Evolutionary Landscape of CpG Island Methylation in X Chromosome Inactivation

Mengning Liu

Corpus Christi College, University of Cambridge

& The Wellcome Trust Sanger Institute

This dissertation is submitted for the degree of

Doctor of Philosophy

October 2008

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

Mengning Liu

To my dearest mum, for giving me unconditional love and the curiosity genes...

Abstract

Dosage compensation in mammals depends upon X Chromosome Inactivation (XCI), the transcriptional silencing of one X in female cells. In human and mouse, CpG island methylation on the inactive X (X_i) has been implicated in the maintenance of XCI and has been used previously to indicate a gene's XCI status. There is evidence that XCI is more complete in mouse than in human, and so I speculated that CpG island methylation on the X_i might also be more extensive in the mouse. In marsupials, by contrast, the small amount of available evidence points to absence of methylation on the X_i. I have studied the methylation of CpG islands on the X chromosomes of human, mouse and opossum in order to provide further evidence for these suggested species differences.

Ninety-one human genes, including several thought to escape from XCI, were assayed together with the mouse orthologues for 52 of these genes. Female and male genomic DNA was digested with *Msp*I or its methylation-sensitive isoschizomer *Hpa*II then tested the ability of the DNA to support amplification of PCR products containing multiple *Msp*I/*Hpa*II sites. Only the three known mouse escapees tested (*Eif2s3*, *Ddx3x*, *Utx*) were hypomethylated in the female, compared to 13 genes in human. This is consistent with the suggestion that gene

silencing is more effective in the mouse. Furthermore, the assay indicated that partial methylation is common among the human genes but rare in the mouse. Bisulphite sequencing of CpG islands from eight human-mouse orthologous pairs has confirmed this difference.

The restriction-PCR assay was then applied to 37 X-linked genes in opossum and CpG island hypermethylation was found to be rare on the female X chromosomes. However, for at least six genes, there was a greater level of methylation in the female sample, which was subsequently confirmed by bisulphite sequencing.

Results from this study support the view that CpG island methylation as a maintenance mechanism for XCI is common in the eutherian mammals. The lower level of methylation in human than mouse is consistent with the suggestion that escape from XCI is associated with a failure to maintain the inactive state. If this is correct, a high level of escape from XCI in marsupials might be anticipated. This study has also provided the first evidence of a possible role of CGI methylation in marsupial XCI.

Acknowledgements

I would firstly like to thank my supervisor Mark Ross, for his encouragement, support, and guidance throughout this project. It was extremely fortunate for me to have such a supervisor, who always had time for me, and taught me so much in every aspect of science, from experimental design to data analysis, from critical thinking to scientific writing. I can not have asked for a better supervisor.

I would also like to thank my second supervisor Nigel Carter, for all the help and support he has given me, without which I could not have finished this project so smoothly after my team left. And thanks to Nigel as the most active member of my thesis committee, who has injected many helpful comments during the entire project. I am also grateful to have the other two members of my thesis committee, Malcolm Ferguson-Smith and Chris Tyler-Smith, for the many lively discussions we had.

Special thanks also to Colette Johnston, for all the experimental skills, precious experience, and inspiring discussions she shared with me; and Alison Coffey, for taking valuable time to proof-read my thesis and the great advice she has offered me. Thanks to Christina Hedberg-Delouka, without her superb temper and organisation skills my PhD would not have progressed so efficiently.

I am indebted to many people at the Sanger Institute for their expertise and kindness: Val Curwen and Steve Searle for computational help; Matthew Francis, Nancy Holroyd, Ruth Taylor, Zahra Abdellah, Carl Skuce, and other members of the faculty small sequencing projects team, for fulfilling my massive sequencing requests with great patience and top quality.

The X team, regrettably no longer existing, will always be fondly remembered. Thanks to my fantastic team-mates Frances Lovell, Christine Burrows, Helen Albert, Daniel Leongamornlert and Yujun Zhang, for all the help and fun. Also thanks to the wonderful people in Team 70 and office F207, for giving me the joy of belonging to a team again. My fellow Sanger PhD students have been a source of delight at all times. Special thanks to Matthew Davis, for sitting next to me in many stressful evenings and weekends; and to Katherine Fawcett, for keeping my spirits up in the final days of my PhD.

Last but not least, I would like to thank my family. Thanks to my parents, for their ongoing love and support and opening my eyes to science; and thanks to my beloved husband, Li Liu, for, really, everything.

Contents

Abstract iii

Acknowledgements v

1 Introduction **1**

1.1 Dosage compensation and X chromosome inactivation (XCI) . . . 1

1.2 The mechanism of XCI 5

1.3 Escape from XCI 12

1.4 XCI in other mammals 19

1.5 Aims of the thesis 24

2 Materials and Methods **26**

2.1 Chemical reagents 26

2.2 Enzymes and commercially prepared kits 27

2.3 Solutions and buffers 28

2.3.1 Buffers 28

2.3.2 Media 29

2.4 Size markers 30

2.5 Tissues and cell lines 31

2.6 Primer sequences 32

2.7 Key World Wide Web addresses 32
2.8 Mammalian cell culture 33
2.8.1 Resuscitating frozen cells 33
2.8.2 Growing and harvesting cells 33
2.8.3 Freezing cells for storage 34
2.9 DNA extraction 35
2.9.1 Phenol/chloroform extraction of genomic DNA from animal tissues 35
2.9.2 Phenol/chloroform extraction of DNA from a reaction mix 36
2.9.3 Mini-preps of genomic DNA from cultured cells 37
2.10 DNA manipulation 37
2.10.1 Purification of DNA 37
2.10.2 Quantification 38
2.10.3 Restriction enzyme digests 39
2.10.4 Agarose gel electrophoresis 39
2.10.5 Bisulphite modification 40
2.11 Polymerase Chain Reaction (PCR) 41
2.11.1 Primer design 42
2.11.2 Restriction-PCR Methylation Assay (RPMA) 42
2.11.3 Bisulphite PCR 44
2.11.4 Colony PCR 45
2.12 Bacterial cloning 46
2.12.1 Ligation of vector and insert 46
2.12.2 Transformation 46
2.13 Plasmid sequencing 47

2.14 Computational analysis 47
2.14.1 cpg 47
2.14.2 BLAST 47
2.14.3 RepeatMasker 48
2.14.4 GAP4 48
2.14.5 Alignment of DNA sequences 48
2.14.6 Multi-species comparative sequence analysis 49
2.14.7 MethTools 49
2.14.8 Perl scripts 49

3 Comparison of CpG island methylation on the mouse and human X chromosomes 51

3.1 Introduction 51
3.1.1 Identification of CGIs 54
3.1.2 Approaches to studying CGI methylation 56
3.1.3 Aims 61
3.2 Identification of target CpG islands 61
3.2.1 Region of interest 61
3.2.2 Identification of target human and mouse CGIs 62
3.3 Assessing CpG island methylation on mouse and human X chromosomes by Restriction-PCR Methylation Assay (RPMA) 63
3.3.1 Optimisation of experimental conditions 65
3.3.2 Primer design 69
3.3.3 Analysis of CGI methylation on the human X chromosome 69
3.3.4 Analysis of CGI methylation on the mouse X chromosome 76

3.4 Discussion 79

4 Comparison of detailed CGI methylation patterns on the human and mouse X chromosomes 84

4.1 Introduction 84

4.1.1 Aims of this chapter 89

4.2 High resolution analysis of methylation levels on the mouse and human X chromosomes by bisulphite sequencing 91

4.2.1 Identification of target CGIs 91

4.2.2 Primer design and PCR 93

4.2.3 Cloning and sequencing 94

4.2.4 Methylation analysis 95

4.2.5 Comparison between human and mouse methylation profiles 97

4.2.6 Comparison between bisulphite sequencing and RPMA results 110

4.3 Discussion 113

5 Investigation of CpG island methylation on the opossum X chromosome 118

5.1 Introduction 118

5.1.1 The grey short-tailed opossum (*M. domestica*) 121

5.1.2 The opossum X chromosome 122

5.1.3 Aims of this chapter 124

5.2 Identification of CGIs on the opossum X chromosome 124

5.3 Assessing CGI methylation on the opossum X chromosomes by RPMA 125

5.3.1 Analysis of CGI methylation on the opossum X chromosome 126
5.4 Assessing CGI methylation on the opossum X chromosomes by bisulphite sequencing 132
5.5 Comparison of methylation status across multiple tissues 143
5.6 Discussion 153

6 Comparative analysis of features of 5' CGIs in human, mouse and opossum 158
6.1 Introduction 158
6.1.1 Aims of this chapter 160
6.2 CGIs in the conserved region of human, mouse and opossum X chromosome 160
6.3 CGIs associated with orthologous genes on human, mouse and opossum XCR 163
6.4 Conservation of 5' CGIs associated with orthologous genes 165
6.5 Discussion 172

7 Discussion 175
7.1 Future directions 179
7.2 Conclusions 183

References 184

List of Tables

2.1 Key World Wide Web addresses used in this study 32

3.1 RPMA results of human CGIs 74

3.2 RPMA results of mouse CGIs 78

4.1 Human and mouse genes chosen for bisulphite sequencing analysis of CGI methylation 92

4.2 Comparison between bisulphite sequencing and RPMA results . . 112

5.1 RPMA results of opossum CGIs 131

6.1 Comparison of XCR gene and CGI features in opossum, human, and mouse. 162

6.2 Comparison of 5' CGIs associated with the orthologous gene set. . 165

List of Figures

1.1 Tortoiseshell cat 5

1.2 Evolutionary strata of human X chromosome 15

3.1 Approaches to studying DNA methylation. 60

3.2 Restriction-PCR Methylation Assay (RPMA) 65

3.3 Optimisation of restriction enzyme digestion conditions 68

3.4 RPMA results of human CGIs with hypomethylation or female-specific methylation 72

3.5 RPMA results of human CGIs with other methylation patterns . . 73

3.6 RPMA results of mouse CGIs with hypomethylation or female-specific methylation of CGI 77

4.1 Bisulphite sequencing 86

4.2 Hypothesised model of CGI methylation patterns on the human and mouse X chromosomes 90

4.3 Graphical representation of CGI methylation. 96

4.4 CGI methylation profiles of human *DDX3X* and mouse *Ddx3x* . . 98

4.5 CGI methylation profiles of human *UTX* and mouse *Utx* 99

4.6 CGI methylation profiles of human *EIF2S3* and mouse *Eif2s3x* . . 100

4.7 CGI methylation profiles of human *OFD1* and mouse *Ofd1* 101
4.8 CGI methylation profiles of human *POLA* and mouse *Pola1* . . . 103
4.9 CGI methylation profiles of human *MSL3L1* and mouse *Msl3l1* . 104
4.10 CGI methylation profiles of human *ATP6AP2* and mouse *Atp6ap2* 105
4.11 CGI methylation profiles of human *PRPS2* and mouse *Prps2* . . . 106
4.12 CGI methylation profiles of human *RAB9A* 107
4.13 CGI methylation profiles of human *CDKL5* 108
4.14 CGI methylation profiles of mouse *Syap1* 109
4.15 CGI methylation profiles of mouse *Hccs* 110
4.16 Variation of CGI methylation on the human and mouse X chromosomes - revised model 117

5.1 South American, grey short-tailed opossum (*Monodelphis domestica*) 122
5.2 RPMA results of three opossum genes with female-specific methylation of CGI 130
5.3 Distribution of assayed CGIs on the opossum X chromosome . . . 132
5.4 CGI methylation profiles of opossum *G6PD* 134
5.5 CGI methylation profiles of opossum *HPRT1* 135
5.6 CGI methylation profiles of opossum *FAAH2* 138
5.7 CGI methylation profiles of opossum *FGF16* 139
5.8 CGI methylation profiles of opossum *IAG2* 140
5.9 CGI methylation profiles of opossum *PGK1* 141
5.10 CGI methylation profiles of opossum *ITM2A* 142
5.11 CGI methylation profiles of opossum *F8A*∼ 143

5.12 Graphical representation of average methylation levels in individual CGI 145
5.13 CGI methylation profiles of opossum *G6PD* in four tissues 146
5.14 CGI methylation profiles of opossum *HPRT1* in three tissues . . . 147
5.15 CGI methylation profiles of opossum *FGF16* in four tissues 148
5.16 CGI methylation profiles of opossum *PGK1* in four tissues 149
5.17 CGI methylation profiles of opossum *ITM2A* in four tissues 150
5.18 CGI methylation profiles of opossum *IAG2* in four tissues 151
5.19 CGI methylation profiles of opossum *F8A*∼ in four tissues 152

6.1 Comparison of sizes of 5' CGIs in opossum, human, and mouse XCR. 163
6.2 Orthologous genes on the opossum, human, and mouse X chromosomes 164
6.3 CGIs associated with the orthologous gene set 165
6.4 Homology between 5' CGIs of the orthologous genes. 166
6.5 VISTA plot of human, mouse and opossum *PGK1* and its CGI . . 168
6.6 VISTA plot of human, mouse and opossum *SOX3* and its CGI . . 169
6.7 VISTA plot of human, mouse and opossum *PGRMC1* and its CGI 170
6.8 VISTA plot of human, mouse and opossum *PIM2* and its CGI . . 171

7.1 Phylogeny and time scale of mammalian species with available sequence information 181

Chapter 1

Introduction

1.1 Dosage compensation and X chromosome inactivation (XCI)

Most animals and plants adopt sexual reproduction, which creates great genetic variation. Many sexually reproducing species consist of two sexes, male and female. In some cases, like in certain species of fish and reptiles, sex is determined by environmental variables, but in most cases, sex determination is genetic. A system of chromosomal sex-determination has evolved independently in many plant and animal groups, where the two sexes have different chromosomal constitutions (Bull, 1983). One sex, termed the heterogametic sex, has two different sex chromosomes, while the other sex, the homogametic sex, has a pair of identical chromosomes. The X and Y chromosomes (when male is heterogametic), or the Z and W chromosomes (when female is heterogametic), are often morphologically and genetically distinct, shaped by sex chromosome evolution.

1.1 Dosage compensation and X chromosome inactivation (XCI)

In the 1960s, Ohno (1967) first proposed an autosomal origin for sex chromosomes. According to this hypothesis, a pair of proto-sex chromosomes arose from a pair of autosomes when a sex-determining allele evolved on one of them, termed the Y or W chromosome (the XY system is discussed thereafter). Starting from a pair of identical autosomes, the sex chromosomes went onto very different paths. Selection favoured restriction of recombination between the Y-linked sex-determining region and its X homologue to avoid sexual ambiguity. This restriction eventually spread throughout the two chromosomes, leading to degeneration of the Y due to accumulation of mutations. The X chromosome, on the other hand, retained its gene content, which comprises an indispensable part of the genome, and had to accommodate its hemizygous state in male. Thus the need for dosage compensation between the two sexes arose.

Just as chromosomal sex-determination has evolved many times in different lineages, so dosage compensation mechanisms have also evolved multiple times. Typical of evolution, different solutions have been used to solve the same problem: to achieve dosage balance between gene expression from the two sexes, as well as balance between expression from the X and autosomes. Dosage compensation mechanisms have been best studied in the fruit fly *Drosophila melanogaster*, the nematode worm *Caenorhabditis elegans*, and in mouse and human. Each group has evolved a drastically different system to achieve dosage compensation. Drosophila has an XY sex-determination system but without a male-determining gene on the Y. Rather, sex determination results from counting the X:autosome ratio. An X:autosome ratio of 1:1 results in development into a female, and an X:autosome ratio of 2:1 gives rise to a male. So an XO individual in Drosophila is a male, not a female as in human. Dosage compensation is achieved by up-

regulating the expression of genes on the single X in male two-fold (reviewed in Baker *et al.*, 1994; Cline and Meyer, 1996). In *C. elegans*, the Y chromosome is lost completely, and sex is determined by the number of X chromosomes. As in Drosophila, the XO individual is a male, but the XX individual is a hermaphrodite (having both male and female reproductive organs). Dosage compensation is targeted at both X chromosomes in the hermaphrodite, reducing the transcription level of each X by half (reviewed in Cline and Meyer, 1996; Meyer, 2000).

