_2PO_4, 300mM NaCl, 20mM imidazole, pH8.0) and batch-eluted with elution buffer (50mM NaH_2PO_4, 300mM NaCl, 250mM imidazole, pH8.0). Recombinant protein-containing fractions were identified by SDS-PAGE (sodium dodecyl sulphate polyacrylamide gel electrophoresis) and Coomassie Blue staining. The peak fractions of the zKdm2 antigens were pooled and dialyzed into BC150 (50mM Hepes pH 7.9, 150mM KCl, 10% Glycerol, fresh 0.5mM DTT) (Klose and Bird, 2004). The peak elution fractions from the ZF-CxxC-PHD and Bio-CAP CxxC-Avi proteins were pooled and digested overnight at 4°C with His-tagged tobacco etch virus (TEV) protease. The protein was then desalted to remove the imidazole and reapplied to a Ni-NTA column (as described above) to remove the cleaved tag and TEV protease. The cleaved CxxC-Avi protein was then desalted into 10 mM Tris pH 8.0 containing 250 mM potassium glutamate in preparation for *in vitro* biotinylation. While the zKdm2 ZF-CxxC-PHD constructs were dialysed into BC100 buffer (50mM Hepes pH 7.9, 100mM KCl, 10% Glycerol, fresh 0.5mM DTT) overnight.

2.10.2 Protein quantification by Bradford reagent

Bradford reagent was used to quantify protein concentration. Bradford reagent was diluted 1 in 4 with milliQ water and 1 mL aliquots were made. A standard curve was made by diluting 0, 1, 3, 6 and 10 µL BSA (1 mg/mL) in the 1 mL aliquots. 1 µL of each sample was also diluted into the 1 mL aliquots. All aliquots were transferred into 1 mL cuvettes and absorbance measured at 595 nm. Protein concentration of the samples was calculated from the standard curve.

2.10.3 *In vitro* biotinylation

Following protein purification and cleavage of the His-tag, the CxxC-Avi construct was *in vitro* biotinylated by addition of His-tagged recombinant BirA to the protein. The reaction was supplemented with 10 mM ATP, 10 mM MgOAc, and 50 μM D-biotin and left to proceed overnight. The following day the reaction was supplemented with 20 mM imidazole and applied to Ni-NTA resin to remove the His-tagged BirA ligase, which bound to the resin. The CxxC-Avi protein, from the flow-through, was then dialysed into BC150 buffer and stored as aliquots at -80°C. The efficiency of biotinylation was verified by mass spectrometry prior to and after *in vitro* biotinylation. *In vitro* biotinylation of the CxxC-Avi protein was performed by Dr Rob Klose. Mass spectrometry to verify protein biotinylation was performed in-house by Dr David Staunton (Department of Biochemistry, University of Oxford).

2.11 **Protein electrophoresis**

2.11.1 Deyolking zebrafish embryos for western blot analysis

Zebrafish embryos were prepared for western blot analysis as described by Link et al (Link et al., 2006). Embryos were manually dechorionated with forceps using a dissecting microscope and then transferred to a 1.5 mL tube with 1 mL deyolking buffer. Embryos were dechorionated on an agarose-coated petri dish if the embryos were younger than bud stage. The embryos were manually dissociated by pipetting with a 200 μL tip followed by shaking for 5 min at 1100 rpm at room temperature (Eppendorf, Thermomixer comfort) to dissolve the yolk. The cell suspension was then pelleted at 300g for 30 s and the supernatant was discarded without disturbing the cell pellet. Samples were then put immediately on dry ice and stored at -20°C or -80°C.

2.11.2 Sample preparation

De-yolked zebrafish cell pellets were boiled at 100°C for 5-15 min (depending on the embryonic stage) with 1X sample buffer prior to loading. Nuclear extract from 293T cells was boiled at 100°C for 5 min with 1X sample buffer prior to loading.

2.11.3 Polyacrylamide gel electrophoresis

Polyacrylamide gels were made fresh using a Bio-Rad Mini-PROTEAN Tetra Cell casting frame and stand. Typically 0.75 mm gels were cast between glass plates with 0.75 mm 10-well combs. A separating gel was poured first and covered with water until it had set. A stacking gel was then poured above the separating gel into which the comb was placed.

Once set, the gels were removed from the casting equipment, placed into the running module and the gel running tank was filled to the appropriate level with 1X SDS PAGE running buffer. The comb was removed carefully and the wells flushed with SDS PAGE running buffer. Sampled were loaded into the wells using long pipette tips (Prot/Elec tips, BioRad). Gels were typically run at 200 V for 45 min – 1 hr or until the loading dye migrated out of the bottom of the gel. Gels were then removed from between the glass plates and the stacking gel discarded.

Table 2.13 *Composition of 6-18 % SDS-PAGE gels*

Media	Separating gel	Stacking gel
Acrylamide	6-18 %	5 %
Tris-HCl	0.4 M (pH 8.8)	125 µM (pH 6.8)
SDS	0.1 %	0.1 %
APS	0.1 %	0.1 %
Temed	0.08 % in 6 % acrylamide 0.06 % in 8-10 % acrylamide 0.04 % in 12-18 % acrylamide	0.1 %

2.11.4 Coomassie blue stain

For visualisation of proteins, SDS PAGE gels were incubated with Coomassie blue stain with rocking for 10 – 15 min. The gel was then rinsed with water and incubated with de-stain solution for around 1 hr or until the background staining had disappeared. Gels were left in water over-night and then imaged on a flat-bed scanner.

2.11.5 Western blotting

For western blotting, proteins were transferred from the poly-acrylamide gel onto an appropriately sized nitrocellulose membrane (typically 0.45 µm or 0.2 µm for histones, Bio Rad) by semi-dry

transfer. For semi-dry transfer, six sheets of Whatman paper and the nitrocellulose membrane were soaked in semi-dry transfer buffer before use. The SDS poly-acrylamide gel was laid on top of the nitrocellulose membrane above three layers of Whatman paper. A further three layers of Whatman were placed above the gel and any air bubbles were removed using a roller. The proteins were transferred to the membrane from the gel at constant 200 mA and a maximum of 23 V for 1.5 hr (BIORAD, Trans-Blot 5D, Semi-Dry Transfer Cell).

Following transfer, the membrane was blocked with 5 % milk powder in 1XPBS buffer containing 0.1 % Tween20 (PBSTw/milk) for 1 hr at room temperature. The membrane was then washed three times with PBSTw/milk, and incubated overnight at 4°C with primary antibody or serum diluted in PBSTw/milk. Dilution factors for primary antibody varied with the antibody used (typically 1:200 – 1:500). The next day the membrane was washed three times for 10 min with PBSTw followed by incubation for 1 hr with the secondary antibody in PBSTw containing 1 % BSA (Sigma Aldrich), typically at 1:5000 dilution. The membrane was then washed three times for 10 min with PBSTw to remove excess secondary antibody.

ECL A and ECL B solutions were mixed well in equal proportions (while being protected from light) and added to the membrane for 1 min at room temperature. The membrane was immediately exposed to film (Fujifilm, Super RX) in a cassette (Amersham Biosciences, Hypercassette) and developed (Xograph Imaging Systems, Compact 4).

2.12 **Electrophoretic mobility shift assay (EMSA)**

2.12.1 EMSA DNA probe

A 216 bp 601 nucleosome positioning sequence was used as a DNA probe for EMSA containing 18 CpG dinucleotides (generated by Dr Jin Zhou).

GGC**CG**CCCTGGAGAATCC**CG**GTGC**CG**AGGC**CG**CTCAATTGGT**CG**TAGACAGCTCTAGCAC**CG**CTTAAA**CG**CA

CGTA**CGCG**CTGTCCCC**CGCG**TTTTAAC**CG**CCAAGGGGATTACTCCCTAGTCTCCAGGCA**CG**TGTCAGATATATA

CATCCTGTGCATGTATTGAACAG**CG**ACCTTGC**CG**GTGCCAGT**CG**GATAGTGTTC**CG**AGCTCCCACTCTAG

2.12.2 *In vitro* methylation of DNA

For *in vitro* methylation, 1 µg of DNA probe was incubated with 1 µg recombinant M.SssI (generated by Dr Jin Zhou) in 1X New England Biolabs (NEB) Buffer 2 containing 1X S-adenosylmethionine (SAM) and incubated in a 20 µL volume at 37°C overnight. Fresh M.SssI (1 µg) and 1X SAM were added to the reaction the next day and incubated at 37°C for an additional 2 hr to overnight. The methylated DNA was then purified using the QIAquick PCR purification kit (QIAGEN, 28106), eluted in 50µl distilled water and the DNA concentration was quantified using a Nanodrop 1000 spectrophotometer (Thermo Scientific) at 260 nm.

To quantify the extent of methylation, 100 ng of methylated probe was digested with the methylation sensitive restriction enzyme, HhaI. The HhaI enzyme cuts at 5'-GCGC-3' sites, but this activity is inhibited by cytosine methylation at the central CpG of the recognition site. HhaI digested and undigested 601 DNA was then resolved on an agarose gel and analysed by ethidium bromide staining. Efficient methylation caused lack of digestion of the methylated probe.

2.12.3 Radioactive labeling of EMSA probe

Methylated and unmethylated EMSA probes were radioactively end labelled using [γ-^{32}P] ATP (Perkin Elmer). 200 ng DNA probe was incubated with 20 units T4 polynucleotide kinase (New England Biolabs), 1X T4 polynucleotide kinase buffer (supplied with the enzyme) and 2 µl of 10 mCi/mL [γ-^{32}P] ATP in a 30 µL volume for 1 hr at 37°C. The reaction was then purified using the QIAquick PCR Purification kit (QIAGEN) and eluted in 50 µL EB buffer. The radioactivity of 1 µL radiolabelled probe was quantified using a Bioscan QC-2000 β-counter as counts per minute per µL (cpm/µL). The probe volume was then adjusted to give 2,000 cpm/µL. Probes were stored at -20°C and were used for up to two weeks after preparation. All work with radioactive substances was performed using appropriate safety precautions.

2.12.4 Electrophoretic mobility shift assay (EMSA)

Up to 2 µg of recombinant protein was incubated in EMSA binding buffer with 25 ng/µL poly-dAdT competitor DNA for 10 min at room temperature. The radioactive end-labelled 601 probe (with 0.025 % bromophenol and 1X Orange G, to allow visualisation of the running band during electrophoresis) was added to the protein in binding buffer and incubated at room temperature for 20 min. The samples were loaded onto a 1.3% agarose gel (made with 0.2X TBE without EDTA, as EDTA may interfere with DNA binding). The gel was typically run in 0.2X TBE (without EDTA) at 100 V at 4°C or until the dye front had migrated approximately 7 cm from the wells. The gel was subsequently dried onto a sheet of DE81 anion exchanger paper (Whatman) using a vacuum driven gel-dryer (Amersham) at 60°C for approximately 1-2 hr. The gel was exposed to a phosphorimager screen overnight. The screen was then scanned using a fluorescent image analyser (FLA-7000, Fujifilm).

2.13 **DNA methylation analysis**

2.13.1 Bisulfite Sequencing

Bisulfite conversion of DNA was performed using the EZ DNA Methylation-Gold Kit (Zymo Research) according to the manufacturer's instructions. 130 µL of the CT conversion reagent was added to 20 µL of the sample to be converted and incubated using the following conditions: 98°C for 10 min, 64°C for 2.5 hr and 4°C for up to 20 hr. In 600 µL binding buffer, the sample was applied to a Zymo-Spin IC column. Following a wash step, the sample was desulfonated on the column, washed once more and eluted in a 10 µL volume of elution buffer. 1-4 µL of the elution volume was used for PCR-amplification using Maxima Hot Start Taq DNA Polymerase (Fermentas), which adds adenines to the 3'-end of PCR products via its deoxynucleotidyl transferase activity. Thus the PCR products were suitable for cloning into the pGEM-T Easy vector by TA cloning using the manufacturer's instructions (pGEM-T Easy Vector System, Promega). White clones were selected using blue-white selection (see 2.8.8), plasmid DNA isolated (QIAGEN Spin Miniprep Kit) and sequenced using T7 or SP6 primers.

Sequenced clones were analysed using the web-based tool QUMA by comparison to unconverted DNA sequence (http://quma.cdb.riken.jp/) (Kumaki et al., 2008).

2.13.2 Biotinylated CxxC Affinity Purification (Bio-CAP)

2.13.2.1 **Preparation of beads**

For each individual Bio-CAP experiment, 25 µL of packed NeutrAvidin Agarose Resin (Thermo Scientific, 29200) or NeutrAvidin-coated magnetic beads (Thermo Scientific, 7815-2104-011150) was washed with BC100 buffer and then incubated with 50 µL of 0.5 µg/ µL biotinylated hKDM2b-CxxC protein diluted in BC150 buffer (or 50 µL of BC100 buffer alone for a 'beads only' control) for 1 hr at 4°C. The conjugated resin/CxxC protein was then washed three times with CAP1000 buffer and once with CAP100 buffer.

2.13.2.2 **Bio-CAP**

Genomic DNA at a concentration of 0.35 mg/mL was sonicated to an average size of 150-350 bp using a Diagenode Bioruptor, using the high setting for 30 s on, 30 s off for 2 hr with a pulse spin every 10 min. Sonicated DNA was then diluted in CAP100 buffer to 16 µg/mL and 100 µL was retained as an Input sample. For each Bio-CAP experiment, 500 µL of diluted sonicated DNA, typically corresponding to 8 µg of DNA (unless otherwise stated), was added to the conjugated CxxC resin. The DNA and resin were incubated at 4°C with rotation for 1 hr. The resin was then collected by centrifugation at 1500 x g for 3 min at 4°C or by magnetisation, and the unbound flow through (FT) material was removed. The resin and any associated DNA was washed twice with 1 mL of CAP100 buffer, before the first elution was performed by adding 50 µL of CAP300 (12.5% glycerol, 0.1% Triton-x-100, 20 mM HEPES pH 7.9, 300 mM NaCl) to the resin and incubating at room temperature for 10 min. Following centrifugation or magnetisation, a 50 µL elution fraction was carefully collected. The elution process was repeated using another 50 µL of CAP300 and the 300 mM elution fractions were pooled (giving a total volume of 100 µL). Subsequent elutions were performed in the same way using buffers with 500 mM, 700 mM and 1 M NaCl sequentially. Each

100 µL elution fraction, and 100 µL of both the Input and FT samples were purified using a PCR purification column (QIAGEN) and DNA was eluted in 50 µL distilled H_2O. For real-time qPCR analysis, Bio-CAP samples were typically diluted 10-fold and 5 µL of this was used per 15 µL reaction (see 2.8.1.5).

For the Bio-CAP spiking experiment, 200 bp regions containing set numbers of CpGs were amplified from human genomic DNA. The methylated human probe was generated by *in vitro* methylation of the 13 CpG probe with M.SssI (NEB) (see 2.12.2) and the hydroxymethylated human probe was generated by amplifying the 13 CpG probe in a PCR reaction containing 5hmC. Approximately 10 pg of each 200 bp probe was added to 8 µg of sonicated mouse genomic DNA and the Bio-CAP assay was then performed in the same way as above. The 200 bp human probes were interrogated by qPCR using primer sets nested within each region.

2.13.2.3 **Magnetic Bio-CAP**

Mouse genomic DNA sonicated to 150-350 bp was incubated with magnetic CxxC resin for 1 hr in the same way as a regular Bio-CAP experiment (2.13.2.2). The magnetic CxxC resin and associated DNA was then collected using a magnetic microcentrifuge tube rack, allowing unbound FT material to be removed. Two 10 min washes were performed in CAP500 (500mM NaCl) to remove non-specifically bound material, followed by two 10 min elution steps with 50 µL CAP1000 (1M NaCl). Elutions were combined, purified by PCR purification (QIAGEN) and analysed by qPCR.

2.13.3 Methylated DNA immunoprecipitation (MeDIP)

MeDIP was performed in the laboratory of Dr Skirmantas Kriaucionis. Genomic DNA was sonicated using a Diagenode Bioruptor to an average size of 250 bp. Prior to immunoprecipitation DNA blunting, dA overhang addition and adaptor ligation was completed according to Illumina library preparation recommendations. Immunoprecipitation was performed using 2 µg of anti-mC antibody (Eurogentec) using a general MeDIP protocol (http://www.epigenome-noe.net/WWW/researchtools/protocol.php?protid=33) with some minor modifications.

2.14 **Massively parallel sequencing**

2.14.1 Picogreen quantification of DNA

To quantify DNA at low concentrations, Picogreen quantification was used. The concentration of input DNA (i.e. sonicated DNA diluted to around 16 µg/mL in CAP100 buffer for Bio-CAP experiments) was determined using a Nanodrop 1000 spectrophotometer (Thermo Scientific) at 260 nm. The input DNA was then diluted to 1 µg/mL in 1xTE and a dilution series of 200 ng/µL 100 ng/µL 40 ng/µL 8 ng/µL was made. 2 µL of each dilution was mixed with 2 µL of Picogreen (diluted 1:200 in 1xTE), fluorescence was detected using a Nanodrop 3300 and a standard curve was generated using the measurements. DNA samples were diluted 10x in 1xTE and 2 µL of 10x diluted or undiluted sample was then mixed with 2 µL diluted Picogreen. Fluorescence was detected using a Nanodrop 3300 and the concentration of the samples inferred from this curve. The two values from the undiluted and diluted DNA sample were averaged to give an estimate of the DNA concentration for each sample.

2.14.2 Library preparation

Bio-CAP DNA was prepared for Solexa 2G sequencing by blunting the DNA with a mixture of T4 DNA polymerase, Klenow DNA polymerase, and T4 PNK (NEB) according to manufacturer's instruction. dA overhangs were then added and Illumina adapters ligated. Adapter-ligated DNA was subject to 18 cycles of PCR before size selection by agarose gel electrophoresis. Amplified DNA was purified using the QIAquick Gel Extraction Kit (QIAGEN). The purified DNA was quantified with both an Agilent Bioanalyzer and Invitrogen Qubit and diluted to a working concentration of 10 nM prior to sequencing.

2.14.3 Bio-CAP sequencing

All Bio-CAP experiments were performed in duplicate (either technical or biological replicates) with matched input controls. Next generation sequencing was performed using two Illumina sequencing platforms: Genome Analyser IIx and HiSeq 1000 yielding 51bp single-end or paired-end reads.

2.14.4 Read alignment

Raw sequencing reads were aligned to the appropriate genome using the Bowtie short-read aligner (v0.12.7, (Langmead et al., 2009)). If the data was paired-end, only one of the ends was used in the alignment. Only uniquely mapping reads, with a maximum of two mismatches across the entire read length were used. The reference genomes used in this thesis are as follows: human, hg19; mouse, mm9; platypus, ornAna1; chicken, galGal3; lizard, anoCar2; frog, xenTro3; zebrafish danRer7. A reference genome was generated for the Tc1 mouse model (TC1: the entire mouse genome plus human chromosome 21) with the help of Dr. Simon J. McGowan (Computational Biology Research Group, Weatherall Institute of Molecular Medicine).

2.14.5 Peak calling

Peaks were typically identified using MACS (v1.4.0) (Zhang et al., 2008) specifying a bandwidth of 300 and an mfold range of 10-30. Binding intervals were filtered by a q-value of 0.01. Wig files were also generated from MACS for visualisation of the alignment.

2.14.6 Data visualisation

Genomic peaks and intervals were visualised using the Integrated Genome Browser (IGB, http://bioviz.org/igb/) software (Nicol et al., 2009).

2.15 **Analysis of massively parallel sequencing data – seven vertebrates**

Data analysis was performed in R and python using bespoke scripts available online (www.cgat.org).

2.15.1 Overlap of NMIs with CGIs

CpG island annotations were downloaded from UCSC Tables and overlapped with NMI intervals using BEDTools intersectBed. 1 bp overlap was required for successful overlap of these features.

ensGene gene models were downloaded from Ensembl for all seven vertebrate species. The transcription site start (TSS) was defined as the +/- 1kb region surrounding the most 5' transcript for each gene. These TSS intervals were overlapped with NMIs from testis Bio-CAP-seq using BEDTools intersectBed, requiring a 1 bp overlap for intersection of these features.

2.15.2 Genomic localization of non-methylated islands

Ensembl (release 66) genes were annotated as having an NMI at their TSS using a single base pair overlap of an NMI with a window extending 1 kb upstream and downstream from the transcript transcription start site (TSS) of the most 5' annotated transcript for each gene.

Multi-tissue RNAseq datasets for platypus, chicken and lizard were used to improve the annotation of gene TSSs (Barbosa-Morais et al., 2012; Brawand et al., 2011; Julien et al., 2012). Where transcript models were not provided by the authors TopHat (Trapnell et al., 2009) and Cufflinks (Trapnell et al., 2010) were used to construct transcript models from short-read data. Similarly, the XTev gene dataset was used to improve the annotation of TSSs in the frog genome (Akkers et al., 2010). NMIs were annotated with respect to both Ensembl (release 66) and RNAseq-based genome annotation as being associated with the following features in a hierarchical manner: protein-coding gene TSS (+/-1kb), gene body, upstream or downstream of a gene (within 5kb of the annotated gene model). Remaining NMIs were annotated as intergenic.

For mouse ES cell and zebrafish 24hpf embryos, published lncRNA models, Ensembl non-coding RNA annotations, tissue-specific H3K4me1 and RNAseq data were used to account for intergenic NMIs in a hierarchical manner. Where short read genomic alignments were not provided by the authors, chromatin mark datasets were aligned to the appropriate reference genome using Bowtie and peaks were called using MACS as above (2.14.5). Throughout, overlap of genomic intervals was assessed using BEDTools (Quinlan and Hall, 2010) and statistical significance calculated using the Genomic Association Tester (GAT) (Ponjavic et al., 2007).

2.15.3 Nucleotide properties of non-methylated islands

CpG observed/expected (CpG O/E) and GC content were calculated as in Gardiner-Garden et al (Gardiner-Garden and Frommer, 1987). Both measures were calculated for each non-methylated island (NMI) and for a control region of the same size 10kb upstream of each NMI interval.

2.15.4 Evolutionary conservation of NMIs

Evolutionary conservation of protein-coding genes was calculated using OPTIC (Heger and Ponting, 2008). Conserved genes (pairwise 1:1 orthologues) were defined as having an NMI or not depending on whether a 1 kb region centred on the gene TSS was overlapped or not overlapped by 1 bp with an NMI. The conservation score was calculated as: n/min(x,y) where n is the number of conserved genes with an NMI at the TSS of both orthologues and x and y are the numbers of conserved genes with a TSS-associated NMI in each species. For three-way conservation, 1:1:1 orthologues from human, mouse and zebrafish were defined as having an NMI in one, two or all species.

2.15.5 Tissue-specific and broad NMIs

NMIs were called from non-methylated DNA profiles in both testis and liver using MACS (v1.4.0)(Zhang et al., 2008). An NMI was defined as tissue-specific if it did not overlap with an NMI in the other tissue. Broad NMI-associated genes were defined as having greater than 90% of their gene length covered by NMIs. Short NMI-associated genes had less than 10% (but greater than 0%) gene coverage by NMIs.

2.15.6 Gene expression analysis

Published tissue-specific RNA-seq data was obtained for testis and liver from Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) for human, mouse, platypus and chicken. DESeq (Anders and Huber, 2010) was used to identify genes differentially expressed between liver and testis for all four species ($p<0.05$, fold change > 2).

2.15.7 Visualisation of genome-wide data

2.15.7.1 **Composite analysis of H3K4me3 genome-wide**

Published H3K4me3 ChIP-seq data was profiled across tissue-specific NMIs from human and mouse (testis and liver) and frog and zebrafish (embryonic stages stage 11-12 and 24 hpf respectively) using sitepro from the CEAS package (Shin et al., 2009).

2.15.7.2 **Venn diagrams for comparison of interval datasets**

Proportional two-way Venn diagrams were generated in R using the 'VennDiagram' package (Chen and Boutros, 2011). The EulerAPE drawing tool was used to generate a three-way Venn diagram (http://www.eulerdiagrams.org/eulerAPE/).

2.15.8 Gene ontology (GO) analysis

Gene Ontology (GO) analysis was performed using a hypergeometic test. Terms were clustered using a modified ReVigo (Supek et al., 2011) script and a representative term from each cluster was plotted using the GO term enrichment score.

2.16 **Analysis of massively parallel sequencing data – Tc1 mouse**

Bio-CAP sequencing was performed in duplicate for testis and liver tissue from Tc1 mouse (received from Claudia Kutter and Duncan Odom). The combined high salt elutions were analysed by massively parallel sequencing.

2.16.1 Regions of human chromosome absent in Tc1 mouse

The regions lost in Tc1 mouse model from human chromosome 21 were chr21:18735370-18868903, chr21:18972558-19718193, chr21:33626046-36366435 and chr21:46866675-47321951. These regions were removed from appropriate bed files prior to use in comparative analyses using the following bedtools command.

bedtools subtract -A -a bedfile.bed -b unmapped_Tc1_regions.bed > bedfile _rmunmap.bed

2.16.2 Identifying differentially methylated NMIs between human and Tc1 mouse

NMIs were identified for Tc1 testis and liver using MACS as was performed previously for the human datasets (2.14.5). A combined dataset of all NMIs was generated by combining all NMIs from human and Tc1 mouse chromosome 21 for each tissue (testis and liver). Bam files were generated for human chromosome 21 from human and Tc1 mouse and the number of reads were normalised relative to one another by downsampling. The number of reads was then counted for human and

Tc1 mouse at all intervals from the combined NMI dataset. All NMIs which exhibited a 1.5 fold change in read number were considered differentially methylated between human and Tc1 mouse.

2.17 **External datasets**

2.17.1 Seven vertebrate external datasets

Computationally predicted CpG islands for all seven species were obtained from the UCSC genome browser (Kent et al., 2002). LncRNA datasets from recent publications for mouse (Belgard et al., 2011; Guttman et al., 2010) (GSE20851, GSE27243) and zebrafish (Pauli et al., 2012; Ulitsky et al., 2011) (GSE32900 and GSE32880) were obtained from GEO (Edgar et al., 2002) or from supplementary material. H3K4me1 datasets for zebrafish 24hpf (Aday et al., 2011) (GSE20600) and mouse ES cell (Stadler et al., 2011) (GSE30206, GSE11172) were obtained from GEO. Mouse (Smagulova et al., 2011) and human (Bernstein et al., 2010; Hammoud et al., 2009) (liver and testis), frog (Akkers et al., 2009) (stage 11-12) and zebrafish (Aday et al., 2011) (24hpf) H3K4me3 datasets were obtained from GEO (GSE24438, GSE19465, GSE15594, GSE14025, GSE20600). H3K4me3 and H3K27Me3 datasets for mouse ES cells (Mikkelsen et al., 2007) (GSE12241) and frog stage 11-12 (Akkers et al., 2009) (GSE14025) were obtained from GEO. RNAseq datasets for human, mouse, platypus and chicken (Brawand et al., 2011; Julien et al., 2012; Stadler et al., 2011) were obtained from GEO (GSE30280, GSE30352 and GSE36120), RNAseq datasets for lizard (Barbosa-Morais et al., 2012) were obtained before publication from Kutter and Odom (now available at GSE41338) and zebrafish RNAseq data were obtained from the EBI (ERP000016, Sample ERS000081). The XTev dataset (Akkers et al., 2010) was obtained from the Veenstra lab website (http://131.174.221.43/gertjanveenstra/genomedata.asp). ChIP-seq data for a number of polycomb factors profiles for mouse ES cells (Ku et al., 2008) were obtained from GEO (GSE13084).

2.17.2 Tc1 mouse external datasets

H3K4me3 ChIP-seq datasets for the human chromosome 21 mouse model, Tc1, were obtained from the European Nucleotide Archive (ENA), study accession ERP001396 (Ward et al., 2013). CEBPA and

HNF4A ChIP-seq datasets were obtained from the ENA, study accession ERP000054 (Schmidt et al., 2010).

2.18 **Bioinformatics tools**

2.18.1 UCSC Genome Bioinformatics

The UCSC Genome Bioinformatics (http://genome.ucsc.edu/) website was used to browse genomes, perform *in silico* PCR and download various genomic features using the Tables functionality.

2.18.2 National Center for Biotechnology Information (NCBI)

The NCBI website (http://www.ncbi.nlm.nih.gov/) was used to search for genes, to perform BLAST (Basic Local Alignment Search Tool) searches of a DNA or protein sequence of interest, to find vertebrate gene homologues and extract DNA and protein sequences.

2.18.3 Ensembl

The Ensembl website (http://www.ensembl.org/) was used to perform BLAT/BLAST searches of a DNA or protein sequence of interest, to download gene models (Ensembl Gene) from BioMart (www.ensembl.org/biomart/martview/) and search for gene homologs.

2.18.4 ExPASy Bioinformatics Resource Portal

ExPASy Bioinformatics Resource Portal was used for *in silico* translation of a DNA sequence (Translate Tool http://web.expasy.org/translate/) and to calculate the expected molecular weight of a given peptide (Compute pI/Mw tool http://web.expasy.org/compute_pi/).

2.18.5 NEBcutter V2.0

NEBcutter V2.0 (http://tools.neb.com/NEBcutter2/) was used for *in silico* digestion of a given DNA sequence with selected restriction enzymes.

2.18.6 ClustalW2

The ClustalW2 utility (http://www.ebi.ac.uk/Tools/msa/clustalw2/) was used to perform multiple sequence alignments of a number of DNA or protein sequences of interest.

2.18.7 Antigenicity Plot

The antigenicity plot program from JaMBW (http://www.bioinformatics.org/JaMBW/3/1/7/) was used to analyse the antigenicity of the zKDM2B antigen (Hopp and Woods, 1981). The resultant chart shows amino acid position along the x axis and antigenic index on the y axis. High values for the antigenic index indicate that the region is likely to be exposed and accessible for antigen binding.

2.18.8 Vector NTI

The software Vector NTI was used to store DNA and protein sequences with user-added annotations. Vector NTI was used for *in silico* manipulation of DNA sequences by restriction endonuclease digestion or translation tools. DNA and protein sequences were compared by multiple sequence alignment or contig tools.

2.18.9 PyMol

PyMol was used to download and manipulate protein structures from the RCSB Protein Data Bank (PBD, www.rcsb.org/pdb/).

2.19 **Buffers, morpholinos, plasmids and BACs**

2.19.1 Common solutions

Table 2.14 *Common solutions*

Buffer	Composition
Phosphate buffered saline (PBS) 1X	140 mM NaCl 3 mM KCl 2 mM KH_2PO_4 10 mM Na_2HPO_4
Tris-acetate EDTA (TAE) 1X	40 mM Tris 20 mM glacial acetic acid 1mM EDTA Adjusted to pH 8.0
TE buffer	10 mM Tris, 0.1 mM EDTA, pH 7.6

2.19.2 Solutions

Solutions were prepared n milliQ water and filtered and autoclaved where appropriate. Buffers were stored at room temperature unless indicated, 4°C (*) or -20°C (†).

Buffer	Composition
Zebrafish embryo husbandry and manipulation	
60 x E3 *	300 mM NaCl 10.2 mM KCl 20 mM $CaCl_2$ 20 mM $MgCl_2$ Adjust to pH 7.2 with NaOH.
Deyolking buffer	55 mM NaCl 1.8 mM KCl 1.25 mM $NaHCO_3$
MS222 *	0.4 % Ethyl-3-amino benzoate methanesulfonate salt Adjust to pH 7.0.
Whole-mount *in situ* hybridisation (WISH)	
PBS + Tween (PBSTw) 1X	1X PBS 0.1 % Tween-20
4 % paraformaldehyde (PFA) *	10 mL 16 % PFA Diluted 4X with 10 mL PBSTw.
Hybe +/+ *	50% Formamide 5x SSC 9.2 mM Citric Acid 0.1 % Tween 0.05 % tRNA (+) 0.005 % Heparin (+)
Hybe -/- *	50% Formamide 5x SSC 9.2 mM Citric Acid 0.1 % Tween
20 x SSC *	3 M NaCl 0.3 M Sodium Citrate
MABTw *	0.1 M maleic Acid 0.15 M NaCl 0.1 % Tween Adjust to pH 7.5.
MAB Block †	0.1 M maleic Acid 0.15 M NaCl 2 % Boehringers Blocking Reagent Adjust to pH 7.5.
AP Buffer	100 mM Tris-HCl pH 9.5 100 mM NaCl 50 mM MgCl2 0.1 % Tween Make fresh.
Bleaching solution	10 % hydrogen peroxide 5 % formamide 0.5 % SSC Make fresh.

50 – 90 % glycerol	50 – 90 % glycerol 50 – 10 % PBSTw
Western Blot	
Semi-dry transfer buffer	Glycine 39 mM Tris 48 mM SDS 0.037 % Methanol 20 %
ECL A *	Luminal (in DMSO) 2.5 mM p-comedic acid (in DMSO) 396 µM Tris-HCl pH 8.5 100 mM
ECL B *	H_20_2 56.32 mM Tris-HCl pH 8.5 100 mM
Electrophoretic mobility shift assay	
Binding buffer (4x) *	16 % Bicol 400 80 mM Hepes pH 7.9 600 mM KCl 4 mM EDTA 2 mM DTT
20X TBE running buffer (1L)	Tris 216g Boric acid 110g EDTA 80ml of 0.5M pH 8.0 (No EDTA for CxxC domain proteins) Adjust to 1L with H2O
BAC mini-prep	
P1 buffer (4°C) *	50 mM Tris (pH 8.0) 10 mM EDTA (pH 8.0) Add 0.1 mg/ml RNaseA prior to use.
P2 buffer	200 mM Sodium hydroxide 1 % SDS
P3 buffer	3 M Potassium acetate Adjust pH to 5.5 with glacial acetic acid
PAGE and western Blotting	
Deyolking buffer	55 mM NaCl 1.8 mM KCl 1.25 mM $NaHCO_3$
Coomassie blue stain	50 % Methanol 10 % Glacial acetic acid 0.1 % Brilliant blue R250 dye
De-stain solution	30 % Methanol 10 % Glacial acetic acid
SDS-PAGE Running Buffer (10X)	250mM Tris 1.92M Glycine 1 % SDS
Sample Buffer (4X)	8 % SDS 0.4 M Tris (pH 6.8) 0.4 M DTT 40 % Glycerol 4 % Bromophenol Blue

Genomic DNA extraction	
gDNA extraction buffer	20 mM Tris HCl (pH 8.0) 10 mM EDTA 100 mM NaCl 0.5 % SDS (add after homogenisation) 200 µg/ml Proteinase K (add fresh)
Sperm extraction buffer	10 mM Tris HCl (pH 8.0) 10 mM EDTA 100 mM NaCl 1 % SDS (add after homogenisation) 200 µg/ml Proteinase K (add fresh) 40 mM DTT (add fresh)
Biotinylated CxxC affinity purification (Bio-CAP)	
CAP100 CAP300 CAP500 CAP700 CAP1000 *	12.5% Glycerol 0.1% Triton-x-100 20 mM HEPES pH 7.9 X mM NaCl Different CAP buffers have increasing salt concentrations: CAP100, X = 100 mM NaCl CAP300, X = 300 mM NaCl CAP500, X = 500 mM NaCl CAP700, X = 700 mM NaCl CAP1000, X = 1000 mM NaCl

2.19.3 Morpholinos

Morpholinos were designed by Gene Tools (http://www.gene-tools.com/) and a BLAST search was performed to ensure that the designed morpholino had only one target in the zebrafish genome (Table 2.15). Morpholinos were designed to bind across or upstream of the ATG translation start sites (aMO) or to bind across a splice junction to disrupt normal splicing (sMO). Lyophilised morpholino stocks were rehydrated in distilled water to make a stock concentration of 50 ng/nL which was aliquoted and stored at -20°C. For injection, morpholinos were diluted in distilled water to a working concentration and injected in the range 1-20 ng. Prior to use, morpholinos were heated to 65°C for 5 min and then placed on ice.

Table 2.15 *Morpholinos used in this thesis*

Gene	GeneID	Morpholino name	Sequence (5' - 3')	Target site
zkdm2aa (1)	571851	sMO-1	CCAGACCGCTGTAAGACACGAGCGA	i1-e2
zkdm2b (10)	100536129 & 562643 *	sMO-2	AAACCATCTGTGTTCTTACCACTGT	e2-i2
zkdm2ab (14)	799441	sMO-3	GCATGATGAAATGAGCCATACTTTA	e3-i3
zkdm2aa (1)	571851	aMO-1	CAGCGGACATCACGTCCATAACCGA	Overlap ATG
zkdm2b (10)	100536129 & 562643 *	aMO-2	ACGCCTCCATGTCCGAGTTGTCCGT	Overlap ATG
zkdm2ab (14)	799441	aMO-3	ACGGGATTTAGCCTCTGACATACTC	Overlap ATG
zkdm2aa (1)	571851	aMO-5	GGCCGTCTGGTGAAAGAAACCTACA	Upstream of ATG
zkdm2b (10)	100536129 & 562643 *	aMO-6	GCGTTATTTCGGGCATTTTTCACGC	Upstream of ATG
zkdm2ab (14)	799441	aMO-7	TAGCCTTCAACTTTCGCTTTCCGCC	Upstream of ATG
zpcgf1	368900	aMO-8	CCATCGGACCTTGCTCCGCCATCTT	Overlap ATG
zkdm2b (10)	100536129 & 562643 *	aMO-9	CACACTTTGATTGCTTGTGCTGAAA	Upstream of ATG
zrnf2	327035	aMO-10	TGGACCGTCTGTGTCATCTTCATAC	Overlap ATG
zrnf2	327035	aMO-11	ATGTTGTGAACACTCTCTCTCCTGC	Upstream of ATG
zpcgf1	368900	aMO-12	TTCTGCATACTTCCGGGTAACGGGA	Upstream of ATG
zkdm2b (10)-SF	100536129 & 562643 *	aMO-13	AGCTGACATGGCTACGGCTTGCTAC	Overlap ATG

* Zebrafish Kdm2b, zKdm2b, on chromosome 10 is currently annotated as two genes on NCBI: the 5' end is "LOC100536129 lysine-specific demethylase 2B-like" (gene ID 100536129) and the 3' end is "kdm2bb lysine (K)-specific demethylase 2Bb" (gene ID 562643). Target sites: i=intron, e=exon

2.19.4 Plasmids

Table 2.16 *ESTs used to clone zebrafish kdm2 genes*

Klose Plasmid Index	EST name	Gene	Name/Clone ID	Source	Vector	Selection
393	1_1	*zkdm2aa* (1)	BC129321	Open Biosystems	pME18S-FL3	Amp
394	1_2	*zkdm2aa* (1)	AL922788	Open Biosystems	pBluescript SK+	Amp
395	1_3	*zkdm2aa* (1)	BC128874	Open Biosystems	pCMV-SPORT6.1	Amp
396	14_1	*zkdm2ab* (14)	EB931586	Open Biosystems	pSPORT-1	Amp
397	14_2	*zkdm2ab* (14)	DR730314	imagenes	p-Express-1	Amp
398	10_1	*zkdm2b* (10)	CK027726	imagenes	p-Express-1	Amp
393	10_2	*zkdm2b* (10)	ZF_mu_249N16	gene services	pSPORT1	Amp
340	10_3	*zkdm2b* (10)	EB782957	Open Biosystems	pCCM114	Amp
341	10_4	*zkdm2b* (10)	BG799427	Open Biosystems	pT7T3D-PacI	Amp

Table 2.17 *Plasmids generated in this thesis*

Klose Plasmid Index	Name	Purpose	Tag	Parent vector	Selection
407	KDM2aa (1) FL			pCMV-SPORT6.1	Amp
408	KDM2b (10) delta 2b			pCCM114 vector	Amp
409	KDM2b (10) FL			pCCM114 vector	Amp
410	KDM2aa (1) CxxC-PHD (from 2.1_3)	PE	His (CT)	177 pNIC28	Kan
411	KDM2b (10) CxxC-PHD (from 2.10_3, delta 2bp)		His (CT)	177 pNIC28	Kan
412	KDM2b (10) CxxC-PHD (from 2.10_4)	PE	His (CT)	177 pNIC28	Kan
413	KDM2ab (14) CxxC-PHD (from 2.14_1)	PE	His (CT)	177 pNIC28	Kan
414	KDM2aa (1) antigen	PE	His (CT)	177 pNIC28	Kan
415	KDM2ab (14) antigen	PE	His (CT)	177 pNIC28	Kan
416	KDM2b (10) antigen	PE	His (CT)	177 pNIC28	Kan
417	KDM2aa (1) aMO Sensitive rescue construct (pCS2+)	IVT	3xFlag (CT)	421 pCS2+	Amp
418	KDM2aa (1) aMO Insensitive rescue construct (pCS2+)	IVT	3xFlag (CT)	421 pCS2+	Amp
419	KDM2ab (14) aMO Sensitive rescue construct (pCS2+)	IVT	3xFlag (CT)	421 pCS2+	Amp
420	KDM2ab (14) aMO Insensitive rescue construct (pCS2+)	IVT	3xFlag (CT)	421 pCS2+	Amp
540	eF1alpha GFP polyA pT2AL200R150G Kan R	BR		1517 miniTol2 GFP – Patient Lab	Kan
541	eF1alpha GFP polyA pT2AL200R150G Kan R	BR			Kan
603	iTol2_amp_eF1alpha-gfp_kan	BR		445	Amp Kan
612	iTol2_amp_eF1alpha-gfp_HApBAC3.6-pGEMT-easy	BR		603	Amp
613	zKDM2B aMO sens pCS2+ 3xFLAG	IVT	3xFlag (CT)	pCS2+ & 409	Amp
614	zKDM2B aMO ins pCS2+ 3xFLAG	IVT	3xFlag (CT)	613	Amp
615	pCAGIpuro_LIC_zKDM2B_FS2	EE	Flag strepII	441 pCAGIpuro LIC	Amp
619	619 pGTvL2 zKDM2B (10) antigen	PE	His/GST	273 pGTVL2	Kan

Purpose: BR = BAC recombineering; IVT = *In vitro* transcription; EE = Eukaryotic Expression; PE = Prokaryotic Expression. NT = N-terminal, CT = C-terminal.

2.19.5 Bacterial artificial chromosomes (BACs)

Table 2.18 *Bacterial artificial chromosomes used in this thesis*

Name	BAC	Genomic clone	Location	Vector
mBAC1	RP23-189D14	RPCIB731D14189Q	chr3:103,598,642-103,803,302	pBACe3.6
mBAC2	RP23-477B14	RPCIB731B14477Q	chr8:74,666,059-74,866,244	pBACe3.6
zBAC2	DKEY-194K4	HUKGB735K04194Q	chr2:29,588,938-29,829,367	pIndigoBAC-536
zBAC3	DKEY-5K13	HUKGB735K135Q	chr13:31,858,931-32,077,313	pIndigoBAC-536

2.19.6 *In situ* hybridisation probes

Table 2.19 In situ *hybridisation probes generated in this thesis*

Name	Gene	F primer		R primer		Location	Nucleotide position
ISH-1	*zkdm2aa* (1)	991	(T3)	992	(T7)	3'UTR	104-914 beyond TAA
ISH-2	*zkdm2ab* (14)	987	(T3)	988	(T7)	3'UTR	218-? beyond TAA
ISH-3	*zkdm2b* (10)	989	(T3)	990	(T7)	between CxxC-Fbox	2053-2415 from ATG
ISH-4	*zkdm2aa* (1)	1242	(T3)	1243	(T7)	between PHD-Fbox	1971-2878 from ATG
ISH-5	*zkdm2ab* (14)	1246	(T3)	1247	(T7)	between PHD-Fbox	1959-2929 from ATG
ISH-6	*zkdm2b* (10)	1244	(T3)	1245	(T7)	3'UTR	75-785 beyond TAG

Table 2.20 *Other* in situ *hybridisation probes used in this thesis*

Patient lab plasmid index	Gene	Linearise (antisense)	Polymerase (antisense)	Ref
166	*gata2*	EcoRI	T7	
236	*ntl*	XhoI	T7	(Schulte-Merker et al., 1992)
239	*bmp2b*	EcoRI	T3	
308	*wnt8a*	EcoRI	T7	(Kelly et al., 1995)

Chapter 3 -- Biotinylated CxxC affinity purification is a robust method to selectively purify non-methylated DNA

3.1 **Introduction**

In vertebrate genomes, the majority of DNA methylation occurs at cytosine residues in a CpG dinucleotide context, with 70-80% of CpGs existing in the methylated state (Bird, 2002; Bird et al., 1985; Ehrlich et al., 1982). CpGs which remain non-methylated are clustered at around 70% of gene promoters, where they appear to be protected from DNA methylation (Deaton and Bird, 2011). It has been known for over thirty years that DNA methylation has a repressive effect on transcription (Doerfler, 1983). A link between DNA methylation at gene promoters and transcriptional silencing (Esteller, 2007; Suzuki and Bird, 2008; Thomassin et al., 2001) has provoked tremendous interest in investigating DNA methylation status. As a consequence, a large number of techniques have been developed for analysing the methylation state of DNA (Laird, 2010), the majority of which detect the presence of methylated CpGs. However, due to the strong enrichment of non-methylated DNA at mammalian gene promoters, we sought to develop a technique for the direct interrogation of non-methylated DNA which was fast, simple and cost-effective.

In this Chapter, the development and characterisation of the technique, biotinylated CxxC-affinity purification (Bio-CAP) will be described.

3.2 **Currently available techniques for DNA methylation analysis**

Many techniques have been developed for analysing the methylation state of DNA (see 1.11). An appropriate methodology was sought here for profiling the non-methylated fraction of the genome. Bisulfite sequencing generates single nucleotide resolution maps of DNA methylation, but is prohibitively expensive for many applications and it can be a challenge to map reads due to a loss of sequence complexity (1.11.1.3). Methods which affinity purify methylated DNA (for example MethylCap and MeDIP-seq) are limited in their detection of non-methylated DNA at genomic loci

with low CpG density, as it is uncertain whether lack of signal can be attributed to lack of DNA methylation or low CpG density and subsequent failure to purify the small amount of methylated DNA. MRE-seq (methylation sensitive restriction enzyme sequencing) directly interrogates the presence of non-methylated DNA, but is biased towards regions of the genome containing restriction sites.

CxxC affinity purification (CAP), first described by Adrian Bird and colleagues, directly purifies the non-methylated fraction of the genome (Illingworth et al., 2008). Since only a small fraction of the vertebrate genome is non-methylated, enrichment of non-methylated DNA coupled to massively parallel sequencing reduces the sequencing depth requirements and enables detection of non-methylated DNA even at CpG sparse regions (Illingworth et al., 2010). Therefore, from the techniques currently available for DNA methylation analysis, the CAP method was chosen for further development. In its current format CAP requires large amounts of input material and an automated chromatography system for elution of DNA. Therefore, optimisation steps were undertaken to create a more streamlined technique which is amenable to small amounts of input material.

3.3 **Developing a modified CAP technique**

3.3.1 CxxC affinity purification

In the original CAP protocol, recombinant His-tagged MBD1 ZF-CxxC protein is immobilised on nickel-charged sepharose beads, and packed into a 1 mL chromatography column (Figure 1.17C) (Illingworth et al., 2008). Genomic DNA, digested with MseI (a frequent cutter, that will digest AT-rich DNA at T^TA_A sites) was passed over the column and allowed to bind to the immobilised ZF-CxxC domain. DNA was then eluted with an increasing NaCl gradient (Illingworth et al., 2008). Due to the size of the column, the CAP elutions were large and required precipitation prior to downstream analysis. These extra DNA handling steps generally lead to a loss of material, and resulted in large amounts of starting material being required. This requirement precludes CAP from being applied to samples where material is limiting, such as patient samples, rare cell-types, or early developmental

stages. Furthermore, traditional CAP is inaccessible for some laboratories as it requires large amounts of recombinant ZF-CxxC protein and an automated chromatography system for processing of samples. Due to these limitations, optimisation steps were undertaken to develop an improved CAP technique.

3.3.2 The ZF-CxxC domain as a non-methylated DNA affinity module

3.3.2.1 **Human KDM2A and KDM2B specifically bind non-methylated DNA *in vitro***

In order to improve the CxxC-affinity purification technique, the first aim was to identify a high affinity non-methylated DNA binding module for a modified CAP technique. The DNA binding properties of two ZF-CxxC domain-containing proteins, human KDM2A and KDM2B, were investigated *in vitro* by electrophoretic mobility shift assay (EMSA, 2.12.4) which interrogates the *in vitro* binding of a protein to a radiolabelled DNA probe. Recombinant KDM2A and KDM2B ZF-CxxC domains (KDM2A-12 and KDM2B-3, see Appendix 2.1-4ii and 2.10.1) were incubated with a radiolabelled 216 bp DNA oligonucleotide probe (2.12.1) which was either *in vitro* methylated using M.SssI or left non-methylated. With increasing protein concentration, the migration of the radiolabelled probe was increasingly retarded for hKDM2B as was previously shown for hKDM2A (Blackledge et al., 2010), demonstrating that both proteins are able to bind non-methylated DNA *in vitro* (Figure 3.1C). Importantly, methylation of the DNA probe abrogated DNA binding of both proteins, with very little or no retardation in band migration observed with the *in vitro* methylated DNA probe (Figure 3.1C). Together, the human KDM2B ZF-CxxC domain appears to be an ideal affinity module for purifying non-methylated DNA, as it has a high affinity for non-methylated CpGs, higher than that of KDM2A (Figure 3.1C). Furthermore, the KDM2B ZF-CxxC domain appeared stable (Figure 3.2B) unlike that of the MBD1 ZF-CxxC domain (used in conventional CAP) which exhibited a reduced functional concentration due to degradation (Rob Klose, unpublished observations).

Interestingly, increasing the hKDM2A or hKDM2B protein concentration appears to cause a step-wise shift in band migration for both KDM2A and KDM2B as opposed to an all-or-nothing shift. This

can likely be attributed to the fact that multiple CpGs present in the DNA probe (18 in total) constitute multiple binding sites for ZF-CxxC domain binding. Therefore, an increase in ZF-CxxC protein concentration may increase the number of ZF-CxxC domains bound to a single probe, leading to a greater retardation in probe migration.

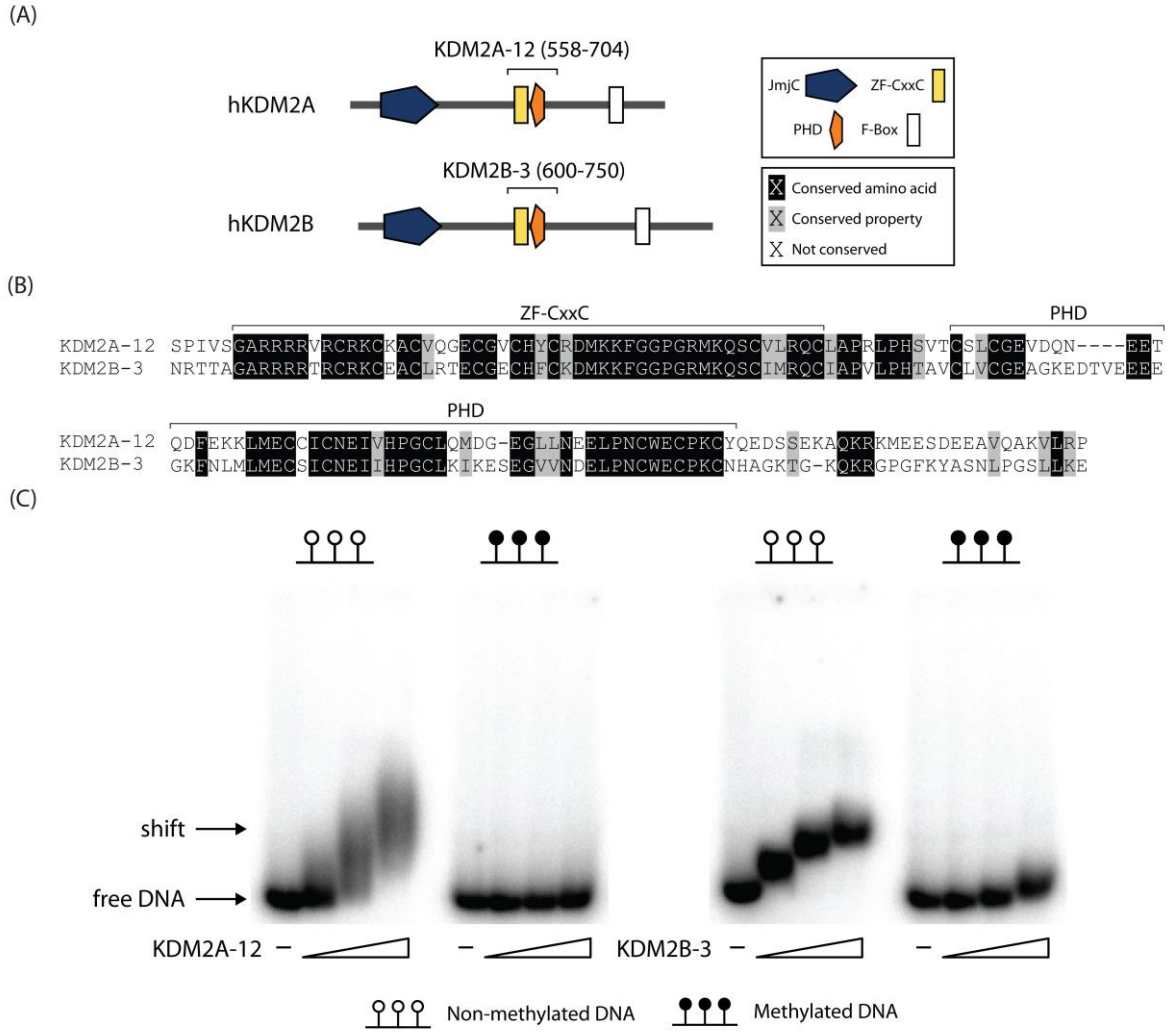


Figure 3.1 Human KDM2A and KDM2B bind to non-methylated DNA

(A) Domain architecture of human KDM2A and KDM2B. The ZF-CxxC and PHD domain containing constructs KDM2A-12 and KDM2B-3 are illustrated above with amino acid sequence indicated in brackets. (B) Multiple sequence alignment of the ZF-CxxC and PHD domain containing region from human KDM2A and KDM2B (constructs KDM2A-12 and KDM2B-3 respectively). Conservation of amino acid sequence is indicated by shading and the ZF-CxxC and PHD domains are highlighted above the alignment. (C) Recombinant human KDM2A and KDM2B specifically shifts a non-methylated 216 bp 601 probe in an electrophoretic mobility shift assay (EMSA) in a 1.3 % agarose gel. Increasing concentration of recombinant protein causes an increased shift (0, 240, 480, 960 ng KDM2A-12 protein and 0, 120, 240 and 480 ng KDM2B-3 protein). Methylation of the 601 probe abrogates binding of recombinant human KDM2A and KDM2B. Both recombinant proteins were purified using an N-terminal His-tag which was cleaved prior to performing the EMSA assay.

3.3.2.2 **Production of a biotinylated ZF-CxxC affinity module**

In the CAP approach, the ZF-CxxC domain needs be fixed to a solid support. The original CAP technique utilised a His-tag to immobilise the MBD1 ZF-CxxC domain onto a Ni-NTA based resin (Illingworth et al., 2008). A His-tag is not an ideal linkage to a solid support as Ni-NTA resin has intrinsic charge properties which may cause non-specific interactions with DNA. Furthermore, the presence of an N-terminal His-tag appeared to be detrimental to non-methylated DNA binding of the KDM2B ZF-CxxC domain (Figure 3.2A and Figure 3.3A, left panel compare His-CxxC with CxxC). To overcome these disadvantages, a 15 amino acid AviTag^{TM} (GLNDIFEAQKIEWHE), which can be *in vitro* biotinylated at a specific lysine residue by recombinant BirA ligase (Beckett et al., 1999; Schatz, 1993), was engineered onto the C-terminus of the KDM2B ZF-CxxC domain (Figure 3.2C, construct KDM2B-4). The interaction between biotin and avidin is one of the strongest known non-covalent interactions in nature (Green, 1963). Therefore, the addition of a biotin moiety onto the hKDM2B affinity module enables strong and specific binding to an avidin-based support. The support chosen was NeutrAvidin resin, a deglycosylated form of avidin, which exhibits a near neutral isoelectric point and hence minimises non-specific interactions with DNA and RNA, while retaining a high affinity for biotin ($K_d = 10^{-15}$ M) .

To produce the improved CAP affinity module, His-CxxC-Avi protein, was expressed in *E. coli* and purified on Ni-NTA resin using the N-terminal His-tag (Figure 3.2B). The His-tag was proteolytically cleaved by His-tagged TEV protease to produce CxxC-Avi, and the cleaved His-tag and His-tagged TEV protease were removed by a second purification on a Ni-NTA resin (Figure 3.2B). Following cleavage, the protein was *in vitro* biotinylated and the biotinylation reaction was monitored by mass spectrometry (see 2.10.3, Figure 3.2D). A fraction of the protein was biotinylated prior to *in vitro* biotinylation (peak at around 19167 Da, Figure 3.2D) due to the endogenous activity of a biotin ligase in *E. coli*. Importantly, *in vitro* biotinylation yielded a homogenously biotinylated species, with complete loss of the unmodified protein (Figure 3.2D, right).

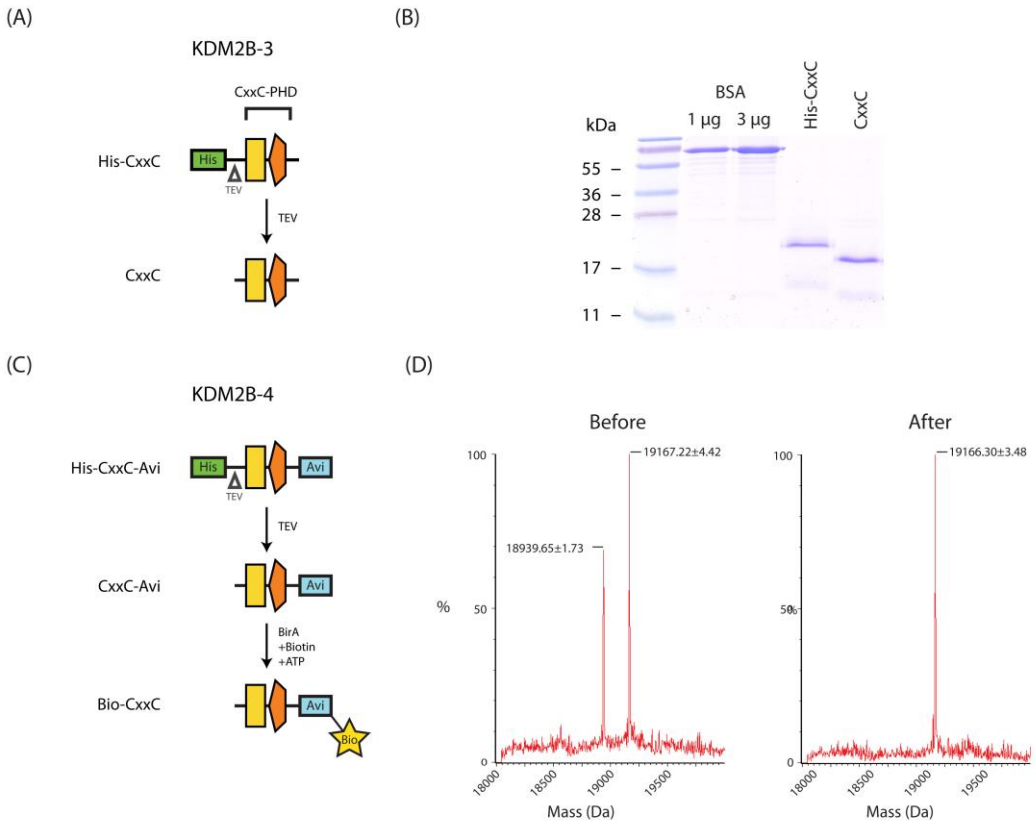


Figure 3.2 Production of human KDM2B ZF-CxxC-PHD construct with AviTag and *in vitro* biotinylation

(A and B) Human KDM2B protein, construct KDM2B-3, containing the ZF-CxxC and PHD domains (amino acids 600-750) was expressed in *E. coli* with an N-terminal His-tag (His-CxxC). The His-tag was removed by TEV cleavage (CxxC). (B) hKDM2B ZF-CxxC protein before and after TEV cleavage was loaded on a polyacrylamide gel along with 1 µg and 3 µg of BSA for quantification of the His-CxxC and CxxC constructs. Proteins were visualised by Coomassie stain. (C) Human KDM2B protein, construct KDM2B-4, with an N-terminal His-tag and a C-terminal AviTag was expressed in *E. coli* (His-CxxC-Avi). Following removal of the His-tag by TEV-cleavage (CxxC-Avi), the AviTag was *in vitro* biotinylated with the enzyme BirA (Bio-CxxC). (D) *In vitro* biotinylation was monitored by mass spectrometry. The expected mass of CxxC-Avi is 18939 Da and Bio-CxxC is 19166 Da.

3.3.2.3 **Non-methylated DNA binding of hKDM2B is unaffected by biotinylation**

The effect of *in vitro* biotinylation on the DNA binding properties of the hKDM2B ZF-CxxC-PHD domain was tested by EMSA. Both the Avi-tagged construct (CxxC-Avi) and the *in vitro* biotinylated construct (Bio-CxxC) demonstrated equivalent binding to the non-methylated DNA probe (Figure 3.3B, left) while no interaction was detected for either construct with the methylated probe (Figure 3.3B, right). Hence, neither the presence of the AviTag nor the biotin moiety negatively impacted the binding of the KDM2B protein to non-methylated DNA. Therefore the biotinylated affinity module,

Bio-CxxC, appeared suitable for use in a modified CAP-like purification strategy which will be referred to as biotinylated CxxC affinity purification (Bio-CAP).

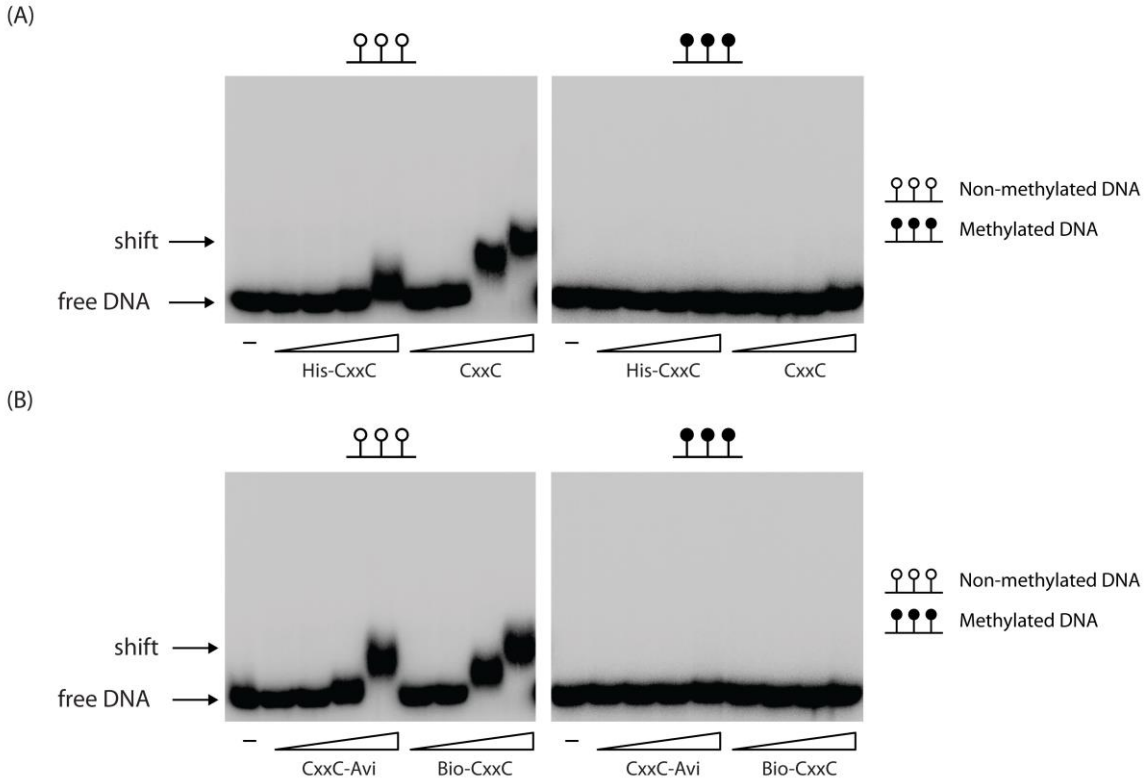


Figure 3.3 A biotinylated ZF-CxxC module retains strong affinity for non-methylated DNA

(A) His-CxxC and CxxC recombinant proteins (construct KDM2B-3) cause a shift in mobility of a non-methylated 216 bp CpG containing 601 DNA probe in a concentration-dependent manner by EMSA. EMSA shift appears to be increased upon removal of the His-tag by TEV cleavage (left panel). His-CxxC and CxxC binding is abrogated by DNA methylation of the CpG probe (right panel). (B) CxxC-Avi recombinant protein (construct KDM2B-4) causes a shift of a non-methylated CpG containing probe, and biotinylation does not appear to affect binding (Bio-CxxC, left panel). CxxC-Avi and Bio-CxxC binding is abrogated by DNA methylation of the CpG probe (right panel).

3.3.3 Biotinylated ZF-CxxC affinity purification (Bio-CAP)

3.3.3.1 **A high affinity DNA binding module makes Bio-CAP a simple, small-scale method**

Having generated a high affinity non-methylated DNA binding module, the next aim was to design a streamlined Bio-CAP protocol for the purification of non-methylated DNA. The nature of the biotin-avidin interaction, together with the choice of solid support resin, means that a simple, small-scale Bio-CAP procedure, including all DNA binding, wash and elution steps, can be performed in standard microcentrifuge tubes in batch (Figure 3.4C). A manual step-wise wash-elution profile was therefore

designed to first disrupt non-specific interactions with methylated DNA or CpG poor DNA prior to elution of strongly bound non-methylated DNA. The small elution volumes used in Bio-CAP eliminates the need for DNA precipitation, and therefore requires less starting material.

3.3.3.2 **Preparation of DNA for Bio-CAP**

In conventional CAP, genomic DNA was digested with MseI to cut bulk AT-rich DNA into small fragments and leave CpG island DNA largely intact. It was estimated that bulk DNA would be digested to an average size 123 bp while CpG island DNA would be 625 bp in size on average (Illingworth et al., 2008). The short AT-rich fragments would be unlikely to have enough CpGs to be retained on the beads, while the longer GC rich fragments would be enriched for CpGs and be retained on the column (Illingworth et al., 2008). While this method allows for a prior enrichment step of GC-rich DNA, it also introduces a bias in fractionation and AT-rich, with non-methylated AT-rich regions excluded from the analysis. Therefore, in order to obtain unbiased DNA fragments of an appropriate size for qPCR analysis and massively-parallel sequencing (100 – 400 bp), genomic DNA was mechanically sheared by sonication (see 2.13.2.2).

3.3.3.3 **Bio-CAP protocol**

A typical Bio-CAP experiment began with around 8 µg of purified genomic DNA (compared to 100 µg for conventional CAP), which was sonicated to around 150 – 350 bp in length with appropriate shearing validated on an agarose gel by electrophoresis (Figure 3.4B). The immobilised affinity module was prepared by incubating 25 µg of recombinant Bio-CxxC protein with 25 µL of packed NeutrAvidin resin. Two stringent washes with CAP1000 buffer (containing 1 M NaCl), were required to remove contaminating bacterial DNA bound to the ZF-CxxC from the expression and purification from *E. coli*. Washes were performed by sedimenting the resin by centrifugation and gentle removal of the wash buffer by pipetting. The beads were re-equilibrated with CAP100 buffer (containing 100 mM NaCl) prior to incubation with the sonicated genomic DNA for 1 hr to allow non-methylated DNA to bind to the ZF-CxxC coated resin. The flow-through (FT) was collected following

sedimentation of the ZF-CxxC resin (Figure 3.4C). The beads were then washed twice with CAP100 buffer to remove any remaining unbound DNA. Material remaining bound to the beads was then subject to sequential elutions with increasing NaCl concentration (300 mM, 500 mM, 700 mM and 1 M NaCl) to disrupt electrostatic interactions between DNA and the ZF-CxxC binding module. Each 50 µL elution step was repeated and combined in order to ensure complete removal of DNA eluted from the beads at each step (Figure 3.4C).

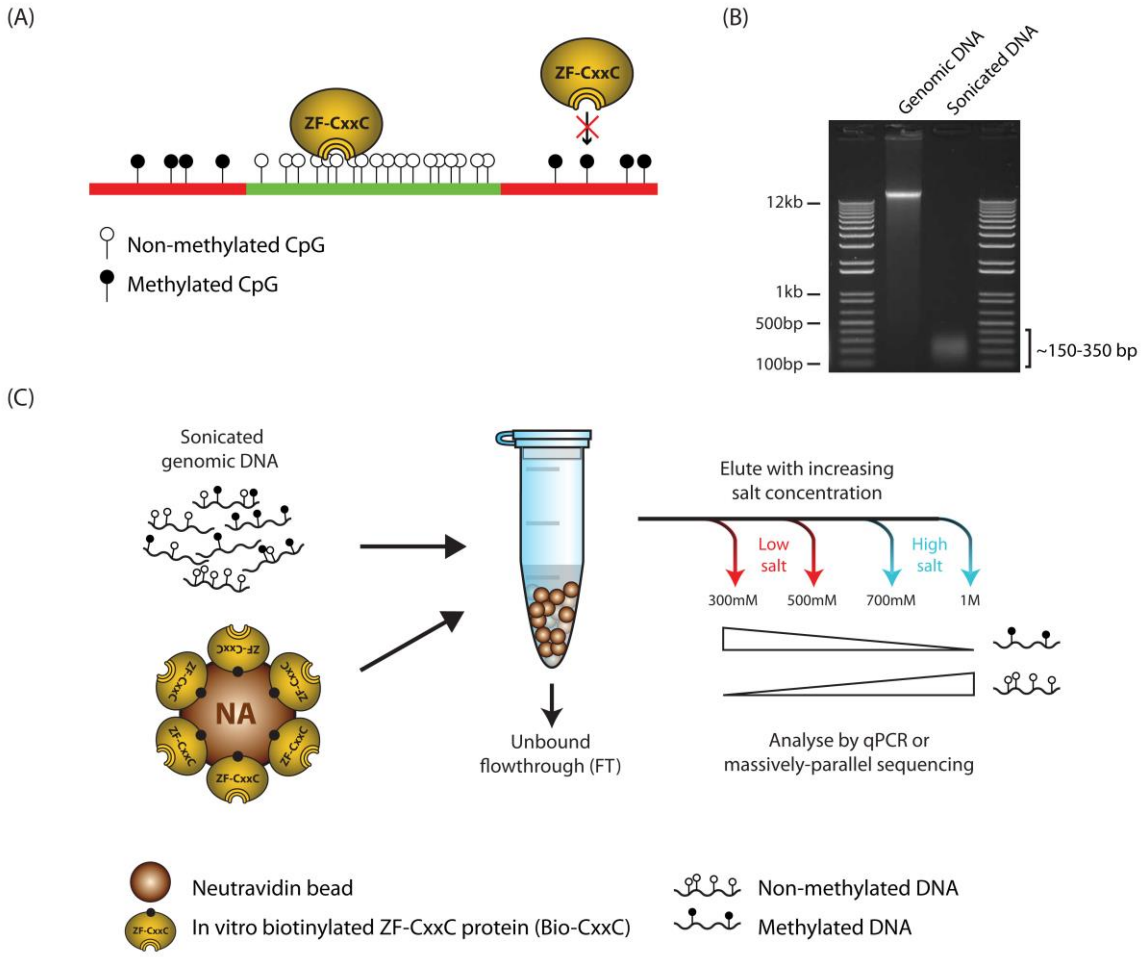


Figure 3.4 The Bio-CAP technique

(A) The ZF-CxxC domain is able to bind to non-methylated CpG dinucleotides. ZF-CxxC DNA binding is abrogated by DNA methylation of cytosine at CpG dinucleotides. (B) For Bio-CAP experiments, genomic DNA was sonicated for two hours to yield DNA fragments of 150-350 bp. Genomic DNA shown was extracted from mouse testis. (C) A schematic of the Bio-CAP procedure. *In vitro* biotinylated recombinant ZF-CxxC protein (Bio-CxxC) is immobilized onto NeutrAvidin beads and incubated with sheared genomic DNA (B), allowing the ZF-CxxC domain to bind to DNA. Unbound DNA is removed in the FT, a series of elutions then follow using increasing salt concentrations. Highly methylated DNA is removed in the early low salt elutions (300 mM and 500 mM) and non-methylated DNA is enriched in the late high salt elutions (700 mM and 1000 mM). DNA from each of these fractions is interrogated by qPCR.

3.3.4 Bio-CAP specifically enriches for CpG island DNA in mouse

To investigate the efficiency of the technique, genomic DNA was extracted from mouse embryonic stem cells (V6.5) and subjected to Bio-CAP. Locus-specific primer pairs designed to gene promoters and bodies of CGI and non-CGI genes (Blackledge et al., 2010) were used to assess enrichment of different genomic regions in each Bio-CAP elution fraction. The *Suv420h1* gene in mouse has a promoter-associated CGI, and qPCR analysis revealed that this promoter region was mostly present in the high salt fractions (700 mM and 1 M), with little or no *Suv420h1* CGI DNA detected in the FT or low salt (300 mM) fractions (Figure 3.5A). In contrast, a region from the *Suv420h1* gene body was not detected in the high salt fractions, but was instead enriched in the FT and 300 mM elution (Figure 3.5A). Calculating the ratio of promoter versus body in each of the fractions revealed that for *Suv420h1*, the 700 mM and 1 M fractions are highly enriched for *Suv420h1* promoter DNA, and hence are "CGI-enriched". Similar qPCR analysis was performed using promoter and body primer sets for *Fabp7*, a gene which lacks a CGI at its promoter. qPCR primer sets from *Fabp7* body region reveal a very similar profile to that observed for *Suv420h1* body. Importantly, the *Fabp7* promoter region, unlike *Suv420h1*, was mostly eluted in the FT and low salt fractions in keeping with the lack of a CGI at this gene promoter (Figure 3.5C). From this analysis, it was clear that the high salt fractions exhibit a substantial enrichment for the *Suv420h1* CGI compared to the gene body regions and non-CGI promoters. This becomes more apparent when signal from the 700 mM and 1 M elutions are combined into a high-salt CGI-enriched fraction (Figure 3.5G and I), and Bio-CAP signal is only observed for the *Suv420h1* CGI promoter.

The *in silico* CpG island prediction algorithm has been used to predict the presence of non-methylated DNA across vertebrate genomes (Gardiner-Garden and Frommer, 1987), however some regions which contain non-methylated CpGs are overlooked by this algorithm. One such example of this is the *Fastkd2* gene, which does not have a computationally detectable CGI at its promoter, despite being identified as non-methylated in mouse sperm by CAP-sequencing (Illingworth et al., 2010). Consistent with this previous observation, Bio-CAP qPCR analysis detects enrichment of the

Fastkd2 promoter in the high salt elutions (Figure 3.5B), with high levels of enrichment over the control body region in the combined 700 mM and 1 M elutions (Figure 3.5H), suggesting that this gene promoter is also non-methylated in mouse ES cells (Figure 3.5B). Therefore, despite the fact that *Fastkd2* gene promoter region falls beneath the detection requirements for bioinformatically defined CpG island detection, Bio-CAP is able to detect the presence of non-methylated CpGs at this locus. Furthermore, the combined 700 mM and 1 M elutions for the *Fastkd2* promoter appeared to be intermediate between that of *Suv420h1* and *Fabp7* perhaps indicating that Bio-CAP may be partially quantitative in the detection of non-methylated DNA at distinct loci.

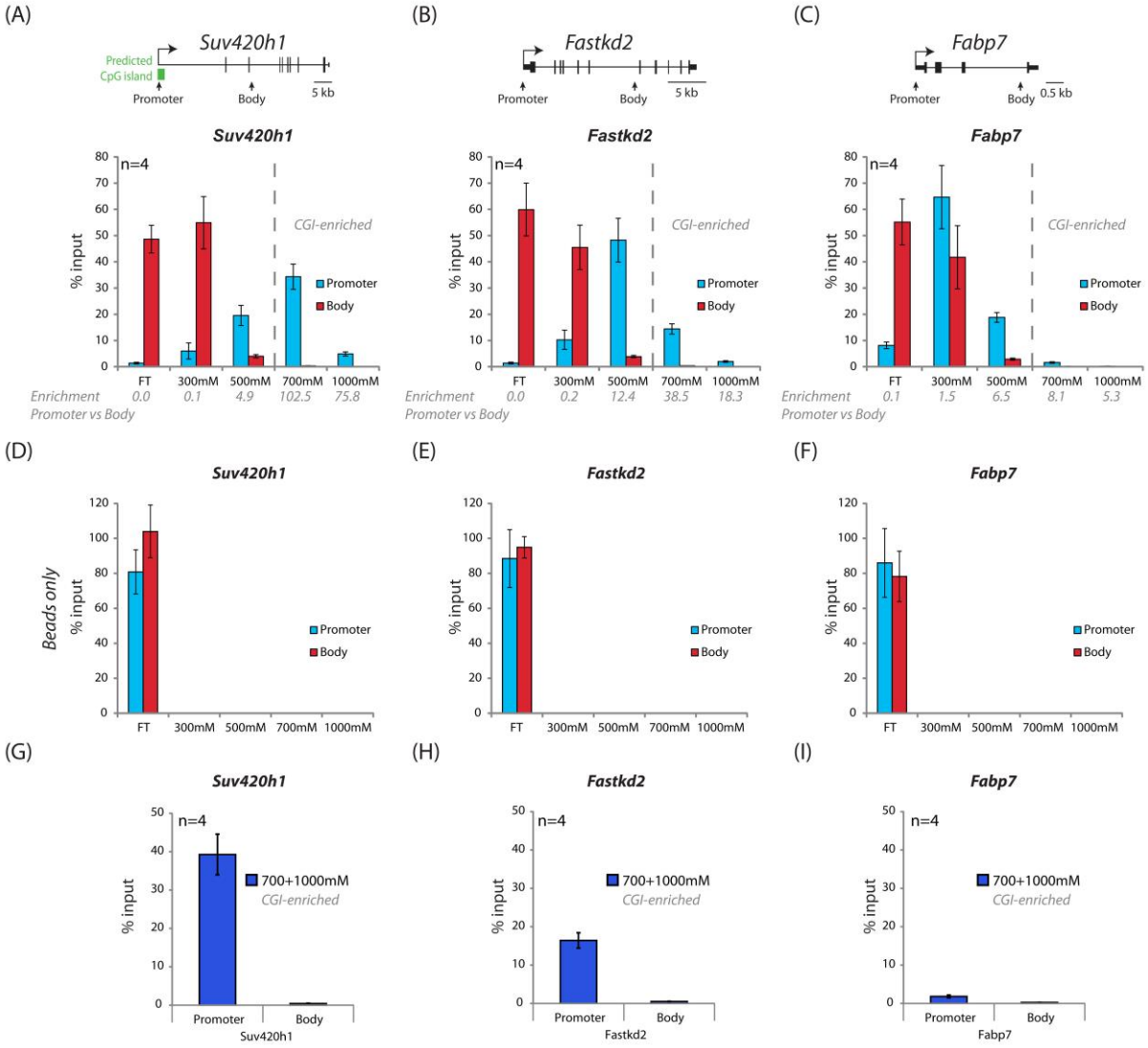


Figure 3.5 Bio-CAP specifically purifies CGI DNA

(A-C) The mouse CGI gene *Suv420h1* and non-CGI genes *Fastkd1* and *Fabp7* are shown schematically with exons shown as black boxes and the transcription start site as an arrow. qPCR primer sets were designed

at the promoter and in the gene body for each gene as indicated by arrows and CpG islands are shown in green. Bio-CAP was performed using mouse V6.5 embryonic stem cell genomic DNA. DNA from the flow through (FT), and four salt elutions (300, 500, 700 and 1000 mM) was subjected to qPCR using primer sets specific to the *Suv420h1*, *Fastkd1* and *Fabp7* promoter (blue) or body region (red). The 700 mM and 1000 mM elutions appear to be "CGI-enriched". (D-F) Bio-CAP 'beads only' controls followed by qPCR were performed as in (A-C) for mouse V6.5 embryonic stem cell gDNA. All DNA was detected in the FT for the 'beads only' controls. (G-I) Data from (A-C) illustrating the detection of "CGI-enriched" 700 + 1000 mM elutions at the promoter and body of *Suv420h1*, *Fastkd1* and *Fabp7*. Where multiple experiments were performed, n is indicated and error bars are standard error of the mean (SEM) = (standard deviation)/(√n).

To verify that the Bio-CAP elution profiles described above were dependent on the specific interaction of the ZF-CxxC domain with non-methylated DNA, a control experiment was performed using NeutrAvidin beads alone without addition of recombinant ZF-CxxC protein. In these beads-only experiments all DNA from the promoter and body of all three genes analysed was detected in the FT fraction, demonstrating that the NeutrAvidin beads alone exhibit no non-specific binding to DNA (Figure 3.5D-F). Therefore collectively, this analysis has revealed that affinity purification of DNA based on the hKDM2B ZF-CxxC domain is able to specifically retain regions of the genome corresponding to CGIs, which are only eluted at high salt concentrations consistent with their non-methylated state. By contrast, non-CGI regions are not retained and either are not bound by the immobilised ZF-CxxC domain (and are detected in the FT), or exhibit a low affinity interaction which is disrupted by low salt elutions. Therefore the modified CAP approach provides a simple and robust method for specific isolation of non-methylated DNA.

3.3.5 CpG density affects the recovery of non-methylated DNA in Bio-CAP

Interestingly, Bio-CAP qPCR analysis of the three mouse genes above suggested that Bio-CAP may be partially quantitative in detecting non-methylated DNA, as the *Fastkd2* low-CpG gene promoter gave intermediate Bio-CAP qPCR signal in the combined high salt fractions compared to the *Suv420h1* CGI promoter and the *Fabp7* non-CGI promoter. Bio-CAP can only be partially quantitative due to the impact of CpG density on recovery of non-methylated DNA. For example, two regions of DNA of equivalent length may be entirely non-methylated, however if one sequence has a higher density of CpG dinucleotides than the other it is likely that it will be more enriched by the Bio-CAP technique. This is because the higher number of non-methylated CpG sites available will create a higher binding

affinity for the Bio-CAP ZF-CxxC module and therefore enable better retention of the DNA sequence on the beads during the Bio-CAP procedure. Furthermore, should a region be mosaically methylated, that region would be expected to be less enriched by Bio-CAP qPCR due to occlusion of methylated CpG substrates and more easy elution from the Bio-CAP affinity matrix.

To analyse the extent to which Bio-CAP can be considered a quantitative technique, and to discover how CpG density influences Bio-CAP read-out, 200 bp regions of DNA were PCR amplified from the human genome and supplemented into mouse ES cell genomic DNA prior to Bio-CAP. Amplicons were designed to be 200 bp in length to match the average sonication length of the mouse genomic DNA. The different human amplicons were selected to contain specific numbers of CpG dinucleotides (0, 2, 6, 13 and 24 CpGs) and were spiked into mouse genomic DNA such that each amplicon would be at approximately the same molar concentration as the mouse genomic DNA. The spiked Bio-CAP elution fractions were interrogated using human-specific primer pairs within the 200 bp PCR product, allowing specific analysis of each human fragment separately. Unsurprisingly, the fragment containing 0 CpGs was found almost exclusively in the FT and 300 mM fractions, indicating no specific interaction of the ZF-CxxC affinity module with this probe (Figure 3.6A). Increasing the number of CpG dinucleotides in the human spiked probes incrementally increased the enrichment of the probes in the higher salt fractions, with a visible shift in the peak elution fraction towards the more stringent elution conditions. This suggests that DNA probes with increased numbers of CpGs are more tightly associated with the ZF-CxxC affinity module, perhaps because the probe is able to interact with multiple immobilised ZF-CxxC moieties. This is reminiscent of the step-wise migration of a non-methylated CpG probe in an EMSA assay with increasing concentrations of ZF-CxxC protein (see Figure 3.1B).

Importantly, there was a near-linear relationship between the number of CpGs in the 200 bp probe and recovery in the high salt (700 mM and 1M) fractions (Figure 3.6B). This implies that the combined high-salt fractions of Bio-CAP may give a partially quantitative read-out of the absolute number of non-methylated CpG dinucleotides present in a given DNA fragment. Furthermore, this

experiment emphasises that Bio-CAP is able to detect the presence of non-methylated DNA at a given locus, even if the region falls below the threshold for traditional *in silico* definition as a CGI. This is demonstrated for the human probe with only 6 CpG dinucleotides in a 200 bp region, which was robustly detected in the "CGI-enriched" fraction (700 mM and 1M) (Figure 3.6A) despite falling below the CpG observed/expected threshold for CGI prediction. The Bio-CAP technique cannot be defined as truly quantitative between different loci, however, as it is directly impacted by the CpG density in a given region. As we do not understand exactly how CpG density and spacing of CpG dinucleotides across a given fragment of DNA will affect ZF-CxxC binding, we cannot quantitatively compare two given fragments with differing sequence composition. However, comparison of the same locus between tissues or developmental stages using Bio-CAP should allow direct comparison of the relative DNA methylation status, though again the effect of mosaic methylation of CpGs has not been explored.

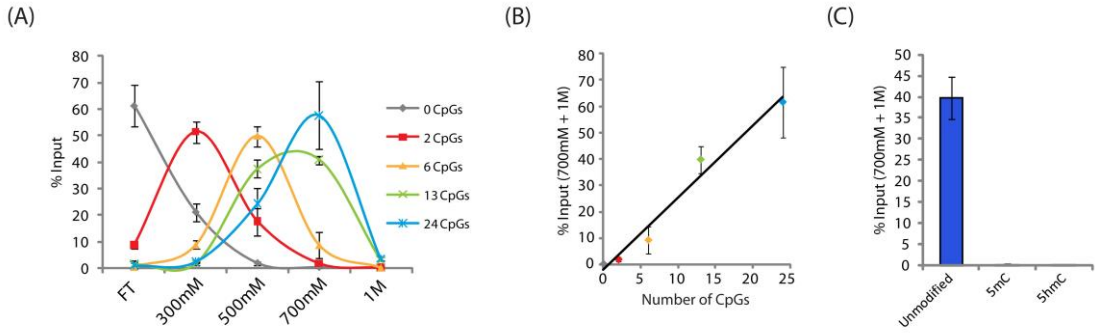


Figure 3.6 Bio-CAP is sensitive to the density and methylation status of CpG dinucleotides

(A) A Bio-CAP experiment with sonicated mouse V6.5 ES cell genomic DNA was spiked with five 200 bp sequences amplified from human DNA each containing a known number of CpGs. Bio-CAP fractions were assayed for the presence of each probe by qPCR using human-specific primers. (B) A scatter graph showing the relationship between number of CpGs in spiked human probes and the percentage input recovery in the high-salt fractions (700 mM + 1 M) from the Bio-CAP experiment shown in (A). A line of best fit is shown. (C) A Bio-CAP experiment with sonicated mouse V6.5 ES cell gDNA was spiked with a 200 bp human probe containing 13 CpGs that were either unmodified, methylated (5mC) or hydroxymethylated (5hmC). Using qPCR analysis, Bio-CAP high-salt fractions (700 mM + 1 M) were assayed for the presence of each of probe variant. All Bio-CAP data are from at least two biological replicates and error bars represent SEM. These experiments were performed by Dr Neil Blackledge.

In order to determine the impact of DNA methylation on the recovery of DNA, the human probe with 13 CpGs was *in vitro* methylated (see 2.13.2). This resulted in complete abrogation of recovery of the probe in the high salt fraction (Figure 3.6C). This demonstrates that DNA methylation

abrogates ZF-CxxC domain binding to DNA in the Bio-CAP experimental set-up. Recently a number of additional modifications of DNA have been identified for cytosines in a CpG dinucleotide context including 5-hydroxymethylcytosine, which is catalysed by oxidation of 5-methylcytosine by ten eleven translocation (TET) enzymes (Almeida et al., 2012a; Almeida et al., 2012b; Ito et al., 2011; Kriaucionis and Heintz, 2009; Song et al., 2012a; Tahiliani et al., 2009). Similarly to the *in vitro* methylated probe, a hydroxymethylated version of the 13 CpG probe demonstrated negligible recovery in the high-salt elutions from Bio-CAP (Figure 3.6C), suggesting that ZF-CxxC interrogation of CpG dinucleotides is also abrogated by hydroxymethylation of cytosine. Together, the spiking experiments reveal that retention of DNA fragments in Bio-CAP is related to the density of non-methylated CpGs and is negatively impacted by cytosine modification. Therefore, Bio-CAP allows for semi-quantitative analysis of the number of non-methylated CpG dinucleotides in a locus-specific manner.

3.4 **The utility of Bio-CAP for analysis of large numbers of low quantity samples**

3.4.1 Bio-CAP is a sensitive technique

Conventional CAP is not amenable to analysis of small amounts of starting material. In contrast, the modified Bio-CAP method uses small elution volumes by due to the utilisation of a high-affinity non-methylated DNA binding module. Therefore, it seemed plausible that Bio-CAP, unlike conventional CAP, may be conducive to DNA methylation analysis of small amounts of starting material.

To examine this possibility, Bio-CAP was performed using only 100 ng of mouse V6.5 ES cell genomic DNA. The combined high-salt fractions were interrogated by qPCR for two CGI promoter genes, *Suv420h1* and *Ncoa2* at promoter and body regions. This analysis revealed that the gene promoter regions of *Suv420h1* and *Ncoa2* were massively enriched compared to gene body regions in the high-salt fractions (Figure 3.7). Similar analysis was performed for two non-CGI genes, *Fabp7* and *Fgf7* which lack non-methylated DNA at their gene promoters. For these genes, little or no enrichment was observed for either the promoter or body regions in the high-salt fractions of Bio-CAP (Figure

3.7B). Importantly, the absolute percentage recovery of promoter DNA using only 100 ng input DNA was comparable to that recovered from a Bio-CAP experiment using 8 µg starting material (compare Figure 3.7A with Figure 3.5A). Together, the Bio-CAP technique is able to detect non-methylated DNA from small amounts of input material, and exhibits consistent recovery of non-methylated DNA across samples with input DNA which differ by almost two orders of magnitude. Bio-CAP may therefore prove to be a valuable technique for the analysis of precious patient samples and furthermore could perhaps be developed as a diagnostic tool should appropriate bio-marker genes with variable DNA methylation status be identified.

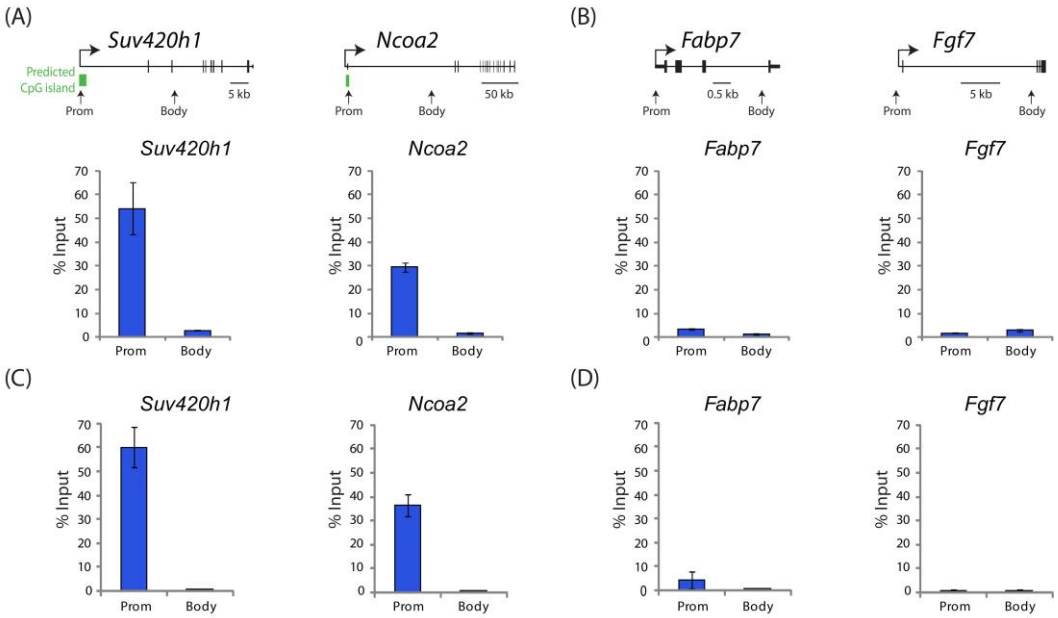


Figure 3.7 Bio-CAP can be performed on low quantities of DNA and can be adapted for magnetic bead-based automation

(A) Bio-CAP was performed using 100 ng of mouse V6.5 ES cell gDNA and the high-salt fractions (700 mM + 1 M) were interrogated by qPCR using promoter and body primer sets specific to two genes with CGI promoters, *Suv420h1* and *Ncoa2*. A schematic of the genes showing the position of the primer sets and CpG island predictions is illustrated above each data set. (B) As for (A) except that primer sets were specific to two genes with non-CGI promoters. (C) A Bio-CAP experiment was performed using NeutrAvidin-coated magnetic beads and the high-salt fraction (1 M) was interrogated by qPCR using promoter and body primer sets specific to two genes with CGI promoters. (D) As for (C) except that primer sets were specific to two genes with non-CGI promoters. All Bio-CAP data are from two biological replicates and error bars represent SEM.

3.4.2 Bio-CAP is amenable to automation

By comparison to the laborious handling time inherent to the CAP technique, Bio-CAP is a simpler, scaled-down protocol which is amenable to parallel processing and potentially automatable using

magnetic separation of samples. In order to test whether Bio-CAP could be adapted for use in a magnetic configuration, Bio-CAP was performed utilising NeutrAvidin coated magnetic particles. Using a streamlined approach, magnetic Bio-CAP was adapted to consist of only two steps: a repeated wash with CAP500 buffer and a repeated elution with CAP1000 buffer. In this format, Bio-CAP takes only 2 hours to perform, therefore rendering it a very rapid method for detection of non-methylated DNA (see 2.13.2.3).

In order to validate that the magnetic Bio-CAP approach is comparable to previous Bio-CAP results, the 1 M elution fractions were analysed by qPCR. Importantly, the magnetic Bio-CAP assay very closely recapitulated the results from the original Bio-CAP assay (compare Figure 3.7C and D with Figure 3.5G and I), with CGI promoter regions highly enriched in the high-salt elutions compared to body regions and non-CGI promoters. Therefore, using magnetic Bio-CAP, up to 16 samples can be processed in parallel using a magnetic centrifuge rack, or the procedure could be scaled up to an automated purification system, such as the Diagenode SX-8G IP-Star®.

3.5 **Bio-CAP coupled to massively parallel sequencing can profile non-methylated DNA genome-wide**

Locus-specific analysis of Bio-CAP elutions by qPCR validated that Bio-CAP is able to purify non-methylated DNA. To generate genome-wide profiles of non-methylated DNA, Bio-CAP high salt elution fractions were coupled to massively parallel sequencing. Bio-CAP was performed using 8 µg sonicated mouse testis genomic DNA, and the elutions were analysed by qPCR prior to sequencing. The combined 700 mM and 1 M elutions yielded around 100 ng of DNA by picogreen quantification, of which around 10 ng of DNA along with 10 ng of input material (sonicated genomic DNA) was used for library preparation and sequencing (performed at the Wellcome Trust Centre for Human Genetics).

Uniquely mapping reads from the sequencing data were aligned to the mouse genome (mm9) using Bowtie (see 2.14.4). Mapping and analysis of Bio-CAP-seq data followed a well-established ChIP-seq

pipeline which is simple compared to the complex analysis required for analysis of bisulfite sequencing data. A first browse across the genome revealed an enrichment of reads at predicted CGIs (for example *Dclre1b*, *Ap4b1m*, *Rsbn1* and *Phtf1*, Figure 3.8A), while no enrichment was observed at non-CGI promoters (see *Gm566* and *Ptpn22*, Figure 3.8A). This confirmed that Bio-CAP specifically purifies non-methylated DNA and does not simply enrich non-specifically for gene promoters genome-wide. Importantly, the sequencing trace for the matched input control for Bio-CAP shows a flat profile with no enrichment at CGIs (Figure 3.8A).

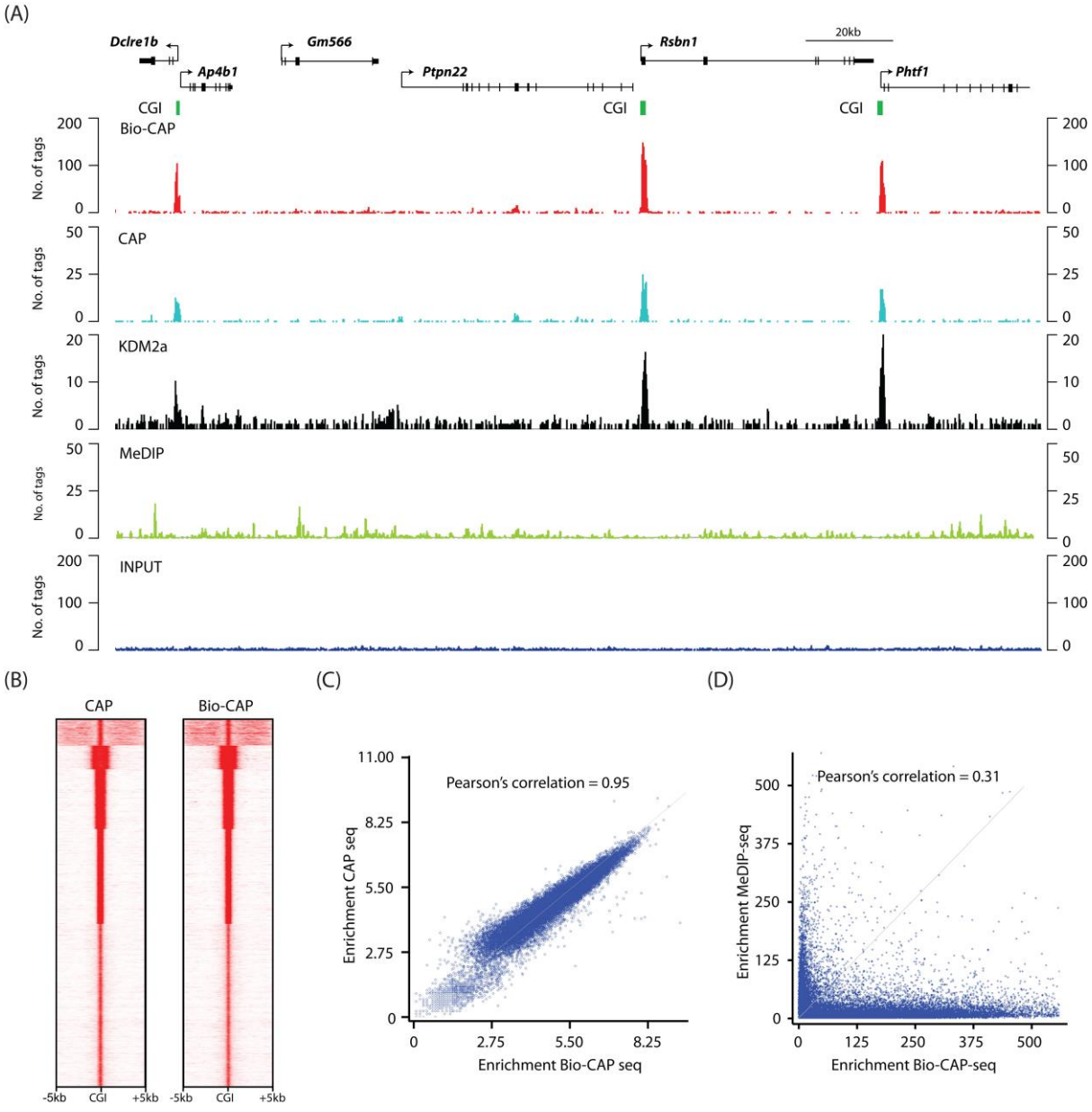


Figure 3.8 Bio-CAP coupled to massively parallel sequencing reveals non-methylated CGIs genome-wide

(A) The high-salt fraction (700 mM + 1 M) from a Bio-CAP experiment with mouse sperm genomic DNA was subjected to massively parallel sequencing and sequencing reads were aligned back to the mouse genome. Bio-CAP sequencing analysis is shown (top panel, red) for an ~180 kb region of chromosome 3 (chr3: 103 598 131–103 782 151), with comparative analysis from conventional CAP sequencing (second panel, light blue), KDM2A ChIP sequencing in mouse ES cells (third panel, black), MeDIP sequencing in mouse ES cells (fourth panel, green) and input material for Bio-CAP (bottom panel, dark blue). Annotated genes in this region are illustrated above the sequencing traces, with arrows indicating transcription start sites and CGIs from *in silico* predictions indicated in green. (B) Clustered heat map comparing CAP-seq and Bio-CAP-seq data sets centered at CGIs including a 5 kb window upstream and downstream. (C) Bio-CAP-seq enrichment plotted against CAP-seq enrichment for all CGIs genome wide. (D) Bio-CAP-seq enrichment plotted against MeDIP-seq enrichment for all CGIs genome wide.

The Bio-CAP non-methylated DNA profiles were compared to available sequencing data from the conventional CAP technique which was used previously to map non-methylated CGIs in mouse sperm (Illingworth et al., 2010). Again, by visualising contiguous regions of the genome (such as that shown in Figure 3.8A), it appears that Bio-CAP correlates strongly with sequencing data from conventional CAP. This correlation held true when all CGIs were analysed genome-wide (Figure 3.8B and C). Bio-CAP is therefore comparable to CAP for the detection of non-methylated DNA genome-wide. Furthermore, ChIP-seq data for KDM2A, a non-methylated DNA binding protein, from mouse ES cells shows a high degree of similarity with the Bio-CAP peaks (Figure 3.8A). Collectively these results show that Bio-CAP coupled to massively parallel sequencing can facilitate genome-wide profiling of non-methylated DNA.

3.5.1 Bio-CAP-seq specifically and efficiently identifies non-methylated DNA genome-wide

Affinity purification techniques have been developed to isolate the methylated fraction of the genome, including methylated DNA immunoprecipitation (MeDIP) and methylated DNA capture (MethylCap), and these methods have been coupled to massively parallel sequencing to visualise profiles of methylated DNA genome-wide (Bock et al., 2010; Brinkman et al., 2010; Down et al., 2008; Weber et al., 2005).

To further validate that Bio-CAP specifically purifies non-methylated DNA genome-wide, Bio-CAP sequencing profiles were generated for mouse ES cells and compared to sequencing traces for

MeDIP (MeDIP was performed by Dr Skirmantas Kriaucionis). Away from CGIs, MeDIP is subject to the same restrictions as Bio-CAP being limited by CpG density, therefore Bio-CAP and MeDIP datasets were compared at CGIs genome-wide. Importantly, no correlation was detected between the two genome-wide techniques, validating that Bio-CAP effectively identifies the non-methylated and not the methylated fractions of the genome. This observation can be visualised at a genomic locus containing a number of CGIs, where all CGIs clearly exhibit a non-methylated Bio-CAP signature but exhibit no MeDIP signal (Figure 3.8A). This comparison also highlights the difficulty visually identifying non-methylated regions of DNA from MeDIP-seq profiles. By comparison, these regions are immediately apparent from the Bio-CAP sequencing traces and can be rapidly identified genome-wide using peak-calling software such as MACS (2.14.5) without the need for the complex computation required for MethylCAP-seq (Down et al., 2008).

3.6 **Discussion**

This chapter demonstrates that Bio-CAP is a versatile and simple method for the purification of non-methylated fragments of DNA from bulk genomic DNA. Importantly, the improved CAP method is amenable to very small volumes of starting material and can be performed rapidly in batch, allowing for parallel processing of samples. Bio-CAP can be coupled to both quantitative PCR for locus-specific analysis or massively-parallel sequencing for genome-wide profiling of non-methylated DNA. Genome-wide sequencing demonstrates that the Bio-CAP technique is very efficient at purifying non-methylated DNA with very little contaminating background signal (Figure 3.8A), allowing even small Bio-CAP peaks of non-methylated DNA to be detected. Therefore, where single nucleotide resolution is not required, this modified CAP technique provides a simple and affordable alternative to whole genome bisulfite sequencing. As the Bio-CAP procedure affinity purifies fragments of non-methylated DNA, the utility of Bio-CAP extends beyond the simple detection of non-methylated DNA and a number of other useful downstream applications for Bio-CAP can be envisaged.

3.6.1 Improved mapping of regulatory regions in poorly annotated genomes or species without a reference genome

Non-methylated regions of DNA are associated with 70% of gene promoters and some enhancer elements in warm-blooded vertebrates (Bock et al., 2012; Deaton and Bird, 2011; Stadler et al., 2011; Xu et al., 2009). Therefore Bio-CAP-seq could be used in poorly annotated genomes of warm-blooded vertebrates to more accurately map promoter regions (and transcription start sites) genome-wide. Furthermore, for species with large and repetitive genomes for which genome sequences are not available, Bio-CAP-seq could be utilised to reduce initial sequencing efforts to identify the majority of functional elements, promoters and enhancers, genome-wide. *De-novo* scaffold assembly could then be used to assemble up to 70% of gene promoters and other regulatory elements. Analysis of these regulatory regions could then be performed by PCR (with primers designed in the sequenced regions) or full gene sequences cloned from cDNA using a forward primer at the start of the putative transcript and an oligo(dT) reverse primer. To capture more sequence either side of non-methylated regions, a shorter sonication time could be used to generate larger DNA fragments; this material could be used in a Bio-CAP experiment and post-sonicated prior to sequencing.

3.6.2 Chromatin Bio-CAP

There is great interest in identifying the chromatin modifications and protein factors which localise at gene promoters. Techniques to isolate specific regions of chromatin have been trialled including a recombination-based method to purify the *Pho5* promoter in yeast (Griesenbeck et al., 2003), purification of telomeric chromatin using tagged nucleic acid probes (PiCh) (Dejardin and Kingston, 2009) or knock-in of a LexA DNA sequence to allow immunoprecipitation using a LexA binding domain (iChIP or ChAP-MS) (Byrum et al., 2012; Hoshino and Fujii, 2009). Bio-CAP provides an opportunity to investigate the chromatin environment at non-methylated regions of the genome. Using a chromatin substrate instead of genomic DNA, the ZF-CxxC affinity module could be used to pull-down non-methylated CpG dinucleotides and thereby isolate CpG island chromatin through a

step-wise wash and elution procedure similar to conventional Bio-CAP. The chromatin associated proteins and post-translational modifications of histone tails could then be interrogated by mass spectrometry. This analysis may give insight into unknown factors which play a role at CGIs in initiating, maintaining or recognising the non-methylated status of DNA.

Purification of non-methylated regions of chromatin may prove more challenging than for genomic DNA due to increased non-specific interactions of chromatin with the NeutrAvidin beads or with the ZF-CxxC protein itself. Furthermore, non-methylated CpG dinucleotides may not be as accessible for binding by the immobilised ZF-CxxC affinity module due to being bound by endogenous ZF-CxxC containing proteins. Recent work has also revealed that ZF-CxxC domain binding to chromatin requires inter-nucleosomal DNA, likely due to the interaction of the ZF-CxxC domain with both faces of DNA (Zhou et al., 2012). As promoter regions and CGIs are thought to be nucleosome-depleted (Fenouil et al., 2012; Valouev et al., 2011), the requirement of the ZF-CxxC domain for linker DNA may not prove to be overly detrimental to CGI chromatin purification.

3.6.3 Detection of CGIs by fluorescent *in situ* hybridisation

There is great interest in visualising the architecture of the eukaryotic nucleus to better understand how chromosomes, nuclear compartments and transcription foci are organised (Schneider and Grosschedl, 2007). RNA-FISH and DNA-FISH can be used to highlight nascent transcription and the location of a gene of interest in the nucleus respectively (Bienko et al., 2013; van Raamsdonk and Tilghman, 2001). DNA-FISH probes can be made from any purified DNA samples, and hence the purified non-methylated DNA from the Bio-CAP protocol provides the material to potentially perform DNA-FISH with Bio-CAP probes, which would provide a unique insight into the positioning of non-methylated DNA in the nucleus.

Chapter 4 -- Non-methylated islands are a shared feature of vertebrate genomes

4.1 **Non-methylated DNA usage appears divergent between diverse vertebrate species**

4.1.1 Identification of HpaII tiny fragments

Vertebrate genomes are pervasively methylated on cytosine bases that exist in the context of a cytosine-guanosine (CpG) dinucleotide. An exception to this generality are contiguous regions of non-methylated DNA. These were first observed in chicken and mouse as a fraction of the genome which is susceptible to methylation sensitive endonuclease digestion (Bird et al., 1985; Cooper et al., 1983). Short fragments of DNA released by this digestion, named HpaII tiny fragments (HTFs), were associated with a number of gene promoters (Bird et al., 1985) suggesting a potential role in gene regulation. Initially it was suggested that HTFs were a feature of most vertebrate genomes (Cooper et al., 1983), however it appeared from subsequent analysis that these genomic elements were absent or scarce in many cold-blooded species including turtle, frog, trout, crocodile and carp (Aïssani and Bernardi, 1991). This suggested that hypomethylated gene promoters may be a unique feature of warm-blooded vertebrate genomes.

4.1.2 CpG island predictions are divergent amongst vertebrate species

Sequencing of a small number of HTFs revealed that they exhibit unique DNA sequence composition compared with other regions of the genome (Bird et al., 1985). In vertebrate genomes, CpG dinucleotides are depleted due to the mutagenic nature of methylated cytosine (Coulondre et al., 1978). However, HTFs exhibited only a minimal depletion of observed over expected CpG ratio (CpG obs/exp) and have a GC content which is often greater than other genomic regions (Bird et al., 1985). HTFs were therefore re-named CpG islands (CGIs) (Bird, 1986; Bird, 1987) and their unique nucleotide content prompted the development of a prediction algorithm to systematically identify regions of DNA with CGI-like properties (Gardiner-Garden, Frommer, 1987).

With the advent of whole-genome sequencing, reference sequences are now available for many species, allowing the location of CGIs to be predicted genome-wide. A CGI prediction track from the UCSC genome browser (Kent et al., 2002) allows easy visualisation of the location of CGIs at the genome-scale (Gardiner-Garden, Frommer, 1987). In agreement with initial work studying a limited number of HTFs, many human and mouse CGIs are promoter-associated (Figure 4.1A-B, seven representative genes). However, scarcity of promoter-associated HTFs in cold-blooded vertebrates (Aïssani and Bernardi, 1991) appears to be supported by a lack of CGI predictions at these same seven promoters in zebrafish (Figure 4.1A-B).

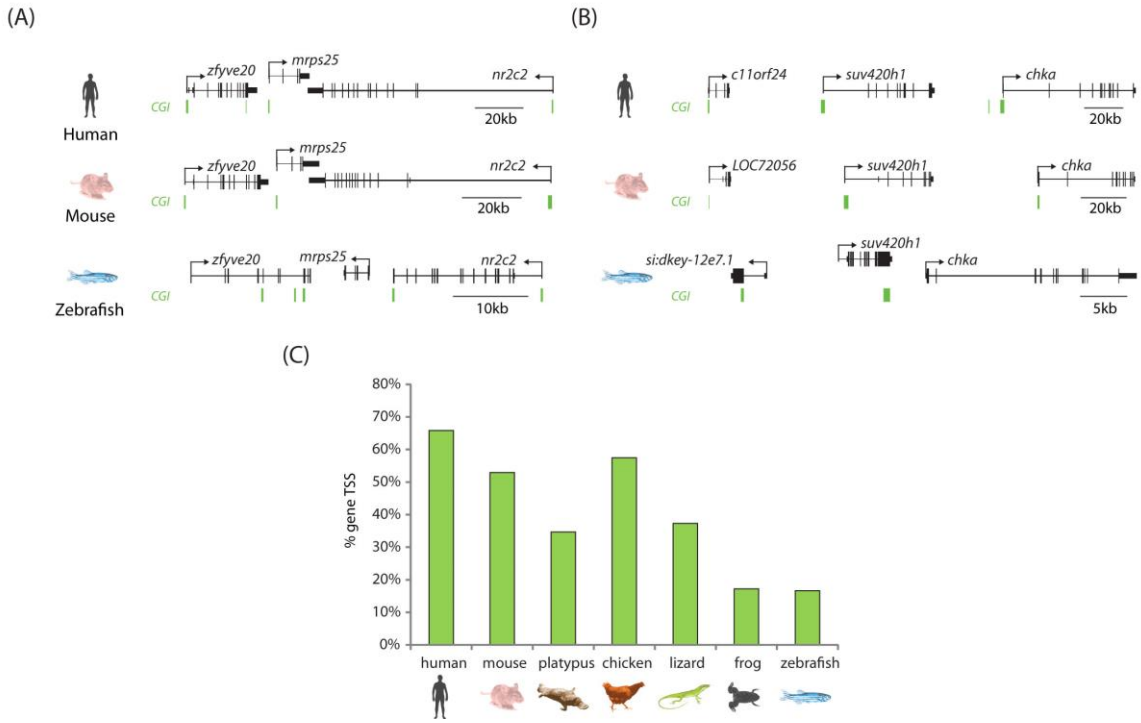


Figure 4.1 CpG island predictions are divergent in cold-blooded vertebrate species

(A and B) Two syntenic regions for human, mouse and zebrafish highlighting the location of CpG islands (green). (A) hg19 chr3:14,985,963-15,145,120; mm9 chr6:92,036,033-92,173,160; danRer7 chr8:55,338,683-55,387,640. (B) hg19 chr11:67,817,721-68,047,458; mm9 chr19:3,708,305-3,896,788; danRer7 chr18:20,881,484-20,933,026. (C) A histogram depicting the overlap of transcription start sites (TSSs; from ensGene, one TSS per gene) with predicted CpG islands (UCSC track) for seven diverse vertebrate species.

Genome-wide, the overlap of CGI-predictions with protein-coding gene transcription start sites (TSSs) from a number of vertebrate species, both warm and cold-blooded, further demonstrated that a lower proportion of gene TSSs in cold-blooded vertebrates (lizard, frog and zebrafish) are associated with predicted CGIs than in warm-blooded vertebrates (human, mouse and chicken)

(Figure 4.1C). Although platypus appears to be an exception to this rule, the poor overlap of TSSs and CGIs can almost certainly be attributed to the poor current genome annotation (the lowest quality of all species interrogated here). Together, these observations suggest that CGI usage in cold-blooded vertebrates may be divergent from warm-blooded vertebrates. Furthermore, it has been suggested that CGIs in cold-blooded vertebrates may be primordial in their evolution and that the emergence of CGIs in warm-blooded vertebrates may be linked to the evolution of the placenta in mammals (Sharif et al., 2010).

In this Chapter, the current dogma that CGIs in cold-blooded vertebrates are divergent will be explored.

4.2 **Recognition of non-methylated DNA appears to be conserved across vertebrates**

4.2.1 Conservation of ZF-CxxC domain-containing proteins across vertebrate species

In mammals, the ZF-CxxC proteins have been shown to play important functions at CGI regions including directing chromatin modifying activities to CGIs (Blackledge et al., 2010; Deaton and Bird, 2011; Long et al., 2013; Thomson et al., 2010). To better understand whether this function of non-methylated DNA is shared across vertebrates, the conservation of the ZF-CxxC proteins was investigated. Surprisingly, despite the apparent divergence in CGI usage, cold-blooded vertebrates appear to have retained at least one copy of most of the eleven ZF-CxxC proteins found in mouse, suggesting that the recognition and interpretation of CGIs is likely to be highly similar across these vertebrate species (Figure 4.2 and Appendix 4.1). This seemed counter-intuitive given that CGIs (nucleation sites for ZF-CxxC domains) are highly divergent in location in cold-blooded vertebrate genomes, and implies that ZF-CxxC proteins are nucleated at non-promoter regions in cold-blooded species (Figure 4.3).

Base composition of gene promoters between vertebrates is known to be highly variable (Aerts et al., 2004; van Heeringen et al., 2011). Considering that CGI algorithms were optimised for warm-blooded vertebrate species (Gardiner-Garden and Frommer, 1987; Takai and Jones, 2002), and that

non-methylated DNA binding proteins appear highly conserved across the vertebrates, it seemed feasible that CGI predictions may fail to accurately predict non-methylated DNA in cold-blooded vertebrates.

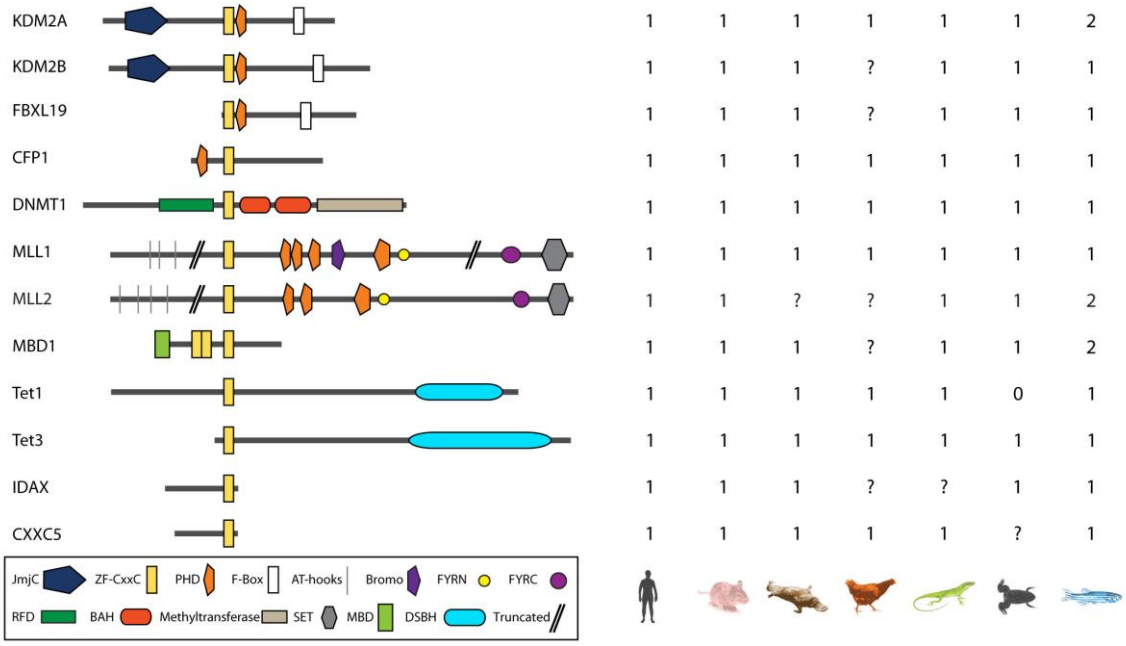


Figure 4.2 ZF-CxxC proteins are conserved across vertebrate species

Twelve ZF-CxxC proteins can be identified from the mouse genome and their domain architecture is depicted (left). Presence of gene orthologues in other vertebrates was analysed by searching the Ensembl and NCBI databases. See Appendix 4.1 for gene names. A "?" refers to lack of gene annotation; "0" means that there is a publication referring to the lack of gene in this species (Xu et al., 2012); 1 or 2 refers to the presence of one or two orthologues in that species.

4.2.2 Evidence for non-methylated DNA at orthologous vertebrate gene promoters

To begin more directly testing whether cold-blooded vertebrate gene promoters lack non-methylated DNA , a locus-specific approach was undertaken to experimentally determine the DNA methylation status of orthologous vertebrate gene promoters exhibiting a CGI prediction which was not shared across all species. The mouse *usp1* gene exhibits a promoter-associated CGI prediction which overlaps the annotated TSS, while all three cold-blooded vertebrate orthologues lack a CGI prediction (Figure 4.4A). As expected, bisulfite analysis in testis tissue for a region within the mouse *usp1* CGI demonstrated that this element is truly non-methylated (Figure 4.4Ai). Intriguingly, bisulfite analysis of a region proximal to the transcription start site for cold-blooded vertebrate *usp1* genes orthologues also revealed the presence of non-methylated DNA for lizard, frog and zebrafish (Figure

4.4Aii-iv).

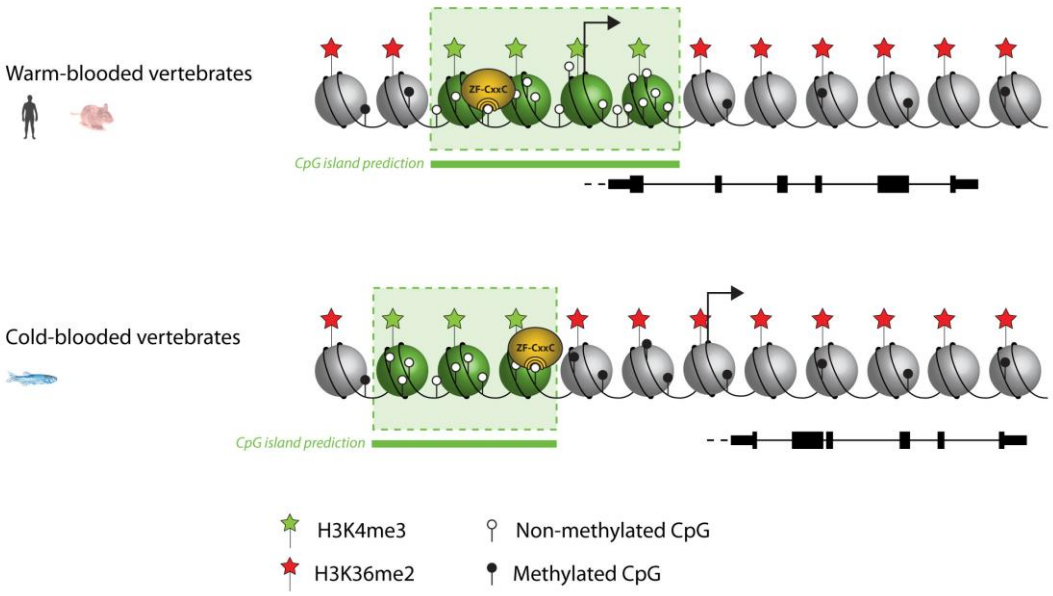


Figure 4.3 A model for CpG island usage in warm and cold-blooded vertebrates

A model depicting a typical gene from warm-blooded vertebrates (such as human and mouse, upper panel) and a cold-blooded vertebrate (such as zebrafish, lower panel). Up to 60-70% of warm-blooded vertebrate gene promoters are associated with a CpG island prediction (upper panel) while CpG island predictions in cold-blooded vertebrates rarely overlap with gene transcription start sites and are mostly found in intergenic regions. As ZF-CxxC proteins are conserved across the vertebrates, this suggests that in cold-blooded vertebrates that ZF-CxxC proteins are not targeted to gene promoter regions. Transcription start sites are depicted by an arrow, nucleosomes are shown as spheres with histone modifications indicated by stars. DNA methylation is indicated by white or black circles with CpG islands and genes models depicted below the DNA.

Unlike *usp1*, the *vbp1* gene does not have an associated CGI at its promoter in mouse (Figure 4.4Bi). However, bisulfite sequencing analysis adjacent to the *Vbp1* TSS of the mouse *vbp1* gene revealed that this gene is in fact non-methylated in mouse testis, despite lacking a CGI-prediction. Again, *vbp1* gene orthologues from cold-blooded vertebrate species also have promoter-associated non-methylated DNA despite the lack of a CGI prediction for frog and zebrafish genes. Together, this analysis indicates that CGI algorithms fail to identify all non-methylated regions in cold-blooded vertebrate species and, in some cases, those in warm-blooded vertebrates also.

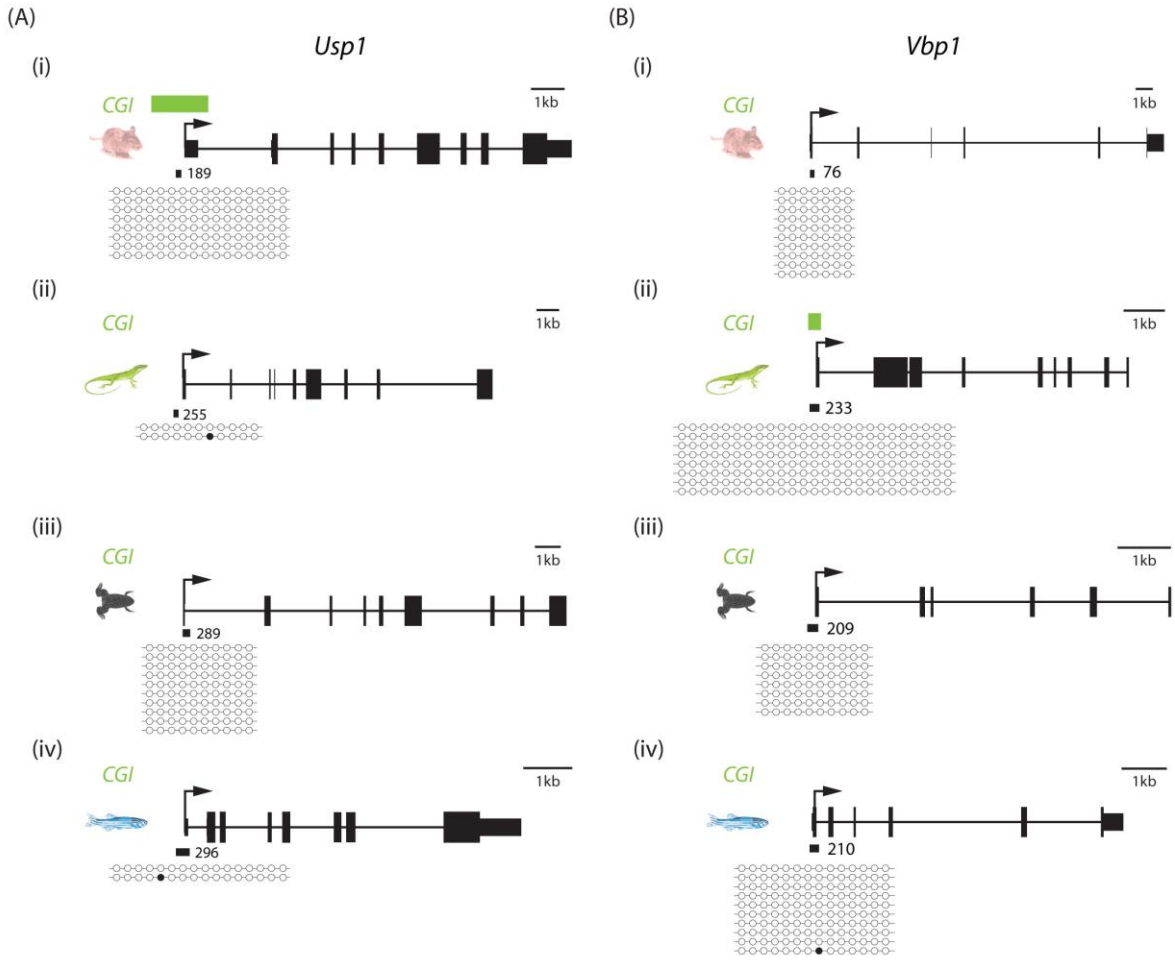


Figure 4.4 Two examples of non-methylated vertebrate gene promoters lacking CpG island predictions

(A) Bisulfite sequencing of the *Usp1* promoter region for mouse (i, mm9, chr4:98,589,851-98,590,039), lizard (ii, anoCar2, chr4:99,968,427-99,968,681), frog (iii, xenTro3, GL172642:3,000,812-3,001,100) and zebrafish (iv, danRer7, chr6:32,103,753-32,104,048). (B) Bisulfite sequencing of the *Vbp1* promoter region for mouse (i, mm9, chrX:72,759,660-72,759,735), lizard (ii, anoCar2, chrUn_GL343202:3,812,679-3,812,911), frog (iii, xenTro3, GL172646:513,380-513,588) and zebrafish (iv, danRer7, chr7:25,061,626-25,061,835). CpG island predictions are shown in green and genes are drawn schematically in black with an arrow representing the transcription start site. Black squares indicate the region interrogated by bisulfite sequencing and the number next to it indicates the length of the analysed region. DNA methylation status for a number of clones is shown below the gene (each line represents a different clone), with an open circle representing a non-methylated cytosine and a filled circle a methylated cytosine.

4.3 **Profiling non-methylated DNA genome-wide for seven vertebrate species**

The prevailing views concerning vertebrate non-methylated DNA elements (Figure 4.3) were called into question by the above experiments examining two specific gene loci. In order to gain a better understanding of the degree of conservation or divergence of non-methylated DNA usage, an experimental approach was undertaken to generate genome-wide maps of non-methylated DNA.

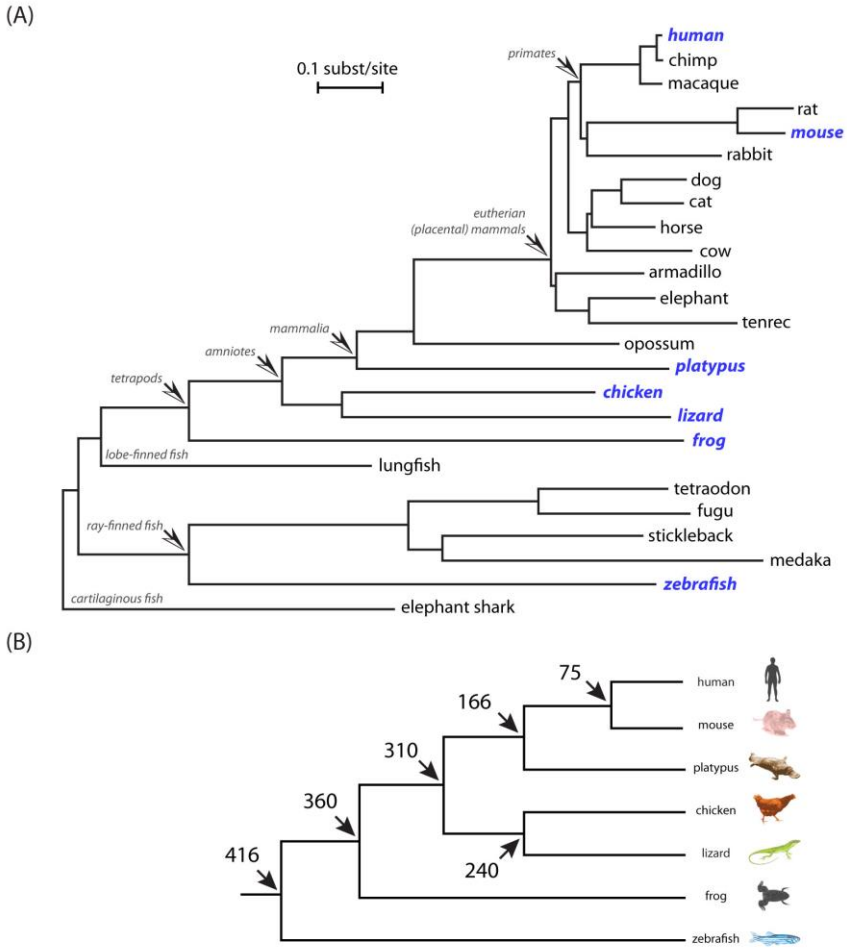


Figure 4.5 The vertebrate phylogenetic tree

(A) A vertebrate phylogenetic tree (adapted from (Amemiya et al., 2013; Miller et al., 2007)) highlighting species analysed by Bio-CAP (bold, italics, blue). (B) A minimal phylogenetic tree for the seven vertebrates analysed in this chapter. Estimated divergence times (million years ago) are indicated at branch-points.

Bio-CAP sequencing was performed for seven diverse vertebrate species encompassing all major branch points of vertebrate evolution, including both warm and cold-blooded species (Figure 4.5A-B). The vertebrate species chosen include two eutherian mammals (human—*Homo sapiens* and house mouse—*Mus musculus*), a monotreme (platypus—*Ornithorhynchus anatinus*), a bird (chicken—*Gallus gallus*), a lizard (green anole lizard—*Anolis carolinensis*), a frog (African clawed frog—*Xenopus tropicalis*), and a teleost fish (zebrafish—*Danio rerio*). Each Bio-CAP experiment was performed in duplicate and validated prior to sequencing by locus-specific qPCR analysis (for sequencing statistics, see Appendix 4.2). Bio-CAP sequencing was initially performed on testis tissue as it is mostly composed of one cell-type, spermatocytes. Furthermore, CGIs are proposed to be defined in the germline where they resist CpG depletion, necessary for their survival over

evolutionary time (Antequera and Bird, 1993; Antequera and Bird, 1999; Saxonov et al., 2006). Therefore, the testis represents an interesting tissue for analysing the genome-wide distribution and evolution of CGIs.

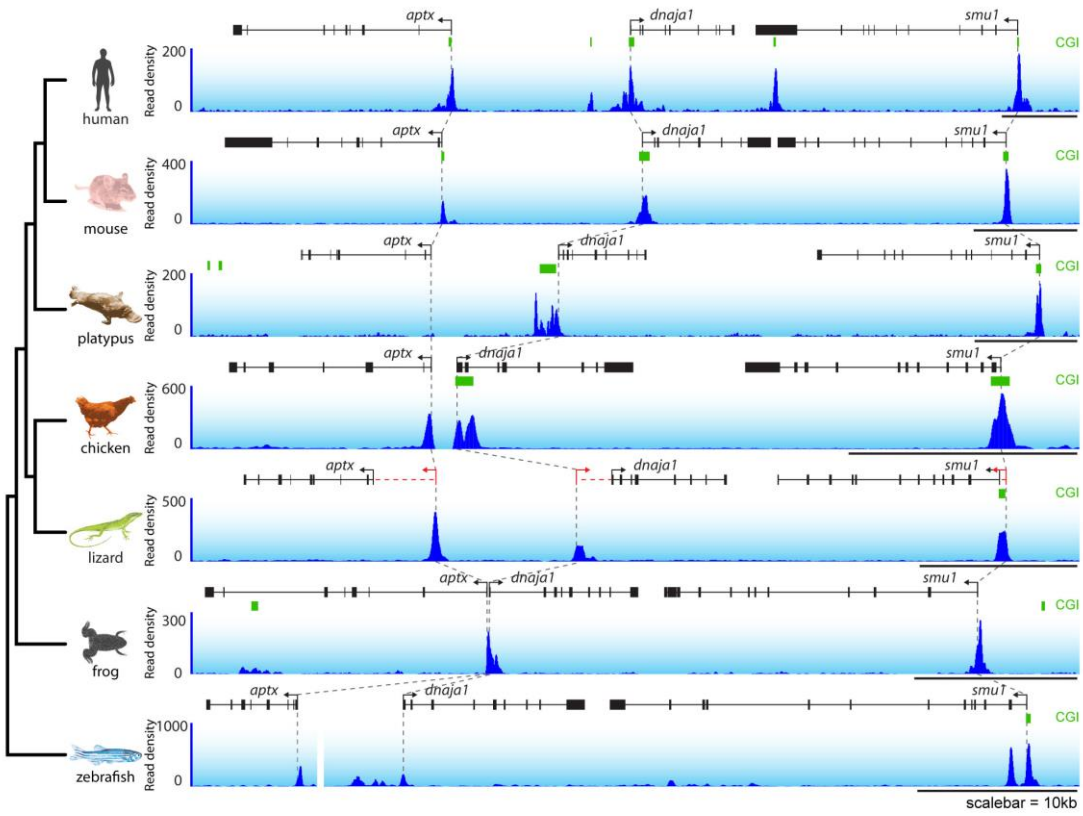


Figure 4.6 Non-methylated islands are a common feature of vertebrate gene promoters

Non-methylated DNA profiles in testis at three representative syntenic regions for seven vertebrate species. Genes are shown in black (improved annotation of gene TSSs using RNA-seq data is shown in red), CpG island predictions in green (CGI), and non-methylated DNA profiles are shown in blue. A phylogenetic tree (left) highlights the evolutionary relationship among the seven species. Dashed grey lines highlight the relationship between the gene TSSs across the species. A gap in the zebrafish profile indicates that *aptx* is found at a separate locus from *dnaja1* and *smu1*.

4.4 Bio-CAP profiles appear highly similar between vertebrates in contrast to CGI predictions

4.4.1 Profiles of non-methylated DNA for seven vertebrates

Initial visualisation of the Bio-CAP sequencing data at a syntenic locus revealed that punctate peaks of non-methylated DNA were coincident with the majority of gene TSSs for the warm-blooded vertebrates human, mouse, platypus and chicken, as would be expected from CGI predictions (Figure 4.6). Strikingly, for the cold-blooded vertebrates it also appeared that peaks of non-methylated DNA

were centred at gene transcription start sites, despite CGI predictions at this syntenic locus being almost entirely divergent from that at the warm-blooded vertebrate genes (Figure 4.6). This potentially surprising observation (given the distribution of CGI predictions at these loci) is in fact consistent with initial observations from locus-specific analyses (Figure 4.4) and preliminary observations made by others (Cross et al., 1991; Feng et al., 2010; Zemach et al., 2010).

4.4.2 Identification of a robust set of Bio-CAP sequencing peaks

Bio-CAP-sequencing profiles exhibited high levels of signal to noise (Figure 4.6 and Appendix 4.3), which allowed high confidence peak calling with a threshold of 6-fold above background using MACS (Zhang et al., 2008) (see 2.14.5). This allowed the identification of a robust set of 11,099 – 40,871 non-methylated Bio-CAP peaks for each of the seven vertebrates in testis tissue (see Appendix 4.4). Importantly, a number of non-methylated peaks were validated by bisulfite sequencing (Appendix 4.6).

4.4.3 CpG island prediction algorithms perform poorly in cold-blooded vertebrates

Visualisation of Bio-CAP sequencing data revealed that non-methylated regions of DNA appear to be highly localised to vertebrate gene promoters (Figure 4.6). This is contrary to CGI predictions which appear to rarely overlap gene TSSs in cold-blooded vertebrates (Figure 4.1A-C) and, from inspection of a number of loci, are more commonly found in gene bodies or intergenic regions in these species (Figure 4.1A-B, Figure 4.6 and Appendix 4.3).

To ascertain the genome-wide relationship between predicted CGIs and Bio-CAP-seq peaks, all CGIs (Kent et al., 2002) and all Bio-CAP peaks were compared (see 2.15.1). For human and mouse, the overlap of predicted CGIs with Bio-CAP non-methylated DNA was good, with 73 % and 87 % of CGIs overlapping Bio-CAP peaks respectively (Figure 4.7A). However, for cold-blooded vertebrates lizard, frog and zebrafish, the overlap of CGI predictions and Bio-CAP-defined non-methylated regions was extremely poor (22-44 % of CGIs overlapped with a Bio-CAP peak in testis), confirming that CGI predictions in these species largely fail to identify regions of non-methylated DNA (Figure 4.7A and

Appendix 4.5). Furthermore, for all species, there were a large number of Bio-CAP peaks which were not identified by CGI predictions (Figure 4.7A). From this comparative analysis it is clear that CGI predictions perform poorly in all species examined here for the identification of non-methylated DNA in testis tissue. Therefore, hereafter non-methylated regions identified by Bio-CAP sequencing will be referred to as non-methylated islands (NMIs) to reflect that these regions were identified by experimental profiling of non-methylated DNA, and are distinct from the CGI elements identified by computational prediction.

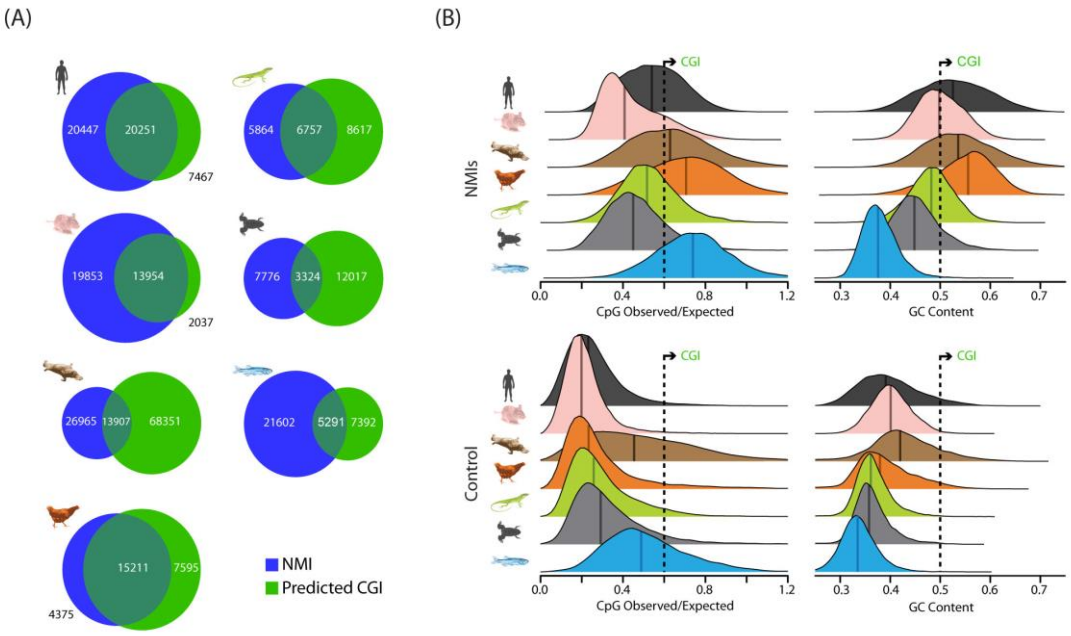


Figure 4.7 CpG island predictions do not accurately identify non-methylated islands of DNA in vertebrate genomes

(A) The genome-wide overlap between CpG islands (green) and non-methylated islands (blue) is depicted as a Venn diagram for each of the species. (B) Nucleotide properties of non-methylated islands and control regions are depicted as density plots. CpG observed/expected (left) and GC content (right) are shown for NMIs and control regions (10 kb upstream of each NMI) of the genome. Median values are shown as dark vertical lines. Thresholds for CpG island prediction are indicated (black dashed line).

4.4.4 Nucleotide properties of genomes are variable between vertebrate species

To determine the reason why CGI prediction algorithms fail to accurately predict non-methylated DNA in vertebrate genomes, the two nucleotide properties used to predict CGIs were interrogated at all testis NMIs. As a control, an equivalent sized region 10 kb region upstream was also analysed. CpG obs/exp and GC content were highly variable for NMIs from different species (Figure 4.7B, upper). For the warm-blooded vertebrates (human, mouse, platypus and chicken), the majority of

non-methylated regions exhibited CpG obs/exp and GC content above the cutoff for CGI prediction (indicated at 0.6 and 0.5 respectively). Conversely, for lizard and frog the majority of NMIs exhibited CpG obs/exp and GC% below the CGI requirements, which may explain the poor prediction of these regions as CGIs. For zebrafish, the majority of NMIs were above the CpG obs/exp cutoff of 0.6, suggesting that these regions should be accurately predicted. However, the GC content was extremely low, as has been observed previously (Cross et al., 1991), explaining why these regions were not predicted despite being CpG dense (Figure 4.7B, upper).

4.5 **Bio-CAP reveals a common strategy of non-methylated DNA usage at vertebrate gene promoters**

In mammalian genomes, previous estimates have suggested that non-methylated regions are a feature of 50-70% of gene promoters (Antequera, 2003; Antequera and Bird, 1993). Intriguingly, visual inspection of three syntenic regions suggested that NMIs are strongly associated with gene promoters in all vertebrate species (Figure 4.6 and Appendix 4.3). In order to determine whether this holds true genome-wide, the proportion of gene TSSs (+/- 1kb, 2.15.2) overlapping an NMI was determined for all seven species. Unsurprisingly, around 70% of human and mouse gene TSSs were associated with a NMI (Figure 4.8A) and similarly chicken, another warm-blooded species, exhibited a large proportion of non-methylated gene promoters. Strikingly, contrary to CGI predictions, around 60% of zebrafish gene TSSs were overlapped by an NMI. The same association appeared initially low for the remaining two cold-blooded species, lizard and frog; however these two species have two of the most poorly annotated genomes. Incorporation of external RNA-seq data (Barbosa-Morais et al., 2012) and XTev gene models for frog (Akkers et al., 2010) increased the proportion of NMI-associated gene promoters to 50-60 % (Figure 4.8A) by improving the annotation of gene TSSs (Figure 4.6, red). A slight improvement can also be observed for platypus and chicken following incorporation of RNA-seq data (Brawand et al., 2011; Julien et al., 2012). Gene annotation however remained poor for platypus and future improvements in the annotation of this species will no doubt reveal that a high proportion of gene promoters are non-methylated.

Importantly, profiles of non-methylated DNA appear highly similar at gene promoters between the seven vertebrate species analysed here, with the peak of non-methylated DNA consistently overlapping the gene TSS (Figure 4.8B). Therefore the size, and potentially the regulatory importance, of these non-methylated elements appeared to be highly similar across all seven vertebrates. Together, these data conclusively show that NMIs are a shared property of vertebrate gene promoters, contrary to previous suggestions that promoter-associated CGIs are a unique feature of warm-blooded vertebrate genomes.

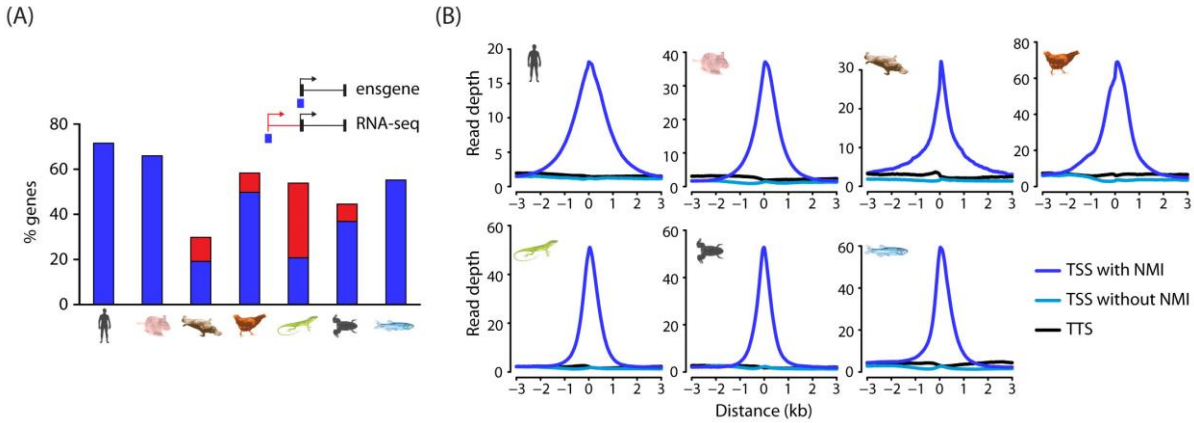


Figure 4.8 Non-methylated islands are a common feature of vertebrate gene transcription start sites

(A) A histogram depicting the proportion of protein-coding transcription start sites (TSSs) which are overlapped by an NMI for all seven species. Blue bars indicate overlap with annotated TSSs and red bars indicate overlap with additional TSSs identified using RNA-seq data (platypus, chicken and lizard) or Xtev gene sets (frog). (B) Profiles of non-methylated DNA were plotted over a 6kb window centred on all TSSs with an NMI (dark blue), without an NMI (blue), and for all transcription termination sites (black). The non-methylated DNA signal peaks at the TSS of gene promoters in all vertebrates.

4.6 **Non-methylated islands are conserved at orthologous gene promoters**

From closer inspection of the syntenic loci, it appeared that NMIs tend to be shared at orthologous gene promoters across all seven vertebrate species (Figure 4.6). In order to investigate the possibility that NMIs may be conserved at orthologous vertebrate genes, gene orthologues (with only one paralogue in each species) were identified for all seven vertebrates in a pairwise manner. Of note, these comparisons were constrained by the least-well annotated genome to control for quality of genome annotation (see 2.15.4). Pairwise comparisons revealed a striking tendency for homologous vertebrate genes to have an NMI at their respective gene promoters (Figure 4.9A). This is highlighted

by the comparison of human gene orthologues with any of the other six vertebrates, which revealed 93-98 % conservation of promoter NMIs (Figure 4.9A). Furthermore, a more extensive three-way comparison of human, mouse and zebrafish gene orthologues revealed conservation of NMIs at the majority of one-to-one orthologous gene promoters, despite 450 million years divergence between these species (Figure 4.9B). Collectively, this remarkable degree of conservation reveals that promoter-associated NMIs appear to be maintained across evolutionary time potentially as part of the mechanisms regulating vertebrate gene expression.

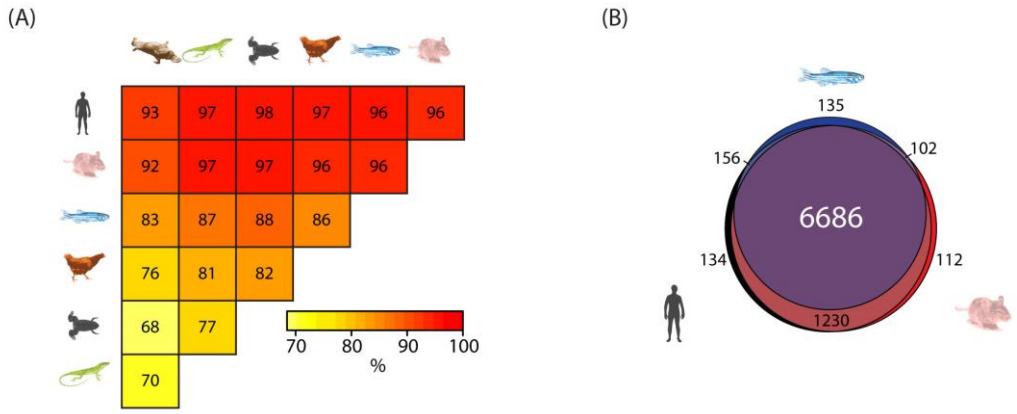


Figure 4.9 Non-methylated islands are a highly conserved epigenetic feature of vertebrate gene promoters

(A) The presence of NMIs at orthologous gene TSSs is illustrated by a pairwise analysis of NMIs at vertebrate gene orthologues. The percentage of NMIs conserved at orthologous gene TSSs was calculated in a pairwise manner and found to be highly statistically significant for all comparisons across the seven vertebrate species ($p<10^{-10}$, hypergeometric test). (B) A proportional Venn diagram illustrating the three-way comparison of NMI presence at conserved human-mouse-zebrafish gene orthologue TSSs.

4.7 **A novel broad class of non-methylated island**

4.7.1 Identification of genes associated with broad regions of non-methylated DNA

On average, non-methylated DNA appeared to peak at gene TSSs and spread symmetrically 1-2 kb in either direction (Figure 4.8B). Punctate NMIs, such as those seen at the promoters of the *tbca* and *wdr41* genes (Appendix 4.3A), most closely conform to the original description of a CGI (Bird et al., 1985; Bird, 1986) and hence can be referred to as 'canonical' NMIs. The *otp* gene, on the other hand, exhibited a distinct profile of non-methylated DNA, where the non-methylated DNA region reaches across the gene body and a substantial distance upstream and downstream of the locus, encompassing around 20 kb in total (Appendix 4.3A). This 'broad' type of NMI was observed for

many other genes, including *sp9* and *irx5* from human to zebrafish (Figure 4.10A-B). Strikingly, the *hox* loci appeared highly hypomethylated across the entire cluster in testis tissue for all seven species, covering greater than a 50 kb region of DNA on average (Figure 4.10C).

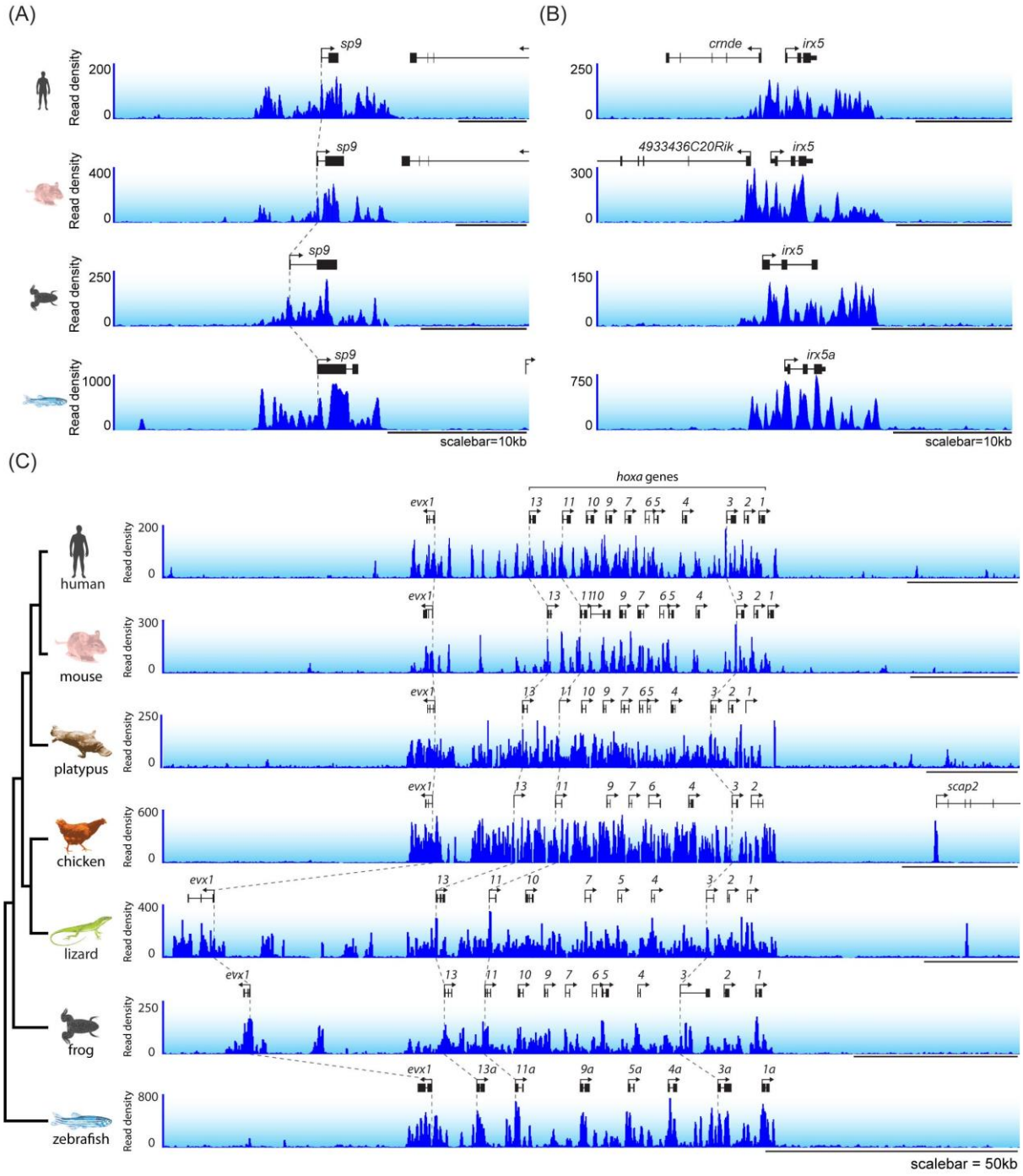


Figure 4.10 A novel class of broad gene-associated non-methylated islands

(A-B) Two examples of a broad region of non-methylated DNA associated with the *sp9* and *irx5* gene for four representative species (human, mouse, frog and fish). Dashed grey lines highlight the location of the gene TSSs across the four species. (C) The *hoxa* gene cluster from all seven vertebrate species is associated with broad regions of non-methylated DNA. Genes are shown in black and non-methylated DNA profiles

are shown in blue and dashed grey lines highlight the relationship between conserved gene TSSs across the species.

4.7.2 Broad NMIs are found at a substantial number of genes and are modified by H3K4me3

This novel broad class of NMI were easy to identify by eye, but initially proved difficult to detect bioinformatically as broad non-methylated regions were often detected as multiple smaller NMIs. Since the vast majority of broad NMIs appeared to be associated with genes and encompass the entire gene locus, broad NMIs were defined as covering greater than 90 % of an associated gene. A control group of short canonical NMIs were required to cover less than 10 %, but greater than 0 %, of the associated gene (see 2.15.5). To investigate the function of broad NMIs, non-methylated DNA was profiled for mouse embryonic stem cells (ES cells), frog embryonic stage 11-12 and zebrafish 24 hours post fertilisation (hpf), for which genome-wide chromatin modification datasets were available (see 2.17.1). Profiling non-methylated DNA over all identified broad NMI genes confirmed that Bio-CAP signal covers the full length of the gene, while Bio-CAP signal was confined to a punctate peak at the 5' end of the associated gene for the control set of short NMI genes (Figure 4.11A).

Canonical promoter-associated NMIs are known to be modified by H3K4me3 (Guenther et al., 2005; Thomson et al., 2010), therefore available genome-wide H3K4me3 ChIP-sequencing data was interrogated to observe whether broad NMIs were similarly modified. Strikingly, H3K4me3 was observed to be enriched across the entire gene body for broad NMIs (Figure 4.11B) closely mirroring the underlying non-methylated DNA (Figure 4.11A). Broad regions of non-methylated DNA therefore appear to direct the placement of the H3K4me3 modification across the entire gene length, suggesting that broad NMIs may directly impact the regulation of an associated gene by embedding it within an active chromatin domain (Guenther et al., 2007; Hon et al., 2009).

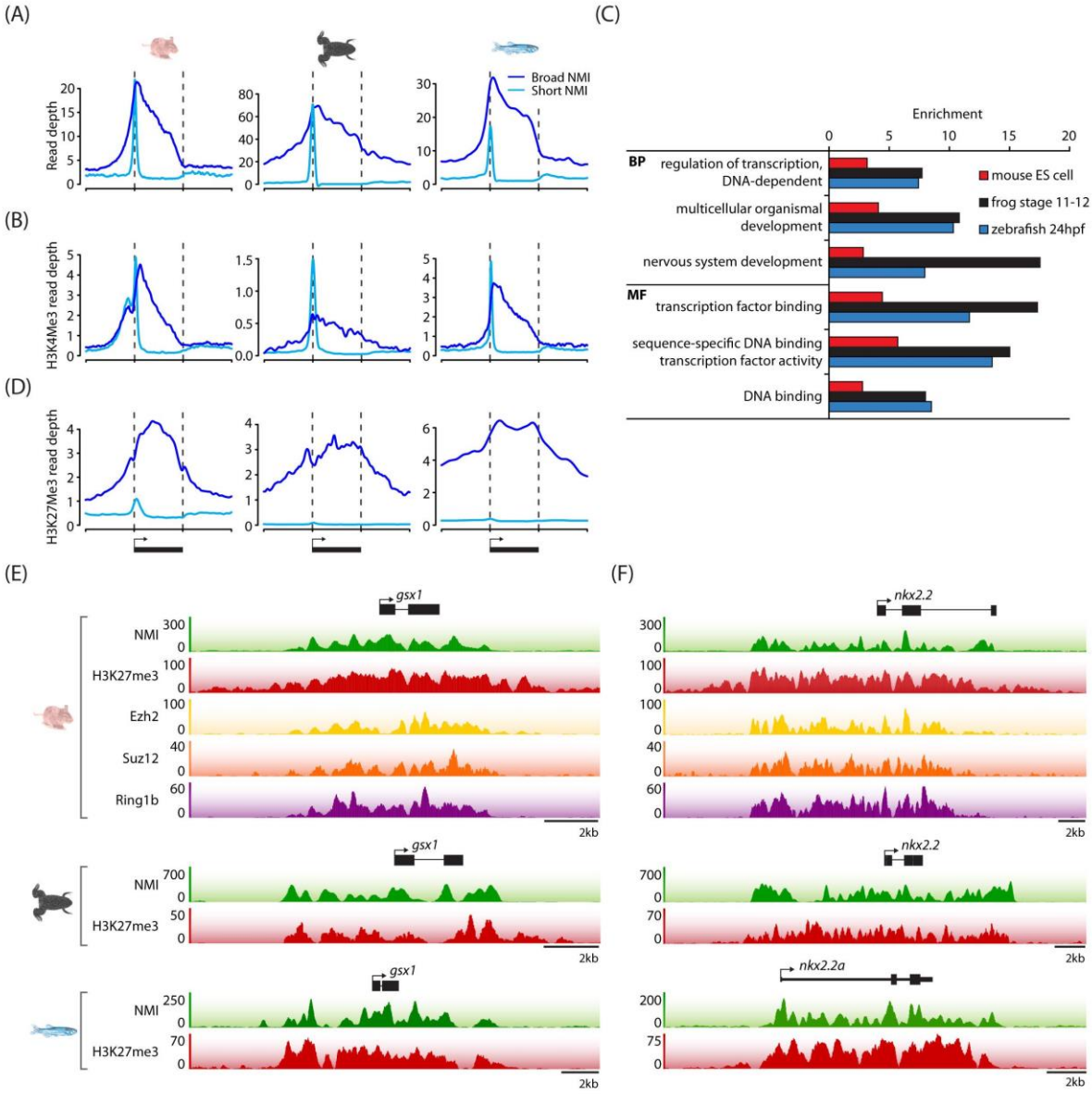


Figure 4.11 Broad non-methylated islands encompass polycomb-regulated developmental genes

(A) Non-methylated DNA profiles are depicted for genes associated with broad NMIs (dark blue) and canonical NMIs (light blue) in mouse embryonic stem (ES) cells, frog stage 11-12 and zebrafish 24 hpf. The profile is scaled to show an averaged gene with one gene length depicted upstream and downstream. 712, 160 and 350 broad NMIs were detected for mouse, frog and zebrafish respectively. (B) H3K4me3 ChIP-seq signal from mouse, frog and zebrafish was plotted as in (B). H3K4me3 profiles reflect the underlying non-methylated DNA profiles. (C) Genes associated with broad NMIs were analysed by Gene Ontology (GO) analysis for mouse ES cell, frog stage 11-12 and zebrafish 24 hpf. Broad NMIs are found to be significantly enriched for GO term categories associated with sequence-specific DNA binding, transcriptional regulation and development. MF = molecular function; BP = biological process. $p < 10\text{-}5$ for all GO terms. (D) H3K27me3 ChIP-seq signal from mouse, frog and zebrafish was plotted for the same gene sets as in (B). The profile is scaled to show an averaged gene with three gene lengths depicted upstream and downstream. As for H3K4me3, H3K27me3 ChIP-seq profiles correspond to the underlying non-methylated DNA profile. (E-F) Representative examples of two broadly non-methylated genes *gsx1* and *nkx2.2* for mouse, frog and zebrafish. In all three species, the broad non-methylated regions (green) are associated with the polycomb repressive mark H3K27me3 (red). In addition, in mouse, polycomb repressive complex 2 (ezh2, yellow and suz12, orange) and polycomb repressive complex 1 (ring1b, purple) components are

associated with the broad non-methylated regions. The y-axis depicts read density. Genes are depicted above the profiles in black.

4.7.3 Genes associated with broad NMIs are enriched for developmentally regulated transcription factors

To determine whether broad NMIs are associated with a specific class of genes, gene ontology (GO) term analysis was performed for broad NMIs genes. Interestingly, broad NMIs were enriched at genes associated with organismal development, DNA binding and regulation of transcription (Figure 4.11C), genes typically associated with polycomb-mediated repression during development (Schwartz and Pirrotta, 2007). Polycomb repressive complexes constitute a highly conserved mechanism for gene repression across metazoans (Morey and Helin, 2010), and were first identified in *Drosophila melanogaster* as important for early embryonic patterning and regulation of *Hox* genes (Jurgens, 1985; Lewis, 1978; Schwartz and Pirrotta, 2007).

Polycomb repressive complex 2 (PRC2) catalyses the placement of the repressive histone modification H3K27me3 via the protein subunit EZH2 (Cao et al., 2002; Czermin et al., 2002; Kuzmichev et al., 2002; Müller et al., 2002). Similar to H3K4me3, the H3K27me3 histone modification was found to be highly enriched over the entire gene body of broad NMI-associated genes suggesting that non-methylated DNA may nucleate polycomb repression (Figure 4.11D). Indeed, 45% and 89% of broad NMIs were associated with H3K27me3 in mouse ES cells and frog embryonic stage 11–12 tissues respectively, a significantly higher proportion than observed for canonical NMIs (Fisher's exact test, odds ratio > 5.3, $p<10^{-35}$). This observation is in agreement with recent literature which describes that clustered CGI predictions are often a good predictor of polycomb-mediated repression (Orlando et al., 2012).

Interestingly, the profile of H3K27me3 across broad NMI genes appeared quite distinct from H3K4me3, as the H3K27me3 modification appeared to spread beyond three gene lengths upstream and downstream of the gene transcription start and end site (Figure 4.11D). This observation held true at individual genes (Figure 4.11E-F), supporting a model of propagation of polycomb marks

away from sites of nucleation (Margueron et al., 2009). Interestingly it appeared that the polycomb proteins themselves may somehow be influenced by the broad non-methylated region as the distribution of the polycomb proteins EZH2, SUZ12 and RING1B in mouse ES cells appeared to mirror the non-methylated signal at the *gsx1* and *nkx2-2* genes (Figure 4.11E-F). Therefore broad NMIs may act to nucleate polycomb proteins during early development to mediate transcriptional repression or poise genes for later activation. In keeping with this suggestion, a mechanistic link has recently been described between non-methylated DNA and the recruitment of polycomb proteins, whereby KDM2B (a ZF-CxxC domain-containing protein) has been identified as a component of a polycomb repressive complex, PRC1 (Farcas et al., 2012; He et al., 2013; Wu et al., 2013). Therefore, in this variant PRC1 complex, the KDM2B ZF-CxxC may be acting to direct polycomb repression to non-methylated regions of the genome.