In contrast to the straightforward solutions evolved in flies and worms, a much more complex dosage compensation mechanism has developed in mammals, namely X chromosome inactivation (XCI). The hypothesis was first proposed by Mary Lyon over 40 years ago: in mammals, the dosage equalisation of gene products from the X chromosome is achieved by transcriptionally inactivating one of the two X chromosomes in female cells during embryogenesis (Lyon, 1961). In her landmark paper, Lyon linked together several curious observations about the X chromosome in mammals. Barr and Bertram (1949) had discovered that a 'nucleolar satellite' was always found in female cells but not in male cells; Ohno (1960; 1959) suggested that the 'Barr body', as it was later called, is a highly condensed X chromosome. Around the same time, it was found that female mice with only one X chromosome (XO) are viable and even fertile (Welshons and Russell, 1959), suggesting that one active X chromosome is sufficient for normal development. Lyon's hypothesis explains the mosaic, rather than intermediate, coat colour of female mice heterozygous for X-linked pigment genes: during very early stages of embryonic development, one X chromosome becomes inactivated in each cell, and the progenies of which can only make the normal or mutant form of the pigment, depending on which X is active, giving rise to patches of colours

(see Figure 1.1 for an example in cat).

At the same time, the idea of XCI was also indicated in a paper on mammalian sex chromosomes by Russell (1961). In an attempt to explain the variegated-type position effects, where the recessive allele is manifested in heterozygote as a result of X;autosome translocation, Russell suggested that only one X is active in females and the extra Xs acquire characteristics of heterochromatin (Russell, 1961). Direct evidence in support of the XCI hypothesis came from study of cloned cells. Davidson and colleagues (1963) studied cell cultures from human females heterozygous for the X-linked gene encoding the enzyme glucose-6-phosphate dehydrogenase (G6PD). They found that each cloned culture, originating from single cells, only expressed one of the two alleles. This proved that only one X chromosome is active in each cell, and showed how the inactivation pattern of every cell is faithfully reproduced in its daughter cells.

Subsequently, Ohno (1967) provided a model for the evolution of XCI in mammals. Following the decay of the Y chromosome, the rate of product output of X-linked genes was doubled to compensate for the hemizygous existence of X-linked genes in the male. Once this step was accomplished, there was no longer the need for two active Xs in the female, and inactivation of one X has evolved. Therefore dosage compensation in mammals includes two separate elements: a) transcriptional silencing of the inactive X (X_i); and b) two-fold upregulation of gene expression on the active X (X_a). The upregulation has only been confirmed in the past 15 years, thanks to advance in gene expression assays (Adler *et al.*, 1997; Lin *et al.*, 2007; Nguyen and Disteche, 2006), but the silencing has been extensively studied over the last half century.

Figure 1.1: Tortoiseshell cat. Manifestation of XCI in coat colour patches. Photo reproduced from Wikimedia Commons (`http://en.wikipedia.org/wiki/Image:Tortoiseshell_tabby_-_TortiePuff.jpg`) under the GNU Free Documentation License.

1.2 The mechanism of XCI

Although many of the details are still poorly understood, decades of active research has now constructed a basic model for the initiation, spreading and maintenance of XCI.

In the somatic cells of eutherians, the choice of the X to be inactivated is random. Early studies suggested a region from where X-inactivation is initiated and spreads across the chromosome (Russell, 1963). A *cis*-acting X-inactivation cen-

tre (XIC) was supported by mouse X;autosome translocations, where inactivation spread from the inactive X into the attached autosome (Russell and Montgomery, 1965), and X chromosome deletions, where inactivation was limited to the XIC-bearing segment (Rastan and Robertson, 1985).

From the region of the human XIC, the *XIST* gene was cloned (Brown *et al.*, 1991). Its expression was only detected in human cells containing the X_i, including female and male cells with multiple X chromosomes, but not seen in cells from XO female or normal male (Brown *et al.*, 1991). By use of somatic cell hybrids containing a single human X chromosome, they confirmed that *XIST* is expressed exclusively from the X_i but not the X_a (Brown *et al.*, 1991). Using the human *XIST* cDNA as a probe, its mouse homologue *Xist* was mapped to the mouse Xic (Borsani *et al.*, 1991). Like its human homologue, expression of the mouse *Xist* is also limited to the inactive X chromosome (Borsani *et al.*, 1991). Its location and unique expression pattern in both human and mouse made *XIST/Xist* the prime candidate as a controlling element of XCI. Both the human and mouse genes contain several tandem repeats and produce large, alternatively spliced, transcripts (Brockdorff *et al.*, 1992; Brown *et al.*, 1992). The matured transcripts lack significant conserved open reading frames (ORF) and localise almost exclusively to the nucleus, suggesting that *XIST/Xist* may not encode a protein but rather function as a structural RNA (Brockdorff *et al.*, 1992; Brown *et al.*, 1992).

The role of XIST as a structural RNA in XCI was strongly supported by the close association of spliced but non-translated XIST RNA with the X_i, observed in fluorescent *in situ* hybridization (FISH) studies (Clemson *et al.*, 1996). More direct evidence of the involvement of *XIST* in XCI came from transgenic studies. The mouse embryonic stem (ES) cell has provided an excellent model to study

the mechanism of XCI. Upon induction to differentiate, the female ES cells carry out random XCI, allowing researchers better access to details of the early events. In addition, these cells can be manipulated to investigate how mutations in individual genes can affect the process of XCI. Using gene targeting to delete large parts of the *Xist* gene, Penny and colleagues (1996) demonstrated that an intact *Xist* is required for *cis* inactivation: in the cells with an X chromosome bearing the null allele of *Xist*, only the X chromosome with the intact *Xist* could be inactivated. Furthermore, *Xist* was shown to be sufficient to induce inactivation when introduced into an autosome (Herzing *et al.*, 1997).

Current evidence suggests that the role of XIC to initiate *cis* inactivation is predominantly carried out by the *XIST* locus, but there are other elements present in the XIC which may play a part in the other role of XIC: the choice of the active X. Since the X chromosome contains many genes important for normal development and proper cell functioning, it is critical to ensure one, and only one, X remains active. In random XCI, the initial step is the random choice of a single future X_a, followed by inactivation of any remaining X chromosome. Before the discovery of *Xist*, Rastan (1983) proposed a simple model, where a *trans*-acting blocking factor randomly blocks the Xic on a single X chromosome and thus protects that chromosome from being silenced. This model was supported by behaviour of embryo-derived (EK) cell lines with X chromosome deletions (Rastan and Robertson, 1985). When cells with a partly-deleted X chromosome were induced to differentiate, either the intact X or the deleted X could be inactivated; but when Xic was included in the deletion, neither X was inactivated (Rastan and Robertson, 1985).

The nature of the *trans*-acting blocking factor remains unclear to this day,

but studies in mice have provided several candidates within the Xic that may respond to the blocking factor. The *Xist* gene itself does not seem to affect choice, as mouse ES cells with one *Xist* deleted still behaved as if they had two Xic and underwent inactivation (Penny *et al.*, 1996). At least in mouse, multiple elements 3' to the *Xist* appear to be involved in the complex process of counting and choice. When a 65-kb region 3' to *Xist* exon 6 was deleted, the X chromosome bearing the deletion was always inactivated, even in XO cells, suggesting that this region is needed to repress *Xist* in *cis* (Clerc and Avner, 1998). From this region, a transcript antisense to *Xist* was discovered by RNA FISH using strand-specific probes in mouse ES cells, and was named *Tsix* (Lee and Lu, 1999). The Tsix RNA, spanning the entire length of *Xist*, is specifically expressed from the active X from the onset of XCI and persists only until the establishment of XCI, consistent with a role in blocking *Xist* function. Such a role was supported by examination of female ES cells heterozygous for a deletion of *Tsix*, in which only the mutation-bearing X was always inactivated (Lee and Lu, 1999). However, restoration of *Tsix* in cells carrying the previously described 65-kb region failed to restore random XCI, suggesting existence of additional functional elements within this region (Morey *et al.*, 2001). Further exploration in the remaining region uncovered a second *cis*-acting element, *Xite* (X inactivation intergenic transcription elements), which lies upstream of *Tsix* and appears to act through promotion of *Tsix* persistence (Ogawa and Lee, 2003). A less clearly defined locus that is involved in the choice process is the X controlling element (*Xce*), which has long been found to affect the randomness of XCI (Cattanach and Isaacson, 1967; Cattanach and Williams, 1972). It has been mapped 3' to *Xist*, but its exact location and nature remain undetermined (Chadwick *et al.*, 2006; Simmler

et al., 1993).

In human, the process of choice is even less understood. An orthologous gene of the mouse *Tsix*, *TSIX*, was also found in the human XIC (Migeon *et al.*, 2001). However, despite being an antisense to *XIST*, the human version of *TSIX* was seen expressed only from the inactive X, together with *XIST*, so does not appear to block XIST function (Migeon *et al.*, 2002). The other locus affecting choice in mouse, *Xite*, does not appear to have an orthologue in human so far.

After choosing the future X_a, *Xist* expression from the future X_i is upregulated and the *Xist* transcripts rapidly spread along and associate with, or 'coat', the chromosome in *cis* (Clemson *et al.*, 1996). The silencing itself is achieved by changing the chromatin of the future X_i to a transcriptionally inactive state via a cascade of chromatin modifications. DNA is packed around a special kind of proteins, the histones, to form chromatin. For transcription to happen, the chromatin needs to be in an 'open' state to allow the transcription initiation complex to access and associate with the DNA. In contrast, transcription is inhibited when the chromatin is in a 'closed' state, tightly packed and gives no access to the factors essential for transcription initiation. Several modifications to certain amino acids in the histone, especially in the tails, can change the conformation of the protein and in turn affects the packing of DNA around it. Histone modifications associated with a particular state of the chromatin can be seen as the histone 'signature' for the state. The 'open' chromatin, or euchromatin, usually has the lysines in histones acetylated, whereas the 'close' chromatin, or heterochromatin, normally has these lysines hypoacetylated (Grunstein, 1997).

It has been noticed that the chromatin of X_i carries epigenetic marks distinct from those of the X_a, but closely resembling the features of constitutive hete-

rochromatin, including hypoacetylation of histones H3 and H4, and enrichment of the histone H2A variant macroH2A (Boggs *et al.*, 1996; Costanzi and Pehrson, 1998; Jeppesen and Turner, 1993). Using mouse ES cells as a model, a number of research groups were able to establish the timing of these events in early development (Goto *et al.*, 2002; Heard *et al.*, 2001; reviewed in Plath *et al.*, 2002). A combination of Xist RNA FISH and immunofluorescence staining with specific antibodies also enabled researchers to detect the fine details and interplay of the various types of histone modification (Goto *et al.*, 2002; Heard *et al.*, 2001; reviewed in Plath *et al.*, 2002). Closely following the spread of Xist RNA along the X_i, one of the earliest events is the methylation of histone H3 lysine 9 (H3K9), together with hypoacetylation of H3K9 and loss of methylation at histone H3 lysine 4 (H3K4) (Heard *et al.*, 2001). Hypoacetylation of H4 was observed shortly after modifications of the H3 tail, and silencing of the two genes examined also followed X-chromosome-wide H3K9 methylation (Heard *et al.*, 2001). These histone marks are maintained in adult cells, as shown by studies of the mouse genes encoding phosphoglycerate kinase 1 (*Pgk1*) and hypoxanthine-guanine phosphoribosyltransferase (*Hprt*) (Goto *et al.*, 2002). At both loci, chromatin on the inactive X was marked by H3K9 methylation and hypoacetylation of histones H3 and H4, whereas the opposite pattern of H3K4 methylation and acetylated H3 and H4 was found on the active X (Goto *et al.*, 2002). Enrichment of macroH2A on the X_i occurred later and was not essential for inactivation, suggesting a redundant role in the maintenance of XCI (Csankovszki *et al.*, 1999; Mermoud *et al.*, 1999). As a result of a series of chromatin modifications, the chromatin of the inactive X is turned into tightly packed heterochromatin, manifesting itself as the condensed 'Barr body' during interphase, and closed to transcription.

Experiments in mouse demonstrated that, although *XIST* is still expressed for a while after the initiation of XCI and may contribute towards the stabilisation of XCI, it is not essential for the maintenance of XCI (Brown and Willard, 1994; Csankovszki *et al.*, 2001, 1999). During mitosis, *XIST* is not expressed and XIST RNA dissociates from the X_i (Clemson *et al.*, 1996); but the inactivation pattern is faithfully replicated through cell generations. The 'memory' mechanism of XCI maintenance is not clear but DNA methylation has been suggested to play a role (Kaslow and Migeon, 1987). Clusters of CpG-rich sequences, or CpG islands (CGI), are present near the 5' transcription start sites of most housekeeping genes (Bird *et al.*, 1985). CGIs are normally unmethylated to allow gene activity, but CGIs on the inactive X are frequently hypermethylated (Tribioli *et al.*, 1992). Observations in mouse embryos and mouse embryonic stem cells confirmed that CGI methylation is a late event in XCI, only occurring after the cascade of histone modification, consistent with a role in the maintenance of XCI (Keohane *et al.*, 1996; Lock *et al.*, 1987). Once established, CGI methylation patterns can be faithfully reproduced after DNA replication, thanks to the preferential methylation at semi-methylated sites by the maintenance-methylating enzyme, DNA methyltransferase 1 (Dnmt1). In support with the maintenance function, treatments with demethylation agents were shown to induce re-activation of genes on the inactive X (Graves, 1982; Jones *et al.*, 1982). CGI methylation and its involvement in XCI is discussed in more details in Chapter 3.

1.3 Escape from XCI

Much insight into the process of XCI has been gained through studying the expression of genes on the X chromosome. One interesting discovery was the exception to the rule: although most genes on the inactive X are stably silenced, a subset of genes actually escape from XCI and are expressed from both the active and inactive X.

Escapees were first discovered in humans. Very soon after Lyon published her hypothesis of XCI, there was suggestion that the XG blood group gene in human is not inactivated on X_i based on failure to identify cells manifesting the recessive allele in heterozygous females (Gorman *et al.*, 1963). A new test for XCI status was then proposed using blood cells from patients with Lesch-Nyhan syndrome (Fialkow, 1970). The syndrome is caused by a deficiency of the enzyme HPRT, but red blood cells from heterozygous female patients were found to have skewed XCI, where the X carrying the mutant allele of *HPRT* is always inactivated, presumably due to cell selection in early development. Three females, each heterozygous for both *HPRT* and *XG*, and each inherited the normal *HPRT* allele and the recessive *XG* allele from their fathers, had their red cells typed as XG(+ve), indicating that the *XG*(+ve) allele on the inactive X was also expressed (Fialkow, 1970). Later a second escapee, encoding the protein steroid sulfatase (STS), was identified using cloned cell lines, which had the same X inactivated (Shapiro *et al.*, 1979). Interestingly, *STS* is closely linked with *XG* and both are located at the extreme distal short arm of the X chromosome (Xp), suggesting a region of genes that escape from XCI.

Most subsequent escapees were discovered using the somatic cell hybrid method.

When hybrid cells are made by merging rodent and human cells, these hybrids will retain most rodent chromosomes but lose the human chromosomes. Hybrid cell lines retaining human X_a or X_i can be established by growing in suitable selective media (Mohandas *et al.*, 1980). This experimental system provides the means to directly test gene expression from morphologically normal X_a and X_i, and was first used to confirm the expression of *STS* from X_i (Mohandas *et al.*, 1980). Using somatic cell hybrids, other genes at the tip of Xp were found to escape XCI, including *CD99* (Goodfellow *et al.*, 1984), *XE7* (Ellison *et al.*, 1992), and *ANT3* (which encodes adenine nucleotide translocator 3) (Schiebel *et al.*, 1993). However, the identification of *ZFX* (which encodes a zinc-finger protein), located in the more proximal part of Xp, and *RPS4X* (which encodes ribosomal protein S4), located in the long arm of X chromosome (Xq), as escapees suggests that genes escaping inactivation are interspersed among genes subject to inactivation (Fisher *et al.*, 1990; Schneider-Gädicke *et al.*, 1989). Recently, Carrel and Willard (2005) created an inactivation profile for the whole human X chromosome by analysing gene expression from human X_a and X_i in a panel of rodent/human somatic cell hybrids. Six hundred and twenty four transcripts, representing ∼95% of all assayable genes on the human X chromosome, were assayed in up to nine cell hybrids, each containing a different X_i. They found about 15% of the transcripts to be expressed from the X_i in all or all but one or two hybrids, thus escaping from XCI. An additional 10% of the transcripts were expressed from the X_i in some hybrids, displaying some degree of escape.

The escapees are not randomly distributed along the human X, but show very strong position effect (Carrel and Willard, 2005). For example, genes on the short arm of the X chromosome appear to be more susceptible to escaping from XCI,

perhaps reflecting the evolutionary history of the sex chromosomes. As described in section 1.1, XCI arose as one of the major consequences of mammalian sex chromosome evolution. By comparing present day X-Y homologues in human, Lahn and Page (1999) proposed that human sex chromosome evolution was punctuated by at least four events, as these genes can be divided into four distinct groups based on the degree of divergence between the X and Y copies. Moreover, genes within each group are clustered together on the X chromosome, and these clusters form an orderly sequence corresponding to the degrees of divergence. Lahn and Page (1999) suggested that the most likely event to account for this phenomenon was chromosomal inversion: each time a large block of chromosome was inverted on the Y and ceased to recombine with the X, followed by decay of genes within this block on the Y and recruitment of these genes into XCI on the X. They termed these blocks evolutionary strata, stratum 1 (S1) being the oldest, and stratum 4 (S4) being the youngest. The sequencing of the human X chromosome provided further support to the evolutionary strata, and helped to detect finer detail of the evolutionary events that created the younger strata (Ross *et al.*, 2005). As a result, the boundaries between the strata were refined and S4 was subdivided into S4 and S5 (Ross *et al.*, 2005). This model of sex chromosome evolution is consistent with the distribution of genes that escape XCI: escapees are only sparse in the older strata, S1 and S2, but abundant in the younger strata, S3-5, with S4 and S5 containing almost entirely escapees (Figure 1.2). At the tip of the short arm lies the short arm pseudoautosomal region (PAR1), where X-Y homology is maintained by an obligatory recombination in male meiosis and all genes escape from XCI.

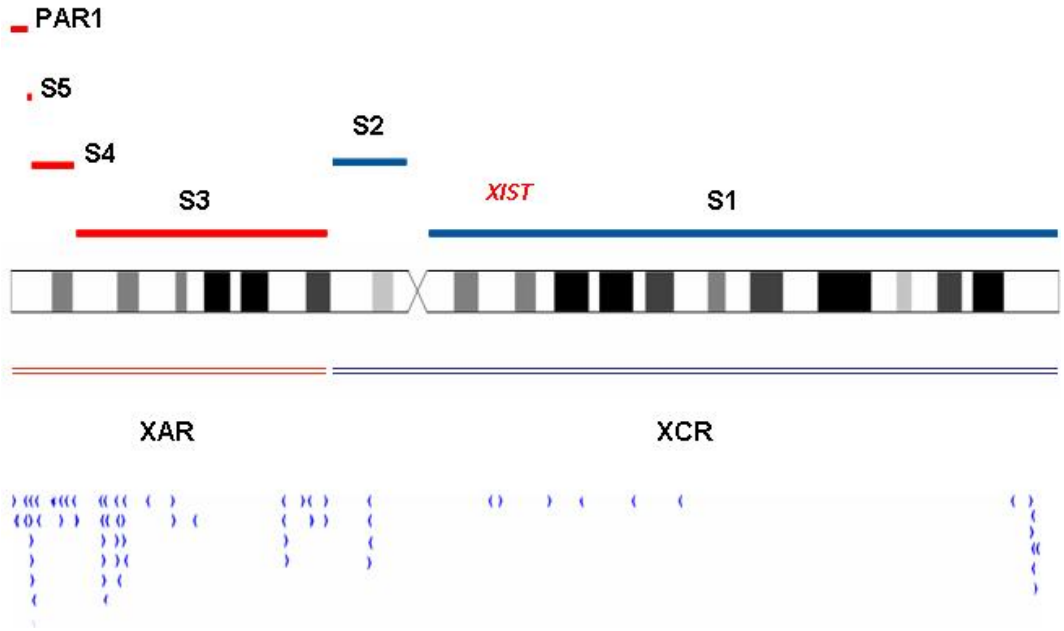

Figure 1.2: Evolutionary strata of human X chromosome. S1-5 are the five evolutionary strata. PAR1 = pseudoautosomal region; XCR = X conserved region; XAR = X added region (added to X and Y from an autosome prior to radiation of the eutherian mammals). Blue triangles represent the positions and orientations of genes that escape from X chromosome inactivation.

Escaping from XCI is not unique to human. The discovery of human escapees led researchers to test the XCI status of their mouse homologues. A convenient experimental system has been developed for mouse to distinguish between gene expression from the active X and the inactive X, using female mouse carrying Searle's X;autosome translocation, T(X;16)16H (thereafter shortened to T16H), in which the normal X is preferentially inactivated (Lyon *et al.*, 1964). By crossing T16H mice with another strain, the F1 female progeny would not only have non-random XCI, but also provide plenty of polymorphisms between the two X chromosomes for examination of allele-specific expression at almost every locus (Adler *et al.*, 1991; Ashworth *et al.*, 1991). Using this system, several human

genes that escape from XCI were found to have mouse orthologues that are also escapees (Agulnik *et al.*, 1994; Ehrmann *et al.*, 1998; Greenfield *et al.*, 1998; Wu *et al.*, 1994). However, more frequently, a gene that escapes from XCI in human is subject to inactivation in mouse (Adler *et al.*, 1991; Ashworth *et al.*, 1991; Kay *et al.*, 1991; reviewed in Disteche *et al.*, 2002). Although a chromosomal-wide XCI profile is not available for mouse, a comparison of XCI status of a homologous gene cluster in human and mouse was in strong support of more complete XCI in mouse (Tsuchiya *et al.*, 2004). When genes surrounding a known escapee in both human and mouse, *SMCX/Smcx* ('selected mouse cDNA on the X', which is now known to encode a histone demethylase), were examined for their XCI status, four additional escapees were found in human but *Smcx* remained the only escapee in this region in mouse.

It remains unclear why genes that escape would differ in different mammalian species and little is known about the XCI status of X-linked genes in other mammals. An initial multi-species comparison study was carried out using the 5' CGI methylation state as an indicator of the gene's XCI status (Jegalian and Page, 1998). According to a previous study of methylation states of 28 CGIs on the human X chromosome, methylation of the CGIs of the inactivated genes and demethylation of the CGIs of the active genes appeared to be a general rule. Jegalian and Page (1998) tested CGI methylation of *ALD* (which encodes the adrenoleukodystrophy protein), a gene known to undergo XCI in human and is presumably also silenced in other eutherian species, in 18 species representing nine eutherian orders. In all cases, methylation was observed in female, where *ALD* was expected to be silenced on the inactive X, and never observed in male, where the single X is active, in support with the correlation between CGI methy-

lation and the gene's activity status. They then tested CGI methylation of three known escapees in human in up to 19 species. Interestingly, each gene was shown to escape in a unique group of species, and no gene escapes in all species. Using the same technique, a second study compared the XCI status of seven X-linked genes in four mammals, and found a similar species-specific variation in escaping from XCI (Yen *et al.*, 2007).

Genes that escape from XCI provide a unique opportunity to study the mechanisms of propagation and maintenance of inactivation. Individual escapees were found to lack epigenetic features of inactivated genes, such as histone modification and DNA methylation (reviewed in Disteche, 2002). However, at the genetic level, no universal 'signature' of inactivated or escapee genes has been identified. No significant difference has been found between CGIs associated with inactivated and escapee genes. Previous studies of X;autosome translocation and XIC transgene showed that XCI can spread along the autosome, suggesting that it does not require genomic elements unique to the X chromosome for the propagation of XCI (Lee and Jaenisch, 1997; Rastan and Robertson, 1985). However, it was also noticed that inactivation of the autosome is not as efficient as inactivation of the X chromosome, which led to the hypothesis that there are 'way stations', possibly made of repeated sequences, along the X chromosome to facilitate spreading of the inactivation signals (Gartler and Riggs, 1983). In 1998, Lyon suggested that the LINE1 (L1 long interspersed elements) repeat family may be a good candidate for the 'way stations', as they are particularly concentrated on the X chromosome in both human and mouse, and previous X;autosome translocation and XIC transgene studies were consistent with XCI signals travelling better in LINE-rich regions. A systematic analysis of the distribution of LINE1 elements

in the human genome found a two-fold enrichment of LINE1 on the X chromosome compared to autosomes (Bailey *et al.*, 2000). They also found a significant increase of LINE1 content around XIC and significantly reduced LINE1 contents in the escape regions.

Escapees may occur from either failure to establish or failure to maintain inactivation. Current evidence is far from conclusive but does favour the latter scenario. Expression analysis of a known escapee, *Smcx*, in single cells from embryonic and adult mouse tissues suggested that this gene is inactivated in early embryonic development followed by re-activation (Lingenfelter *et al.*, 1998), perhaps due to progressive loss of CGI methylation (Disteche *et al.*, 2002). It is also not clear at which level inactivation is regulated. The multi-species comparison study carried out by Jegalian and Page (1998) suggested a gene-by-gene or cluster-by-cluster development of XCI following Y-gene decay. The presence of blocks of escapees in human indicates existence of chromatin domains of different XCI status (Carrel *et al.*, 1999), but studies in mouse showed that the escapee domain can be as small as a single gene (Tsuchiya *et al.*, 2004). The CTCF protein has been implicated in the enhancer-blocking function of chromatin insulators (Bell *et al.*, 1999; Lobanenkov *et al.*, 1990; Mukhopadhyay *et al.*, 2004). In a recent study, CTCF binding was detected at the 5' end of several escape genes adjacent to inactivated genes, but not genes embedded in large escaping domains (Filippova *et al.*, 2005).

1.4 XCI in other mammals

Most of our current knowledge about XCI comes from studies in human and mouse, both belonging to the extant mammal group Eutheria, or the 'placental' mammals. There are two other extant groups of mammals, the Metatheria, or marsupials, and the Prototheria, or monotremes. For convenience, the names eutherians, marsupials and monotremes are used in this thesis. Diverged from the monotremes $\sim$ 240 million years ago (MYA) (Murphy *et al.*, 2004), the eutherians and marsupials are together known as Theria, and the two lineages split $\sim$ 180 MYA (Murphy *et al.*, 2004). Marsupial and monotreme mammals have many very distinct features, some almost non-mammal like, yet they share some fundamental mammalian characteristics with the eutherians. Comparisons between these most distantly related mammals give us an insight of the rules and variations of mammalian genetics, and allow us to reconstruct the evolutionary past of the mammal lineage.

There are three orders and more than 200 species of marsupial mammals, most found in Australia, with one order found in South America, and only one species in North America (Wilson and Reeder, 2005). Like the eutherians, the marsupials are fur-bearing animals, give birth to live young, and feed the young with milk secreted from their mammary glands, but they have very distinct strategy of reproduction. Although they give live birth, the marsupials lack long gestation periods like the eutherians. Whereas eutherian young are born relatively developed, the marsupial young are born at a very immature stage and complete development attached to a teat, often protected in a pouch (the marsupium).

This kind of variation on a common mammalian theme is also seen in the ge-

netics of the sex chromosome. Marsupial and eutherian mammals share the same XY sex determination system. Comparative studies showed that genes located on the long arm of the human X chromosome are also X-linked in marsupials, indicating a common origin of the sex chromosomes in therian mammals (Spencer *et al.*, 1991b). However, the homology is limited to a subset of genes on the eutherian X chromosome - genes located on the distal short arm of the human X were found to be autosomal in marsupials (Spencer *et al.*, 1991a). Also, the marsupial X chromosome, representing 2-3% of the haploid genome, is much smaller than the eutherian X, which makes up 5% of the haploid genome (Graves and Watson, 1991). It has been postulated that the marsupial X chromosome may reflect the ancient X in the common ancestor of therian mammals, whereas a translocation from autosome added a large block of genes to the eutherian X before the radiation of eutherian mammals (Graves, 1995). The older and younger portions of present day eutherian X have been termed the X conserved region (XCR) and the X added region (XAR) (Figure 1.2). The marsupial Y chromosome, even smaller than the eutherian Y, has also lost most of its gene contents, thus the need for dosage compensation.

Like the eutherians, marsupial mammals also achieve dosage compensation through XCI (Cooper *et al.*, 1971; Richardson *et al.*, 1971; Sharman, 1971), but the exact mechanism of XCI appears to be different. In the marsupials XCI is imprinted, where the paternal X is always silenced (Cooper *et al.*, 1971; Richardson *et al.*, 1971; Sharman, 1971). In the eutherians, XCI is random in somatic cells, and imprinted XCI has only been observed in mouse and bovine extraembryonic tissues (Takagi and Sasaki, 1975; Xue *et al.*, 2002). There has therefore been speculation that imprinted XCI is the ancestral form of XCI in mammals (Huynh and

Lee, 2005). At the molecular level, the master switch gene that initiates random XCI in eutherians, *XIST*, is not present on the marsupial X (Hore *et al.*, 2007; Mikkelsen *et al.*, 2007). A recent search for sequences homologous to the human and mouse *XIST* and the flanking region in a number of vertebrates suggested that this region is disrupted in marsupials and monotremes (Hore *et al.*, 2007). This study also proposed that this unstable region underwent great expansion in the early eutherian evolution and provided start materials for the evolution of the XIC. The recently completed draft genomic sequences of a marsupial (short-tailed opossum) and a monotreme (platypus) have also confirmed the lack of *XIST* on both X chromosomes (Mikkelsen *et al.*, 2007; Warren *et al.*, 2008). The molecular events leading to the initiation and spreading of XCI in marsupials is not understood, but a link has been suggested to the meiotic sex chromosome inactivation (MSCI) in spermatogenesis (Hornecker *et al.*, 2007; Namekawa *et al.*, 2007). MSCI, previously only found in eutherian mammals, was recently identified in the grey short-tailed opossum (Hornecker *et al.*, 2007; Namekawa *et al.*, 2007). In opossum MSCI, the only X chromosome is silenced before fusing with the Y chromosome through a dense plate, and the silencing stably persists after meiosis, suggesting that the inactive state of the paternal X in a zygote may be inherited from the state in the gamete (Hornecker *et al.*, 2007; Namekawa *et al.*, 2007).

Furthermore, XCI in the marsupials appears to be less stable than XCI in the eutherians. Whereas in human and mouse only a small proportion of genes escape XCI (Carrel and Willard, 2005; Disteche *et al.*, 2002), expression of the repressed paternal allele was detected for all examined genes in marsupials (Cooper *et al.*, 1993). Studies of X-linked electrophoretic variants, *i.e.* allozymes, in a number of

marsupial species showed that the expression of paternal allele is very common but highly variable, displaying different patterns for different genes, tissues, and species. One of the most widely studied X-linked loci is *G6PD*, which undergoes stable random inactivation in human and mouse. In the marsupials, the paternal allele of *G6PD* is preferentially inactivated, but various degrees of paternal expression were seen in almost all tissues examined in the Virginian opossum, *Didelphis virginiana* (Samollow *et al.*, 1987). An even higher extent of de-repression of the paternal allele of *G6PD* was found in cultured cells arising from a number of Virginian opossum tissues (Migeon *et al.*, 1989). Again the extent of paternal expression varied greatly among different tissues and even different cell types in the same tissue, from no expression to full expression (Migeon *et al.*, 1989). Another widely studied gene, *PGK1*, also X-inactivated in human and mouse, displays a very different pattern in marsupials from that of *G6PD*. Cooper and colleagues (1977) observed partial expression of the paternal allele of *PGK1* in two Australian marsupials, the eastern grey kangaroo (*Macropus giganteus*) and the pretty-face wallaby (*M. parryi*), but Samollow and colleagues (1987) found only maternal expression in the American marsupial Virginian opossum. The unstable XCI in marsupials has been attributed to lack of maintenance by DNA methylation. In eutherian mammals, the 5' CGIs are usually methylated on the inactive, but not the active, X chromosome (Tribioli *et al.*, 1992). In marsupials limited evidence suggested that CGIs are not methylated on either X chromosome (Kaslow and Migeon, 1987; Loebel and Johnston, 1996). CGI methylation on the marsupial X chromosome is discussed in more detail in Chapter 5.