4.8 **Discussion**

4.8.1 Non-methylated islands of DNA are a shared feature of vertebrate gene promoters

Although the DNA methylation system is highly conserved amongst vertebrates, reliance on CGI maps previously indicated that this epigenetic system had significantly diverged between vertebrates, and even acquired unique properties during mammalian evolution (Aïssani and Bernardi, 1991; Sharif et al., 2010). Recent low-resolution studies have suggested that non-methylated DNA profiles may be more conserved than previously realised (Feng et al., 2010; Zemach et al., 2010). However, a lack of experimentally-determined regions of non-methylated DNA outside of eutherian mammals has hindered our understanding of the usage of this epigenetic system amongst non-mammalian vertebrates.

To address these outstanding questions, high resolution genome-wide profiles of non-methylated DNA were generated for seven diverse vertebrate species. 'Canonical', punctate non-methylated islands of DNA (NMIs) were observed to be a shared feature of 50-70 % of vertebrate gene

promoters, overturning previously held views that CGIs are a unique feature of warm-blooded vertebrate promoters. Furthermore, it appears that promoter-associated NMIs have been subject to selective pressure over evolutionary time to be maintained at orthologous gene promoters as a required feature for the accurate regulation of many vertebrate genes.

4.8.2 Differences in body temperature may impact the ability to predict CpG islands in cold versus warm-blooded vertebrates

Given the high degree of similarity in NMI location at gene TSSs in all vertebrate genomes analysed here, it was intriguing that CGI prediction algorithms perform so poorly at identifying these regions. The apparent reduction in fidelity of the CGI algorithm at the cold- to warm-blooded vertebrate transition may be related to the fact that spontaneous deamination, the process by which CpGs are depleted in methylated genomes (Coulondre et al., 1978), is a temperature sensitive reaction (Lindahl, 1993; Yang et al., 2000). Different rates of CpG loss may therefore have occurred in cold- versus warm-blooded vertebrate lineages leading to different nucleotide composition of both the genome as a whole and the non-methylated island regions of that genome. In hindsight, given the variable nucleotide composition between vertebrate genomes (Figure 4.7B, lower), it is not surprising that sequence-based CGI prediction algorithms are unable to identify non-methylated regions of the genome for all vertebrate species.

4.8.3 A class of broad non-methylated island at developmentally regulated genes

In addition to 'canonical' NMIs, a novel class of broad NMI was identified which appears to be strongly associated with polycomb-repressed developmentally regulated genes. These broad regions proved challenging to identify using bioinformatic tools as broad regions were often made up of several Bio-CAP peaks. Broad NMIs were therefore identified based on the percentage overlap of a gene, and inevitably this method will enrich for shorter genes. In order to identify all broadly non-methylated regions, a more sophisticated approach will be required. Interestingly, this observation may also indicate that broadly non-methylated regions are in fact not contiguously non-methylated

but represent clusters of non-methylated islands, perhaps each associated with an individual feature (Branciamore et al., 2010).

In mouse ES cells, chromatin modifying activities are targeted to regions of non-methylated DNA to create a unique chromatin regulatory environment (Blackledge et al., 2010; Long et al., 2013; Thomson et al., 2010). Therefore, perhaps developmental genes associated with broad NMIs have a more complex regulatory regime, with a larger number of non-methylated alternative promoters or intragenic enhancers compared to canonical NMI genes. Alternatively, developmentally regulated genes may require the integration of more transcription factors and signaling pathways for the correct coordination of transcription. Subsequently, a broader region of non-methylated DNA may be required to accommodate more binding sites for these factors, or simply be a consequence of the DNA being inaccessible for DNA methylation. Overall, it appears that these broad non-methylated elements may contribute to the complex regulation and tight temporal control of developmental genes which is required for expression at the correct developmental stages and in a tissue-specific manner.

4.8.4 CpG island predictions methylated in testis may be non-methylated in another tissue

In this chapter, a large number of CGI predictions were shown to be methylated in testis tissue for all of the seven vertebrate species (Figure 4.7A). This highlights that the CGI prediction algorithm performs poorly at accurately identifying all NMIs in testis. However it is worth noting that DNA methylation is known to vary between different tissues and developmental stages (as will be discussed further in the following chapter). Therefore some of the CGI predictions which are not non-methylated in testis may in fact be non-methylated in other tissues. Importantly, this also highlights the fact that CGI predictions can never successfully predict non-methylated DNA for all cell lineages, as a different subset of regions will be non-methylated across cell-types. Therefore there is real utility in experimentally profiling non-methylated DNA for many different tissues or developmental stages (Ziller et al., 2013) in order to gain a more complete understanding of regions

in vertebrate genomes which can be non-methylated, many of which may fall below CGI prediction thresholds. Furthermore, in order to fully appreciate the non-methylated DNA landscape of a specific cell-type of interest, experimental profiling will always be required.

Chapter 5 -- Dynamic DNA methylation at distal regulatory elements

The belief that DNA methylation is a static modification and unvarying between cell types has recently been challenged. Many studies have now demonstrated that for human and mouse DNA is in fact differentially methylated during differentiation and between distinct tissue types (Farthing et al., 2008; Illingworth et al., 2010; Maunakea et al., 2010; Mohn et al., 2008; Straussman et al., 2009; Weber et al., 2007). It remains unclear, however, whether dynamic DNA methylation, the gain or loss of non-methylated DNA, is a shared epigenetic strategy for the regulation of genome function across vertebrate species.

5.1 **Profiling non-methylated islands in liver tissue for seven vertebrate species**

5.1.1 Bio-CAP profiles of non-methylated DNA for liver are similar to those from testis

To examine whether the location of non-methylated DNA varies between tissues in all vertebrates, genome-wide non-methylated DNA profiles were generated by Bio-CAP-seq for liver tissue for human, mouse, platypus, chicken, lizard, frog and zebrafish and compared with the testis Bio-CAP-seq data generated in the previous chapter. Liver tissue was chosen as it is a fully differentiated, relatively homogenous tissue composed mainly of hepatocytes (Odom et al., 2006), and there are also publically available genome-wide gene expression and chromatin modification datasets for several of the species studied here (Barbosa-Morais et al., 2012; Bernstein et al., 2010; Brawand et al., 2011; Smagulova et al., 2011). Visualisation of the non-methylated DNA profiles generated for liver Bio-CAP-seq revealed that, like testis, strong punctate peaks of non-methylated DNA usually coincide with gene promoters (Figure 5.1). Indeed, a very similar proportion of protein-coding TSSs were associated with NMIs in liver as were observed testis, with between 50-70 % of gene TSSs having an associated NMI depending on the species (compare Figure 4.8A and Figure 5.2A), suggesting that the utilisation of non-methylated DNA at gene TSSs is a shared feature across vertebrate tissues.

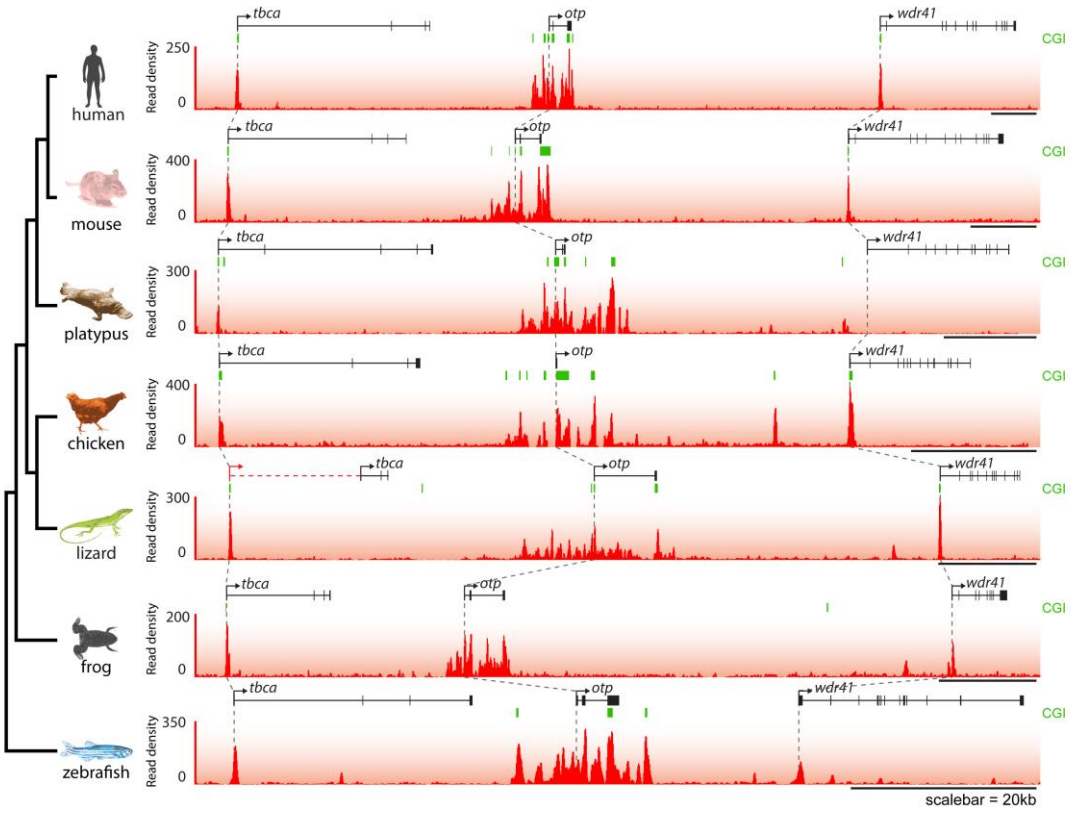


Figure 5.1 Profiles of non-methylated DNA for liver tissue

Non-methylated DNA profiles at a syntenic locus containing three genes *tbca*, *otp* and *wdr41* for seven vertebrate species. Genes are shown in black (improved annotation of gene TSSs using RNA-seq data is shown in red), CpG island predictions in green (CGI), and non-methylated DNA profiles are shown in red. A phylogenetic tree (left) highlights the evolutionary relationship among the seven species. Dashed grey lines highlight the relationship between the gene TSSs across the species. The equivalent sequencing traces for testis tissue can be seen in Appendix 4.3A.

5.1.2 Substantial changes in non-methylated DNA occur between tissues

To investigate whether changes in DNA methylation are detected between testis and liver tissues, peaks were identified for liver Bio-CAP sequencing using MACS (see 2.14.5) (Zhang et al., 2008) and compared to peaks previously identified for testis (see Appendix 4.4 for numbers of identified Bio-CAP peaks). Interestingly, although a large number of NMIs were shared between tissues, a considerable number of tissue-specific NMIs were detected for all seven species (Figure 5.2B). Intriguingly, there appeared to be a shift from a majority of unique NMIs being testis-specific for human and mouse to a larger proportion being liver-specific in lizard, frog and zebrafish (Figure 5.2B). The relevance of this is unclear; nevertheless it appears that changes in DNA methylation occurring during lineage specification are highly species-specific. Importantly, in agreement with

observations from mammalian species (Farthing et al., 2008; Illingworth et al., 2010; Maunakea et al., 2010; Mohn et al., 2008; Straussman et al., 2009; Weber et al., 2007), vertebrates appear to utilise differential methylation to epigenetically regulate a subset of NMIs.

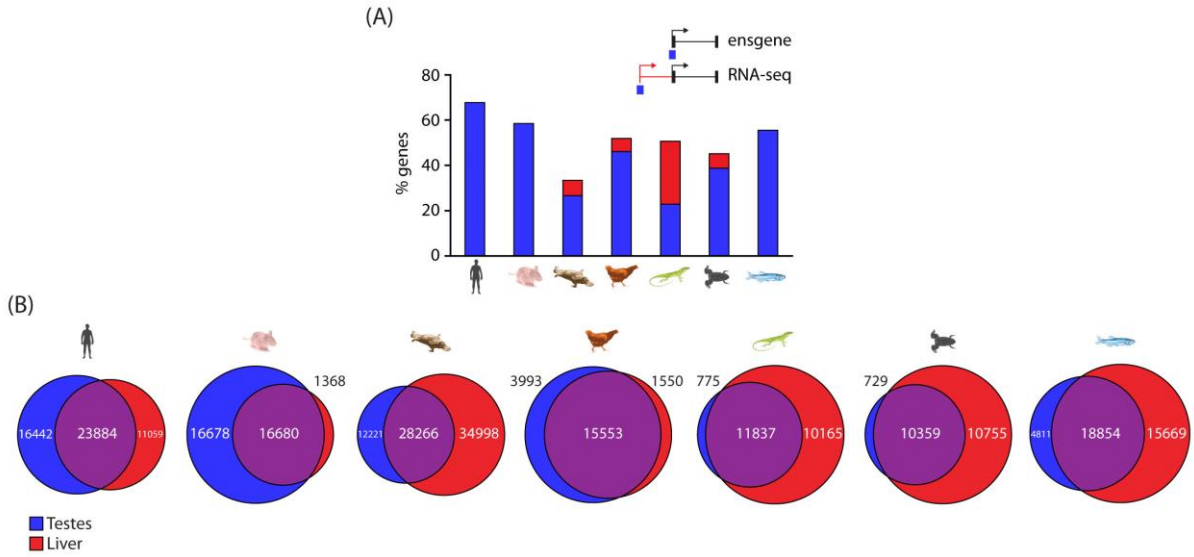


Figure 5.2 Non-methylated islands are associated with gene promoters in vertebrate liver

(A) A histogram depicting the proportion of protein-coding transcription start sites (TSSs) which are overlapped by an NMI in liver tissue for all seven species. Blue bars indicate overlap with annotated TSSs and red bars indicate overlap with additional TSSs identified using RNA-seq data (platypus, chicken and lizard) or Xtev gene sets (frog). (B) Venn diagrams depicting the overlap between Bio-CAP peaks identified in testis and liver tissues for seven vertebrate species.

To examine the properties of these differentially methylated NMIs more closely, heat maps were generated for a 10 kb window centred around each NMI interval, with NMIs clustered as shared or unique between tissues and then ranked by interval length (Figure 5.3A). Visualisation of all NMIs in this manner clearly demonstrated that shared NMIs were enriched for Bio-CAP-seq signal in both tissues, whereas unique NMIs from either testis or liver exhibited Bio-CAP-seq signal only in the appropriate tissue (Figure 5.3A). Therefore, non-methylated DNA profiles from testis and liver revealed a class of 'plastic' NMIs that exhibit differential methylation in distinct tissue-types.

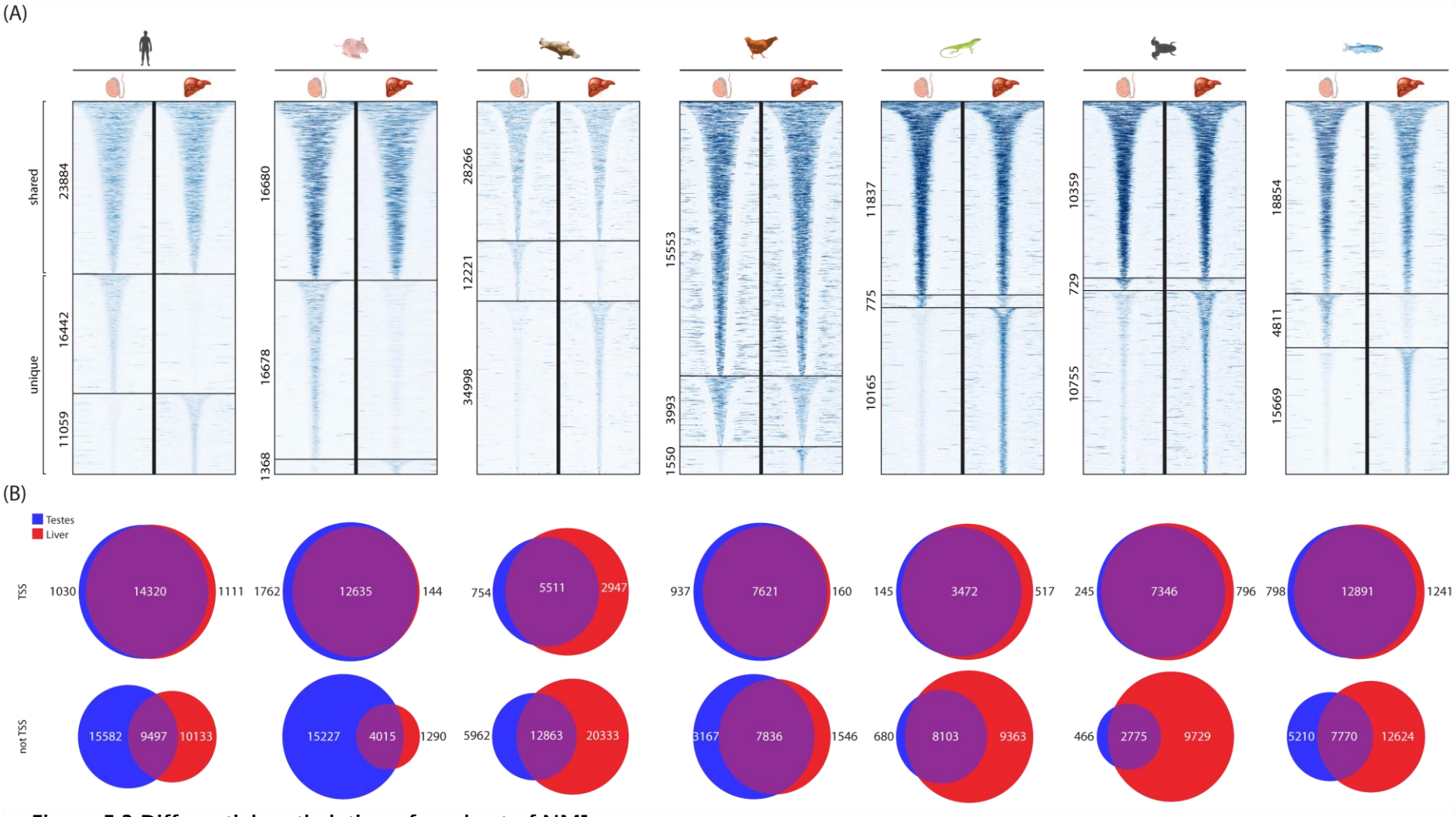


Figure 5.3 Differential methylation of a subset of NMIs

(A) All vertebrate genomes have a subset of NMIs that are subject to differential methylation as illustrated by a heat map of non-methylated DNA signal from testis and liver in human, mouse, platypus, chicken, lizard, frog and zebrafish. In each case, NMIs are ranked according to length and clustered as shared (upper) or unique (lower) between the two tissues. A 5kb window centred at the NMI is shown and read density is indicated by colour intensity. (B) The overlap of NMIs identified in liver and testis is depicted by Venn diagrams for NMIs associated with protein-coding TSSs (upper) and for NMIs away from TSSs (lower). NMIs at TSSs are generally non-methylated in both tissues whereas differentially methylated NMIs tend to be found away from TSSs.

5.1.3 Validation of differential methylation between testis and liver

In order to validate that the unique NMIs were truly differentially methylated between, a number of the testis or liver unique NMIs were validated by locus-specific bisulfite sequencing for mouse and zebrafish (Figure 5.4). Interestingly, differentially methylated NMIs exhibited a tendency to be mosaically methylated, as demonstrated by the *Patl2* gene promoter in mouse and the 3' exonic NMI in the zebrafish *akap11* gene (Figure 5.4Ai and Biii). A similar observation has been described recently for mouse ES cells and neuronal progenitors (Stadler et al., 2011) whereby low methylated regions (LMRs), exhibiting around 30 % DNA methylation, tended to be differentially methylated between tissues (Stadler et al., 2011). Therefore, partial methylation may be an indicator of regions which are susceptible to DNA methylation changes between tissues. However, since it is not possible to infer the methylation status of individual CpGs using the Bio-CAP technique, this observation was not pursued further here.

5.2 **Dynamic DNA methylation occurs away from gene transcription start sites**

5.2.1 Non-methylated islands at gene TSSs are shared between testis and liver

A similar proportion of gene TSSs in testis and liver were observed to be associated with an NMI for each species (Figure 4.8A and Figure 5.2A). This suggested either that the same subset of gene promoters remained non-methylated in both tissues or that a similar number of different gene TSSs were non-methylated in liver compared to testis. To determine whether DNA methylation varies at gene promoters, all NMIs located within 1 kb of a gene TSS were compared for liver and testis. This analysis revealed a high degree of overlap for testis and liver NMIs located at the 5' end of genes in all species (Figure 5.3B, upper). The correlation was not as strong for platypus, but this is almost certainly due to the poor annotation of gene TSSs for this species. Therefore, the majority of vertebrate gene promoter-associated NMIs appear to remain non-methylated between two very distinct tissues. Interestingly, only minimal DNA methylation changes were detected at gene promoters for 30 diverse cell-types in human, suggesting that mammalian gene promoters tend to

be non-methylated across the majority of cell-types (Ziller et al., 2013). Although whether this observation holds true for other vertebrate species remains to be seen.

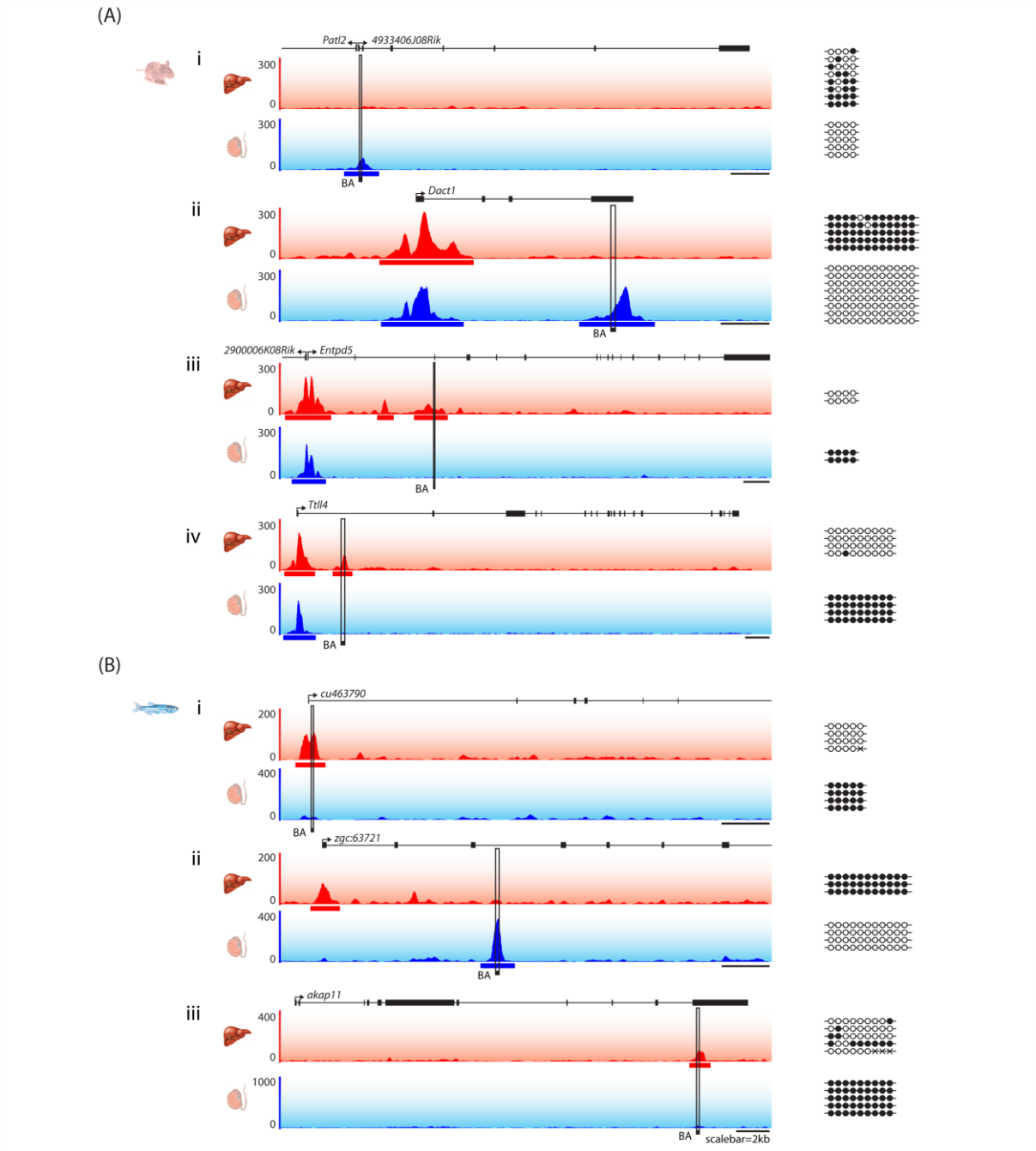


Figure 5.4 Validation of differentially methylated NMIs between liver and testis in mouse and zebrafish by bisulfite sequencing

(A, i-iv) Mouse NMIs unique to liver or testis were analysed by bisulfite sequencing to verify that these regions were indeed differentially methylated. Traces of non-methylated DNA are shown for differentially methylated regions in mouse liver (red) and testis (blue) with NMIs shown as bars under the traces. The y-axis depicts read density. Bisulfite PCR amplicons are shown as a black rectangle (BA). Empty and filled circles represent non-methylated and methylated CpG dinucleotides, respectively. (B, i-iii) Zebrafish NMIs unique to liver or testis were validated by bisulfite sequencing as in (A).

An important caveat to the analysis presented here is that the presence or absence of a Bio-CAP peak is compared as a binary measure; therefore some more subtle or quantitative changes in NMI DNA methylation levels may have been missed. This is important to consider as previous studies have identified subtle changes in levels of DNA methylation, including changes in the size or shape of a non-methylated region (Doi et al., 2009; Irizarry et al., 2009; Molaro et al., 2011), which may have functional importance in gene regulation. However, irrespective of subtle changes in the size of the non-methylated domain, NMIs at vertebrate gene promoters appear to remain non-methylated across distinct tissue-types, potentially as a means to demarcate a subset of genes as being subject to a particular mode of gene regulation.

5.2.2 Differential methylation of gene promoters is associated with differential gene expression

Most NMIs at gene TSSs were shared between tissues. However, in all species tested a small subset of promoter-associated NMIs were observed to be unique to one tissue (Figure 5.3B, upper). *De novo* placement of DNA methylation at gene promoters is known to be linked with long-term silencing of gene expression (Bird, 2002; Jones, 2012; Jones et al., 1998; Nan et al., 1998; Razin, 1998). Therefore, using RNA-seq datasets available for human, mouse, platypus and chicken, gene expression was monitored for the genes which underwent DNA methylation changes between testis and liver (Brawand et al., 2011; Julien et al., 2012; Stadler et al., 2011). This analysis revealed that genes which were differentially expressed between the two tissue-types were significantly enriched for having a differentially methylated promoter-associated NMI (Figure 5.5). Therefore, the presence of DNA methylation at a gene TSS closely correlates with gene repression and, although an infrequent event, DNA methylation mediated silencing appears to be a common epigenetic strategy for regulating a subset of vertebrate genes.

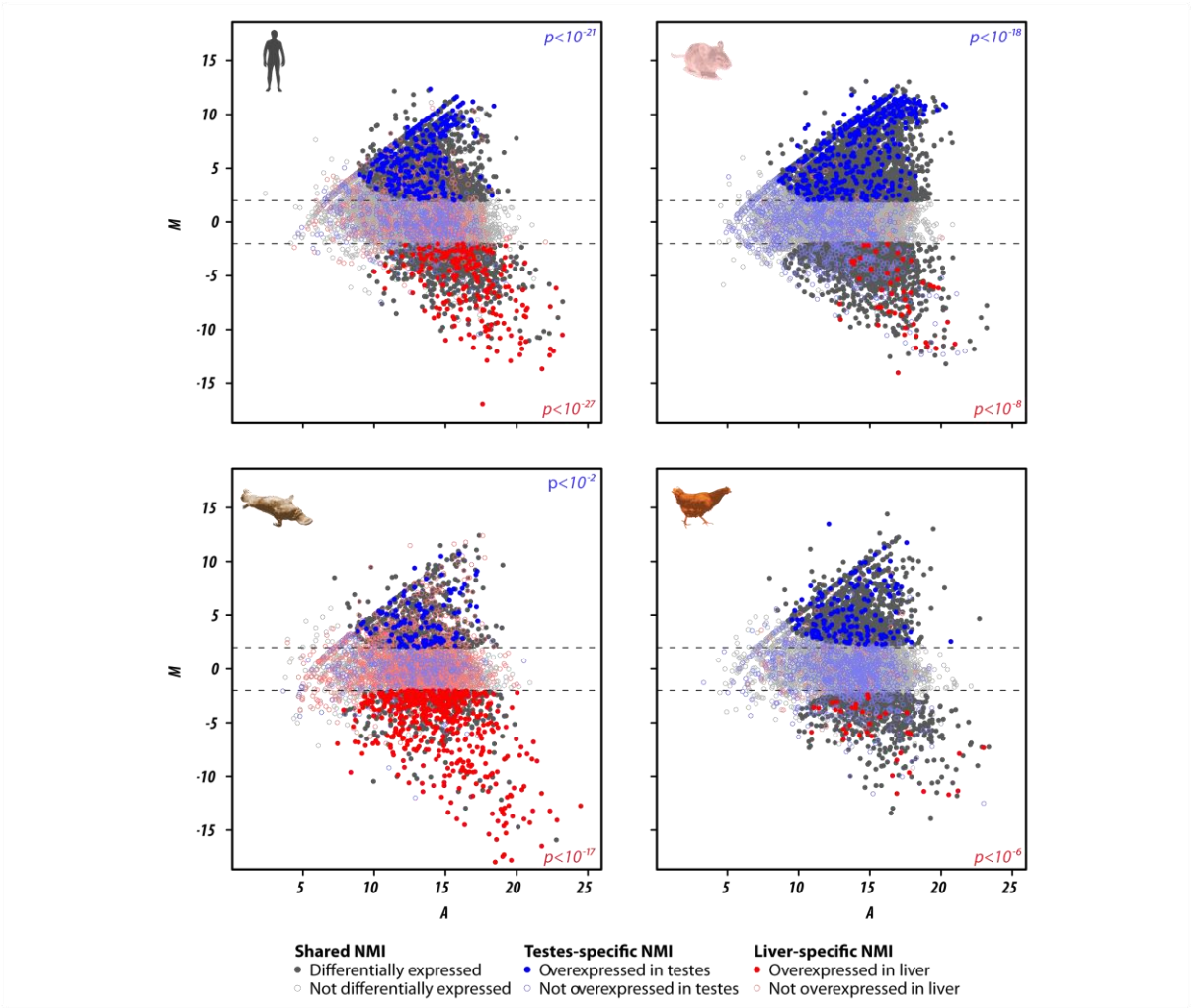


Figure 5.5 Genes with TSS-associated testis or liver specific NMIs are over-represented for increased differential expression in the same tissue

MA plots depicting expression differences for genes with TSS-associated NMIs from liver and testis for human, mouse, platypus and chicken. Genes are coloured according to whether they share an NMI in both liver and testis (grey) or have an NMI only in liver (red) or testis (blue). Genes are further distinguished as being differentially expressed or overexpressed in a tissue-specific manner (dark, filled circle) or not (light, open circles). The log mean expression of the gene in both liver and testis is displayed on the x axis (A) and the log ratio of gene expression is displayed on the y axis (M). The dotted lines indicate a fold change threshold of 2. Genes with tissue-specific NMIs were significantly over-represented in the set of genes which had increased differential expression in 7 out of 8 cases (Fisher's exact test, human testis $p < 10^{-21}$, liver $p < 10^{-27}$; mouse testis $p < 10^{-18}$, liver $p < 10^{-8}$; platypus testis $p < 10^{-2}$, liver $p < 10^{-17}$; chicken liver $p < 10^{-6}$).

5.2.3 Tissue-specific non-methylated islands occur away from gene TSSs

To examine DNA methylation changes occurring away from gene promoters, NMIs that were greater than 1 kb away from TSSs were compared for testis and liver tissues. Interestingly, non-TSS NMIs exhibited a much lower degree of overlap between testis and liver tissues, indicating that the majority of differential methylation occurs away from gene promoters. This observation may be

even more striking than presented, as a number of the non-TSS NMIs (Figure 5.3B, lower, purple) may actually occur at unannotated or alternative gene start sites which generally would be shared between tissues. Furthermore, the NMIs which happen to be shared between these two particular tissues may be methylated in another tissue not profiled here. Together, these observations reveal a common usage of constitutive NMIs across vertebrates at gene promoters while differentially methylated NMIs appear to be found away from TSSs.

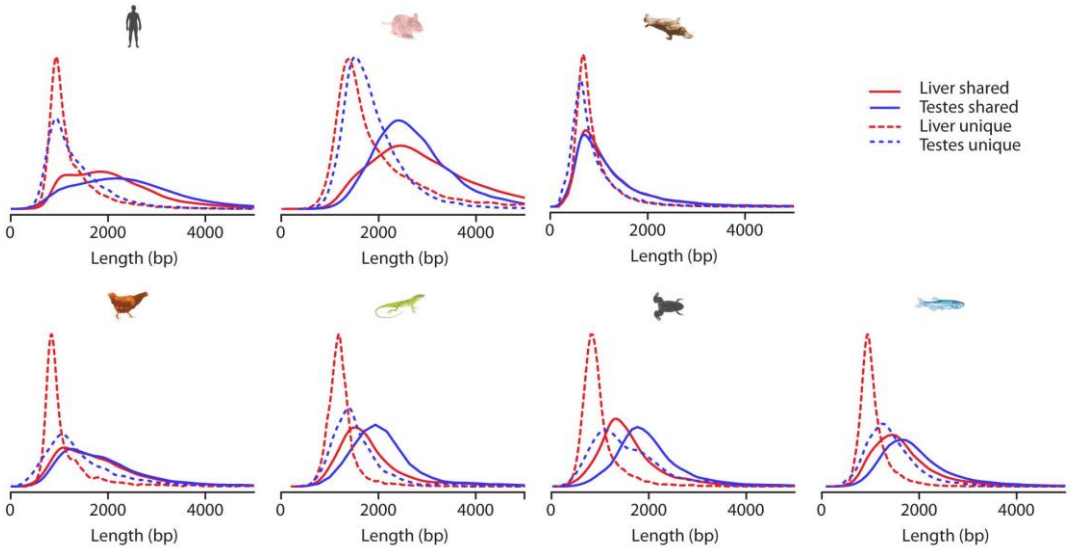


Figure 5.6 Length distributions for shared and unique NMIs for seven vertebrates

NMI length distribution plots for shared (Shared NMIs, solid line) or unique (Unique NMIs, dashed line) NMIs from testis (blue) or liver (red). Shared NMIs tend to be longer than tissue-specific unique NMIs.

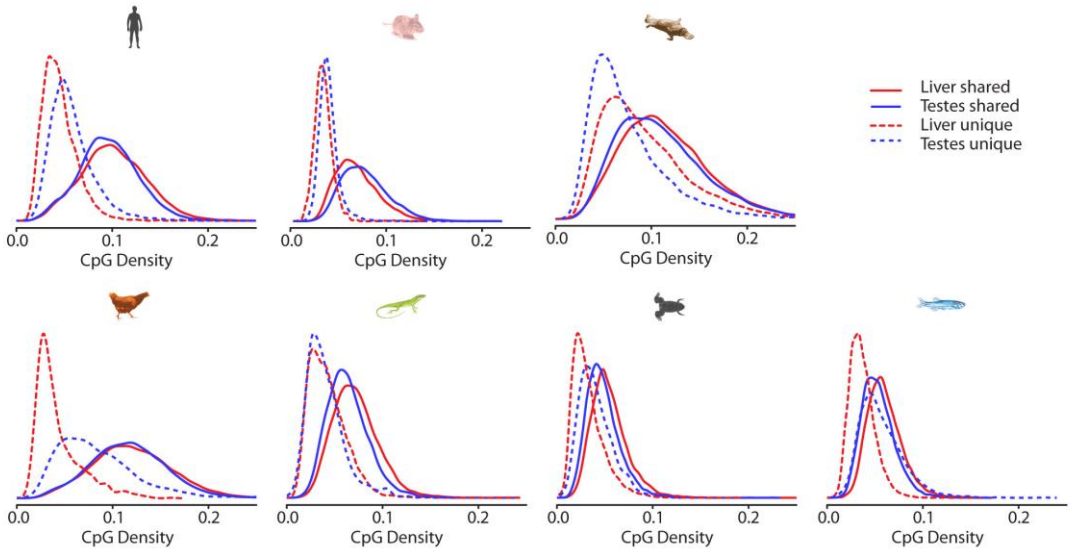


Figure 5.7 Unique vertebrate NMIs have lower CpG density than shared NMIs

CpG density distribution plots for shared (solid line) or unique (dashed line) NMIs from testis (blue) or liver (red). Shared NMIs tend to have higher CpG density than unique NMIs.

5.3 **Differentially methylated NMIs have distinct CpG density and length compared to constitutively non-methylated NMIs**

Interestingly, it was apparent from the heat maps of non-methylated DNA from testis and liver that NMIs clearly clustered into two distinct types. The shared NMIs appeared longer than unique NMIs, with more intense Bio-CAP signal, suggesting that more non-methylated DNA was purified at shared NMIs compared to those which are unique to testis or liver (Figure 5.3A). The lower Bio-CAP signal at unique NMIs could be attributed to mosaic DNA methylation, as was detected for individual loci (Figure 5.4), or a reduced CpG density at these regions. To further investigate the properties of unique NMIs, the length and CpG density properties were calculated for all vertebrate NMIs which were shared or unique between testis and liver tissue. Frequency distributions plots revealed that unique NMIs (dashed lines) were shorter and less CpG dense overall than shared NMIs (bold lines, Figure 5.6 and Figure 5.7). Therefore, differentially methylated NMIs exhibit unique nucleotide and length properties compared to shared TSS-associated NMIs.

The clear disparity between the nucleotide and length properties of shared and tissue-specific NMIs suggests that that these features may be important for the function of these elements. Two possible explanations can be envisaged to explain the reduced size and CpG density of plastic differentially methylated NMIs. Firstly, over evolutionary time differentially methylated NMIs may have been more frequently methylated in the germline compared to TSS NMIs (Weber et al., 2007) and thus increasingly eroded in CpG density due to accumulation of deamination mutations at methylated cytosines (Coulondre et al., 1978). Selective pressure for the retention of NMIs in particular tissues or lineages may have prevented the complete loss of CpG density at these differentially methylated NMIs compared to intergenic regions (Weber et al., 2007). Alternatively, lower CpG density may have been selected for during evolution as a necessary feature to facilitate differential methylation at NMIs unique for a particular tissue. In this scenario, high CpG density may be a signal for permanent lack of methylation and would therefore preclude dynamic methylation at CpG dense regions whereas intermediate CpG density may be optimal for switching between methylation states

between tissue-types. Whichever may be the case, given that DNA sequence remains unchanged between tissue-types, it is probable that DNA methylation changes observed at plastic NMIs are made possible by the nucleotide composition, but are initiated by changes in the local chromatin environment, recruitment of DNA-binding factors or demethylase enzymes during differentiation (Costa et al., 2013; Strunnikova et al., 2005; Thomassin et al., 2001).

5.4 **Intergenic NMIs are associated with distal regulatory regions**

In order to understand how the differential methylation observed at non-TSS NMIs may contribute to genome function, intergenic NMIs (greater than 5 kb from a protein-coding, Figure 5.8A) were compared to annotated genomic features. For these analyses, Bio-CAP-seq data generated in the previous chapter for mouse V6.5 ES cells and zebrafish embryonic stage 24 hpf were used, for which genome-wide transcription and chromatin modification datasets were available for comparison (see 2.17.1). Initially, non-coding RNA genes were interrogated to identify whether NMIs are associated with non-coding gene TSSs in the same way as was observed for coding genes (Figure 4.8A and Figure 5.2A). A large proportion of long non-coding RNA (lncRNA) TSSs identified in mouse and zebrafish, 45 and 64 % respectively, had a TSS-associated NMI (Figure 5.8B) (Belgard et al., 2011; Guttman et al., 2010; Pauli et al., 2012; Ulitsky et al., 2011). This is perhaps unsurprising as lncRNA genes are transcribed by RNA Polymerase II (RNAPII) (Guttman et al., 2009; Wang and Chang, 2011) and thus are likely to be regulated in a similar manner to coding RNAs. However, other classes of non-coding RNAs were not as highly enriched for TSS-associated NMIs (Appendix 5.1).

The strong correlation with lncRNA TSSs, and overlap with other annotated non-coding TSSs, accounted for a significant number of intergenic NMIs in mouse ES cells and zebrafish 24 hpf, but by no means all (Figure 5.8C and Appendix 5.1). Functional annotation was therefore sought for the remaining intergenic NMIs in a hierarchical manner using other available external datasets. Enhancer regions have been shown to be associated with specific histone marks, most notably the H3K4me1 modification (Heintzman et al., 2009; Heintzman et al., 2007; Koch et al., 2007; Zentner et al., 2011).

Furthermore, non-productive transcription has been reported to occur at active enhancer elements (Natoli and Andrau, 2012). Of the remaining intergenic NMIs, a large proportion were marked with H3K4me1 ChIP-seq signal in both mouse and zebrafish (Figure 5.8C), and a further fraction were associated with the 5' end of transcripts identified from recently published RNA-seq data, representing either unannotated genes, or non-productive transcription typical of enhancers (Natoli and Andrau, 2012). In total, 46 and 57 % of non-genic NMIs from mouse ES cells and zebrafish 24 hpf were associated with functional annotation typical of either non-protein-coding genes or enhancers, suggesting that differential methylation may act to epigenetically modify distal gene regulatory elements.

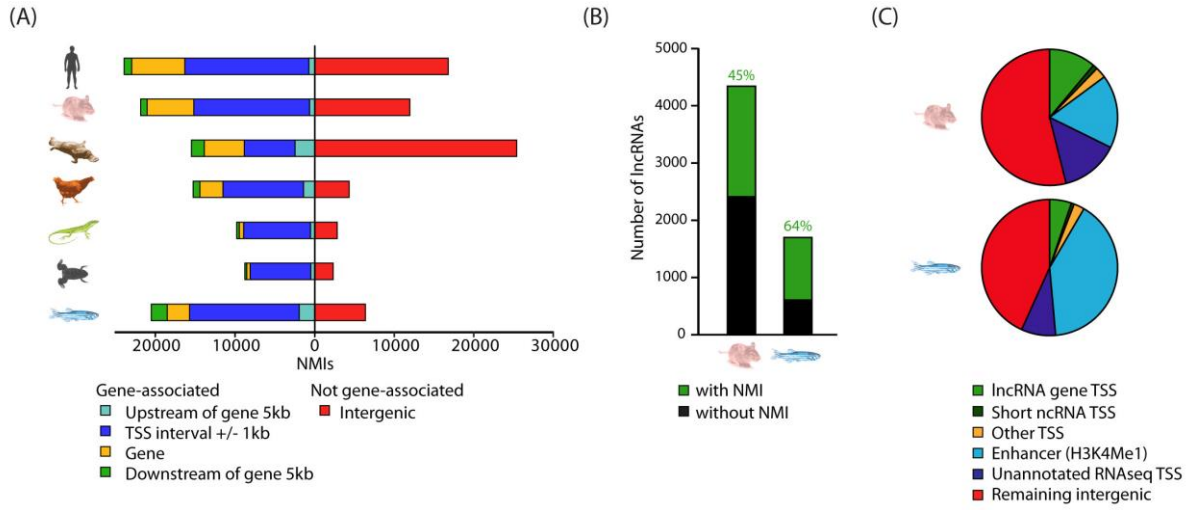


Figure 5.8 Intergenic NMIs are associated with distal regulatory elements, non-coding RNAs, and unannotated transcripts

(A) Most NMIs are associated with known protein-coding genes (left) but a substantial proportion are located within intergenic regions of the genome (right). (B) NMIs (green) are found at 45% and 64% of all known long non-coding RNA (lncRNA) TSSs (black) in mouse and zebrafish respectively. (C) A pie chart depicting the proportion of intergenic NMIs (>5kb from a protein-coding gene) associated with different genomic features in mouse embryonic stem (ES) cells and zebrafish 24hpf embryos. The association was performed hierarchically in the following order: lncRNA TSSs, other non-coding RNA TSSs (miRNAs, rRNAs, snRNAs, or snoRNAs), other TSSs (pseudogenes and processed transcripts), putative enhancer mark H3K4me1 and novel RNA-seq TSSs. This analysis indicates that intergenic NMIs mark novel transcriptional units or regulatory elements.

5.5 **Differential NMIs act as an epigenetic switch**

Given that the majority of differential DNA methylation of NMIs occurs away from known gene promoters (Figure 5.3B), and that intergenic NMIs are enriched for regulatory elements such as distal enhancers (Figure 5.8C), an interesting possibility is that differential methylation away from

gene promoters may function as part of an epigenetically driven regulatory process. Overlap of available testis and liver H3K4me3 datasets with unique human testis NMIs indeed revealed that testis specific H3K4me3 signal is detectable at these sites with negligible H3K4me3 signal detected from liver (Figure 5.9A, upper left). The reverse held true for the liver unique peaks which only exhibited H3K4me3 signal from the liver H34me3 ChIP-seq dataset (Figure 5.9A, upper right). The same trend was also observed using available external datasets for mouse (Figure 5.9A, lower). Type-1 ZF-CxxC domain-containing proteins (Long et al., 2013) are known to bind specifically to non-methylated DNA (Blackledge et al., 2010; Thomson et al., 2010). A member of this family, CFP1, is part of a SET1-containing complex which catalyses placement of the H3K4me3 histone modification (Lee and Skalnik, 2005), which is associated with active chromatin. Therefore, it is highly probable that unique NMIs, differentially methylated between tissue-types, act to direct lineage-specific placement of the H3K4me3 chromatin mark.

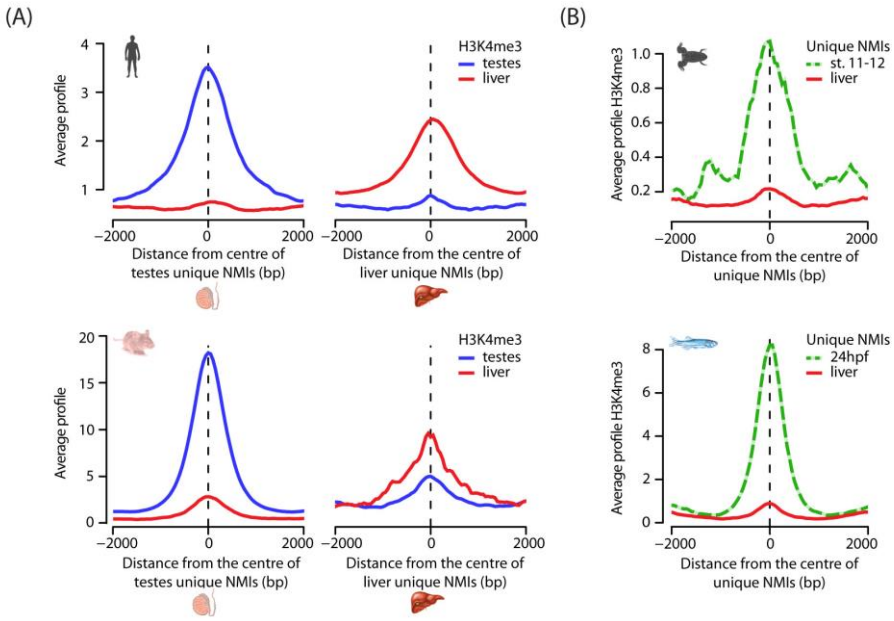


Figure 5.9 Chromatin modification at NMIs depends on their underlying DNA methylation state

(A) H3K4me3 read density from testis (blue) and liver (red) is profiled over testis unique (left) and liver unique (right) NMIs for human (upper) and mouse (lower) and displayed as an average profile. At differentially methylated loci, the histone H3K4me3 modification is found preferentially in the tissue with the non-methylated NMI. (B) The H3K4me3 signal (profiled in frog stage 11-12 embryos and zebrafish 24hpf) is present specifically at unique NMIs from frog stage 11-12 and zebrafish 24hpf (green) and not at unique NMIs from the liver (red).

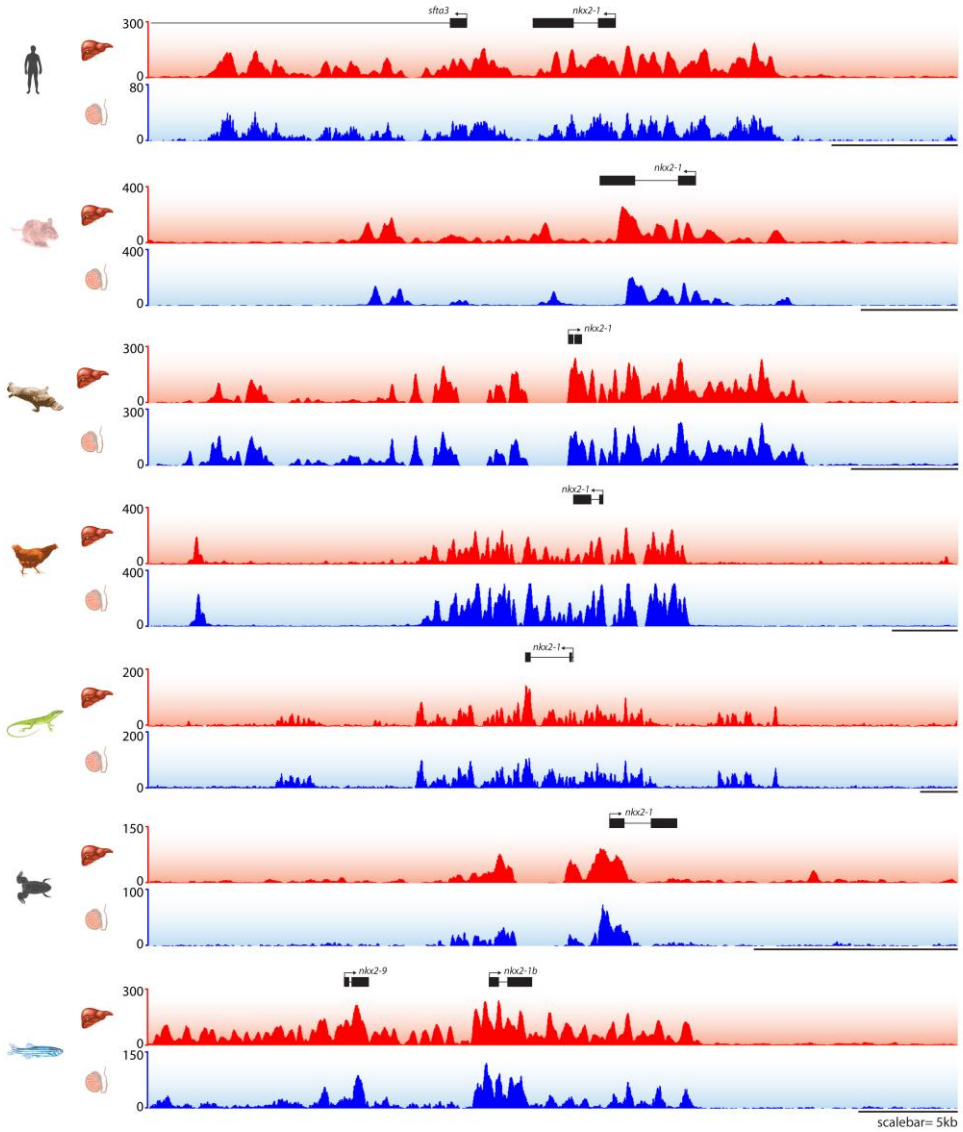


Figure 5.10 A broad non-methylated island region at the *nkx2-1* gene is shared across tissues for all seven vertebrate species

The nkx2-1 gene is associated with an extended region of non-methylated DNA in both liver and testis for a number of vertebrate species. hg19 chr14:36,970,102-37,002,381; mm9 chr12:57,604,491-57,664,542; ornAna1 chr4:23,088,393-23,126,926; galGal3 chr5:38,981,295-39,043,103; anoCar2 chr1:24,593,153-24,701,127; xenTro3 GL172781:1727255-1747537; danRer7 chr17:38,412,338-38,453,558.

Importantly, this observation was not unique to warm-blooded vertebrates. H3K4me3 dataset were available for frog and zebrafish embryonic tissues stage 11-12 and 24 hpf respectively (see 2.17.1), for which matching Bio-CAP-seq datasets had already been generated. Comparison of unique embryonic NMIs and unique liver NMIs for frog and zebrafish revealed that the H3K4me3 modification was only enriched at the embryonic stage unique NMIs, with no H3K4me3 signal detected for the liver unique NMIs which are methylated in the frog and zebrafish embryo (Figure 5.9C). Thus the chromatin modification state of tissue-specific NMIs appears to change in an

epigenetic switch-like mechanism correlating with the underlying DNA methylation state. Therefore this epigenetically plastic set of NMIs, identified for all species, appears to affect the local chromatin landscape at distal regulatory elements and perhaps drive functional diversity across different tissue types in an otherwise fixed DNA sequence.

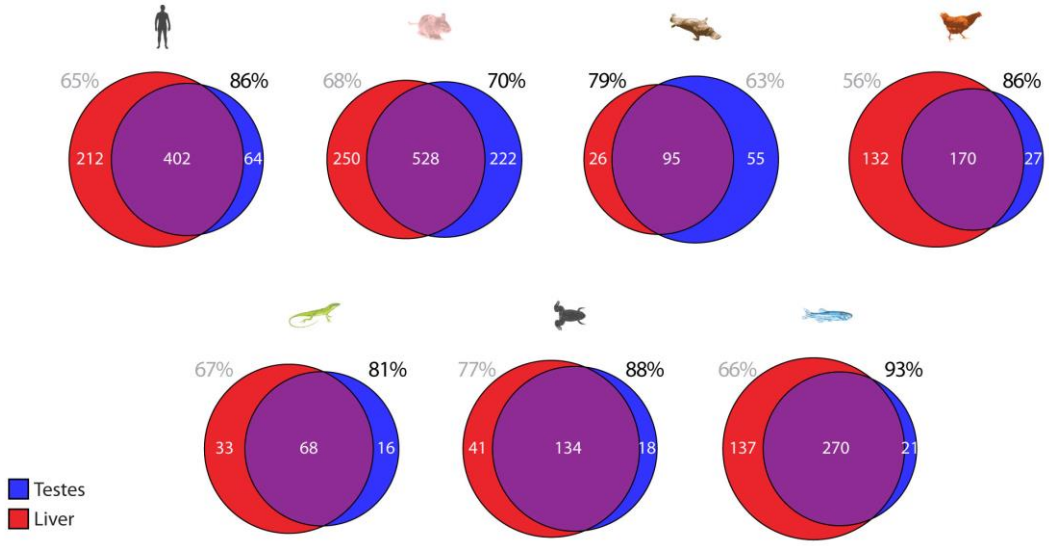


Figure 5.11 Broad NMIs are maintained between diverse tissue types

Genes whose gene body is greater than 90% overlapped by NMI intervals were identified for testis and liver. Venn diagrams depict the degree of overlap for testis and liver broad NMI genes.

5.6 **Broad non-methylated islands appear to be maintained across tissues**

NMIs overlapping gene TSSs appear to remain non-methylated across distinct tissue-types (Figure 5.3B, upper). This raises the question of whether the class of broad NMIs identified in Chapter 4 also remains broadly non-methylated across multiple tissues. Visual inspection of gene-associated broad NMIs revealed that the non-methylated DNA signature appeared to be maintained between testis and liver (for example the *otp* and *nkx2.1* genes, Figure 5.1 and Figure 5.10). In order to expand this comparison genome-wide, genes exhibiting broad non-methylated regions were identified for liver as before (with greater than 90% of the coding region covered with Bio-CAP signal, 4.7.2). Comparison of broad NMI gene lists for testis and liver revealed that a significant proportion of this class of gene remain broadly non-methylated between distinct tissue-types for all species analysed (Figure 5.11). Therefore, broad NMI genes appear to be protected from DNA methylation in different

tissues, similar to canonical promoter-associated NMIs. This suggests that in contrast to distal regulatory elements, broad NMIs generally are not modulated by dynamic DNA methylation.

5.7 **Discussion**

5.7.1 A third class of 'plastic' NMI

In addition to the canonical and broad NMIs described in the previous chapter, here a third 'plastic' class of NMIs was described. Plastic NMIs exhibited lower CpG density than canonical NMIs, were generally found associated with distal gene regulatory elements and were differentially methylated between tissues. Together these three classes of NMI appear to constitute a highly conserved epigenetic logic for the utilisation of non-methylated DNA across vertebrate evolution.

5.7.2 DNA methylation changes mostly occur away from gene promoters

In keeping with previous reports, DNA methylation changes at gene promoters were observed to be a rare event, with the majority of TSS-associated NMIs remaining non-methylated in both testis and liver tissues (Illingworth et al., 2010; Maunakea et al., 2010; Meissner et al., 2008; Ziller et al., 2013). However, for the small number of cases where a promoter NMI was methylated in one tissue, DNA methylation was associated with repression of transcription (Futscher et al., 2002; Song et al., 2005). This gene silencing may be mediated by methyl-binding proteins following DNA methylation gain (Bogdanovic and Veenstra, 2009; Clouaire and Stancheva, 2008) or DNA methylation may be acquired once the associated gene has already entered the repressed state (Jones, 2012). Interestingly, a number of NMIs away from gene TSSs were shared between testis and liver. It is likely that a proportion of these will in fact overlap alternative start sites upstream or within gene bodies (Maunakea et al., 2010), unannotated genes (Illingworth et al., 2010) or mis-annotated lncRNAs. Furthermore, shared intergenic NMIs which do not overlap any of the features listed may in fact be differentially methylated in other tissues not profiled here (Ziller et al., 2013).

Unlike at gene TSSs, a large proportion of non-TSS NMIs undergo DNA methylation changes between testis and liver. The concomitant placement of H3K4me3 modification at these sites, observed

previously in mammals (Schmidl et al., 2009; Stadler et al., 2011), suggests that the presence of non-methylated DNA may act to positively regulate enhancer-like elements and non-coding RNA genes. Regulation of these elements in this manner will likely impact the expression of nearby genes. Together, differential methylation of intergenic NMIs during development and differentiation, in distinct tissue-types, may play an important role in the differential regulation of distal enhancers and the expression of lineage-specific genes across vertebrate species.

5.7.3 Distinct features of shared and unique vertebrate non-methylated islands

Interestingly, certain features appear to vary between the constitutively non-methylated 'canonical' class and the differentially methylated 'plastic' class of NMIs. These differences may help us begin to understand what it is that drives or predisposes a genomic locus to DNA methylation changes. Firstly, CpG density generally appeared lower for tissue-specific NMIs for all species (Figure 5.7), which is reflected in the lower Bio-CAP signal intensity observed for these regions (Figure 5.3A) due to fewer CpG substrates for the ZF-CxxC dependent affinity purification (Bio-CAP). In keeping with this, it has been reported that CpG poor regions are most often silenced by DNA methylation changes as opposed to polycomb-mediated repression (Xie et al., 2013). Partial DNA methylation could also be responsible for the reduced Bio-CAP signal as low methylated regions (10-50 % methylated) have been reported be differentially methylated between cell-types in the mouse (Stadler et al., 2011). Secondly, tissue-specific NMIs were observed to be much shorter than NMIs which were shared between testis and liver (Figure 5.6) (Stadler et al., 2011; Ziller et al., 2013). This suggests that the encroachment or loss of DNA methylation at dynamically methylated regions may be aided by the small size of the non-methylated region which will exhibit many fewer CpG sites than a constitutive NMI (Hon et al., 2012). Together, differentially methylated regions appear to be shorter in length, have a reduced CpG density and exhibit mosaic DNA methylation. All of these features may cause a sensitisation to signals which influence DNA methylation profiles and a predisposition to differential methylation between distinct tissue-types. How these and other factors

influence the non-methylated DNA landscape will be explored in more detail in the following chapter.

5.7.4 DNA methylation changes at the edges of non-methylated islands

Interestingly, NMI length was also observed to vary between tissue types for some species at constitutively non-methylated NMIs (Figure 5.6). In this analysis, sperm NMIs were observed to be longer than those from a non-germline tissue at all shared NMIs for lizard, frog and zebrafish, extending an observation previously made for human (Molaro et al., 2011). This change in NMI size is reminiscent of CpG island shores (Doi et al., 2009), and a number of reports have inversely linked changes in DNA methylation at the edges of non-methylated regions with the placement of active histone modifications (Hodges et al., 2011) and activation of associated genes (Doi et al., 2009; Hodges et al., 2011; Irizarry et al., 2009). Whether variation in NMI width is functional, and is important for or merely responsive to gene expression changes remains to be seen.

5.7.5 The drivers of DNA methylation change

DNA methylation changes occurring at non-TSS NMIs during development will involve either the loss or gain of DNA methylation. *De novo* placement of DNA methylation is poorly characterised at non-promoter elements, but may be linked to an increase in nucleosome occupancy, for example following transcription factor binding loss (You et al., 2011), as it has been suggested that nucleosomes may be the preferred substrate for *de novo* methyltransferases (Chodavarapu et al., 2010). This remains contentious, however, and many other reports claim that linker DNA and chromatin remodelling is required for *de novo* DNA methylation (Felle et al., 2011), in keeping with a dramatic reduction in DNA methylation in mice lacking the chromatin remodeller, *Lsh* (Dennis et al., 2001; Geiman et al., 2001). Therefore perhaps loss of protective DNA binding events, which shielded a specific DNA sequence from *de novo* methylation (Feldman et al., 2006b; Fuks et al., 2001; Hervouet et al., 2009; Plath et al., 2003; Velasco et al., 2010), would expose a stretch of DNA leading to acquisition of DNA methylation. It would be interesting therefore to observe whether the

differentially methylated NMIs identified here contain tissue-specific DNA binding motifs within their sequence, and subsequent loss of these DNA binding factors during differentiation may drive DNA methylation changes between tissues.

For NMIs which lose DNA methylation between cell-types, either passive or active demethylation may drive this change. Active DNA demethylation has been proposed to be mediated by the TET enzymes which catalyse the oxidation of 5-methylcytosine (5mC) to 5-hydroxymethylcytosine (5hmC) (Kriaucionis and Heintz, 2009; Tahiliani et al., 2009). Interestingly, 5hmC has been detected at cis-regulatory elements (Ficz et al., 2011; Pastor et al.; Wu et al., 2011a) suggesting that the TET enzymes may be directed to methylated DNA, perhaps through recruitment by a transcription factor (Costa et al., 2013) to drive active demethylation of lineage-specific NMIs. Conversely, loss of the DNA binding protein which recruits the TET enzymes may cause a gain of DNA methylation. Other *in vitro* and *in vivo* evidence supports a role for sequence specific binding events driving loss of DNA methylation (Stadler et al., 2011; Thomassin et al., 2001), although whether this is due to passive demethylation over multiple cell divisions or active demethylation is unclear. Passive demethylation is proposed to be replication-dependent and occur when maintenance DNA methylation is inhibited, perhaps through the binding of specific DNA-binding factors (Hsieh, 1999; Lin and Hsieh, 2001; Lin et al., 2000; Matsuo et al., 1998). Importantly, it has been shown that this can occur away from gene promoter regions and therefore could constitute a mechanism for driving differential methylation of distal regulatory elements (Lin et al., 2000; Stadler et al., 2011; Thomassin et al., 2001). However, despite the knowledge that large-scale DNA methylation changes occur between distinct tissues, it remains incompletely understood how these changes are defined and mediated during differentiation and development.

Chapter 6 -- Introduction of exogenous DNA sequence uncovers traits necessary for non-methylated island formation

As described in the previous chapter, the mechanisms by which patterns of non-methylated DNA are defined and modulated in distinct tissues are incompletely understood. Factors contributing to DNA methylation patterns are likely to include both the nucleation of DNA binding factors and the underlying DNA sequence composition. Indeed there is some evidence to support a role for both of these factors in defining DNA methylation patterns, with reports that transcription factor binding can direct a DNA sequence to become non-methylated (Stadler et al., 2011), while increased CpG observed over expected ratio and high GC content (the criteria for CpG islands) is partially predictive of non-methylated DNA in warm-blooded vertebrates (Gardiner-Garden and Frommer, 1987) (this thesis, Chapter 4). However, the extent to which these and other features contribute to the non-methylated state is unknown.

In this chapter, the mechanisms by which non-methylated DNA is defined will be explored, utilising two model systems which introduce novel DNA sequences into the mouse nucleus.

6.1 **Introduction of heterologous DNA sequences to investigate establishment of non-methylated DNA**

It is difficult to address experimentally how non-methylated DNA profiles are established at the genome-scale as many of the contributing factors may be currently unknown. To investigate the extent to which nucleotide sequence contributes to DNA methylation status, novel DNA sequences can be introduced into a heterologous nuclear environment. This methodology has been used previously to determine that introduction of short sequences containing a so-called methylation-determining region (MDR) can drive a DNA sequence to become non-methylated (Lienert et al., 2011). These short sequences may be subject to site of integration effects from the surrounding chromatin, an effect which could be overcome by integration of a much larger DNA fragment. A number of mouse models exist which have incorporated much larger portions of foreign DNA, and

which could be exploited to determine non-methylated profiles of a large exogenous DNA sequence. Such models include the humanised mouse model in which the mouse alpha-globin locus was replaced with that of human (Wallace et al., 2007), a mouse containing two human chromosome fragments (Tomizuka et al., 2000) and a number of mouse models for Down's syndrome which contain large fragments of human chromosome 21 (Kazuki et al., 2001; O'Doherty et al., 2005; Smith et al., 1997).

6.1.1 A transchromosomic Down's syndrome mouse model (Tc1)

The transchromosomic mouse model (Tc1) represents the most intact model for Down's syndrome to date as it has an almost entire copy of human chromosome 21 which encodes 198 human genes (O'Doherty et al., 2005). Human chromosome 21 in the Tc1 mouse, which has regions syntenic to sections of mouse chromosomes 10, 16 and 17 (Antonarakis et al., 2004), is freely segregating, present in 24-55 % of cells and transmitted through the female germline (O'Doherty et al., 2005). Previous analysis of human chromosome 21 in this model has shown that transcription factor binding and histone modifications are largely dictated by DNA sequence (Wilson et al., 2008) and therefore information within the genetic sequence of human chromosome 21 appears to be sufficient to recapitulate the human transcriptional profile in the heterologous mouse nuclear environment (Wilson et al., 2008). The Tc1 mouse model therefore provides an opportunity to determine whether DNA sequence can similarly dictate non-methylated DNA profiles irrespective of the regulatory milieu. The Tc1 mouse model was therefore used to investigate the methylation status of 42 Mb of human DNA sequence in the mouse heterologous nuclear environment.

6.2 **Profiling DNA methylation in the Tc1 mouse model**

6.2.1 Bio-CAP sequencing and alignment to the Tc1 mouse

Non-methylated DNA was profiled by Bio-CAP sequencing for Tc1 mouse testis and liver genomic DNA, as described previously (2.13.2.2). These tissues were chosen as Bio-CAP sequencing data had previously been generated for the same tissues from wildtype mouse and human (Chapter 4 and

Chapter 5), and furthermore multiple external genome-wide ChIP-seq datasets were available for comparison (Schmidt et al., 2010; Ward et al., 2013; Wilson et al., 2008). The Bio-CAP sequencing data was aligned to a composite Tc1 genome containing all mouse chromosomes (19 autosomes plus sex chromosomes) and human chromosome 21 (2.14.4). Read mapping was restricted so that individual reads could only align once to suppress highly similar sequences in the mouse genome from aligning to human chromosome 21. Around 92% of human chromosome 21 is present in the Tc1 mouse, therefore for downstream analysis four regions which appear to have been lost during chromosome transfer (including two previously reported regions (O'Doherty et al., 2005)) were excluded (see 2.16.1). Importantly, comparison of non-methylated DNA profiles from Tc1 mouse with wildtype mouse revealed that the presence of a human chromosome in the mouse nucleus does not grossly impact the size and distribution of mouse NMIs in testis tissue for all mouse chromosomes (Appendix 6.1). This is in agreement with normal H3K4me3 features for all mouse chromosomes described previously (Wilson et al., 2008).

6.2.2 Non-methylated islands are recapitulated at gene promoters in the Tc1 mouse model

In order to observe whether non-methylated DNA profiles are similar for human chromosome 21 between the human and mouse regulatory environments, non-methylated DNA profiles were visualised for both testis and liver tissues (Figure 6.1Ai and Bi). Similar to the endogenous profiles in human (Hs-chr21, dark red and dark blue), human chromosome 21 in the Tc1 mouse (Tc1-chr21, bright red and bright blue) exhibited strong punctate peaks of non-methylated DNA coincident with gene promoters (Figure 6.1Ai and Bi). Indeed, a very similar proportion of genes on human chromosome 21 were overlapped by non-methylated DNA in both the human and mouse nucleus (Table 6.1).

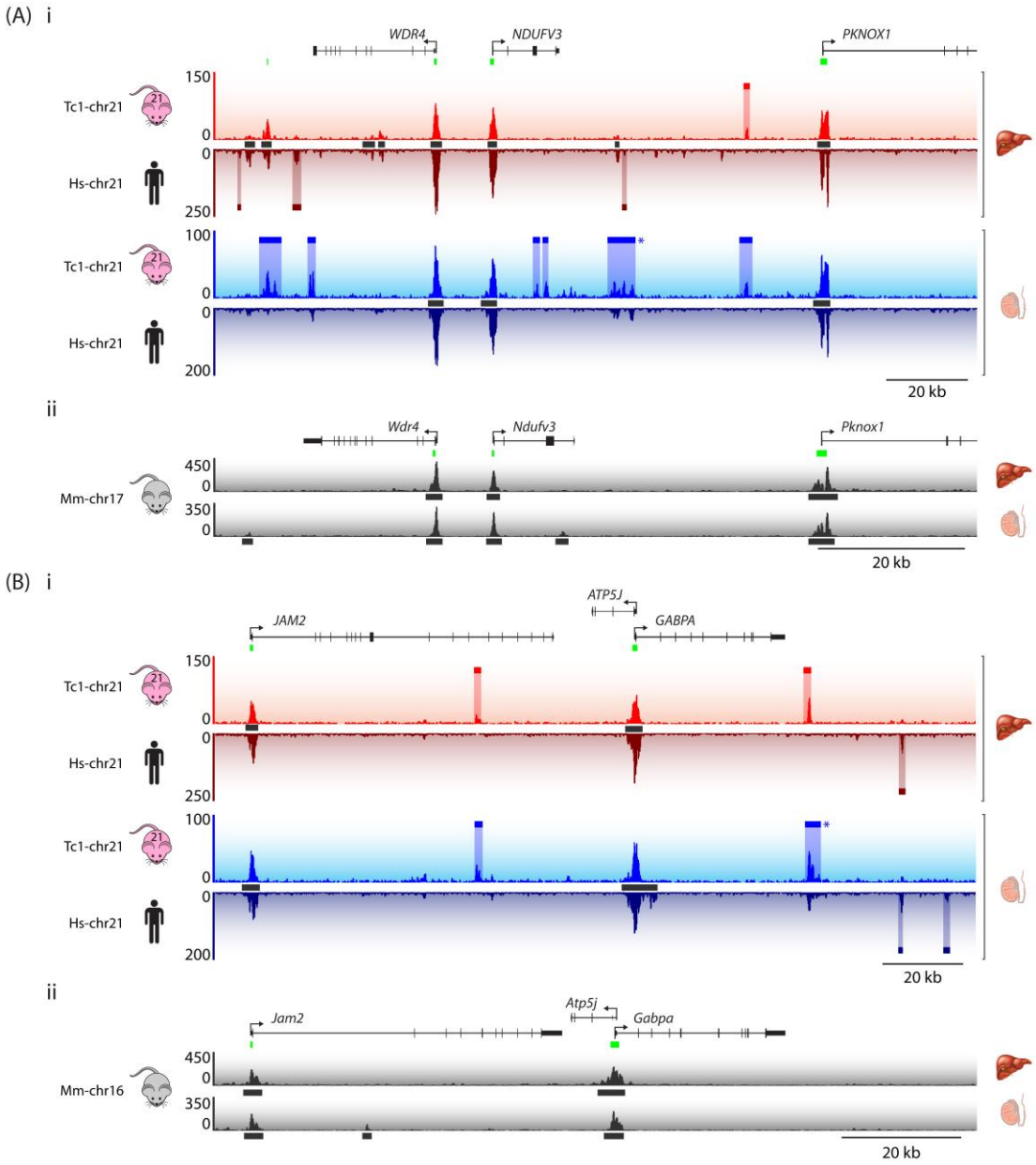


Figure 6.1 Non-methylated DNA profiles from human chromosome 21 in the human and mouse nuclear environment and orthologous regions in the mouse

(A-B, i) Bio-CAP sequencing profiles of non-methylated DNA for two loci from human chromosome 21. Data from liver is shown in red and testis in blue, with Tc1 mouse profiles (brighter shade, Tc1-chr21) shown above the inverted endogenous human profile (darker shade, Hs-chr21). NMIs shared between human and Tc1 for a given tissue are indicated as grey bars between the two traces. NMIs unique to either Tc1 or human (1.5 fold change in normalised read-count) are indicated by a coloured bar with shading to highlight the location. An asterisk highlights two unique NMIs in Tc1-chr21 which appear to exhibit a small amount of non-methylated DNA in Hs-chr21. (A-B, ii) Bio-CAP sequencing profiles for the mouse syntenic region to (i) in wildtype mouse. NMIs are indicated by a grey bar beneath the sequencing trace. Genes are shown above the traces in black, with exons shown as vertical bars and transcription start site as an arrow. CpG island predictions are shown as green bars. The vertical scale on the traces represents Bio-CAP read density. A 20 kb scale bar is shown.

To map NMI location chromosome-wide, and identify regions of differential methylation between the human or Tc1 mouse nuclear environment, peaks were called from Bio-CAP non-methylated

DNA profiles using MACS (v1.4.0) (Zhang et al., 2008) (2.14.5). Peaks were combined for human and Tc1 mouse and were determined to be differentially methylated if they exhibited a greater than 1.5 fold change in Bio-CAP read count (see 2.16.1). This analysis highlighted that a large number of NMIs on human chromosome 21 were shared between human and Tc1 mouse (Figure 6.2A, upper). Furthermore, the majority of NMIs within 1kb of a gene TSS were identified in both nuclear environments (Figure 6.2B, upper). Together, this analysis revealed that human gene promoters do not undergo massive aberrant DNA methylation or demethylation in the mouse nucleus, and instead accurately recapitulate the non-methylated status of gene TSSs from human.

	TSS overlapping an NMI (%)	
	Tissue	
Species	Testes	Liver
Hs	56.3%	62.3%
Tc1	59.3%	64.3%

Table 6.1 *Proportion of human chromosome 21 genes with an NMI overlapping the transcription start site*

The 198 genes retained on chromosome 21 in Tc1 mouse were interrogated for overlap of an NMI at their TSS for Tc1 mouse and human testis and liver tissues.

6.2.3 Chromosome 21 gene promoters in mouse more closely resemble the endogenous human non-methylated state than they do mouse orthologous promoters

For the small number of chromosome 21 genes in the Tc1 mouse which were observed to be differentially methylated compared to human (Figure 6.2B, upper), it is possible that their methylation state reflects that of the orthologous mouse promoter rather than the human counterpart in the same tissue-type. To test this hypothesis, 1:1 mouse orthologues were identified for genes on human chromosome 21. The presence of a TSS-associated NMI was then compared for Tc1-chr21 human orthologues versus either human genes or orthologous mouse genes in their endogenous environment. Interestingly, more differences were observed when comparing TSS-associated NMIs between Tc1-chr21 and the equivalent orthologous gene from wildtype mouse (Mm-chr16, Mm-chr10 or Mm-chr18) than the same comparison with TSS-associated NMIs from Hs-chr21 (Figure 6.3A-B). Therefore, at gene promoters, DNA sequence appears to be able to dictate

the local DNA methylation status, irrespective of the regulatory environment, and define whether a gene promoter is NMI-associated or not. In other words, despite the high degree of conservation of vertebrate NMIs at gene promoters reported in Chapter 4 (Figure 4.9), it appears that a small number of mouse orthologues of NMI-associated human genes have lost the *'strong'* NMI signature associated with TSS NMIs and therefore are methylated in the equivalent mouse tissue (Figure 6.3C).

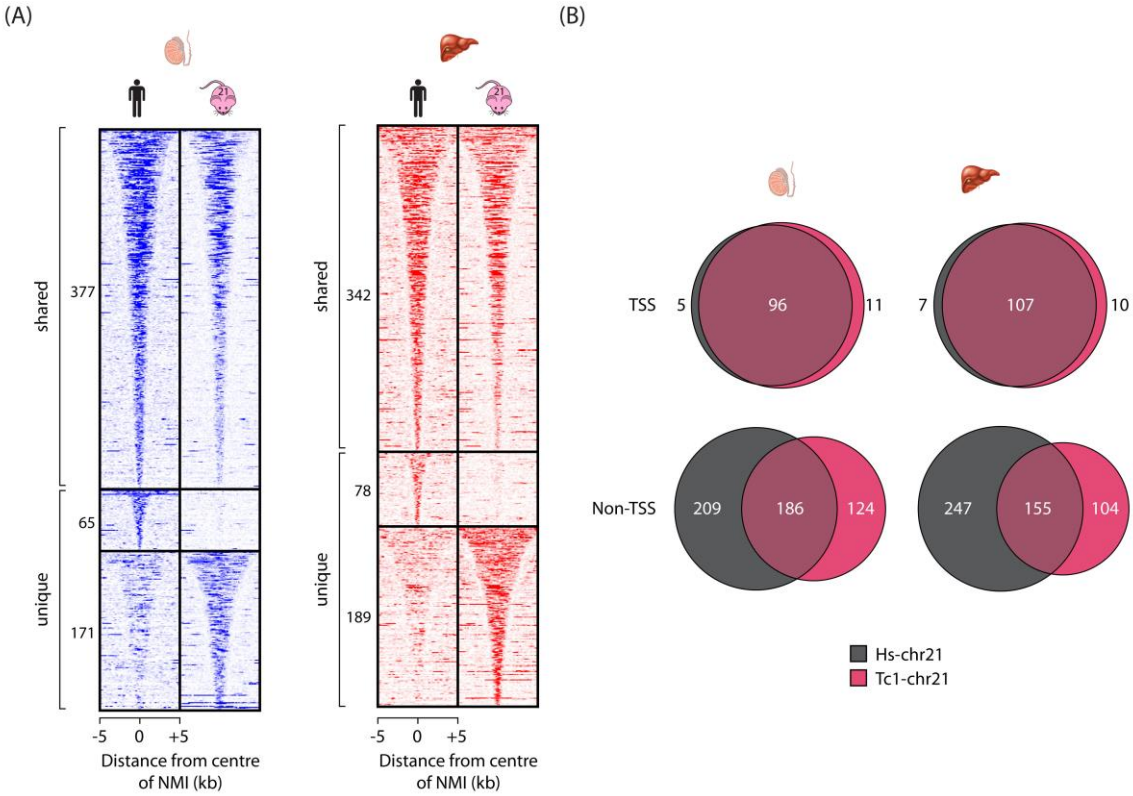


Figure 6.2 TSS-associated non-methylated islands on human chromosome 21 are maintained in the Tc1 mouse

(A) A subset of NMIs on human chromosome 21 are subject to differential methylation as illustrated by heatmaps of non-methylated DNA signal from human chromosome 21 in the human versus mouse nucleus for testis (left) and liver tissue (right). In each case, NMIs are ranked according to length and clustered as shared (upper) or unique (lower) between the two nuclear environments. A 10 kb window centred at each NMI is shown and read density is indicated by colour intensity. (B) Venn diagrams illustrating the overlap of NMIs from human (Hs, grey) and Tc1 mouse (Tc1, pink) on human chromosome 21. Overlaps are depicted for NMIs associated with protein-coding TSSs (within 1 kb of a protein-coding TSS, upper) and for NMIs away from TSSs (greater than 1 kb from a protein-coding TSS). NMIs at TSSs are generally non-methylated in both nuclear environments, while the majority differentially methylated NMIs occur away from gene TSSs.

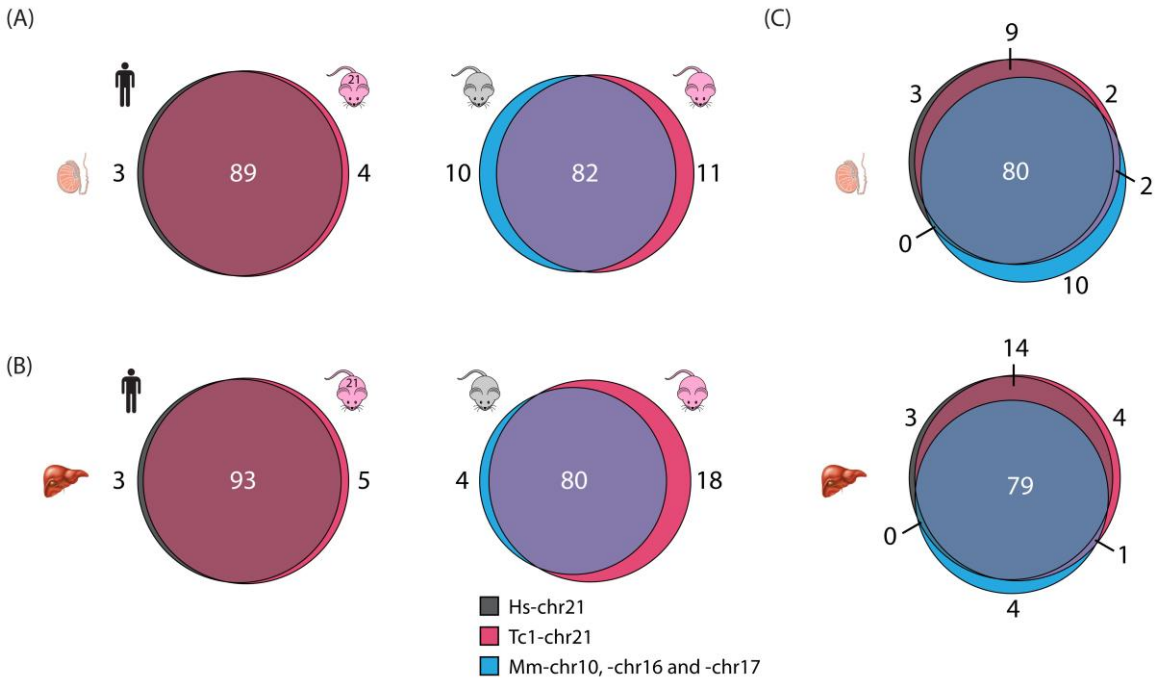


Figure 6.3 Human chromosome 21 TSSs in Tc1 mouse more closely recapitulate the methylation status of human gene TSSs than mouse gene orthologue TSSs

(A) Venn diagrams depict the overlap of gene TSSs associated with an NMI from Tc1-chr21 with TSSs associated with an NMI from endogenous Hs-chr21 (left) or from mouse gene orthologues (Mm-chr10, -16 and -16) for testis tissue (right). Gene TSSs were only analysed if an orthologous gene exists in both human and mouse. (B) Venn diagrams depict overlap of gene TSSs associated with an NMI for liver tissue similar to (A). (C) Three-way Venn diagram overlaps for gene TSSs associated with an NMI from human chromosome 21 in Tc1 mouse (Tc1-chr21, pink), human (Hs-chr21, grey) or from mouse gene orthologues (blue, Mm-chr10, -16 and -16) for testis (upper) or liver (lower).