In human and mouse, XCI is established *via* a cascade of histone modifications, including hypoacetylation of core histones H3 and H4 (Boggs *et al.*, 1996;

Jeppesen and Turner, 1993). To exploit whether the same mechanism is employed in the marsupials, Wakefield and colleagues (1997) labelled metaphase chromosomes of Tammar wallaby fibroblasts with an antibody against acetylated histone H4. The autosomes were all brightly labelled, but the two X chromosomes showed a striking difference in the extent of labelling. While one X in female cells and the single X in male cells were brightly labelled, like the autosomes, the other X in female cells showed significantly weaker labelling, consistent with the inactive X being underacetylated. To date, histone hypoacetylation remains the only molecular feature known to be common to both eutherian and marsupial XCI.

The monotreme mammals, which are the most distantly related to the eutherians, share even less similarities with them. There are currently only five known species of monotremes, the duck-bill platypus, and four species of echidnas, all found in Australia (Wilson and Reeder, 2005). Like the marsupials, the monotremes are also most famous for their distinct mode of reproduction, and are sometimes referred to as the 'egg-laying mammals'. Despite the reptilian-like birth, the young are hatched at a less developed stage than birds and reptiles, and typical of mammals, rely on the milk of their mothers during the initial period of their life. As monotreme females lack nipples, the young suck milk directly through the abdominal skin.

The genetics of monotreme mammals reflects their unique evolutionary position. Although monotremes employ an XY sex determination system, their sex chromosomes are highly unusual, consisting of multiple sex chromosomes that form a translocation chain during meiosis (Rens *et al.*, 2004). The monotreme sex chromosomes appear to have different evolutionary origins from those of the marsupials and eutherians. X-linked genes in the therian mammals map to auto-

somes in platypus, while considerable homology is shared between the platypus X chromosome and the chicken Z chromosome (Veyrunes *et al.*, 2008). In addition, the male-determining gene common to the therian mammals, *SRY*, is not present in the monotremes (Wallis *et al.*, 2007). It is not clear whether X-linked genes in monotreme mammals are dosage compensated, but given the different origin of the monotreme sex chromosomes, there is no obvious reason to expect that XCI will occur as the mechanism in monotremes.

1.5 Aims of the thesis

Genes that escape from XCI present a unique opportunity to improve our understanding of the extraordinary biological process of X chromosome inactivation. To date, the XCI status has only been characterised for genes on the human X chromosome and a small number of genes on the mouse X chromosome. It is desirable to catalogue the XCI status of a greater number of genes in multiple species, as comparisons between different species will greatly help in understanding how the landscape of silencing has evolved.

To achieve this, it is critical to have a method to assess the XCI status of individual genes in multiple species. Currently, the vast majority of XCI data in human were generated by the somatic hybrid method and most studies in mouse have made use of the T16H translocation. Both experimental systems are difficult to establish and not suitable for extending to other species. In human and mouse, CGI methylation on the inactive X has been implicated in the maintenance of XCI and has been used previously to indicate a gene's XCI status. Therefore, I have chosen to study the 5' CGI methylation states of a larger number of genes

as surrogate for their XCI status.

The four main aims in this thesis are:

1. To test whether CGI methylation serves as a good indicator for a gene's XCI status by comparison of methylation state with known XCI profile for a number of genes in human and mouse (Chapter 3).

2. To search for differences in CGI methylation of X-linked genes between human and mouse; and to explore whether such differences, if existing, are parallel to the differences in XCI profiles between the two species (Chapters 3 and 4).

3. To investigate the existence or absence of CGI methylation of X-linked genes in marsupials, as the small amount of available evidence points to lack of methylation on the X_i (Chapter 5).

4. To compare characteristics of CGIs in the homologous region of X chromosome between human, mouse, and opossum (Chapter 6).

Chapter 2

Materials and Methods

Materials

2.1 Chemical reagents

All common chemicals were purchased from Sigma and Invitrogen unless specified below or in the text.

Amersham Biosciences	2'-deoxynucleoside 5' triphosphates
Invitrogen	Agarose
	Knockout™ DMEM
	Foetal Bovine Serum (FBS)
	Trypsin-EDTA
Melford	5-bromo-4-chloro-3-indolyl-β-D-galactoside (X-Gal)
New England Biolabs	Bovine Serum Albumin (BSA)
Sigma	β-mercaptoethanol
	Dimethyl sulfoxide (DMSO)

Sodium acetate (NaOAc)

Sodium dodecyl sulfate (SDS)

Formamide

Phenol

Phenol:chloroform:isoamylalcohol, 25:24:1

Chloroform:isoamylalcohol, 24:1

Phosphate buffered saline, pH7.2 (PBS)

Minimum Essential Medium Eagle (MEME)

Penicillin/streptomycin/L-glutamine

2.2 Enzymes and commercially prepared kits

All restriction endonucleases were purchased from New England Biolabs (NEB).

Applied Biosystems	AmpliTaq® DNA polymerase
	AmpliTaq Gold® DNA polymerase
BD Biosciences	BD Advantage™ 2 PCR kit
Eppendorf	Phase Lock Gel (PLG)
Invitrogen	Proteinase K
Promega	LigaFast™ rapid DNA ligation system
QIAGEN	DNeasy® blood & tissue kit
	Gel extraction kit
	PCR purification kit
Zymo Research	EZ DNA Methylation Kit™

2.3 Solutions and buffers

Solutions used in this thesis are listed in alphabetical order below. Final concentrations of reagents are given for all solutions. Unless otherwise specified, solutions were made in nanopure water.

2.3.1 Buffers

6x Agarose Gel Loading Dye

0.25% Bromophenol blue

0.25% Xylene cyanol

25% Glycerol

10x NEB PCR buffer

670 mM Tris-HCl (pH 8.8)

166 mM $(NH_4)_2SO_4$ (enzyme grade)

67 mM $MgCl_2$

Adjust pH to 8.8 with HCl

10x TBE

890 mM Tris Base

890 mM Borate

20 mM EDTA (pH 8.0)

1x TE

10 mM Tris-HCl (pH 7.4)

1 mM EDTA

1x $T_{0.1}E$

10 mM Tris-HCl (pH 7.5)

0.1 mM EDTA

1x SET

10 mM Tris-HCl (pH 7.5)

100 mM $NaCl_2$

1mM EDTA (pH 8.0)

28% Sucrose/Cresol red solution

1x $T_{0.1}E$

28% (w/v) Sucrose

0.008% (w/v) Cresol red

2.3.2 Media

All media were prepared in nanopure water and either autoclaved or filter-sterilised prior to use. When used for bacterial growth, 15 mg/ml bacto-agar were added to the appropriate media. Where appropriate ampicillin (dissolved in 1 M sodium bicarbonate, stored at -20 °C) was added to media at a final concentration of 75

μg/ml.

LB

10 mg/ml Bacto-tryptone

5 mg/ml yeast extract

10 mg/ml NaCl

2% w/v dextrose (pH 7.4)

LB/amp/X-Gal plates

LB media

15 g/l agar

75 μg/ml ampicillin

100 μg/ml X-Gal

2.4 Size markers

E-Gel® Low Range Quantitative DNA Ladder (Invitrogen)

Consists of five linear double-stranded DNA fragments of the following sizes in bp: 100, 200, 400, 800, 2000. Electrophoresis of 10 μl of the ladder results in bands containing 5, 10, 20, 40, and 100 ng of DNA, respectively.

1 kb ladder (1 mg/ml) (Invitrogen)

Contains 1 to 12 repeats of a 1018 bp fragment and vector fragments from 75 bp to 1636 bp to produce the following sized fragments in bp: 75, 134, 154, 201, 220, 298, 344, 394, 506/517, 1018, 1636, 2036, 3054, 4072, 5090, 6108, 7125, 8144,

9162, 10180, 11198, 12216. The 1636 bp band contains 10% of the mass applied to the gel.

100 bp DNA Ladder (Invitrogen)

Consists of 15 blunt ended fragments between 100 and 1500 bp in multiples of 100 with an additional fragment at 2072 bp.

2.5 Tissues and cell lines

Human fibroblasts

Female human fibroblast line GM01122 and male human fibroblast line GM01237 (Coriell Cell Repositories: NIGMS Human Genetic Cell Repository), obtained from a brother-sister pair, were cultured in MEME supplemented with 100 U penicillin, 0.1 mg/ml streptomycin, 2 mM L-glutamine, and 15% FBS.

Mouse embryonic fibroblasts

Female mouse embryonic fibroblast line MEF4 and male mouse embryonic fibroblast line MEF2, prepared by Dr Colette Johnston, were cultured in Knockout™ DMEM supplemented with 100 U penicillin, 0.1 mg/ml streptomycin, 2 mM L-glutamine, and 7% FBS.

Opossum tissue samples

Opossum tissue samples were obtained from Dr Sarah Mackay in Glasgow University. Liver, spleen, kidney and heart samples were snap frozen in liquid nitrogen immediately post-mortem, shipped in dry ice and stored at -70 °C.

2.6 Primer sequences

All primer sequences used in this thesis are listed in Appendices I and II. Primers were synthesised by either Sigma or Invitrogen.

2.7 Key World Wide Web addresses

Table 2.1: Key World Wide Web addresses used in this study

Website	Address
EBI- ClustalW2	http://www.ebi.ac.uk/Tools/clustalw2/index.html
EBI - MUSCLE	http://www.ebi.ac.uk/Tools/muscle/index.html
Ensembl genome browser	http://www.ensembl.org
Gap4	http://staden.sourceforge.net/manual/gap4_unix_toc.html
GeneDoc	http://www.nrbsc.org/
NCBI	http://www.ncbi.nlm.nih.gov/
Primer3	http://frodo.wi.mit.edu/primer3/input.htm
RepeatMasker	http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker
The Human Epigenome Project	http://www.epigenome.org
The Wellcome Trust Sanger Institute	http://www.sanger.ac.uk/
UCSC genome browser	http://genome.ucsc.edu/
Vega	http://vega.sanger.ac.uk/
VISTA	http://genome.lbl.gov/vista/index.shtml

Methods

2.8 Mammalian cell culture

The mammalian cells used in this study and their culture conditions are detailed in section 2.5.

2.8.1 Resuscitating frozen cells

1. A vial containing a frozen aliquot of the cells was taken from liquid nitrogen and rapidly defrosted in a 37 °C water bath.

2. Using a plastic Pasteur pipette, contents of the vial were quickly transferred to a 75 cm^2 flask with warm media.

3. Cells were monitored for growth and split after 2-3 days (see section 2.8.2).

2.8.2 Growing and harvesting cells

All cell lines were adherent and were grown at 37°C, 5% CO_2 in 75 cm^2 flasks (Cornings). Cells were passaged when their density reached greater than 80% confluency. These fibroblasts were typically split at 1:3 and required splitting every 3-4 days.

1. The media from the culture flask was aspirated into 1% Virkon.

2. The cells were washed twice in 10 ml PBS.

3. 2 ml of pre-warmed Trypsin-EDTA were added.

4. The flask was incubated at 37 °C for 2 minutes and the cells were dislodged by sharply tapping the flask 2-3 times.

5. Trypsinisation was stopped by addition of 8 ml of pre-warmed growth media with gentle pipetting to resuspend the cells.

6. 4 ml of cell suspension were transferred to each fresh flask containing 12 ml pre-warmed growth media.

If necessary, cells were counted using a haemocytometer (Sigma) with a 0.1 mm sample depth and light microscope (Olympus).

For harvesting, cells were transferred to a 50 ml Falcon tube after step 5 and pelleted by centrifugation at 1200 rpm for 5 minutes. Cell pellets were washed with 5 ml PBS, re-pelleted, and resuspended in appropriate buffers or stored at -70 °C.

2.8.3 Freezing cells for storage

1. Cells from one 75 cm^2 flask were harvested as described in section 2.8.2 and resuspended in 2 ml of freezing medium containing 75% growth medium, 20% FBS, and 5% DMSO.

2. 1 ml of cell suspension was aliquoted into each cryo vial (Nunc).

3. Vials were frozen at 1°C per minute down to -70°C for 24 hours.

4. Once frozen the cells were transferred into liquid nitrogen for long-term storage.

2.9 DNA extraction

2.9.1 Phenol/chloroform extraction of genomic DNA from animal tissues

The DNA solutions were mixed with organic solvents using a blood tube rotator (Stuart) at 20 rpm for 10 minutes. All centrifugation steps were carried out in a Sorvall Legend RT centrifuge at 3000 rpm for 5 minutes. All steps were performed at room temperature unless otherwise stated.

1. A slice of tissue was weighed in a 1.5 ml eppendorf tube. Up to 500 mg of tissue was used for one preparation.

2. The tissue slice was ground to a fine powder in liquid nitrogen using a mortar and pestle and gently resuspended in 4.5 ml 1x SET, 250 μl 10% SDS and 100 μl proteinase K (10 mg/ml) in a 50 ml Falcon tube.

3. The mixture was shaken at 90 rpm at 55 °C overnight in a shaking incubator.

4. 5 ml phenol (TE saturated) were added to the mixture and mixed by gentle inversion. The sample was rolled for 10 minutes to mix, followed by 5 minutes of centrifugation to separate the two phases.

5. Using a Pasteur pipette, the aqueous layer was carefully transferred to a fresh 50 ml Falcon tube containing 10 ml phenol:chloroform:isoamylalcohol (25:24:1).

6. 5 ml 1x SET were added to the organic layer in the original tube, mixed by gentle inversion and submitted to centrifugation.

7. The aqueous layer was added to the phenol:chloroform:isoamylalcohol mix from the previous step, mixed by gentle inversion and submitted to centrifugation.

8. The aqueous layer was combined with 10 ml chloroform:isoamylalcohol (24:1), mixed by gentle inversion and submitted to centrifugation.

9. The extraction with chloroform:isoamylalcohol was repeated (steps 7 and 8), and the final aqueous layer was transferred to a fresh tube containing 25 ml of ice cold absolute ethanol and 1 ml 3M NaOAC (pH 6).

10. The tube was hold horizontally and gently tipped from side to side to mix, and then incubated at -20 °C for one hour or longer.

11. Glass hooks for spooling DNA were made by holding an unplugged glass Pasteur pipette in a gas flame until the glass melts and bends into a hook.

12. DNA was spooled out using a sterile glass hook, dipped into a 1.5 ml eppendorf tube containing 70% ethanol to rinse, then rinsed again in a second tube containing fresh 70% ethanol.

13. The excess liquid was drained off and the DNA was dissolved in 500 μl $T_{0.1}E$.

2.9.2 Phenol/chloroform extraction of DNA from a reaction mix

The phase-lock gel (PLG) heavy was used to facilitate extraction. DNA was extracted from 100-500 μl of reaction mix in one 1.5 ml PLG tube.

1. PLG tubes were spun at 13000 rpm for 1 minute immediately prior to use.
2. The sample and an equal volume of phenol:chloroform:isoamylalcohol (25:24:1) were added to the top of the gel and mixed by vigorously shaking the tube.
3. The tube was spun at 13000 rpm for 5 minutes to separate the phases.
4. The aqueous layer above gel was transferred to a fresh eppendorf tube for ethanol precipitation (section 2.10.1).

2.9.3 Mini-preps of genomic DNA from cultured cells

Genomic DNA was extracted from cultured cells using the DNeasy blood & tissue kit (QIAGEN) in accordance with the manufacturer's protocol. One fresh or frozen pellet containing 1x 10^6 cells (section 2.8.2) was used for each preparation.

2.10 DNA manipulation

2.10.1 Purification of DNA

Ethanol precipitation

1. 0.1 volumes of 3M sodium acetate and 2.5 volumes of 100% ethanol were added to the DNA sample.
2. The samples were well mixed by vortexing and incubated at -20 °C for 30 minutes.
3. The DNA was pelleted by centrifugation for 20 minutes at 13000 rpm using a bench top centrifuge.

4. The pellet was washed with 1 volume of 70% ethanol, followed by centrifugation for 5 minutes at 13000 rpm.

5. DNA pellets were air-dried and resuspended in $T_{0.1}E$.

Direct PCR product purification

PCR products (10 to 50 μl) were purified for subsequent analysis using the PCR purification kit (QIAGEN) in accordance with the manufacturer's protocol.

Agarose gel purification

1. The appropriately sized DNA fragment was excised from the gel using a clean scalpel.

2. The gel slice was weighed in a 1.5 ml eppendorf tube.

3. Purification proceeded using the gel extraction kit (QIAGEN) in accordance with the manufacturer's instructions.

2.10.2 Quantification

DNA was quantified by applying the Beer-Lambert equation relating absorbance, and extinction co-efficient to DNA concentration. Absorbance readings were measured at 260 nm and the extinction coefficients used were 50 for dsDNA and 33 for ssDNA. Absorbance readings were taken on either Eppendorf Biophotometer (Eppendorf) or NanoDrop spectrophotometer (NanoDrop Technologies).

2.10.3 Restriction enzyme digests

1. Restriction digests of genomic DNA (up to 20 μg) were completed using the appropriate buffer and 10-50 units of enzyme per μg DNA (10 units per μg DNA for bisulphite treatment, 50 units per μg DNA for the restriction-PCR methylation assay).

2. Samples were incubated at 37°C overnight.

2.10.4 Agarose gel electrophoresis

1. An agarose gel was prepared (2.5% for most PCR amplified products and 1% for fragments over 1 kb) by melting agarose in 1x TBE. Ethidium bromide was added to a concentration of 250 ng/ml.

2. The appropriate amount of loading buffer was added to the DNA depending on volume.

3. The appropriate amount of samples and appropriate size markers (see section 2.4) were loaded into each well.

4. Electrophoresis was performed in 1x TBE with a voltage ranging between 50-200 volts (5 volts per cm from electrode to electrode) for the time required to obtain satisfactory separation, typically 45 minutes.

5. DNA was visualised under UV light on a transilluminator and digitally photographed using LabWorks Image Acquisition and Analysis Software (UVP Bioimaging Systems).

2.10.5 Bisulphite modification

Bisulphite modification of DNA was performed in a 96-well microtitre plate (Costar Thermowell^{TM} M-type) in a PTC-225 (MJ Research) thermal cycler. All reagents (except $T_{0.1}E$) were from the EZ DNA methylation kit (Zymo Research). Samples were processed in batches of 10.

1. 20 μg of genomic DNA was digested with HindIII (section 2.10.3) and purified using phenol/chloroform extraction (section 2.9.2) followed by ethanol precipitation (section 2.10.1). Digested DNA was diluted to 27 ng/μl in $T_{0.1}E$.

2. One tube of CT Conversion Reagent was dissolved in 750 μl water and 210 μl M-Dilution buffer, and mixed thoroughly by vortexing every 1-2 minutes for 10 minutes.

3. In each of wells 1-10 of row A of a 96-well plate, 45 μl of digested DNA were combined with 5 μl M-Dilution buffer.

4. The plate was incubated in a pre-heated thermal cycler at 37 °C for 15 minutes.

5. The 50 μl of sample in row A were split into five aliquots of 10 μl in rows A-E, to which 20 μl of prepared CT Conversion Reagent were quickly added.

6. Bisulphite modification was performed with the cycling profile: 20 cycles at 95 °C for 30 seconds and 50 °C for 15 minutes.

7. Following bisulphite modification the plate was immediately incubated on ice for 10 minutes.

8. Samples from each column (*e.g.* well 1 from rows A-E) were pooled into a 1.5 ml eppendorf tube, to which 400 μl of M-Binding buffer were added.

9. Desulphonation was carried out in accordance with the manufacturer's instructions.

10. 50 μl $T_{0.1}E$ were added to the 10 μl eluted DNA sample.

11. Concentrations of the samples were measured using a NanoDrop spectrophotometer set at ssDNA (section 2.10.2). Samples were diluted to 10 ng/μl with $T_{0.1}E$ prior to PCR (section 2.11.3). If the concentration was lower than 10 ng/μl, the sample was used directly in PCR.

2.11 Polymerase Chain Reaction (PCR)

PCR was performed in a 96-well microtitre plate or in 0.2 ml tubes (ABgene) in a PTC-225 thermal cycler (MJ Research).

1. A premix sufficient for the number of planned reactions was prepared, allowing for a 1x reaction mix once the DNA template was added. The volume and contents of the final reaction are described below.

2. PCR amplications were performed with the specific cycling profile as described below.

3. Reaction products were visualised by agarose gel electrophoresis and stained with ethidium bromide (section 2.10.4).

2.11.1 Primer design

Primers were designed using Primer3. Where possible primers were selected using the following parameters:

- Melting temperature between 57°C and 63°C
- (G+C) content between 30-80%
- Length between 18 - 22 bp
- Less than 2°C difference in melting temperature between the two primers

Wherever necessary, the parameters were relaxed to the following:

- Melting temperature between 55°C and 65°C
- (G+C) content between 30-80%
- Length between 18 - 27 bp
- Less than 2°C difference in melting temperature between the two primers

Primer pairs that would amplify fragments containing HindIII cleavage sites (AAGCTT) were excluded. BLAST searches (section 2.14.2) were performed to test the specificity of the primer pair.

2.11.2 Restriction-PCR Methylation Assay (RPMA)

RPMA protocol a)

1. PCR mix was prepared containing the following: 1x Advantage 2 PCR buffer, 5% DMSO, 0.4 mM of each deoxyribonucleoside triphosphate (dATP,

dCTP, dGTP, dTTP), 0.67 μM of each primer, 0.5x Advantage 2 polymerase mix and 0.75 U of AmpliTaq Gold DNA polymerase, in a final reaction volume of 10 μl and aliquoted into wells of a 96-well microtitre plate.

The primers used were designed to amplify 150-250 bp fragments containing (ideally 2-3) HpaII/MspI cleavage sites (CCGG) within 5' CpG islands. The expected amplified sequences were checked for number of CCGG sites and absence of HindIII cleavage (AAGCTT) sites using a Perl script (see Appendix I for sequences).

2. Initially the PCR mix was made without the deoxyribonucleoside triphosphate and the polymerases. 5 μl of restriction-enzyme-digested genomic DNA (10 ng/μl) (section 2.10.3) was added to the previously prepared PCR mix and incubated at 100 °C for 5 minutes.

3. Then the rest of ingredients were added and PCR was performed using the cycling profile: 94 °C for 2 minutes, followed by 35 cycles at 94 °C for 30 seconds, 65 °C for 30 seconds and 72 °C for 1 minute.

4. Success of amplification was assessed using gel electrophoresis and ethidium bromide staining (section 2.10.4).

RPMA protocol b)

1. PCR mix was prepared containing the following: 1x GeneAmp PCR Gold buffer, 1.5 mM $MgCl_2$, 0.2 mM of each deoxyribonucleoside triphosphate, 0.4 μM of each primer, and 0.75 U of AmpliTaq Gold DNA polymerase,

in a final reaction volume of 20 μl and aliquoted into wells of a 96-well microtitre plate.

The primers used were as described in protocol a).

2. 5 μl of restriction-enzyme-digested genomic DNA (10 ng/μl) (section 2.10.3) was added to the previously prepared PCR mix.

3. PCR was performed using the cycling profile: 95 °C for 10 minutes, followed by 35 cycles at 95 °C for 30 seconds, 60 °C for 30 seconds and 72 °C for 1 minute, and finally followed by 1 extension cycle at 72 °C for 2 minutes.

4. Success of amplification was assessed using gel electrophoresis and ethidium bromide staining (section 2.10.4).

2.11.3 Bisulphite PCR

1. PCR mix was prepared containing the following: 1x GeneAmp PCR Gold buffer, 1.5 mM $MgCl_2$, 0.2 mM of each deoxyribonucleoside triphosphate, 0.4 μM of each primer, and 0.75 U of AmpliTaq Gold DNA polymerase, in a final reaction volume of 10 μl and aliquoted into wells of a 96-well microtitre plate.

   The primers used were designed to amplify 400-800 bp fragments within 5' CpG islands. Primers were designed to *in silico* bisulphite-converted sequences and contained no CpGs (see Appendix II for sequences).

2. 5 μl of bisulphite modified genomic DNA (section 2.10.5) was used as template and added to the previously prepared PCR mix.

3. PCR was performed using the cycling profile: 95 °C for 10 minutes, followed by 3 cycles at 94 °C for 45 seconds, 60 °C for 1 minute and 72 °C for 1 minute 30 seconds, 3 cycles at 94 °C for 45 seconds, 59 °C for 1 minute and 72 °C for 1 minute 30 seconds, 3 cycles at 94 °C for 45 seconds, 58 °C for 1 minute and 72 °C for 1 minute 30 seconds, and 35 cycles at 94 °C for 45 seconds, 57 °C for 1 minute and 72 °C for 1 minute, where 5 seconds was added to the 72 °C extension step each cycle, and finally followed by 1 extension cycle at 72 °C for 15 minutes.

4. Success of amplification was assessed using gel electrophoresis and ethidium bromide staining (section 2.10.4).

2.11.4 Colony PCR

1. PCR mix was prepared containing the following: 1x NEB PCR buffer, 7 ng/μl of β-mercaptoethanol, 330 μg/ml of BSA, 0.5 mM of each deoxyribonucleoside triphosphate, 0.75 μM of each primer, and 1 U of AmpliTaq DNA polymerase, in a final reaction volume of 15 μl and aliquoted into wells of a 96-well microtitre plate.

   The primers used were M13 forward and reverse (see Appendix II for sequences).

2. Following bacterial transformation (section 2.12.2) individual white colonies were picked using a sterile inoculation loop (VWR) and dipped into 15 μl of the previously prepared PCR mix.

3. PCR was performed using the cycling profile: 94 °C for 5 minutes, followed

by 30 cycles at 94 °C for 30 seconds, 60 °C for 30 seconds and 72 °C for 1 minute 30 seconds, and finally followed by 1 extension cycle at 72 °C for 5 minutes.

4. Success of amplification was assessed using gel electrophoresis and ethidium bromide staining (section 2.10.4).

2.12 Bacterial cloning

2.12.1 Ligation of vector and insert

The pGEM®-T Easy vector was used in all cloning in this thesis. The inserts were all PCR products amplified using Taq polymerase so no A-tailing was necessary.

1. A 10 μl ligation reaction was performed containing 3 units of T4 DNA ligase, 1x ligation buffer, 20 ng vector, and 2 μl insert DNA.

2. The reaction was mixed and incubated for 20 minutes at room temperature or overnight at 4 °C.

2.12.2 Transformation

1. 50 μl of competent Mach1^{TM} cells (Invitrogen) were thawed on ice.

2. 2 μl of ligation mix (section 2.12.1) were added to the cells and mixed by gently flicking the tube.

3. The mixture was incubated on ice for 30 minutes.

4. The cells were heat-shocked at 42 °C in a water bath for 45 seconds, followed by a 2 minute incubation on ice.

5. 500 μl of LB broth were added to each transformation reaction, which was then shaken at 225 rpm at 37 °C for an hour.

6. The transformation mix was then plated out in varying quantities (typically 50 μl) onto agar plates containing ampicillin and X-Gal.

7. The plates were incubated overnight at 37 °C.

8. Success of transformation was determined by colony PCR (section 2.11.4).

2.13 Plasmid sequencing

All plasmid end-sequencing was performed by the Sanger Institute Faculty Small Sequencing Projects (FSSP) team.

2.14 Computational analysis

2.14.1 cpg

The presence of CpG islands were predicted by Val Curwen using the program cpg written by Gos Micklem (unpublished, www.ccbi.cam.ac.uk/software/cpg).

2.14.2 BLAST

Web-based BLAST analysis was performed at Ensembl (`http://www.ensembl.org/`) or Vega (`http://vega.sanger.ac.uk/`).

2.14.3 RepeatMasker

In the bisulphite sequencing analysis, prior to *in silico* bisulphite conversion and primer design, human and mouse repeat sequences were masked using RepeatMasker (Smit and Green, unpublished) (`http://www.repeatmasker.org`).

The repeats were soft-masked with lower case letters in the output file, and marked with underline for subsequent *in silico* bisulphite conversion and primer design.

2.14.4 GAP4

In the bisulphite sequencing analysis, the quality of DNA sequences was assessed using Gap4, a program for sequence assembly and editing. The contig editor used phred confidence values to calculate the confidence of the consensus sequence and identify places requiring visual trace inspection. Vector sequences were automatically located and tagged in the process of assembly. Additional information about Gap4 can be found at: `http://staden.sourceforge.net/overview.html`.

2.14.5 Alignment of DNA sequences

DNA sequences were aligned using the program ClustalW2 (Larkin *et al.*, 2007) or MUSCLE (Labarga *et al.*, 2007) via a web-based server at the EBI (`http://www.ebi.ac.uk/`). User-defined parameters were left at their default settings. For bisulphite sequencing results the alignments were then manually edited and presented using the program GeneDoc (Nicholas 1997, unpublished) (`http://www.nrbsc.org/`).

2.14.6 Multi-species comparative sequence analysis

Web-based multi-species comparative sequence analysis was performed using the VISTA server (Frazer *et al.*, 2004) (`http://genome.lbl.gov/vista/`).

2.14.7 MethTools

Methylation analysis was carried out using MethTools, a collection of open source Perl scripts that make graphical representation of bisulphite sequencing results and calculate methylation densities (Grunau *et al.*, 2000) (`http://genome.imb-jena.de/methtools/`). Customised modifications were made to the original scripts as appropriate and noted in the text. The modified scripts are included in Appendix IV.