6.2.4 NMIs away from TSSs on human chromosome 21 are often misinterpreted in the mouse nuclear environment

Similar to human chromosome 21 in its endogenous regulatory environment (a human nucleus), smaller Bio-CAP peaks were observed away from gene promoters, in gene bodies and intergenic regions, for chromosome 21 in the mouse nucleus (Figure 6.1Ai and Bi, Tc1-chr21). Intriguingly, these promoter-distal NMIs often appeared to be differentially methylated between the two nuclear environments (Figure 6.1Ai and Bi). Indeed, a large number of differentially methylated NMIs were observed chromosome-wide in both testis and liver tissue (Figure 6.2A, unique). In agreement with observations made from two genomic loci (Figure 6.1Ai and Bi), a large number of DNA methylation changes appeared to occur away from gene TSSs (Figure 6.2B, lower). Again, it may be possible that the non-methylated DNA profiles observed for Tc1-chr21 may resemble the syntenic regions from the mouse genome, instead of recapitulating the endogenous human profiles. However, this

appeared not to be the case as non-TSS NMIs in Tc1 mouse were not mirrored by similar NMI peaks in the syntenic region in mouse (Figure 6.1Aii and Bii). It is conceivable that human and mouse conserved non-genic sequences (Keightley et al., 2005) have diverged sufficiently so that the sequence motifs necessary for the generation of non-methylated DNA profiles in the mouse nuclear environment have been lost from human chromosome 21. Together, these observations suggest that DNA sequence alone is not sufficient to accurately define the non-methylated status of non-TSS NMIs in the Tc1 mouse model, and therefore other factors such as transcription factors, other DNA-binding factors or chromatin modifications may play a role in the interpretation of the NMI signature.

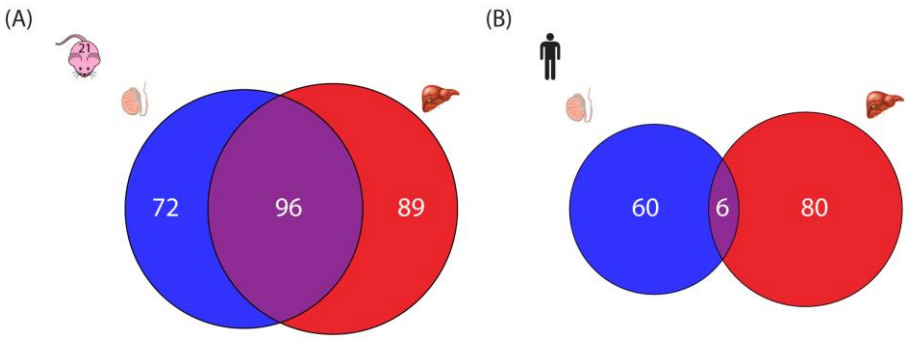


Figure 6.4 NMIs incorrectly methylated on chromosome 21 are not the same in liver and testis tissue

(A) NMIs unique to Tc1 mouse on chromosome 21 (i.e. not identified in human) for liver and testis tissue were compared and are depicted as a Venn diagram. The Tc1 mouse appears to make similar mistakes in both tissues. (B) NMIs unique to human on chromosome 21 (i.e. not identified in Tc1 mouse) for liver and testis tissue were compared and are depicted as a Venn diagram. The Tc1 mouse appears to fail to detect very different NMIs in liver and testis.

6.2.5 Different NMIs are differentially methylated in testis and liver

To examine whether the same alterations in DNA methylation profiles were occurring in testis and liver, unique NMIs for human and for Tc1 mouse were compared for each tissue. Interestingly, it appears that different processes may be occurring at human NMIs which are lost in the mouse nucleus (human-unique NMIs) versus novel NMIs which appear in the mouse nucleus (Tc1-unique NMIs). NMIs which were lost from human chromosome 21 in Tc1 testis and liver are almost completely non-overlapping, suggesting that these NMIs may be dictated by tissue-specific DNA-binding factors in human which are lacking in Tc1 mouse (Figure 6.4B). NMIs unique to Tc1, on the

other hand, appeared to overlap to a higher degree for the two tissues profiled, suggesting that there may be a greater input from nucleotide composition in defining the non-methylated state at Tc1-unique NMIs shared between liver and testis (Figure 6.4A). Alternatively, a transcription factor shared in both tissues in mouse may be responsible for the novel Tc1 NMIs identified in both testis and liver. Overall, human-specific NMIs on chromosome 21 were almost completely distinct for testis and liver, while novel NMIs in the Tc1 mouse tended to overlap between the two tissues.

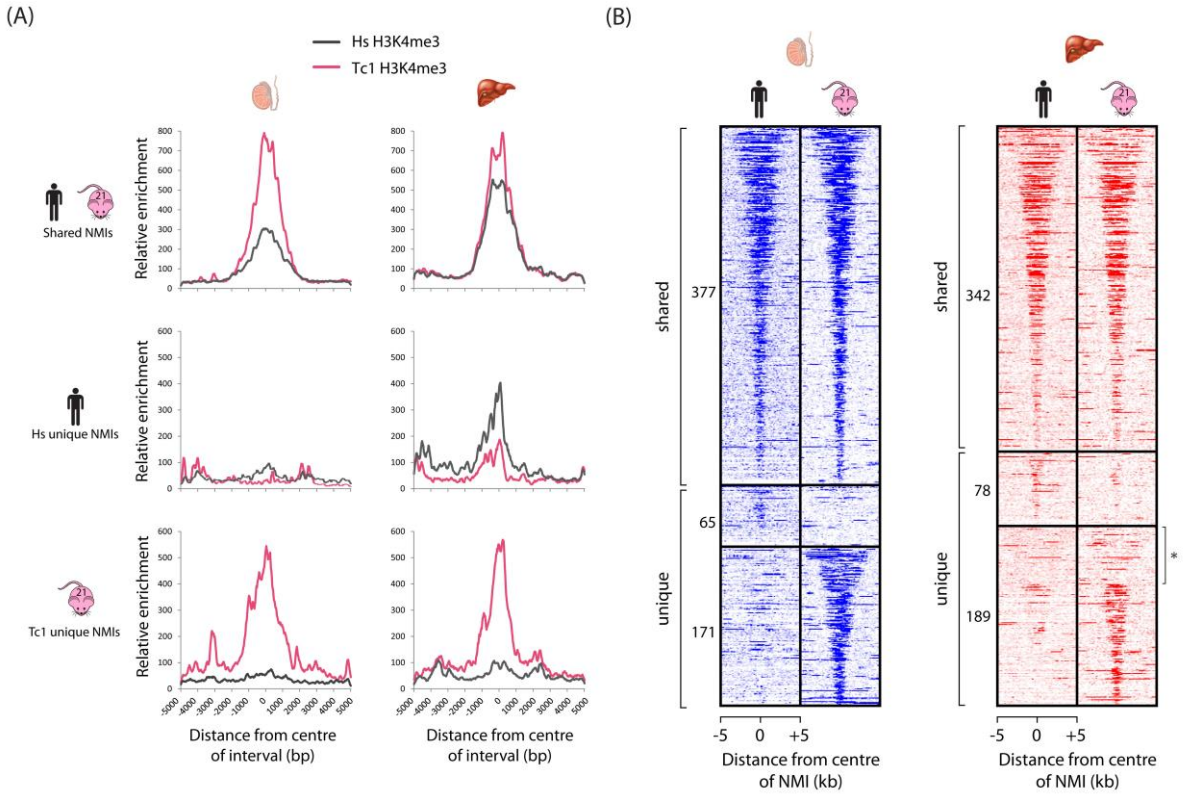


Figure 6.5 The H3K4me3 modification mirrors differential DNA methylation of human chromosome 21

(A) H3K4me3 read density from human (grey) or Tc1 mouse (pink) is profiled over NMIs which are shared between Hs and Tc1 mouse (top), unique to human (middle) or unique to Tc1 mouse (bottom) for testis (left) or liver (right). At differentially methylated loci, which are unique to human or Tc1 mouse, the H3K4me3 modification is found preferentially in the species with the non-methylated NMI. (B) Heatmaps of H3K4me3 ChIP-seq signal from human chromosome 21 in the human versus mouse nucleus for testis (left) and liver tissue (right). In each case, H3K4me3 ChIP-seq signal is plotted over the same intervals as for Figure 6.2A and read density is indicated by colour intensity. NMIs which are subject to differential methylation appear to be marked by H3K4me3 in the tissue in which they are non-methylated, apart from a subset of broader NMIs in liver (marked by an asterisk, *).

6.3 **Changes to the human non-methylated DNA profile impacts chromatin architecture**

6.3.1 The H3K4me3 histone modification segregates with unique NMIs for human and Tc1 mouse

An interesting link can be drawn between differential methylation observed here to occur on chromosome 21 between two distinct species and that observed in the previous chapter to occur between different tissue-types for the same species. In both cases, the majority of DNA methylation changes occurred away from gene TSSs (Figure 5.3B and Figure 6.2B). Previously, differentially methylated NMIs were observed to act as an epigenetic switch to perhaps dynamically regulate the local chromatin environment (Figure 5.8). To examine if the same held true for inter-species differential methylation, available H3K4me3 datasets were profiled over unique and shared NMIs for human and Tc1 mouse (Ward et al., 2013). Intriguingly, a similar pattern was observed whereby NMIs unique to either Hs-chr21 or Tc1-chr21 exhibited unique enrichment for the H3K4me3 modification (Figure 6.5A). Importantly, this switch appeared to occur at the majority of differentially methylated loci, as seen on a heatmap profiling H3K4me3 across shared and unique NMIs (Figure 6.5B).

Interestingly, closer inspection of equivalent heatmaps for non-methylated DNA and H3K4me3 (Figure 6.2A versus Figure 6.5B) revealed that at Tc1-unique NMIs a small amount of non-methylated DNA was detectable for the same positions on chromosome 21 in human (Figure 6.2A). However, no H3K4me3 was detectable for Hs-chr21 at Tc1 unique NMIs (Figure 6.5A). This suggests that in the Tc1 mouse, regions which exhibit low levels of non-methylated DNA in the human nucleus may be erroneously amplified in the mouse regulatory environment (Figure 6.1A-B, two example loci for testis, highlighted by an asterisk). Furthermore, only when fully non-methylated in the Tc1 mouse are these regions sufficient to nucleate detectable levels of H3K4me3 (Figure 6.5B). Together, it appears that differential methylation occurring on human chromosome 21 in distinct nuclear environments can direct a switch-like change in the local chromatin architecture. As the H3K4me3 histone modification is associated with active chromatin and transcriptional initiation (Guenther et

al., 2007), it will be extremely interesting to observe whether associated changes in gene expression are observed near unique NMIs in the Tc1 mouse.

6.4 **Transcription factor binding may be instructive in defining regions of non-methylated DNA**

To investigate what might be driving the observed DNA methylation changes on chromosome 21, the nucleation of DNA binding factors to differentially methylated NMIs was explored. ChIP-seq datasets were available for two liver-specific transcription factors from human and Tc1 liver tissue: CCAAT/enhancer-binding protein alpha (CEBPA) and hepatocyte nuclear factor 4 alpha (HNFA) (Ward et al., 2013). As might be expected, a large number of differences in transcription factor occupancy were observed between human and Tc1 mouse, perhaps due to changes in transcription factor binding site preferences (Figure 6.6A). Strikingly, a number of the species-specific binding events for either or both CEBPA and HNF4A were observed to coincide with a species-specific NMI. For example, a human-unique CEBPA and HNF4A binding event in intron 12 of the *WDR4* gene was observed to coincide with a unique peak of non-methylated DNA in human mouse liver (Figure 6.6B). Similarly, a Tc1-unique NMI at the TSS of the *LOC339622* gene overlapped a Tc1-unique HNF4A binding event in liver (Figure 6.6C). Only a minority of species-specific NMIs were observed to coincide with a species-specific transcription factor binding event for human (18.4 %) and Tc1 mouse (17.0 %) (Figure 6.6D and E). However, it is actually quite striking that such a large proportion of differentially methylated NMIs can potentially be accounted for by only two transcription factors. In fact, a much broader correlation between transcription factor binding and non-methylated DNA may have been observed if many more transcription factors ChIP-seq datasets had been available. Overall, the correlation between species-specific NMIs and species-specific transcription factor binding events suggests that transcription factor binding could be related to the non-methylated status of DNA as has been suggested previously for mammals (Stadler et al., 2011; Thomassin et al., 2001).

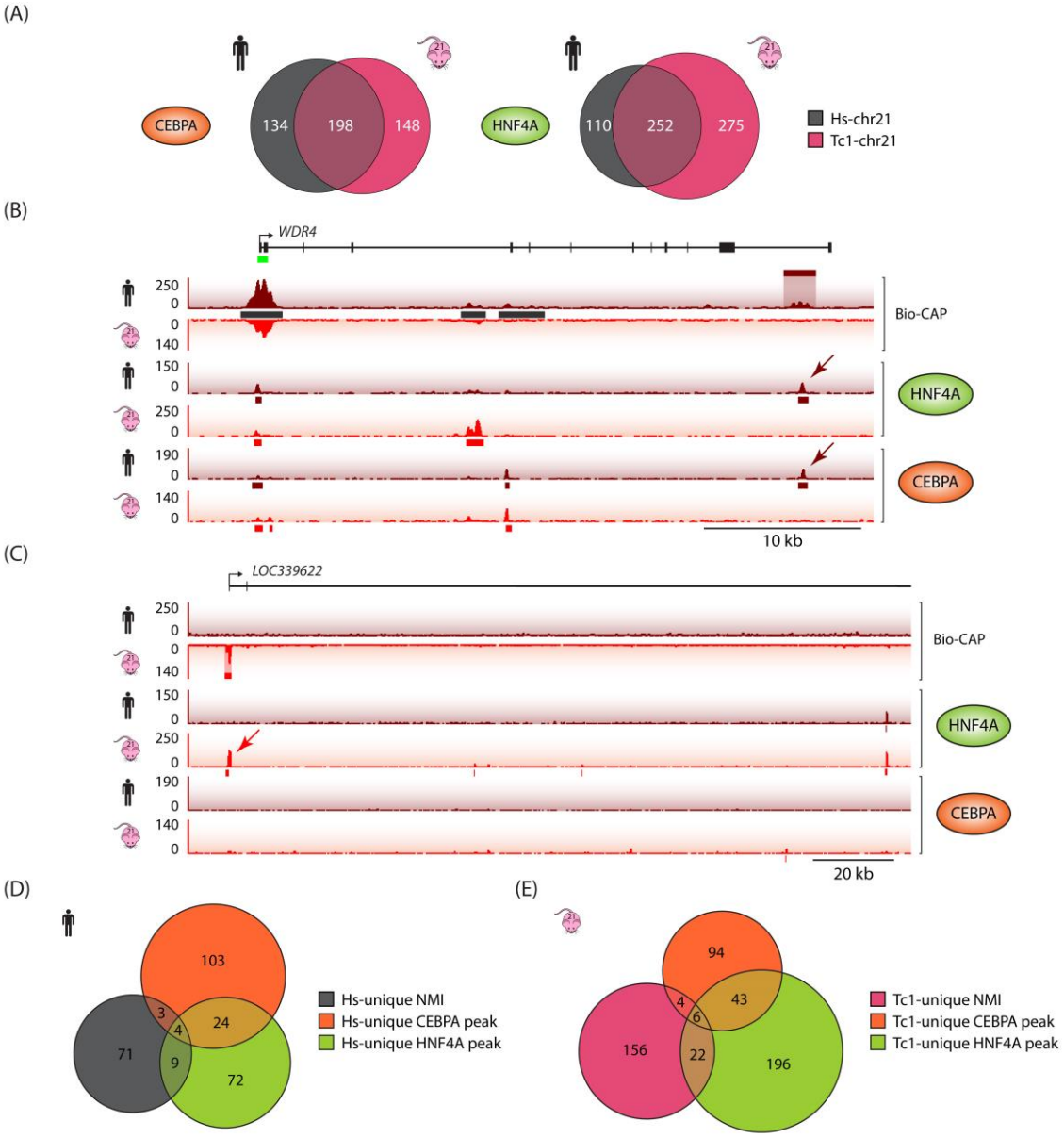


Figure 6.6 Transcription factor binding sites vary between the human and mouse regulatory environment

(A) Venn diagrams illustrating the overlap of ChIP-seq peaks for the transcription factors CEBPA (left) and HNF4A (right) between human and Tc1 mouse on human chromosome 21. Peaks were identified from ChIP-seq data from two replicate datasets and only replicated intervals were used for analysis. (B-C) Examples of unique NMIs from human (B) or Tc1 mouse (C) which exhibit a peak of HNF4A and/or CEBPA binding (arrow) at the same location only in the species exhibiting a unique NMI. (D) A Venn diagram depicting the overlap of human unique NMIs (Hs-unique NMI, grey) with CEBPA and HNF4A ChIP seq intervals (orange and green) which were unique to human and not identified in Tc1 mouse. (E) A Venn diagram depicting the same analysis as (D) for features unique to the Tc1 mouse. Datasets for CEBPA and HNF4A from human and Tc1 mouse were publically available (ENA - ERP000054).

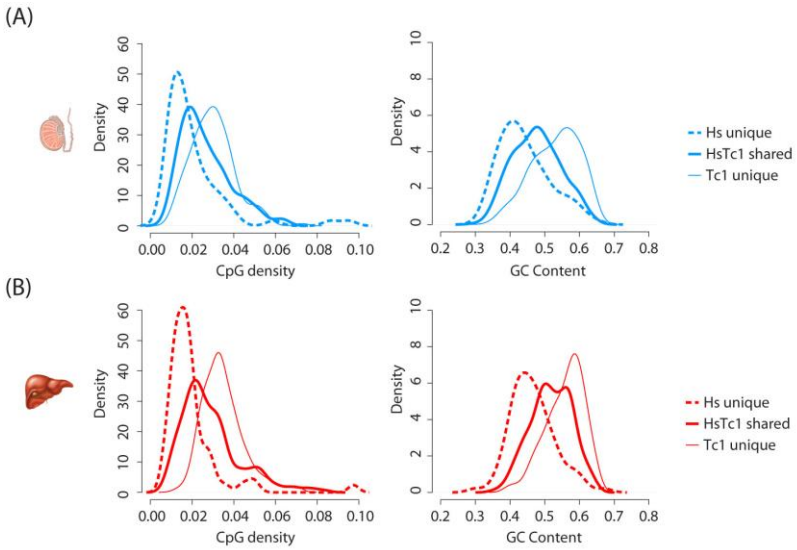


Figure 6.7 Nucleotide properties vary at NMIs which are shared or unique in different regulatory environments

(A) CpG density (left) and GC content (right) are depicted as density plots for non-methylated islands from human chromosome 21 which are shared between human and Tc1 mouse (HsTc1 shared, bold line) or are unique to either human (Hs unique, dashed line) or Tc1 mouse (Tc1 unique, thin line) in testis tissue. (B) As for (A), nucleotide properties of NMIs from liver are depicted as density plots.

6.5 **Shared or unique NMIs between human and Tc1 mouse exhibit distinct nucleotide composition**

The link between transcription factor binding and non-methylated DNA dynamics appears compelling, but may not account for all changes in DNA methylation on human chromosome 21 in the human versus mouse regulatory environment. In order to investigate the potential role of sequence composition in defining DNA methylation status, CpG density and GC content were investigated. Intriguingly, NMIs unique to either human or Tc1 mouse have distinct nucleotide properties compared to shared NMIs. The class of NMIs unique to Hs-chr21 exhibited an overall reduced CpG density and GC content (Figure 6.7A-B, dashed line). Conversely, Tc1-unique NMIs tended to encode DNA sequences which are enriched for CpG density and GC content (Figure 6.7A-B, thin line), suggesting that these regions may be incorrectly recognised in the Tc1 nucleus as having an NMI signature due to their strong nucleotide properties. Together, NMIs on human chromosome 21 which are incorrectly methylated or non-methylated in the Tc1 nucleus exhibit nucleotide properties which are divergent from those which are accurately non-methylated (Figure 6.7A-B, bold line). This may suggest that the human and mouse nuclear environments interpret CpG density and

GC content differently. Alternatively, the loss of human-specific DNA-binding events across chromosome 21 in the Tc1 mouse regulatory landscape may release CpG dense regions to become non-methylated, while CpG poor regions, which may be bound in human by protective DNA-binding factors to keep them non-methylated, become methylated in the Tc1 mouse.

6.6 **Introduction of zebrafish DNA into the mouse nuclear environment**

6.6.1 Generating stable mouse ES cell lines containing zebrafish BAC DNA

Non-methylated profiles of DNA were observed to be faithfully recapitulated for human gene promoters in the mouse nuclear environment, despite 75 million years of evolutionary distance (Figure 6.2B, upper). To observe whether similar levels of promoter methylation fidelity could be achieved over even longer distances of evolutionary time, 200 kb of zebrafish DNA was introduced into the mouse genome (Figure 6.8). Therefore the DNA substrate and nuclear regulatory environment are around 416 million years diverged (Ponting, 2008). A zebrafish bacterial artificial chromosome (BAC) was engineered to contain a Kanamycin/Neomycin selectable marker and a transposable element to aid genomic integration (Suster et al., 2009) (see 2.9.6, Figure 6.8A and B). Stable mouse ES cell lines were generated by transfection of the BAC (Figure 6.8C) and selection with G418, followed by isolation of individual colonies which were screened for successful BAC integration into the mouse genome (see 2.9.10). Three positive clones were expanded and genomic DNA from these lines was extracted and subjected to Bio-CAP sequencing (2.13.2.2).

6.6.2 The zebrafish promoter-associated non-methylated signature is recognised in the mouse nucleus

The DKEY-5K13 BAC (named zBAC3 here) corresponded to chr13:31,858,931-32,077,313 from the zebrafish genome, and contained seven genes, six of which exhibited a TSS-associated NMI in zebrafish 24 hpf tissue, with two further intergenic NMIs detected upstream of the *prkcha* and *hif1a* genes (Figure 6.9A) (almost identical Bio-CAP profiles were observed for liver and testis tissue, data not shown). All three stable ES cell clones had integrated at least one copy of the BAC, as detected

by massively parallel sequencing (Figure 6.9B), although zBAC3-17 appeared to have lost parts of the BAC at or around the *hif1a* gene. zBAC3 DNA appeared to closely recapitulate the zebrafish non-methylated DNA profiles in mouse ES cells, with five of the six TSS-associated NMIs and both of the intergenic NMIs (3 and 5) being identified as non-methylated in the zBAC3-9 clone. Bio-CAP signal was detected at gene promoters for the two remaining clones zBAC-16 and 17 (Figure 6.9B). Interestingly, four novel non-TSS NMIs were identified for the zBAC3-9 clone (labelled 7-10, Figure 6.9B).

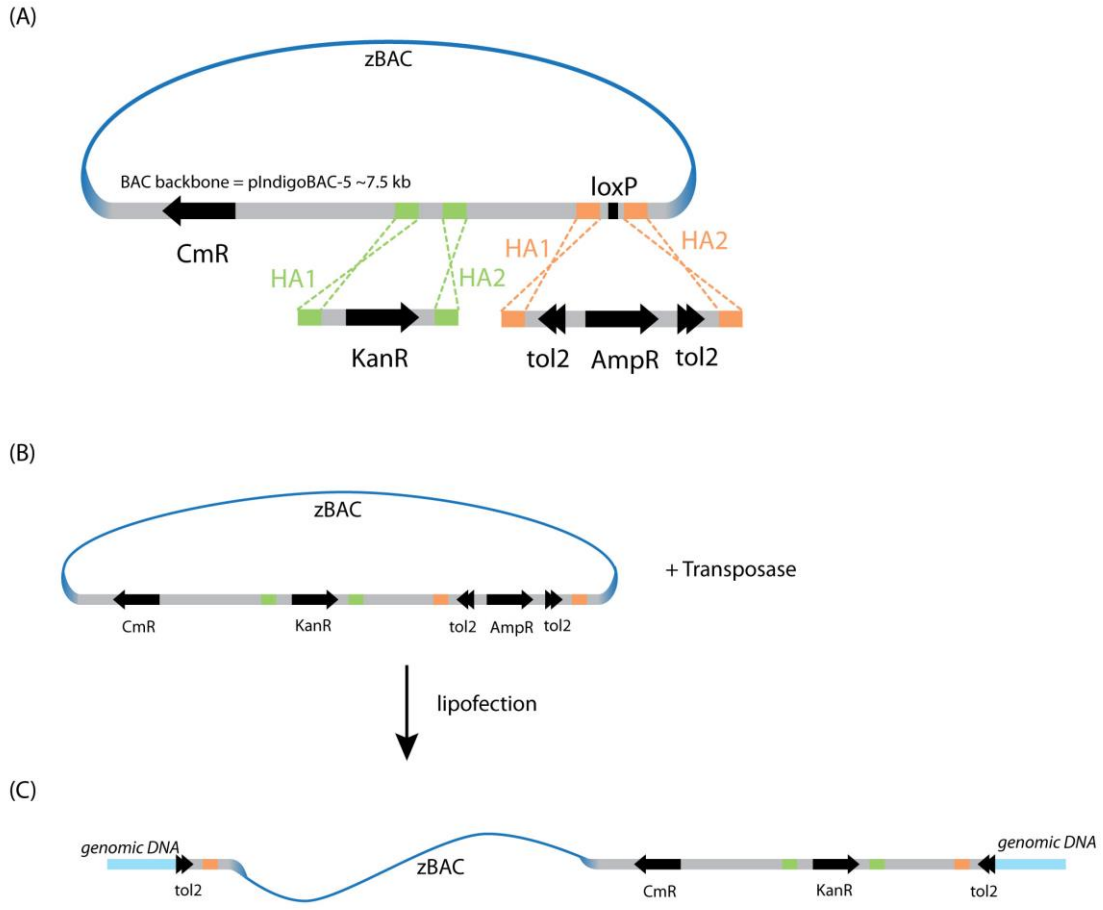


Figure 6.8 Recombineering a zebrafish bacterial artificial chromosome for insertion into the mouse genome

(A) Zebrafish bacterial artificial chromosomes (zBACs) were selected from the DanioKey (DKEY) BAC library in which around 200 kb of zebrafish genomic DNA had been cloned into the pIndigoBAC536 vector backbone. Primers were designed to amplify inverted tol2 arms flanking the Ampicillin resistance cassette (AmpR) with homology arms to the pIndigoBAC536 vector surrounding the single loxP site (orange, HA1 and HA2). Primers were also designed to amplify the Kanamycin resistance cassette (KanR) with homology arms elsewhere in the pIndigoBAC536 vector (green, HA1 and HA2). (B) Following the two recombineering steps, the zBAC vector backbone contained the newly inserted KanR cassette and AmpR cassette flanked

by inverted tol2 arms. (C) The Tol2 arms enabled transposition of the zBAC into the mouse genome following transfection of the zBAC along with DNA encoding the tol2 transposase enzyme.

6.6.3 Novel zebrafish NMIs in the mouse nucleus exhibit distinct nucleotide properties

To determine what might be driving the non-methylated DNA profiles of zebrafish DNA in the mouse nuclear environment, underlying nucleotide properties were investigated for NMIs identified from 24 hpf zebrafish and novel non-methylated regions from the zBAC3-9 clone. GC content appeared to share the same distribution for the endogenous (1-6) versus novel (7-10) zebrafish NMIs whereas CpG density and length appeared to be generally lower for the novel NMIs identified in the mouse nucleus (Figure 6.9D, black and Appendix 6.2). These observations are likely to be skewed, however, as the novel NMIs were all intergenic compared to a majority of TSS-associated NMIs identified for zebrafish 24 hpf. Nucleotide properties of NMIs within the syntenic region of the mouse genome were also analysed (Figure 6.9C and Appendix 6.2). In general, mouse ES cell NMIs from this region (NMIs 11-22) exhibited higher CpG density, GC content and length compared to those from zebrafish (Figure 6.9D). Therefore, it is possible that gene promoters in zebrafish exhibit similar enough nucleotide properties to the NMIs in mouse for them to be recognised as non-methylated domains, alternatively it may be the presence of the transcriptional machinery, or the act of transcription at gene promoters, which enables zebrafish promoter-associated NMIs to be recapitulated in the mouse nucleus. However, the novel zebrafish NMIs (Figure 6.9B and D, black) exhibited even lower CpG density and GC content than the zebrafish gene promoters; therefore it seems unlikely that nucleotide composition can lead to these elements becoming non-methylated in the mouse nucleus. Instead, perhaps random acquisition of mouse transcription factor binding sites in the zebrafish genome may act to drive NMI formation when introduced into the mouse nucleus.

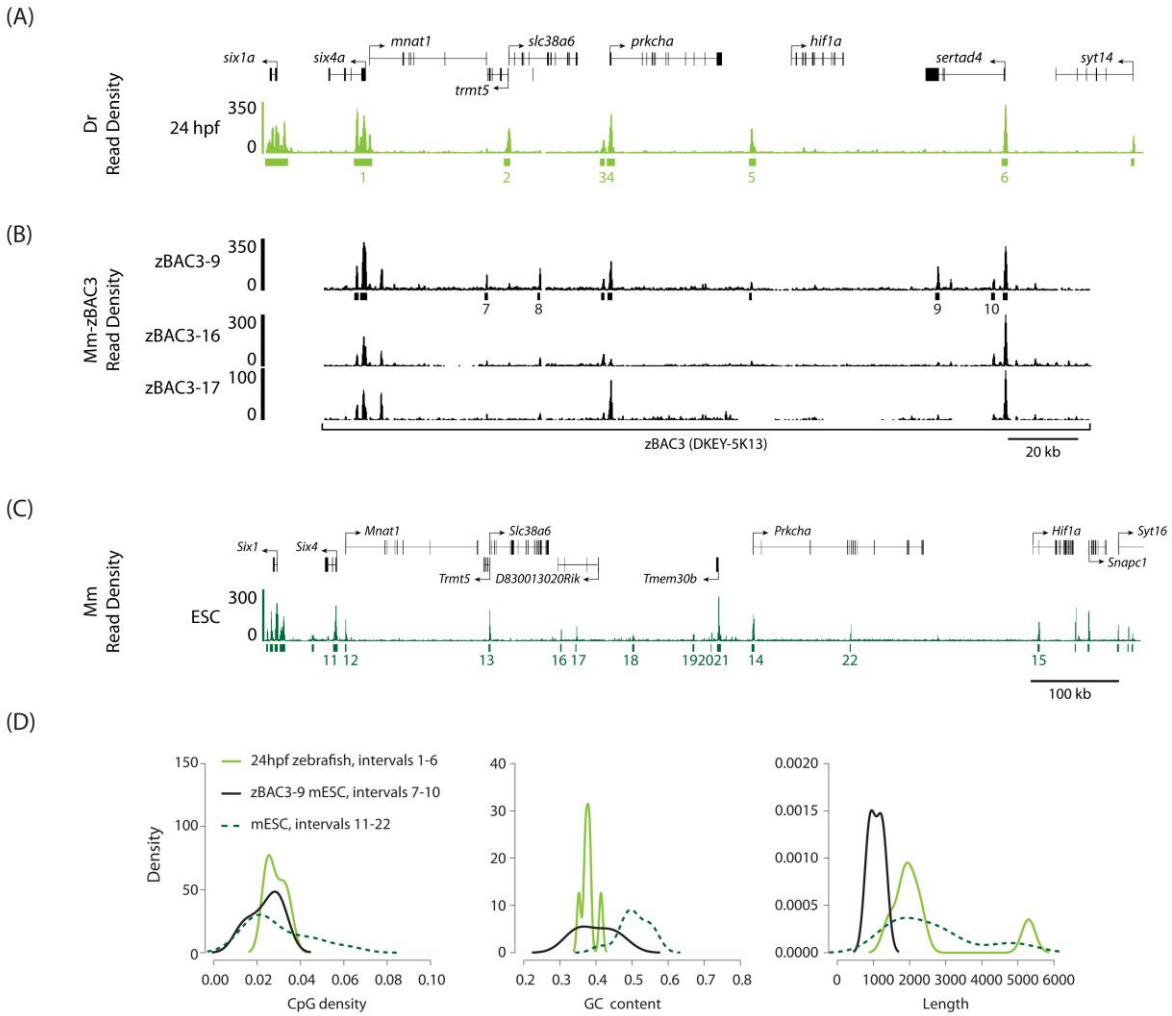


Figure 6.9 Zebrafish non-methylated islands are partially recapitulated in the mouse nucleus

(A) Non-methylated DNA profiles for zebrafish embryonic 24 hours post-fertilisation (hpf) at the genomic region chr13 : 31,836,400 - 32,092,497, encompassing the region corresponding to the bacterial artificial chromosome DKEY-5K13 (named zBAC3). Peak intervals are depicted for 24 hpf zebrafish Bio-CAP sequencing (green, labelled 1-6) over the region corresponding to the BAC (chr13:31,858,931-32,077,313). (B) Non-methylated DNA profiles for zBAC3 DNA which was incorporated into the mouse genome. Three distinct clones are depicted: zBAC3-9, zBAC3-16 and zBAC3-17. Peak intervals which are distinct from those detected in the endogenous environment (zebrafish 24 hpf) are depicted for clone zBAC3-9, labelled 7-10. (C) Non-methylated DNA profiles for mouse ES cell is shown across the syntenic locus for zBAC3 (chr12:74,028,645-75,149,698), from the Six4 to Hif1a gene. Bio-CAP peaks 11-15 correspond to orthologous zebrafish promoter NMIs, while peaks 16-22 have no clear orthologous regions in zebrafish. (D) CpG density, GC content and length are depicted as density plots for NMIs from the zBAC3 region in zebrafish embryonic stage 24 hours post fertilisation (hpf, light green line) or are unique to zBAC3-9 integrated into mouse ESCs (black line). Density distributions are also depicted for intervals 11-22 from mouse ES cells (dark green dashed line).

6.7 **Discussion**

6.7.1 Non-methylated DNA maps for exogenous DNA sequences in the mouse nucleus

The factors which define patterns of non-methylated DNA and which are able to direct differential methylation of the same DNA sequence remain unclear. Here, profiles of non-methylated DNA were generated for exogenous DNA sequences introduced into a heterologous regulatory environment, enabling investigation of important determinants of non-methylated DNA. Analysis of 42 Mb of human and 200 kb of zebrafish DNA sequence in Tc1 mouse and mouse ES cells respectively provides tentative insights into the establishment of DNA methylation profiles between a regulatory environment which has diverged from the DNA sequence elements being regulated. Overall, it appears that different modes of action are at play at gene promoters, which appear to be accurately recognised and appropriately non-methylated, versus non-TSS NMIs which are less faithfully recapitulated in a foreign regulatory environment.

6.7.2 Maintenance of the non-methylated state at gene promoters

Strikingly, for both heterologous systems presented here, promoter NMIs appeared to recapitulate endogenous non-methylated DNA profiles suggesting that very strong signals have been maintained at vertebrate gene TSSs to define them as non-methylated. The local nucleotide composition may play an important role in dictating the non-methylated state at gene promoters, perhaps by virtue of CpG dinucleotide density. Indeed, the TSS-associated class of NMIs have an elevated CpG density compared with non-TSS NMIs, therefore one possibility is that these regions nucleate a high concentration of ZF-CxxC-containing non-methylated DNA binding proteins which act to demarcate these regions to be constitutively non-methylated. Alternatively, the presence of the mouse transcriptional machinery at human or zebrafish gene promoters may protect TSS NMIs from methylation. Whatever the main determinant, the strong NMI signature appears to have been maintained in diverse vertebrate species over evolutionary time as a conserved feature of vertebrate gene promoters.

6.7.3 Non-methylated islands away from gene promoters do not recapitulate the endogenous methylation state

Interestingly, a large number of DNA methylation alterations were observed on human chromosome 21 between the human and mouse nucleus (Figure 6.2A), with the majority of these changes occurring away from gene TSSs (Figure 6.2B). An interesting link can be drawn here to observations made in the previous chapter in which differentially methylated NMIs between distinct tissue-types were observed to occur predominantly away from gene TSSs (Figure 5.3B). Analysis of these differentially methylated NMIs revealed that NMIs unique to either human or Tc1 mouse appeared to experience distinct influences from nucleotide composition and transcription factor binding to define their DNA methylation state.

6.7.3.1 **A proportion of human NMIs are not recapitulated in the Tc1 mouse**

A number of NMIs were identified on human chromosome 21 which were observed to be non-methylated in human but were not detected in the Tc1 mouse (Hs-unique, Figure 6.2B). Interestingly, these NMIs were almost completely different between testis and liver tissue (Figure 6.4B) suggesting that these NMIs in human may be bound by testis or liver specific DNA binding factors which initiate and maintain their non-methylated state. These NMIs exhibit low CpG density and GC content (Figure 6.7) and therefore perhaps, without the appropriate DNA binding factor in the Tc1 mouse, these regions become methylated and lose H3K4me3 modification. Examples of associated Hs-unique transcription factor binding events and Hs-unique NMIs were observed for HNF4A and CEBPA in human liver (Figure 6.6B and D). Recent experimental evidence has shown that the introduction of a CTCF or REST binding site can lead to the loss of DNA methylation at a given region (Stadler et al., 2011). However, these binding factors exhibit strong DNA binding activity and have relatively long recognition sequences (Bruce et al., 2009; Phillips and Corces, 2009), therefore it remains to be seen whether transcription factors with shorter consensus sequences are also able to drive loss of DNA methylation. Overall, the analysis here suggests that DNA binding can drive loss of DNA methylation at Hs-unique NMIs (model Figure 6.10A).

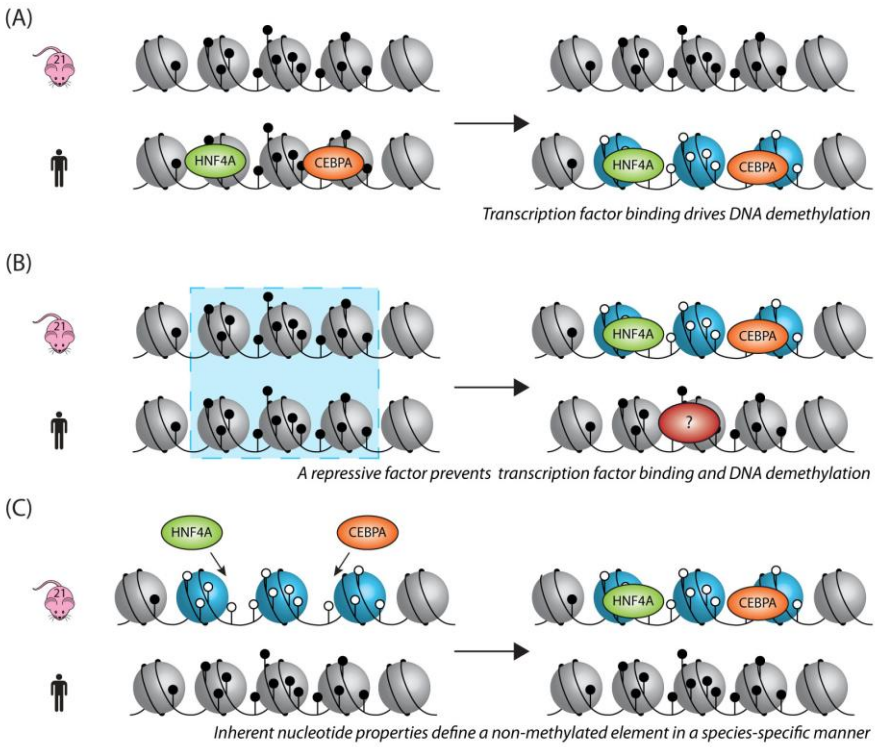


Figure 6.10 Potential models for differential methylation of non-TSS NMIs between human and Tc1 mouse

(A-C) Models to explain the differential methylation observed on human chromosome 21 between the human and mouse nucleus. (A) Transcription factors in the human nucleus may be able to bind DNA sequences which are not bound in the equivalent tissue-type in the Tc1 mouse nucleus (left). This binding event may drive the region to become non-methylated in the human nucleus (right). (B) A given DNA sequence may be susceptible to becoming non-methylated in both the human and mouse nucleus (left, blue box). However, there may be a repressive factor present in human (labelled "?") which prevents this region from becoming non-methylated (right), while the region is free to become non-methylated in mouse in the absence of the repressive factor. (C) Thirdly, a given region of human DNA may become non-methylated in the mouse nucleus, perhaps due to recognition of sequence properties (such as CpG density) in the mouse nuclear environment which were not detected in human (left). This novel non-methylated region may then be more open and accessible for binding of transcription factors in the mouse compared to the same DNA region in human (right).

6.7.3.2 **Novel NMIs at heterologous DNA sequences in the mouse nucleus**

A large number of Tc1-unique NMIs were detected on human chromosome 21 in the mouse nucleus. However, on closer inspection, these NMIs appeared to in fact to be non-methylated in the equivalent tissue for human but to a much lesser extent (Figure 6.2A). Therefore a pre-existing partially non-methylated signature in human may be amplified in the mouse nucleus. The partially non-methylated regions in human may be poised to become activated in a different tissue in human, or when presented with a certain stimulus. Perhaps there is a human-specific activity which is suppressing these regions from becoming fully non-methylated in human as they exhibit features

(high CpG density and GC content) normally associated with constitutively non-methylated regions. These regions may therefore become non-methylated in Tc1 mouse, where this repressive signal (perhaps a DNA-binding factor or chromatin modification landscape) is presumably not present (model, Figure 6.10B). An alternative possibility is that in human, the nucleotide composition at these loci is only sufficient to cause the region to become partially non-methylated whereas in the mouse nucleus (where NMIs tend to have lower CpG density and GC content, see Figure 4.7B) the nucleotide composition is sufficient to drive them to become fully non-methylated (model, Figure 6.10C). As a large number of Tc1-unique NMIs were observed to overlap between testis and liver (Figure 6.4A), either option is possible as tissue-specific activities would not be required to direct the observed DNA methylation differences.

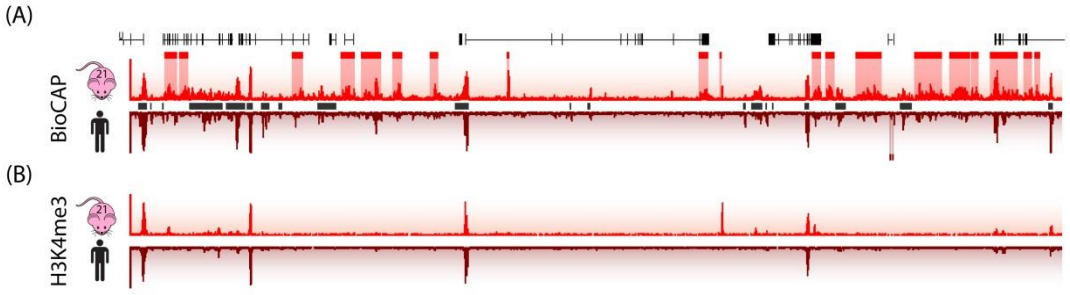


Figure 6.11 Subtelomeric broad hypomethylated regions in Tc1 liver are not marked by H3K4me3

(A) Bio-CAP non-methylated DNA profiles for Tc1 and human liver (chr21:46,265,920-46,868,370) depicting a large number of hypomethylated regions for Tc1 mouse illustrated by the large number of unique Tc1 NMIs (red bars above sequencing traces). Shared NMIs are indicated as grey bars between the Tc1 and human Bio-CAP traces. (B) H3K4me3 sequencing profiles for human and Tc1 liver. The broadly hypomethylated regions of DNA are not broadly modified by H3K4me3.

6.7.3.3 **Broadly hypomethylated regions in the Tc1 mouse do not acquire H3K4me3**

Of note, a subset of Tc1-unique NMIs appeared to escape H3K4me3 modification in liver (Figure 6.5B, asterisk). This is perhaps surprising as non-methylated DNA is known to recruit ZF-CxxC proteins, including CFP1 which mediates the placement of H3K4me3 (Thomson et al., 2010). These hypomethylated regions were clustered on the right arm of chromosome 21 near the telomere. Therefore it is possible that other heterochromatic features may compensate for the loss of DNA methylation which may then preclude modification of H3K4me3 (Figure 6.11). These chromatin

changes may include compaction, H3K9me3, H3K27me3 and histone hypoacetylation, to prevent activation of the subtelomeric region of the chromosome. Future work will include investigation of these possibilities.

6.7.4 The contribution of DNA sequence in defining non-methylated DNA

Fragments of *E. coli* DNA, with CpG density equivalent to that at mouse promoters, introduced into the mouse genome were shown to establish a hypomethylated region (Lienert et al., 2011). Therefore, in some instances CpG density appears to be sufficient to establish the non-methylated state independent of transcription factor binding events. This may be the case at gene promoters where a high CpG density may contribute to the establishment of the non-methylated state between tissues and different nuclear environments, although this is confounded by the presence of many transcription factor binding events. For NMIs with weaker nucleotide signature (lower CpG density and GC content), their DNA sequence properties may render them susceptible to changes in DNA methylation. In other words, these regions may have sufficient CpG density to functionally constitute an NMI, however the CpG content is low enough to be readily methylated upon loss of an activating factor or gain or a repressive factor. This is evidenced in the Tc1 mouse model, when in the absence of human specific transcription factors, a number of NMIs with low CpG density became methylated in the Tc1 mouse (Figure 6.7A-B, dashed line).

Conversely, in some cases, nucleotide composition may be sufficient to drive the non-methylated state in distinct nuclear environments. There is evidence for this again in the Tc1 mouse as Tc1-unique NMIs exhibited a high CpG density (Figure 6.7A-B, thin line), suggesting that perhaps the mouse nuclear environment has a different threshold for nucleotide composition for establishing the non-methylated state. Alternatively, a repressive factor may not be present in the mouse to suppress non-methylated DNA at these loci. Together, CpG density above a defined threshold appears sufficient to establish a NMI, while intermediate nucleotide properties appear to facilitate

the establishment of a non-methylated domain, but these regions appear to have increased susceptibility to DNA methylation.

6.7.5 The role of DNA-binding factors in defining non-methylated DNA

When an exogenous DNA is introduced into a heterologous regulatory environment, it will be exposed to a "non-self" transcription factor repertoire, which may be responsible for driving some of the DNA methylation differences observed here (Figure 6.2A). Firstly, a different complement of transcription factors may be expressed in equivalent tissues in human versus mouse and therefore create a distinct transcription factor binding landscape. This could be addressed by comparing available RNA-seq data from equivalent tissues from human and mouse (Brawand et al., 2011; Stadler et al., 2011). Alternatively, over evolutionary time, a subset of human and mouse transcription factors may have diverged at their DNA binding interface and have acquired slightly different consensus binding sequences (Figure 6.6A). In either case, differential transcription factor binding events may be important for driving differential methylation between Hs-chr21 and Tc1-chr21.

Counter to this argument, a large proportion of novel transcription factor binding sites for both Hs-chr21 and Tc1-chr21 did not overlap a novel NMI (Figure 6.6D and E). For these cases, it may be that the DNA binding events occur at CpG poor regions which, even if non-methylated, would not be detectable by Bio-CAP-seq. Alternatively, Tc1-specific transcription factor binding events may occur in shared NMIs at DNA-binding consensus sequences not utilised in human. Interestingly, transcription factor binding sites tend to be very common, occurring up to thousands of times throughout a genome, however only a minority tend to be bound in given tissue. Those sites which are utilised are often found within a CpG island (Fernandez et al., 2003; Zeller et al., 2006), suggesting either that the successful binding of the transcription factor was able to drive demethylation (Figure 6.10A) or that the binding event required that the locus be non-methylated (Figure 6.10C). Regardless, it is unlikely that all transcription factor binding events will be able to

drive the assembly of an NMI, and similarly that all NMIs are dictated by a DNA binding event. However, for pioneer transcription factors, which are able to bind on nucleosomes and often to methylated DNA (Cairns, 2009; Sérandour et al., 2011; Wiench et al., 2011; Xu et al., 2009), this event may provide the mechanism for driving a change in the DNA methylation landscape in a distinct tissue-type or heterologous regulatory milieu.

6.7.6 Future directions

6.7.6.1 **Gene expression changes on human chromosome 21**

As mentioned previously, RNA-seq datasets are available for human and Tc1 mouse; therefore it will be interesting to observe whether loss or gain of non-TSS NMIs in the Tc1 mouse impacts nearby gene expression. Functional annotation of these NMIs (using available histone modification data for human) would then reveal whether these NMIs normally act as distal regulatory elements in any tissue context for human. Furthermore, for genes which are incorrectly repressed, it would be very interesting to observe whether they have acquired polycomb repression perhaps due to lack of an appropriate activating factor, by analysing genome-wide profiles of H3K27me3 in Tc1 testis and liver tissues.

6.7.6.2 **What happens to the repetitive portion of chromosome 21 in the mouse nucleus?**

The Tc1 analysis performed in this chapter was limited to interrogation of the non-repetitive portion of the mouse genome. Odom and colleagues recently revealed that the Tc1 mouse exhibits active histone modifications at primate-specific repetitive elements and that these novel active regions are associated with activation of nearby genes (Ward et al., 2013). Locus-specific analysis revealed an association of these activated regions with DNA methylation changes. Therefore a more thorough chromosome-wide analysis of DNA methylation in repetitive regions may reveal more global links between DNA methylation, chromatin modification and transcription in the repetitive fraction of the genome.

6.7.6.3 **Can overexpression of an exogenous factor drive DNA methylation change?**

To investigate the role of transcription factors in the endogenous regulation of NMIs, a given transcription factor could be depleted and DNA methylation interrogated at sites previously bound by the transcription factor. However, loss of transcription factor activity may also have secondary effects on DNA methylation due to changes in expression of genes regulated by the depleted transcription factor. Furthermore, once a non-methylated domain is established it may be maintained by a distinct mechanism and no longer require the DNA binding capacity of the initiating transcription factor. Alternatively, overexpression of an exogenous DNA binding factor from a distantly related species (for example zebrafish) in mouse ES cells may be more informative to observe whether novel NMIs appear genome-wide at putative DNA binding sites for this factor. The data generated in this chapter may already be able to address this given that the *Six4a* gene on zBAC3 encodes a homeobox transcription factor. If this gene is expressed from the zBAC3 integration site, and the zebrafish DNA binding motif is sufficiently different to that of the mouse protein, then novel NMIs on mouse chromosomes could be interrogated to see whether they have an underlying homeobox motif. In future work, this factor could be tagged to enable ChIP experiments to verify DNA binding at these loci. Further to this, motif-searching could more generally be applied to the available Tc1 and zBAC3 data to determine whether novel NMIs are enriched for mouse or zebrafish transcription factor binding sites.

Together, the work presented in this chapter provides an intriguing insight into the regulation of DNA methylation patterns in vertebrate genomes. The future work proposed here will help to uncover further links between transcription, chromatin modifications and DNA binding factors in the regulation of the non-methylated DNA landscape.

Chapter 7 -- Investigating the functional importance of non-methylated islands during early zebrafish development

7.1 **Introduction**

7.1.1 Understanding the biological importance of non-methylated DNA

In this thesis, non-methylated islands at gene promoters have been shown to be highly conserved across vertebrate species (Chapter 4) and to be similarly recognised and established across millions of years of evolution (Chapter 6). Many players in non-methylated island function have been identified, a number of which have chromatin-modifying activities (Blackledge et al., 2010; Thomson et al., 2010). However, the mechanism by which they contribute to regulatory function, normal gene expression, and how this is modulated during development, remains unclear.

The importance of DNA methylation and NMIs has been investigated during development using mouse as a model system. Mice mutant for DNA methyltransferase (DNMT) activity are embryonic lethal (Li et al., 1992; Okano et al., 1999; Okano and Li, 2002), suggesting that DNA methylation is essential for early development. Surprisingly however, mouse ES cells continue to proliferate and self-renew in the absence of DNA methylation (Tsumura et al., 2006) despite failure to differentiate into embryonic lineages (Sakaue et al., 2010). Therefore, DNA methylation appears dispensable for normal cellular homeostasis, but is essential during development, likely for the establishment of imprinting and for appropriate silencing of pluripotency factors during differentiation (Kelsey and Feil, 2013; Paulsen and Ferguson-Smith, 2001; Smith and Meissner, 2013). Similarly, available mutant mouse models for type-1 ZF-CxxC proteins (Long et al., 2013) exhibit embryonic lethality (Goldsworthy et al., 2013; Li et al., 1992; Yagi et al., 1998), suggesting that the appropriate regulation of NMIs is also essential for early development. The early embryonic lethality of DNMT and ZF-CxxC mutant mouse models makes it difficult to study the role of DNA methylation during early development due to internal embryonic development in placental mammals. Therefore, zebrafish was explored as an alternative viable model for studying the importance of NMIs during

early development due to the high degree of gene conservation, ease of gene knockdown and the well-characterised and rapid external embryonic developmental programme (Kimmel et al., 1995).

7.1.2 Zebrafish as a model system to study early chromatin-based processes

The zebrafish is an excellent model for studying very early embryogenesis as large numbers of zebrafish embryos can be obtained, which develop rapidly and externally (see Appendix 1.1 - Appendix 1.3). Embryos can be genetically manipulated by injection of DNA, mRNA or morpholinos and are transparent for ease of visualisation of phenotypes (Hogan et al., 2009). During early development, organisms make the transition from reliance on maternally provided gene products to those transcribed from the zygotic genome (Tadros and Lipshitz, 2009). This process, called zygotic genome activation (ZGA) or mid-blastula transition (MBT) occurs in mouse at the 2-cell stage, making the analysis of this process incredibly challenging in mammals. MBT is delayed in the zebrafish until 10 cell divisions are complete. Therefore, the greater number of cells available at this stage in zebrafish and the ease of collection of a large number of embryos makes zebrafish a great model for the study of early events such as ZGA.

The zebrafish embryo has proven to be a good model for studying chromatin-based processes (Ignatius et al., 2013; Qi et al., 2010; Rai et al., 2008; Young et al., 2006; Yu et al., 2012) due to the high degree of conservation of chromatin-modifiers (Mudbhary and Sadler, 2011) and genome-wide patterns of histone modification reminiscent of those observed in mammals (Lindeman et al., 2010; Vastenhouw et al., 2010). Importantly for the study of NMI function, recent work analysing genomic CpG methylation (Feng et al., 2010; Zemach et al., 2010) and work from this thesis (Chapter 4, Figure 4.6) has revealed that most zebrafish promoters are embedded in non-methylated regions of DNA. Furthermore the entire family of ZF-CxxC-domain containing proteins appears to be encoded in the zebrafish genome (Figure 4.2) and histone modification profiles similar to mammalian CpG islands have been identified (Aday et al., 2011; Lindeman et al., 2010). Knockdown of members of the DNA methylation machinery or ZF-CxxC domain-containing proteins have also been shown previously to

lead to developmental defects (Anderson et al., 2009; Tittle et al., 2011; Wan et al., 2011; Young et al., 2006).

Due to their well-characterised role in regulating the chromatin landscape and activity of NMIs in mouse ES cells (Blackledge et al., 2010; Farcas et al., 2012), the ZF-CxxC domain-containing KDM2 histone lysine demethylase enzymes will be explored as a means to investigate the role of NMIs during zebrafish development. The data shown in this chapter is preliminary and as such further investigation will be required to confirm any conclusions made.

Human orthologue	Name (this thesis)	NCBI	Ensembl	Locus (Ensembl)
KDM2A (KDM2A)	***zkdm2aa*** (zKdm2aa)	kdm2aa 571851	kdm2aa ENSDARG00000059653	chr1: 45,294,184-45,325,450
	zkdm2ab (zKdm2ab)	kdm2a 799441	CT955967.1 ENSDARG00000078133	chr14: 22,756,760-22,780,686
KDM2B (KDM2B)	***zkdm2ba*** (zKdm2ba)	kdm2ba 406446	kdm2ba ENSDARG00000036593	chr8: 42,291,602-42,314,330
	zkdm2b (zKdm2b)	kdm2bb 562643	kdm2bb ENSDARG00000046010	chr10: 43,194,428-43,202,258 (3' end only)

Table 7.1 *Zebrafish kdm2 gene nomenclature*

A table outlining the nomenclature for human and zebrafish *kdm2* gene orthologues, with protein nomenclature in brackets.

7.2 **Investigating the conservation and functionality of zebrafish *kdm2* genes**

7.2.1 Three *KDM2* gene orthologues in the zebrafish

Syntenic analysis of human *KDM2A* and *KDM2B* revealed two paralogous genes for each gene which were likely generated following the teleost-specific whole genome duplication to give rise to four *KDM2* gene orthologues (Appendix 7.1 and Appendix 7.2) (Catchen et al., 2009; Taylor et al., 2003). The two *KDM2A* orthologues, z*kdm2aa* and z*kdm2ab*, and two *KDM2B* orthologues, *zkdm2ba* and *zkdm2bb* were identified from the Ensembl database with predicted transcripts *ENSDART00000106684, ENSDART00000111839 ENSDART00000053158* and *ENSDART00000113426* respectively. In these annotated transcripts, *zkdm2ba* appeared to lack the JmjC domain and *zkdm2bb* both the JmjC and ZF-CxxC domains (Appendix 7.3A-B). Investigation of available EST

libraries revealed that *zkdm2ba* was correctly annotated, however *zkdm2bb* was mis-annotated and the full length transcript in fact encoded both a JmjC and ZF-CxxC domain (Appendix 7.3). Due to the conserved domain architecture with human KDM2A and KDM2B, the zebrafish genes z*kdm2aa*, *zkdm2ab* and *zkdm2bb* genes were chosen for further investigation (Figure 7.1A). For simplicity, *zkdm2bb* will be referred to as *zkdm2b* from here on. Full length cDNAs were not available for *zkdm2aa*, *zkdm2ab* and *zkdm2b*, therefore expressed sequence tag (EST) clones were obtained to build full length *zkdm2* cDNAs (see Appendix 2.1-2).

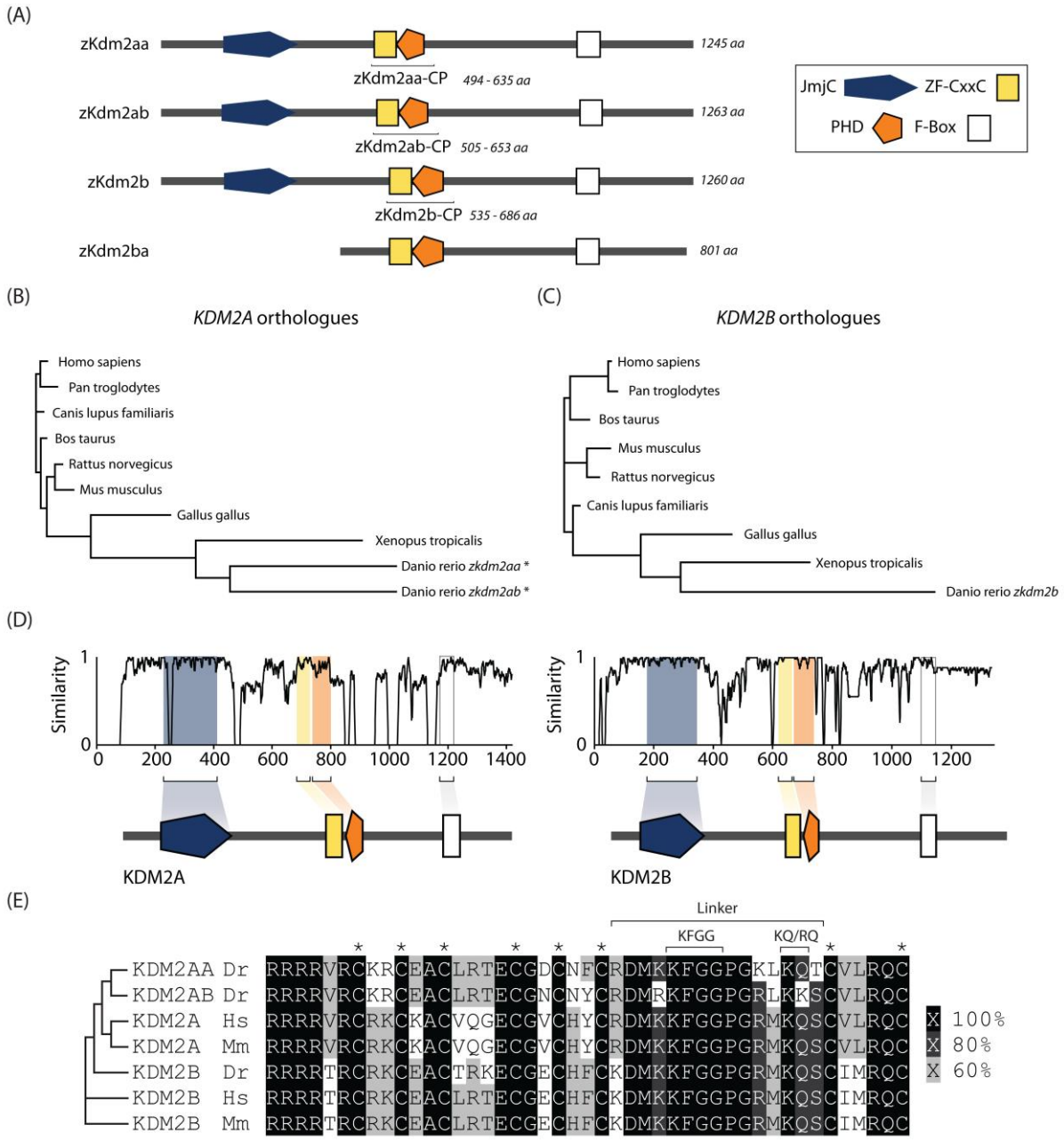


Figure 7.1 Three conserved *kdm2* orthologues in zebrafish

(A) Domain architecture of zebrafish zKdm2aa, zKdm2ab and zKdm2b, with amino acid (aa) length. (B-C) A phylogenetic tree generated from an alignment of *KDM2A* (B) and *KDM2B* (C) gene orthologues in human (*Homo sapiens*), chimpanzee (*Pan troglodytes*), dog (*Canis lupus familiaris*), cow (*Bos taurus*), rat (*Rattus norvegicus*), mouse (*Mus musculus*), chicken (*Gallus gallus*), frog (*Xenopus tropicalis*) and zebrafish (*Danio rerio*). The *KDM2A* gene has been duplicated in zebrafish to give rise to *zkdm2aa* (on chromosome 1) and *zkdm2ab* (on chromosome 14). (D) Similarity plots from the alignment of KDM2A (left) and KDM2B proteins (right) from the species listed in (B-C). There is considerable similarity between the amino acid sequence of vertebrate KDM2 proteins, especially in the functional domains (highlighted). (E) A multiple sequence alignment of KDM2A and KDM2B ZF-CxxC domains from human (Hs), mouse (Mm) and zebrafish (Dr) illustrating that the zKdm2aa and zKdm2ab zebrafish paralogues cluster with KDM2A from human and mouse while zKdm2b from zebrafish does not. Zn-coordinating cysteines are indicated with an asterisk and the linker region, important for non-methylated CpG binding, is highlighted.

7.2.2 All zebrafish zKdm2 proteins can specifically bind non-methylated DNA

To examine whether zebrafish zKdm2 gene orthologues have conserved functionality, and whether they will be a good model system for studying the functionality of NMIs during development, sequence conservation was assessed by comparison of the zebrafish zKdm2 proteins to a number of other vertebrate KDM2 gene orthologues. A striking degree of similarity was observed at the amino acid level amongst KDM2A and KDM2B orthologues (Figure 7.1B-D), especially within functional domains (Figure 7.1D-E, highlighted). Closer inspection of the zebrafish zKdm2aa, zKdm2ab and zKdm2b ZF-CxxC domains revealed conservation of all the important residues for non-methylated DNA binding, strongly suggestive that the zebrafish zKdm2 ZF-CxxC domains may be functional non-methylated DNA binding domains (Figure 7.1E).

To investigate whether the zebrafish zKdm2 protein orthologues can in fact bind to non-methylated DNA, zebrafish ZF-CxxC domains were analysed by electrophoretic mobility shift assay (EMSA), a technique which examines the capacity of a protein to bind to a DNA substrate *in vitro*. Recombinant proteins encompassing the zKdm2aa, zKdm2ab and zKdm2b ZF-CxxC domains were expressed and purified from *E. coli* using an N-terminal His-tag (Figure 7.1A and Figure 7.2A) (Blackledge et al., 2010; Farcas et al., 2012). The tag was then cleaved with TEV protease (Figure 7.2B), as the His-tag was shown previously to be detrimental to DNA binding for mammalian ZF-CxxC proteins (Figure 3.3A). All three zebrafish ZF-CxxC domains demonstrated a concentration-dependent binding to a non-methylated DNA probe *in vitro* (2.12.1) (Figure 7.2C, left). Importantly, binding was inhibited by

in vitro methylation of the DNA probe (2.12.2) (Figure 7.2C, right). Therefore the zebrafish ZF-CxxC domain-containing proteins are able to bind specifically to non-methylated DNA *in vitro*, as was predicted from sequence conservation. Due to the highly similar profiles of non-methylated DNA observed in zebrafish and mouse (Chapter 4, Figure 4.6), it is extremely likely that zebrafish zKdm2 enzymes will localise to NMIs genome-wide in a manner similar to that observed in mouse (Blackledge et al., 2010; Farcas et al., 2012). Therefore, investigation of the zKdm2 family in the zebrafish embryo should provide an insight into the importance of NMI function during early development.

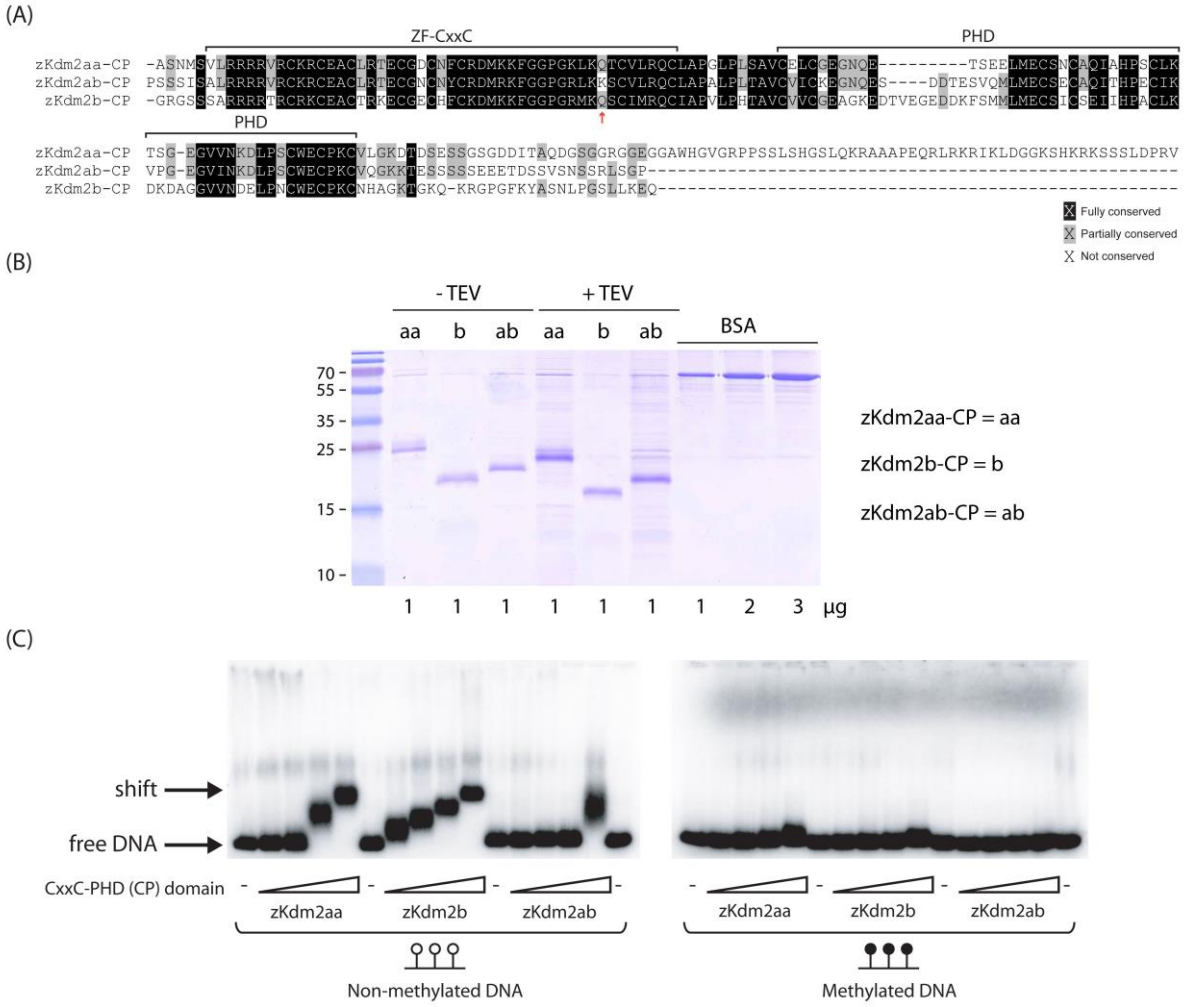


Figure 7.2 Zebrafish zKdm2 ZF-CxxC domains bind specifically to non-methylated DNA *in vitro*

(A) Multiple sequence alignment of the ZF-CxxC and PHD domain containing constructs (CP) from zebrafish zKdm2aa, zKdm2ab and zKdm2b which were used for analysis of DNA binding specificity *in vitro*. Conservation of amino acid sequence is indicated by shading and the ZF-CxxC and PHD domains are highlighted above the alignment. (B) A Coomassie-stained polyacrylamide gel of recombinant zKdm2aa-

CP (aa), zKdm2ab-CP (ab) and zKdm2b-CP (b) protein before and after TEV-protease cleavage of the N-terminal His-tag used to purify the protein. 1, 2 and 3 µg of BSA was loaded as a standard for protein concentration. (C) Recombinant zebrafish zKdm2aa, zKdm2ab and zKdm2b specifically shifts a non-methylated 216 bp 601 probe in an electrophoretic mobility shift assay (EMSA) in a 1.3 % agarose gel. Increasing concentration of recombinant protein causes an increased shift, as more protein is able to bind to the 18 available CpG dinucleotides in the probe (10, 30, 90, 270 ng zKdm2aa-CP, zKdm2ab-CP and zKdm2b-CP protein). Methylation of the 601 probe abrogates binding of zKdm2aa-CP, zKdm2ab-CP and zKdm2b-CP protein. All three were purified using an N-terminal His-tag which was cleaved with TEV-protease (B) prior to performing the EMSA assay.

7.3 **Zebrafish *zkdm2* genes are maternally expressed and ubiquitous during early development**

To investigate whether zebrafish z*kdm2aa*, z*kdm2ab* and *zkdm2b* genes are expressed during early development, temporal gene expression was analysed for important stages of early development including blastula, the onset of epiboly, early and late somitogenesis and organogenesis at 2, 4, 12, 15 and 26 hours post fertilisation (hpf) respectively (Kimmel et al., 1995). cDNA was generated from the RNA isolated from 40 embryos in biological triplicates and analysed by quantitative reverse transcription PCR (qRT-PCR). Gene expression was normalised to three house-keeping genes (*β-actin*, *tbp* and *b2m*) and the expression of all three genes was scaled to 1 for the first time-point. z*kdm2aa*, z*kdm2ab* and *zkdm2b* appeared to be expressed throughout early development with a peak of around four-fold increased expression occurring at 4 hpf (Figure 7.3A). Interestingly, this peak in expression occurred immediately following the zebrafish MBT, when zygotic transcription begins and maternal mRNAs are degraded (Giraldez et al., 2006; Tadros and Lipshitz, 2009). Therefore, the peak in expression may be a result of the surge in zygotic transcription occurring at this time or alternatively may suggest that these genes play an important role during this time in development and are therefore upregulated. Importantly, gene expression was detectable before MBT at 2 hpf, implying that *zkdm2* transcripts are maternally deposited in the zebrafish embryo as little or no zygotic transcription occurs before 3 hpf (Tadros and Lipshitz, 2009).

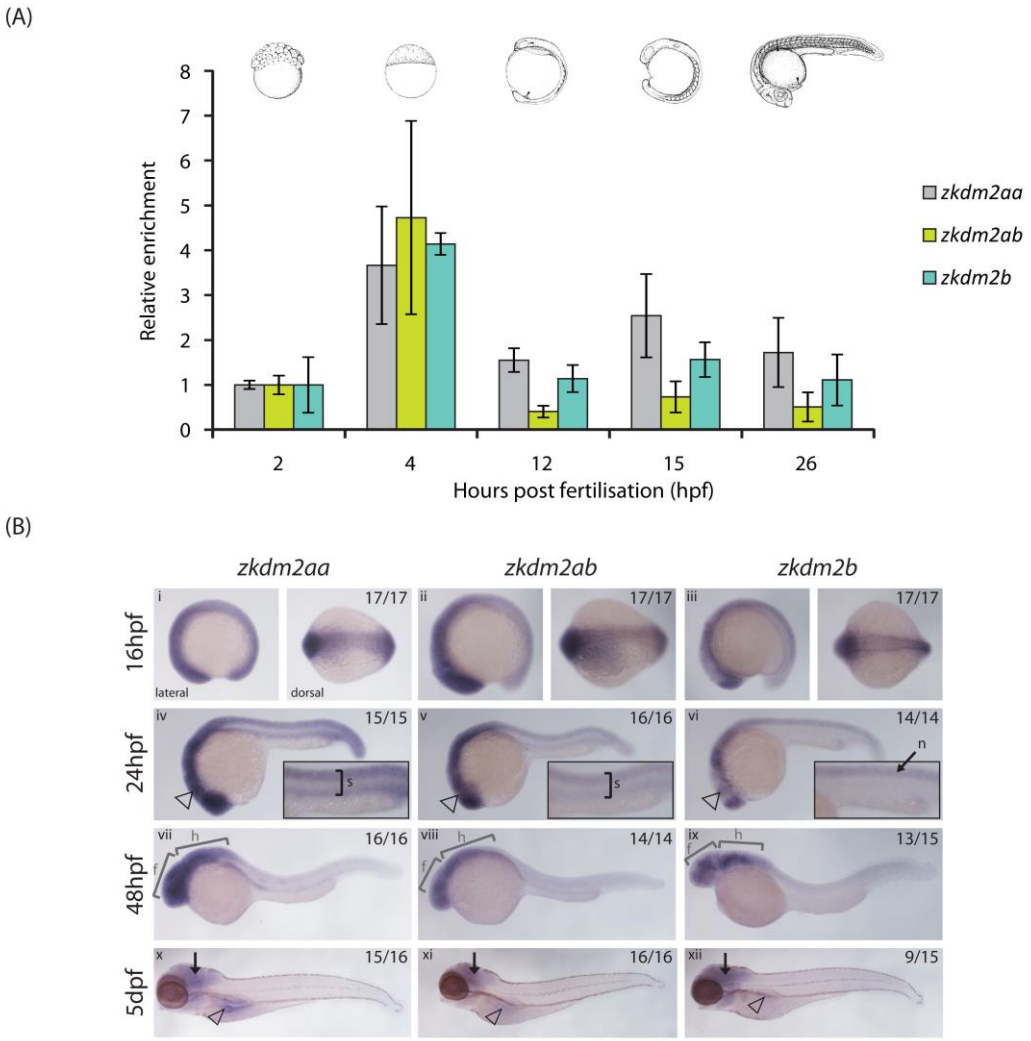


Figure 7.3 Zebrafish *zkdm2aa*, *zkdm2ab* and *zkdm2b* gene expression during early development

(A) All three zebrafish *zkdm2* orthologues are expressed during early embryogenesis as detected by qPCR. Each experiment was repeated in triplicate and error bars represent standard deviation. (B) *zkdm2* gene expression was monitored spatially by *in situ* hybridisation. Gene expression was ubiquitous at 16hpf (i-iii) and become more restricted to the head regions during development. At 24hpf, gene expression was seen in the brain (open arrow head) and in the somites of *zkdm2aa* and *zkdm2ab* stained embryos (iv, vi s), and in the neural tube for *zkdm2b* (v n). At 48hpf, expression was seen in the forebrain (f) and hindbrain (h) for all three *zkdm2* orthologues (vii-ix). By 5dpf, gene expression was restricted to the head (vertical arrow). A stained region above the yolk (open arrowhead) is likely to be background from the swim bladder. Hours post fertilisation (hpf), somites (s), neural tube (n), forebrain (f), hindbrain (h).

To examine the spatial distribution of *zkdm2* gene expression during early development *in situ* hybridisation (ISH) was performed for early zebrafish embryos (Table 2.19, ISH-1, 2 and 3). During somitogenesis at 16 hpf, and earlier (data not shown), all three *zkdm2* gene transcripts were ubiquitously expressed throughout the embryo (Figure 7.3B i-iii). By 24hpf, strong gene expression was observed in the head region for all three genes (open arrowhead). Distinct expression patterns were revealed in the tail at 24hpf, with *zkdm2aa* and *zkdm2ab* exhibiting somite-specific gene

expression (Figure 7.3B iv-v, **s**) while *zkdm2b* appeared to be expressed in the neural tube (Figure 7.3B vi, **n**). Later in development, by 48hpf, all three genes were expressed in the zebrafish forebrain (**f**) and hindbrain (**h**) (Figure 7.3B vii-ix) and this expression persisted to 5dpf (Figure 7.3B x-xii, vertical arrow), while trunk and tail expression had diminished. Staining was also observed in a region above the yolk at 5 days (Figure 7.3B x-xii, open arrowhead), though this is likely an artefact of dye accumulation in the swim bladder. Importantly, similar expression patterns were observed using a second antisense ISH probe complementary to a different part of the *zkdm2* transcripts (Table 2.19, ISH-4, 5 and 6) while sense probe controls generated for ISH-1, 2 and 3 gave no signal (data not shown). Together, gene expression analysis revealed that the *zkdm2* family are maternally deposited and expressed throughout the embryo during early embryogenesis. This may indicate that these enzymes are required during early development, perhaps for the functionality of NMIs during early development.

7.4 **Splice morpholino knockdown of zygotic *zkdm2* in zebrafish embryos**

To explore whether the *zkdm2* genes do in fact play important roles during early development, each was depleted by morpholino mediated knockdown. Morpholinos (MOs) are short (typically 25 bp in length) chemically stabilised oligomers which are anti-sense to their target mRNA molecule and either block appropriate splicing or translation depending on their binding position on the target mRNA or pre-mRNA (Bill et al., 2009). Splice-blocking morpholinos target nascent mRNAs and prevent recognition of the overlapped splice consensus sequence by the spliceosome and generally lead to incorrect splicing, production of an aberrant mRNA and reduction in protein levels (Bill et al., 2009). Splice-blocking morpholinos were designed to bind across early splice junctions in the *zkdm2aa*, *zkdm2ab* and *zkdm2b* genes (sMO-1, 2 and 3, Table 2.15) and their function was validated by RT-PCR (Appendix 7.6). *zkdm2aa*, *zkdm2ab* and *zkdm2b* splice morpholinos were injected at multiple concentrations, to titrate the morpholino dose, and only at relatively high levels of morpholino (15 or 30 ng) were late embryonic cardiac and vascular defects observed (see Appendix

7.7 for more details). However, given that the *zkdm2* genes are expressed ubiquitously during early development, these phenotypes were observed relatively late. It is therefore possible that the maternal store of *zkdm2* mRNA (Figure 7.3A) which was deposited and pre-processed in the oocyte is able to compensate for the reduction in expression from nascent transcription occurring after MBT, as maternal mRNA is resistant to splice-blocking morpholinos (Bill et al., 2009).

7.5 **Knockdown of maternal and zygotic zebrafish *zkdm2* affects early development and embryonic patterning**

In order to block both maternally-deposited and zygotic mRNAs (transcribed following MBT), two translation-blocking morpholinos, which bind upstream of or across the ATG site of *zkdm2aa*, *zkdm2ab* and *zkdm2b* were obtained. One of the *zkdm2aa* translation-blocking morpholinos exhibited no developmental phenotype and so *zkdm2aa* will not be discussed further. Knockdown of the two remaining genes, *zkdm2ab* and *zkdm2b*, gave early developmental defects and their phenotypes and impact on early patterning will be discussed in turn.

7.5.1 Knockdown of *zkdm2ab* causes developmental defects in zebrafish embryos

Injection of antisense *zkdm2ab* translation blocking morpholinos caused early developmental defects. At shield stage, corresponding to the onset of gastrulation, *zkdm2ab* morphant embryos appeared to be undergoing appropriate gastrulation movements as a shield, generated by internalization and convergent movement of blastoderm cells (see Appendix 1.5C), was visible on the dorsal side of the embryo (Figure 7.4A, arrow). However, the epibolising blastomeres appeared to have become detached from the yolk cell suggesting that contacts between the embryonic blastomeres and the yolk syncytial layer (YSL) essential for blastoderm epiboly (Appendix 1.5C) may have been lost (Figure 7.4Aiii, asterisk) (Chen and Kimelman, 2000; Cheng et al., 2004; Solnica-Krezel and Driever, 1994; Trinkaus, 1951). Whatever the cause of the epiboly defect this phenotype does not appear to cause developmental arrest as embryos were observed at a later stage to have completed epiboly and gastrulation movements. At 22hpf, two morphant phenotypes of differing

severity were observed. In the first class, embryos appeared to have developed a normal body plan, however craniofacial development appeared abnormal and the head region appeared darker with the eye and brain regions less well defined, indicative of brain necrosis (Figure 7.4Aiv). Furthermore, tail development appeared affected by *zkdm2ab* knockdown as the tail was shorter and kinked with irregular somitic patterning. The second phenotype was much more severe; embryos appeared to have a discernible head region and eye, however the body axis appeared shortened with minimal trunk tissue and the remaining tail tissue appeared to curve away from the yolk cell towards the dorsal side of the embryo (Figure 7.4Av). This is reminiscent of dorsalised embryos, which can be obtained by Bmp inhibition (Leung et al., 2003; Phillips et al., 2006; Yu et al., 2008) as well as disruption of epiboly (Strahle and Jesuthasan, 1993).