2.14.8 Perl scripts

The following Perl scripts were generated to aid analysis of large datasets. These scripts were included in Appendix V.

extract_cpg_seq Extracts sequence of CpG island from the Ensembl or Vega database

parse_primers Checks RPMA primers for absence of HindIII cleavage sites and counts number of HpaII/MspI cleavage sites

primer_invitrogen Converts FASTA format primer file to a primer list compatible with Invitrogen primer plate order form

CGI_CGcount Calculates GC content of the XCR of human, mouse and opossum

CGI_compare_1 Searches for 5' CpG island and orthologous genes for

human, mouse, and opossum XCR genes

CGI_compare_2 Extracts sequence of desired CpG islands from a list

Chapter 3

Comparison of CpG island methylation on the mouse and human X chromosomes

3.1 Introduction

The vertebrate genomes are depleted of the dinucleotide CpG, which is only observed at 0.2-0.25 of the frequency expected from base composition (Russell *et al.*, 1976; Swartz *et al.*, 1962). The majority of the CpGs are methylated at the carbon-5-position of the cytosine (Bird and Taggart, 1980; Ehrlich *et al.*, 1982) and their rarity has been attributed to the tendency for 5mCpG to mutate by deamination to TpG (and CpA in the complementary strand) (Coulondre *et al.*, 1978). However, when genomic DNA from a number of vertebrates was digested with 5mCpG-sensitive restriction enzymes, it was found that about 1% of the genome, the HTF (*Hpa*II tiny fragments) fraction, is highly rich in unmethylated

Chapter 3

Comparison of CpG island methylation on the mouse and human X chromosomes

3.1 Introduction

The vertebrate genomes are depleted of the dinucleotide CpG, which is only observed at 0.2-0.25 of the frequency expected from base composition (Russell *et al.*, 1976; Swartz *et al.*, 1962). The majority of the CpGs are methylated at the carbon-5-position of the cytosine (Bird and Taggart, 1980; Ehrlich *et al.*, 1982) and their rarity has been attributed to the tendency for 5mCpG to mutate by deamination to TpG (and CpA in the complementary strand) (Coulondre *et al.*, 1978). However, when genomic DNA from a number of vertebrates was digested with 5mCpG-sensitive restriction enzymes, it was found that about 1% of the genome, the HTF (*Hpa*II tiny fragments) fraction, is highly rich in unmethylated

*Hpa*II and *Hha*I sites (Cooper *et al.*, 1983). Bird and colleagues (1985) then mapped three randomly chosen HTFs to 'islands' of DNA that are CpG-rich and unmethylated in a range of mouse tissues. At the same time, several studies identified unmethylated regions at the 5' end of genes (McKeon *et al.*, 1982; Stein *et al.*, 1983) and sequence analysis revealed that these regions are also CpG-rich (Tykocinski and Max, 1984). Having reviewed the mounting evidence, Bird (1986) proposed to define HTF-like sequences simply by their GC content and CpG frequency, and predicted that such HTF-like islands would be unmethylated. These CpG-rich islands were found near the transcription origin of many genes, especially housekeeping genes (Bird, 1986). This trend was confirmed by Larsen and colleagues, whose extensive sequence analysis study found that all known human housekeeping genes and some tissue-specific genes are associated with CpG islands (CGIs) covering their transcriptional start (Larsen *et al.*, 1992).

CGI methylation has been implicated in the transcriptional silencing of genes. Keshet and colleagues (1985) built constructs of the hamster *aprt* and *tk* genes with different portions methylated, and studied the effects of methylation on their expression in transfected mouse cells. They found that methylation in the 5' promoter region, which was identified by other researchers as a CGI (Lewis, 1986; Tykocinski and Max, 1984), is sufficient to inhibit transcription for both genes. Studies of genes on the X chromosome in mammals provided further support for the inhibitory role of 5' CGI methylation. As an effect of XCI, the same gene presents an active and an inactive copy in the same cell environment, and experimental systems have been developed to distinguish the two: human-rodent somatic cell hybrid containing only the active or inactive human X chromosome, and *Mus musculus* x *M. caroli* mouse, in which the *M. caroli* X chromosome is

always inactivated. With the help of a de-methylation agent, 5-azacytidine, researchers induced re-expression of human *HPRT* and mouse *Hprt* genes from the inactive X chromosome (Graves, 1982; Mohandas *et al.*, 1980). Characterisation of methylation profiles of the 5' region of the human *HPRT* and mouse *Hprt* genes revealed that this region is methylated on the inactive X, but unmethylated on the active and re-activated X in females, and as expected, unmethylated on the single active X in males (Lock *et al.*, 1986; Wolf *et al.*, 1984; Yen *et al.*, 1984). Indeed, hypermethylated 5' CGIs, although very rare in mammalian genomes, are commonly seen on the inactive X chromosome (Tribioli *et al.*, 1992). In healthy individuals, hypermethylated 5' CGIs have also been identified for imprinted genes, which only express from one chromosome depending on the parental origin (Stöger *et al.*, 1993; Tremblay *et al.*, 1995). For example, methylation of the 5' region was found in sperms but not oocytes for the maternally expressed *H19* gene in mouse, and this pattern is maintained in the pre-implantation embryo (Tremblay *et al.*, 1995). Disruption of normal CGI methylation states can have serious health consequences. A systematic study of methylation states of more than a thousand CGIs in nearly a hundred primary human tumours found tumour-type-specific aberrant methylation patterns (Costello *et al.*, 2000). Using a candidate gene approach, another report showed that one or more of the 12 key cancer genes investigated had promoter hypermethylation in each of the 15 major tumour types studied, and there exist unique patterns for different tumour types (Esteller *et al.*, 2001). The exact mechanism by which CGI methylation induces transcriptional silencing remains unclear but it is possible that methylation at 5' CGI prevents initiation of transcription by blocking access of key DNA-binding factors (Comb and Goodman, 1990).

In the search for regulatory factors of XCI, CGI methylation, which provides effective in *cis* silencing and can be stably inherited through cell division, became an attractive candidate. Initially, Riggs (1975) proposed a model in which DNA methylation regulates the activity of the X inactivation centre and thus plays a role in initiating XCI. However, the discovery that the opossum *G6PD* gene, X-linked and dosage-compensated like its mouse companion, is hypomethylated at its 5' CGI even in females led researchers to believe that DNA methylation stabilises rather than initiates the process (Kaslow and Migeon, 1987). Study of early embryonic development in mouse confirmed that methylation takes place after the initiation of XCI, supporting the maintenance role of CGI methylation in XCI (Lock *et al.*, 1987).

3.1.1 Identification of CGIs

In their initial description of the HTF islands, Bird and colleagues (1985) found several distinct features of these islands: they are GC-rich, not deficient of CpG (with similar frequencies of CpG and GpC), and unmethylated. These islands, later known as CpG islands, were identified by isolation and analysis of small fragments generated by methylation-sensitive restriction enzyme digestion (Antequera and Bird, 1993; Bird *et al.*, 1985). In his landmark paper, Bird (1986) formulated a simple test for HTF-like sequences based on CpG frequency: if a sequence has GC content of over 50% and has similar numbers of CpG and GpC, both near the expected frequency from the sequence composition, the region is HTF-like and is predicted to be unmethylated. Using this test, he found that CGIs are frequently located at the 5' region of genes, most of which are house-

keeping (Bird, 1986). This strong association with gene was confirmed in larger studies of more CGIs and made CGI a popular marker in the search of novel gene sequences (Antequera and Bird, 1993; Larsen *et al.*, 1992).

With the rapid accumulation of sequence data in public databases, computer algorithms were soon developed to search for CGIs. The first algorithm was proposed by Gardiner-Garden and Frommer (1987) who defined a CGI as a GC-rich region (GC content >50%) greater than 200 bp in length and having an observed CpG / expected CpG ratio (Obs_{CpG}/Exp_{CpG}) of greater than or equal to 0.6. Using these criteria, they screened vertebrate genomic sequences containing genes transcribed by RNA polymerase II in GenBank (a database of all publicly available DNA sequences). Most of the CGIs identified are associated with the 5' end of genes and most of the 5' CGIs extend well into genes. Consistent with the other investigations, they also found all housekeeping genes in their survey to have 5' CGIs. The CGIs are highly CpG-rich, with average GC content over 60 and Obs_{CpG}/Exp_{CpG} over 0.8. In addition, although 200 bp was used as the cut-off size of CGI, most of the CGIs identified are greater than 500 bp.

These criteria have been widely applied in prediction of CGIs, but many repetitive elements, for example human Alus, may also fulfil the criteria, leading to a high rate of false positive predictions (Schmid, 1998). In an attempt to more accurately identify CGIs, more stringent parameters than the original Gardiner-Garden and Frommer definition were favoured in some more recent studies. The Vertebrate Genome Annotation (Vega) database, which contains high quality manual annotation of finished vertebrate genome sequences, uses a higher cut-off size of 400 bp, optimised using known CGI sequences (Gos Micklem, personal communications). A slightly stricter set of parameters were proposed by Takai

and Jones (2002) after analysing CGIs in human chromosomes 21 and 22. They defined a CGI as a stretch of DNA of greater than 500 bp length with a GC content equal to or greater than 55% and $\mathrm{Obs}_{\mathrm{CpG}}/\mathrm{Exp}_{\mathrm{CpG}}$ of 0.65 or more. CGIs identified by the above two methods are more likely to be associated with the 5' regions of genes.

3.1.2 Approaches to studying CGI methylation

As DNA methylation is recognised as an important component in gene regulation, the field of methylation is attracting growing interest of researchers around the world, especially in cancer biology. A great range of methodologies to study the phenomenon have been developed in the past two decades. To date, most efforts in resolving states of DNA methylation were enabled by methods based on restriction enzyme digestion or bisulphite conversion. The various methods reviewed below are summarised in Figure 3.1.

In the early days of DNA methylation research, genomic DNA was digested with methylation-sensitive restriction enzymes and Southern blotting was used to provide readout of methylation status. In the first studies to examine *de novo* methylation of 5' CGI as a potential mechanism of oncogenesis, DNA samples from primary retinoblastomas and unaffected controls were digested with *Sac*I and the methylation-sensitive *Hpa*II, and a restriction map was established by Southern hybridisation of a probe covering the 5' region of the retinoblastoma gene (*RB*) (Greger *et al.*, 1989; Sakai *et al.*, 1991). If all *Hpa*II sites are methylated, the probe would detect a large fragment bounded by SacI sites, while smaller fragments would indicate some *Hpa*II sites being unmethylated, thus

cleaved. Using methylation-sensitive restriction enzymes with different recognition sites, a number of CpGs were assayed, and hypermethylation at the *RB* CGI was found in one out of 21 (Greger *et al.*, 1989) and five out of 56 retinoblastomas (Sakai *et al.*, 1991). Employing a similar approach, abnormal hypermethylation in tumours were identified at the 5' CGI of more tumour-suppressor genes, including the von Hippel-Lindau (*VHL*) gene (Herman *et al.*, 1994) and the *p16* gene (Merlo *et al.*, 1995). This approach is simple and powerful for examining DNA methylation patterns, but only a subset of CpGs are suitable for assaying, limited by the variety of methylation-sensitive restriction enzymes. Furthermore, using Southern hybridisation as a detection method relies on availability of suitable probes and demands large quantity of test DNA, which may not always be accessible.

The advent of the polymerase chain reaction (PCR) has provided a powerful tool for molecular biology and several methods combining restriction enzymes with PCR have been employed in methylation analysis. The sensitivity of PCR allowed Singer-Sam and colleagues (1990) to study the timing of DNA methylation on the X chromosome in early mouse embryos, when the embryo consists of fewer than a thousand cells. DNA extracted from individual mouse embryos was digested with *Hpa*II and used as template in PCR amplification with a pair of primers enclosing an *Hpa*II site in the 5' CGI of the X-linked *Pgk1* gene. A PCR product is only generated when the site is protected by methylation. They found that this site became 40% methylated in females at about the time of XCI (Singer-Sam *et al.*, 1990a). An alternative to the candidate approach is methylation fingerprinting, which is popular for studying aberrant methylation patterns in tumours (Frigola *et al.*, 2002; Gonzalgo *et al.*, 1997; Huang *et al.*, 1997). In

methylation-sensitive arbitrarily-primed PCR (MS-AP-PCR), for example, DNA from tumour and control samples are digested with *Rsa*I and *Hpa*II, and arbitrary primers in combination with low stringency reaction conditions are used to amplify random sequences containing methylated sites (Gonzalgo *et al.*, 1997). The primers were designed with a 3' end complementary to the recognition sequence of *Hpa*II to enhance the chance that an amplified product contains *Hpa*II sites, and DNA digested with the methylation-resistant isoschizomer of *Hpa*II, *Msp*I, was used as a control to distinguish between amplified products as a result of methylation at *Hpa*II sites and lack of *Hpa*II sites. These PCR-based restriction techniques are simple, fast, and sensitive, but can still only analyse CpGs located in methylation-sensitive restriction enzyme recognition sites.

The second class of methods 'fixes' the methylation pattern in DNA sequence by selective deamination of unmethylated cytosine to uracil using sodium bisulphite (Frommer *et al.*, 1992). Bisulphite-converted DNA can then be sequenced to give a detailed methylation profile of individual CpG dinucleotides. Bisulphite sequencing gives the most comprehensive methylation profiles and remains the gold standard in DNA methylation studies, but it is also the most laborious and thus limited in the scale of study. This technique is discussed in more detail in Chapter 4. PCR-based alternatives to sequencing have lower resolution but are more rapid: methylation-specific PCR (MSP) primers or unmethylated-DNA-specific PCR (USP) primers can be used to preferentially amplify hyper- or hypomethylated regions (Herman *et al.*, 1996; Kawakami *et al.*, 2004). Where accurate quantification of methylation of individual sites is required, methylation-sensitive single nucleotide primer extension (MS-SNuPE), which uses radioactive C or T, can alternatively be applied (Gonzalgo and Jones, 1997).

In recent years, microarrays have gained popularity as a high-throughput detection method in DNA methylation research. Methylation-specific oligonucleotide (MSO) microarrays were designed to contain sets of oligonucleotides that distinguish between bisulphite-converted TpGs and methylation-protected CpGs, and were used to screen methylation at selected CpGs in tumour samples (Adorján *et al.*, 2002; Gitan *et al.*, 2002). Alternatively, hyper- or hypomethylated fractions may be enriched prior to hybridisation to microarrays. In differential methylation hybridisation (DMH), fragmented genomic DNA are ligated to adaptors followed by digestion with methylation-sensitive restriction enzymes, and the intact hypermethylated fragments are enriched by PCR amplification with primers specific to the adaptors (Huang *et al.*, 1999). Similarly, hypomethylated proportions can be enriched if the genomic DNA is initially fragmented by methylation-sensitive restriction enzyme, and the methylated fragments are eliminated by an enzyme that cleaves DNA containing methylcytosine (Schumacher *et al.*, 2006). More recently, a promising new technique has been developed to enrich methylated DNA by immunoprecipitation. In this methylation DNA immunoprecipitation (MeDIP) approach, genomic DNA is fragmented, denatured, and incubated with a methylcytosine-specific antibody, and methylated DNA fragments are recovered through immunoprecipitation (Weber *et al.*, 2005). MeDIP-enriched DNA and control genomic DNA can be labelled with different fluorescent dyes and hybridised to the same microarray, and methylation levels can be measured by calculating the ratio between the two signals. Combining MeDIP with two different arrays, a whole genome tiling array and a CGI array, Weber and colleagues (2005) confirmed that the inactive X is globally hypomethylated but hypermethylated at CGIs. These array-based methods are very powerful high-throughput

methods that yield a great amount of information in a single experiment and are ideal for whole genome profiling, but they require extensive optimisation and careful statistical analysis, and are limited in resolution.

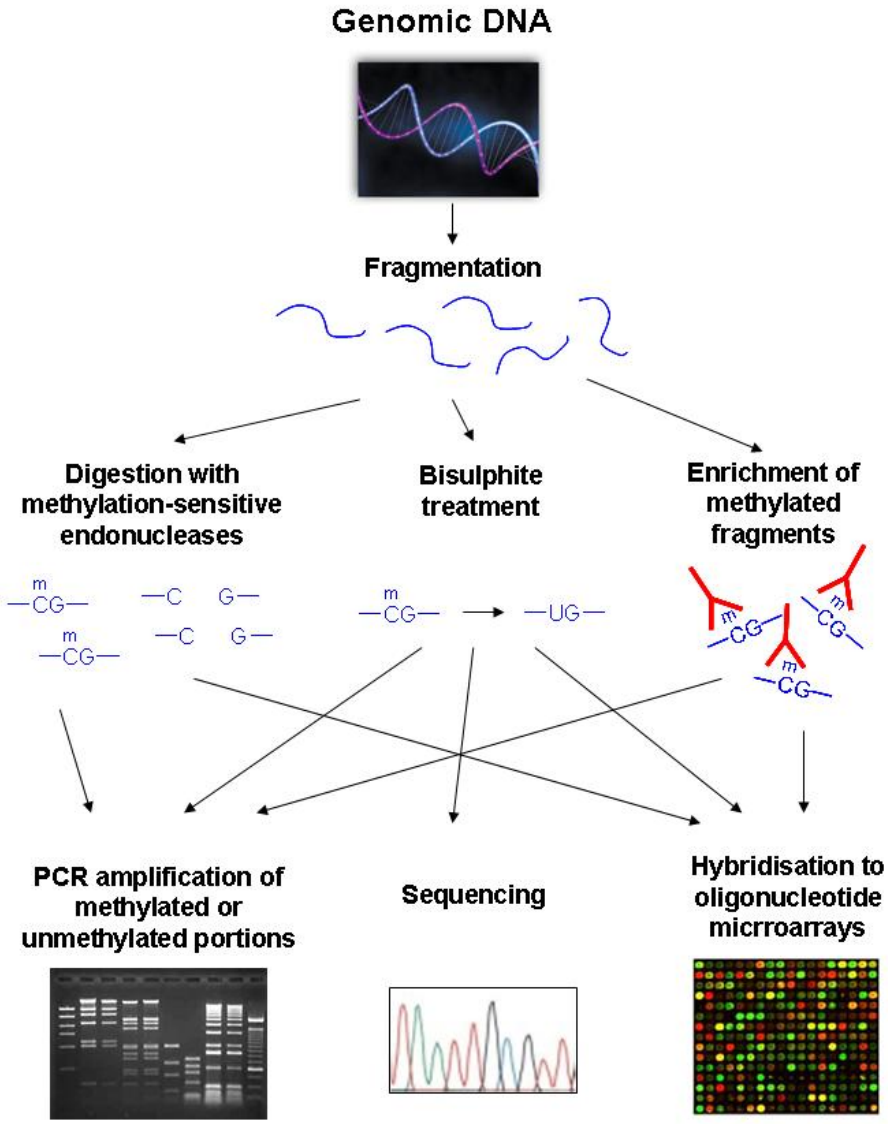

Figure 3.1: Approaches to studying DNA methylation.

3.1.3 Aims

The work described in this chapter had two aims: the first was to assess the use of CGI methylation state as an indication of a gene's XCI status in human; the second was to extend this analysis to the orthologous genes in the mouse in order to see if there are gross differences between the two species that might explain variability in escape from XCI.

3.2 Identification of target CpG islands

3.2.1 Region of interest

Recent studies revealed that more than 15% of genes on the human X chromosome escape from XCI (Carrel and Willard, 2005), but the distribution of such 'escapees' is not uniform, reflecting their evolutionary history. As described in Chapter 1, the human X chromosome can be divided into five evolutionary strata (Lahn and Page, 1999; Ross *et al.*, 2005): S1 and S2 in the XCR, which is common to X chromosomes in all mammalian groups (Graves, 1995); and S3, S4, S5 in the XAR, which was translocated from an autosome to the sex chromosomes in the eutherian lineage (Figure 1.2) (Graves 1995). These strata show an increasing level of sequence identity between X and Y chromosomes from S1 (lowest) to S5 (highest). At the distal end of the XAR, beyond S5, lies PAR1, where X and Y still recombine. Escapees are very rare in S1 and S2, but constitute a good proportion of genes in S3 and make up the entire set of genes tested in S4, S5 and PAR1, supporting a differential recruitment of X-linked genes into XCI (Carrel and Willard, 2005).

For the purpose of this study, the candidate region needs to contain a large number of inactivated and escapee genes and have orthologues on the mouse X chromosome, so any difference of XCI statuses between human and mouse are most likely to be seen. The first requirement limits the candidate region to XAR. In the XAR, most genes in S4 and S5 do not have orthologues on the mouse X chromosome, and comparison between the human and chicken genomes suggests that these genes were present in the original XAR and were lost in the mouse lineage (Ross *et al.*, 2005). Therefore, I have initially chosen to study genes within S3.

3.2.2 Identification of target human and mouse CGIs

Information about human genes used in this study was obtained from the Vega database (v19, based on the NCBI36 build) for the high quality annotation. There were not enough mouse genes annotated in Vega at the time, so the Ensembl annotation (v38, based on the NCBIM35 build) was used for the mouse genes. CGIs were predicted by Val Curwen using the cpg program (Gos Micklem, unpublished) with the following parameters: a minimum length of 400 bp, a minimal $\mathrm{Obs_{CpG}}/\mathrm{Exp_{CpG}}$ ratio of 0.6 and a minimal GC content of 50%.

On the human X chromosome, 147 protein-coding genes were found in S3, 91 (62%) of which had a predicted CGI close to the 5' end. Three pairs of genes, transcribed in opposite directions, had a CGI overlapping with the first exons of both genes, so 88 unique CGIs were available to study. CGI sizes ranged between 400 bp and 3 kb, with an average size of about 1 kb. Two CGIs were upstream from the transcription start site of the genes. The remaining CGIs all overlapped

exon1, the vast majority extending into intron1.

Orthologous genes in mouse for each of the 147 human S3 genes were identified based on Ensembl annotation. Putative orthologues were identified for 131 genes, of which 126 were also X-linked in mouse. Fifty-four of the 126 genes had a predicted 5' CGI. Only one gene, *CXorf23*, has a human orthologue without 5' CGI, and is not suitable for this study. The other 53 genes were all included in this study. CGI sizes ranged between 400 bp and 2 kb, with an average size of about 900 bp. Three CGIs were entirely upstream of the gene start. All the other CGIs cover the entire exon1 and extend into intron 1.

In total, 88 human CGIs and 53 mouse CGIs were selected for this study. Their sequences, with 100 bp flanking sequence on each side, were extracted from the appropriate database (Vega for human, Ensembl for mouse) via Ensembl Perl API using a Perl script (Appendix V).

3.3 Assessing CpG island methylation on mouse and human X chromosomes by Restriction-PCR Methylation Assay (RPMA)

For the purpose of rapidly assaying a large number of CGIs initially, I decided to use a restriction-PCR approach, which I have termed RPMA (restriction PCR methylation assay). RPMA is similar to the method used by Singer-Sam and colleagues (1990a; 1990b) but includes DNA digested by the methylation-resistant *Msp*I to serve as a control for successful digestion. Female and male genomic DNAs are digested with *Msp*I or its methylation-sensitive isoschizomer *Hpa*II,

and then the digested DNA is tested for its ability to act as template for amplification of PCR products containing multiple *Msp*I/*Hpa*II sites from predicted CpG islands. This method was employed by Jegalian and Page (1998) to assay CGI methylation of four X-linked genes in 19 species. This is the only large multi-species XCI study to date, from which the authors concluded that CGI methylation accompanies XCI in a wide range of eutherians and that XCI evolved on a gene-by-gene or cluster-by-cluster basis.

Human and mouse genomic DNA was extracted from female and male cultured fibroblast cells, digested with *Hind*III, *Hpa*II, or *Msp*I, and then used as template for PCR amplification. Since the predicted amplicons contained no *Hind*III cleavage sites, the *Hind*III-digested DNA was expected always to support amplification of PCR products, and thus acted as a positive control. By contrast, DNA digested with the methylation-resistant *Msp*I was expected always to be cleaved between the PCR primers and to fail to yield any PCR products. This allowed control for effectiveness of restriction digestion. Cleavage by the methylation-sensitive *Hpa*II is blocked by methylation of the CpG in the CCGG recognition site, and so a PCR product was only expected when all the CpG dinucleotides in the amplicon were methylated. An outline of the experimental strategy and the expected agarose gel electrophoresis pictures for hypo- and hypermethylated CGIs are shown in Figure 3.2.

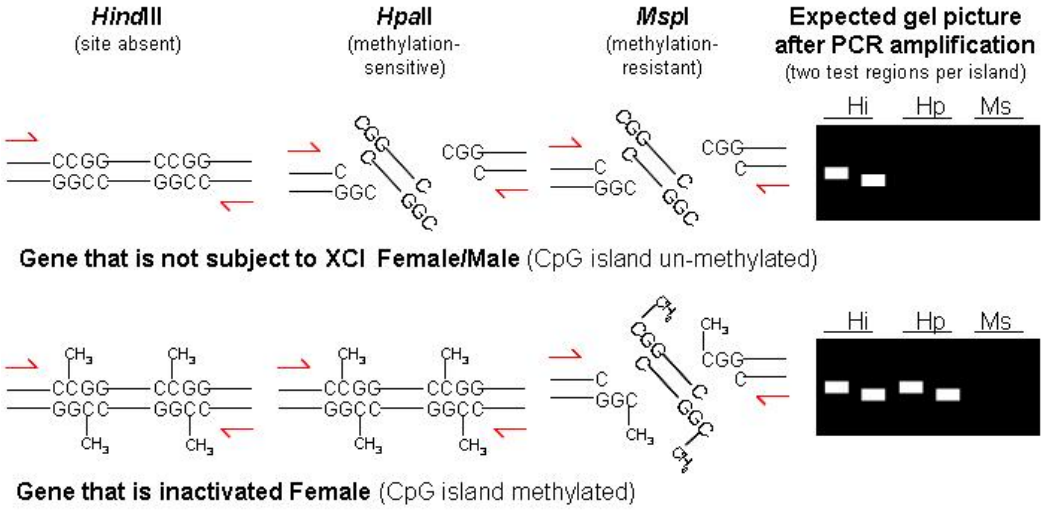

Figure 3.2: Restriction-PCR Methylation Assay (RPMA). Female and male genomic DNAs digested with *Hind*III (*Hi*, positive control), *Hpa*II (*Hp*), or *Msp*I (*Ms*, negative control) were used as templates for PCR amplification. A strong band in the female *Hp* lane (comparable to the female *Hi* lane) is expected if the CGI is hypermethylated and the gene undergoes XCI, whereas absence of product in the female *Hp* lane is expected when the CGI is hypomethylated and the gene escapes from XCI. Wherever possible, two regions were tested in each CGI to give better representation.

3.3.1 Optimisation of experimental conditions

PCR conditions were optimised to ensure successful and specific amplification of the GC-rich sequences. A set of reaction conditions have been previously optimised for amplification of GC-rich region: 5% DMSO was included to prohibit secondary structure formation and to help destabilise the double-helix structure; 50% of the dGTP was substituted with 7-deaza-dGTP to prevent compressions; and a mixture of Advantage™ 2 polymerase mix and AmpliTaq® polymerase was found to facilitate PCR amplification of a wide range of product sizes (credit to Christine Burrows and Dr Tamsin Eades). These PCR conditions were tested on four published primer pairs (Jegalian and Page 1998) using human and mouse

genomic DNA digested with *Hind*III, *Hpa*II, or *Msp*I, but it failed to amplify the *ZFX* CGI in both human and mouse, and gave only very weak signals for the CGIs of *SMCX* and *RPS4X*, again in both human and mouse.

I then tested the effects of a higher primer concentration, a hot-start step of heating the template and primers at 100 °C for 5 minutes, and an alternative cycling profile as used by Jegalian and Page (1998) (with a higher annealing temperature). The higher primer concentration and the hot-start step were found to improve PCR amplification. The alternative cycling profile reduced appearance of non-specific products but also gave weaker signals of the desired products. To find out if the PCR results were affected by the restriction enzyme digestion step, these experiments were repeated on undigested genomic DNA and digested DNA that underwent a heat inactivation of restriction enzymes at 65 °C for 20 minutes. Heat inactivation reduced appearance of non-specific products, but the signals were not as clean and strong as when using undigested DNA as template. On the base of these results, a higher primer concentration and the hot-start step were incorporated into my original PCR conditions for subsequent experiments.

There are two possible ways that the restriction enzyme digestion step may affect PCR amplification: template availability may be reduced by over-digestion; or chemicals present in the digestion reaction may interfere with PCR conditions. To address the former issue, restriction enzyme digestion conditions were optimised to ensure complete but not over digestion of genomic DNA. Human DNA was digested with *Hind*III, to test for over-digestion, or *Msp*I, to test for incomplete digestion, at 100 units of enzyme per μg DNA concentration for different lengths of time. The PCR results from templates generated by 4-, 8-, or 16-hour digestion were indistinguishable (Figure 3.3a) for all four amplicons, and

for convenience a 16-hour (overnight) digestion was performed for all subsequent experiments. Next the human DNA was digested with *Hpa*II or *Msp*I at four different enzyme concentrations, 10, 20, 50, or 100 units of enzyme per μg DNA, to test for incomplete digestion. In all cases, no PCR products at expected amplicon size were detected, but 10 and 20 units of enzyme per μg DNA resulted in a number of non-specific products for the *SMCX* CGI, so the optimal concentration was decided to be 50 units of enzyme per μg genomic DNA (Figure 3.3b). To examine the possible negative effects of chemicals present in the digestion reaction on PCR, the digested DNA was purified from heat-inactivated digestion reaction by organic solvent extraction followed by ethanol precipitation. PCR using the purified digested DNA gave comparable results with PCR using undigested DNA, so the purification step was used in all subsequent experiments.

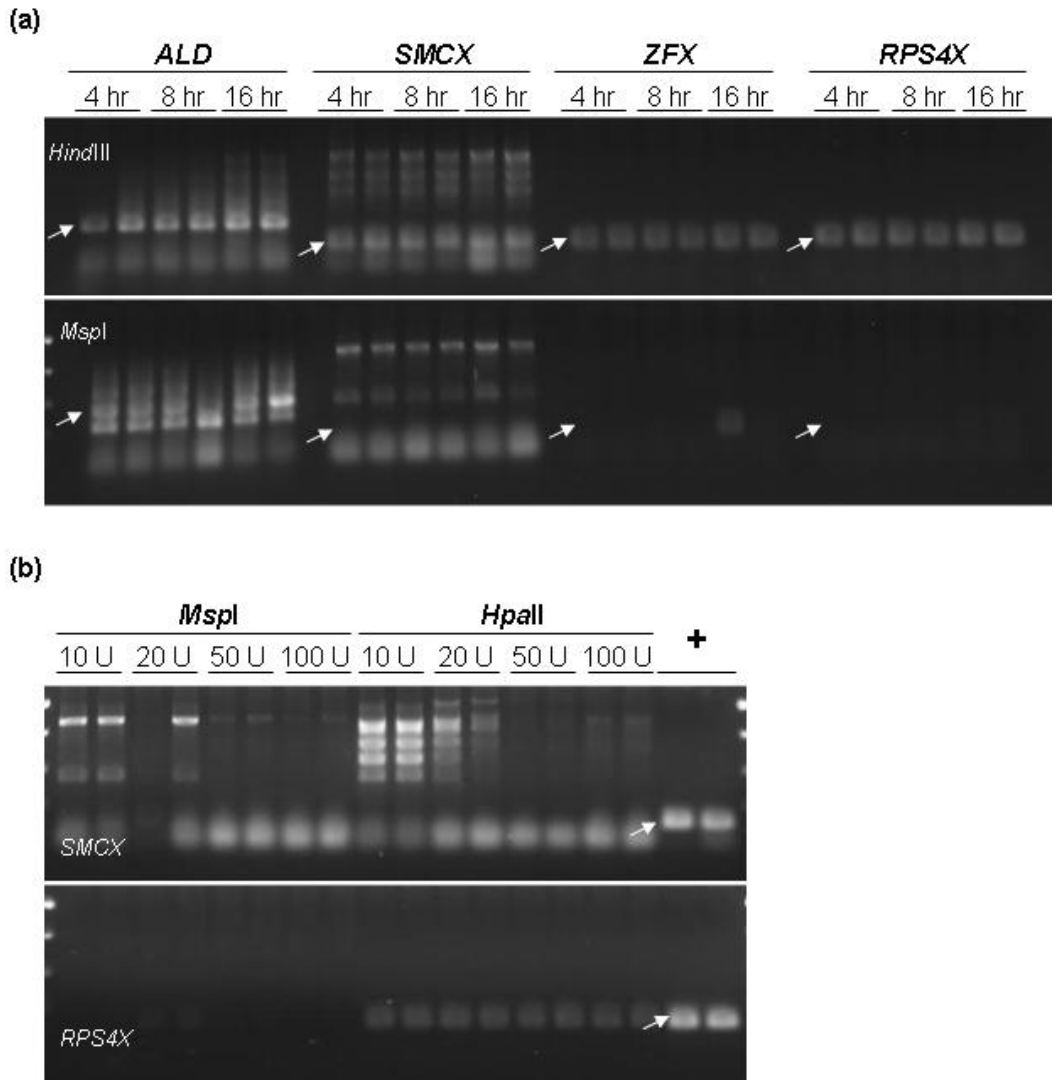

Figure 3.3: Optimisation of restriction enzyme digestion conditions. Expected product sizes are indicated by arrows. a) PCR results from templates generated by 4-, 8-, or 16-hour digestion. A PCR product is expected after *Hind*III but not *Msp*I digestion. b) PCR results from templates generated by digestion with 10, 20, 50, or 100 units of enzyme per μg DNA. '+' are positive controls. No products are expected from either digestion.