Due to their severity and apparent pleiotropic effects, it could be argued that the developmental defects described for the *zkdm2ab* morpholino are due to non-specific toxicity effects of morpholino injection as a number of these phenotypes, including shortened or gnarled tails, body curvature and small heads, are common to the overall deformed “monster”-like appearance commonly observed during morpholino-based screening (Bedell et al., 2011). However, very similar phenotypes were obtained with a second translation-blocking morpholino targeting *zkdm2ab* which binds upstream of the mRNA ATG site (Figure 7.4B). Furthermore, given that *zkdm2ab* is known to be expressed early during development (Figure 7.3A), it is probable that the early morphant phenotypes are due to specific loss of *zkdm2ab* during early development. Similarly, the later brain necrosis phenotype is consistent with the strong levels of *zkdm2ab* expression in the head-region following gastrulation (Figure 7.3B). Confirmation of phenotype specificity will await rescue of the described defects with co-injection of morpholino-resistant mRNA.

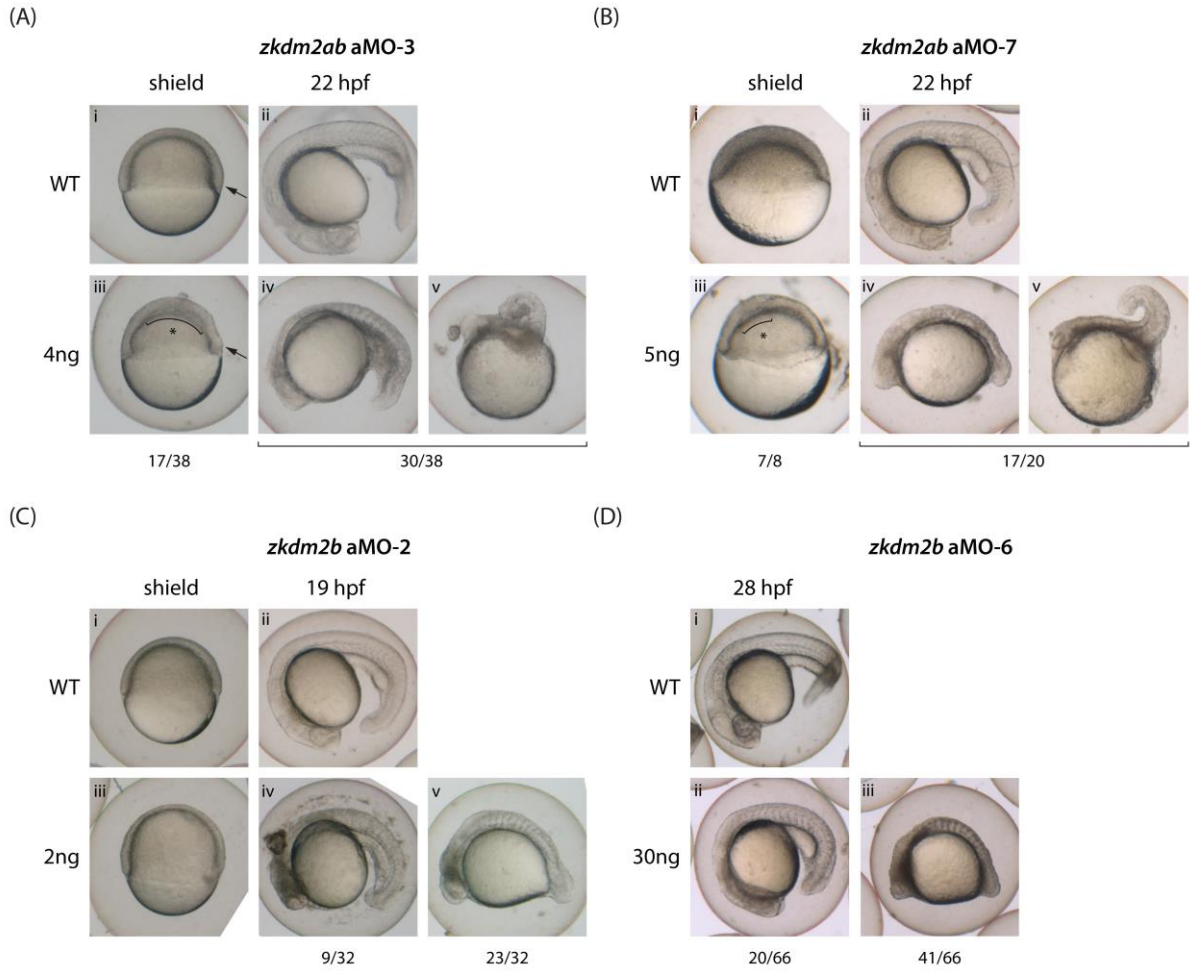


Figure 7.4 *zkdm2* genes are required for early embryonic development

Translation-blocking morpholinos were injected for *zkdm2ab* (A-B) and *zkdm2b* (C-D) and embryos were imaged at the onset of gastrulation (shield) and a later stage after gastrulation had completed (19-28 hpf). 4 ng of aMO3 (A) or 5ng of aMO7 (B) were injected into the 1-2 cell zebrafish embryo which block translation from *zkdm2ab* mRNA. At shield stage (arrow indicates shield formation), at the onset of gastrulation, the deep cell blastomeres appeared to have lost adhesion to the yolk cell (asterisk). At 22 hours post fertilisation (hpf), head necrosis and tail defects were apparent (A iv-v and B iv-v). 2 ng of aMO2 (C) or 30 ng of aMO6 (D) were injected into the 1-2 cell zebrafish embryo. After gastrulation had completed (19 or 28 hpf), morphant embryos exhibited head necrosis and aberrant tail development (C iv-v and D ii-iii).

7.5.2 Dorso-ventral patterning is affected following *zkdm2ab* knock-down

Given that developmental abnormalities could be detected from early gastrulation for *zkdm2ab* knockdown, it is likely that *zkdm2ab* provides an important function during very early development and therefore may affect early embryonic patterning. To investigate whether early developmental patterning is affected in *zkdm2ab* morphant embryos, the expression of a number of patterning genes was monitored in the early embryo by *in situ* hybridization at the onset of gastrulation. Shield stage was chosen to study patterning genes, as the fate of precursor cells which will give rise to

different cell-types and tissues is established at this stage and morphant epiboly defects are apparent by this time (Figure 7.4Aiii and Biii).

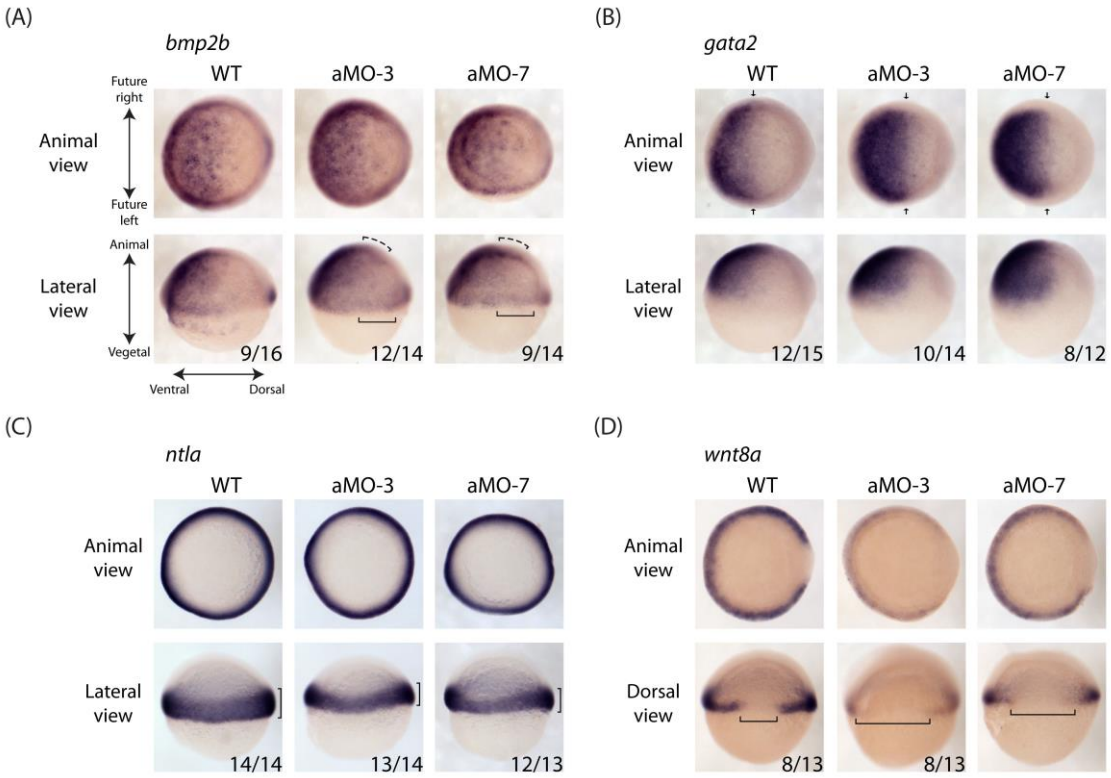


Figure 7.5 Patterning gene expression in *zkdm2ab* morphant zebrafish embryos

In situ hybridisation analysis of early patterning gene expression for *bmp2b* (A), *gata2* (B), *ntla* (C), and *wnt8a* (D) at shield stage of zebrafish embryonic development for wildtype and *zkdm2ab* morphant embryos. Translation blocking morpholinos were designed to bind across (aMO-3), or just upstream of (aMO-7) the *zkdm2ab* translation start site ATG. Shield stage embryos are displayed with dorsal to the right with images shown from the animal pole (upper) and lateral (lower) aspects. The number of embryos showing the same phenotype as the image shown are indicated.

The expression pattern of two ventral (*bmp2b*, *gata2*), a ventral mesendoderm (*wnt8a*) and a mesendoderm (*ntla*) specific gene were examined. Expression of *bmp2b*, essential for inducing ventral cell-types and for normal dorso-ventral patterning in the zebrafish embryo (Schier and Talbot, 2005), is normally observed in the ventral half of the early gastrula zebrafish embryo (Figure 7.5A, WT). In the *zkdm2ab* morphant embryos, for both aMO-3 and aMO-7, the *bmp2b* expression domain was observed to have expanded slightly towards the dorsal side of the embryo and along the margin (Figure 7.5A, aMO-3 and aMO-7, dashed and solid brackets respectively). Interestingly, the ventral non-neural ectoderm marker *gata2*, which is responsive to Bmp signalling during early

development (Dee et al., 2007), appeared to be up-regulated with *zkdm2ab* knockdown, with increased staining and a slight increase in the expression domain (Figure 7.5B). This is suggestive that Bmp signalling may be increased in *zkdm2ab* embryos, although in order to show this conclusively, other targets of Bmp signalling should be examined, for example *eve1* would be expected to have increased expression (Seebald and Szeto, 2011) while *gsc* and *chordin* expression levels would be reduced (Leung et al., 2003). Alternatively, a reporter line which drives expression of GFP under the control of a Bmp responsive element could be used to monitor Bmp activity. This element is activated by Smad1/5/8 activity and thus provides a read-out of canonical Bmp signalling in the developing zebrafish embryo (Monteiro et al., 2008).

Wnt8a is strongly expressed at the ventrolateral margin during gastrulation where it provides a posteriorising signal (Erter et al., 2001). In *zkdm2ab* morphant embryos, expression levels of *wnt8a* appeared to decrease and furthermore, the proportion of the dorsal margin which is devoid of *wnt8a* expression was expanded. Therefore, Wnt signalling may be decreased in *zkdm2ab* deficient embryos. This also suggests that the global ventralisation is not occurring in the *zkdm2ab* morphant embryos since despite an apparent up regulation of *bmp2b* and *gata2*, *wnt8a* appears to be reduced at this stage. The relationship between Bmp and Wnt signalling at this stage is complex (Agathon et al., 2003; Szeto and Kimelman, 2004), therefore more patterning genes would need to be analysed to understand the patterning defects observed here. The T-box gene notail, *ntla*, is essential for mesoderm formation and notochord development (Feldman et al., 1998; Gritsman et al., 1999; Schulte-Merker and Smith, 1995) and is expressed in all margin cells in the late blastula and early gastrula embryo and later in the tail bud (Schulte-Merker et al., 1992; Talbot et al., 1995). In *zkdm2ab* morphant embryos, the *ntla* domain appeared slightly reduced suggesting that nodal signalling and mesoderm induction (Harvey et al., 2010) may be affected in these embryos. Once again, in order to validate that nodal signalling is reduced, a luciferase reporter line under the control of a Smad-binding element could be exploited (Jonk et al., 1998). The down-regulation of nodal ligands *cyclops* and *squint* and direct target genes, for example *mixer* and *gsc* would provide

further evidence for down-regulation of nodal signalling. Together, there is evidence for developmental patterning being affected in *zkd2mab* morpholino-injected embryos and the observed changes in expression domains are likely to have knock-on consequences for later development.

7.5.3 Knockdown of *zkdm2b* causes developmental defects

Similar to *zkdm2ab*, *zkdm2b* morphant embryos displayed epiboly defects at high morpholino concentrations (15 ng, data not shown), however embryos injected with this concentration of morpholino did not survive to 24 hpf, therefore a lower concentration of morpholino was used in subsequent experiments. Despite the lack of observable morphological phenotype at early gastrula, embryos injected with a lower dose of morpholino displayed brain necrosis and tail defects similar to *zkdm2ab* morphant embryos, with two phenotypic classes observed. In the first class, the tails of the morphant embryos appeared shorter, perhaps due to developmental delay (Figure 7.4Civ), whereas for the more severe phenotype, the majority of tail tissue was lacking suggesting that the embryo has been dorsalised (Figure 7.4Civ). Again, these phenotypes were recapitulated using a second anti-sense translation-blocking morpholino (Figure 7.4D), which indicates that these phenotypes are specific to *zkdm2b* knockdown, although rescue of the morphant phenotypes with a morpholino-insensitive mRNA will be required to demonstrate absolute specificity. Morphological defects observed in the brain and tail following gastrulation were once again indicative of an earlier defect during development and the potential importance of *zkdm2* gene function in the brain-region.

7.5.4 Patterning gene expression appears abnormal in *zkdm2b* morphant embryos

In order to investigate whether patterning gene expression was affected following *zkdm2b* knockdown, ventral (*bmp2b*) and marginal (*ntla*) specific gene expression was monitored at shield stage for two different concentrations of a *zkdm2b* translation-blocking morpholino (aMO-2). Similar to *zkdm2ab* knockdown, the *bmp2b* expression domain was observed to increase towards the dorsal side of the embryo in *zkdm2b* morphant embryos (Figure 7.6A), first along the margin at the lower

morpholino dose and across the entire embryo at the higher dose (Figure 7.6A). Therefore, Bmp signaling may be up-regulated in *zkdm2b* morphant embryos, although again this would need to be verified. *Ntla* expression appeared unaffected at the lower morpholino dose (Figure 7.6B). However, at a higher dose of morpholino, ectopic *ntla* expression was observed to spread away from the margin towards the animal pole, for a proportion of morphant embryos. This suggests that the mesodermal population may have expanded, although ectodermal markers would need to be checked to see whether this has occurred at the expense of ectodermal cells (Figure 7.6B). Together, *zkdm2b* knockdown appears to affect patterning of the early embryo, and the more severe patterning defects at the higher morpholino dose may ultimately be responsible for the lethality of the higher *zkdm2b* morpholino dose.

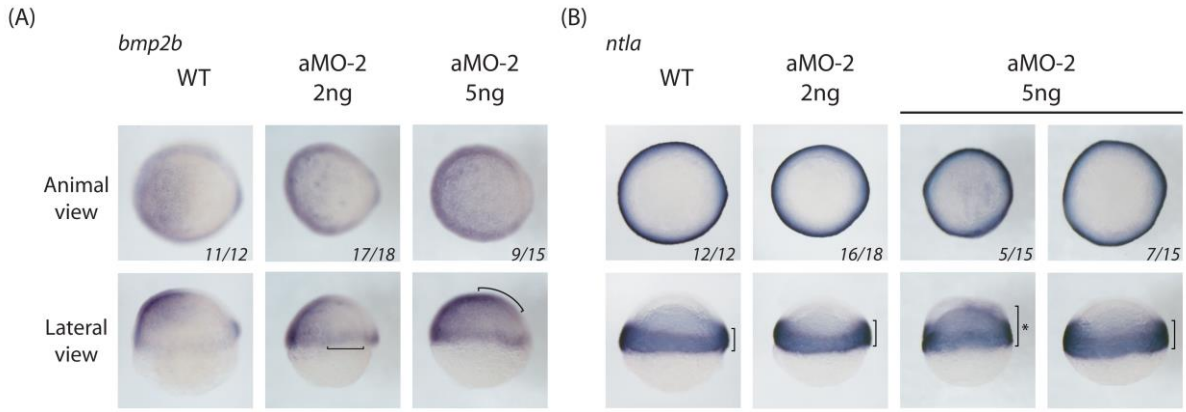


Figure 7.6 Expression of early patterning genes is affected in *zkdm2b* morphant embryos at the onset of gastrulation

(A) *In situ* hybridisation analysis of *bmp2b* gene expression at shield stage of zebrafish embryonic development for wildtype and *zkdm2b* morphant embryos. (B) *In situ* hybridisation analysis of *ntla* gene expression at shield stage of zebrafish embryonic development for wildtype and *zkdm2b* morphant embryos. Shield stage embryos are displayed with dorsal to the right with images shown from the animal pole (upper) and lateral (lower) aspects. The amount of morpholino injected and the number of embryos showing the same phenotype are indicated. Embryos are displayed as in (A).

7.5.5 Simultaneous knockdown of *zkdm2ab* and *zkdm2b* induces blastula arrest

To investigate whether *zkdm2ab* and *zkdm2b* are able to partially compensate for one another *zkdm2b* (aMO-2) and *zkdm2ab* (aMO-3) morpholinos were injected consecutively into 1-cell embryos. The phenotypes described previously for *zkdm2ab* or *zkdm2b* morphant embryos were once again observed for singly injected control embryos (compare Figure 7.4 and Figure 7.7).

Interestingly, the doubly injected embryos appeared to exhibit a stronger phenotype than either of the morpholinos injected alone, as around 50 % of embryos had arrested at the blastula stage at 5.3 hpf and were still arrested at 23 hpf or had died (Figure 7.7). Given that both zKdm2ab and zKdm2b are expected to bind to NMIs genome-wide, this suggests either that *zkdm2ab* and *zkdm2b* are able to partially compensate for one another or that they both play important and related roles in regulating NMI function during very early development.

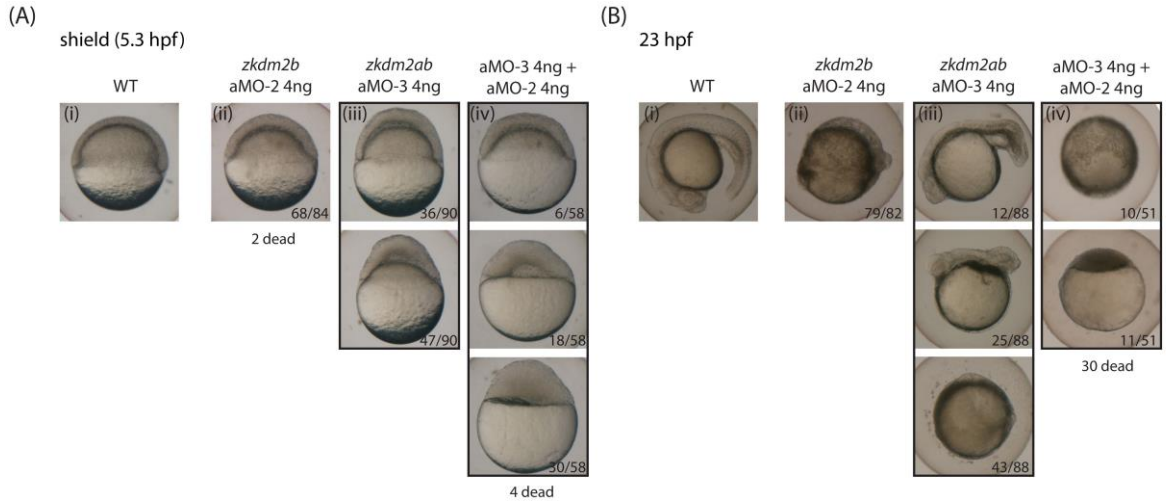


Figure 7.7 Co-injection of *zkdm2ab* and *zkdm2b* translation-blocking morpholinos causes arrest at blastula

(A-B) Embryos were injected at the 1-cell stage with 4ng *zkdm2b* translation-blocking morpholino aMO-2 (ii) or 4ng *zkdm2ab* translation-blocking morpholino aMO-3 (iii) or injected consecutively with both morpholinos (iv). Embryos injected with both morpholinos exhibited a more severe phenotype than those injected with only aMO-2 or aMO-3 alone.

7.6 **Exploring polycomb-mediated repression in zebrafish**

In addition to its proposed role as a histone demethylase, KDM2B has recently been shown to target polycomb repressive complex 1 (PRC1) to NMIs via its ZF-CxxC DNA binding domain in mouse ES cells (Farcas et al., 2012; He et al., 2013; Wu et al., 2013). Since this complex appears to be conserved from *Drosophila* to mouse (Farcas et al., 2012; Lagarou et al., 2008; Sanchez et al., 2007), *zkdm2b* might also be involved in polycomb repression in zebrafish. In order to investigate this possibility, morpholinos were designed to other proteins present in the KDM2B-containing PRC1 complex to see whether they also affect early development as observed for *zkdm2b* knockdown embryos. Since KDM2A was not detected following RING1B pull-down (the catalytic subunit of PRC1), suggesting

that KDM2A does not exist in a PRC1 complex (Farcas et al., 2012), zebrafish *zkdm2ab* will not be discussed in this section.

7.6.1 Knock-down of KDM2B-PRC1 components in the zebrafish

The variant KDM2B-containing PRC1 complex contains another unique subunit, PCGF1 (Farcas et al., 2012). Zebrafish encodes one *pcgf1* gene (Le Faou et al., 2011) to which two translation blocking morpholinos were designed (2.19.3, ENSDARG00000001249). At 26 hpf, *pcgf1* morpholino-injected embryos exhibited similar phenotypes to those described for *zkdm2b* morphants, with head necrosis and either mild or severe tail defects; a proportion of embryos appeared to lack an extended tail, while the remainder exhibited a kink in the tail (Figure 7.8, data shown for aMO-8 only, aMO-12 exhibited the same phenotype). Further characterization will include ascertaining whether *pcgf1* and *zkdm2b* knockdown yields similar phenotypes at the molecular level, for example whether the expression of a similar set of genes is affected in *zkdm2b* and *pcgf1* morphant embryos.

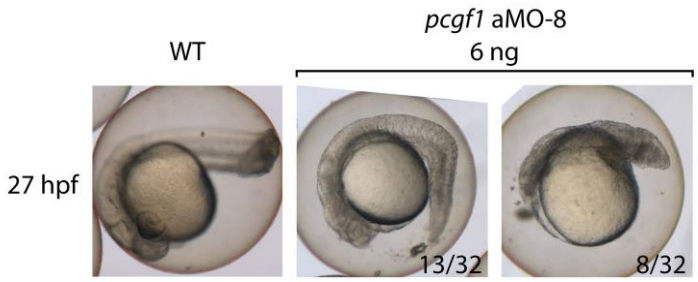


Figure 7.8 Knockdown of PRC1 component *pcgf1*

Knockdown of *pcfg1* with a translation-blocking morpholino, aMO-8. At 27 hpf, embryos exhibit two main phenotypes. In the first phenotypic class, embryos show necrosis in the brain region and sharply bent tails (middle panel) while in the second phenotypic class embryos appear more have short tails, and appear broadly necrotic (right panel).

To determine whether the phenotypes described for *zkdm2b* and *pcgf1* resemble knockdown of all PRC1 complexes, the only *Ring1* orthologue in zebrafish, *rnf2* (Le Faou et al., 2011), was targeted by two translation-blocking morpholinos (aMO-10 and aMO-11). Contrary to previous reports in which *Rnf2* was implicated in pectoral fin development (van der Velden et al., 2012) or haematopoietic stem cell development (Yu et al., 2012), this analysis indicated that both brain and tail development was affected by *Rnf2* deficiency (Figure 7.9). This phenotype is likely to be specific as both morpholinos yielded the same phenotype and loss of Rnf2 protein was detected by western blot

analysis (Figure 7.9B). Therefore the defects in pectoral fin development or loss of haematopoietic stem cells reported previously for knock-down of Rnf2 may be due to a reduction and not a full depletion of *Rnf2* (van der Velden et al., 2012; Yu et al., 2012). As before, rescue of the observed phenotypes with mRNA injection will be required to prove that the defects observed can be attributed to a depletion of Rnf2. It is perhaps surprising that knockdown of the function of all PRC1 complexes with the *rnf2* morpholinos exhibited a weaker morphological phenotype than knockdown of the Kdm2b-PRC1 complex alone (compare Figure 7.4C and D with Figure 7.9A). However, this could be explained by the presence of maternally-deposited Rnf2 protein, detectable until the 15 somite stage (van der Velden et al., 2012), which may partially compensate for knockdown of maternal and zygotic *rnf2*. Overall, the observed *zkdm2b* morpholino phenotypes appeared similar to those for PRC1 components *pcgf1* and *rnf2*, with head necrosis and tail defects observed in all cases. Co-immunoprecipitation of zebrafish Rnf2 with injected tagged zPcgf1 or zKdm2b would provide further evidence these proteins may act together in the zebrafish embryo, as in the mouse, to mediate repression of a subset of target genes.

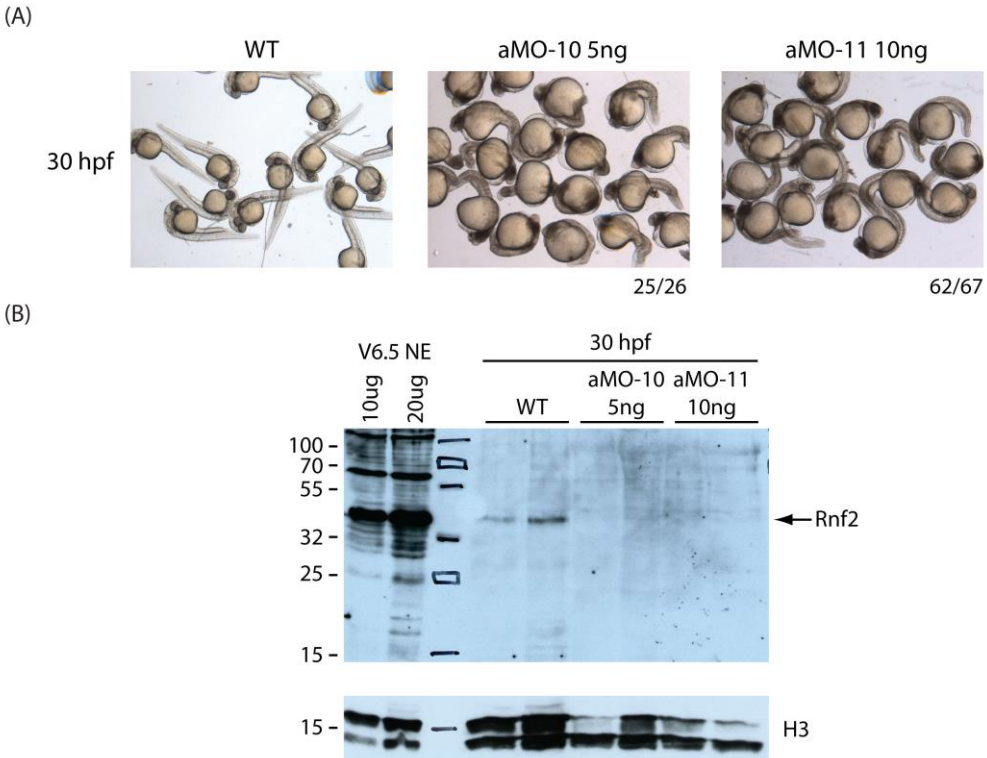


Figure 7.9 Knockdown of PRC1 component *rnf2*

(A) Images of wildtype or morphant zebrafish embryos at 30 hpf. Embryos were injected at the 1-2 cell stage with 5 ng aMO-10 or 10 ng aMO-11 translation-blocking *rnf2* morpholinos. Embryos exhibit brain necrosis and defects in tail development, which appear shorter and bent. (B) A western blot of wildtype and *rnf2* morphant embryos probed with monoclonal mouse RING1B antibody which cross-reacts with zebrafish Rnf2 protein. The equivalent of 7.5 embryos was loaded per lane. The membrane was re-probed without stripping with anti-H3 antibody as a loading control as a different secondary antibody was used.

7.6.2 Identification of a short-form of *zkdm2b* in zebrafish

A short form of KDM2B has been described for mouse (Fukuda et al., 2011), therefore the *zkdm2b* locus was analysed to determine whether a second isoform also exists in zebrafish. Transcription of the mouse short form initiates within an intron of the full length gene, and visualisation of the zebrafish locus revealed a peak of non-methylated DNA (Bio-CAP) and enrichment of the H3K4me3 modification, reminiscent of an alternative transcription start site, in intron 12 of *zkdm2b* (Figure 7.10A). *In silico* translation of all three potential reading frames for *zkdm2b* intron 12, and alignment to the 18 unique amino acids from the mouse KDM2B short form (mKDM2B-SF) revealed a region of high sequence similarity. This region overlapped directly with the observed TSS-like chromatin locus (Figure 7.10B). To determine if this homologous region represented an alternative start exon for a short *zkdm2b* isoform, RT-PCR primers were designed in the novel exon and the downstream annotated exon 13. Strikingly, an RT-PCR product was detected using a mixed cDNA pooled from a number of zebrafish embryonic stages, suggesting that a novel splicing event does occur from the putative novel exon to the following downstream exon in *zkdm2b* (Figure 7.10C-D, product A-N). Primer walking further upstream from the alternative exon revealed that a 5' UTR could be detected over 100 bp upstream of the alternative exon (Figure 7.10D, product E-N). Cloning and sequencing of a PCR product from across the splice junction revealed that the splicing event occurs in frame and thus a functional protein is likely to be made (Appendix 7.8A-C, product D-N). *In silico* translation of the proposed mRNA identified that the translational start ATG is likely to coincide with that from the mouse sequence (Appendix 7.8D). Together, analysis of the zebrafish *zkdm2b* genomic locus revealed that, similar to mouse, *zkdm2b* has a second isoform which initiates downstream of the JmjC domain-containing exons. Further analysis will be required to determine whether a zKdm2b

short-form protein product can be detected, and to determine whether this protein is functionally redundant with the long form, or has specific alternative functions.

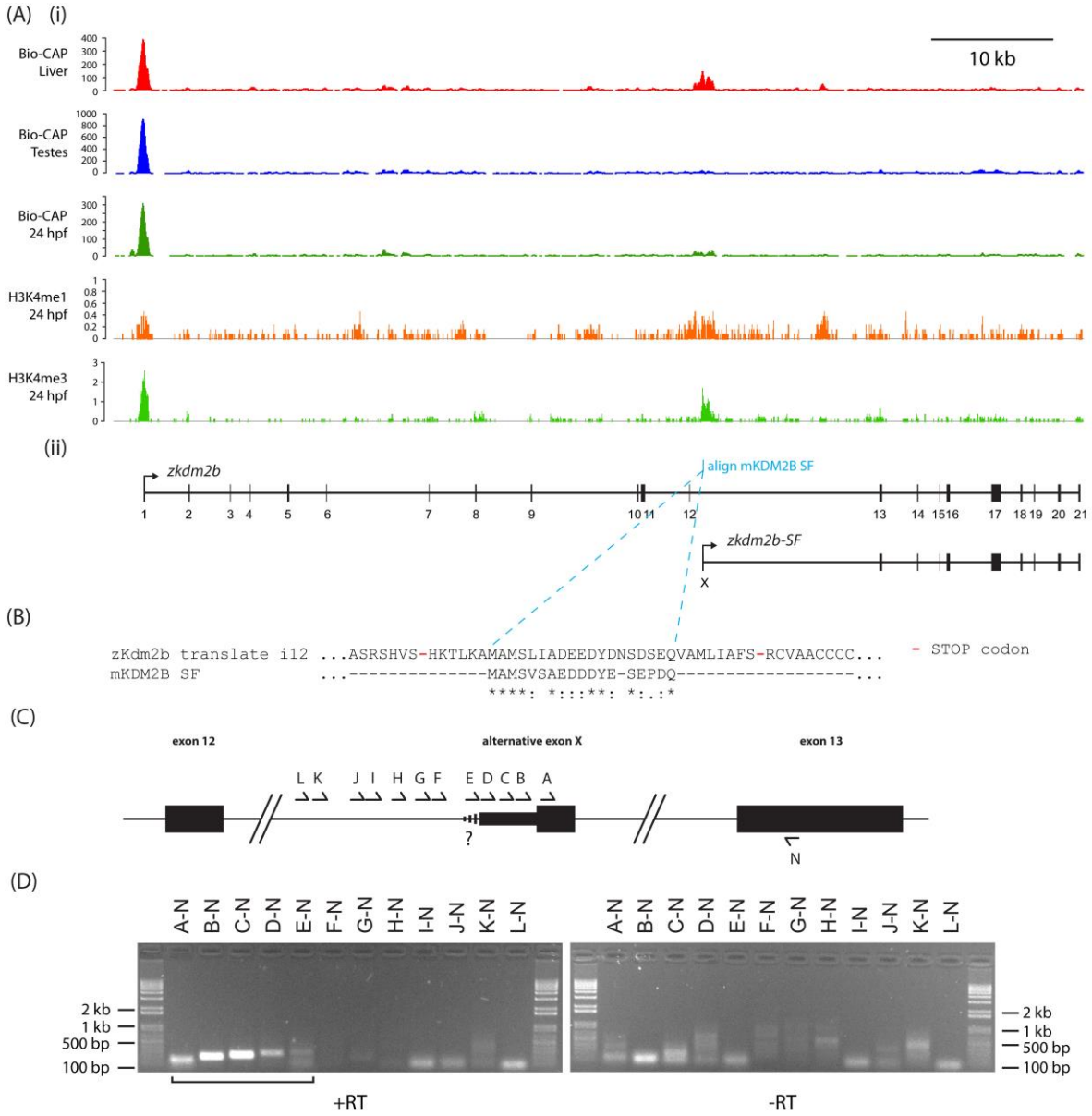


Figure 7.10 Evidence for an alternative exon in *zkdm2b*

(A) (i) The zebrafish *zkdm2b* gene spans around 80 kb with 21 exons. The transcription start site at exon 1 has a peak of H3K4me3 and H3K4me1 histone marks (Utah Bioinformatics Shared Resource, ChIP-Seq 24hpf embryo) and also is associated with non-methylated DNA (Bio-CAP) in all tissues analysed. There is a second peak of H3K4me3 and Bio-CAP (in liver and 24 hpf embryo) found at the 5′ end of intron 12. There is a short form of KDM2B in mouse (mKDM2B-SF), suggesting that zebrafish may have a short form of zKdm2b. (B) Sequence alignment of the unique amino acids from mKDM2B-SF with an *in silico* translation of zKdm2b intron 12. There is a good alignment with frame 3 translation of the intron flanked by two stop codons. (C) RT-PCR primer sets for exon 13 and in intron 12 at the putative alternative exon X. (D) RT-PCR using mixed zebrafish cDNA (from a number of embryonic stages) with and without reverse transcriptase (+/- RT). RT-PCR reveals that splicing occurs from the putative alternative exon (primers A-E, indicated by black brackets) to the downstream exon 13 and that the UTR of the alternative exon reaches to primer E (left). There are no specific PCR products for the –RT control (right).

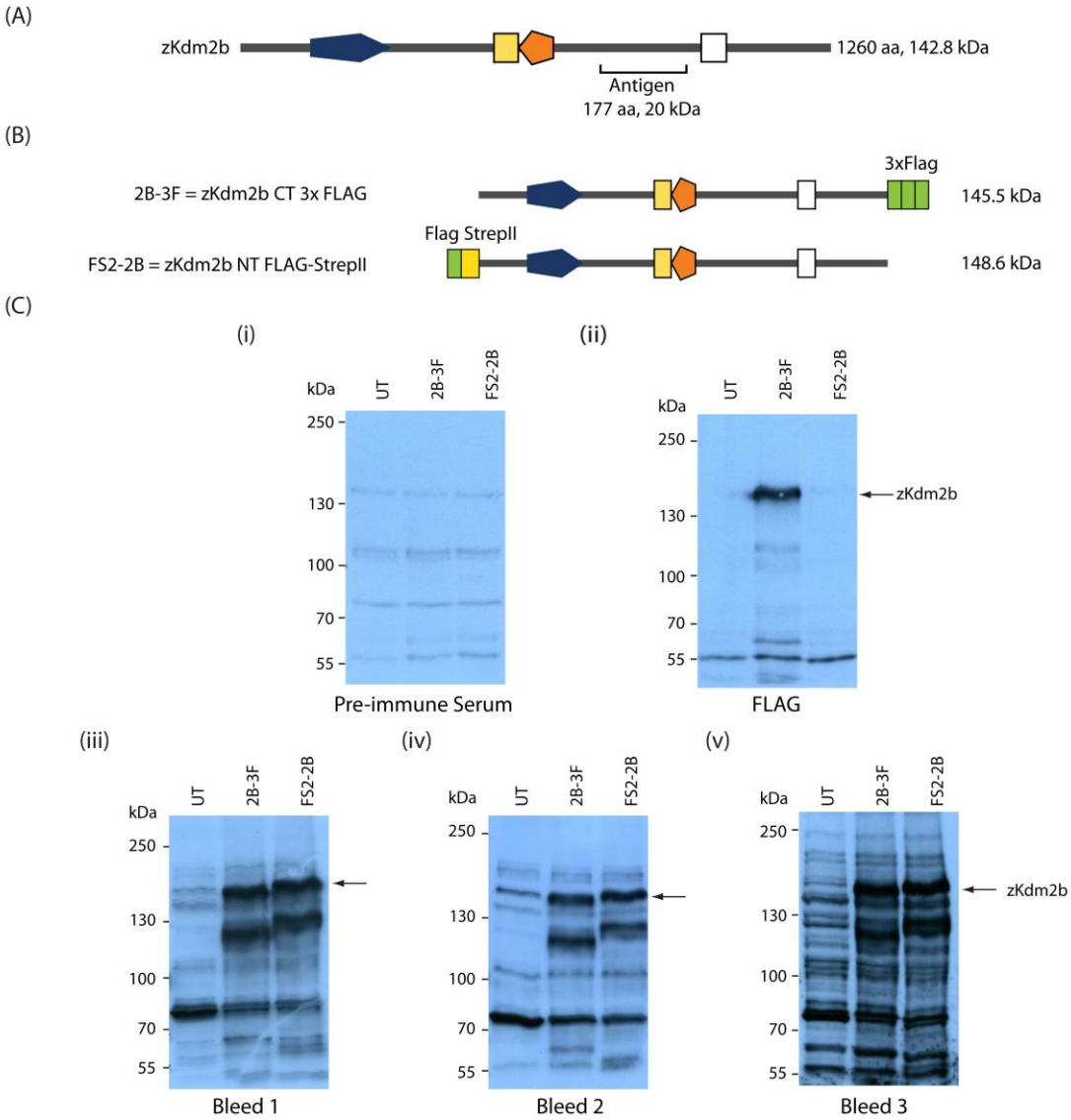


Figure 7.11 Generating a zKdm2b antibody

(A) A schematic illustrating the region of zKdm2b selected for antigen production which is equivalent to the mouse KDM2B antigen used to successfully generate a KDM2B antibody for mouse (Farcas et al., 2012). (B) Schematics of tagged zKdm2b proteins, C-terminal 3xFLAG tagged full length zKdm2b (2B-3F) and N-terminal FLAG-StrepII tagged full length zKdm2b (FS2-2B). (C) Nuclear extract from untransfected (UT), 2B-3F and FS2-2B transiently transfected 293T cells were probed with pre-immune serum (i), anti-FLAG antibody (ii) and immunised rabbit sera (iii-v). The rabbit sera recognise the zKdm2b protein following immunization (arrows, iii-v).

7.7 **Generation of an anti-zKdm2b antibody for zebrafish**

Antibody reagents are scarce for zebrafish proteins, however an antibody against zKdm2b was desirable for many reasons including to determine whether a zKdm2b short form exists in the zebrafish embryo, to what extent zKdm2b is maternally deposited in the embryo, and furthermore to validate that the *zkdm2b* morpholinos specifically deplete the zKdm2b protein in the embryo. Previously, antibodies for mouse KDM2A and KDM2B were raised using antigens designed to the

poorly conserved region between the PHD and F-box domains to prevent cross-reactivity between these two family members. Alignment of the zebrafish zKdm2 proteins similarly revealed a lack of conservation of this region (Appendix 7.9A, green bracket), and was deemed to be antigenic (antigenicity plot, Appendix 7.9B, 2.18.7). The zKdm2b antigen was therefore designed to this region (Figure 7.11A), cloned into a prokaryotic expression vector (Appendix 2.1-4iii), and recombinant protein was expressed in *E. coli*. Following purification using an N-terminal His-tag (2.10.1, Appendix 7.10A), recombinant protein was quantified by Bradford reagent (2.10.2), analysed for excess degradation by SDS-PAGE (Appendix 7.10B) and sent for antibody production (PTU/BS Scottish National Blood Transfusion Service).

To test whether the antibody was able to recognise the zebrafish zKdm2b protein, full length *zkdm2b* was cloned into a mammalian expression vector to generate an N-terminal Flag-StrepII tagged construct (Appendix 2.15, FS2-2B). This construct, along with a C-terminal 3xFlag tagged construct (Appendix 2.1-3ii, 2B-3F), was expressed in human embryonic kidney 293T cells by transient transfection (2.2.3). Nuclear extracts from untransfected and transfected 293T cells were separated by SDS-PAGE and the zKdm2b proteins were probed with an anti-Flag antibody, along with pre-immune serum, and each of three bleed sera by western blot analysis. The anti-Flag primary antibody revealed that Flag-tagged zKdm2b migrated at its expected position at around 145-148 kDa (Figure 7.11Eii, arrow), although the Flag antibody appeared to recognise the 3xFlag tag much more efficiently than the Flag-StrepII tag. All three bleed sera were observed to have reactivity at the same position as that detected by the anti-Flag antibody, suggesting that antibodies present within the sera were able to detect zKdm2b (Figure 7.11Eiii-v, arrow). Importantly, no signal was detected for the pre-immune serum, and therefore the zKdm2b reactivity appeared specific to an immune response mounted after antigen injection. Furthermore, no signal was detected for the untransfected control lane at the 145-148 kDa position, and therefore the reactivity is specific to the over-expressed protein (Figure 7.11Ciii-v, arrow). Therefore, it appears that antibodies are present in the three sera which are able to detect the zKdm2b protein. Future work will include affinity-

purification of the antibody, which will then be tested using zebrafish embryo extracts by western blot and chromatin immunoprecipitation. This reagent will prove extremely useful in future analysis of the zebrafish zKdm2b protein and its role in early development.

7.8 **Discussion**

In this chapter, the function of zebrafish *zkdm2* gene orthologues was investigated by expression analysis and morpholino-mediated knockdown. The initial results presented here suggest that these enzymes may serve an important function during early development, and raises many questions for future lines of enquiry. The roles played by the four zebrafish *zkdm2* gene paralogues, implications of the data thus far and future directions will be discussed below.

7.8.1 Functional divergence of the four *zkdm2a* and *zkdm2b* zebrafish paralogues

During teleost evolution, a whole genome duplication event is proposed to have occurred (Santini et al., 2009), and following this genome-scale doubling, a large number of duplicated genes were subsequently lost during evolution. All four genes from the *zkdm2* gene family persist in the zebrafish genome although each pair of paralogues appears to have diverged from one another, and potentially acquired distinct as well as shared functions. For example, the zKdm2ab ZF-CxxC domain exhibits a lower affinity for non-methylated DNA than that of zKdm2aa, likely due to a mutation from KQ to KK in the DNA binding loop (Figure 7.2B, orange arrow), which may correspond to a functional divergence between the two proteins (Figure 7.2C). It is plausible that the resultant loss in binding affinity was tolerated during evolution due to the presence of a second paralogous zKdm2aa protein with a canonical KQ binding loop. Of the duplicated KDM2B paralogues, *zkdm2ba* has lost the catalytic JmjC domain and presumably the enzymatic activity found in KDM2B. *zkdm2ba* was therefore initially dismissed as being more similar to the *Fbxl19* gene family (Figure 1.13) rather than the other full-length *zkdm2* genes. However, with the discovery of a short form for *zkdm2b*, which resembles *zkdm2ba*, it is possible that *zkdm2ba* and the *zkdm2b* short-form have similar functions.

Therefore, perhaps the importance of *zkdm2ba* should be reconsidered and investigated in early zebrafish development.

7.8.2 *zkdm2* genes have crucial functions during early development

Mutations in Kdm2b have been shown to play critical roles in embryogenesis, causing neural defects and exencephaly in a mouse model or lethal multi-organ developmental dysplasia in cattle harbouring a Kdm2b missense mutation (Fukuda et al., 2011; Testoni et al., 2012). Links to differentiation, cell fate decisions and reprogramming (Dong et al., 2013; Du et al., 2013; Liang et al., 2012) also suggest that *Kdm2* genes may play an important role during early embryonic development. The preliminary evidence presented here further indicates that *Kdm2* gene function may be required during embryogenesis since early gene expression and developmental defects were observed following *zkdm2ab* and *zkdm2b* morpholino knockdown. Together, the timing of these defects, and the known role of *Kdm2* genes in promoter function (Blackledge et al., 2010; Farcas et al., 2012), suggests that these enzymes may have an important role at the onset of zygotic transcription, at the mid-blastula transition (Tadros and Lipshitz, 2009). In support of this, double morphant embryos appeared to arrest at blastula (3 hpf), at the time when MBT should occur. By impacting the chromatin architecture at NMIs, the zKdm2ab and zKdm2b proteins may facilitate the appropriate expression of early patterning and developmentally regulated genes. Depletion of a zKdm2 protein may therefore prevent the formation of appropriate chromatin architecture at NMIs and dysregulate gene regulatory networks during early development.

7.8.2.1 **Chromatin modifications and DNA methylation during early zebrafish development**

In support of the idea that establishment of appropriate chromatin architecture is essential for MBT, recent work investigating chromatin changes during early zebrafish development have suggested that chromatin modifications become detectable at or just following the zebrafish MBT. Typically, the H3K4me3 modification is observed before other histone modifications (Andersen et al., 2012;

Vastenhouw et al., 2010), and as H3K4me3 is also modulated by ZF-CxxC domain-containing proteins this may suggest that the depletion of H3K36me2 at NMIs, by zKdm2 proteins, may follow similar dynamics. Furthermore, two recent studies have reported that genome-scale DNA methylation changes occur from fertilisation up to the zebrafish MBT to remodel the DNA methylation landscape of the maternal genome to match the paternal genome (Jiang et al., 2013; Potok et al., 2013). Together, DNA methylation changes and the delayed appearance of histone modifications suggest that the chromatin landscape is remodelled during early development in the lead-up to the onset of zygotic transcription. This remodelling may include changes in chromatin architecture at NMIs, which would in part be mediated by the ZF-CxxC domain-containing proteins from the zKdm2 family.

7.8.2.2 **Speculating about the role of zKdm2ab and zKdm2b during early development**

The H3K36me2 histone modification, the substrate of KDM2A, is one of the most abundant histone modifications in mammalian cells as it is found on 30-50 % of histone H3 (Garcia et al., 2008; Peters et al., 2003; Robin et al., 2007). From studies in yeast, H3K36me2 is thought to be repressive to transcription (Carrozza et al., 2005; Li et al., 2009), and therefore through the action of KDM2A, NMIs are depleted of this modification and this depletion may be considered 'permissive' to transcriptional activation (Blackledge et al., 2010). Therefore, during the DNA methylation changes observed to occur during early zebrafish development, zKdm2ab may play an important role in patterning the chromatin architecture at newly formed NMIs. Without zKdm2ab, H3K36 histone methylation may not be fully depleted from NMIs and this may have knock-on consequences for gene expression. Given that a large proportion of genes in the zebrafish are associated with an NMI (Chapter 4), this may be expected to have broad implications for a large number of genes and gene regulatory networks.

KDM2B has been implicated in targeting a variant PRC1 complex to non-methylated DNA (Farcas et al., 2012; He et al., 2013; Wu et al., 2013). The H3K27me3 modification, placed by PRC2, was detected in the zebrafish embryo by MBT but not before (Andersen et al., 2012), suggesting that

polycomb domains are established during early development in preparation for zygotic transcription at 3 hpf. Therefore, knockdown of *zkdm2b* in zebrafish may deplete one mechanism for polycomb targeting in the embryo. Consequently, a subset of genes which should be silenced or poised by polycomb repression at MBT may be aberrantly expressed due to a lack of silencing. Once again, this would be expected to impact the coordinated expression of important master regulators of gene regulatory networks and have knock-on effects in early development.

7.8.3 *zkdm2* gene expression during early development

The spatial expression patterns of *zkdm2ab* and *zkdm2b* suggests that during early development the function of these proteins is required ubiquitously, perhaps to direct appropriate chromatin architecture at all NMIs in all cells. Later during development, expression seems to become restricted to the brain region, suggesting that the zKdm2 enzymes may have an important function in neurons during later embryonic development. This, and the neuronal cell death detected in *zkdm2* morphant embryos, is consistent with a KDM2B mutant mouse model which displays neural defects and exencephaly (Fukuda et al., 2011). Despite not giving a quantitative estimate of gene expression levels, nor revealing the relative expression levels of the three *zkdm2* genes, qRT-PCR gene expression analysis revealed that the three genes have a peak in expression at MBT, perhaps consistent with an important role around the time of zygotic genome activation in defining appropriate chromatin architecture at NMIs.

7.8.4 What are the causes of the observed morphant phenotypes?

The morphant phenotypes observed following gastrulation, at around 24 hpf, appear pleiotropic as they affect many organs and tissues of the zebrafish embryo. Therefore, should these phenotypes prove to be specific to *zkdm2* knockdown, they are very likely to be due to a combination of many effects. If transcription at the mid-blastula transition is indeed affected by *zkdm2* knockdown, coordination of expression of a large number of genes would be misregulated and the accumulation of many small defects could sum to the phenotypes observed later in development. For example,

both epiboly and embryonic patterning defects were observed at the onset of gastrulation and the potential importance of these observed changes will be discussed briefly. Firstly, *zkdm2ab* knockdown caused epiboly defects which may be due to reduction in connections between the YSL and enveloping layer (EVL) at the margin (Betchaku and Trinkaus, 1978), as it has been suggested that the YSL may provide the driving force for epiboly, pulling the embryonic blastomeres vegetally (Trinkaus, 1984) (see Appendix 1.4C). Furthermore, defective epiboly, such as failure to close the blastopore properly at the end of epiboly, has been reported to lead to defective tail development (Strahle and Jesuthasan, 1993), similar to the phenotypes observed for *zkdm2* morphant embryos.

As development progresses the Bmp, Wnt, Nodal and other signalling pathways elaborate and refine the early expression patterns detectable at the onset of gastrulation. Therefore, errors in patterning, although subtle at shield stage, could go on to generate drastic phenotypic changes, as observed for the *zkdm2ab* morphant embryos. *Wnt8a* expression was observed to be reduced following *zkdm2ab* knockdown, which may also contribute to tail abnormalities, as w*nt8a* signals maintain presomitic mesoderm in the tail bud (Thorpe et al., 2005) and the Wnt8 target gene *cdx4* is essential for tail development (Davidson et al., 2003; Golling et al., 2002; Shimizu et al., 2005). As for the neurological phenotypes, a potential increase in Bmp signalling and expansion of *gata2* expression suggests there may be an expansion of non-neural ectoderm in the early embryo (Neave et al., 1997; Read et al., 1998). Potential loss of ectodermal tissue may then lead to diminished neural fates, in keeping with the brain necrosis observed later in *zkdm2ab* knockdown embryos. Intriguingly, a reduction in Bmp signaling has been reported to cause tail defects similar to those observed for *zkdm2ab* (Leung et al., 2003; Phillips et al., 2006) and *zkdm2b* (Yu et al., 2008), therefore misregulation of Bmp signaling may impact on tail development in these morphant embryos. It is worth emphasising that while it is likely that these expression changes may contribute to the observed developmental defects, it is likely that the expression of many more genes not monitored here will be affected by *zkdm2* knockdown.

7.8.5 Future directions

The above discussion regarding the function of the zKdm2 proteins during embryonic development has been hypothetical based on preliminary observations from *zkdm2* expression and knockdown studies. Further work is required to determine whether chromatin modification changes are directed by the zKdm2 enzymes during early development and whether these changes can be linked to fluctuations in gene expression and embryonic developmental defects. Due to the mechanistic link with polycomb repression in the mouse, future work will focus on zKdm2b, and the investigation of whether zKdm2b plays an important role in defining the chromatin landscape during the first hours of development in the zebrafish and whether dysregulation of this important process impacts gene expression.

7.8.5.1 **Analysis of the chromatin landscape before and after the mid-blastula transition in *zkdm2b* wildtype and morphant embryos**

In order to understand more clearly how the chromatin landscape changes leading up to MBT, specific histone modifications (including H3K4me3, H3K36me2, H3K36me3, H3K27me3 and H3K9me3) will be monitored by western blot, normalising carefully for quantities of histone using a histone H3 loading control. Other chromatin-binding factors will also be monitored such as Rnf2 and zKdm2b to see when during development they can be detected bound to chromatin. Once a careful time course of chromatin dynamics during development is known, this can then be monitored following *zkdm2b* knockdown to observe whether global chromatin changes can be detected. Should changes be observed, these could then be profiled more closely genome-wide by ChIP-sequencing to observe whether locus-specific changes are occurring.

7.8.5.2 **Genome-wide analysis of gene expression and histone modifications following *zkdm2b* knockdown**

A number of recent publications have profiled transcriptional changes before and after MBT and generally determined that zygotic transcription only initiates following MBT, although processing of

maternally deposited mRNAs, such as poly-adenylation, has been detected prior to MBT (Aanes et al., 2011; Pauli et al., 2012; Vesterlund et al., 2011). It would therefore be very informative to analyse global gene expression levels by RNA-sequencing to uncover whether a group of genes become de-repressed at MBT when *zkdm2* is depleted. Gene expression changes could be compared with genome-wide profiles of the polycomb-repressive H3K27me3 modification to observe whether genes normally repressed by polycomb are aberrantly de-repressed in *zkdm2b* morphant embryos. Furthermore, changes in Rnf2 genome-wide binding profiles could be compared to gene expression changes, with the hypothesis that loss of zKdm2b inhibits correct targeting of Rnf2 to target NMI promoters and therefore prevents appropriate polycomb-mediated repression at MBT. The zebrafish model system is potentially more powerful than a cell culture system for these experiments as the chromatin environment is established *de novo* during early development (Andersen et al., 2012; Vastenhouw et al., 2010). Therefore perturbations of chromatin modifying enzymes may yield stronger gene expression changes during early zebrafish development than for the ES cell system in which the chromatin architecture is already established with potentially multiple layers of repression acting at polycomb-repressed targets. Therefore the perturbation of one targeting system may only have minor effects on target gene expression (Farcas et al., 2012). Overall, this analysis will be very powerful for investigating in an *in vivo* system how zKdm2b function, non-methylated DNA, polycomb repression and gene regulation are interlinked.

7.8.5.3 **Rescue of deficient phenotypes with zKdm2b mutant constructs**

Having determined the effects of *zkdm2b* knockdown on the chromatin landscape and gene expression during development, the zebrafish provides a powerful tool to investigate which domains or amino acid residues are important for zKdm2b function. *zkdm2b* morphant embryos can be complemented with both wildtype and mutant mRNA to rescue the observed knockdown phenotypes. Mutations in the ZF-CxxC domain, for example, could reveal whether the binding to non-methylated DNA is essential for zKdm2b function. Recently, C-terminal leucine-rich repeats have

been shown to be important for the formation of the PRC1 complex (Wu et al., 2013), therefore it would be interesting to examine whether zebrafish *zkdm2b* with LRR mutations is able to rescue the *zkdm2b* morphant phenotypes. Over-expression of wildtype and mutant constructs will also be informative as to whether the dosage of zKdm2b is important during development. For example, injection of a construct able to bind to the PRC1 complex, but lacking a ZF-CxxC domain may be expected to phenocopy *zkdm2b* depletion as it would titrate away the zKdm2b PRC1 complex from chromatin. Together these experiments will provide further insight into the function of zKdm2b during zebrafish embryogenesis and more generally may uncover novel features of polycomb-mediated repression in the regulation of gene expression during development.

Chapter 8 – Discussion

8.1 **A novel non-methylated DNA landscape at the outset of the vertebrate lineage**

At the emergence of the vertebrates, a novel genome-wide pervasive DNA methylation profile was adopted which was completely distinct from the mosaic DNA methylation patterns observed in invertebrate ancestors (Bird, 1995; Tweedie et al., 1997). Global DNA methylation is proposed to have provided a novel mechanism for transcriptional silencing and noise-reduction and therefore may have permitted the expansion of gene number which was required for the evolution of vertebrates (Bird, 1995). Despite conservation of the DNA methylation system, computationally determined CGI maps, used as a proxy for non-methylated DNA, had previously indicated that patterns of non-methylated DNA may be significantly divergent between vertebrate species (Aïssani and Bernardi, 1991; Sharif et al., 2010). In this thesis, experimental profiling of non-methylated DNA in seven diverse vertebrate species, covering all major vertebrate evolutionary branch points, has overturned this belief and non-methylated islands (NMIs) have been shown to constitute a common feature of vertebrate gene promoters.

A simple explanation for this observation is that NMIs arose in an early vertebrate ancestor and were consequently inherited by all living vertebrate species (Antequera, 2004; Antequera and Bird, 1999; Bird, 1987). However, high resolution maps of the non-methylated DNA landscape are lacking for species positioned close to the invertebrate-vertebrate transition (Figure 8.1). At the genome-scale, invertebrate chordate species which immediately precede the vertebrates (lancet and sea squirt) and primordial vertebrate species including the jawless fish (lampreys and hagfish), have been predicted or experimentally shown to have mosaic (Albalat et al., 2012; Feng et al., 2010; Suzuki et al., 2007; Zemach et al., 2010) or global (Tweedie et al., 1997) methylation patterns respectively. Higher resolution profiling of non-methylated DNA for these species may therefore help to reveal whether the transition from fractional to global methylation occurred rapidly or gradually at the invertebrate-vertebrate transition. Therefore, exciting future work stemming from

the results presented here includes further investigation of the origins of pervasive DNA methylation and NMIs, and the selective advantage these new genomic features may have conferred during vertebrate evolution.

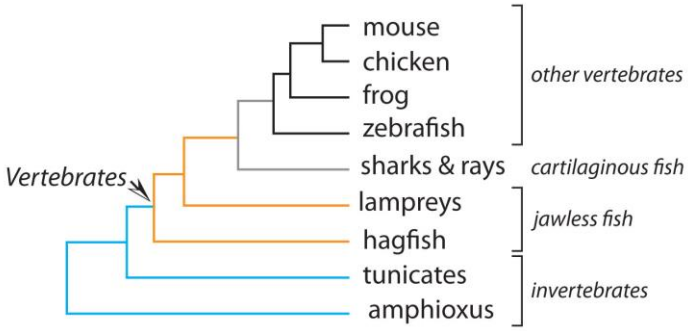


Figure 8.1 A phylogenetic tree of vertebrate evolution, including representative living vertebrate and invertebrate species

A phylogenetic tree of the invertebrate-vertebrate transition. Invertebrate chordates are shown in blue, jawless fish in orange, cartilaginous fish in grey and other vertebrates in black. The phylogeny of the jawless fish is disputed (Ota and Kuratani, 2007).

8.2 **Non-methylated DNA as a platform for gene regulation**

This thesis has revealed for the first time that genome-wide patterns of non-methylated DNA are highly similar between divergent vertebrate species. The function of NMIs is still not fully understood despite many years of investigation, although they appear to play an important role as a platform for gene regulation (Blackledge and Klose, 2011). A large family of proteins have been identified which are able to localise to NMIs and recruit chromatin modifying activities. These proteins appear to cluster into two classes, the first of which appear to place activating histone modifications or remove repressive chromatin marks, such as CFP, MLL1/2, KDM2A, and the TET1/3 enzymes (Blackledge et al., 2010; Guenther et al., 2005; Thomson et al., 2010; Williams et al., 2011). The opposing class of ZF-CxxC proteins are proposed to recruit polycomb repressive complex 1 and the SIN3A co-repressor complex via KDM2B and TET1 respectively to oppose transcription (Ku et al., 2008; Lynch et al., 2012; Mendenhall et al., 2010; Williams et al., 2011). This ability to recruit opposing transcriptional activities to NMIs has led to the suggestion that NMIs act to facilitate a bistable chromatin state, which is not instructive but responsive to the expression status of the associated gene (Klose et al., 2013). Therefore, NMIs appear to demarcate regions which are subject

to transcriptional regulation, and the ZF-CxxC proteins act to place chromatin modifications that may fine-tune the expression state of the associated gene. In keeping with this idea, the acquisition of DNA methylation at gene promoter NMIs was observed to be a rare event between distinct cell-types (Chapter 5). This is presumably because DNA methylation would lock in a repressed state, as is observed for silenced genes on the inactive X chromosome (Jones, 2012; Lock et al., 1987), whereas ZF-CxxC mediated transcriptional inhibition may be readily counteracted by gene induction (Klose et al., 2013).

In Chapter 5, intergenic NMIs were observed to exhibit features of gene regulatory elements, and were often differentially methylated between tissues, a feature which had previously been observed for mammals (Farthing et al., 2008; Illingworth et al., 2010; Maunakea et al., 2010; Mohn et al., 2008; Straussman et al., 2009; Weber et al., 2007), but not carefully examined in other vertebrates. Distal regulatory elements are thought to interact directly with promoters via looping (Ptashne, 1986; Tolhuis et al., 2002; Vernimmen et al., 2009). Interruption of these interactions, by insulator elements for example (Gaszner and Felsenfeld, 2006), would ablate the activating input from the enhancer (Vernimmen et al., 2011). It is therefore an interesting possibility that the open nature of NMI promoters and distal regulatory elements renders them particularly competent to interact. The gain of DNA methylation observed at enhancer sequences during differentiation may reduce their capacity to interact with promoter elements and regulate transcription. These possibilities could be investigated using new high resolution imaging techniques to visualize promoter-enhancer interactions (by DNA-FISH) or chromosome capture techniques (de Wit and de Laat, 2012).

Interestingly, long stretches of non-methylated DNA were observed at developmental genes and transcription factor genes (Chapter 4). This class of gene is likely to be subject to complex gene regulation and so association with a broad NMI may act to demarcate the whole locus as accessible for integration of a large number of inputs from transcription factors and potentially a large number of enhancer elements. As repression may be the default state for NMIs (Ku et al., 2008; Lynch et al., 2012), at least during development, a broad NMI may also recruit a considerable amount of

polycomb-mediated repression, and therefore these genes may require a greater amount of activating signal in order to turn on the associated gene. This could be important for developmentally regulated genes as precocious expression in an inappropriate tissue or developmental stage could have disastrous repercussions for development. Future investigation of the broad class of NMIs will help further reveal the biological function of these genomic elements.

8.3 **How changes in DNA methylation state are coordinated – lessons from exogenous DNA sequences**

When comparing distinct tissue types in different vertebrate species, the majority of differentially methylated NMIs were observed to occur away from gene promoters and exhibited lower CpG and GC content than constitutively non-methylated NMIs (Chapter 5). Intriguingly, the nucleotide properties of NMIs were extremely variable between vertebrates suggesting that species-specific mechanisms may have evolved to specify non-methylated DNA profiles in different vertebrates (Chapter 4). In order to investigate the properties that define NMIs, exogenous DNA sequences from human and zebrafish were transferred into the mouse nucleus and the fidelity of non-methylated DNA patterning was interrogated (Chapter 6). Intergenic NMIs exhibited the greatest variation in DNA methylation state between the parental and host nuclear environments, suggesting that non-TSS NMIs are more susceptible to alterations in the regulatory milieu, while promoter-associated NMIs have more robust “belt-and-braces” mechanisms for defining the non-methylated state, perhaps including recruitment of multiple general transcription factors, the transcriptional machinery or transcription itself to maintain the promoter region as free of DNA methylation. Conversely, intergenic NMIs may be reliant on a smaller number of lineage-specific DNA binding factors, productive interaction with DNA sequences in *cis* or a change in the local chromatin landscape. Any or all of these factors may not be present in the foreign nuclear environment, and lead to loss of the NMI. Exciting future lines of research will include an investigation into the link between gene transcriptional status, local chromatin modifications and the presence of NMIs.

8.4 **Regulating the timing of gene expression onset at zygotic genome activation**

Knockdown of zebrafish KDM2 gene homologues, shown to specifically bind non-methylated DNA *in vitro*, caused developmental defects and misregulation of patterning genes in early zebrafish development (Chapter 7). This suggests that the chromatin landscape at NMIs may be important for early development. The KDM2 family is known to nucleate at NMIs genome-wide in mouse ES cells (Blackledge et al., 2010; Farcas et al., 2012) where both KDM2A and B are proposed to catalyse the removal of the repressive H3K36me2 modification (Blackledge et al., 2010; He et al., 2008) while KDM2B is also able to recruit polycomb repressive complexes (Farcas et al., 2012; He et al., 2013; Wu et al., 2013). The onset of zygotic gene expression, the transition from reliance on maternal RNAs to zygotic transcripts, is tightly linked with developmental timing. In the zebrafish this occurs at the mid-blastula transition (MBT) following 10 cell divisions and immediately prior to epiboly movements (Appendix 1.4). The pluripotency factors Pou5f1 (also known as Oct4), Nanog and SoxB1 have been implicated in driving the expression of the first wave of transcription at MBT in zebrafish (Lee et al., 2013) while the chromatin landscape at MBT has also been implicated in defining the timing of genomic transcriptional competency in the zebrafish (Andersen et al., 2012; Jiang et al., 2013; Lindeman et al., 2011; Potok et al., 2013; Vastenhouw et al., 2010).

Combining these models, transcription factors, such as Pou5f1, Nanog and SoxB1, may ultimately coordinate the gene regulatory network at MBT, while appropriate histone modification profiles at NMIs may be required for preparing and maintaining (Figure 8.2D-F) or suppressing gene expression (Figure 8.2A-C) (Vastenhouw and Schier, 2012). For example, zKdm2b may normally act to target polycomb silencing to certain NMIs during development, and depletion of *zkdm2b* may reduce appropriate targeting of polycomb factors (Figure 8.2A) and lead to precocious expression of genes prior to MBT (Figure 8.2B). Similarly, an increase in H3K36me2 at NMI promoters due to depletion of zebrafish z*kdm2a* and z*kdm2b* genes may lead to encroachment of this transcriptionally repressive modification into NMIs (Figure 8.2D), leading to inappropriate silencing of genes which should be activated at MBT (Figure 8.2E).

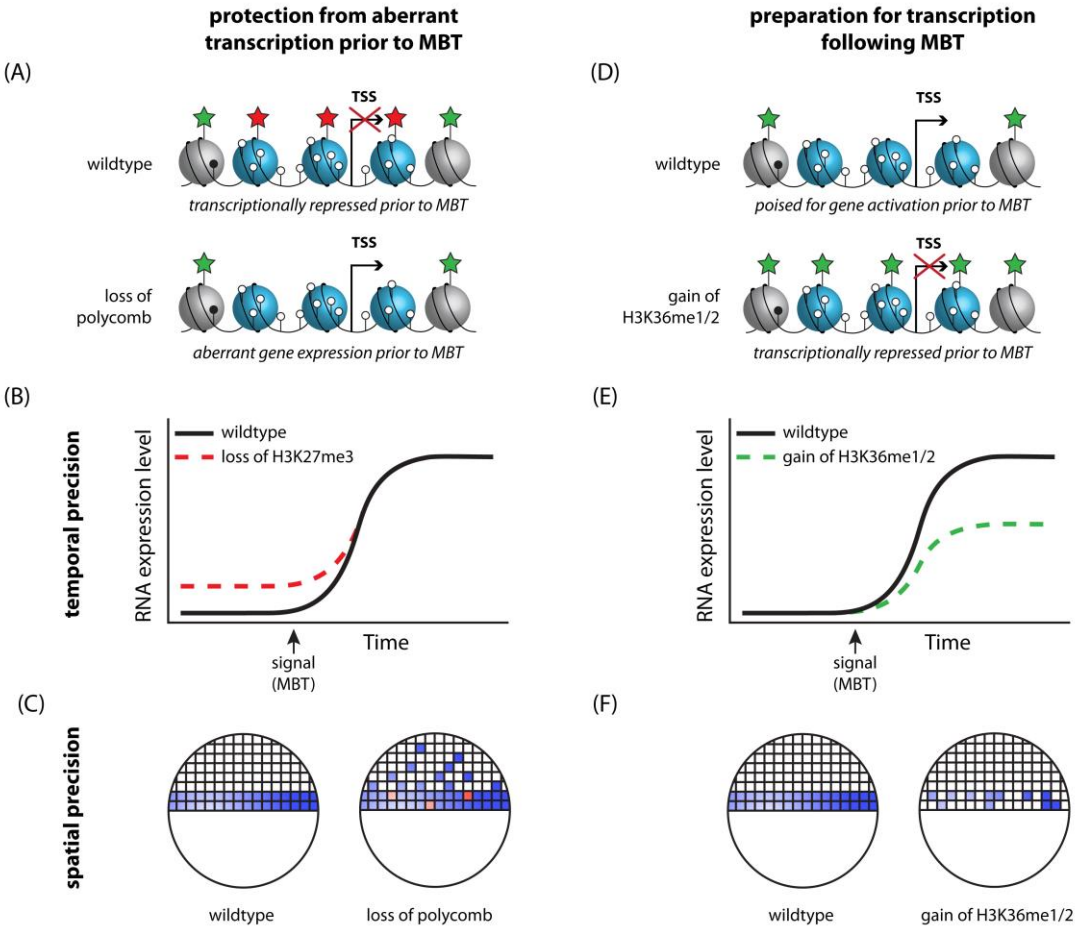


Figure 8.2 Potential roles of H3K27me3 placement and H3K36me1/2 removal at non-methylated islands during early zebrafish embryogenesis

(A-B) Loss of H3K27me3 at NMIs during early development (for example in *zkdm2b* morphant embryos) may result in the derepression and low-level expression of developmental genes not normally expressed in pluripotent blastomeres prior to MBT. (C) During lineage specification at shield stage, loss of H3K27me3 may result in the derepression of developmentally-controlled genes in cells which would normally suppress expression of those genes (ectopic blue cells). Loss of repression of pluripotency factors or genes from other lineages may also lead to failure to activate genes in some cells (non-expressing red cells). (D-E) Failure to demethylate H3K36me1/2 at NMIs during development (for example in *zkdm2ab* morphant embryos) may lead to encroachment of H3K36me1/2 and reduced efficiency of gene activation at MBT. (F) Gain of H3K36me1/2 at NMIs may lead to the loss or stochastic silencing of developmental gene expression during embryogenesis. (A and D) Nucleosomes are shown as spheres, a white circle represents non-methylated CpG while a black filled circle represents methylated DNA. Red and green stars represent the H3K27me3 and H3K36me1/2 histone modifications respectively. (C and F) Schematics of shield stage embryos expressing a developmentally regulated gene. Modified from (Vastenhouw and Schier, 2012).

The accurate coordination of gene regulatory networks and the establishment of cell-fate maps is essential for early patterning and cell fate decisions (de-Leon and Davidson, 2007; Reik, 2007). Therefore loss or gain of chromatin-based gene repression may result in the ectopic expression (Figure 8.2C) or stochastic silencing (Figure 8.2F) of developmentally-controlled genes in some cells of the developing embryo. Misregulation of gene expression at MBT could therefore provide one

explanation for the phenotypes observed in the zebrafish *zkdm2* knockdown embryos, as morphological and patterning defects were observed soon after MBT. Furthermore, depletion of both z*kdm2a* and z*kdm2b* orthologues appears to cause developmental arrest prior to epiboly, suggesting that severe changes in NMI chromatin architecture can phenocopy failure to initiate transcription at MBT (Lee et al., 2013). Transcriptional profiling in embryos depleted of z*kdm2b* soon after zygotic genome activation will provide exciting new insights into the role of NMI chromatin architecture in the complex process of zygotic genome activation.

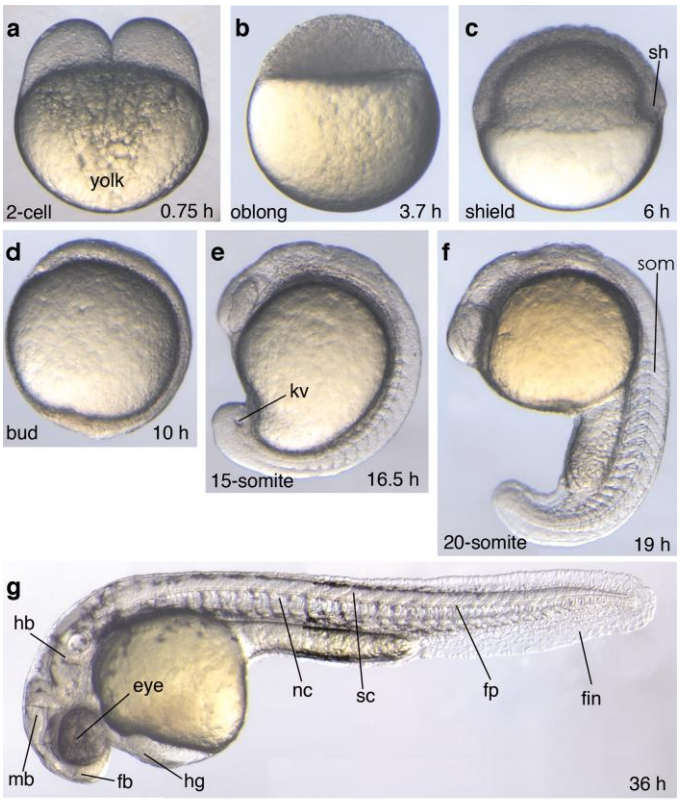
Chapter 9 – Appendices

Chapter 1

Appendix 1.1 The zebrafish model system to study developmental biology

Zebrafish is a well-established model for the study of embryonic development and vertebrate gene function (Howe et al., 2013). Embryos develop externally and rapidly, with the zebrafish body shaped by 24hpf (hours post fertilisation) (Kimmel et al., 1995). A large number of embryos can be obtained from a single pair mating, facilitating accelerated genetic studies by gene knockdown or overexpression, by antisense depletion or mRNA injection and transgenesis, with fluorescent reporter injection. A key feature for the study of early development is the transparent nature of zebrafish embryos which can be easily visualised at the single-cell level for much of early development, allowing ease of visualisation of many developmental processes using only a compound microscope. A well-annotated genome sequence is available (Howe et al., 2013), and importantly at least 70 % of human genes have at least one zebrafish orthologue, suggesting that the genetic circuitry is highly conserved between teleosts and man. Together these features and the ability to perform high throughput screening makes the zebrafish a good model for the study of human disease (Lieschke and Currie, 2007).

Appendix 1.2 Zebrafish embryogenesis



Schier AF, Talbot WS. 2005.
Annu. Rev. Genet. 39:561–613

Images of zebrafish embryos at indicated stages and developmental ages in hours post fertilisation (h). The blastodisc divides to give two cells (A), and synchronous divisions continue to give rise to a blastula of cells above the yolk cell (B). Epiboly movements then occur whereby embryonic blastomeres spread across the yolk cell, moving towards the vegetal pole (bottom of yolk cell, C). At 6h, gastrulation movements begin with involution movements forming the shield (sh, C). Epiboly movements continue until the yolk cell is completely enveloped at the bud stage (D), gastrulation movements continue during this time to

shape the embryo along an anteroposterior axis. The process of somitogenesis then occurs (E), the tail forms and extends away from the head and trunk region (F). A recognizable fish body plan is apparent at 36 h (G). Embryos are oriented with animal-pole to the top (A-B); animal-pole to the top, dorsal to the right (C); anterior to the top, dorsal to the right (D-F); anterior to the left, dorsal to the top (G). sh, embryonic shield; kv, Kupffer's vesicle; som, somite; hg, hatching gland; fb, forebrain; mb, midbrain; hb, hindbrain; nc, notochord; sc, spinal cord; fp, floor plate. Figure taken from (Schier and Talbot, 2005).

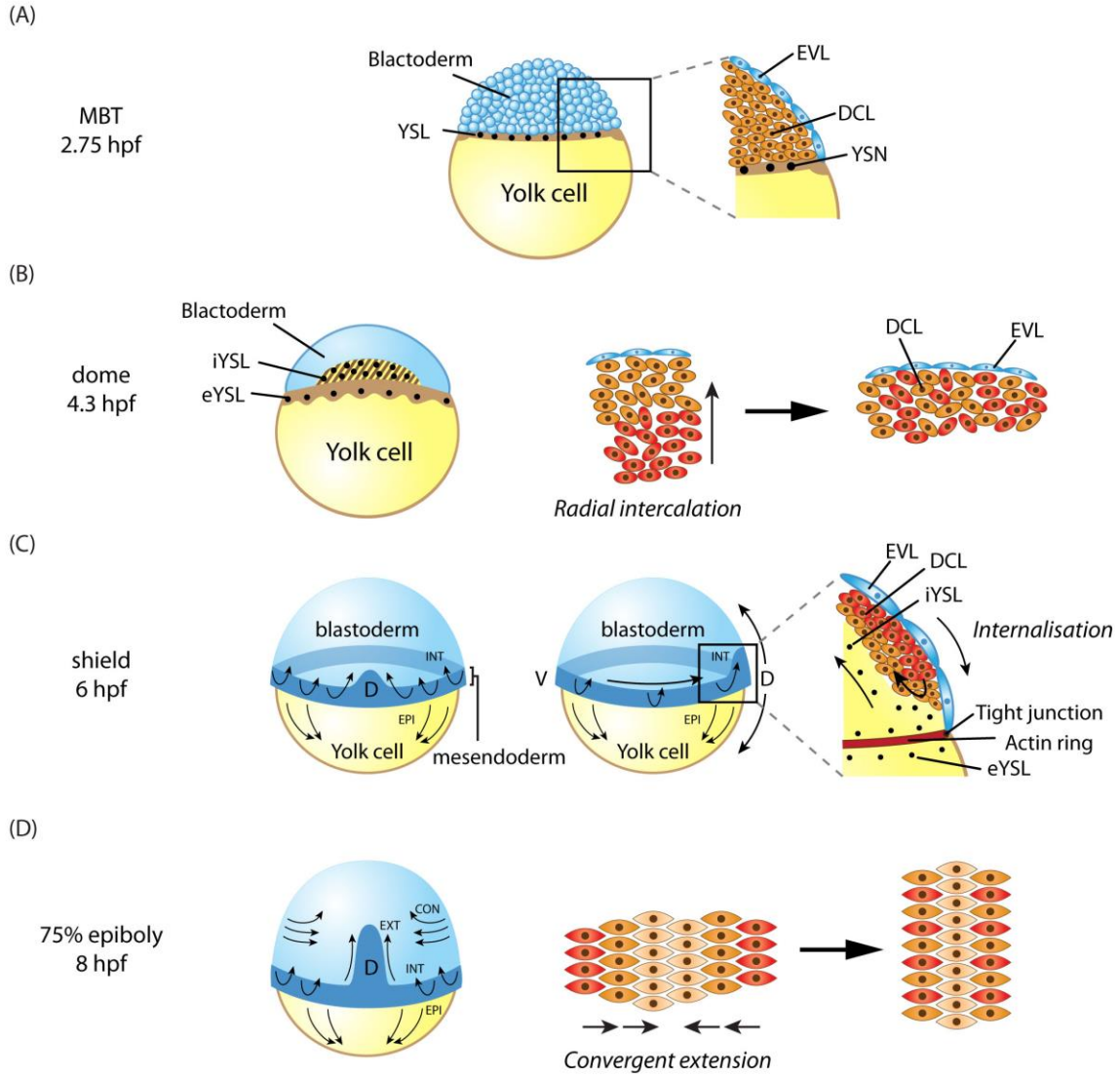
Appendix 1.3 Zebrafish development

Soon after fertilisation cytoplasmic streaming generates a large blastodisc (a single cell of the embryo) on top of the yolk cell, and during the following 3 hours rapid and synchronous divisions without cell growth generate a blastula embryo consisting of around 1000 cells atop the yolk cell. At the 512 cell stage (2.75hpf), the midblastula transition (MBT) initiates whereby zygotic genome activation occurs and the cell cycle lengthens and becomes asynchronous (Kane and Kimmel, 1993). Cells at the blastoderm margin adjacent to the yolk maintain cytoplasmic bridges to the yolk cell and at this stage the marginal blastomeres drop their cytoplasm and nuclei into the yolk to generate a thin multinucleate structure, the yolk syncytial layer (YSL, Appendix 1.5A) (Kimmel and Law, 1985). At this stage, the embryo proper consists of superficial blastomeres which form the enveloping layer (EVL) and the internal blastomeres which are within this outer layer called the deep cell layer (DCL) (Appendix 1.5A). At late blastula, cell movements begin to shape the embryo and accurate coordination of these movements is important for development as cell fate decisions are made concomitantly. Epibolic movements are the first to occur with the spreading of the YSL, EVL and DCL towards the vegetal pole involving radial intercalation movements of the DCL (Appendix 1.5B, right). Epibolising blastomeres completely wrap the yolk cell by the end of gastrulation. At 50% epiboly it is possible to generate a fate map for the zebrafish embryo (Kimmel et al., 1995; Schier and Talbot, 2005) and it is at this stage that gastrulation movements begin. Blastomeres at the margin (equatorial position of the embryo) internalise to form the shield at the future dorsal-most side of the embryo (Appendix 1.5C). Internalisation, convergence and extension movements of the DCL cells remodel the gastrulating embryo into a vertebrate body plan (Appendix 1.5D). These movements continue until tailbud stage when an elongated embryo is formed with distinct axes (10hpf, Appendix 1.2D). During the following 14 hours the somites are formed and by 24hpf a beating heart is visible (Appendix 1.2E-G).

Appendix 1.4 Mid-blastula transition

All animals make the transition during development from reliance on maternally provided gene products to those transcribed from the zygotic genome (Tadros and Lipshitz, 2009). This zygotic genome activation (ZGA) event occurs in zebrafish at the mid-blastula stage (hence, mid-blastula transition) concomitant with the degradation of maternal stores of mRNA directed by miR-430 (Giraldez et al., 2006). An early theory for explaining the onset of MBT invoked the dilution of a transcriptional repressor during divisions prior to MBT (Newport and Kirschner, 1982). Many other ideas have since been proposed (Tadros and Lipshitz, 2009), including the suggestion that a particular chromatin or DNA methylation landscape may be required prior to transcriptional activation (Andersen et al., 2012; Jiang et al., 2013; Lindeman et al., 2011; Potok et al., 2013; Vastenhouw et al., 2010).

Appendix 1.5 Epiboly and gastrulation movements of the zebrafish embryo



(A) Cross-section of blastula embryo at the mid-blastula transition (MBT, 2.75 hpf), animal pole to the top. The YSL forms by the fusing of blastomeres with the yolk cell. The embryonic blastoderm can be identified as two layers, the deep cell layer (DCL) and a single-cell enveloping layer (EVL). (B) Cross-section at dome stage (4.3 hpf), the yolk syncytial nuclei (YSN) can now be subdivided into those in front of the blastoderm margin (eYSL) and those lying beneath the blastoderm (iYSL). At this stage epiboly commences and the blastomeres undergo a process of spreading and thinning over the yolk cell, radial intercalation. (C) A dorsal (left) and lateral (middle) view of the zebrafish embryo at the onset of gastrulation (shield, 6hf). Epiboly movements continue (EPI, arrows), while marginal blastomeres (at the equatorial position of the embryo) begin to internalize (INT, arrows) and migrate towards the animal pole. Internalisation cell movements are detailed (right). The EVL is connected to an actin ring (which moves with the blastoderm epibolising margin) via tight junctions (right). (D) A dorsal view of an embryo mid-way through gastrulation (75% epiboly, 8 hpf). Epiboly and internalisation movements continue until the blastoderm covers the entire yolk cell. Blastomeres move towards the dorsal side of the embryo by convergence movements (CON, arrows) and are spread along the animal-vegetal axis by extension (EXT, arrows). YSL, yolk syncytial layer; YSN, yolk syncytial nuclei; iYSL, internal yolk syncytial layer; eYSL, external yolk syncytial layer; D, dorsal; V, ventral; EVL, enveloping layer; DCL, deep cell layer; INT, internalisation; EPI, epiboly; CON, convergence; EXT, extension. Adapted from (Carvalho and Heisenberg, 2010; Rohde and Heisenberg, 2007).

Chapter 2

Appendix 2.1 Preparation of DNA constructs

In all cases, the sequence integrity of all constructs resulting from PCR amplification in this section was verified by Sanger sequencing (2.8.9).

1. Zebrafish KDM2 expressed sequence tags (ESTs)

Zebrafish ESTs were obtained to generate full length cDNA constructs for zKdm2aa, zKdm2ab and zKdm2b (see Table 2.16).

2. Generation of full length zebrafish KDM2 complementary DNAs (cDNAs)

i. *zkdm2aa*

Sequencing of the three available zKdm2aa ESTs (Table 2.16) revealed that they cover the following nucleotide sequences of zKdm2aa: *2.1_1* 1-774 bp, *2.1_2* 276-1150 bp, *2.1_3* 999-3737 bp (end of gene). EST 2.1_1 was digested with restriction endonucleases AvrII and HindIII, 2.1_2 with HindIII and SphI and 2.1_3 with AvrII and SphI. The two excised inserts from 2.1_1 and 2.2_2 and the vector from 2.1_3 were resolved by gel electrophoresis and purified by gel extraction using QIAquick Gel Extraction Kit (QIAGEN). A three-way ligation was performed to generate a full length cDNA for zKdm2aa.

ii. *zkdm2ab*

Sequencing of EST 2.14_1 revealed that this was in fact a full length cDNA for the KDM2AB gene, therefore no further cloning was required.

iii. *zkdm2b*

Sequencing of the ESTs for *zkdm2b* revealed that two ESTs covered the full length of the gene: 2.10_1 1-1523 and 2.10_3 1515-3783 bp. The 3' end of *zkdm2b* was released from the EST 2.10_3 by digestion with EagI and inserted via ligation into a single EagI site in the 2.10_1 plasmid. Clones were screened for correct orientation of ligation to identify a full length *zkdm2b* cDNA. Unfortunately, a 2bp deletion was detected at position 1793-1794 bp of the putative full length *zkdm2b* resulting in a frame shift. This deletion was present in the 2.10_3 EST, therefore a second cloning step was performed, whereby the full length *zkdm2b* vector missing 2 bp sequence (named Δ2bp) and EST 2.10_4 were digested with NotI and BstBI restriction endonucleases. The insert from 2.10_4 (corresponding to *zkdm2b* 1519-1899 bp) was inserted into the Δ2bp vector via ligation to generate a full length cDNA for *zkdm2b*.

iv. Polymorphisms

A number of polymorphisms were detected between zebrafish reference transcript sequences (danRer7) and the sequence of the ESTs, a number of which affected the amino acid sequence (Appendix 7.3, Appendix 7.4 and Appendix 7.5).

3. Zebrafish KDM2 *in vitro* transcription constructs

i. pCS2+ plasmid

The pCS2+ plasmid contains a *Cytomegalovirus* (CMV) promoter and SP6 promoter in the forward orientation and a T7 promoter in the reverse orientation surrounding a multiple cloning site. This vector can therefore be used for *in vitro* RNA synthesis to produce appropriately processed mRNA for injection into zebrafish embryos, or can be used for transient over expression in mammalian cells.

ii. Full length zebrafish KDM2 constructs

Full length zebrafish KDM2 cDNAs were cloned into the pCS2+ vector. Appropriate restriction sites were identified in the multiple cloning site in pCS2+ which did not cut in the *zkdm2* gene to be cloned. Primers were then designed to amplify the full length *zkdm2* gene (generated in 2) including a 5' overhang containing the appropriate restriction site for ligation into pCS2+. The reverse primer also included a 3xFLAG tag followed by a stop codon. Both *zkdm2aa* and *zkdm2ab* were amplified with primers containing EcoRI and XbaI restriction sites, while both forward and reverse *zkdm2b* primers contained StuI sites. The *zkdm2* genes were amplified by PCR, and the PCR products and the pCS2+ vector were digested with the appropriate restriction enzymes. Vector and inserts were resolved by gel electrophoresis and purified using the QIAquick Gel Extraction Kit (QIAGEN) and the PCR product inserted via ligation. StuI is a blunt cutter and so the efficiency of ligation was low and the orientation of the insert was determined by digestion with BamHI to identify a positive clone.

Table 1 *Primers used for cloning zkdm2 genes into pCS2+ vector*

Klose Primer Index	F/R	Restriction site	C-terminal tag	ATG morpholino	Gene	Klose plasmid index
1188	F	EcoRI		Sensitive	*zkdm2aa* (1)	417
1179	R	XbaI	3xFLAG			
1190	F	EcoRI		Sensitive	*zkdm2ab* (14)	419
1183	R	XbaI	3xFLAG			
2438	F	StuI		Sensitive	*zkdm2b* (10)	613
1181	R	StuI	3xFLAG			

iii. Morpholino insensitive rescue constructs

In order to rescue the observed phenotypes in zebrafish injected with ATG-blocking morpholinos, *zkdm2 in vitro* expression constructs were generated which were mutated in the ATG morpholino binding site. The forward primer was designed to amplify from the start ATG, and the third (and sometimes the first or second) codon positions were altered to change the nucleotide sequence without affecting the protein sequence. The reverse primer remained the same. The cloning was performed as for ii.

Table 2 *Primers used for cloning zkdm2 rescue constructs*

Klose Primer Index	F/R	Restriction site	C-terminal tag	ATG morpholino	Gene	Klose plasmid index
1215	F	EcoRI		Insensitive	*zkdm2aa* (1)	418
1179	R	XbaI	3xFLAG			
1217	F	EcoRI		Insensitive	*zkdm2ab* (14)	420
1183	R	XbaI	3xFLAG			
2439	F	StuI		Insensitive	*zkdm2b* (10)	Failed
1181	R	StuI	3xFLAG			

The cloning of the *zkdm2b* morpholino insensitive pCS2+ construct proved challenging, and so an alternative approach was taken. The *zkdm2b* ATG morpholino sensitive construct (plasmid 613) was cut with EcoRI as was a PCR product amplifying the 5'-end of *zkdm2b* including an EcoRI site in the forward primer and the codon mismatches (in the wobble positions) designed above to disrupt the ATG morpholino binding site. The reverse primer was downstream of an EcoRI site present in the *zkdm2b* gene. The PCR product and plasmid 613 were digested with EcoRI, resolved by electrophoresis, purified and ligated to generate a morpholino insensitive *zkdm2b* expression construct (plasmid 614).

Table 3 *Primers used for alternative cloning of the zkdm2 rescue construct*

Klose Primer Index	F/R	Restriction site	C-terminal tag	ATG morpholino	Gene	Klose plasmid index
2463	F	EcoRI	-	Insensitive	*zkdm2b* (10)	614
2464	R	-	-			

4. Prokaryotic expression constructs

i. pNIC28 prokaryotic expression vector

The prokaryotic expression vector pNIC28 (a gift from U. Oppermann) was modified previously to express an N-terminal 6-his tag followed by a tobacco etch virus (TEV) protease cleavage site and was adapted for ligation independent cloning (LIC) as described previously (Blackledge, Zhou et al). pNIC28 is cut with BsaI prior to LIC processing, as described in 2.8.6.

ii. zKdm2 ZF-CxxC PHD expression constructs for EMSA

ZF-CxxC containing constructs encoding amino acids 544-648 of human KDM2A was cloned previously by Dr Jin Zhou, (Blackledge et al., 2010). Similar constructs were PCR amplified from the appropriate cDNA corresponding to amino acids 600-750 of human KDM2B, 494-635 of zebrafish KDM2AA (chr1), 506-653 of zebrafish KDM2AB (chr14), 535-686 of zebrafish KDM2B (chr10). These PCR products were LIC-processed and inserted via ligation independent cloning (LIC) into the adapted pNIC28 prokaryotic expression vector (see 2.8.6). The recombinant protein was expressed and purified in BL21 *E. coli* (see 2.10.1) and used in electrophoretic mobility shift assay (EMSA, see 2.12).

Table 4 *Primers for amplification of LIC-adapted ZF-CxxC constructs*

Klose Primer Index	F/R	Template	Gene	Klose plasmid index
985	F	EST 2.1_3	*zkdm2aa* (1)	410
986	R			
899	F	EST 2.14_1	*zkdm2ab* (14)	413
904	R			
983	F	2.10_4	*zkdm2b* (10)	412
984	R			

iii. *zkdm2* antigen expression constructs for antibody production

Constructs for zKdm2 antigens were designed to the same region of the zKdm2 proteins as was used previously to make antibodies for mouse KDM2A and KDM2B. This region lies between the conserved ZF-CxxC-PHD and F-box regions, and in zebrafish corresponds to amino acids 753-915 for zKdm2aa, 713-875 for zKdm2ab and 770-946 for zKdm2b. These regions were amplified from full length *zkdm2* plasmids using primers containing LIC overhangs. These PCR products were LIC-processed and inserted via ligation independent cloning (LIC) into the pNIC28 prokaryotic expression vector (see 2.8.6).

Table 5 *Primers for amplification of zkdm2 antigen constructs*

Klose Primer Index	F/R	Template	Gene	Klose plasmid index
1284	F	410	*zkdm2aa* (1)	414
1285	R			
1286	F	396	*zkdm2ab* (14)	415
1287	R			
1282	F	412	*zkdm2b* (10)	416
1283	R			

The zKdm2b antigen expressed very well, but the majority of the protein did not bind to the Ni-NTA affinity resin and remained in the flow through. As the protein was difficult to purify, the LIC-processed PCR product for *zkdm2b* was inserted via ligation independent cloning (LIC) into the pGTVL2 plasmid (273) which has an N-terminal His-GST tag. Fusion of the antigen to GST improved the yield over the His only tagged protein, perhaps improving solubility.

iv. Expression construct of Avi-tagged human KDM2B for Bio-CAP

Rob Klose PCR amplified amino acids 600-750 of human KDM2B (ZF-CxxC and PHD domain) and engineered an AviTag at the C terminus. This PCR product was LIC-processed and inserted via ligation independent cloning (LIC) into the adapted pNIC28 prokaryotic expression vector (see 2.8.6).

5. Eukaryotic expression constructs

i. Zebrafish eukaryotic expression vector

Zebrafish *zkdm2b* was inserted via LIC cloning (2.8.6) into the pCAGIpuro LIC vector (441) for eukaryotic expression. The pCAGIpuro LIC vector has been adapted to add an N-terminal Flag Strep II tag onto proteins LIC-cloned into this vector and has a pCAG promoter suitable for eukaryotic expression.

Table 6 *Primers for cloning zkdm2b into a eukaryotic expression vector*

Klose Primer Index	F/R	Template	Gene	Klose plasmid index
2465	F	410	*zkdm2b* (10)	615
2466	R			

6. Cassettes for BAC recombineering

BAC recombineering was used to introduce a cassette of interest into the backbone of a bacterial artificial chromosome (BAC). Zebrafish BACs with the same backbone originating from the pIndigoBAC536 plasmid were ordered to allow the same recombineering strategy between different BACs. An inverted Tol2 construct flanking ampicillin selection marker was obtained from Stefan Schulte-Merker (Bussmann and Schulte-Merker, 2011; Suster et al., 2009) (plasmid 445).

For the zebrafish BAC, zBAC2, an inverted Tol2 cassette and a Kan/Neo selection cassette were introduced into the backbone. The Kan/Neo cassette allowed selection in bacteria and mouse ES cells using Kanamycin and G418 respectively.

Table 7 *Primers for generating a recombineering cassette to introduce iTol2 amp into zBAC2*

Klose Primer Index	Stefan-Schulte Merker Primer	F/R	Template	Plasmid name
1730	pIndigoBAC_HA1_iTol2_fw	F	445	iTol2_amp
1731	pIndigoBAC_HA2_iTol2_rev	R		

Analytical PCR was performed to verify successful recombineering. Homology arm 1 was amplified using primers 1732-1733 (pIndigoBAC_HA1_control_fw – amp_HA1_control_rev) and homology arm 2 using primers 1734-1735 (amp_HA2_control_fw – pIndigoBAC_HA2_control_. Primers were taken from a protocol by Stefan Schulte-Merker (Bussmann and Schulte-Merker, 2011) (http://www.hubrecht.eu/research/schulte-merker/documents/bacrecombineering_stepbystep.pdf).

Table 8 *Primers for generating a recombineering cassette to introduce Kan/Neo into zBAC2*

Klose Primer Index	F/R	Template	Plasmid name
1815	F	461	RGKamphNGFP
1816	R		

Analytical PCR was performed to verify successful recombineering. Homology arm 1 was amplified using primers 1817-1819 and homology arm 2 using primers 1818-1820.

Chapter 3

There are no appendices for Chapter 3.

Appendix 4.1 Vertebrate ZF-CxxC gene orthologues

	Human	Mouse	Platypus	Chicken	Lizard	Frog	Zebrafish
KDM2A	ENSG00000173120	ENSMUSG00000054611	ENSOANG00000005993	ENSGALG00000013286	ENSACAG00000007630	ENSXETG00000015793	ENSDARG00000059653 *kdm2aa*; ENSDARG00000078133 kdm2ab
KDM2B	ENSG00000089094	ENSMUSG00000029475	ENSOANG00000006859	ENSGALG00000004225	ENSACAG00000001871	ENSXETG00000020966	ENSDARG00000046010 ‡
FBXL19	ENSG00000099364	ENSMUSG00000030811	ENSOANG00000030194	?	ENSACAG00000007651	ENSXETG00000001927	ENSDARG00000076929
CFP1	ENSG00000154832	ENSMUSG00000024560	ENSOANG00000014594	?	ENSACAG00000006087	ENSXETT00000047467 †	ENSDARG00000025718
DNMT1	ENSG00000130816	ENSMUSG00000004099	ENSOANG00000012255	ENSGALG00000028997	ENSACAG00000003948	ENSXETG00000022035	ENSDARG00000030756
MLL1/KMT2A	ENSG00000118058	ENSMUSG00000002028	ENSOANG00000003760	ENSGALG00000006818	ENSACAG00000007014	ENSXETG00000010087	ENSDARG00000004537
MLL2/MLL4/WBP7	ENSG00000105663	ENSMUSG00000006307	?	?	ENSACAG00000016679	ENSXETG00000017567	ENSDARG00000062962 *mll4a*; ENSDARG00000060697 *mll4b*
MBD1	ENSG00000141644	ENSMUSG00000024561	ENSOANG00000014593	?	ENSACAG00000006091	ENSXETT00000064259 †	ENSDARG00000025699 *mdb1*; ENSDARG00000088407 *mbd1 (2 of 2)*
TET1	ENSG00000138336	ENSMUSG00000047146	ENSOANG00000015695 *	ENSGALG00000004073	ENSACAG00000010131	?	ENSDARG00000075230
TET3	NM_001097187.1	ENSMUSG00000034832 *	ENSOANG00000004099 *	ENSGALG00000013436 *	ENSACAG00000001632 *	NM_001097187.1	ENSDARG00000062646 *
IDAX (CxxC4)	ENSG00000168772	ENSMUSG00000044365	ENSOANG00000008235	?	?	ENSXETG00000014100	ENSDARG00000077817
CXXC5	ENSG00000171604	ENSMUSG00000046668	ENSOANG00000009530	ENSGALG00000000884	ENSACAG00000002186	?	ENSDARG00000078865

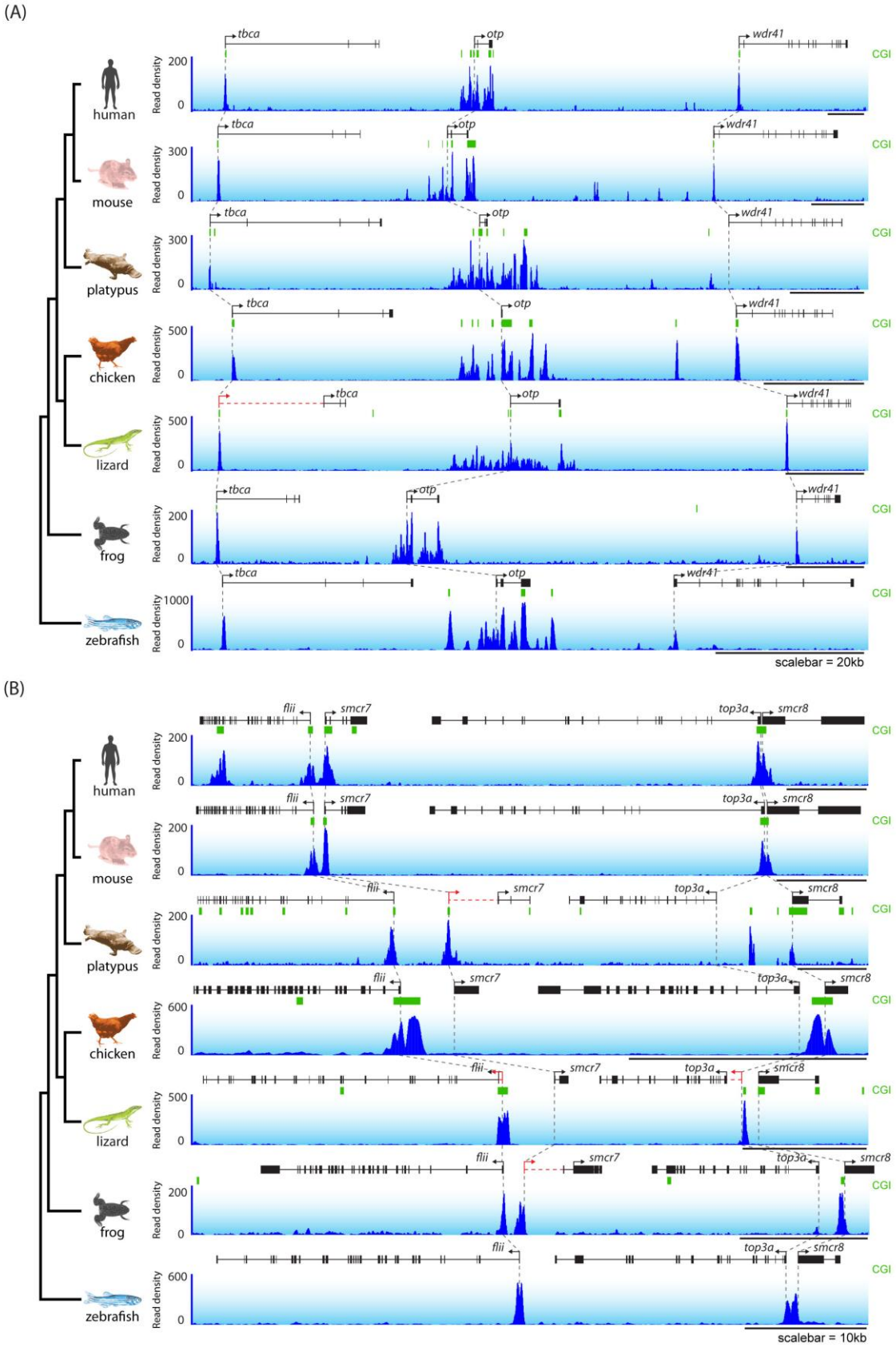
A table listing the Ensembl gene IDs for all annotated ZF-CxxC proteins from human, mouse, platypus, chicken, lizard, frog and zebrafish. * no annotated ZF-CxxC domain, † frog gene ENSXETG00000021938 contains the orthologues for both CFP1 and MBD1 (different transcripts), ‡ contains only the 3' end of the gene.

Appendix 4.2 Bio-CAP sequencing statistics

			Reads		
		sample	Total	Mapped	Unique mapped
Species	Human	Input (testis)	64,886,866	47,604,119	43,706,654
		Testes	53,637,385	36,568,199	27,630,493
		Liver	68,094,500	50,722,292	37,630,783
	Mouse	Input (testis)	56,973,944	37,674,053	33,599,701
		Testes	41,624,336	24,946,136	18,897,043
		Liver	61,630,866	44,725,313	37,638,942
		ES cell	35,292,148	22,506,507	21,328,749
	Platypus	Input (testis)	64,982,542	45,689,505	43,605,199
		Testes	72,706,367	34,146,707	30,695,377
		Liver	70,763,280	31,221,768	27,853,293
	Chicken	Input (testis)	63,312,725	46,978,928	43,306,854
		Testes	55,709,171	36,332,203	23,641,494
		Liver	76,503,324	58,646,349	48,408,635
	Lizard	Input (testis)	41,187,232	29,717,442	29,182,278
		Testes	39,245,777	27,424,149	25,170,801
		Liver	37,597,309	17,745,828	16,778,478
	Frog	Input (testis)	38,418,567	20,374,974	19,639,692
		Testes	39,484,021	23,206,617	20,191,025
		Liver	70,328,659	44,226,210	37,552,538
		Stage 11-12	38,901,649	30,553,529	10,788,278
	Zebrafish	Input (testis)	56,207,545	34,302,018	32,384,749
		Testes	58,697,552	38,487,629	30,258,963
		Liver	57,105,500	37,975,697	30,874,004
		24 hpf	36,280,355	20,687,841	17,506,436
		Total	*1,299,571,620*	*842,464,013*	*708,270,459*

A table listing the total number of sequencing reads for duplicate experiments for each species and tissue. Numbers of reads mapped and uniquely mapped to the appropriate reference genome are also listed.

Appendix 4.3 Non-methylated islands are a common feature of vertebrate gene promoters



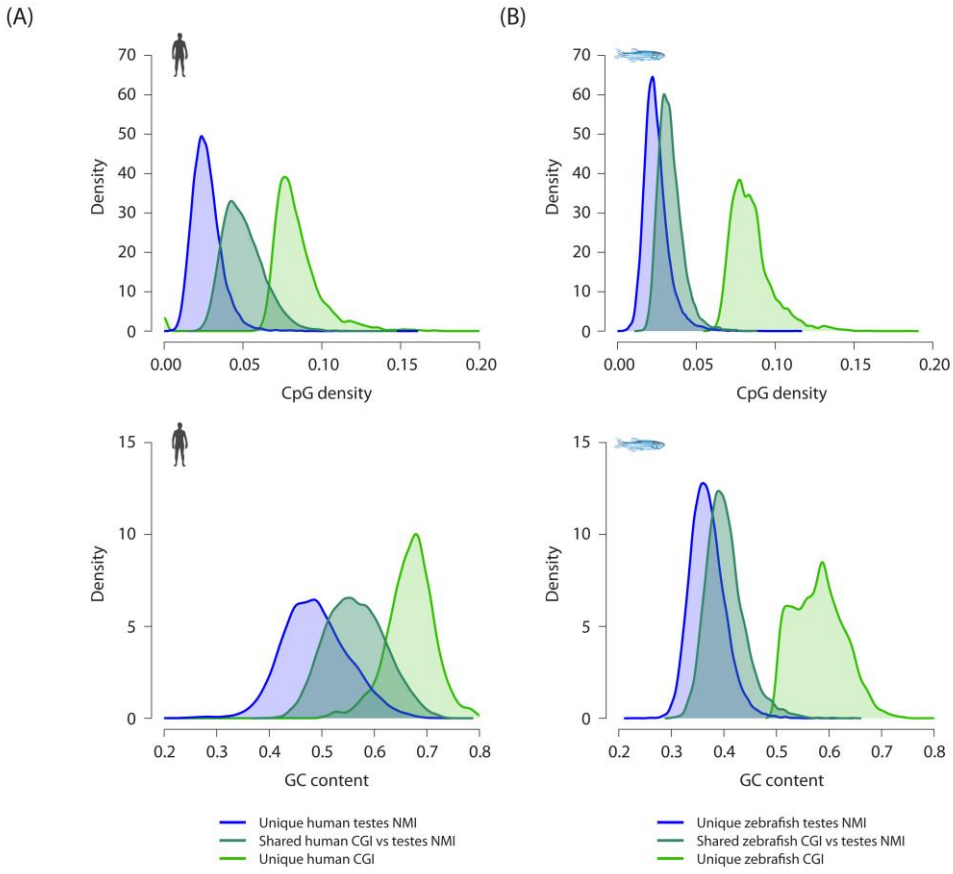
(A-B) Non-methylated DNA profiles in testis at two representative syntenic regions for seven vertebrate species. Genes are shown in black (improved annotation of gene TSSs using RNA-seq data is shown in red), CpG island predictions in green (CGI), and non-methylated DNA profiles are shown in blue. A phylogenetic tree (left) highlights the evolutionary relationship among the seven species. Dashed grey lines highlight the relationship between the gene TSSs across the species.

Appendix 4.4 Bio-CAP sequencing peak-calling

		Peaks				
		testis	**liver (male)**	**ES cell**	**stage 11-12**	**24 hpf**
Species	**human**	40697	35325			
	mouse	33806	18142	26115		
	platypus	40871	63946			
	chicken	19585	17279			
	lizard	12620	22684			
	frog	11099	21392		12910	
	zebrafish	26892	34983			28111

Numbers of peaks identified by MACS with the peak-calling threshold at 6-fold above background.

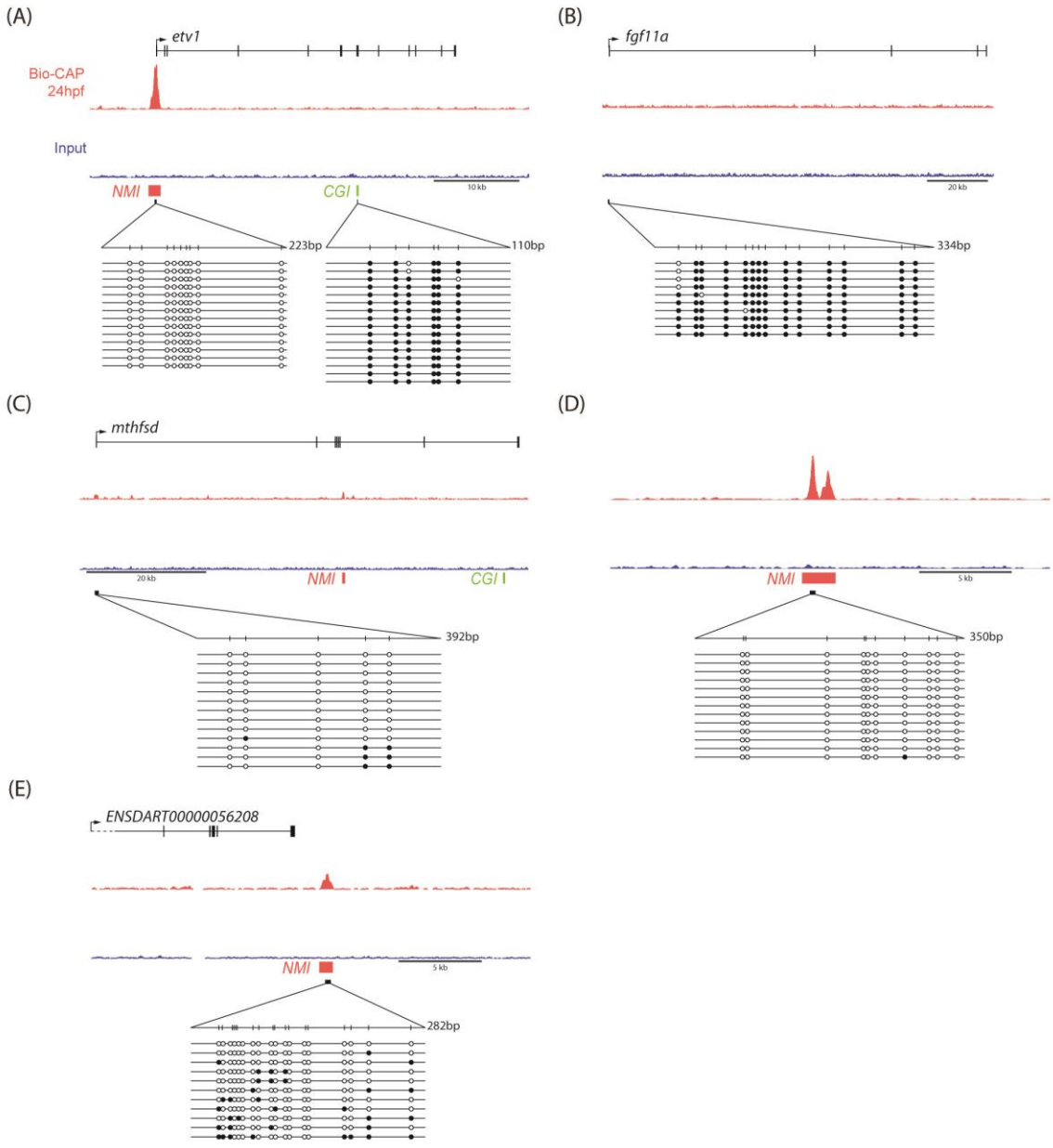
Appendix 4.5 Differential nucleotide composition of testis NMIs and predicted CGIs from human and zebrafish



(A) Frequency distribution plots of CpG density (upper) and GC content (lower) for intervals in human which are shared between or unique to CGI predictions and testis NMIs. (B) Frequency distribution plots of CpG density (upper) and GC content (lower) for intervals in zebrafish which are shared between or unique to CGI predictions and testis NMIs.

CGIs not overlapping an NMI (bright green) exhibit elevated CpG density and GC content, while unique NMIs not overlapping a CGI prediction exhibit lower CpG density and GC content than for intervals overlapped by both a CGI and an NMI. This therefore reveals why CGI predictions generally fail to identify NMIs, because not all NMIs fall within the nucleotide properties of CGIs. This distinction is greater for zebrafish compared to human.

Appendix 4.6 Bisulfite validation of Bio-CAP-seq peaks



(A) Two CGIs from the *etv2* gene, one of which overlaps a Bio-CAP peak (chr15 : 31,656,198 - 31,699,292), (B) the *fgf11* gene promoter (chr7 : 23,553,975 - 23,705,445), (C) the *mthfsd* gene promoter (chr18:30,804,636-30,911,755) and (D and E) two intergenic Bio-CAP peaks (chr1 : 21,120,139 - 21,168,393

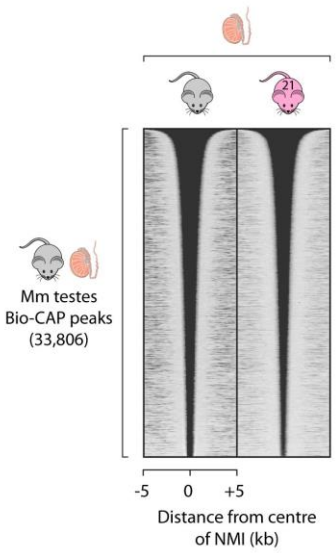
and chr8 : 2,025,545 - 2,056,774). The Bio-CAP-seq trace is shown in red and the matched input in blue with predicted CpG islands depicted in green.

Appendix 5.1 The presence of intergenic NMIs at a number of classes of non-coding transcription start sites

	Number of NMIs	
No. intergenic NMIs overlapping:	mouse ES cell	zebrafish 24hpf
miRNA	99	53
snRNA	46	22
snoRNA	28	13
rRNA	6	6
TOTAL intergenic NMIs	6071	6835

Intergenic NMIs from mouse ES cells and zebrafish 24hpf embryos were scored for overlap with the primary transcript TSSs of ncRNAs (miRNA, snRNA, snoRNA and rRNA).

Appendix 6.1 Introduction of human chromosome 21 into the mouse nucleus does not affect the non-methylated status of endogenous NMIs from mouse chromosomes



A heatmap of Bio-CAP signal from wildtype (left) and Tc1 mouse (right) testis tissues centred at all NMIs from wildtype mouse testis (33,806 peaks), ranked by length. The presence of human chromosome 21 does not appear to affect the size and distribution of NMIs in the Tc1 mouse.

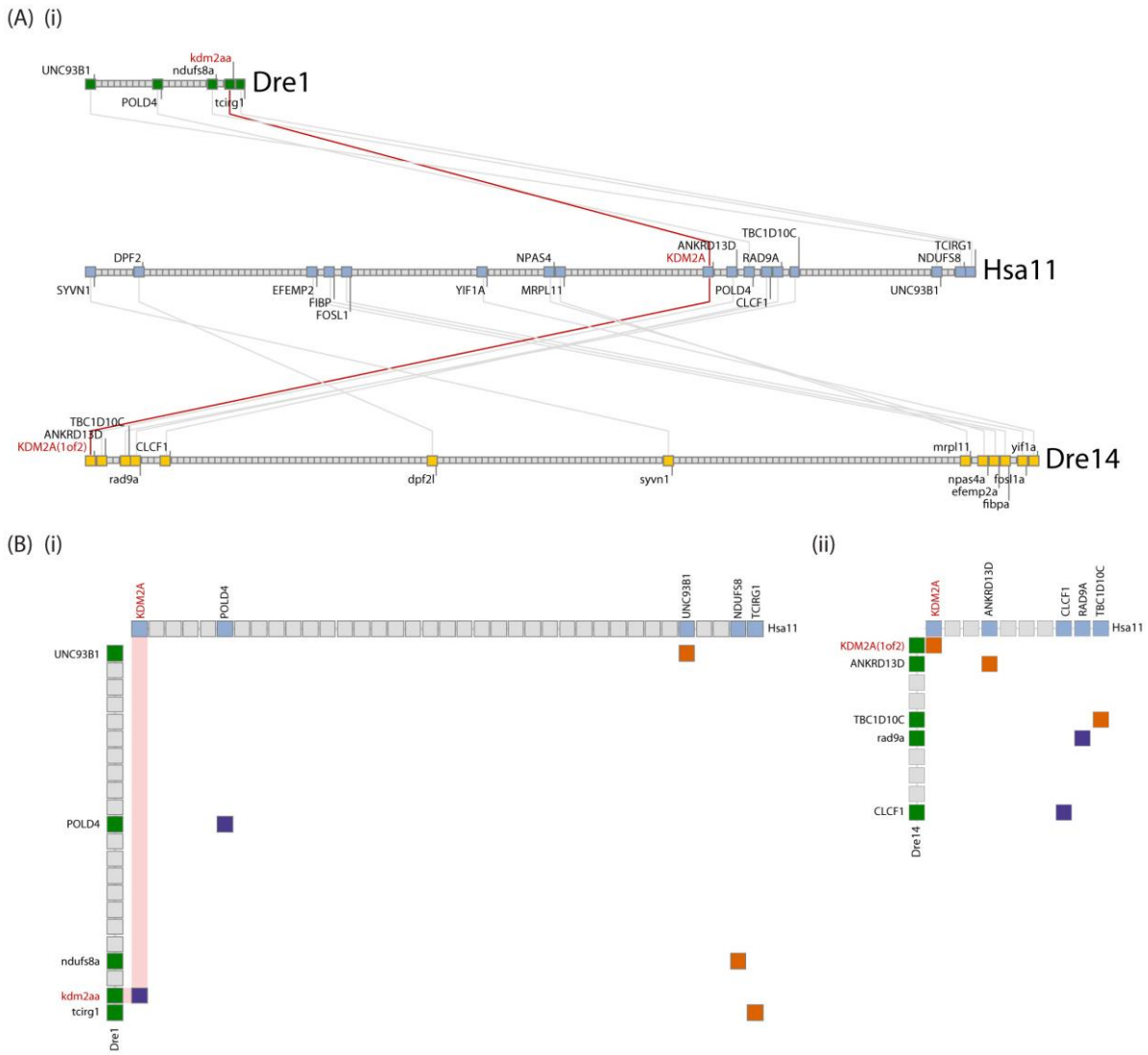
Appendix 6.2 Nucleotide properties of NMIs from zBAC3 in the zebrafish embryo or mouse ES cells

Interval	Chr	Start	End	GC %	Length	CpG count	CpG density
1	chr13	31867525	31872821	41.39	5296	172	0.0325
2	chr13	31910343	31912214	35.28	1871	48	0.0257
3	chr13	31937720	31939155	38.12	1435	35	0.0244
4	chr13	31939623	31941927	37.02	2304	68	0.0295
5	chr13	31980158	31982253	37.28	2095	51	0.0243
6	chr13	32052292	32054133	38.24	1841	64	0.0348
7	chr13	31904837	31905792	33.72	955	13	0.0136
8	chr13	31919889	31920744	42.34	855	25	0.0292
9	chr13	32033361	32034655	46.37	1294	39	0.0301
10	chr13	32049288	32050469	36.58	1181	26	0.0220

A table describing the nucleotide properties of non-methylated islands identified from a zebrafish BAC incorporated into the mouse genome.

Chapter 7

Appendix 7.1 Synteny analysis of human *KDM2A* reveals *zkdm2aa* and *zkdm2ab* as zebrafish gene orthologues



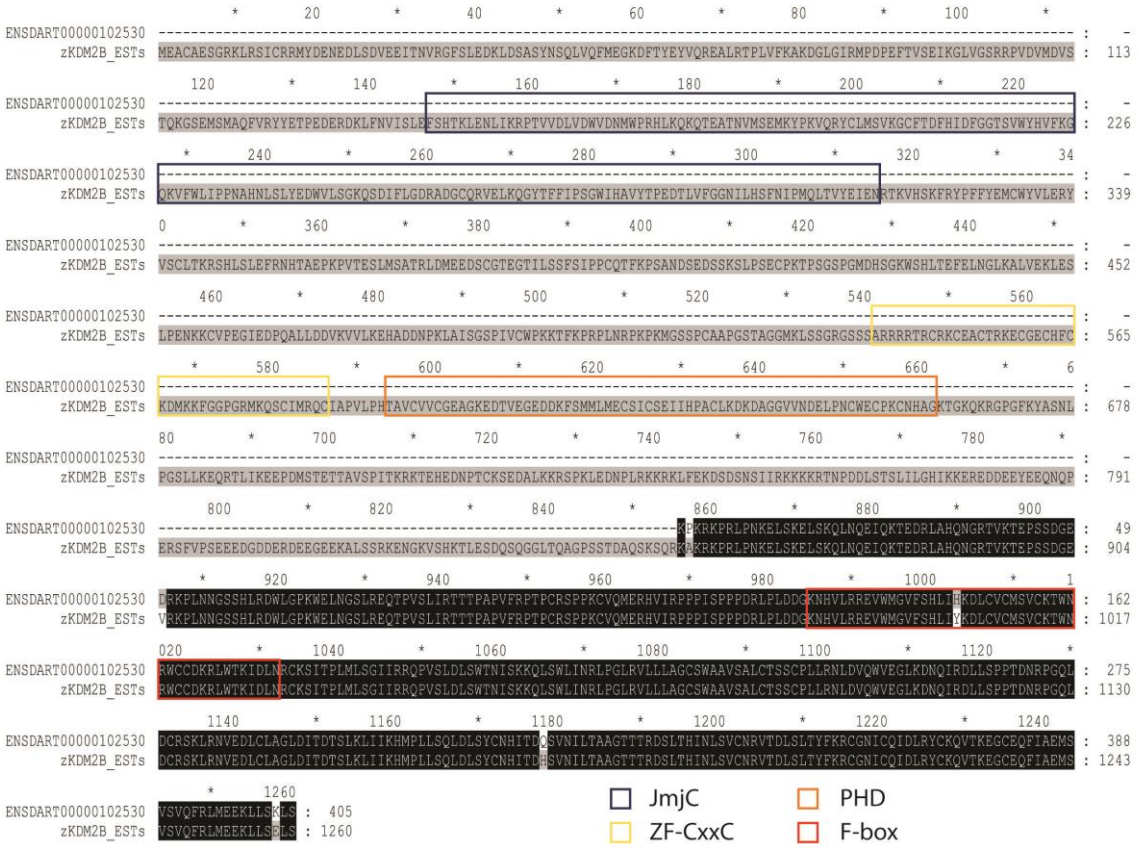
(A) Composite gene trace illustrating the synteny between the human KDM2A gene on human chromosome 11 and *zkdm2aa* and *zkdm2ab* on zebrafish chromosomes 1 and 14 respectively (B) Gene homology matrix of human KDM2A and neighbouring genes compared to the zebrafish *zkdm2aa* (i) and *zkdm2ab* (ii) gene loci. Plots were generated using the Synteny Database (http://syntenydb.uoregon.edu/synteny_db/)

Appendix 7.2 Synteny analysis of human *KDM2B* reveals *zkdm2ba* and *zkd2mbb* as gene orthologues from the zebrafish genome

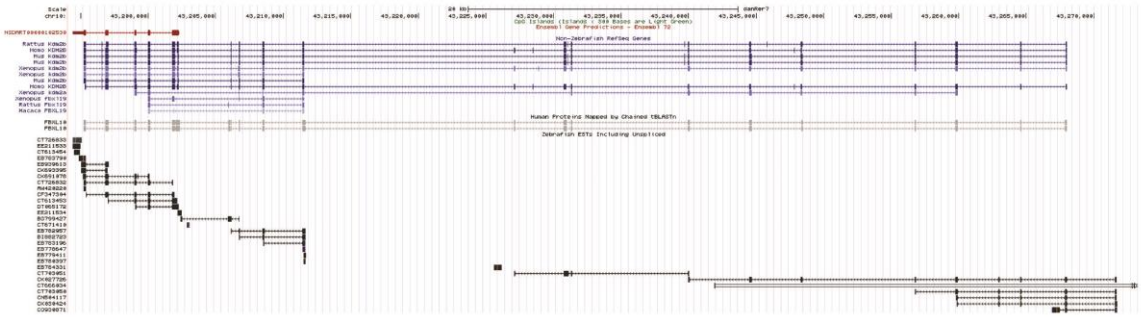
(A) Composite gene trace illustrating the synteny between the human KDM2B gene on human chromosome 12 and *zkdm2ba* and *zkdm2bb* on zebrafish chromosomes 10 and 8 respectively. (B) Gene homology matrix of human *KDM2B* and neighbouring genes compared to the zebrafish *zkdm2ba* (i) and *zkdm2bb* (ii) gene loci. Plots were generated using the Synteny Database (http://syntenydb.uoregon.edu/synteny_db/)

Appendix 7.3 Alignment of translated *zkdm2b* ESTs and Ensembl ENSDART00000102530

(A)

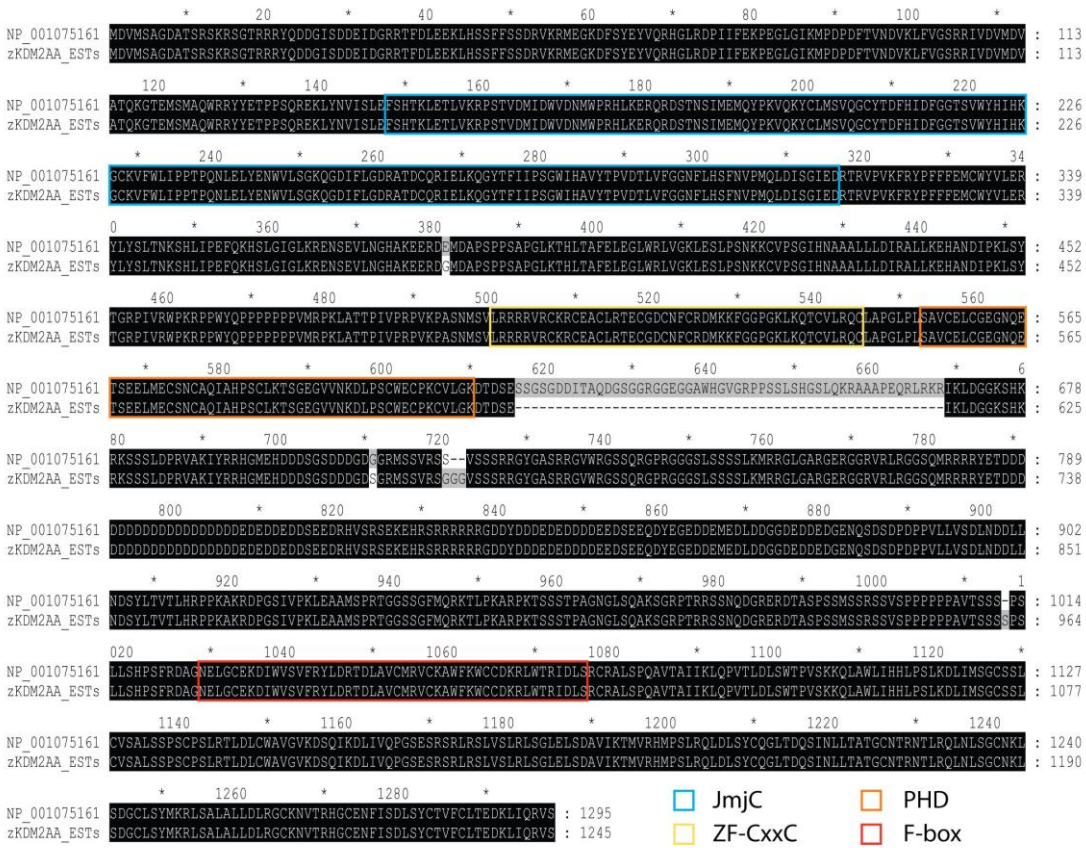


(B)



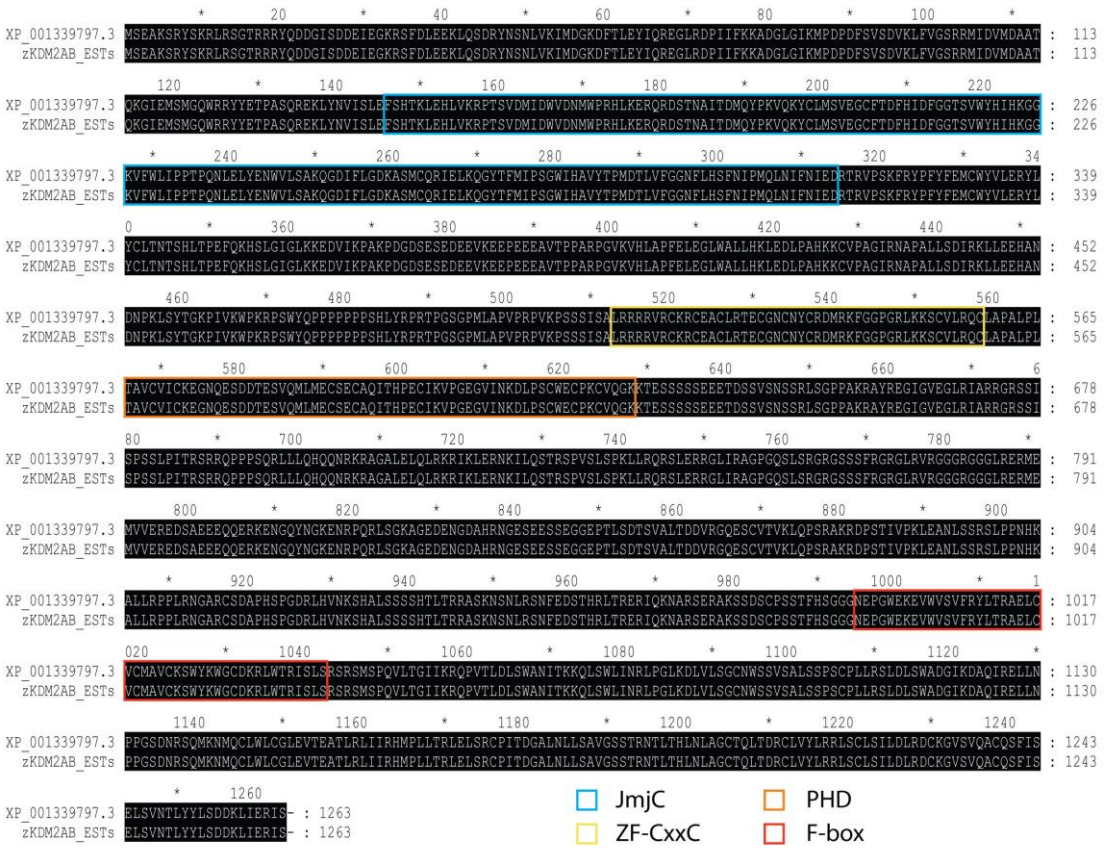
(A) Alignment of amino acid sequence of zKdm2b from this thesis (translation of the cloned *zkdm2b* construct, plasmid 409 generated from ESTs) with the sequence from Ensembl (ENSDART00000102530). (B) A snapshot from the UCSC genome-browser (Zv9, chr10:43,194,095-43,273,265) of the *zkdm2b* locus showing the Ensembl transcript (red), other RefSeq genes (blue) and spliced zebrafish ESTs (black). Illustrating that EST evidence can be detected for the majority of Kdm2b exons predicted by homology to other species. The *zkdm2b* transcript is transcribed right-to-left.

Appendix 7.4 Alignment of translated *zkdm2aa* ESTs with NCBI NP_001075161



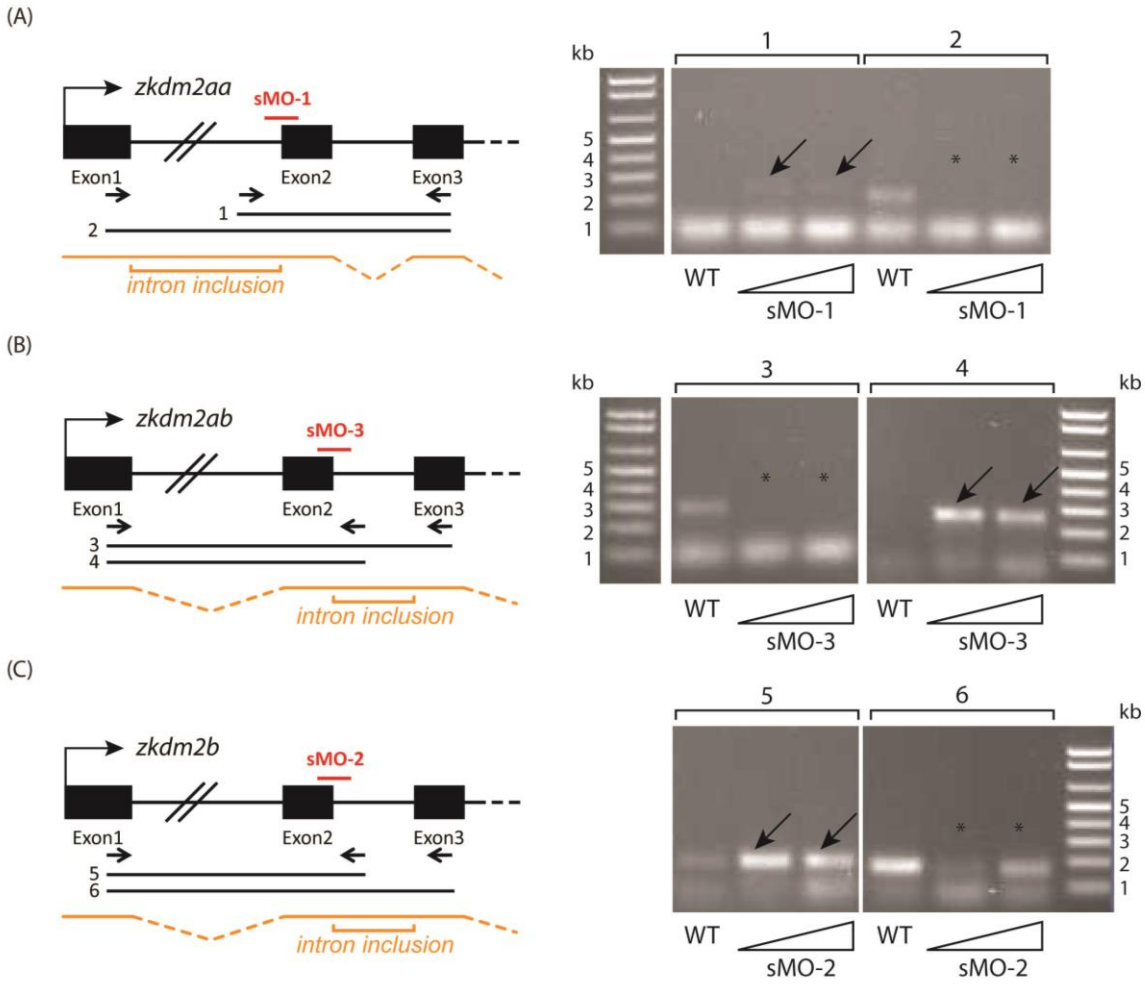
Alignment of amino acid sequence of zKdm2aa from this thesis (translation of the cloned *zkdm2aa* construct, plasmid 407 generated from ESTs) with the sequence from NCBI (NP_001075161).

Appendix 7.5 Alignment of translated *zkdm2ab* ESTs with NCBI XP_001339797.3



Alignment of amino acid sequence of zKdm2ab from this thesis (translation of the *zkdm2aa* EST construct, plasmid 396) with the sequence from NCBI (XP_001339797.3).

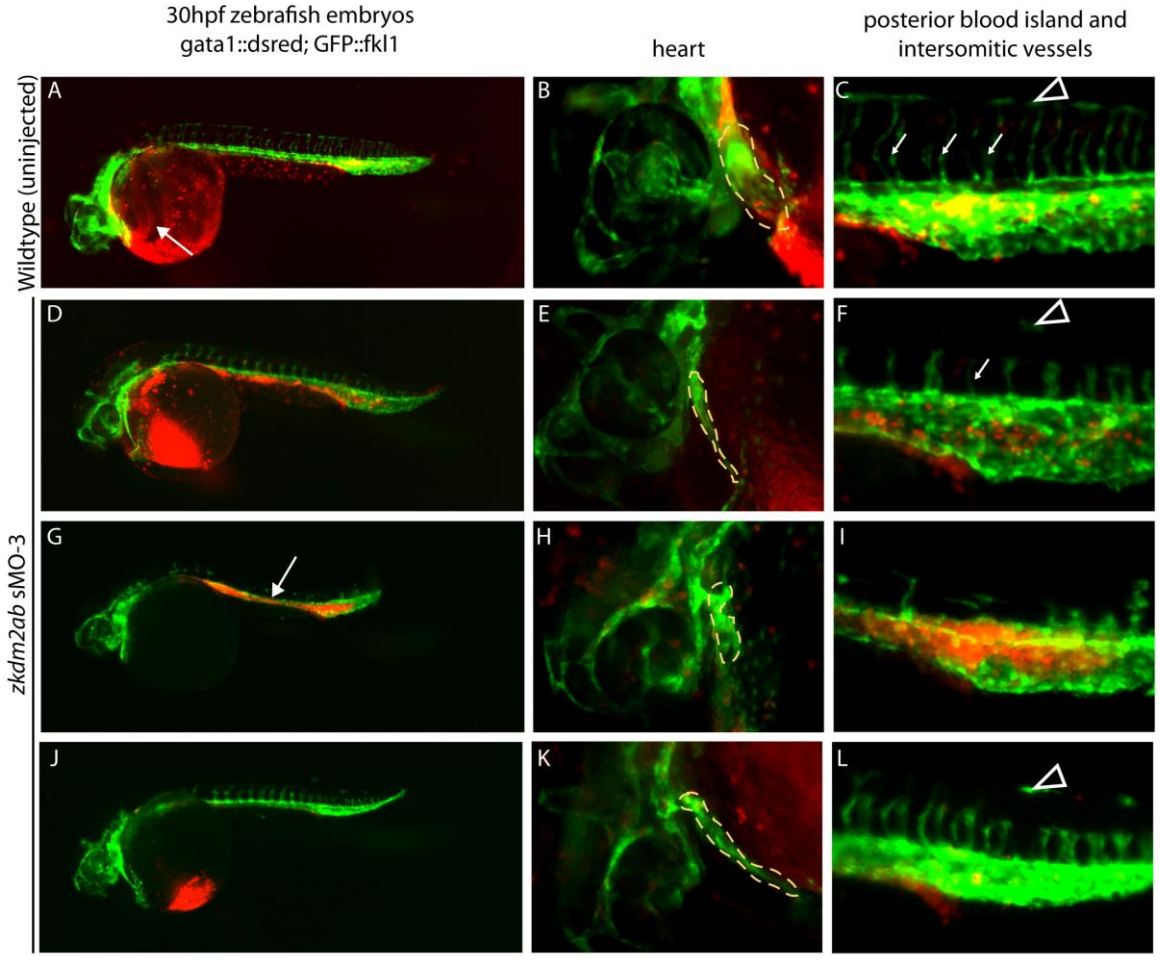
Appendix 7.6 Validation of *zkdm2* splice morpholinos by RT-PCR



(A-C) qRT-PCR was performed on embryos injected with two doses of splice-blocking morpholino for each morpholino, *zkdm2aa* (A aMO-1), *zkdm2ab* (B aMO-3) and *zkdm2b* (C aMO-2). Schematics of genes *zkdm2aa*, *zkdm2ab* and *zkdm2b* are shown with splice-morpholino binding sites and primers used for RT-PCR indicated (left). RT-PCR products were separated by gel electrophoresis. RT-PCR products observed in WT embryos were lost upon sMO injection (RT-PCR products 2, 3 and 6, asterisk). RT-PCR products including intronic sequence become detectable following splice-morpholino injection (RT-PCR products 1, 4 and 5, arrow). Splice morpholinos can cause exon skipping, intron inclusion or use of cryptic splice sites, all of which can introduce an early stop codon, lead to a frameshift, or cause the mRNA to be targeted for nonsense mediated decay. The RT-PCR analysis suggests that the splice-blocking morpholinos have caused intron inclusion of intron 1 for *zkdm2aa* (sMO-1) and inclusion of intron 2 for *zkdm2ab* (sMO-3) and *zkdm2b* (sMO-2) (orange).

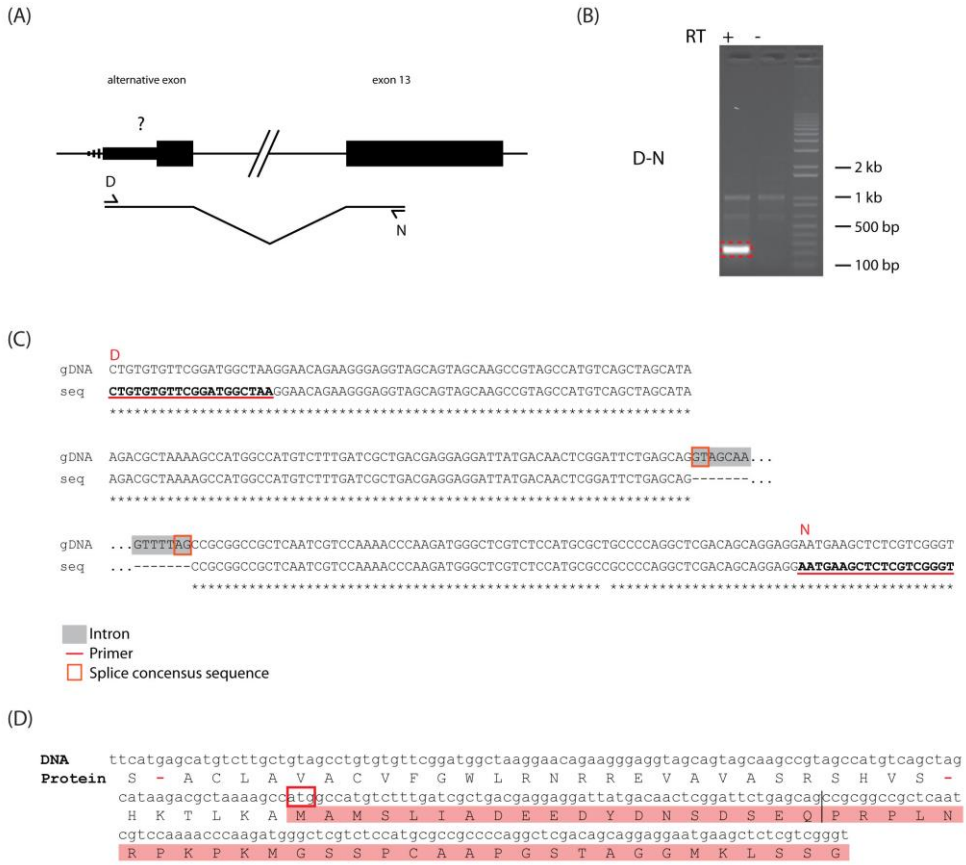
Appendix 7.7 *zkdm2ab* morphant embryos exhibit heart and vascular defects and aberrant blood cell accumulation

In order to easily visualise the *zkdm2ab* splice morpholino phenotype, sMO-3 was injected into the gata1:dsred; fkl1:GFP transgenic line. The gata1 promoter drives dsred expression in erythroid cells while the flk1 promoter drives gfp expression in the vasculature. Example images are shown for *zkdm2ab* morphant embryos which exhibit pooling of erythrocytes in the trunk or on the yolk, posterior defects such as bent tails, vasculature defects and string hearts at 30 hpf. No earlier morphological phenotypes were observed for any of the titrated morphant embryos (5, 10, 15 and 30 ng splice morpholino).



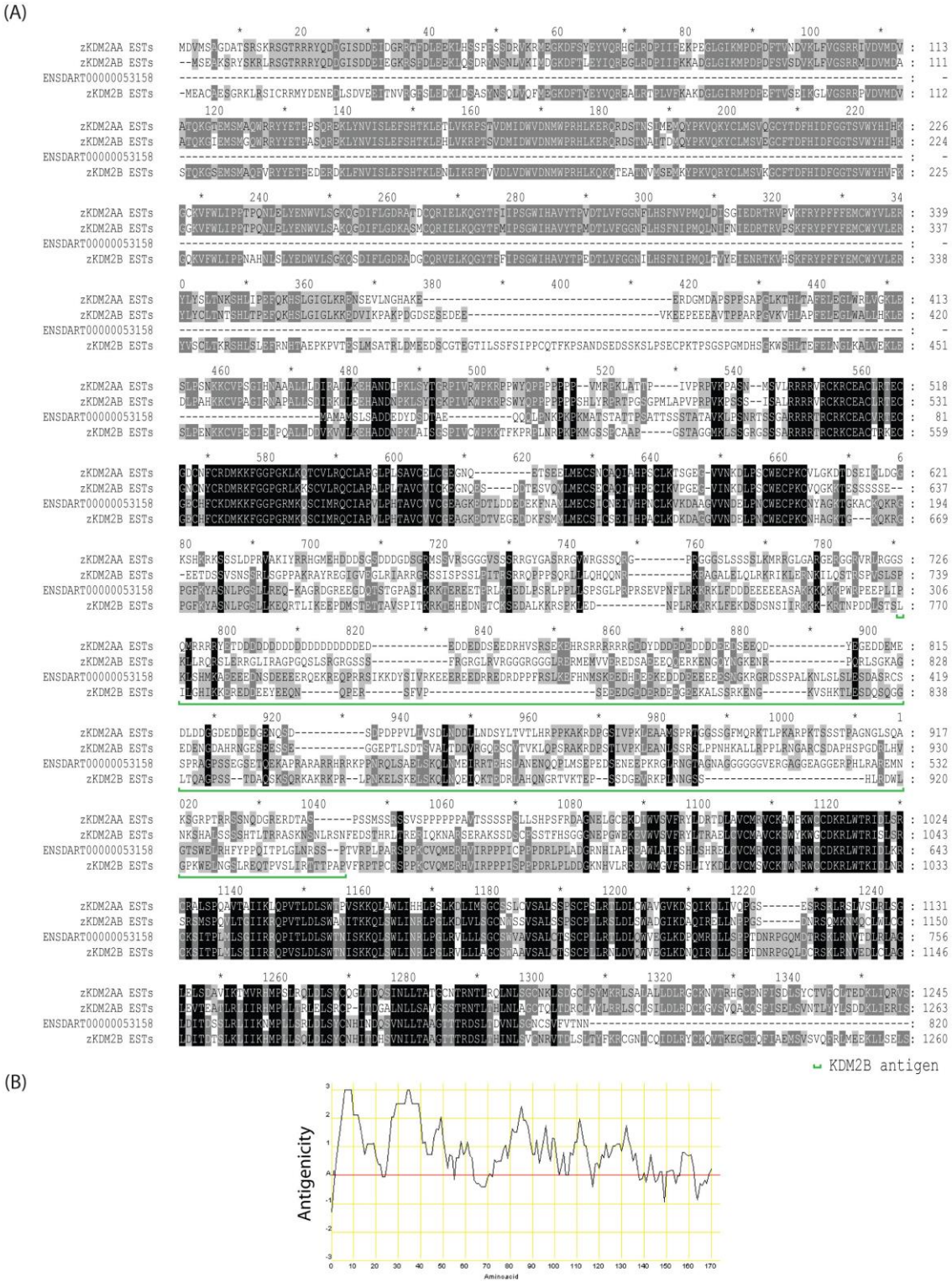
(A) Wildtype embryos have strong circulation (arrow, A) expanded hearts (B) and ISVs (2 C) which extend to the dorsal side of the embryo and have fused to form the DLAV (1 C). The PBI is well structured and has cells circulating through it (3 C). *zkdm2ab* morphants have similar but variable phenotypes, four examples of which have been shown here. Hearts in the morphant embryo are thin and string-like (E, H) or aberrant in shape (K) and despite the heart beating, there is no or little blood flow (from watching the live embryos). Blood is seen to blood along the trunk in the main vessels or in the PBI (arrowheads D, J) or on the yolk (arrowhead G). ISVs in the morphant embryos are stunted in length (2 F,IL) and fail to form the DLAV (1 F,I,L). Furthermore the PBI appears less structured than in the wildtype animal (3 F, L) or the vasculature here is dense (3 I).

Appendix 7.8 Sequencing the *zkdm2b* alternative start site



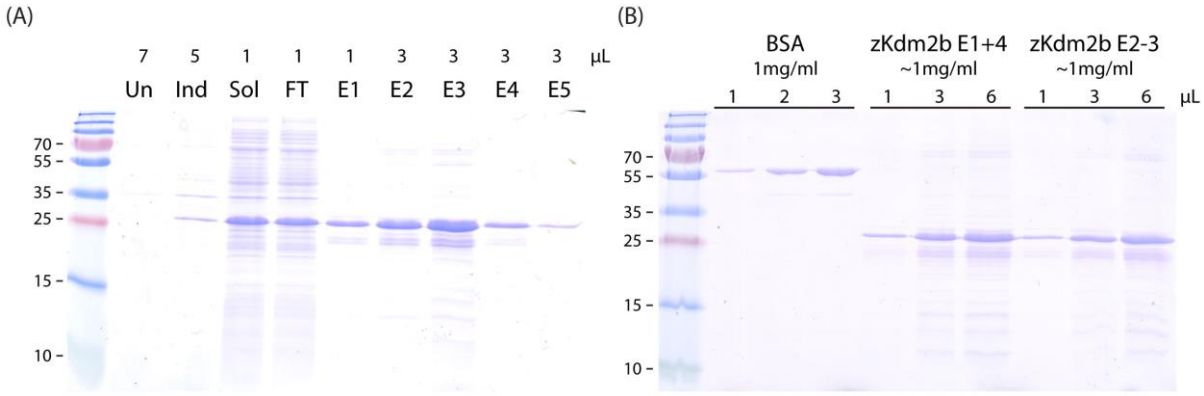
(A) To determine the splice junctions for the alternative exon in *zkdm2b*, the qRT-PCR product from D-N (alternative exon to exon 13) was TA cloned and sequenced. (B) Amplification of D-N, indicating the excised band (red dashed line) to be TA cloned. PCRs were performed with cDNA made with and without reverse transcriptase (RT + and -). (C) Alignment of sequenced D-N PCR product (seq) with genomic DNA. Introns are highlighted in grey with consensus splice sequences boxed in orange, introns have been truncated as indicated by an ellipsis (...). Primers are highlighted in red. Both splice junctions are in frame (i.e. the end and beginning of exon 12 and 13 encode codon triplets. (D) Intron 12 from primer E was translated in frame with the splicing event to exon 13. This allows prediction of the start ATG (red box) for translation as the first ATG following a STOP codon (red dash). The putative coding sequence is highlighted in pink and the splice junction is indicated by a vertical line.

Appendix 7.9 Alignment of zKdm2 proteins, highlighting the zKdm2b antigen



(A) Alignment of zKdm2aa, zKdm2ab, zKdm2ba (Ensembl sequence, ENSDART00000053158) and zKdm2b amino acid sequence highlighting the zKdm2b antigen (green). (B) Antigenicity plot for the zKdm2b antigen (http://www.bioinformatics.org/JaMBW/3/1/7/).

Appendix 7.10 Production of the zKdm2b antigen



(A) Expression and purification of His-tagged zKdm2b antigen from *E. coli* using a Ni-NTA purification strategy. Samples were taken before and after induction with IPTG (Un and Ind respectively), after cell lysis and removal of insoluble pellet (Sol), the flow-through from the Ni-NTA column (FT) and a number of 250 mM imidazole elutions (E1 - E5). Volumes loaded are indicated. (B) Dialysed elutions 1+4 and 2-3 were quantified by Bradford reagent and diluted to approximately 1 mg/ml. SDS-PAGE gel of 1, 2 and 3 µg BSA alongside ~1, 3 and 6 µg of zKdm2b elutions 1+4 and 2-3.

Chapter 10 -- References

Aanes, H., Winata, C.L., Lin, C.H., Chen, J.P., Srinivasan, K.G., Lee, S.G., Lim, A.Y., Hajan, H.S., Collas, P., Bourque, G., Gong, Z., Korzh, V., Alestrom, P., and Mathavan, S. (2011). Zebrafish mRNA sequencing deciphers novelties in transcriptome dynamics during maternal to zygotic transition. Genome Res *21*, 1328-1338.

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

Abramowitz, L.K., and Bartolomei, M.S. (2012). Genomic imprinting: recognition and marking of imprinted loci. Current opinion in genetics & development *22*, 72-78.

Achour, M., Jacq, X., Ronde, P., Alhosin, M., Charlot, C., Chataigneau, T., Jeanblanc, M., Macaluso, M., Giordano, A., Hughes, A.D., Schini-Kerth, V.B., and Bronner, C. (2007). The interaction of the SRA domain of ICBP90 with a novel domain of DNMT1 is involved in the regulation of VEGF gene expression. Oncogene *27*, 2187-2197.

Aday, A.W., Zhu, L.J., Lakshmanan, A., Wang, J., and Lawson, N.D. (2011). Identification of cis regulatory features in the embryonic zebrafish genome through large-scale profiling of H3K4me1 and H3K4me3 binding sites. Dev Biol *357*, 450-462.

Aerts, S., Thijs, G., Dabrowski, M., Moreau, Y., and De Moor, B. (2004). Comprehensive analysis of the base composition around the transcription start site in Metazoa. BMC Genomics *5*, 34.

Agathon, A., Thisse, C., and Thisse, B. (2003). The molecular nature of the zebrafish tail organizer. Nature *424*, 448-452.

Aïssani, B., and Bernardi, G. (1991). CpG islands: features and distribution in the genomes of vertebrates. Gene *106*, 173-183.

Akkers, R.C., van Heeringen, S.J., Jacobi, U.G., Janssen-Megens, E.M., Francoijs, K.J., Stunnenberg, H.G., and Veenstra, G.J. (2009). A hierarchy of H3K4me3 and H3K27me3 acquisition in spatial gene regulation in Xenopus embryos. Dev Cell *17*, 425-434.

Akkers, R.C., van Heeringen, S.J., Manak, J.R., Green, R.D., Stunnenberg, H.G., and Veenstra, G.J.C. (2010). ChIP-chip designs to interrogate the genome of Xenopus embryos for transcription factor binding and epigenetic regulation. PloS one *5*, e8820.

Albalat, R., Marti-Solans, J., and Canestro, C. (2012). DNA methylation in amphioxus: from ancestral functions to new roles in vertebrates. Briefings in functional genomics *11*, 142-155.

Allen, M.D., Grummitt, C.G., Hilcenko, C., Min, S.Y., Tonkin, L.M., Johnson, C.M., Freund, S.M., Bycroft, M., and Warren, A.J. (2006). Solution structure of the nonmethyl-CpG-binding CXXC domain of the leukaemia-associated MLL histone methyltransferase. EMBO J *25*, 4503-4512.

Almeida, R.D., Loose, M., Sottile, V., Matsa, E., Denning, C., Young, L., Johnson, A.D., Gering, M., and Ruzov, A. (2012a). 5-hydroxymethyl-cytosine enrichment of non-committed cells is not a universal feature of vertebrate development. Epigenetics *7*, 383-389.

Almeida, R.D., Sottile, V., Loose, M., De Sousa, P.A., Johnson, A.D., and Ruzov, A. (2012b). Semi-quantitative immunohistochemical detection of 5-hydroxymethyl-cytosine reveals conservation of its tissue distribution between amphibians and mammals. Epigenetics *7*, 137-140.

Amemiya, C.T., Alfoldi, J., Lee, A.P., Fan, S., Philippe, H., MacCallum, I., Braasch, I., Manousaki, T., Schneider, I., Rohner, N., Organ, C., Chalopin, D., Smith, J.J., Robinson, M., Dorrington, R.A., Gerdol, M., Aken, B., Biscotti, M.A., Barucca, M., Baurain, D., Berlin, A.M., Blatch, G.L., Buonocore, F., Burmester, T., Campbell, M.S., Canapa, A., Cannon, J.P., Christoffels, A., De Moro, G., Edkins, A.L., Fan, L., Fausto, A.M., Feiner, N., Forconi, M., Gamieldien, J., Gnerre, S., Gnirke, A., Goldstone, J.V., Haerty, W., Hahn, M.E., Hesse, U., Hoffmann, S., Johnson, J., Karchner, S.I., Kuraku, S., Lara, M., Levin, J.Z., Litman, G.W., Mauceli, E., Miyake, T., Mueller, M.G., Nelson, D.R., Nitsche, A., Olmo, E., Ota, T., Pallavicini, A., Panji, S., Picone, B., Ponting, C.P., Prohaska, S.J., Przybylski, D., Saha, N.R., Ravi, V., Ribeiro, F.J., Sauka-Spengler, T., Scapigliati, G., Searle, S.M.J., Sharpe, T., Simakov, O., Stadler, P.F., Stegeman, J.J., Sumiyama, K., Tabbaa, D., Tafer, H., Turner-Maier, J., van Heusden, P., White, S., Williams, L., Yandell, M., Brinkmann, H., Volff, J.-N., Tabin, C.J., Shubin, N., Schartl, M., Jaffe, D.B., Postlethwait,

J.H., Venkatesh, B., Di Palma, F., Lander, E.S., Meyer, A., and Lindblad-Toh, K. (2013). The African coelacanth genome provides insights into tetrapod evolution. Nature *496*, 311-316.

Amir, R.E., Van den Veyver, I.B., Wan, M., Tran, C.Q., Francke, U., and Zoghbi, H.Y. (1999). Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. Nature genetics *23*, 185-188.

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome biology *11*, R106.

Andersen, I.S., Reiner, A.H., Aanes, H., Alestrom, P., and Collas, P. (2012). Developmental features of DNA methylation during activation of the embryonic zebrafish genome. Genome Biol *13*, R65.

Anderson, R.M., Bosch, J.A., Goll, M.G., Hesselson, D., Dong, P.D.S., Shin, D., Chi, N.C., Shin, C.H., Schlegel, A., Halpern, M., and Stainier, D.Y.R. (2009). Loss of Dnmt1 catalytic activity reveals multiple roles for DNA methylation during pancreas development and regeneration. Developmental Biology *334*, 213-223.

Antequera, F. (2003). Structure, function and evolution of CpG island promoters. Cell Mol Life Sci *60*, 1647-1658.

Antequera, F. (2004). Genomic specification and epigenetic regulation of eukaryotic DNA replication origins. EMBO J *23*, 4365-4370.

Antequera, F., and Bird, A. (1993). Number of CpG islands and genes in human and mouse. Proc Natl Acad Sci U S A *90*, 11995-11999.

Antequera, F., and Bird, A. (1999). CpG islands as genomic footprints of promoters that are associated with replication origins. Current biology : CB *9*, R661-667.

Antequera, F., Macleod, D., and Bird, A.P. (1989). Specific protection of methylated CpGs in mammalian nuclei. Cell *58*, 509-517.

Antequera, F., Tamame, M., Villanueva, J.R., and Santos, T. (1984). DNA methylation in the fungi. J Biol Chem *259*, 8033-8036.

Antonarakis, S.E., Lyle, R., Dermitzakis, E.T., Reymond, A., and Deutsch, S. (2004). Chromosome 21 and down syndrome: from genomics to pathophysiology. Nat Rev Genet *5*, 725-738.

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

Arita, K., Isogai, S., Oda, T., Unoki, M., Sugita, K., Sekiyama, N., Kuwata, K., Hamamoto, R., Tochio, H., Sato, M., Ariyoshi, M., and Shirakawa, M. (2012). Recognition of modification status on a histone H3 tail by linked histone reader modules of the epigenetic regulator UHRF1. Proc Natl Acad Sci U S A *109*, 12950-12955.

Armstrong, K.M., Bermingham, E.N., Bassett, S.A., Treloar, B.P., Roy, N.C., and Barnett, M.P.G. (2011). Global DNA methylation measurement by HPLC using low amounts of DNA. Biotechnology Journal *6*, 113-117.

Avvakumov, G.V., Walker, J.R., Xue, S., Li, Y., Duan, S., Bronner, C., Arrowsmith, C.H., and Dhe-Paganon, S. (2008). Structural basis for recognition of hemi-methylated DNA by the SRA domain of human UHRF1. Nature *455*, 822-825.

Bach, C., Mueller, D., Buhl, S., Garcia-Cuellar, M.P., and Slany, R.K. (2009). Alterations of the CxxC domain preclude oncogenic activation of mixed-lineage leukemia 2. Oncogene *28*, 815-823.

Ball, M.P., Li, J.B., Gao, Y., Lee, J.H., LeProust, E.M., Park, I.H., Xie, B., Daley, G.Q., and Church, G.M. (2009). Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. Nat Biotechnol *27*, 361-368.

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

Barbosa-Morais, N.L., Irimia, M., Pan, Q., Xiong, H.Y., Gueroussov, S., Lee, L.J., Slobodeniuc, V., Kutter, C., Watt, S., Colak, R., Kim, T., Misquitta-Ali, C.M., Wilson, M.D., Kim, P.M., Odom, D.T., Frey, B.J., and Blencowe, B.J. (2012). The evolutionary landscape of alternative splicing in vertebrate species. Science *338*, 1587-1593.

Barker, D., Schafer, M., and White, R. (1984). Restriction Sites Containing Cpg Show a Higher Frequency of Polymorphism in Human DNA. Cell *36*, 131-138.

Bartolomei, M.S., Zemel, S., and Tilghman, S.M. (1991). Parental imprinting of the mouse H19 gene. Nature *351*, 153-155.

Baubec, T., Ivanek, R., Lienert, F., and Schubeler, D. (2013). Methylation-dependent and -independent genomic targeting principles of the MBD protein family. Cell *153*, 480-492.

Baylin, S.B., and Jones, P.A. (2011). A decade of exploring the cancer epigenome - biological and translational implications. Nat Rev Cancer *11*, 726-734.

Becker, M., Muller, S., Nellen, W., Jurkowski, T.P., Jeltsch, A., and Ehrenhofer-Murray, A.E. (2012). Pmt1, a Dnmt2 homolog in Schizosaccharomyces pombe, mediates tRNA methylation in response to nutrient signaling. Nucleic Acids Res *40*, 11648-11658.

Beckett, D., Kovaleva, E., and Schatz, P.J. (1999). A minimal peptide substrate in biotin holoenzyme synthetase-catalyzed biotinylation. Protein Science *8*, 921-929.

Bedell, V.M., Westcot, S.E., and Ekker, S.C. (2011). Lessons from morpholino-based screening in zebrafish. Briefings in functional genomics *10*, 181-188.

Belgard, T.G., Marques, A.C., Oliver, P.L., Abaan, H.O., Sirey, T.M., Hoerder-Suabedissen, A., Garcia-Moreno, F., Molnar, Z., Margulies, E.H., and Ponting, C.P. (2011). A transcriptomic atlas of mouse neocortical layers. Neuron *71*, 605-616.

Bernstein, B.E., Kamal, M., Lindblad-Toh, K., Bekiranov, S., Bailey, D.K., Huebert, D.J., McMahon, S., Karlsson, E.K., Kulbokas, E.J., 3rd, Gingeras, T.R., Schreiber, S.L., and Lander, E.S. (2005). Genomic maps and comparative analysis of histone modifications in human and mouse. Cell *120*, 169-181.

Bernstein, B.E., Stamatoyannopoulos, J.A., Costello, J.F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M.A., Beaudet, A.L., Ecker, J.R., Farnham, P.J., Hirst, M., Lander, E.S., Mikkelsen, T.S., and Thomson, J.A. (2010). The NIH Roadmap Epigenomics Mapping Consortium. Nat Biotechnol *28*, 1045-1048.

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

Bestor, T.H., and Ingram, V.M. (1983). Two DNA methyltransferases from murine erythroleukemia cells: purification, sequence specificity, and mode of interaction with DNA. Proc Natl Acad Sci U S A *80*, 5559-5563.

Betchaku, T., and Trinkaus, J.P. (1978). Contact relations, surface activity, and cortical microfilaments of marginal cells of the enveloping layer and of the yolk syncytial and yolk cytoplasmic layers of Fundulus before and during epiboly. Journal of Experimental Zoology *206*, 381-426.

Beumer, J., Parise, R., Newman, E., Doroshow, J., Synold, T., Lenz, H.-J., and Egorin, M. (2008). Concentrations of the DNA methyltransferase inhibitor 5-fluoro-2'-deoxycytidine (FdCyd) and its cytotoxic metabolites in plasma of patients treated with FdCyd and tetrahydrouridine (THU). Cancer Chemother Pharmacol *62*, 363-368.

Bian, C., Xu, C., Ruan, J., Lee, K.K., Burke, T.L., Tempel, W., Barsyte, D., Li, J., Wu, M., Zhou, B.O., Fleharty, B.E., Paulson, A., Allali-Hassani, A., Zhou, J.Q., Mer, G., Grant, P.A., Workman, J.L., Zang, J., and Min, J. (2011). Sgf29 binds histone H3K4me2/3 and is required for SAGA complex recruitment and histone H3 acetylation. EMBO J *30*, 2829-2842.

Bienko, M., Crosetto, N., Teytelman, L., Klemm, S., Itzkovitz, S., and van Oudenaarden, A. (2013). A versatile genome-scale PCR-based pipeline for high-definition DNA FISH. Nat Meth *10*, 122-124.

Bill, B.R., Petzold, A.M., Clark, K.J., Schimmenti, L.A., and Ekker, S.C. (2009). A primer for morpholino use in zebrafish. Zebrafish *6*, 69-77.

Bilodeau, S., Kagey, M.H., Frampton, G.M., Rahl, P.B., and Young, R.A. (2009). SetDB1 contributes to repression of genes encoding developmental regulators and maintenance of ES cell state. Genes & development *23*, 2484-2489.

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

Bird, A. (2007). Perceptions of epigenetics. Nature *447*, 396-398.

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

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

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

Bird, A.P. (1987). CpG islands as gene markers in the vertebrate nucleus. Trends in Genetics *3*, 342-347.

Bird, A.P. (1995). Gene number, noise reduction and biological complexity. Trends Genet *11*, 94-100.

Bird, A.P., Taggart, M.H., Nicholls, R.D., and Higgs, D.R. (1987). Non-methylated CpG-rich islands at the human alpha-globin locus: implications for evolution of the alpha-globin pseudogene. EMBO J *6*, 999-1004.

Bird, A.P., and Wolffe, A.P. (1999). Methylation-Induced Repression— Belts, Braces, and Chromatin. Cell *99*, 451-454.

Birke, M., Schreiner, S., Garcia-Cuellar, M.P., Mahr, K., Titgemeyer, F., and Slany, R.K. (2002). The MT domain of the proto-oncoprotein MLL binds to CpG-containing DNA and discriminates against methylation. Nucleic Acids Res *30*, 958-965.

Blackledge, N.P., and Klose, R. (2011). CpG island chromatin: a platform for gene regulation. Epigenetics *6*, 147-152.

Blackledge, N.P., Zhou, J.C., Tolstorukov, M.Y., Farcas, A.M., Park, P.J., and Klose, R.J. (2010). CpG islands recruit a histone H3 lysine 36 demethylase. Molecular Cell *38*, 179-190.

Bock, C., Beerman, I., Lien, W.-H., Smith, Zachary D., Gu, H., Boyle, P., Gnirke, A., Fuchs, E., Rossi, Derrick J., and Meissner, A. (2012). DNA Methylation Dynamics during In Vivo Differentiation of Blood and Skin Stem Cells. Molecular Cell *47*, 633-647.

Bock, C., Tomazou, E.M., Brinkman, A.B., Muller, F., Simmer, F., Gu, H.C., Jager, N., Gnirke, A., Stunnenberg, H.G., and Meissner, A. (2010). Quantitative comparison of genome-wide DNA methylation mapping technologies. Nat Biotechnol *28*, 1106-U1196.

Bogdanovic, O., Long, S.W., Heeringen, S.J.V., Res, G., Brinkman, A.B., Stunnenberg, H.G., Jones, P.L., and Veenstra, G.J.C. (2011). Temporal uncoupling of the DNA methylome and transcriptional repression during embryogenesis. Genome Research.

Bogdanovic, O., and Veenstra, G.J. (2009). DNA methylation and methyl-CpG binding proteins: developmental requirements and function. Chromosoma *118*, 549-565.

Borgel, J., Guibert, S., Li, Y.F., Chiba, H., Schubeler, D., Sasaki, H., Forne, T., and Weber, M. (2010). Targets and dynamics of promoter DNA methylation during early mouse development. Nature genetics *42*, 1093-U1090.

Borowczyk, E., Mohan, K.N., D'Aiuto, L., Cirio, M.C., and Chaillet, J.R. (2009). Identification of a region of the DNMT1 methyltransferase that regulates the maintenance of genomic imprints. Proceedings of the National Academy of Sciences *106*, 20806-20811.

Bostick, M., Kim, J.K., Estève, P.-O., Clark, A., Pradhan, S., and Jacobsen, S.E. (2007). UHRF1 Plays a Role in Maintaining DNA Methylation in Mammalian Cells. Science *317*, 1760-1764.

Bourc'his, D., and Bestor, T.H. (2004). Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. Nature *431*, 96-99.

Bourc'his, D., Xu, G.L., Lin, C.S., Bollman, B., and Bestor, T.H. (2001). Dnmt3L and the establishment of maternal genomic imprints. Science *294*, 2536-2539.

Boyes, J., and Bird, A. (1991). DNA methylation inhibits transcription indirectly via a methyl-CpG binding protein. Cell *64*, 1123-1134.

Branciamore, S., Chen, Z.-X., Riggs, A.D., and Rodin, S.N. (2010). CpG island clusters and pro-epigenetic selection for CpGs in protein-coding exons of HOX and other transcription factors. Proceedings of the National Academy of Sciences *107*, 15485-15490.

Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csárdi, G., Harrigan, P., Weier, M., Liechti, A., Aximu-Petri, A., Kircher, M., Albert, F.W., Zeller, U., Khaitovich, P., Grützner, F., Bergmann, S., Nielsen, R., Pääbo, S., and Kaessmann, H. (2011). The evolution of gene expression levels in mammalian organs. Nature *478*, 343-348.

Brinkman, A.B., Simmer, F., Ma, K., Kaan, A., Zhu, J., and Stunnenberg, H.G. (2010). Whole-genome DNA methylation profiling using MethylCap-seq. Methods *52*, 232-236.

Bruce, A.W., López-Contreras, A.J., Flicek, P., Down, T.A., Dhami, P., Dillon, S.C., Koch, C.M., Langford, C.F., Dunham, I., Andrews, R.M., and Vetrie, D. (2009). Functional diversity for REST (NRSF) is defined by in vivo binding affinity hierarchies at the DNA sequence level. Genome Research *19*, 994-1005.

Brzeski, J., and Jerzmanowski, A. (2003). Deficient in DNA methylation 1 (DDM1) defines a novel family of chromatin-remodeling factors. J Biol Chem *278*, 823-828.

Bussmann, J., and Schulte-Merker, S. (2011). Rapid BAC selection for tol2-mediated transgenesis in zebrafish. Development *4332*, 4327-4332.

Byrum, S.D., Raman, A., Taverna, S.D., and Tackett, A.J. (2012). ChAP-MS: A Method for Identification of Proteins and Histone Posttranslational Modifications at a Single Genomic Locus. Cell Reports *2*, 198-205.

Cairns, B.R. (2009). The logic of chromatin architecture and remodelling at promoters. Nature *461*, 193-198.

Camporeale, G., Shubert, E.E., Sarath, G., Cerny, R., and Zempleni, J. (2004). K8 and K12 are biotinylated in human histone H4. European journal of biochemistry / FEBS *271*, 2257-2263.

Campos, C., Valente, L.M.P., and Fernandes, J.M.O. (2012). Molecular evolution of zebrafish dnmt3 genes and thermal plasticity of their expression during embryonic development. Gene *500*, 93-100.

Cao, R., Wang, L., Wang, H., Xia, L., Erdjument-Bromage, H., Tempst, P., Jones, R.S., and Zhang, Y. (2002). Role of Histone H3 Lysine 27 Methylation in Polycomb-Group Silencing. Science *298*, 1039-1043.

Cardoso, C., Lutz, Y., Mignon, C., Compe, E., Depetris, D., Mattei, M.G., Fontes, M., and Colleaux, L. (2000). ATR-X mutations cause impaired nuclear location and altered DNA binding properties of the XNP/ATR-X protein. Journal of medical genetics *37*, 746-751.

Carlone, D.L., Lee, J.H., Young, S.R., Dobrota, E., Butler, J.S., Ruiz, J., and Skalnik, D.G. (2005). Reduced genomic cytosine methylation and defective cellular differentiation in embryonic stem cells lacking CpG binding protein. Mol Cell Biol *25*, 4881-4891.

Carlone, D.L., and Skalnik, D.G. (2001). CpG binding protein is crucial for early embryonic development. Mol Cell Biol *21*, 7601-7606.

Carrozza, M.J., Li, B., Florens, L., Suganuma, T., Swanson, S.K., Lee, K.K., Shia, W.J., Anderson, S., Yates, J., Washburn, M.P., and Workman, J.L. (2005). Histone H3 methylation by Set2 directs deacetylation of coding regions by Rpd3S to suppress spurious intragenic transcription. Cell *123*, 581-592.

Carvalho, L., and Heisenberg, C.-P. (2010). The yolk syncytial layer in early zebrafish development. Trends in Cell Biology *20*, 586-592.

Catchen, J.M., Conery, J.S., and Postlethwait, J.H. (2009). Automated identification of conserved synteny after whole-genome duplication. Genome Res *19*, 1497-1505.

Cedar, H., and Bergman, Y. (2009). Linking DNA methylation and histone modification: patterns and paradigms. Nat Rev Genet *10*, 295-304.

Challen, G.A., Sun, D., Jeong, M., Luo, M., Jelinek, J., Berg, J.S., Bock, C., Vasanthakumar, A., Gu, H., Xi, Y., Liang, S., Lu, Y., Darlington, G.J., Meissner, A., Issa, J.-P.J., Godley, L.A., Li, W., and Goodell, M.A. (2012). Dnmt3a is essential for hematopoietic stem cell differentiation. Nature genetics *44*, 23-31.

Chang, P.Y., Hom, R.A., Musselman, C.A., Zhu, L., Kuo, A., Gozani, O., Kutateladze, T.G., and Cleary, M.L. (2010). Binding of the MLL PHD3 finger to histone H3K4me3 is required for MLL-dependent gene transcription. J Mol Biol *400*, 137-144.

Chantalat, S., Depaux, A., Hery, P., Barral, S., Thuret, J.Y., Dimitrov, S., and Gerard, M. (2011). Histone H3 trimethylation at lysine 36 is associated with constitutive and facultative heterochromatin. Genome Res *21*, 1426-1437.

Charalambous, M., da Rocha, S.T., and Ferguson-Smith, A.C. (2007). Genomic imprinting, growth control and the allocation of nutritional resources: consequences for postnatal life. Current opinion in endocrinology, diabetes, and obesity *14*, 3-12.

Chedin, F., Lieber, M.R., and Hsieh, C.L. (2002). The DNA methyltransferase-like protein DNMT3L stimulates de novo methylation by Dnmt3a. Proc Natl Acad Sci U S A *99*, 16916-16921.

Chen, H., and Boutros, P.C. (2011). VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. BMC Bioinformatics *12*, 35.

Chen, L., MacMillan, A.M., Chang, W., Ezaz-Nikpay, K., Lane, W.S., and Verdine, G.L. (1991). Direct identification of the active-site nucleophile in a DNA (cytosine-5)-methyltransferase. Biochemistry *30*, 11018-11025.

Chen, L., MacMillan, A.M., and Verdine, G.L. (1993). Mutational separation of DNA binding from catalysis in a DNA cytosine methyltransferase. Journal of the American Chemical Society *115*, 5318-5319.

Chen, R.Z., Akbarian, S., Tudor, M., and Jaenisch, R. (2001). Deficiency of methyl-CpG binding protein-2 in CNS neurons results in a Rett-like phenotype in mice. Nature genetics *27*, 327-331.

Chen, S., and Kimelman, D. (2000). The role of the yolk syncytial layer in germ layer patterning in zebrafish. Development *127*, 4681-4689.

Chen, T., and Li, E. (2006). Establishment and maintenance of DNA methylation patterns in mammals. Current topics in microbiology and immunology *301*, 179-201.

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

Chen, T., Ueda, Y., Dodge, J.E., Wang, Z., and Li, E. (2003). Establishment and maintenance of genomic methylation patterns in mouse embryonic stem cells by Dnmt3a and Dnmt3b. Mol Cell Biol *23*, 5594-5605.

Chen, Z.X., Mann, J.R., Hsieh, C.L., Riggs, A.D., and Chedin, F. (2005). Physical and functional interactions between the human DNMT3L protein and members of the de novo methyltransferase family. Journal of cellular biochemistry *95*, 902-917.

Cheng, J.C., Miller, A.L., and Webb, S.E. (2004). Organization and function of microfilaments during late epiboly in zebrafish embryos. Developmental dynamics : an official publication of the American Association of Anatomists *231*, 313-323.

Cheng, X. (1995). DNA modification by methyltransferases. Current opinion in structural biology *5*, 4-10.

Chodavarapu, R.K., Feng, S., Bernatavichute, Y.V., Chen, P.-Y., Stroud, H., Yu, Y., Hetzel, J.A., Kuo, F., Kim, J., Cokus, S.J., Casero, D., Bernal, M., Huijser, P., Clark, A.T., Kramer, U., Merchant, S.S., Zhang, X., Jacobsen, S.E., and Pellegrini, M. (2010). Relationship between nucleosome positioning and DNA methylation. Nature *466*, 388-392.

Chow, J., and Heard, E. (2009). X inactivation and the complexities of silencing a sex chromosome. Current Opinion in Cell Biology *21*, 359-366.

Chuang, L.-Y., Yang, C.-H., Lin, M.-C., and Yang, C.-H. (2012). CpGPAP: CpG island predictor analysis platform. BMC Genetics *13*, 13.

Chuang, L.S., Ian, H.I., Koh, T.W., Ng, H.H., Xu, G., and Li, B.F. (1997). Human DNA-(cytosine-5) methyltransferase-PCNA complex as a target for p21WAF1. Science *277*, 1996-2000.

Cierpicki, T., Risner, L.E., Grembecka, J., Lukasik, S.M., Popovic, R., Omonkowska, M., Shultis, D.D., Zeleznik-Le, N.J., and Bushweller, J.H. (2010). Structure of the MLL CXXC domain-DNA complex and its functional role in MLL-AF9 leukemia. Nat Struct Mol Biol *17*, 62-68.

Cirio, M.C., Ratnam, S., Ding, F., Reinhart, B., Navara, C., and Chaillet, J.R. (2008). Preimplantation expression of the somatic form of Dnmt1 suggests a role in the inheritance of genomic imprints. BMC developmental biology *8*, 9.

Clouaire, T., de Las Heras, J.I., Merusi, C., and Stancheva, I. (2010). Recruitment of MBD1 to target genes requires sequence-specific interaction of the MBD domain with methylated DNA. Nucleic Acids Res *38*, 4620-4634.

Clouaire, T., and Stancheva, I. (2008). Methyl-CpG binding proteins: specialized transcriptional repressors or structural components of chromatin? Cell Mol Life Sci *65*, 1509-1522.

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

Clynes, D., Higgs, D.R., and Gibbons, R.J. (2013). The chromatin remodeller ATRX: a repeat offender in human disease. Trends Biochem Sci *38*, 461-466.

Cokus, S.J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C.D., Pradhan, S., Nelson, S.F., Pellegrini, M., and Jacobsen, S.E. (2008). Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature *452*, 215-219.

Comb, M., and Goodman, H.M. (1990). CpG methylation inhibits proenkephalin gene expression and binding of the transcription factor AP-2. Nucleic acids research *18*, 3975-3982.

Cooper, D., Mort, M., Stenson, P., Ball, E., and Chuzhanova, N. (2010). Methylation-mediated deamination of 5-methylcytosine appears to give rise to mutations causing human inherited disease in CpNpG trinucleotides, as well as in CpG dinucleotides. Human Genomics *4*, 406 - 410.

Cooper, D.N., Taggart, M.H., and Bird, A.P. (1983). Unmethlated domains in vertebrate DNA. Nucleic acids research *11*, 647-658.

Cortázar, D., Kunz, C., Saito, Y., Steinacher, R., and Schär, P. (2007). The enigmatic thymine DNA glycosylase. DNA Repair *6*, 489-504.

Costa, Y., Ding, J., Theunissen, T.W., Faiola, F., Hore, T.A., Shliaha, P.V., Fidalgo, M., Saunders, A., Lawrence, M., Dietmann, S., Das, S., Levasseur, D.N., Li, Z., Xu, M., Reik, W., Silva, J.C.R., and Wang, J. (2013). NANOG-dependent function of TET1 and TET2 in establishment of pluripotency. Nature *495*, 370-374.

Coulondre, C., Miller, J.H., Farabaugh, P.J., and Gilbert, W. (1978). Molecular basis of base substitution hotspots in Escherichia coli. Nature *274*, 775-780.

Cross, S., Kovarik, P., Schmidtke, J., and Bird, A. (1991). Non-methylated islands in fish genomes are GC-poor. Nucleic Acids Res *19*, 1469-1474.

Cross, S.H., Charlton, J.A., Nan, X., and Bird, A.P. (1994). Purification of CpG islands using a methylated DNA binding column. Nature genetics *6*, 236-244.

Cross, S.H., Meehan, R.R., Nan, X., and Bird, A. (1997). A component of the transcriptional repressor MeCP1 shares a motif with DNA methyltransferase and HRX proteins. Nature genetics *16*, 256-259.

Cuadrado, M., Sacristan, M., and Antequera, F. (2001). Species-specific organization of CpG island promoters at mammalian homologous genes. EMBO reports *2*, 586-592.

Cuozzo, C., Porcellini, A., Angrisano, T., Morano, A., Lee, B., Pardo, A.D., Messina, S., Iuliano, R., Fusco, A., Santillo, M.R., Muller, M.T., Chiariotti, L., Gottesman, M.E., and Avvedimento, E.V. (2007). DNA Damage, Homology-Directed Repair, and DNA Methylation. PLoS Genet *3*, e110.

Czermin, B., Melfi, R., McCabe, D., Seitz, V., Imhof, A., and Pirrotta, V. (2002). Drosophila Enhancer of Zeste/ESC Complexes Have a Histone H3 Methyltransferase Activity that Marks Chromosomal Polycomb Sites. Cell *111*, 185-196.

Dahl, C., and Guldberg, P. (2003). DNA methylation analysis techniques. Biogerontology *4*, 233-250.

Davidson, A.J., Ernst, P., Wang, Y., Dekens, M.P., Kingsley, P.D., Palis, J., Korsmeyer, S.J., Daley, G.Q., and Zon, L.I. (2003). cdx4 mutants fail to specify blood progenitors and can be rescued by multiple hox genes. Nature *425*, 300-306.

de-Leon, S.B.-T., and Davidson, E.H. (2007). Gene Regulation: Gene Control Network in Development. Annual review of biophysics and biomolecular structure *36*, 191-212.

de Wit, E., and de Laat, W. (2012). A decade of 3C technologies: insights into nuclear organization. Genes & development *26*, 11-24.

Dean, W., Santos, F., Stojkovic, M., Zakhartchenko, V., Walter, J., Wolf, E., and Reik, W. (2001). Conservation of methylation reprogramming in mammalian development: aberrant reprogramming in cloned embryos. Proc Natl Acad Sci U S A *98*, 13734-13738.

Deaton, A.M., and Bird, A. (2011). CpG islands and the regulation of transcription. Genes Dev *25*, 1010-1022.

Deaton, A.M., Webb, S., Kerr, A.R., Illingworth, R.S., Guy, J., Andrews, R., and Bird, A. (2011). Cell type-specific DNA methylation at intragenic CpG islands in the immune system. Genome Res *21*, 1074-1086.

DeChiara, T.M., Robertson, E.J., and Efstratiadis, A. (1991). Parental imprinting of the mouse insulin-like growth factor II gene. Cell *64*, 849-859.

Dee, C.T., Gibson, A., Rengifo, A., Sun, S.K., Patient, R.K., and Scotting, P.J. (2007). A change in response to Bmp signalling precedes ectodermal fate choice. The International journal of developmental biology *51*, 79-84.

Dejardin, J., and Kingston, R.E. (2009). Purification of Proteins Associated with Specific Genomic Loci. Cell *136*, 175-186.

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

Dhayalan, A., Rajavelu, A., Rathert, P., Tamas, R., Jurkowska, R.Z., Ragozin, S., and Jeltsch, A. (2010). The Dnmt3a PWWP domain reads histone 3 lysine 36 trimethylation and guides DNA methylation. J Biol Chem *285*, 26114-26120.

Dhayalan, A., Tamas, R., Bock, I., Tattermusch, A., Dimitrova, E., Kudithipudi, S., Ragozin, S., and Jeltsch, A. (2011). The ATRX-ADD domain binds to H3 tail peptides and reads the combined methylation state of K4 and K9. Human Molecular Genetics *20*, 2195-2203.

Di Lorenzo, A., and Bedford, M.T. (2011). Histone arginine methylation. FEBS Letters *585*, 2024-2031.

Di Ruscio, A., Ebralidze, A.K., Benoukraf, T., Amabile, G., Goff, L.A., Terragni, J., Figueroa, M.E., De Figueiredo Pontes, L.L., Alberich-Jorda, M., Zhang, P., Wu, M., D'Alo, F., Melnick, A., Leone, G., Ebralidze, K.K., Pradhan, S., Rinn, J.L., and Tenen, D.G. (2013). DNMT1-interacting RNAs block gene-specific DNA methylation. Nature *503*, 371-376.

Doerfler, W. (1983). DNA methylation and gene activity. Annual review of biochemistry *52*, 93-124.

Doherty, A.S., Bartolomei, M.S., and Schultz, R.M. (2002). Regulation of Stage-Specific Nuclear Translocation of Dnmt1o during Preimplantation Mouse Development. Developmental Biology *242*, 255-266.

Doi, A., Park, I.H., Wen, B., Murakami, P., Aryee, M.J., Irizarry, R., Herb, B., Ladd-Acosta, C., Rho, J., Loewer, S., Miller, J., Schlaeger, T., Daley, G.Q., and Feinberg, A.P. (2009). Differential methylation of tissue- and cancer-

specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts. Nature genetics *41*, 1350-1353.

Dong, A., Yoder, J.A., Zhang, X., Zhou, L., Bestor, T.H., and Cheng, X. (2001). Structure of human DNMT2, an enigmatic DNA methyltransferase homolog that displays denaturant-resistant binding to DNA. Nucleic Acids Res *29*, 439-448.

Dong, R., Yao, R., Du, J., Wang, S., and Fan, Z. (2013). Depletion of histone demethylase KDM2A enhanced the adipogenic and chondrogenic differentiation potentials of stem cells from apical papilla. Experimental cell research.

Down, T.A., Rakyan, V.K., Turner, D.J., Flicek, P., Li, H., Kulesha, E., Graf, S., Johnson, N., Herrero, J., Tomazou, E.M., Thorne, N.P., Backdahl, L., Herberth, M., Howe, K.L., Jackson, D.K., Miretti, M.M., Marioni, J.C., Birney, E., Hubbard, T.J.P., Durbin, R., Tavare, S., and Beck, S. (2008). A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. Nat Biotechnol *26*, 779-785.

Drane, P., Ouararhni, K., Depaux, A., Shuaib, M., and Hamiche, A. (2010). The death-associated protein DAXX is a novel histone chaperone involved in the replication-independent deposition of H3.3. Genes Dev *24*, 1253-1265.

Du, J., Ma, Y., Ma, P., Wang, S., and Fan, Z. (2013). Demethylation of epiregulin gene by histone demethylase FBXL11 and BCL6 corepressor inhibits osteo/dentinogenic differentiation. Stem cells *31*, 126-136.

Ebert, A., Lein, S., Schotta, G., and Reuter, G. (2006). Histone modification and the control of heterochromatic gene silencing in Drosophila. Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology *14*, 377-392.

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

Edgar, R., Domrachev, M., and Lash, A.E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res *30*, 207-210.

Edmunds, J.W., Mahadevan, L.C., and Clayton, A.L. (2008). Dynamic histone H3 methylation during gene induction: HYPB/Setd2 mediates all H3K36 trimethylation. EMBO J *27*, 406-420.

Ehrlich, M. (2003). The ICF syndrome, a DNA methyltransferase 3B deficiency and immunodeficiency disease. Clinical Immunology *109*, 17-28.

Ehrlich, M., Gama-Sosa, M.A., Huang, L.-H., Midgett, R.M., Kuo, K.C., McCune, R.A., and Gehrke, C. (1982). Amount and distribution of 5-methylcytosine in human DNA from different types of tissues or cells. Nucleic acids research *10*, 2709-2721.

Ehrlich, M., Sanchez, C., Shao, C., Nishiyama, R., Kehrl, J., Kuick, R., Kubota, T., and Hanash, S.M. (2008). ICF, an immunodeficiency syndrome: DNA methyltransferase 3B involvement, chromosome anomalies, and gene dysregulation. Autoimmunity *41*, 253-271.

Elango, N., and Yi, S.V. (2008). DNA Methylation and Structural and Functional Bimodality of Vertebrate Promoters. Molecular Biology and Evolution *25*, 1602-1608.

Elsasser, S.J., Huang, H., Lewis, P.W., Chin, J.W., Allis, C.D., and Patel, D.J. (2012). DAXX envelops a histone H3.3-H4 dimer for H3.3-specific recognition. Nature *491*, 560-565.

Epsztejn-Litman, S., Feldman, N., Abu-Remaileh, M., Shufaro, Y., Gerson, A., Ueda, J., Deplus, R., Fuks, F., Shinkai, Y., Cedar, H., and Bergman, Y. (2008). De novo DNA methylation promoted by G9a prevents reprogramming of embryonically silenced genes. Nat Struct Mol Biol *15*, 1176-1183.

Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., Ku, M., Durham, T., Kellis, M., and Bernstein, B.E. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. Nature *473*, 43-49.

Ernst, P., Fisher, J.K., Avery, W., Wade, S., Foy, D., and Korsmeyer, S.J. (2004). Definitive hematopoiesis requires the mixed-lineage leukemia gene. Dev Cell *6*, 437-443.

Ernst, P., and Vakoc, C.R. (2012). WRAD: enabler of the SET1-family of H3K4 methyltransferases. Briefings in functional genomics *11*, 217-226.

Erter, C.E., Wilm, T.P., Basler, N., Wright, C.V.E., and Solnica-Krezel, L. (2001). Wnt8 is required in lateral mesendodermal precursors for neural posteriorization in vivo. Development *128*, 3571-3583.

Esteller, M. (2007). Epigenetic gene silencing in cancer: the DNA hypermethylome. Human Molecular Genetics *16*, R50-R59.

Estève, P.-O., Chin, H.G., Smallwood, A., Feehery, G.R., Gangisetty, O., Karpf, A.R., Carey, M.F., and Pradhan, S. (2006). Direct interaction between DNMT1 and G9a coordinates DNA and histone methylation during replication. Genes & development *20*, 3089-3103.

Eustermann, S., Yang, J.-C., Law, M.J., Amos, R., Chapman, L.M., Jelinska, C., Garrick, D., Clynes, D., Gibbons, R.J., Rhodes, D., Higgs, D.R., and Neuhaus, D. (2011). Combinatorial readout of histone H3 modifications specifies localization of ATRX to heterochromatin. Nat Struct Mol Biol *18*, 777-782.

Farcas, A.M., Blackledge, N.P., Sudbery, I., Long, H.K., McGouran, J.F., Rose, N.R., Lee, S., Sims, D., Cerase, A., Sheahan, T.W., Koseki, H., Brockdorff, N., Ponting, C.P., Kessler, B.M., and Klose, R.J. (2012). KDM2B links the Polycomb Repressive Complex 1 (PRC1) to recognition of CpG islands. eLife *1*, e00205.

Farthing, C.R., Ficz, G., Ng, R.K., Chan, C.F., Andrews, S., Dean, W., Hemberger, M., and Reik, W. (2008). Global mapping of DNA methylation in mouse promoters reveals epigenetic reprogramming of pluripotency genes. PLoS Genet *4*, e1000116.

Fatemi, M., Hermann, A., Gowher, H., and Jeltsch, A. (2002). Dnmt3a and Dnmt1 functionally cooperate during de novo methylation of DNA. European journal of biochemistry / FEBS *269*, 4981-4984.

Fatemi, M., and Wade, P.A. (2006). MBD family proteins: reading the epigenetic code. Journal of cell science *119*, 3033-3037.

Feldman, B., Gates, M.A., Egan, E.S., Dougan, S.T., Rennebeck, G., Sirotkin, H.I., Schier, A.F., and Talbot, W.S. (1998). Zebrafish organizer development and germ-layer formation require nodal-related signals. Nature *395*, 181-185.

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

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

Felle, M., Hoffmeister, H., Rothammer, J., Fuchs, A., Exler, J.H., and Langst, G. (2011). Nucleosomes protect DNA from DNA methylation in vivo and in vitro. Nucleic Acids Res *39*, 6956-6969.

Feng, S., Cokus, S.J., Zhang, X., Chen, P.Y., Bostick, M., Goll, M.G., Hetzel, J., Jain, J., Strauss, S.H., Halpern, M.E., Ukomadu, C., Sadler, K.C., Pradhan, S., Pellegrini, M., and Jacobsen, S.E. (2010). Conservation and divergence of methylation patterning in plants and animals. Proc Natl Acad Sci U S A *107*, 8689-8694.

Fenouil, R., Cauchy, P., Koch, F., Descostes, N., Cabeza, J.Z., Innocenti, C., Ferrier, P., Spicuglia, S., Gut, M., Gut, I., and Andrau, J.-C. (2012). CpG islands and GC content dictate nucleosome depletion in a transcription-independent manner at mammalian promoters. Genome Research *22*, 2399-2408.

Ferguson-Smith, A.C. (2011). Genomic imprinting: the emergence of an epigenetic paradigm. Nat Rev Genet *12*, 565-575.

Ferguson-Smith, A.C., Cattanach, B.M., Barton, S.C., Beechey, C.V., and Surani, M.A. (1991). Embryological and molecular investigations of parental imprinting on mouse chromosome 7. Nature *351*, 667-670.

Fernandez, P.C., Frank, S.R., Wang, L., Schroeder, M., Liu, S., Greene, J., Cocito, A., and Amati, B. (2003). Genomic targets of the human c-Myc protein. Genes & development *17*, 1115-1129.

Ficz, G., Branco, M.R., Seisenberger, S., Santos, F., Krueger, F., Hore, T.A., Marques, C.J., Andrews, S., and Reik, W. (2011). Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. Nature *473*, 398-402.

Filion, G.J., Zhenilo, S., Salozhin, S., Yamada, D., Prokhortchouk, E., and Defossez, P.A. (2006). A family of human zinc finger proteins that bind methylated DNA and repress transcription. Mol Cell Biol *26*, 169-181.

Fouse, S.D., Nagarajan, R.O., and Costello, J.F. (2010). Genome-scale DNA methylation analysis. Epigenomics *2*, 105-117.

Frauer, C., Rottach, A., Meilinger, D., Bultmann, S., Fellinger, K., Hasenoder, S., Wang, M., Qin, W., Soding, J., Spada, F., and Leonhardt, H. (2011). Different binding properties and function of CXXC zinc finger domains in Dnmt1 and Tet1. PLoS One *6*, e16627.

Friedman, S., and Ansari, N. (1992). Binding of the EcoRII methyltransferase to 5-fluorocytosine-containing DNA. Isolation of a bound peptide. Nucleic Acids Res *20*, 3241-3248.

Frommer, M., McDonald, L.E., Millar, D.S., Collis, C.M., Watt, F., Grigg, G.W., Molloy, P.L., and Paul, C.L. (1992). A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. Proceedings of the National Academy of Sciences *89*, 1827-1831.

Fujita, N., Watanabe, S., Ichimura, T., Tsuruzoe, S., Shinkai, Y., Tachibana, M., Chiba, T., and Nakao, M. (2003). Methyl-CpG binding domain 1 (MBD1) interacts with the Suv39h1-HP1 heterochromatic complex for DNA methylation-based transcriptional repression. J Biol Chem *278*, 24132-24138.

Fuks, F., Burgers, W.A., Godin, N., Kasai, M., and Kouzarides, T. (2001). Dnmt3a binds deacetylases and is recruited by a sequence-specific repressor to silence transcription. EMBO J *20*, 2536-2544.

Fuks, F., Hurd, P.J., Deplus, R., and Kouzarides, T. (2003). The DNA methyltransferases associate with HP1 and the SUV39H1 histone methyltransferase. Nucleic Acids Res *31*, 2305-2312.

Fukuda, T., Tokunaga, A., Sakamoto, R., and Yoshida, N. (2011). Fbxl10/Kdm2b deficiency accelerates neural progenitor cell death and leads to exencephaly. Molecular and cellular neurosciences *46*, 614-624.

Futscher, B.W., Oshiro, M.M., Wozniak, R.J., Holtan, N., Hanigan, C.L., Duan, H., and Domann, F.E. (2002). Role for DNA methylation in the control of cell type specific maspin expression. Nature genetics *31*, 175-179.

Gal-Yam, E.N., Egger, G., Iniguez, L., Holster, H., Einarsson, S., Zhang, X., Lin, J.C., Liang, G., Jones, P.A., and Tanay, A. (2008). Frequent switching of Polycomb repressive marks and DNA hypermethylation in the PC3 prostate cancer cell line. Proc Natl Acad Sci U S A *105*, 12979-12984.

Gao, Z., Zhang, J., Bonasio, R., Strino, F., Sawai, A., Parisi, F., Kluger, Y., and Reinberg, D. (2012). PCGF homologs, CBX proteins, and RYBP define functionally distinct PRC1 family complexes. Mol Cell *45*, 344-356.

Garcia, B.A., Thomas, C.E., Kelleher, N.L., and Mizzen, C.A. (2008). Tissue-specific expression and post-translational modification of histone H3 variants. J Proteome Res *7*, 4225-4236.

Gardiner-Garden, M., and Frommer, M. (1987). CpG Islands in vertebrate genomes. Journal of Molecular Biology *196*, 261-282.

Gaszner, M., and Felsenfeld, G. (2006). Insulators: exploiting transcriptional and epigenetic mechanisms. Nat Rev Genet *7*, 703-713.

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

Ge, Y.Z., Pu, M.T., Gowher, H., Wu, H.P., Ding, J.P., Jeltsch, A., and Xu, G.L. (2004). Chromatin targeting of de novo DNA methyltransferases by the PWWP domain. Journal of Biological Chemistry *279*, 25447-25454.

Gearhart, M.D., Corcoran, C.M., Wamstad, J.A., and Bardwell, V.J. (2006). Polycomb group and SCF ubiquitin ligases are found in a novel BCOR complex that is recruited to BCL6 targets. Mol Cell Biol *26*, 6880-6889.

Geiman, T.M., and Muegge, K. (2000). Lsh, an SNF2/helicase family member, is required for proliferation of mature T lymphocytes. Proc Natl Acad Sci U S A *97*, 4772-4777.

Geiman, T.M., Tessarollo, L., Anver, M.R., Kopp, J.B., Ward, J.M., and Muegge, K. (2001). Lsh, a SNF2 family member, is required for normal murine development. Biochimica et biophysica acta *1526*, 211-220.

Gibbons, R.J., Bachoo, S., Picketts, D.J., Aftimos, S., Asenbauer, B., Bergoffen, J., Berry, S.A., Dahl, N., Fryer, A., Keppler, K., Kurosawa, K., Levin, M.L., Masuno, M., Neri, G., Pierpont, M.E., Slaney, S.F., and Higgs, D.R. (1997). Mutations in transcriptional regulator ATRX establish the functional significance of a PHD-like domain. Nature genetics *17*, 146-148.

Gibbons, R.J., and Higgs, D.R. (2010). ATRX: Taming tandem repeats. Cell Cycle *9*, 4605-4606.

Gibbons, R.J., McDowell, T.L., Raman, S., O'Rourke, D.M., Garrick, D., Ayyub, H., and Higgs, D.R. (2000). Mutations in ATRX, encoding a SWI/SNF-like protein, cause diverse changes in the pattern of DNA methylation. Nature genetics *24*, 368-371.

Gibbons, R.J., Pellagatti, A., Garrick, D., Wood, W.G., Malik, N., Ayyub, H., Langford, C., Boultwood, J., Wainscoat, J.S., and Higgs, D.R. (2003). Identification of acquired somatic mutations in the gene encoding chromatin-remodeling factor ATRX in the [alpha]-thalassemia myelodysplasia syndrome (ATMDS). Nature genetics *34*, 446-449.

Ginno, P.A., Lott, P.L., Christensen, H.C., Korf, I., and Chedin, F. (2012). R-loop formation is a distinctive characteristic of unmethylated human CpG island promoters. Mol Cell *45*, 814-825.

Giraldez, A.J., Mishima, Y., Rihel, J., Grocock, R.J., Van Dongen, S., Inoue, K., Enright, A.J., and Schier, A.F. (2006). Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs. Science *312*, 75-79.

Glover, A.B., and Leyland-Jones, B. (1987). Biochemistry of azacitidine: a review. Cancer treatment reports *71*, 959-964.

Goldsworthy, M., Absalom, N.L., Schroter, D., Matthews, H.C., Bogani, D., Moir, L., Long, A., Church, C., Hugill, A., Anstee, Q.M., Goldin, R., Thursz, M., Hollfelder, F., and Cox, R.D. (2013). Mutations in Mll2, an H3K4 methyltransferase, result in insulin resistance and impaired glucose tolerance in mice. PLoS One *8*, e61870.

Goll, M.G., and Bestor, T.H. (2005). Eukaryotic cytosine methyltransferases. Annual review of biochemistry *74*, 481-514.

Goll, M.G., Kirpekar, F., Maggert, K.A., Yoder, J.A., Hsieh, C.L., Zhang, X., Golic, K.G., Jacobsen, S.E., and Bestor, T.H. (2006). Methylation of tRNAAsp by the DNA methyltransferase homolog Dnmt2. Science *311*, 395-398.

Golling, G., Amsterdam, A., Sun, Z., Antonelli, M., Maldonado, E., Chen, W., Burgess, S., Haldi, M., Artzt, K., Farrington, S., Lin, S.Y., Nissen, R.M., and Hopkins, N. (2002). Insertional mutagenesis in zebrafish rapidly identifies genes essential for early vertebrate development. Nature genetics *31*, 135-140.

Gowher, H., and Jeltsch, A. (2001). Enzymatic properties of recombinant Dnmt3a DNA methyltransferase from mouse: the enzyme modifies DNA in a non-processive manner and also methylates non-CpG [correction of non-CpA] sites. J Mol Biol *309*, 1201-1208.

Gowher, H., and Jeltsch, A. (2002). Molecular enzymology of the catalytic domains of the Dnmt3a and Dnmt3b DNA methyltransferases. J Biol Chem *277*, 20409-20414.

Gowher, H., Liebert, K., Hermann, A., Xu, G., and Jeltsch, A. (2005a). Mechanism of stimulation of catalytic activity of Dnmt3A and Dnmt3B DNA-(cytosine-C5)-methyltransferases by Dnmt3L. J Biol Chem *280*, 13341-13348.

Gowher, H., Stockdale, C.J., Goyal, R., Ferreira, H., Owen-Hughes, T., and Jeltsch, A. (2005b). De novo methylation of nucleosomal DNA by the mammalian Dnmt1 and Dnmt3A DNA methyltransferases. Biochemistry *44*, 9899-9904.

Grandjean, V., Yaman, R., Cuzin, F., and Rassoulzadegan, M. (2007). Inheritance of an epigenetic mark: the CpG DNA methyltransferase 1 is required for de novo establishment of a complex pattern of non-CpG methylation. PLoS One *2*, e1136.

Grau, D.J., Chapman, B.A., Garlick, J.D., Borowsky, M., Francis, N.J., and Kingston, R.E. (2011). Compaction of chromatin by diverse Polycomb group proteins requires localized regions of high charge. Genes Dev *25*, 2210-2221.

Green, N.M. (1963). Avidin. 1. The Use of (14-C)Biotin for Kinetic Studies and for Assay. The Biochemical journal *89*, 585-591.

Grewal, S.I., and Elgin, S.C. (2007). Transcription and RNA interference in the formation of heterochromatin. Nature *447*, 399-406.

Griesenbeck, J., Boeger, H., Strattan, J.S., and Kornberg, R.D. (2003). Affinity purification of specific chromatin segments from chromosomal loci in yeast. Mol Cell Biol *23*, 9275-9282.

Gritsman, K., Zhang, J., Cheng, S., Heckscher, E., Talbot, W.S., and Schier, A.F. (1999). The EGF-CFC protein one-eyed pinhead is essential for nodal signaling. Cell *97*, 121-132.

Gu, T.P., Guo, F., Yang, H., Wu, H.P., Xu, G.F., Liu, W., Xie, Z.G., Shi, L., He, X., Jin, S.G., Iqbal, K., Shi, Y.G., Deng, Z., Szabo, P.E., Pfeifer, G.P., Li, J., and Xu, G.L. (2011). The role of Tet3 DNA dioxygenase in epigenetic reprogramming by oocytes. Nature *477*, 606-610.

Guenther, M.G., Jenner, R.G., Chevalier, B., Nakamura, T., Croce, C.M., Canaani, E., and Young, R.A. (2005). Global and Hox-specific roles for the MLL1 methyltransferase. Proc Natl Acad Sci U S A *102*, 8603-8608.

Guenther, M.G., Levine, S.S., Boyer, L.A., Jaenisch, R., and Young, R.A. (2007). A chromatin landmark and transcription initiation at most promoters in human cells. Cell *130*, 77-88.

Guttman, M., Amit, I., Garber, M., French, C., Lin, M.F., Feldser, D., Huarte, M., Zuk, O., Carey, B.W., Cassady, J.P., Cabili, M.N., Jaenisch, R., Mikkelsen, T.S., Jacks, T., Hacohen, N., Bernstein, B.E., Kellis, M., Regev, A., Rinn, J.L., and Lander, E.S. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature *458*, 223-227.

Guttman, M., Garber, M., Levin, J.Z., Donaghey, J., Robinson, J., Adiconis, X., Fan, L., Koziol, M.J., Gnirke, A., Nusbaum, C., Rinn, J.L., Lander, E.S., and Regev, A. (2010). Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat Biotechnol *28*, 503-510.

Guy, J., Hendrich, B., Holmes, M., Martin, J.E., and Bird, A. (2001). A mouse Mecp2-null mutation causes neurological symptoms that mimic Rett syndrome. Nature genetics *27*, 322-326.

Guzman, L.M., Belin, D., Carson, M.J., and Beckwith, J. (1995). Tight regulation, modulation, and high-level expression by vectors containing the arabinose PBAD promoter. Journal of Bacteriology *177*, 4121-4130.

Hackenberg, M., Previti, C., Luque-Escamilla, P., Carpena, P., Martinez-Aroza, J., and Oliver, J. (2006). CpGcluster: a distance-based algorithm for CpG-island detection. BMC Bioinformatics *7*, 446.

Hagarman, J.A., Motley, M.P., Kristjansdottir, K., and Soloway, P.D. (2013). Coordinate Regulation of DNA Methylation and H3K27me3 in Mouse Embryonic Stem Cells. PLoS ONE *8*, e53880.

Hajkova, P., Jeffries, S.J., Lee, C., Miller, N., Jackson, S.P., and Surani, M.A. (2010). Genome-Wide Reprogramming in the Mouse Germ Line Entails the Base Excision Repair Pathway. Science *329*, 78-82.

Hammoud, S.S., Nix, D.A., Zhang, H., Purwar, J., Carrell, D.T., and Cairns, B.R. (2009). Distinctive chromatin in human sperm packages genes for embryo development. Nature *460*, 473-478.

Hanck, T., Schmidt, S., and Fritz, H.-J. (1993). Sequence-specific and mechanism-based crosslinking of Dcm DNA cytosine-C5 methyltransferase of E.coli K-12 to synthetic oligonucleotides containing 5-fluoro-2'-deoxycytidine. Nucleic acids research *21*, 303-309.

Harris, R.A., Wang, T., Coarfa, C., Nagarajan, R.P., Hong, C., Downey, S.L., Johnson, B.E., Fouse, S.D., Delaney, A., Zhao, Y., Olshen, A., Ballinger, T., Zhou, X., Forsberg, K.J., Gu, J., Echipare, L., O'Geen, H., Lister, R., Pelizzola, M., Xi, Y., Epstein, C.B., Bernstein, B.E., Hawkins, R.D., Ren, B., Chung, W.Y., Gu, H., Bock, C., Gnirke, A., Zhang, M.Q., Haussler, D., Ecker, J.R., Li, W., Farnham, P.J., Waterland, R.A., Meissner, A., Marra, M.A., Hirst, M., Milosavljevic, A., and Costello, J.F. (2010). Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. Nat Biotechnol *28*, 1097-1105.

Harvey, S.A., Tümpel, S., Dubrulle, J., Schier, A.F., and Smith, J.C. (2010). no tail integrates two modes of mesoderm induction. Development *137*, 1127-1135.

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

Hashimshony, T., Zhang, J., Keshet, I., Bustin, M., and Cedar, H. (2003). The role of DNA methylation in setting up chromatin structure during development. Nature genetics *34*, 187-192.

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

Hawkins, R.D., Hon, G.C., Lee, L.K., Ngo, Q., Lister, R., Pelizzola, M., Edsall, L.E., Kuan, S., Luu, Y., Klugman, S., Antosiewicz-Bourget, J., Ye, Z., Espinoza, C., Agarwahl, S., Shen, L., Ruotti, V., Wang, W., Stewart, R., Thomson, J.A., Ecker, J.R., and Ren, B. (2010). Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. Cell stem cell *6*, 479-491.

He, J., Kallin, E.M., Tsukada, Y., and Zhang, Y. (2008). The H3K36 demethylase Jhdm1b/Kdm2b regulates cell proliferation and senescence through p15(Ink4b). Nat Struct Mol Biol *15*, 1169-1175.

He, J., Nguyen, A.T., and Zhang, Y. (2011a). KDM2b/JHDM1b, an H3K36me2-specific demethylase, is required for initiation and maintenance of acute myeloid leukemia. Blood *117*, 3869-3880.

He, J., Shen, L., Wan, M., Taranova, O., Wu, H., and Zhang, Y. (2013). Kdm2b maintains murine embryonic stem cell status by recruiting PRC1 complex to CpG islands of developmental genes. Nature cell biology *15*, 373-384.

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

Heger, A., and Ponting, C.P. (2008). OPTIC: orthologous and paralogous transcripts in clades. Nucleic Acids Res *36*, D267-270.

Heintzman, N.D., Hon, G.C., Hawkins, R.D., Kheradpour, P., Stark, A., Harp, L.F., Ye, Z., Lee, L.K., Stuart, R.K., Ching, C.W., Ching, K.A., Antosiewicz-Bourget, J.E., Liu, H., Zhang, X.M., Green, R.D., Lobanenkov, V.V., Stewart, R., Thomson, J.A., Crawford, G.E., Kellis, M., and Ren, B. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature *459*, 108-112.

Heintzman, N.D., Stuart, R.K., Hon, G., Fu, Y., Ching, C.W., Hawkins, R.D., Barrera, L.O., Van Calcar, S., Qu, C., Ching, K.A., Wang, W., Weng, Z., Green, R.D., Crawford, G.E., and Ren, B. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nature genetics *39*, 311-318.

Heitz, E. (1928). Das Heterochromatin der Moose. Jahrb Wiss Botanik, 762–818.

Hendrich, B., and Bird, A. (1998). Identification and characterization of a family of mammalian methyl-CpG binding proteins. Mol Cell Biol *18*, 6538-6547.

Hendrich, B., Guy, J., Ramsahoye, B., Wilson, V.A., and Bird, A. (2001). Closely related proteins MBD2 and MBD3 play distinctive but interacting roles in mouse development. Genes Dev *15*, 710-723.

Hendrich, B., Hardeland, U., Ng, H.H., Jiricny, J., and Bird, A. (1999). The thymine glycosylase MBD4 can bind to the product of deamination at methylated CpG sites. Nature *401*, 301-304.

Hendrich, B., and Tweedie, S. (2003). The methyl-CpG binding domain and the evolving role of DNA methylation in animals. Trends Genet *19*, 269-277.

Hermann, A., Schmitt, S., and Jeltsch, A. (2003). The human Dnmt2 has residual DNA-(cytosine-C5) methyltransferase activity. J Biol Chem *278*, 31717-31721.

Hervouet, E., Vallette, F.M., and Cartron, P.-F. (2009). Dnmt3/transcription factor interactions as crucial players in targeted DNA methylation. Epigenetics *4*, 487-499.

Ho, K.L., McNae, I.W., Schmiedeberg, L., Klose, R.J., Bird, A.P., and Walkinshaw, M.D. (2008). MeCP2 binding to DNA depends upon hydration at methyl-CpG. Mol Cell *29*, 525-531.

Hodges, E., Molaro, A., Dos Santos, C.O., Thekkat, P., Song, Q., Uren, P.J., Park, J., Butler, J., Rafii, S., McCombie, W.R., Smith, A.D., and Hannon, G.J. (2011). Directional DNA methylation changes and complex intermediate states accompany lineage specificity in the adult hematopoietic compartment. Mol Cell *44*, 17-28.

Hogan, B., Verkade, H., Lieschke, G., and Heath, J. (2009). Manipulation of Gene Expression During Zebrafish Embryonic Development Using Transient Approaches. In Wnt Signaling, E. Vincan, ed. (Humana Press), pp. 273-300.

Holliday, R., and Pugh, J.E. (1975). DNA Modification Mechanisms and Gene Activity during Development. Science *187*, 226-232.

Hon, G.C., Hawkins, R.D., Caballero, O.L., Lo, C., Lister, R., Pelizzola, M., Valsesia, A., Ye, Z., Kuan, S., Edsall, L.E., Camargo, A.A., Stevenson, B.J., Ecker, J.R., Bafna, V., Strausberg, R.L., Simpson, A.J., and Ren, B. (2012). Global DNA hypomethylation coupled to repressive chromatin domain formation and gene silencing in breast cancer. Genome Research *22*, 246-258.

Hon, G.C., Hawkins, R.D., and Ren, B. (2009). Predictive chromatin signatures in the mammalian genome. Human Molecular Genetics *18*, R195-R201.

Honda, S., Lewis, Z.A., Shimada, K., Fischle, W., Sack, R., and Selker, E.U. (2012). Heterochromatin protein 1 forms distinct complexes to direct histone deacetylation and DNA methylation. Nat Struct Mol Biol *19*, 471-477.

Hopp, T.P., and Woods, K.R. (1981). Prediction of protein antigenic determinants from amino acid sequences. Proc Natl Acad Sci U S A *78*, 3824-3828.

Horsthemke, B., and Wagstaff, J. (2008). Mechanisms of imprinting of the Prader-Willi/Angelman region. American journal of medical genetics Part A *146A*, 2041-2052.

Hoshino, A., and Fujii, H. (2009). Insertional chromatin immunoprecipitation: A method for isolating specific genomic regions. J Biosci Bioeng *108*, 446-449.

Hotchkiss, R.D. (1948). The quantitative separation of purines, pyrimidines, and nucleosides by paper chromatography. J Biol Chem *175*, 315-332.

Howard, G., Eiges, R., Gaudet, F., Jaenisch, R., and Eden, A. (2007). Activation and transposition of endogenous retroviral elements in hypomethylation induced tumors in mice. Oncogene *27*, 404-408.

Howe, K., Clark, M.D., Torroja, C.F., Torrance, J., Berthelot, C., Muffato, M., Collins, J.E., Humphray, S., McLaren, K., Matthews, L., McLaren, S., Sealy, I., Caccamo, M., Churcher, C., Scott, C., Barrett, J.C., Koch, R., Rauch, G.-J., White, S., Chow, W., Kilian, B., Quintais, L.T., Guerra-Assuncao, J.A., Zhou, Y., Gu, Y., Yen, J., Vogel, J.-H., Eyre, T., Redmond, S., Banerjee, R., Chi, J., Fu, B., Langley, E., Maguire, S.F., Laird, G.K., Lloyd, D., Kenyon, E., Donaldson, S., Sehra, H., Almeida-King, J., Loveland, J., Trevanion, S., Jones, M., Quail, M., Willey, D., Hunt, A., Burton, J., Sims, S., McLay, K., Plumb, B., Davis, J., Clee, C., Oliver, K., Clark, R., Riddle, C., Eliott, D., Threadgold, G., Harden, G., Ware, D., Mortimer, B., Kerry, G., Heath, P., Phillimore, B., Tracey, A., Corby, N., Dunn, M., Johnson, C., Wood, J., Clark, S., Pelan, S., Griffiths, G., Smith, M., Glithero, R., Howden, P., Barker, N., Stevens, C., Harley, J., Holt, K., Panagiotidis, G., Lovell, J., Beasley, H., Henderson, C., Gordon, D., Auger, K., Wright, D., Collins, J., Raisen, C., Dyer, L., Leung, K., Robertson, L., Ambridge, K., Leongamornlert, D., McGuire, S., Gilderthorp, R., Griffiths, C., Manthravadi, D., Nichol, S., Barker, G., Whitehead, S., Kay, M., Brown, J., Murnane, C., Gray, E., Humphries, M., Sycamore, N., Barker, D., Saunders, D., Wallis, J., Babbage, A., Hammond, S., Mashreghi-Mohammadi, M., Barr, L., Martin, S., Wray, P., Ellington, A., Matthews, N., Ellwood,

M., Woodmansey, R., Clark, G., Cooper, J., Tromans, A., Grafham, D., Skuce, C., Pandian, R., Andrews, R., Harrison, E., Kimberley, A., Garnett, J., Fosker, N., Hall, R., Garner, P., Kelly, D., Bird, C., Palmer, S., Gehring, I., Berger, A., Dooley, C.M., Ersan-Urun, Z., Eser, C., Geiger, H., Geisler, M., Karotki, L., Kirn, A., Konantz, J., Konantz, M., Oberlander, M., Rudolph-Geiger, S., Teucke, M., Osoegawa, K., Zhu, B., Rapp, A., Widaa, S., Langford, C., Yang, F., Carter, N.P., Harrow, J., Ning, Z., Herrero, J., Searle, S.M.J., Enright, A., Geisler, R., Plasterk, R.H.A., Lee, C., Westerfield, M., de Jong, P.J., Zon, L.I., Postlethwait, J.H., Nusslein-Volhard, C., Hubbard, T.J.P., Crollius, H.R., Rogers, J., and Stemple, D.L. (2013). The zebrafish reference genome sequence and its relationship to the human genome. Nature *advance online publication*.

Howell, C.Y., Bestor, T.H., Ding, F., Latham, K.E., Mertineit, C., Trasler, J.M., and Chaillet, J.R. (2001). Genomic Imprinting Disrupted by a Maternal Effect Mutation in the Dnmt1 Gene. Cell *104*, 829-838.

Howlett, S.K., and Reik, W. (1991). Methylation levels of maternal and paternal genomes during preimplantation development. Development *113*, 119-127.

Hsieh, C.L. (1994). Dependence of transcriptional repression on CpG methylation density. Mol Cell Biol *14*, 5487-5494.

Hsieh, C.L. (1999). Evidence that protein binding specifies sites of DNA demethylation. Mol Cell Biol *19*, 46-56.

Ideraabdullah, F.Y., Vigneau, S., and Bartolomei, M.S. (2008). Genomic imprinting mechanisms in mammals. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis *647*, 77-85.

Ignatius, M.S., Unal Eroglu, A., Malireddy, S., Gallagher, G., Nambiar, R.M., and Henion, P.D. (2013). Distinct functional and temporal requirements for zebrafish Hdac1 during neural crest-derived craniofacial and peripheral neuron development. PLoS One *8*, e63218.

Iida, T., Suetake, I., Tajima, S., Morioka, H., Ohta, S., Obuse, C., and Tsurimoto, T. (2002). PCNA clamp facilitates action of DNA cytosine methyltransferase 1 on hemimethylated DNA. Genes Cells *7*, 997-1007.

Ikegami, K., Iwatani, M., Suzuki, M., Tachibana, M., Shinkai, Y., Tanaka, S., Greally, J.M., Yagi, S., Hattori, N., and Shiota, K. (2007). Genome-wide and locus-specific DNA hypomethylation in G9a deficient mouse embryonic stem cells. Genes to Cells *12*, 1-11.

Illingworth, R., Kerr, A., DeSousa, D., Jørgensen, H., Ellis, P., Stalker, J., Jackson, D., Clee, C., Plumb, R., Rogers, J., Humphray, S., Cox, T., Langford, C., and Bird, A. (2008). A Novel CpG Island Set Identifies Tissue-Specific Methylation at Developmental Gene Loci. PLoS Biol *6*, e22.

Illingworth, R.S., Gruenewald-Schneider, U., Webb, S., Kerr, A.R., James, K.D., Turner, D.J., Smith, C., Harrison, D.J., Andrews, R., and Bird, A.P. (2010). Orphan CpG islands identify numerous conserved promoters in the mammalian genome. PLoS Genet *6*.

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

Iqbal, K., Jin, S.G., Pfeifer, G.P., and Szabo, P.E. (2011). Reprogramming of the paternal genome upon fertilization involves genome-wide oxidation of 5-methylcytosine. Proc Natl Acad Sci U S A *108*, 3642-3647.

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

Issa, J.-P.J., and Kantarjian, H.M. (2009). Targeting DNA Methylation. Clinical Cancer Research *15*, 3938-3946.

Ito, S., Shen, L., Dai, Q., Wu, S.C., Collins, L.B., Swenberg, J.A., He, C., and Zhang, Y. (2011). Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. Science *333*, 1300-1303.

Iwase, S., Xiang, B., Ghosh, S., Ren, T., Lewis, P.W., Cochrane, J.C., Allis, C.D., Picketts, D.J., Patel, D.J., Li, H., and Shi, Y. (2011). ATRX ADD domain links an atypical histone methylation recognition mechanism to human mental-retardation syndrome. Nat Struct Mol Biol *18*, 769-776.

Iyer, L.M., Abhiman, S., and Aravind, L. (2011). Natural history of eukaryotic DNA methylation systems. Progress in molecular biology and translational science *101*, 25-104.

Iyer, L.M., Tahiliani, M., Rao, A., and Aravind, L. (2009). Prediction of novel families of enzymes involved in oxidative and other complex modifications of bases in nucleic acids. Cell Cycle *8*, 1698-1710.

Jacobs, S.A., Taverna, S.D., Zhang, Y., Briggs, S.D., Li, J., Eissenberg, J.C., Allis, C.D., and Khorasanizadeh, S. (2001). Specificity of the HP1 chromo domain for the methylated N-terminus of histone H3. EMBO J *20*, 5232-5241.

Jeddeloh, J.A., Stokes, T.L., and Richards, E.J. (1999). Maintenance of genomic methylation requires a SWI2/SNF2-like protein. Nature genetics *22*, 94-97.

Jeltsch, A., Nellen, W., and Lyko, F. (2006). Two substrates are better than one: dual specificities for Dnmt2 methyltransferases. Trends Biochem Sci *31*, 306-308.

Jia, D., Jurkowska, R.Z., Zhang, X., Jeltsch, A., and Cheng, X. (2007). Structure of Dnmt3a bound to Dnmt3L suggests a model for de novo DNA methylation. Nature *449*, 248-251.

Jiang, L., Zhang, J., Wang, J.J., Wang, L., Zhang, L., Li, G., Yang, X., Ma, X., Sun, X., Cai, J., Zhang, J., Huang, X., Yu, M., Wang, X., Liu, F., Wu, C.I., He, C., Zhang, B., Ci, W., and Liu, J. (2013). Sperm, but not oocyte, DNA methylome is inherited by zebrafish early embryos. Cell *153*, 773-784.

Jin, C., and Felsenfeld, G. (2007). Nucleosome stability mediated by histone variants H3.3 and H2A.Z. Genes & development *21*, 1519-1529.

Jin, C., Zang, C., Wei, G., Cui, K., Peng, W., Zhao, K., and Felsenfeld, G. (2009). H3.3/H2A.Z double variant-containing nucleosomes mark 'nucleosome-free regions' of active promoters and other regulatory regions. Nature genetics *41*, 941-945.

Johnson, T.B., and Coghill, R.D. (1925). RESEARCHES ON PYRIMIDINES. C111. THE DISCOVERY OF 5-METHYL-CYTOSINE IN TUBERCULINIC ACID, THE NUCLEIC ACID OF THE TUBERCLE BACILLUS1. Journal of the American Chemical Society *47*, 2838-2844.

Jones, P.A. (1999). The DNA methylation paradox. Trends Genet *15*, 34-37.

Jones, P.A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet *13*, 484-492.

Jones, P.A., and Taylor, S.M. (1980). Cellular differentiation, cytidine analogs and DNA methylation. Cell *20*, 85-93.

Jones, P.L., Jan Veenstra, G.C., Wade, P.A., Vermaak, D., Kass, S.U., Landsberger, N., Strouboulis, J., and Wolffe, A.P. (1998). Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription. Nature genetics *19*, 187-191.

Jonk, L.J.C., Itoh, S., Heldin, C.-H., ten Dijke, P., and Kruijer, W. (1998). Identification and Functional Characterization of a Smad Binding Element (SBE) in the JunB Promoter That Acts as a Transforming Growth Factor-β, Activin, and Bone Morphogenetic Protein-inducible Enhancer. Journal of Biological Chemistry *273*, 21145-21152.

Jorgensen, H.F., Ben-Porath, I., and Bird, A.P. (2004). Mbd1 is recruited to both methylated and nonmethylated CpGs via distinct DNA binding domains. Mol Cell Biol *24*, 3387-3395.

Josse, J., Kaiser, A.D., and Kornberg, A. (1961). Enzymatic synthesis of deoxyribonucleic acid. VIII. Frequencies of nearest neighbor base sequences in deoxyribonucleic acid. J Biol Chem *236*, 864-875.

Julien, P., Brawand, D., Soumillon, M., Necsulea, A., Liechti, A., Schütz, F., Daish, T., Grützner, F., and Kaessmann, H. (2012). Mechanisms and Evolutionary Patterns of Mammalian and Avian Dosage Compensation. PLoS Biol *10*, e1001328.

Jurgens, G. (1985). A group of genes controlling the spatial expression of the bithorax complex in Drosophila. Nature *316*, 153-155.

Jurkowski, T.P., and Jeltsch, A. (2011). On the evolutionary origin of eukaryotic DNA methyltransferases and Dnmt2. PLoS One *6*, e28104.

Kakutani, T., Jeddeloh, J.A., Flowers, S.K., Munakata, K., and Richards, E.J. (1996). Developmental abnormalities and epimutations associated with DNA hypomethylation mutations. Proc Natl Acad Sci U S A *93*, 12406-12411.

Kakutani, T., Jeddeloh, J.A., and Richards, E.J. (1995). Characterization of an Arabidopsis thaliana DNA hypomethylation mutant. Nucleic Acids Res *23*, 130-137.

Kaludov, N.K., and Wolffe, A.P. (2000). MeCP2 driven transcriptional repression in vitro: selectivity for methylated DNA, action at a distance and contacts with the basal transcription machinery. Nucleic Acids Res *28*, 1921-1928.

Kane, D.A., and Kimmel, C.B. (1993). The zebrafish midblastula transition. Development *119*, 447-456.

Kaneda, M., Okano, M., Hata, K., Sado, T., Tsujimoto, N., Li, E., and Sasaki, H. (2004). Essential role for de novo DNA methyltransferase Dnmt3a in paternal and maternal imprinting. Nature *429*, 900-903.

Kangaspeska, S., Stride, B., Metivier, R., Polycarpou-Schwarz, M., Ibberson, D., Carmouche, R.P., Benes, V., Gannon, F., and Reid, G. (2008). Transient cyclical methylation of promoter DNA. Nature *452*, 112-115.

Kareta, M.S., Botello, Z.M., Ennis, J.J., Chou, C., and Chedin, F. (2006). Reconstitution and mechanism of the stimulation of de novo methylation by human DNMT3L. J Biol Chem *281*, 25893-25902.

Kass, S.U., Landsberger, N., and Wolffe, A.P. (1997). DNA methylation directs a time-dependent repression of transcription initiation. Current biology : CB *7*, 157-165.

Kazuki, Y., Shinohara, T., Tomizuka, K., Katoh, M., Ohguma, A., Ishida, I., and Oshimura, M. (2001). Germline transmission of a transferred human chromosome 21 fragment in transchromosomal mice. J Hum Genet *46*, 600-603.

Keightley, P.D., Kryukov, G.V., Sunyaev, S., Halligan, D.L., and Gaffney, D.J. (2005). Evolutionary constraints in conserved nongenic sequences of mammals. Genome Res *15*, 1373-1378.

Kelly, G.M., Greenstein, P., Erezyilmaz, D.F., and Moon, R.T. (1995). Zebrafish wnt8 and wnt8b share a common activity but are involved in distinct developmental pathways. Development *121*, 1787-1799.

Kelsey, G., and Feil, R. (2013). New insights into establishment and maintenance of DNA methylation imprints in mammals. Philosophical Transactions of the Royal Society B: Biological Sciences *368*.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Res *12*, 996-1006.

Keyes, W.M., Pecoraro, M., Aranda, V., Vernersson-Lindahl, E., Li, W., Vogel, H., Guo, X., Garcia, E.L., Michurina, T.V., Enikolopov, G., Muthuswamy, S.K., and Mills, A.A. (2011). ΔNp63α Is an Oncogene that Targets Chromatin Remodeler Lsh to Drive Skin Stem Cell Proliferation and Tumorigenesis. Cell stem cell *8*, 164-176.

Kim, D.H., Nelson, H.H., Wiencke, J.K., Christiani, D.C., Wain, J.C., Mark, E.J., and Kelsey, K.T. (2001). Promoter methylation of DAP-kinase: association with advanced stage in non-small cell lung cancer. Oncogene *20*, 1765-1770.

Kim, J., Kollhoff, A., Bergmann, A., and Stubbs, L. (2003). Methylation-sensitive binding of transcription factor YY1 to an insulator sequence within the paternally expressed imprinted gene, Peg3. Hum Mol Genet *12*, 233-245.

Kim, S.H., Kang, Y.K., Koo, D.B., Kang, M.J., Moon, S.J., Lee, K.K., and Han, Y.M. (2004). Differential DNA methylation reprogramming of various repetitive sequences in mouse preimplantation embryos. Biochem Biophys Res Commun *324*, 58-63.

Kim, S.J., Kang, H.S., Chang, H.L., Jung, Y.C., Sim, H.B., Lee, K.S., Ro, J., and Lee, E.S. (2008). Promoter hypomethylation of the N-acetyltransferase 1 gene in breast cancer. Oncology reports *19*, 663-668.

Kimmel, C.B., Ballard, W.W., Kimmel, S.R., Ullmann, B., and Schilling, T.F. (1995). Stages of embryonic development of the zebrafish. Developmental Dynamics *203*, 253-310.

Kimmel, C.B., and Law, R.D. (1985). Cell lineage of zebrafish blastomeres. III. Clonal analyses of the blastula and gastrula stages. Dev Biol *108*, 94-101.

Klimasauskas, S., Kumar, S., Roberts, R.J., and Cheng, X. (1994). HhaI methyltransferase flips its target base out of the DNA helix. Cell *76*, 357-369.

Klose, R.J., and Bird, A.P. (2006). Genomic DNA methylation: the mark and its mediators. Trends Biochem Sci *31*, 89-97.

Klose, R.J., Cooper, S., Farcas, A.M., Blackledge, N.P., and Brockdorff, N. (2013). Chromatin Sampling—An Emerging Perspective on Targeting Polycomb Repressor Proteins. PLoS Genet *9*, e1003717.

Klose, R.J., Kallin, E.M., and Zhang, Y. (2006). JmjC-domain-containing proteins and histone demethylation. Nat Rev Genet *7*, 715-727.

Klose, R.J., and Zhang, Y. (2007). Regulation of histone methylation by demethylimination and demethylation. Nat Rev Mol Cell Biol *8*, 307-318.

Koch, C.M., Andrews, R.M., Flicek, P., Dillon, S.C., Karaoz, U., Clelland, G.K., Wilcox, S., Beare, D.M., Fowler, J.C., Couttet, P., James, K.D., Lefebvre, G.C., Bruce, A.W., Dovey, O.M., Ellis, P.D., Dhami, P., Langford, C.F., Weng, Z., Birney, E., Carter, N.P., Vetrie, D., and Dunham, I. (2007). The landscape of histone modifications across 1% of the human genome in five human cell lines. Genome Res *17*, 691-707.

Kondo, E., Gu, Z., Horii, A., and Fukushige, S. (2005). The thymine DNA glycosylase MBD4 represses transcription and is associated with methylated p16(INK4a) and hMLH1 genes. Mol Cell Biol *25*, 4388-4396.

Kondo, Y., Shen, L., Yan, P.S., Huang, T.H., and Issa, J.P. (2004). Chromatin immunoprecipitation microarrays for identification of genes silenced by histone H3 lysine 9 methylation. Proc Natl Acad Sci U S A *101*, 7398-7403.

Kouzarides, T. (2007). Chromatin Modifications and Their Function. Cell *128*, 693-705.

Kouzminova, E., and Selker, E.U. (2001). dim-2 encodes a DNA methyltransferase responsible for all known cytosine methylation in Neurospora. EMBO J *20*, 4309-4323.

Krajewski, W.A. (2000). Histone hyperacetylation facilitates chromatin remodelling in a Drosophila embryo cell-free system. Molecular & general genetics : MGG *263*, 38-47.

Kriaucionis, S., and Bird, A. (2003). DNA methylation and Rett syndrome. Hum Mol Genet *12 Spec No 2*, R221-227.

Kriaucionis, S., and Heintz, N. (2009). The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. Science *324*, 929-930.

Krueger, F., Kreck, B., Franke, A., and Andrews, S.R. (2012). DNA methylome analysis using short bisulfite sequencing data. Nat Methods *9*, 145-151.

Ku, M., Koche, R.P., Rheinbay, E., Mendenhall, E.M., Endoh, M., Mikkelsen, T.S., Presser, A., Nusbaum, C., Xie, X., Chi, A.S., Adli, M., Kasif, S., Ptaszek, L.M., Cowan, C.A., Lander, E.S., Koseki, H., and Bernstein, B.E. (2008). Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains. PLoS Genet *4*, e1000242.

Kumaki, Y., Oda, M., and Okano, M. (2008). QUMA: quantification tool for methylation analysis. Nucleic Acids Res *36*, W170-175.

Kunert, N., Marhold, J., Stanke, J., Stach, D., and Lyko, F. (2003). A Dnmt2-like protein mediates DNA methylation in Drosophila. Development *130*, 5083-5090.

Kurihara, Y., Kawamura, Y., Uchijima, Y., Amamo, T., Kobayashi, H., Asano, T., and Kurihara, H. (2008). Maintenance of genomic methylation patterns during preimplantation development requires the somatic form of DNA methyltransferase 1. Developmental Biology *313*, 335-346.

Kuzmichev, A., Nishioka, K., Erdjument-Bromage, H., Tempst, P., and Reinberg, D. (2002). Histone methyltransferase activity associated with a human multiprotein complex containing the Enhancer of Zeste protein. Genes & development *16*, 2893-2905.

La Salle, S., and Trasler, J.M. (2006). Dynamic expression of DNMT3a and DNMT3b isoforms during male germ cell development in the mouse. Dev Biol *296*, 71-82.

Lachner, M., O'Carroll, D., Rea, S., Mechtler, K., and Jenuwein, T. (2001). Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. Nature *410*, 116-120.

Lagarou, A., Mohd-Sarip, A., Moshkin, Y.M., Chalkley, G.E., Bezstarosti, K., Demmers, J.A., and Verrijzer, C.P. (2008). dKDM2 couples histone H2A ubiquitylation to histone H3 demethylation during Polycomb group silencing. Genes Dev *22*, 2799-2810.

Laird, P.W. (2010). Principles and challenges of genome-wide DNA methylation analysis. Nat Rev Genet *11*, 191-203.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R.A., Muzny, D.M., Scherer, S.E., Bouck, J.B., Sodergren, E.J., Worley, K.C., Rives, C.M., Gorrell, J.H., Metzker, M.L., Naylor, S.L., Kucherlapati, R.S., Nelson, D.L., Weinstock, G.M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D.R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H.M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R.W., Federspiel, N.A., Abola, A.P., Proctor, M.J., Myers, R.M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D.R., Olson, M.V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G.A., Athanasiou, M., Schultz, R., Roe, B.A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W.R., de la Bastide, M., Dedhia, N., Blocker, H., Hornischer,

K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J.A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D.G., Burge, C.B., Cerutti, L., Chen, H.C., Church, D., Clamp, M., Copley, R.R., Doerks, T., Eddy, S.R., Eichler, E.E., Furey, T.S., Galagan, J., Gilbert, J.G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L.S., Jones, T.A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W.J., Kitts, P., Koonin, E.V., Korf, I., Kulp, D., Lancet, D., Lowe, T.M., McLysaght, A., Mikkelsen, T., Moran, J.V., Mulder, N., Pollara, V.J., Ponting, C.P., Schuler, G., Schultz, J., Slater, G., Smit, A.F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y.I., Wolfe, K.H., Yang, S.P., Yeh, R.F., Collins, F., Guyer, M.S., Peterson, J., Felsenfeld, A., Wetterstrand, K.A., Patrinos, A., Morgan, M.J., de Jong, P., Catanese, J.J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y.J., and International Human Genome Sequencing, C. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860-921.

Lane, N., Dean, W., Erhardt, S., Hajkova, P., Surani, A., Walter, J., and Reik, W. (2003). Resistance of IAPs to methylation reprogramming may provide a mechanism for epigenetic inheritance in the mouse. Genesis *35*, 88-93.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol *10*, R25.

Lauberth, Shannon M., Nakayama, T., Wu, X., Ferris, Andrea L., Tang, Z., Hughes, Stephen H., and Roeder, Robert G. (2013). H3K4me3 Interactions with TAF3 Regulate Preinitiation Complex Assembly and Selective Gene Activation. Cell *152*, 1021-1036.

Laurent, L., Wong, E., Li, G., Huynh, T., Tsirigos, A., Ong, C.T., Low, H.M., Kin Sung, K.W., Rigoutsos, I., Loring, J., and Wei, C.-L. (2010). Dynamic changes in the human methylome during differentiation. Genome Research *20*, 320-331.

Law, J.A., and Jacobsen, S.E. (2010). Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nat Rev Genet *11*, 204-220.

Law, M.J., Lower, K.M., Voon, H.P., Hughes, J.R., Garrick, D., Viprakasit, V., Mitson, M., De Gobbi, M., Marra, M., Morris, A., Abbott, A., Wilder, S.P., Taylor, S., Santos, G.M., Cross, J., Ayyub, H., Jones, S., Ragoussis, J., Rhodes, D., Dunham, I., Higgs, D.R., and Gibbons, R.J. (2010). ATR-X syndrome protein targets tandem repeats and influences allele-specific expression in a size-dependent manner. Cell *143*, 367-378.

Le Faou, P., Völkel, P., and Angrand, P.-O. (2011). The zebrafish genes encoding the Polycomb repressive complex (PRC) 1. Gene *475*, 10-21.

Lechner, M.S., Schultz, D.C., Negorev, D., Maul, G.G., and Rauscher, F.J., 3rd (2005). The mammalian heterochromatin protein 1 binds diverse nuclear proteins through a common motif that targets the chromoshadow domain. Biochem Biophys Res Commun *331*, 929-937.

Lee, J.H., and Skalnik, D.G. (2005). CpG-binding protein (CXXC finger protein 1) is a component of the mammalian Set1 histone H3-Lys4 methyltransferase complex, the analogue of the yeast Set1/COMPASS complex. J Biol Chem *280*, 41725-41731.

Lee, J.H., Tate, C.M., You, J.S., and Skalnik, D.G. (2007). Identification and characterization of the human Set1B histone H3-Lys4 methyltransferase complex. J Biol Chem *282*, 13419-13428.

Lee, M.T., Bonneau, A.R., Takacs, C.M., Bazzini, A.A., DiVito, K.R., Fleming, E.S., and Giraldez, A.J. (2013). Nanog, Pou5f1 and SoxB1 activate zygotic gene expression during the maternal-to-zygotic transition. Nature *advance online publication*.

Lehmann, L., Ferrari, R., Vashisht, A.A., Wohlschlegel, J.A., Kurdistani, S.K., and Carey, M. (2012). Polycomb repressive complex 1 (PRC1) disassembles RNA polymerase II preinitiation complexes. J Biol Chem *287*, 35784-35794.

Lehnertz, B., Ueda, Y., Derijck, A.A., Braunschweig, U., Perez-Burgos, L., Kubicek, S., Chen, T., Li, E., Jenuwein, T., and Peters, A.H. (2003). Suv39h-mediated histone H3 lysine 9 methylation directs DNA methylation to major satellite repeats at pericentric heterochromatin. Current biology : CB *13*, 1192-1200.

Lei, H., Oh, S.P., Okano, M., Juttermann, R., Goss, K.A., Jaenisch, R., and Li, E. (1996). De novo DNA cytosine methyltransferase activities in mouse embryonic stem cells. Development *122*, 3195-3205.

Leonhardt, H., Page, A.W., Weier, H.U., and Bestor, T.H. (1992). A targeting sequence directs DNA methyltransferase to sites of DNA replication in mammalian nuclei. Cell *71*, 865-873.

Leung, T., Bischof, J., Söll, I., Niessing, D., Zhang, D., Ma, J., Jäckle, H., and Driever, W. (2003). bozozok directly represses bmp2b transcription and mediates the earliest dorsoventral asymmetry of bmp2b expression in zebrafish. Development *130*, 3639-3649.

Lewis, E.B. (1978). A gene complex controlling segmentation in Drosophila. Nature *276*, 565-570.

Lewis, J.D., Meehan, R.R., Henzel, W.J., Maurer-Fogy, I., Jeppesen, P., Klein, F., and Bird, A. (1992). Purification, sequence, and cellular localization of a novel chromosomal protein that binds to methylated DNA. Cell *69*, 905-914.

Lewis, P.W., Elsaesser, S.J., Noh, K.M., Stadler, S.C., and Allis, C.D. (2010). Daxx is an H3.3-specific histone chaperone and cooperates with ATRX in replication-independent chromatin assembly at telomeres. Proc Natl Acad Sci U S A *107*, 14075-14080.

Lewis, Z.A., Honda, S., Khlafallah, T.K., Jeffress, J.K., Freitag, M., Mohn, F., Schubeler, D., and Selker, E.U. (2009). Relics of repeat-induced point mutation direct heterochromatin formation in Neurospora crassa. Genome Res *19*, 427-437.

Li, B., Jackson, J., Simon, M.D., Fleharty, B., Gogol, M., Seidel, C., Workman, J.L., and Shilatifard, A. (2009). Histone H3 lysine 36 dimethylation (H3K36me2) is sufficient to recruit the Rpd3s histone deacetylase complex and to repress spurious transcription. J Biol Chem *284*, 7970-7976.

Li, E. (2002). Chromatin modification and epigenetic reprogramming in mammalian development. Nat Rev Genet *3*, 662-673.

Li, E., Beard, C., and Jaenisch, R. (1993). Role for DNA methylation in genomic imprinting. Nature *366*, 362-365.

Li, E., Bestor, T.H., and Jaenisch, R. (1992). Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. Cell *69*, 915-926.

Li, H., Rauch, T., Chen, Z.X., Szabo, P.E., Riggs, A.D., and Pfeifer, G.P. (2006). The histone methyltransferase SETDB1 and the DNA methyltransferase DNMT3A interact directly and localize to promoters silenced in cancer cells. J Biol Chem *281*, 19489-19500.

Li, X., Ito, M., Zhou, F., Youngson, N., Zuo, X., Leder, P., and Ferguson-Smith, A.C. (2008). A maternal-zygotic effect gene, Zfp57, maintains both maternal and paternal imprints. Dev Cell *15*, 547-557.

Liang, G., He, J., and Zhang, Y. (2012). Kdm2b promotes induced pluripotent stem cell generation by facilitating gene activation early in reprogramming. Nature cell biology *14*, 457-466.

Lienert, F., Wirbelauer, C., Som, I., Dean, A., Mohn, F., and Schubeler, D. (2011). Identification of genetic elements that autonomously determine DNA methylation states. Nature genetics *43*, 1091-1097.

Lieschke, G.J., and Currie, P.D. (2007). Animal models of human disease: zebrafish swim into view. Nature Reviews Genetics *8*, 353-367.

Lim, D.H.K., and Maher, E.R. (2009). Human imprinting syndromes. Epigenomics *1*, 347-369.

Lin, I.G., and Hsieh, C.L. (2001). Chromosomal DNA demethylation specified by protein binding. EMBO reports *2*, 108-112.

Lin, I.G., Tomzynski, T.J., Ou, Q., and Hsieh, C.L. (2000). Modulation of DNA binding protein affinity directly affects target site demethylation. Mol Cell Biol *20*, 2343-2349.

Lindahl, T. (1993). Instability and decay of the primary structure of DNA. Nature *362*, 709-715.

Lindeman, L.C., Andersen, I.S., Reiner, A.H., Li, N., Aanes, H., Ostrup, O., Winata, C., Mathavan, S., Muller, F., Alestrom, P., and Collas, P. (2011). Prepatterning of developmental gene expression by modified histones before zygotic genome activation. Dev Cell *21*, 993-1004.

Lindeman, L.C., Winata, C.L., Aanes, H., Mathavan, S., Alestrom, P., and Collas, P. (2010). Chromatin states of developmentally-regulated genes revealed by DNA and histone methylation patterns in zebrafish embryos. The International journal of developmental biology *54*, 803-813.

Link, V., Shevchenko, A., and Heisenberg, C.P. (2006). Proteomics of early zebrafish embryos. BMC developmental biology *6*, 1.

Lipps, H.J., and Rhodes, D. (2009). G-quadruplex structures: in vivo evidence and function. Trends Cell Biol *19*, 414-422.

Lister, R., O'Malley, R.C., Tonti-Filippini, J., Gregory, B.D., Berry, C.C., Millar, A.H., and Ecker, J.R. (2008). Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell *133*, 523-536.

Lister, R., Pelizzola, M., Dowen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q.M., Edsall, L., Antosiewicz-Bourget, J., Stewart, R., Ruotti, V., Millar, A.H., Thomson, J.A., Ren, B., and Ecker, J.R. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. Nature *462*, 315-322.

Liu, X., Gao, Q., Li, P., Zhao, Q., Zhang, J., Li, J., Koseki, H., and Wong, J. (2013). UHRF1 targets DNMT1 for DNA methylation through cooperative binding of hemi-methylated DNA and methylated H3K9. Nat Commun *4*, 1563.

Liu, Y., Toh, H., Sasaki, H., Zhang, X., and Cheng, X. (2012). An atomic model of Zfp57 recognition of CpG methylation within a specific DNA sequence. Genes Dev *26*, 2374-2379.

Lock, L.F., Takagi, N., and Martin, G.R. (1987). Methylation of the Hprt gene on the inactive X occurs after chromosome inactivation. Cell *48*, 39-46.

Long, H.K., Blackledge, N.P., and Klose, R.J. (2013). ZF-CxxC domain-containing proteins, CpG islands and the chromatin connection. Biochemical Society transactions *41*, 727-740.

Lucifero, D., La Salle, S., Bourc'his, D., Martel, J., Bestor, T.H., and Trasler, J.M. (2007). Coordinate regulation of DNA methyltransferase expression during oogenesis. BMC developmental biology *7*, 36.

Luteijn, M.J., and Ketting, R.F. (2013). PIWI-interacting RNAs: from generation to transgenerational epigenetics. Nat Rev Genet *14*, 523-534.

Lyko, F., Ramsahoye, B.H., and Jaenisch, R. (2000). Development: DNA methylation in Drosophila melanogaster. Nature *408*, 538-540.

Lynch, M.D., Smith, A.J., De Gobbi, M., Flenley, M., Hughes, J.R., Vernimmen, D., Ayyub, H., Sharpe, J.A., Sloane-Stanley, J.A., Sutherland, L., Meek, S., Burdon, T., Gibbons, R.J., Garrick, D., and Higgs, D.R. (2012). An interspecies analysis reveals a key role for unmethylated CpG dinucleotides in vertebrate Polycomb complex recruitment. EMBO J *31*, 317-329.

Lyst, M.J., Ekiert, R., Ebert, D.H., Merusi, C., Nowak, J., Selfridge, J., Guy, J., Kastan, N.R., Robinson, N.D., de Lima Alves, F., Rappsilber, J., Greenberg, M.E., and Bird, A. (2013). Rett syndrome mutations abolish the interaction of MeCP2 with the NCoR/SMRT co-repressor. Nature neuroscience *16*, 898-902.

Lyst, M.J., Nan, X., and Stancheva, I. (2006). Regulation of MBD1-mediated transcriptional repression by SUMO and PIAS proteins. EMBO J *25*, 5317-5328.

Macleod, D., Clark, V.H., and Bird, A. (1999). Absence of genome-wide changes in DNA methylation during development of the zebrafish. Nature genetics *23*, 139-140.

Mancini, D.N., Singh, S.M., Archer, T.K., and Rodenhiser, D.I. (1999). Site-specific DNA methylation in the neurofibromatosis (NF1) promoter interferes with binding of CREB and SP1 transcription factors. Oncogene *18*, 4108-4119.

Margot, J.B., Cardoso, M.C., and Leonhardt, H. (2001). Mammalian DNA methyltransferases show different subnuclear distributions. Journal of cellular biochemistry *83*, 373-379.

Margueron, R., Justin, N., Ohno, K., Sharpe, M.L., Son, J., Drury Iii, W.J., Voigt, P., Martin, S.R., Taylor, W.R., De Marco, V., Pirrotta, V., Reinberg, D., and Gamblin, S.J. (2009). Role of the polycomb protein EED in the propagation of repressive histone marks. Nature *461*, 762-767.

Margueron, R., Trojer, P., and Reinberg, D. (2005). The key to development: interpreting the histone code? Current opinion in genetics & development *15*, 163-176.

Martens, J.H., O'Sullivan, R.J., Braunschweig, U., Opravil, S., Radolf, M., Steinlein, P., and Jenuwein, T. (2005). The profile of repeat-associated histone lysine methylation states in the mouse epigenome. EMBO J *24*, 800-812.

Martin, C., and Zhang, Y. (2005). The diverse functions of histone lysine methylation. Nat Rev Mol Cell Biol *6*, 838-849.

Martin Caballero, I., Hansen, J., Leaford, D., Pollard, S., and Hendrich, B.D. (2009). The methyl-CpG binding proteins Mecp2, Mbd2 and Kaiso are dispensable for mouse embryogenesis, but play a redundant function in neural differentiation. PLoS One *4*, e4315.

Matsuo, K., Silke, J., Georgiev, O., Marti, P., Giovannini, N., and Rungger, D. (1998). An embryonic demethylation mechanism involving binding of transcription factors to replicating DNA. EMBO J *17*, 1446-1453.

Matzke, M., Kanno, T., Daxinger, L., Huettel, B., and Matzke, A.J. (2009). RNA-mediated chromatin-based silencing in plants. Curr Opin Cell Biol *21*, 367-376.

Maunakea, A.K., Nagarajan, R.P., Bilenky, M., Ballinger, T.J., D'Souza, C., Fouse, S.D., Johnson, B.E., Hong, C., Nielsen, C., Zhao, Y., Turecki, G., Delaney, A., Varhol, R., Thiessen, N., Shchors, K., Heine, V.M., Rowitch, D.H., Xing, X., Fiore, C., Schillebeeckx, M., Jones, S.J., Haussler, D., Marra, M.A., Hirst, M., Wang, T., and Costello, J.F. (2010). Conserved role of intragenic DNA methylation in regulating alternative promoters. Nature *466*, 253-257.

Mayer, W., Niveleau, A., Walter, J., Fundele, R., and Haaf, T. (2000). Demethylation of the zygotic paternal genome. Nature *403*, 501-502.

McDowell, T.L., Gibbons, R.J., Sutherland, H., O'Rourke, D.M., Bickmore, W.A., Pombo, A., Turley, H., Gatter, K., Picketts, D.J., Buckle, V.J., Chapman, L., Rhodes, D., and Higgs, D.R. (1999). Localization of a putative transcriptional regulator (ATRX) at pericentromeric heterochromatin and the short arms of acrocentric chromosomes. Proc Natl Acad Sci U S A *96*, 13983-13988.

Meehan, R.R., Lewis, J.D., McKay, S., Kleiner, E.L., and Bird, A.P. (1989). Identification of a mammalian protein that binds specifically to DNA containing methylated CpGs. Cell *58*, 499-507.

Meehan, R.R., Pennings, S., and Stancheva, I. (2001). Lashings of DNA methylation, forkfuls of chromatin remodeling. Genes Dev *15*, 3231-3236.

Meissner, A., Mikkelsen, T.S., Gu, H., Wernig, M., Hanna, J., Sivachenko, A., Zhang, X., Bernstein, B.E., Nusbaum, C., Jaffe, D.B., Gnirke, A., Jaenisch, R., and Lander, E.S. (2008). Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature *454*, 766-770.

Mendenhall, E.M., Koche, R.P., Truong, T., Zhou, V.W., Issac, B., Chi, A.S., Ku, M., and Bernstein, B.E. (2010). GC-Rich Sequence Elements Recruit PRC2 in Mammalian ES Cells. PLoS Genet *6*, e1001244.

Merlo, A., Herman, J.G., Mao, L., Lee, D.J., Gabrielson, E., Burger, P.C., Baylin, S.B., and Sidransky, D. (1995). 5' CpG island methylation is associated with transcriptional silencing of the tumour suppressor p16/CDKN2/MTS1 in human cancers. Nature medicine *1*, 686-692.

Meyer, C., Kowarz, E., Hofmann, J., Renneville, A., Zuna, J., Trka, J., Ben Abdelali, R., Macintyre, E., De Braekeleer, E., De Braekeleer, M., Delabesse, E., de Oliveira, M.P., Cave, H., Clappier, E., van Dongen, J.J., Balgobind, B.V., van den Heuvel-Eibrink, M.M., Beverloo, H.B., Panzer-Grumayer, R., Teigler-Schlegel, A., Harbott, J., Kjeldsen, E., Schnittger, S., Koehl, U., Gruhn, B., Heidenreich, O., Chan, L.C., Yip, S.F., Krzywinski, M., Eckert, C., Moricke, A., Schrappe, M., Alonso, C.N., Schafer, B.W., Krauter, J., Lee, D.A., Zur Stadt, U., Te Kronnie, G., Sutton, R., Izraeli, S., Trakhtenbrot, L., Lo Nigro, L., Tsaur, G., Fechina, L., Szczepanski, T., Strehl, S., Ilencikova, D., Molkentin, M., Burmeister, T., Dingermann, T., Klingebiel, T., and Marschalek, R. (2009). New insights to the MLL recombinome of acute leukemias. Leukemia *23*, 1490-1499.

Mi, S., and Roberts, R.J. (1993). The DNA binding affinity of Hhal methylase is increased by a single amino acid substitution in the catalytic center. Nucleic acids research *21*, 2459-2464.

Mikkelsen, T.S., Ku, M., Jaffe, D.B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.K., Koche, R.P., Lee, W., Mendenhall, E., O'Donovan, A., Presser, A., Russ, C., Xie, X., Meissner, A., Wernig, M., Jaenisch, R., Nusbaum, C., Lander, E.S., and Bernstein, B.E. (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature *448*, 553-560.

Millar, C.B., Guy, J., Sansom, O.J., Selfridge, J., MacDougall, E., Hendrich, B., Keightley, P.D., Bishop, S.M., Clarke, A.R., and Bird, A. (2002). Enhanced CpG mutability and tumorigenesis in MBD4-deficient mice. Science *297*, 403-405.

Miller, W., Rosenbloom, K., Hardison, R.C., Hou, M., Taylor, J., Raney, B., Burhans, R., King, D.C., Baertsch, R., Blankenberg, D., Kosakovsky Pond, S.L., Nekrutenko, A., Giardine, B., Harris, R.S., Tyekucheva, S., Diekhans, M., Pringle, T.H., Murphy, W.J., Lesk, A., Weinstock, G.M., Lindblad-Toh, K., Gibbs, R.A., Lander, E.S., Siepel, A., Haussler, D., and Kent, W.J. (2007). 28-Way vertebrate alignment and conservation track in the UCSC Genome Browser. Genome Research *17*, 1797-1808.

Milne, T.A., Kim, J., Wang, G.G., Stadler, S.C., Basrur, V., Whitcomb, S.J., Wang, Z., Ruthenburg, A.J., Elenitoba-Johnson, K.S., Roeder, R.G., and Allis, C.D. (2010). Multiple interactions recruit MLL1 and MLL1 fusion proteins to the HOXA9 locus in leukemogenesis. Mol Cell *38*, 853-863.

Min, J., Zhang, Y., and Xu, R.M. (2003). Structural basis for specific binding of Polycomb chromodomain to histone H3 methylated at Lys 27. Genes Dev *17*, 1823-1828.

Mitson, M., Kelley, L.A., Sternberg, M.J., Higgs, D.R., and Gibbons, R.J. (2011). Functional significance of mutations in the Snf2 domain of ATRX. Hum Mol Genet *20*, 2603-2610.

Mohn, F., Weber, M., Rebhan, M., Roloff, T.C., Richter, J., Stadler, M.B., Bibel, M., and Schubeler, D. (2008). Lineage-specific polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors. Mol Cell *30*, 755-766.

Molaro, A., Hodges, E., Fang, F., Song, Q., McCombie, W.R., Hannon, Gregory J., and Smith, Andrew D. (2011). Sperm Methylation Profiles Reveal Features of Epigenetic Inheritance and Evolution in Primates. Cell *146*, 1029-1041.

Monteiro, R.M., de Sousa Lopes, S.M.C., Bialecka, M., de Boer, S., Zwijsen, A., and Mummery, C.L. (2008). Real time monitoring of BMP Smads transcriptional activity during mouse development. genesis *46*, 335-346.

Moretti, P., and Zoghbi, H.Y. (2006). MeCP2 dysfunction in Rett syndrome and related disorders. Current opinion in genetics & development *16*, 276-281.

Morey, L., and Helin, K. (2010). Polycomb group protein-mediated repression of transcription. Trends in biochemical sciences *35*, 323-332.

Mortusewicz, O., Schermelleh, L., Walter, J., Cardoso, M.C., and Leonhardt, H. (2005). Recruitment of DNA methyltransferase I to DNA repair sites. Proc Natl Acad Sci U S A *102*, 8905-8909.

Mudbhary, R., and Sadler, K.C. (2011). Epigenetics, development, and cancer: Zebrafish make their mark. Birth Defects Research Part C: Embryo Today: Reviews *93*, 194-203.

Muegge, K. (2005). Lsh, a guardian of heterochromatin at repeat elements. Biochemistry and Cell Biology-Biochimie Et Biologie Cellulaire *83*, 548-554.

Muers, M.R., Sharpe, J.A., Garrick, D., Sloane-Stanley, J., Nolan, P.M., Hacker, T., Wood, W.G., Higgs, D.R., and Gibbons, R.J. (2007). Defining the Cause of Skewed X-Chromosome Inactivation in X-Linked Mental Retardation by Use of a Mouse Model. The American Journal of Human Genetics *80*, 1138-1149.

Müller, J., Hart, C.M., Francis, N.J., Vargas, M.L., Sengupta, A., Wild, B., Miller, E.L., O'Connor, M.B., Kingston, R.E., and Simon, J.A. (2002). Histone Methyltransferase Activity of a Drosophila Polycomb Group Repressor Complex. Cell *111*, 197-208.

Muntean, A.G., Tan, J., Sitwala, K., Huang, Y., Bronstein, J., Connelly, J.A., Basrur, V., Elenitoba-Johnson, K.S., and Hess, J.L. (2010). The PAF complex synergizes with MLL fusion proteins at HOX loci to promote leukemogenesis. Cancer Cell *17*, 609-621.

Myant, K., and Stancheva, I. (2008). LSH cooperates with DNA methyltransferases to repress transcription. Mol Cell Biol *28*, 215-226.

Myant, K., Termanis, A., Sundaram, A.Y., Boe, T., Li, C., Merusi, C., Burrage, J., de Las Heras, J.I., and Stancheva, I. (2011). LSH and G9a/GLP complex are required for developmentally programmed DNA methylation. Genome Res *21*, 83-94.

Nady, N., Lemak, A., Walker, J.R., Avvakumov, G.V., Kareta, M.S., Achour, M., Xue, S., Duan, S., Allali-Hassani, A., Zuo, X., Wang, Y.X., Bronner, C., Chedin, F., Arrowsmith, C.H., and Dhe-Paganon, S. (2011). Recognition of multivalent histone states associated with heterochromatin by UHRF1 protein. J Biol Chem *286*, 24300-24311.

Nagano, T., Mitchell, J.A., Sanz, L.A., Pauler, F.M., Ferguson-Smith, A.C., Feil, R., and Fraser, P. (2008). The Air Noncoding RNA Epigenetically Silences Transcription by Targeting G9a to Chromatin. Science *322*, 1717-1720.

Nakamura, T., Liu, Y.J., Nakashima, H., Umehara, H., Inoue, K., Matoba, S., Tachibana, M., Ogura, A., Shinkai, Y., and Nakano, T. (2012). PGC7 binds histone H3K9me2 to protect against conversion of 5mC to 5hmC in early embryos. Nature *486*, 415-419.

Nakata, S., Sugio, K., Uramoto, H., Oyama, T., Hanagiri, T., Morita, M., and Yasumoto, K. (2006). The methylation status and protein expression of CDH1, p16(INK4A), and fragile histidine triad in nonsmall cell lung carcinoma: epigenetic silencing, clinical features, and prognostic significance. Cancer *106*, 2190-2199.

Nan, X., Campoy, F.J., and Bird, A. (1997). MeCP2 is a transcriptional repressor with abundant binding sites in genomic chromatin. Cell *88*, 471-481.

Nan, X., Hou, J., Maclean, A., Nasir, J., Lafuente, M.J., Shu, X., Kriaucionis, S., and Bird, A. (2007). Interaction between chromatin proteins MECP2 and ATRX is disrupted by mutations that cause inherited mental retardation. Proc Natl Acad Sci U S A *104*, 2709-2714.

Nan, X., Meehan, R.R., and Bird, A. (1993). Dissection of the methyl-CpG binding domain from the chromosomal protein MeCP2. Nucleic Acids Res *21*, 4886-4892.

Nan, X., Ng, H.H., Johnson, C.A., Laherty, C.D., Turner, B.M., Eisenman, R.N., and Bird, A. (1998). Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. Nature *393*, 386-389.

Natoli, G., and Andrau, J.C. (2012). Noncoding transcription at enhancers: general principles and functional models. Annu Rev Genet *46*, 1-19.

Neave, B., Holder, N., and Patient, R. (1997). A graded response to BMP-4 spatially coordinates patterning of the mesoderm and ectoderm in the zebrafish. Mechanisms of development *62*, 183-195.

Newport, J., and Kirschner, M. (1982). A major developmental transition in early xenopus embryos: I. characterization and timing of cellular changes at the midblastula stage. Cell *30*, 675-686.

Ng, H.-H., and Adrian, B. (1999). DNA methylation and chromatin modification. Current opinion in genetics & development *9*, 158-163.

Nicol, J.W., Helt, G.A., Blanchard, S.G., Jr., Raja, A., and Loraine, A.E. (2009). The Integrated Genome Browser: free software for distribution and exploration of genome-scale datasets. Bioinformatics *25*, 2730-2731.

Nishiyama, A., Yamaguchi, L., Sharif, J., Johmura, Y., Kawamura, T., Nakanishi, K., Shimamura, S., Arita, K., Kodama, T., Ishikawa, F., Koseki, H., and Nakanishi, M. (2013). Uhrf1-dependent H3K23 ubiquitylation couples maintenance DNA methylation and replication. Nature.

O'Doherty, A., Ruf, S., Mulligan, C., Hildreth, V., Errington, M.L., Cooke, S., Sesay, A., Modino, S., Vanes, L., Hernandez, D., Linehan, J.M., Sharpe, P.T., Brandner, S., Bliss, T.V.P., Henderson, D.J., Nizetic, D., Tybulewicz, V.L.J., and Fisher, E.M.C. (2005). An Aneuploid Mouse Strain Carrying Human Chromosome 21 with Down Syndrome Phenotypes. Science *309*, 2033-2037.

Oakeley, E.J. (1999). DNA methylation analysis: a review of current methodologies. Pharmacology & Therapeutics *84*, 389-400.

Odom, D.T., Dowell, R.D., Jacobsen, E.S., Nekludova, L., Rolfe, P.A., Danford, T.W., Gifford, D.K., Fraenkel, E., Bell, G.I., and Young, R.A. (2006). Core transcriptional regulatory circuitry in human hepatocytes. Mol Syst Biol *2*, 2006 0017.

Ohki, I., Shimotake, N., Fujita, N., Jee, J., Ikegami, T., Nakao, M., and Shirakawa, M. (2001). Solution structure of the methyl-CpG binding domain of human MBD1 in complex with methylated DNA. Cell *105*, 487-497.

Okamoto, I., Otte, A.P., Allis, C.D., Reinberg, D., and Heard, E. (2004). Epigenetic dynamics of imprinted X inactivation during early mouse development. Science *303*, 644-649.

Okamura, K., Matsumoto, K.A., and Nakai, K. (2010). Gradual transition from mosaic to global DNA methylation patterns during deuterostome evolution. BMC Bioinformatics *11 Suppl 7*, S2.

Okano, M., Bell, D.W., Haber, D.A., and Li, E. (1999). DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell *99*, 247-257.

Okano, M., and Li, E. (2002). Genetic Analyses of DNA Methyltransferase Genes in Mouse Model System. The Journal of Nutrition *132*, 2462S-2465S.

Okano, M., Xie, S., and Li, E. (1998a). Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases. Nature genetics *19*, 219-220.

Okano, M., Xie, S., and Li, E. (1998b). Dnmt2 is not required for de novo and maintenance methylation of viral DNA in embryonic stem cells. Nucleic Acids Res *26*, 2536-2540.

Okitsu, C.Y., and Hsieh, C.L. (2007). DNA methylation dictates histone H3K4 methylation. Mol Cell Biol *27*, 2746-2757.

Ooi, S.K., Qiu, C., Bernstein, E., Li, K., Jia, D., Yang, Z., Erdjument-Bromage, H., Tempst, P., Lin, S.P., Allis, C.D., Cheng, X., and Bestor, T.H. (2007). DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. Nature *448*, 714-717.

Orlando, D.A., Guenther, M.G., Frampton, G.M., and Young, R.A. (2012). CpG island structure and trithorax/polycomb chromatin domains in human cells. Genomics.

Oswald, J., Engemann, S., Lane, N., Mayer, W., Olek, A., Fundele, R., Dean, W., Reik, W., and Walter, J. (2000). Active demethylation of the paternal genome in the mouse zygote. Current biology : CB *10*, 475-478.

Ota, K.G., and Kuratani, S. (2007). Cyclostome embryology and early evolutionary history of vertebrates. Integrative and Comparative Biology *47*, 329-337.

Palmer, L.E., Rabinowicz, P.D., O'Shaughnessy, A.L., Balija, V.S., Nascimento, L.U., Dike, S., de la Bastide, M., Martienssen, R.A., and McCombie, W.R. (2003). Maize genome sequencing by methylation filtration. Science *302*, 2115-2117.

Panning, B., and Jaenisch, R. (1996). DNA hypomethylation can activate Xist expression and silence X-linked genes. Genes Dev *10*, 1991-2002.

Park, S., Osmers, U., Raman, G., Schwantes, R.H., Diaz, M.O., and Bushweller, J.H. (2010). The PHD3 domain of MLL acts as a CYP33-regulated switch between MLL-mediated activation and repression. Biochemistry *49*, 6576-6586.

Pasini, D., Malatesta, M., Jung, H.R., Walfridsson, J., Willer, A., Olsson, L., Skotte, J., Wutz, A., Porse, B., Jensen, O.N., and Helin, K. (2010). Characterization of an antagonistic switch between histone H3 lysine 27 methylation and acetylation in the transcriptional regulation of Polycomb group target genes. Nucleic acids research *38*, 4958-4969.

Pastor, W.A., Pape, U.J., Huang, Y., Henderson, H.R., Lister, R., Ko, M., McLoughlin, E.M., Brudno, Y., Mahapatra, S., Kapranov, P., Tahiliani, M., Daley, G.Q., Liu, X.S., Ecker, J.R., Milos, P.M., Agarwal, S., and Rao, A. Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells. Nature *473*, 394-397.

Pauli, A., Valen, E., Lin, M.F., Garber, M., Vastenhouw, N.L., Levin, J.Z., Fan, L., Sandelin, A., Rinn, J.L., Regev, A., and Schier, A.F. (2012). Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. Genome Res *22*, 577-591.

Paulsen, M., and Ferguson-Smith, A.C. (2001). DNA methylation in genomic imprinting, development, and disease. The Journal of Pathology *195*, 97-110.

Pelizzola, M., Koga, Y., Urban, A.E., Krauthammer, M., Weissman, S., Halaban, R., and Molinaro, A.M. (2008). MEDME: An experimental and analytical methodology for the estimation of DNA methylation levels based on microarray derived MeDIP-enrichment. Genome Research *18*, 1652-1659.

Peters, A.H., Kubicek, S., Mechtler, K., O'Sullivan, R.J., Derijck, A.A., Perez-Burgos, L., Kohlmaier, A., Opravil, S., Tachibana, M., Shinkai, Y., Martens, J.H., and Jenuwein, T. (2003). Partitioning and plasticity of repressive histone methylation states in mammalian chromatin. Mol Cell *12*, 1577-1589.

Peterson, C.L., and Laniel, M.A. (2004). Histones and histone modifications. Current biology : CB *14*, R546-551.

Pfau, R., Tzatsos, A., Kampranis, S.C., Serebrennikova, O.B., Bear, S.E., and Tsichlis, P.N. (2008). Members of a family of JmjC domain-containing oncoproteins immortalize embryonic fibroblasts via a JmjC domain-dependent process. Proc Natl Acad Sci U S A *105*, 1907-1912.

Phillips, B.T., Kwon, H.-J., Melton, C., Houghtaling, P., Fritz, A., and Riley, B.B. (2006). Zebrafish msxB, msxC and msxE function together to refine the neural–nonneural border and regulate cranial placodes and neural crest development. Developmental Biology *294*, 376-390.

Phillips, J.E., and Corces, V.G. (2009). CTCF: Master Weaver of the Genome. Cell *137*, 1194-1211.

Plath, K., Fang, J., Mlynarczyk-Evans, S.K., Cao, R., Worringer, K.A., Wang, H., de la Cruz, C.C., Otte, A.P., Panning, B., and Zhang, Y. (2003). Role of histone H3 lysine 27 methylation in X inactivation. Science *300*, 131-135.

Ponjavic, J., Ponting, C.P., and Lunter, G. (2007). Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. Genome Res *17*, 556-565.

Ponting, C.P. (2008). The functional repertoires of metazoan genomes. Nat Rev Genet *9*, 689-698.

Popp, C., Dean, W., Feng, S.H., Cokus, S.J., Andrews, S., Pellegrini, M., Jacobsen, S.E., and Reik, W. (2010). Genome-wide erasure of DNA methylation in mouse primordial germ cells is affected by AID deficiency. Nature *463*, 1101-U1126.

Potok, M.E., Nix, D.A., Parnell, T.J., and Cairns, B.R. (2013). Reprogramming the maternal zebrafish genome after fertilization to match the paternal methylation pattern. Cell *153*, 759-772.

Proffitt, J.H., Davie, J.R., Swinton, D., and Hattman, S. (1984). 5-Methylcytosine is not detectable in Saccharomyces cerevisiae DNA. Mol Cell Biol *4*, 985-988.

Prokhortchouk, A., Hendrich, B., Jorgensen, H., Ruzov, A., Wilm, M., Georgiev, G., Bird, A., and Prokhortchouk, E. (2001). The p120 catenin partner Kaiso is a DNA methylation-dependent transcriptional repressor. Genes Dev *15*, 1613-1618.

Prokhortchouk, A., Sansom, O., Selfridge, J., Caballero, I.M., Salozhin, S., Aithozhina, D., Cerchietti, L., Meng, F.G., Augenlicht, L.H., Mariadason, J.M., Hendrich, B., Melnick, A., Prokhortchouk, E., Clarke, A., and Bird, A. (2006). Kaiso-deficient mice show resistance to intestinal cancer. Mol Cell Biol *26*, 199-208.

Ptashne, M. (1986). Gene regulation by proteins acting nearby and at a distance. Nature *322*, 697-701.

Puschendorf, M., Terranova, R., Boutsma, E., Mao, X., Isono, K., Brykczynska, U., Kolb, C., Otte, A.P., Koseki, H., Orkin, S.H., van Lohuizen, M., and Peters, A.H. (2008). PRC1 and Suv39h specify parental asymmetry at constitutive heterochromatin in early mouse embryos. Nature genetics *40*, 411-420.

Qi, H.H., Sarkissian, M., Hu, G.Q., Wang, Z., Bhattacharjee, A., Gordon, D.B., Gonzales, M., Lan, F., Ongusaha, P.P., Huarte, M., Yaghi, N.K., Lim, H., Garcia, B.A., Brizuela, L., Zhao, K., Roberts, T.M., and Shi, Y. (2010). Histone H4K20/H3K9 demethylase PHF8 regulates zebrafish brain and craniofacial development. Nature *466*, 503-507.

Qian, C., Li, S., Jakoncic, J., Zeng, L., Walsh, M.J., and Zhou, M.-M. (2008). Structure and Hemimethylated CpG Binding of the SRA Domain from Human UHRF1. Journal of Biological Chemistry *283*, 34490-34494.

Quenneville, S., Verde, G., Corsinotti, A., Kapopoulou, A., Jakobsson, J., Offner, S., Baglivo, I., Pedone, P.V., Grimaldi, G., Riccio, A., and Trono, D. (2011). In embryonic stem cells, ZFP57/KAP1 recognize a methylated hexanucleotide to affect chromatin and DNA methylation of imprinting control regions. Mol Cell *44*, 361-372.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841-842.

Raabe, E.H., Abdurrahman, L., Behbehani, G., and Arceci, R.J. (2001). An SNF2 factor involved in mammalian development and cellular proliferation. Developmental dynamics : an official publication of the American Association of Anatomists *221*, 92-105.

Raddatz, G., Guzzardo, P.M., Olova, N., Fantappie, M.R., Rampp, M., Schaefer, M., Reik, W., Hannon, G.J., and Lyko, F. (2013). Dnmt2-dependent methylomes lack defined DNA methylation patterns. Proc Natl Acad Sci U S A *110*, 8627-8631.

Rai, K., Chidester, S., Zavala, C.V., Manos, E.J., James, S.R., Karpf, A.R., Jones, D.A., and Cairns, B.R. (2007). Dnmt2 functions in the cytoplasm to promote liver, brain, and retina development in zebrafish. Genes Dev *21*, 261-266.

Rai, K., Huggins, I.J., James, S.R., Karpf, A.R., Jones, D.A., and Cairns, B.R. (2008). DNA Demethylation in Zebrafish Involves the Coupling of a Deaminase, a Glycosylase, and Gadd45. Cell *135*, 1201-1212.

Rai, K., Jafri, I.F., Chidester, S., James, S.R., Karpf, A.R., Cairns, B.R., and Jones, D.A. (2010). Dnmt3 and G9a Cooperate for Tissue-specific Development in Zebrafish. Journal of Biological Chemistry *285*, 4110-4121.

Rai, K., Nadauld, L.D., Chidester, S., Manos, E.J., James, S.R., Karpf, A.R., Cairns, B.R., and Jones, D.A. (2006). Zebra Fish Dnmt1 and Suv39h1 Regulate Organ-Specific Terminal Differentiation during Development. Molecular and cellular biology *26*, 7077-7085.

Ramirez-Carrozzi, V.R., Braas, D., Bhatt, D.M., Cheng, C.S., Hong, C., Doty, K.R., Black, J.C., Hoffmann, A., Carey, M., and Smale, S.T. (2009). A Unifying Model for the Selective Regulation of Inducible Transcription by CpG Islands and Nucleosome Remodeling. Cell *138*, 114-128.

Ratnam, S., Mertineit, C., Ding, F., Howell, C.Y., Clarke, H.J., Bestor, T.H., Chaillet, J.R., and Trasler, J.M. (2002). Dynamics of Dnmt1 Methyltransferase Expression and Intracellular Localization during Oogenesis and Preimplantation Development. Developmental Biology *245*, 304-314.

Razin, A. (1998). CpG methylation, chromatin structure and gene silencing[mdash]a three-way connection. EMBO J *17*, 4905-4908.

Rea, S., Eisenhaber, F., O'Carroll, D., Strahl, B.D., Sun, Z.W., Schmid, M., Opravil, S., Mechtler, K., Ponting, C.P., Allis, C.D., and Jenuwein, T. (2000). Regulation of chromatin structure by site-specific histone H3 methyltransferases. Nature *406*, 593-599.

Read, E.M., Rodaway, A.R., Neave, B., Brandon, N., Holder, N., Patient, R.K., and Walmsley, M.E. (1998). Evidence for non-axial A/P patterning in the nonneural ectoderm of Xenopus and zebrafish pregastrula embryos. The International journal of developmental biology *42*, 763-774.

Reik, W. (2007). Stability and flexibility of epigenetic gene regulation in mammalian development. Nature *447*, 425-432.

Reik, W., Dean, W., and Walter, J. (2001). Epigenetic reprogramming in mammalian development. Science *293*, 1089-1093.

Reik, W., and Murrell, A. (2000). Genomic imprinting: Silence across the border. Nature *405*, 408-409.

Rhee, I., Jair, K.-W., Yen, R.-W.C., Lengauer, C., Herman, J.G., Kinzler, K.W., Vogelstein, B., Baylin, S.B., and Schuebel, K.E. (2000). CpG methylation is maintained in human cancer cells lacking DNMT1. Nature *404*, 1003-1007.

Riddihough, G., and Zahn, L.M. (2010). Epigenetics. What is epigenetics? Introduction. Science *330*, 611.

Riggs, A.D. (1975). X inactivation, differentiation, and DNA methylation. Cytogenetics and cell genetics *14*, 9-25.

Robin, P., Fritsch, L., Philipot, O., Svinarchuk, F., and Ait-Si-Ali, S. (2007). Post-translational modifications of histones H3 and H4 associated with the histone methyltransferases Suv39h1 and G9a. Genome Biol *8*, R270.

Rohde, L.A., and Heisenberg, C.P. (2007). Zebrafish gastrulation: Cell movements, signals, and mechanisms. Int Rev Cytol *261*, 159-+.

Rosenfeld, J.A., Wang, Z., Schones, D.E., Zhao, K., DeSalle, R., and Zhang, M.Q. (2009). Determination of enriched histone modifications in non-genic portions of the human genome. BMC Genomics *10*, 143.

Ross, J.P., Suetake, I., Tajima, S., and Molloy, P.L. (2010). Recombinant mammalian DNA methyltransferase activity on model transcriptional gene silencing short RNA-DNA heteroduplex substrates. The Biochemical journal *432*, 323-332.

Rothbart, S.B., Krajewski, K., Nady, N., Tempel, W., Xue, S., Badeaux, A.I., Barsyte-Lovejoy, D., Martinez, J.Y., Bedford, M.T., Fuchs, S.M., Arrowsmith, C.H., and Strahl, B.D. (2012). Association of UHRF1 with methylated H3K9 directs the maintenance of DNA methylation. Nat Struct Mol Biol *19*, 1155-1160.

Russo, V.E.A., Martienssen, R.A., and Riggs, A.D. (1996). Epigenetic Mechanisms of Gene Regulation. Cold Spring Harbor Laboratory Press, Woodbury.

Ruthenburg, A.J., Allis, C.D., and Wysocka, J. (2007). Methylation of lysine 4 on histone H3: intricacy of writing and reading a single epigenetic mark. Mol Cell *25*, 15-30.

Ruzov, A., Dunican, D.S., Prokhortchouk, A., Pennings, S., Stancheva, I., Prokhortchouk, E., and Meehan, R.R. (2004). Kaiso is a genome-wide repressor of transcription that is essential for amphibian development. Development *131*, 6185-6194.

Ryazanova, A.Y., Abrosimova, L.A., Oretskaya, T.S., and Kubareva, E.A. (2012). Diverse Domains of (Cytosine-5)-DNA Methyltransferases: Structural and Functional Characterization (InTech).

Sado, T., Okano, M., Li, E., and Sasaki, H. (2004). De novo DNA methylation is dispensable for the initiation and propagation of X chromosome inactivation. Development *131*, 975-982.

Sakaue, M., Ohta, H., Kumaki, Y., Oda, M., Sakaide, Y., Matsuoka, C., Yamagiwa, A., Niwa, H., Wakayama, T., and Okano, M. (2010). DNA Methylation Is Dispensable for the Growth and Survival of the Extraembryonic Lineages. Current Biology *20*, 1452-1457.

Sanchez, C., Sanchez, I., Demmers, J.A., Rodriguez, P., Strouboulis, J., and Vidal, M. (2007). Proteomics analysis of Ring1B/Rnf2 interactors identifies a novel complex with the Fbxl10/Jhdm1B histone demethylase and the Bcl6 interacting corepressor. Mol Cell Proteomics *6*, 820-834.

SanMiguel, P., Tikhonov, A., Jin, Y.K., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., Springer, P.S., Edwards, K.J., Lee, M., Avramova, Z., and Bennetzen, J.L. (1996). Nested retrotransposons in the intergenic regions of the maize genome. Science *274*, 765-768.

Santini, F., Harmon, L., Carnevale, G., and Alfaro, M. (2009). Did genome duplication drive the origin of teleosts? A comparative study of diversification in ray-finned fishes. BMC Evolutionary Biology *9*, 194.

Santos, F., Hendrich, B., Reik, W., and Dean, W. (2002). Dynamic reprogramming of DNA methylation in the early mouse embryo. Dev Biol *241*, 172-182.

Sarraf, S.A., and Stancheva, I. (2004). Methyl-CpG binding protein MBD1 couples histone H3 methylation at lysine 9 by SETDB1 to DNA replication and chromatin assembly. Mol Cell *15*, 595-605.

Sasai, N., Nakao, M., and Defossez, P.A. (2010). Sequence-specific recognition of methylated DNA by human zinc-finger proteins. Nucleic Acids Res *38*, 5015-5022.

Saxonov, S., Berg, P., and Brutlag, D.L. (2006). A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc Natl Acad Sci U S A *103*, 1412-1417.

Scacheri, P.C., Davis, S., Odom, D.T., Crawford, G.E., Perkins, S., Halawi, M.J., Agarwal, S.K., Marx, S.J., Spiegel, A.M., Meltzer, P.S., and Collins, F.S. (2006). Genome-Wide Analysis of Menin Binding Provides Insights into MEN1 Tumorigenesis. PLoS Genet *2*, e51.

Schatz, P.J. (1993). Use of peptide libraries to map the substrate specificity of a peptide-modifying enzyme: a 13 residue consensus peptide specifies biotinylation in Escherichia coli. Bio/technology (Nature Publishing Company) *11*, 1138-1143.

Schermelleh, L., Haemmer, A., Spada, F., Rosing, N., Meilinger, D., Rothbauer, U., Cardoso, M.C., and Leonhardt, H. (2007). Dynamics of Dnmt1 interaction with the replication machinery and its role in postreplicative maintenance of DNA methylation. Nucleic Acids Res *35*, 4301-4312.

Schier, A.F., and Talbot, W.S. (2005). Molecular genetics of axis formation in zebrafish. Annual Review of Genetics *39*, 561-613.

Schlesinger, Y., Straussman, R., Keshet, I., Farkash, S., Hecht, M., Zimmerman, J., Eden, E., Yakhini, Z., Ben-Shushan, E., Reubinoff, B.E., Bergman, Y., Simon, I., and Cedar, H. (2007). Polycomb-mediated methylation on Lys27 of histone H3 pre-marks genes for de novo methylation in cancer. Nature genetics *39*, 232-236.

Schmidl, C., Klug, M., Boeld, T.J., Andreesen, R., Hoffmann, P., Edinger, M., and Rehli, M. (2009). Lineage-specific DNA methylation in T cells correlates with histone methylation and enhancer activity. Genome Res *19*, 1165-1174.

Schmidt, D., Wilson, M.D., Ballester, B., Schwalie, P.C., Brown, G.D., Marshall, A., Kutter, C., Watt, S., Martinez-Jimenez, C.P., Mackay, S., Talianidis, I., Flicek, P., and Odom, D.T. (2010). Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. Science *328*, 1036-1040.

Schmitges, Frank W., Prusty, Archana B., Faty, M., Stützer, A., Lingaraju, Gondichatnahalli M., Aiwazian, J., Sack, R., Hess, D., Li, L., Zhou, S., Bunker, Richard D., Wirth, U., Bouwmeester, T., Bauer, A., Ly-Hartig, N., Zhao, K., Chan, H., Gu, J., Gut, H., Fischle, W., Müller, J., and Thomä, Nicolas H. (2011). Histone Methylation by PRC2 Is Inhibited by Active Chromatin Marks. Molecular cell *42*, 330-341.

Schmittgen, T.D., and Livak, K.J. (2008). Analyzing real-time PCR data by the comparative C(T) method. Nature protocols *3*, 1101-1108.

Schneider, R., and Grosschedl, R. (2007). Dynamics and interplay of nuclear architecture, genome organization, and gene expression. Genes & development *21*, 3027-3043.

Schoeftner, S., Sengupta, A.K., Kubicek, S., Mechtler, K., Spahn, L., Koseki, H., Jenuwein, T., and Wutz, A. (2006). Recruitment of PRC1 function at the initiation of X inactivation independent of PRC2 and silencing. EMBO J *25*, 3110-3122.

Schulte-Merker, S., Ho, R.K., Herrmann, B.G., and Nusslein-Volhard, C. (1992). The protein product of the zebrafish homologue of the mouse T gene is expressed in nuclei of the germ ring and the notochord of the early embryo. Development *116*, 1021-1032.

Schulte-Merker, S., and Smith, J.C. (1995). Mesoderm formation in response to Brachyury requires FGF signalling. Current biology : CB *5*, 62-67.

Schultz, D.C., Ayyanathan, K., Negorev, D., Maul, G.G., and Rauscher, F.J. (2002). SETDB1: a novel KAP-1-associated histone H3, lysine 9-specific methyltransferase that contributes to HP1-mediated silencing of euchromatic genes by KRAB zinc-finger proteins. Genes & development *16*, 919-932.

Schwartz, Y.B., and Pirrotta, V. (2007). Polycomb silencing mechanisms and the management of genomic programmes. Nat Rev Genet *8*, 9-22.

Seebald, J.L., and Szeto, D.P. (2011). Zebrafish eve1 regulates the lateral and ventral fates of mesodermal progenitor cells at the onset of gastrulation. Developmental Biology *349*, 78-89.

Selker, E.U., Tountas, N.A., Cross, S.H., Margolin, B.S., Murphy, J.G., Bird, A.P., and Freitag, M. (2003). The methylated component of the Neurospora crassa genome. Nature *422*, 893-897.

Sérandour, A.A., Avner, S., Percevault, F., Demay, F., Bizot, M., Lucchetti-Miganeh, C., Barloy-Hubler, F., Brown, M., Lupien, M., Métivier, R., Salbert, G., and Eeckhoute, J. (2011). Epigenetic switch involved in activation of pioneer factor FOXA1-dependent enhancers. Genome Research *21*, 555-565.

Serre, D., Lee, B.H., and Ting, A.H. (2010). MBD-isolated Genome Sequencing provides a high-throughput and comprehensive survey of DNA methylation in the human genome. Nucleic acids research *38*, 391-399.

Sharif, J., Endo, T.a., Toyoda, T., and Koseki, H. (2010). Divergence of CpG island promoters: a consequence or cause of evolution? Development, growth & differentiation *52*, 545-554.

Sharif, J., Muto, M., Takebayashi, S.-i., Suetake, I., Iwamatsu, A., Endo, T.A., Shinga, J., Mizutani-Koseki, Y., Toyoda, T., Okamura, K., Tajima, S., Mitsuya, K., Okano, M., and Koseki, H. (2007). The SRA protein Np95 mediates epigenetic inheritance by recruiting Dnmt1 to methylated DNA. Nature *450*, 908-912.

Shay, J.W., Reddel, R.R., and Wright, W.E. (2012). Cancer and Telomeres-An ALTernative to Telomerase. Science *336*, 1388-1390.

Shen, J.C., Rideout, W.M., 3rd, and Jones, P.A. (1994). The rate of hydrolytic deamination of 5-methylcytosine in double-stranded DNA. Nucleic Acids Res *22*, 972-976.

Shi, Y., Lan, F., Matson, C., Mulligan, P., Whetstine, J.R., Cole, P.A., Casero, R.A., and Shi, Y. (2004). Histone demethylation mediated by the nuclear amine oxidase homolog LSD1. Cell *119*, 941-953.

Shimizu, T., Bae, Y.K., Muraoka, O., and Hibi, M. (2005). Interaction of Wnt and caudal-related genes in zebrafish posterior body formation. Dev Biol *279*, 125-141.

Shin, H., Liu, T., Manrai, A.K., and Liu, X.S. (2009). CEAS: cis-regulatory element annotation system. Bioinformatics *25*, 2605-2606.

Shirane, K., Toh, H., Kobayashi, H., Miura, F., Chiba, H., Ito, T., Kono, T., and Sasaki, H. (2013). Mouse Oocyte Methylomes at Base Resolution Reveal Genome-Wide Accumulation of Non-CpG Methylation and Role of DNA Methyltransferases. PLoS Genet *9*, e1003439.

Silverstein, R.A., and Ekwall, K. (2005). Sin3: a flexible regulator of global gene expression and genome stability. Current genetics *47*, 1-17.

Simmen, M.W., Leitgeb, S., Charlton, J., Jones, S.J., Harris, B.R., Clark, V.H., and Bird, A. (1999). Nonmethylated transposable elements and methylated genes in a chordate genome. Science *283*, 1164-1167.

Simon, J.A., and Kingston, R.E. (2009). Mechanisms of polycomb gene silencing: knowns and unknowns. Nat Rev Mol Cell Biol *10*, 697-708.

Simon, J.A., and Kingston, R.E. (2013). Occupying chromatin: Polycomb mechanisms for getting to genomic targets, stopping transcriptional traffic, and staying put. Mol Cell *49*, 808-824.

Singer, M.J., Marcotte, B.A., and Selker, E.U. (1995). DNA methylation associated with repeat-induced point mutation in Neurospora crassa. Mol Cell Biol *15*, 5586-5597.

Singh, P., Li, Arthur X., Tran, Diana A., Oates, N., Kang, E.-R., Wu, X., and Szabó, Piroska E. (2013). De Novo DNA Methylation in the Male Germ Line Occurs by Default but Is Excluded at Sites of H3K4 Methylation. Cell Reports *4*, 205-219.

Smagulova, F., Gregoretti, I.V., Brick, K., Khil, P., Camerini-Otero, R.D., and Petukhova, G.V. (2011). Genome-wide analysis reveals novel molecular features of mouse recombination hotspots. Nature *472*, 375-378.

Smallwood, S.A., and Kelsey, G. (2012). De novo DNA methylation: a germ cell perspective. Trends in Genetics *28*, 33-42.

Smallwood, S.A., Tomizawa, S.-i., Krueger, F., Ruf, N., Carli, N., Segonds-Pichon, A., Sato, S., Hata, K., Andrews, S.R., and Kelsey, G. (2011). Dynamic CpG island methylation landscape in oocytes and preimplantation embryos. Nature genetics *43*, 811-814.

Smith, D.J., Stevens, M.E., Sudanagunta, S.P., Bronson, R.T., Makhinson, M., Watabe, A.M., O'Dell, T.J., Fung, J., Weier, H.-U.G., Cheng, J.-F., and Rubin, E.M. (1997). Functional screening of 2 Mb of human chromosome 21q22.2 in transgenic mice implicates minibrain in learning defects associated with Down syndrome. Nature genetics *16*, 28-36.

Smith, S.S., Kaplan, B.E., Sowers, L.C., and Newman, E.M. (1992). Mechanism of human methyl-directed DNA methyltransferase and the fidelity of cytosine methylation. Proceedings of the National Academy of Sciences *89*, 4744-4748.

Smith, T.L., Collins, T., and McGowan, R. (2011). Expression of the dnmt3 genes in zebrafish development: similarity to Dnmt3a and Dnmt3b. Development genes and evolution *220*, 347-353.

Smith, Z.D., Chan, M.M., Mikkelsen, T.S., Gu, H., Gnirke, A., Regev, A., and Meissner, A. (2012). A unique regulatory phase of DNA methylation in the early mammalian embryo. Nature *484*, 339-344.

Smith, Z.D., and Meissner, A. (2013). DNA methylation: roles in mammalian development. Nat Rev Genet *14*, 204-220.

Solnica-Krezel, L., and Driever, W. (1994). Microtubule arrays of the zebrafish yolk cell: organization and function during epiboly. Development *120*, 2443-2455.

Song, C.X., Yi, C., and He, C. (2012a). Mapping recently identified nucleotide variants in the genome and transcriptome. Nat Biotechnol *30*, 1107-1116.

Song, F., Smith, J.F., Kimura, M.T., Morrow, A.D., Matsuyama, T., Nagase, H., and Held, W.A. (2005). Association of tissue-specific differentially methylated regions (TDMs) with differential gene expression. Proc Natl Acad Sci U S A *102*, 3336-3341.

Song, J., Rechkoblit, O., Bestor, T.H., and Patel, D.J. (2011). Structure of DNMT1-DNA complex reveals a role for autoinhibition in maintenance DNA methylation. Science *331*, 1036-1040.

Song, J., Teplova, M., Ishibe-Murakami, S., and Patel, D.J. (2012b). Structure-based mechanistic insights into DNMT1-mediated maintenance DNA methylation. Science *335*, 709-712.

Spada, F., Haemmer, A., Kuch, D., Rothbauer, U., Schermelleh, L., Kremmer, E., Carell, T., Langst, G., and Leonhardt, H. (2007). DNMT1 but not its interaction with the replication machinery is required for maintenance of DNA methylation in human cells. J Cell Biol *176*, 565-571.

Stadler, M.B., Murr, R., Burger, L., Ivanek, R., Lienert, F., Schöler, A., Wirbelauer, C., Oakeley, E.J., Gaidatzis, D., Tiwari, V.K., and Schübeler, D. (2011). DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature *480*, 490-495.

Stancheva, I., El-Maarri, O., Walter, J., Niveleau, A., and Meehan, R.R. (2002). DNA methylation at promoter regions regulates the timing of gene activation in Xenopus laevis embryos. Dev Biol *243*, 155-165.

Stancheva, I., and Meehan, R.R. (2000). Transient depletion of xDnmt1 leads to premature gene activation in Xenopus embryos. Genes Dev *14*, 313-327.

Steensma, D.P., Higgs, D.R., Fisher, C.A., and Gibbons, R.J. (2004). Acquired somatic ATRX mutations in myelodysplastic syndrome associated with alpha thalassemia (ATMDS) convey a more severe hematologic phenotype than germline ATRX mutations. Blood *103*, 2019-2026.

Strahl, B.D., and Allis, C.D. (2000). The language of covalent histone modifications. Nature *403*, 41-45.

Strahle, U., and Jesuthasan, S. (1993). Ultraviolet irradiation impairs epiboly in zebrafish embryos: evidence for a microtubule-dependent mechanism of epiboly. Development *119*, 909-919.

Straussman, R., Nejman, D., Roberts, D., Steinfeld, I., Blum, B., Benvenisty, N., Simon, I., Yakhini, Z., and Cedar, H. (2009). Developmental programming of CpG island methylation profiles in the human genome. Nat Struct Mol Biol *16*, 564-571.

Strunnikova, M., Schagdarsurengin, U., Kehlen, A., Garbe, J.C., Stampfer, M.R., and Dammann, R. (2005). Chromatin inactivation precedes de novo DNA methylation during the progressive epigenetic silencing of the RASSF1A promoter. Mol Cell Biol *25*, 3923-3933.

Sun, X.J., Wei, J., Wu, X.Y., Hu, M., Wang, L., Wang, H.H., Zhang, Q.H., Chen, S.J., Huang, Q.H., and Chen, Z. (2005). Identification and characterization of a novel human histone H3 lysine 36-specific methyltransferase. J Biol Chem *280*, 35261-35271.

Supek, F., Bosnjak, M., Skunca, N., and Smuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. PloS one *6*, e21800.

Suster, M.L., Sumiyama, K., and Kawakami, K. (2009). Transposon-mediated BAC transgenesis in zebrafish and mice. BMC genomics *10*, 477.

Suzuki, M.M., and Bird, A. (2008). DNA methylation landscapes: provocative insights from epigenomics. Nat Rev Genet *9*, 465-476.

Suzuki, M.M., Kerr, A.R.W., De Sousa, D., and Bird, A. (2007). CpG methylation is targeted to transcription units in an invertebrate genome. Genome Research *17*, 625-631.

Sved, J., and Bird, A. (1990). The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. Proc Natl Acad Sci U S A *87*, 4692-4696.

Swartz, M.N., Trautner, T.A., and Kornberg, A. (1962). Enzymatic synthesis of deoxyribonucleic acid. XI. Further studies on nearest neighbor base sequences in deoxyribonucleic acids. J Biol Chem *237*, 1961-1967.

Syeda, F., Fagan, R.L., Wean, M., Avvakumov, G.V., Walker, J.R., Xue, S., Dhe-Paganon, S., and Brenner, C. (2011). The replication focus targeting sequence (RFTS) domain is a DNA-competitive inhibitor of Dnmt1. J Biol Chem *286*, 15344-15351.

Szeto, D.P., and Kimelman, D. (2004). Combinatorial gene regulation by Bmp and Wnt in zebrafish posterior mesoderm formation. Development *131*, 3751-3760.

Tachibana, M., Matsumura, Y., Fukuda, M., Kimura, H., and Shinkai, Y. (2008). G9a/GLP complexes independently mediate H3K9 and DNA methylation to silence transcription. EMBO J *27*, 2681-2690.

Tachibana, M., Sugimoto, K., Nozaki, M., Ueda, J., Ohta, T., Ohki, M., Fukuda, M., Takeda, N., Niida, H., Kato, H., and Shinkai, Y. (2002). G9a histone methyltransferase plays a dominant role in euchromatic histone H3 lysine 9 methylation and is essential for early embryogenesis. Genes Dev *16*, 1779-1791.

Tachibana, M., Ueda, J., Fukuda, M., Takeda, N., Ohta, T., Iwanari, H., Sakihama, T., Kodama, T., Hamakubo, T., and Shinkai, Y. (2005). Histone methyltransferases G9a and GLP form heteromeric complexes and are both crucial for methylation of euchromatin at H3-K9. Genes Dev *19*, 815-826.

Tadros, W., and Lipshitz, H.D. (2009). The maternal-to-zygotic transition: a play in two acts. Development *136*, 3033-3042.

Tahiliani, M., Koh, K.P., Shen, Y., Pastor, W.A., Bandukwala, H., Brudno, Y., Agarwal, S., Iyer, L.M., Liu, D.R., Aravind, L., and Rao, A. (2009). Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. Science *324*, 930-935.

Takai, D., and Jones, P.A. (2002). Comprehensive analysis of CpG islands in human chromosomes 21 and 22. Proc Natl Acad Sci U S A *99*, 3740-3745.

Takeshita, K., Suetake, I., Yamashita, E., Suga, M., Narita, H., Nakagawa, A., and Tajima, S. (2011). Structural insight into maintenance methylation by mouse DNA methyltransferase 1 (Dnmt1). Proc Natl Acad Sci U S A *108*, 9055-9059.

Talbert, P.B., and Henikoff, S. (2010). Histone variants--ancient wrap artists of the epigenome. Nat Rev Mol Cell Biol *11*, 264-275.

Talbot, W.S., Trevarrow, B., Halpern, M.E., Melby, A.E., Farr, G., Postlethwait, J.H., Jowett, T., Kimmel, C.B., and Kimelman, D. (1995). A homeobox gene essential for zebrafish notochord development. Nature *378*, 150-157.

Tamaru, H., and Selker, E.U. (2001). A histone H3 methyltransferase controls DNA methylation in Neurospora crassa. Nature *414*, 277-283.

Tamaru, H., Zhang, X., McMillen, D., Singh, P.B., Nakayama, J., Grewal, S.I., Allis, C.D., Cheng, X., and Selker, E.U. (2003). Trimethylated lysine 9 of histone H3 is a mark for DNA methylation in Neurospora crassa. Nature genetics *34*, 75-79.

Tang, J., Wu, S., Liu, H., Stratt, R., Barak, O.G., Shiekhattar, R., Picketts, D.J., and Yang, X. (2004). A novel transcription regulatory complex containing death domain-associated protein and the ATR-X syndrome protein. J Biol Chem *279*, 20369-20377.

Tang, Y., Gao, X.D., Wang, Y., Yuan, B.F., and Feng, Y.Q. (2012). Widespread existence of cytosine methylation in yeast DNA measured by gas chromatography/mass spectrometry. Analytical chemistry *84*, 7249-7255.

Tavares, L., Dimitrova, E., Oxley, D., Webster, J., Poot, R., Demmers, J., Bezstarosti, K., Taylor, S., Ura, H., Koide, H., Wutz, A., Vidal, M., Elderkin, S., and Brockdorff, N. (2012). RYBP-PRC1 Complexes Mediate H2A Ubiquitylation at Polycomb Target Sites Independently of PRC2 and H3K27me3. Cell *148*, 664-678.

Taylor, J.S., Braasch, I., Frickey, T., Meyer, A., and Van de Peer, Y. (2003). Genome Duplication, a Trait Shared by 22,000 Species of Ray-Finned Fish. Genome Research *13*, 382-390.

Taylor, S.M., and Jones, P.A. (1979). Multiple new phenotypes induced in 10T12 and 3T3 cells treated with 5-azacytidine. Cell *17*, 771-779.

Testoni, S., Bartolone, E., Rossi, M., Patrignani, A., Bruggmann, R., Lichtner, P., Tetens, J., Gentile, A., and Drogemuller, C. (2012). KDM2B is implicated in bovine lethal multi-organic developmental dysplasia. PLoS One *7*, e45634.

Thomassin, H., Flavin, M., Espinas, M.L., and Grange, T. (2001). Glucocorticoid-induced DNA demethylation and gene memory during development. EMBO J *20*, 1974-1983.

Thomson, J.P., Skene, P.J., Selfridge, J., Clouaire, T., Guy, J., Webb, S., Kerr, A.R., Deaton, A., Andrews, R., James, K.D., Turner, D.J., Illingworth, R., and Bird, A. (2010). CpG islands influence chromatin structure via the CpG-binding protein Cfp1. Nature *464*, 1082-1086.

Thorpe, C.J., Weidinger, G., and Moon, R.T. (2005). Wnt/β-catenin regulation of the Sp1-related transcription factor sp5l promotes tail development in zebrafish. Development *132*, 1763-1772.

Tittle, R.K., Sze, R., Ng, A., Nuckels, R.J., Swartz, M.E., Anderson, R.M., Bosch, J., Stainier, D.Y.R., Eberhart, J.K., and Gross, J.M. (2011). Uhrf1 and Dnmt1 are required for development and maintenance of the zebrafish lens. Developmental Biology *350*, 50-63.

Tolhuis, B., Palstra, R.-J., Splinter, E., Grosveld, F., and de Laat, W. (2002). Looping and Interaction between Hypersensitive Sites in the Active β-globin Locus. Molecular Cell *10*, 1453-1465.

Tomizuka, K., Shinohara, T., Yoshida, H., Uejima, H., Ohguma, A., Tanaka, S., Sato, K., Oshimura, M., and Ishida, I. (2000). Double trans-chromosomic mice: Maintenance of two individual human chromosome fragments containing Ig heavy and kappa loci and expression of fully human antibodies. Proceedings of the National Academy of Sciences of the United States of America *97*, 722-727.

Tran, R.K., Henikoff, J.G., Zilberman, D., Ditt, R.F., Jacobsen, S.E., and Henikoff, S. (2005). DNA methylation profiling identifies CG methylation clusters in Arabidopsis genes. Current biology : CB *15*, 154-159.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105-1111.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol *28*, 511-515.

Trinkaus, J.P. (1951). A study of the mechanism of epiboly in the egg of Fundulus heteroclitus. Journal of Experimental Zoology *118*, 269-319.

Trinkaus, J.P. (1984). Mechanism of Fundulus Epiboly: A Current View. American Zoologist *24*, 673-688.

Tse, C., Sera, T., Wolffe, A.P., and Hansen, J.C. (1998). Disruption of Higher-Order Folding by Core Histone Acetylation Dramatically Enhances Transcription of Nucleosomal Arrays by RNA Polymerase III. Molecular and cellular biology *18*, 4629-4638.

Tsukada, Y., Fang, J., Erdjument-Bromage, H., Warren, M.E., Borchers, C.H., Tempst, P., and Zhang, Y. (2006). Histone demethylation by a family of JmjC domain-containing proteins. Nature *439*, 811-816.

Tsumura, A., Hayakawa, T., Kumaki, Y., Takebayashi, S., Sakaue, M., Matsuoka, C., Shimotohno, K., Ishikawa, F., Li, E., Ueda, H.R., Nakayama, J., and Okano, M. (2006). Maintenance of self-renewal ability of mouse embryonic stem cells in the absence of DNA methyltransferases Dnmt1, Dnmt3a and Dnmt3b. Genes Cells *11*, 805-814.

Tweedie, S., Charlton, J., Clark, V., and Bird, A. (1997). Methylation of genomes and genes at the invertebrate-vertebrate boundary. Mol Cell Biol *17*, 1469-1475.

Tykocinski, M.L., and Max, E.E. (1984). CG dinucleotide clusters in MHC genes and in 5' demethylated genes. Nucleic Acids Res *12*, 4385-4396.

Ueda, Y., Okano, M., Williams, C., Chen, T., Georgopoulos, K., and Li, E. (2006). Roles for Dnmt3b in mammalian development: a mouse model for the ICF syndrome. Development *133*, 1183-1192.

Ulitsky, I., Shkumatava, A., Jan, C.H., Sive, H., and Bartel, D.P. (2011). Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. Cell *147*, 1537-1550.

Valouev, A., Johnson, S.M., Boyd, S.D., Smith, C.L., Fire, A.Z., and Sidow, A. (2011). Determinants of nucleosome organization in primary human cells. Nature *474*, 516-520.

van der Velden, Y.U., Wang, L., van Lohuizen, M., and Haramis, A.P. (2012). The Polycomb group protein Ring1b is essential for pectoral fin development. Development *139*, 2210-2220.

van Heeringen, S.J., Akhtar, W., Jacobi, U.G., Akkers, R.C., Suzuki, Y., and Veenstra, G.J. (2011). Nucleotide composition-linked divergence of vertebrate core promoter architecture. Genome Res *21*, 410-421.

van Raamsdonk, C.D., and Tilghman, S.M. (2001). Optimizing the detection of nascent transcripts by RNA fluorescence in situ hybridization. Nucleic acids research *29*, e42.

Vastenhouw, N.L., and Schier, A.F. (2012). Bivalent histone modifications in early embryogenesis. Curr Opin Cell Biol *24*, 374-386.

Vastenhouw, N.L., Zhang, Y., Woods, I.G., Imam, F., Regev, A., Liu, X.S., Rinn, J., and Schier, A.F. (2010). Chromatin signature of embryonic pluripotency is established during genome activation. Nature *464*, 922-926.

Veenstra, G.J., and Wolffe, A.P. (2001). Constitutive genomic methylation during embryonic development of Xenopus. Biochimica et biophysica acta *1521*, 39-44.

Velasco, G., Hubé, F., Rollin, J., Neuillet, D., Philippe, C., Bouzinba-Segard, H., Galvani, A., Viegas-Péquignot, E., and Francastel, C. (2010). Dnmt3b recruitment through E2F6 transcriptional repressor mediates germ-line gene silencing in murine somatic tissues. Proceedings of the National Academy of Sciences *107*, 9281-9286.

Venkatesh, S., Smolle, M., Li, H., Gogol, M.M., Saint, M., Kumar, S., Natarajan, K., and Workman, J.L. (2012). Set2 methylation of histone H3 lysine 36 suppresses histone exchange on transcribed genes. Nature *489*, 452-455.

Venolia, L., and Gartler, S.M. (1983). Comparison of transformation efficiency of human active and inactive X-chromosomal DNA. Nature *302*, 82-83.

Vermeulen, M., Mulder, K.W., Denissov, S., Pijnappel, W.W., van Schaik, F.M., Varier, R.A., Baltissen, M.P., Stunnenberg, H.G., Mann, M., and Timmers, H.T. (2007). Selective anchoring of TFIID to nucleosomes by trimethylation of histone H3 lysine 4. Cell *131*, 58-69.

Vernimmen, D., Lynch, M.D., De Gobbi, M., Garrick, D., Sharpe, J.A., Sloane-Stanley, J.A., Smith, A.J., and Higgs, D.R. (2011). Polycomb eviction as a new distant enhancer function. Genes Dev *25*, 1583-1588.

Vernimmen, D., Marques-Kranc, F., Sharpe, J.A., Sloane-Stanley, J.A., Wood, W.G., Wallace, H.A., Smith, A.J., and Higgs, D.R. (2009). Chromosome looping at the human alpha-globin locus is mediated via the major upstream regulatory element (HS -40). Blood *114*, 4253-4260.

Vesterlund, L., Jiao, H., Unneberg, P., Hovatta, O., and Kere, J. (2011). The zebrafish transcriptome during early development. BMC developmental biology *11*, 30.

Visnes, T., Doseth, B., Pettersen, H.S., Hagen, L., Sousa, M.M.L., Akbari, M., Otterlei, M., Kavli, B., Slupphaug, G., and Krokan, H.E. (2009). Uracil in DNA and its processing by different DNA glycosylases. Philosophical Transactions of the Royal Society B: Biological Sciences *364*, 563-568.

Vongs, A., Kakutani, T., Martienssen, R.A., and Richards, E.J. (1993). Arabidopsis thaliana DNA methylation mutants. Science *260*, 1926-1928.

Voo, K.S., Carlone, D.L., Jacobsen, B.M., Flodin, A., and Skalnik, D.G. (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. Mol Cell Biol *20*, 2108-2121.

Waddington, C.H. (2012). The Epigenotype. International Journal of Epidemiology *41*, 10-13.

Wade, P.A., Gegonne, A., Jones, P.L., Ballestar, E., Aubry, F., and Wolffe, A.P. (1999). Mi-2 complex couples DNA methylation to chromatin remodelling and histone deacetylation. Nature genetics *23*, 62-66.

Wagschal, A., Sutherland, H.G., Woodfine, K., Henckel, A., Chebli, K., Schulz, R., Oakey, R.J., Bickmore, W.A., and Feil, R. (2008). G9a Histone Methyltransferase Contributes to Imprinting in the Mouse Placenta. Molecular and cellular biology *28*, 1104-1113.

Wallace, H.A.C., Marques-Kranc, F., Richardson, M., Luna-Crespo, F., Sharpe, J.A., Hughes, J., Wood, W.G., Higgs, D.R., and Smith, A.J.H. (2007). Manipulating the Mouse Genome to Engineer Precise Functional Syntenic Replacements with Human Sequence. Cell *128*, 197-209.

Wan, X., Hu, B., Liu, J.-x., Feng, X., and Xiao, W. (2011). Zebrafish mll Gene Is Essential for Hematopoiesis. Journal of Biological Chemistry *286*, 33345-33357.

Wang, K.C., and Chang, H.Y. (2011). Molecular mechanisms of long noncoding RNAs. Mol Cell *43*, 904-914.

Wang, L., Brown, J.L., Cao, R., Zhang, Y., Kassis, J.A., and Jones, R.S. (2004). Hierarchical recruitment of polycomb group silencing complexes. Mol Cell *14*, 637-646.

Wang, T., Chen, K., Zeng, X., Yang, J., Wu, Y., Shi, X., Qin, B., Zeng, L., Esteban, M.A., Pan, G., and Pei, D. (2011). The histone demethylases Jhdm1a/1b enhance somatic cell reprogramming in a vitamin-C-dependent manner. Cell stem cell *9*, 575-587.

Wang, Y., and Leung, F.C.C. (2004). An evaluation of new criteria for CpG islands in the human genome as gene markers. Bioinformatics *20*, 1170-1177.

Wang, Z., Song, J., Milne, T.A., Wang, G.G., Li, H., Allis, C.D., and Patel, D.J. (2010). Pro isomerization in MLL1 PHD3-bromo cassette connects H3K4me readout to CyP33 and HDAC-mediated repression. Cell *141*, 1183-1194.

Ward, Michelle C., Wilson, Michael D., Barbosa-Morais, Nuno L., Schmidt, D., Stark, R., Pan, Q., Schwalie, Petra C., Menon, S., Lukk, M., Watt, S., Thybert, D., Kutter, C., Kirschner, K., Flicek, P., Blencowe, Benjamin J., and Odom, Duncan T. (2013). Latent Regulatory Potential of Human-Specific Repetitive Elements. Molecular cell *49*, 262-272.

Waters, T.R., and Swann, P.F. (2000). Thymine-DNA glycosylase and G to A transition mutations at CpG sites. Mutation Research/Reviews in Mutation Research *462*, 137-147.

Watt, F., and Molloy, P.L. (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*, 1136-1143.

Weber, M., Davies, J.J., Wittig, D., Oakeley, E.J., Haase, M., Lam, W.L., and Schubeler, D. (2005). Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. Nature genetics *37*, 853-862.

Weber, M., Hellmann, I., Stadler, M.B., Ramos, L., Paabo, S., Rebhan, M., and Schubeler, D. (2007). Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. Nature genetics *39*, 457-466.

Weyrich, A. (2001). Preparation of Genomic DNA from Mammalian Sperm. In Current Protocols in Molecular Biology (John Wiley & Sons, Inc.).

White, H.E., Durston, V.J., Harvey, J.F., and Cross, N.C. (2006). Quantitative analysis of SNRPN(correction of SRNPN) gene methylation by pyrosequencing as a diagnostic test for Prader-Willi syndrome and Angelman syndrome. Clinical chemistry *52*, 1005-1013.

Wiench, M., John, S., Baek, S., Johnson, T.A., Sung, M.-H., Escobar, T., Simmons, C.A., Pearce, K.H., Biddie, S.C., Sabo, P.J., Thurman, R.E., Stamatoyannopoulos, J.A., and Hager, G.L. (2011). DNA methylation status predicts cell type-specific enhancer activity. EMBO J *30*, 3028-3039.

Wilkinson, L.S., Davies, W., and Isles, A.R. (2007). Genomic imprinting effects on brain development and function. Nature reviews Neuroscience *8*, 832-843.

Williams, K., Christensen, J., Pedersen, M.T., Johansen, J.V., Cloos, P.A., Rappsilber, J., and Helin, K. (2011). TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. Nature *473*, 343-348.

Wilson, G.G., and Murray, N.E. (1991). Restriction and modification systems. Annu Rev Genet *25*, 585-627.

Wilson, M.D., Barbosa-Morais, N.L., Schmidt, D., Conboy, C.M., Vanes, L., Tybulewicz, V.L.J., Fisher, E.M.C., Tavaré, S., and Odom, D.T. (2008). Species-Specific Transcription in Mice Carrying Human Chromosome 21. Science *322*, 434-438.

Wong, E., Yang, K., Kuraguchi, M., Werling, U., Avdievich, E., Fan, K., Fazzari, M., Jin, B., Brown, A.M., Lipkin, M., and Edelmann, W. (2002). Mbd4 inactivation increases Cright-arrowT transition mutations and promotes gastrointestinal tumor formation. Proc Natl Acad Sci U S A *99*, 14937-14942.

Wood, A.J., and Oakey, R.J. (2006). Genomic imprinting in mammals: emerging themes and established theories. PLoS Genet *2*, e147.

Wossidlo, M., Arand, J., Sebastiano, V., Lepikhov, K., Boiani, M., Reinhardt, R., Scholer, H., and Walter, J. (2010). Dynamic link of DNA demethylation, DNA strand breaks and repair in mouse zygotes. EMBO J *29*, 1877-1888.

Wossidlo, M., Nakamura, T., Lepikhov, K., Marques, C.J., Zakhartchenko, V., Boiani, M., Arand, J., Nakano, T., Reik, W., and Walter, J. (2011). 5-Hydroxymethylcytosine in the mammalian zygote is linked with epigenetic reprogramming. Nat Commun *2*, 241.

Wu, H., D'Alessio, A.C., Ito, S., Wang, Z., Cui, K., Zhao, K., Sun, Y.E., and Zhang, Y. (2011a). Genome-wide analysis of 5-hydroxymethylcytosine distribution reveals its dual function in transcriptional regulation in mouse embryonic stem cells. Genes Dev *25*, 679-684.

Wu, H., D'Alessio, A.C., Ito, S., Xia, K., Wang, Z., Cui, K., Zhao, K., Sun, Y.E., and Zhang, Y. (2011b). Dual functions of Tet1 in transcriptional regulation in mouse embryonic stem cells. Nature *473*, 389-393.

Wu, J.C., and Santi, D.V. (1987). Kinetic and catalytic mechanism of HhaI methyltransferase. Journal of Biological Chemistry *262*, 4778-4786.

Wu, X., Johansen, J.V., and Helin, K. (2013). Fbxl10/Kdm2b recruits polycomb repressive complex 1 to CpG islands and regulates H2A ubiquitylation. Mol Cell *49*, 1134-1146.

Wysocka, J., Swigut, T., Xiao, H., Milne, T.A., Kwon, S.Y., Landry, J., Kauer, M., Tackett, A.J., Chait, B.T., Badenhorst, P., Wu, C., and Allis, C.D. (2006). A PHD finger of NURF couples histone H3 lysine 4 trimethylation with chromatin remodelling. Nature *442*, 86-90.

Wyszynski, M.W., Gabbara, S., and Bhagwat, A.S. (1992). Substitutions of a cysteine conserved among DNA cytosine methylases result in a variety of phenotypes. Nucleic acids research *20*, 319-326.

Xie, W., Schultz, Matthew D., Lister, R., Hou, Z., Rajagopal, N., Ray, P., Whitaker, John W., Tian, S., Hawkins, R.D., Leung, D., Yang, H., Wang, T., Lee, Ah Y., Swanson, Scott A., Zhang, J., Zhu, Y., Kim, A., Nery, Joseph R., Urich, Mark A., Kuan, S., Yen, C.-a., Klugman, S., Yu, P., Suknuntha, K., Propson, Nicholas E., Chen, H., Edsall, Lee E., Wagner, U., Li, Y., Ye, Z., Kulkarni, A., Xuan, Z., Chung, W.-Y., Chi, Neil C., Antosiewicz-Bourget, Jessica E., Slukvin, I., Stewart, R., Zhang, Michael Q., Wang, W., Thomson, James A., Ecker, Joseph R., and Ren, B. (2013). Epigenomic Analysis of Multilineage Differentiation of Human Embryonic Stem Cells. Cell *153*, 1134-1148.

Xin, Z., Tachibana, M., Guggiari, M., Heard, E., Shinkai, Y., and Wagstaff, J. (2003). Role of Histone Methyltransferase G9a in CpG Methylation of the Prader-Willi Syndrome Imprinting Center. Journal of Biological Chemistry *278*, 14996-15000.

Xu, C., Bian, C., Lam, R., Dong, A., and Min, J. (2011a). The structural basis for selective binding of non-methylated CpG islands by the CFP1 CXXC domain. Nat Commun *2*, 227.

Xu, G.L., Bestor, T.H., Bourc'his, D., Hsieh, C.L., Tommerup, N., Bugge, M., Hulten, M., Qu, X., Russo, J.J., and Viegas-Pequignot, E. (1999). Chromosome instability and immunodeficiency syndrome caused by mutations in a DNA methyltransferase gene. Nature *402*, 187-191.

Xu, J., Watts, J.A., Pope, S.D., Gadue, P., Kamps, M., Plath, K., Zaret, K.S., and Smale, S.T. (2009). Transcriptional competence and the active marking of tissue-specific enhancers by defined transcription factors in embryonic and induced pluripotent stem cells. Genes & development *23*, 2824-2838.

Xu, Y., Wu, F., Tan, L., Kong, L., Xiong, L., Deng, J., Barbera, A.J., Zheng, L., Zhang, H., Huang, S., Min, J., Nicholson, T., Chen, T., Xu, G., Shi, Y., Zhang, K., and Shi, Y.G. (2011b). Genome-wide regulation of 5hmC, 5mC, and gene expression by Tet1 hydroxylase in mouse embryonic stem cells. Mol Cell *42*, 451-464.

Xu, Y., Xu, C., Kato, A., Tempel, W., Abreu, J.G., Bian, C., Hu, Y., Hu, D., Zhao, B., Cerovina, T., Diao, J., Wu, F., He, H.H., Cui, Q., Clark, E., Ma, C., Barbara, A., Veenstra, G.J., Xu, G., Kaiser, U.B., Liu, X.S., Sugrue, S.P., He, X., Min, J., Kato, Y., and Shi, Y.G. (2012). Tet3 CXXC Domain and Dioxygenase Activity Cooperatively Regulate Key Genes for Xenopus Eye and Neural Development. Cell *151*, 1200-1213.

Yagi, H., Deguchi, K., Aono, A., Tani, Y., Kishimoto, T., and Komori, T. (1998). Growth Disturbance in Fetal Liver Hematopoiesis of Mll-Mutant Mice. Blood *92*, 108-117.

Yang, H., Fitz-Gibbon, S., Marcotte, E.M., Tai, J.H., Hyman, E.C., and Miller, J.H. (2000). Characterization of a Thermostable DNA Glycosylase Specific for U/G and T/G Mismatches from the Hyperthermophilic ArchaeonPyrobaculum aerophilum. Journal of Bacteriology *182*, 1272-1279.

Yang, X.-J. (2004). Lysine acetylation and the bromodomain: a new partnership for signaling. BioEssays *26*, 1076-1087.

Yoder, J.A., and Bestor, T.H. (1998). A candidate mammalian DNA methyltransferase related to pmt1p of fission yeast. Hum Mol Genet *7*, 279-284.

Yokoyama, A., Somervaille, T.C., Smith, K.S., Rozenblatt-Rosen, O., Meyerson, M., and Cleary, M.L. (2005). The menin tumor suppressor protein is an essential oncogenic cofactor for MLL-associated leukemogenesis. Cell *123*, 207-218.

Yoo, C.B., and Jones, P.A. (2006). Epigenetic therapy of cancer: past, present and future. Nat Rev Drug Discov *5*, 37-50.

Yoon, H.G., Chan, D.W., Reynolds, A.B., Qin, J., and Wong, J. (2003). N-CoR mediates DNA methylation-dependent repression through a methyl CpG binding protein Kaiso. Mol Cell *12*, 723-734.

You, J.S., Kelly, T.K., De Carvalho, D.D., Taberlay, P.C., Liang, G., and Jones, P.A. (2011). OCT4 establishes and maintains nucleosome-depleted regions that provide additional layers of epigenetic regulation of its target genes. Proceedings of the National Academy of Sciences *108*, 14497-14502.

Young, S.R., Mumaw, C., Marrs, J.A., and Skalnik, D.G. (2006). Antisense targeting of CXXC finger protein 1 inhibits genomic cytosine methylation and primitive hematopoiesis in zebrafish. J Biol Chem *281*, 37034-37044.

Yu, B.D., Hess, J.L., Horning, S.E., Brown, G.A., and Korsmeyer, S.J. (1995). Altered Hox expression and segmental identity in Mll-mutant mice. Nature *378*, 505-508.

Yu, M., Mazor, T., Huang, H., Huang, H.T., Kathrein, K.L., Woo, A.J., Chouinard, C.R., Labadorf, A., Akie, T.E., Moran, T.B., Xie, H., Zacharek, S., Taniuchi, I., Roeder, R.G., Kim, C.F., Zon, L.I., Fraenkel, E., and Cantor, A.B. (2012). Direct recruitment of polycomb repressive complex 1 to chromatin by core binding transcription factors. Mol Cell *45*, 330-343.

Yu, P.B., Hong, C.C., Sachidanandan, C., Babitt, J.L., Deng, D.Y., Hoyng, S.A., Lin, H.Y., Bloch, K.D., and Peterson, R.T. (2008). Dorsomorphin inhibits BMP signals required for embryogenesis and iron metabolism. Nat Chem Biol *4*, 33-41.

Zangi, R., Arrieta, A., and Cossío, F.P. (2010). Mechanism of DNA Methylation: The Double Role of DNA as a Substrate and as a Cofactor. Journal of Molecular Biology *400*, 632-644.

Zeleznik-Le, N.J., Harden, A.M., and Rowley, J.D. (1994). 11q23 translocations split the "AT-hook" cruciform DNA-binding region and the transcriptional repression domain from the activation domain of the mixed-lineage leukemia (MLL) gene. Proc Natl Acad Sci U S A *91*, 10610-10614.

Zeller, K.I., Zhao, X., Lee, C.W.H., Chiu, K.P., Yao, F., Yustein, J.T., Ooi, H.S., Orlov, Y.L., Shahab, A., Yong, H.C., Fu, Y., Weng, Z., Kuznetsov, V.A., Sung, W.-K., Ruan, Y., Dang, C.V., and Wei, C.-L. (2006). Global mapping of c-Myc binding sites and target gene networks in human B cells. Proceedings of the National Academy of Sciences *103*, 17834-17839.

Zemach, A., McDaniel, I.E., Silva, P., and Zilberman, D. (2010). Genome-wide evolutionary analysis of eukaryotic DNA methylation. Science *328*, 916-919.

Zemach, A., and Zilberman, D. (2010). Evolution of eukaryotic DNA methylation and the pursuit of safer sex. Current biology : CB *20*, R780-785.

Zentner, G.E., Tesar, P.J., and Scacheri, P.C. (2011). Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. Genome Res *21*, 1273-1283.

Zhang, H., Zhang, X., Clark, E., Mulcahey, M., Huang, S., and Shi, Y.G. (2010a). TET1 is a DNA-binding protein that modulates DNA methylation and gene transcription via hydroxylation of 5-methylcytosine. Cell Res *20*, 1390-1393.

Zhang, X., Yazaki, J., Sundaresan, A., Cokus, S., Chan, S.W., Chen, H., Henderson, I.R., Shinn, P., Pellegrini, M., Jacobsen, S.E., and Ecker, J.R. (2006). Genome-wide high-resolution mapping and functional analysis of DNA methylation in arabidopsis. Cell *126*, 1189-1201.

Zhang, Y., Jurkowska, R., Soeroes, S., Rajavelu, A., Dhayalan, A., Bock, I., Rathert, P., Brandt, O., Reinhardt, R., Fischle, W., and Jeltsch, A. (2010b). 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*, 4246-4253.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., and Liu, X.S. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol *9*, 17.

Zhang, Y., Ng, H.H., Erdjument-Bromage, H., Tempst, P., Bird, A., and Reinberg, D. (1999). Analysis of the NuRD subunits reveals a histone deacetylase core complex and a connection with DNA methylation. Genes Dev *13*, 1924-1935.

Zhao, X., Ueba, T., Christie, B.R., Barkho, B., McConnell, M.J., Nakashima, K., Lein, E.S., Eadie, B.D., Willhoite, A.R., Muotri, A.R., Summers, R.G., Chun, J., Lee, K.F., and Gage, F.H. (2003). Mice lacking methyl-CpG binding protein 1 have deficits in adult neurogenesis and hippocampal function. Proc Natl Acad Sci U S A *100*, 6777-6782.

Zhou, J.C., Blackledge, N.P., Farcas, A.M., and Klose, R.J. (2012). Recognition of CpG island chromatin by KDM2A requires direct and specific interaction with linker DNA. Mol Cell Biol *32*, 479-489.

Zhu, H., Geiman, T.M., Xi, S., Jiang, Q., Schmidtmann, A., Chen, T., Li, E., and Muegge, K. (2006). Lsh is involved in de novo methylation of DNA. EMBO J *25*, 335-345.

Zilberman, D., Gehring, M., Tran, R.K., Ballinger, T., and Henikoff, S. (2007). Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. Nature genetics *39*, 61-69.

Zilberman, D., and Henikoff, S. (2007). Genome-wide analysis of DNA methylation patterns. Development *134*, 3959-3965.

Ziller, M.J., Gu, H., Muller, F., Donaghey, J., Tsai, L.T., Kohlbacher, O., De Jager, P.L., Rosen, E.D., Bennett, D.A., Bernstein, B.E., Gnirke, A., and Meissner, A. (2013). Charting a dynamic DNA methylation landscape of the human genome. Nature *500*, 477-481.

Zou, X., Ma, W., Solov'yov, I.A., Chipot, C., and Schulten, K. (2012). Recognition of methylated DNA through methyl-CpG binding domain proteins. Nucleic Acids Res *40*, 2747-2758.