3.3.2 Primer design

Primers were designed to amplify 150-250 bp products containing multiple *Hpa*II/*Msp*I cleavage sites (CCGG) lying within the predicted CGI. Since single cleavage site may be protected by rare methylation, leading to over-representation of methylation, and a high number of sites may result in digestion despite hyper-methylation, leading to under-representation of methylation, so I aimed at including 2-4 cleavage sites in each amplicon. The amplicon sizes were designed to be larger than those used in Jegalian and Page's study (1998) to ensure good separation of PCR product and primer-dimer in gel electrophoresis. In order to increase the predictive power of the method, two pairs of primers enclosing regions with no overlapping CCGG sites were designed for each CGI. For three of the 88 human CGIs and four of the 53 mouse CGIs, only one pair of primers could be designed. Primers were tested in PCR amplification using male genomic DNA. Only three pairs of human primers failed the test, one of which is the only pair of primers available for a CGI. The expected products were successfully amplified for all other human primers and all mouse primers. Therefore in total, 87 human CGIs and 53 mouse CGIs were possible to assay. All primer combinations, their predicted amplicon sizes, and the number of enclosed CCGG sites are listed in Appendix I.

3.3.3 Analysis of CGI methylation on the human X chromosome

The 87 human CGIs were all successfully assayed in female and male human fibroblasts. Of these, 82 CGIs had data for both amplicons and five others had

results from a single amplicon.

The CGIs on the single X chromosome in male cells are expected to be unmethylated, like the CGIs on autosomes, and serve as a convenient indicator of the reliability of the assay. Indeed, the vast majority of CGIs showed a pattern consistent with hypomethylation in male samples, as expected (Table 3.1, Figure 3.4). For 58 (67%) CGIs, no PCR products were amplified using the *Hpa*II digested material, despite strong PCR amplification from the *Hind*III digested control. Another 16 (18%) CGIs had a complete lack of signal in the *Hpa*II lane for one amplicon, and a faint signal in the *Hpa*II lane for the other amplicon, still in support with overall hypomethylation. Six CGIs had a faint signal in the *Hpa*II lane for both amplicons, possibly indicating low level methylation. Only Seven (8%) CGIs had a signal in the *Hpa*II lane that is comparable with the signal from the *Hind*III control. These are discussed in more detail below. In contrast, many CGIs showed a pattern consistent with hypermethylation in female, confirming that the assay has the power to detect hypermethylation (Table 3.1, Figure 3.4). These observations indicate that the methylation state inferred from the assay data were reliable.

The CGIs that are hypomethylated in male can be divided into three categories according to their methylation patterns in the female samples.

Thirteen (15%) CGIs were clearly hypomethylated in both female and male samples, as demonstrated by the lack of signal in the *Hpa*II lane for both amplicons of each CGI (Figure 3.4, a-c).

Forty-nine (56%) CGIs showed female-specific methylation in at least one amplicon (Figure 3.4, d-f). In the male samples, at least one amplicon had no signal in the *Hpa*II lane, and the other amplicon was either absent (3 cases), or

had no signal (34 cases) or a faint signal (15 cases) in the *Hpa*II lane. In the female samples, 15 CGIs were clearly hypermethylated at both amplicons, each having a strong signal in the *Hpa*II lane indistinguishable from the signal from the *Hind*III control. The lack of signals from the *Msp*I control eliminated the possibility of incomplete digestion. An additional 22 CGIs had one amplicon with the similar clear hypermethylation pattern, and the other amplicon was either absent (3 cases), had a strong signal in the *Hpa*II lane but also a faint signal from the *Msp*I control (10 cases), suggesting some incomplete digestion, or had a faint or no signal in the *Hpa*II lane. For the other 12 CGIs, one or both amplicons had a strong signal in the *Hpa*II lane but also a faint or very faint signal from the *Msp*I control. In one case (*RP11-450P7.3*), one amplicon is hypermethylated in female but hypomethylated in male, and the other amplicon is hypermethylated in both sexes.

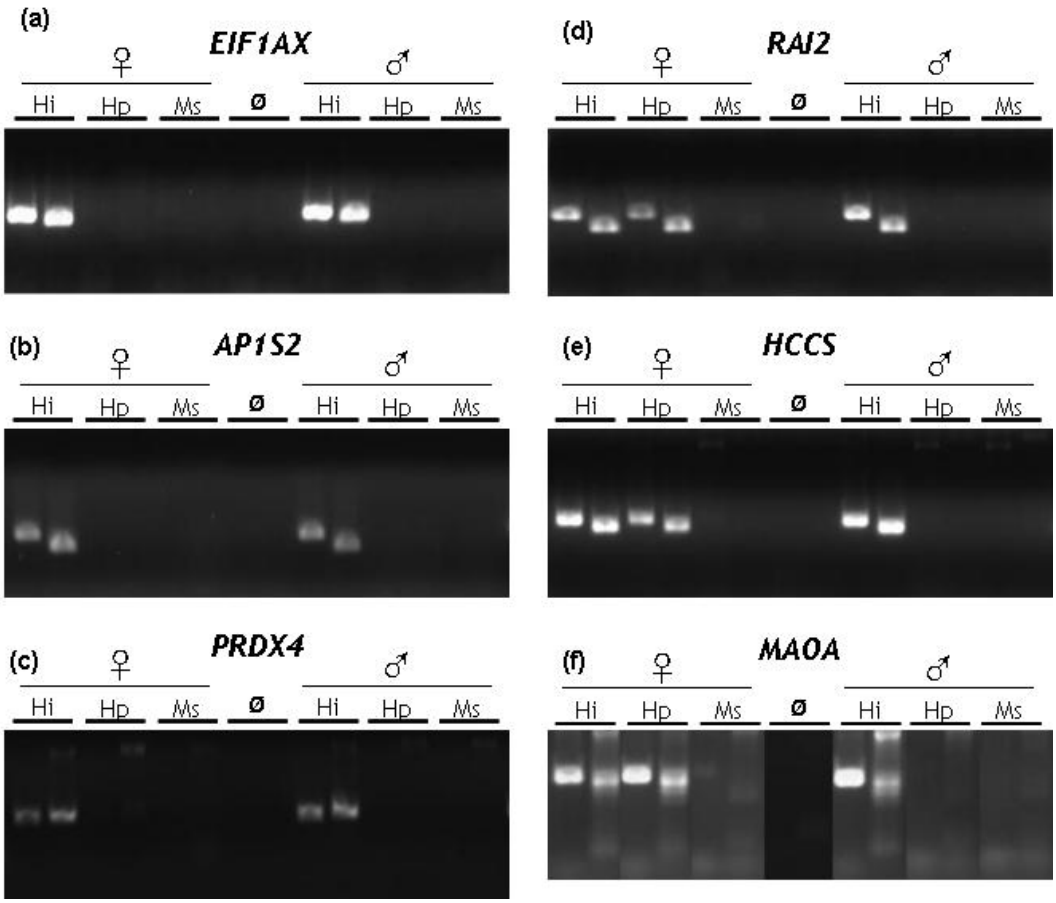

Figure 3.4: RPMA results of human CGIs with hypomethylation or female-specific methylation. (a-c) are examples of CGIs with hypomethylation in both female and male samples. (d-f) are examples of CGIs with female-specific methylation. Lanes in (f) have been re-organised as the original loading pattern had the two amplicons on separate gels.

In 19 cases (22%), the methylation state of the CGI was less well defined. In these cases faint bands were observed in the female *Hpa*II lane (Figure 3.5, a-c). The RPMA approach could be indicating an intermediate state of methylation of these CGIs but would not be able to distinguish between the two possible situations, which are a) a small percentage of heavily methylated CGIs, and b) an intermediate level of methylation in a higher percentage of CGIs. These cases

can be further divided into two broad categories: for eight CGIs, one amplicon was hypomethylated but the other was difficult to score (Figure 3.5a); for another 11 CGIs, the methylation state could not be decided for any amplicon (Figures 3.5, b-c). RPMA was repeated for most of the intermediate cases and the results were either identical or very similar, indicating that this is not an experimental artefact but an accurate representation of an intermediate state.

In the remaining six cases, both male and female CGIs are hypermethylated (Figure 3.5d).

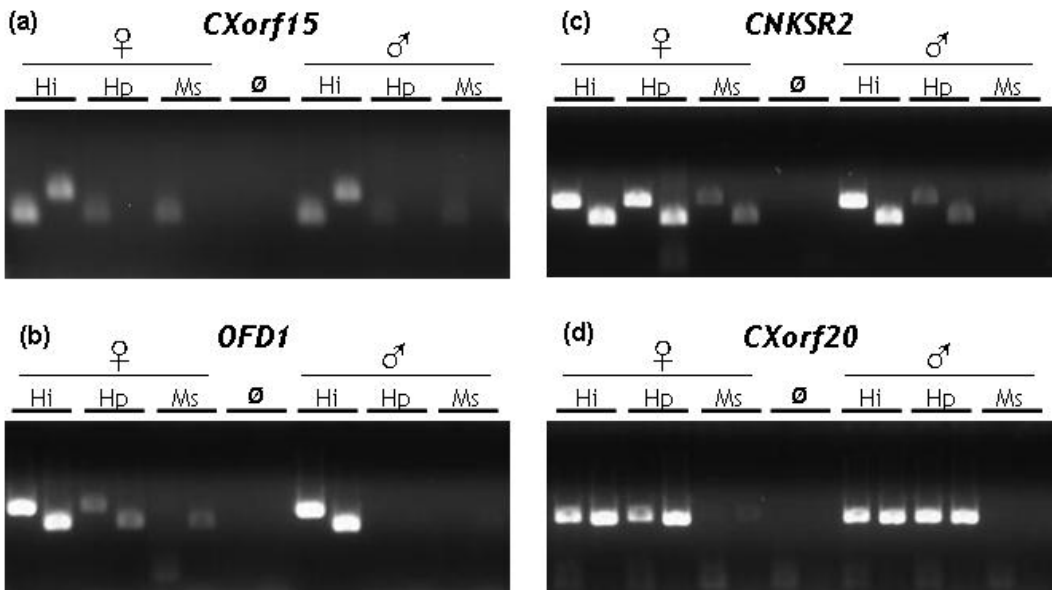

Figure 3.5: RPMA results of human CGIs with other methylation patterns. (a-c) are examples of CGIs for which the methylation state is difficult to define. (d) is an example of hypermethylation in both female and male samples.

Table 3.1: RPMA results of human CGIs. '+' = strong band; 'f' = faint band; 'vf' = very faint band; '-' = no visible band. In the 'XCI state' columns, the 'prediction' column contains XCI status prediction based on RPMA results, and the 'Lit' column contains XCI status recorded in Carrel and Willard (2005), where the numbers of somatic cell hybrids retaining X_i that showed gene expression is presented. XCI status is colour-coded: pink for X-inactivated, violet for escapee, and yellow for undetermined. One CGI is shared by each pair of the following: *RP1-93D11.1* and *RP13-928P6.1*, *RP13-314C10.2* and *RP11-793H5.3*, *OFD* and *TRAPPC2*.

	CGI	RPMA primer pair 1			RPMA primer pair 2			Female-	XCI state	
Gene	size	CCGG	Female	Male	CCGG	Female	Male	specific		
	(bp)	sites	Hi Hp Ms	Hi Hp Ms	sites	Hi Hp Ms	Hi Hp Ms	methylation?	prediction	Lit.
Female-specific methylation in at least one amplicon										
HCCS	481	5	+ + -	+ - -	2	+ + -	+ - -	Y	i	0 / 9
MSL3L1	1151	3	+ + -	+ - -	6	+ + -	+ - -	Y	i	3 / 9
PDZK10	669	3	+ + -	+ - -	3	+ + -	+ - -	Y	i	2 / 9
RP1-93D11.1	742	2	+ + -	+ - -	4	+ + -	+ - -	Y	i	0 / 9
RAI2	1260	6	+ + -	+ - -	3	+ + -	+ - -	Y	i	0 / 9
SCML2	1250	2	+ + -	+ - -	2	+ + -	+ - -	Y	i	0 / 9
MBTPS2	612	4	+ + -	+ - -	2	+ + -	+ - -	Y	i	0 / 9
ACOT9	539	3	+ + -	+ - -	5	+ + -	+ - -	Y	i	2 / 9
PDK3	930	2	+ + -	+ - -	2	+ + -	+ - -	Y	i	0 / 9
POLA	653	2	+ + -	+ - -	3	+ + -	+ - -	Y	i	0 / 9
TSPAN7	989	3	+ + -	+ - -	2	+ + -	+ f -	Y	i	1 / 9
MID1IP1	1857	2	+ + -	+ - -	3	+ + -	+ - -	Y	i	0 / 9
CHST7	1847	3	+ + -	+ - -	4	+ + -	+ - -	Y	i	0 / 9
PHF16	1417	5	+ + -	+ - -	4	+ + -	+ - -	Y	i	0 / 9
RGN	517	3	+ + -	+ - -	2	+ + -	+ - -	Y	i	1 / 9
FAM51A1	565	2	+ + -	+ f -	5	+ - -	+ - -	Y	i	9 / 9
PDHA1	752	2	+ + -	+ - -	7	+ - -	+ - -	Y	i	0 / 6
PTCHD1	2496	3	+ + -	+ - -	3	+ - -	+ - -	Y	i	N/A
BCOR	662	2	+ - -	+ - -	1	+ + -	+ f -	Y	i	2 / 9
SLC9A7	955	2	+ + -	+ - -	7	+ - -	+ - -	Y	i	1 / 9
PRPS2	1176	2	+ f -	+ - -	3	+ + -	+ - -	Y	i	0 / 5
ATP6AP2	965	1	+ + -	+ - -	2	+ f -	+ - -	Y	i	1 / 9
CASK	1151	1	+ + -	+ - -	3	+ f -	+ - -	Y	i	0 / 6
TMSB4X	1491	4	+ vf -	+ - -	1	+ + -	+ f -	Y	i	N/A
KLHL15	1258	5	+ + -	+ - -				Y	i	0 / 9
LANCL3	1077	3	+ + -	+ - -				Y	i	1 / 9
RPGR	977	8	+ + -	+ - -	1	fail	fail	Y	i	0 / 6
RP11-393H10.2	1146	2	+ + vf	+ vf -	3	+ + -	+ - -	Y	i	0 / 9
RPS6KA3	2683	3	+ + -	+ - -	2	+ + vf	+ - -	Y	i	0 / 9
RP11-450P7.3	3047	3	+ + -	+ - -	2	+ + f	+ + f	Y	i	N/A
RP13-314C10.2	724	3	+ + -	+ - -	3	+ + f	+ f -	Y	i	N/A
ARX	1047	2	+ + -	+ - -	2	+ + f	+ f -	Y	i	N/A
PRRG1	1087	1	+ + vf	+ vf vf	1	+ + -	+ - -	Y	i	0 / 9
XK	614	2	+ + f	+ f vf	4	+ + -	+ - -	Y	i	N/A
TCTE1L	676	3	+ + -	+ - -	2	+ + f	+ - -	Y	i	0 / 9
MAOA	1115	2	+ + vf	+ - -	5	+ + -	+ - -	Y	i	9 / 9
ZNF673	588	1	+ + f	+ - -	1	+ + -	+ - -	Y	i	0 / 9
ARHGAP6	2210	4	+ f -	+ - -	2	+ + vf	+ - -	Y	i	1 / 9
PIGA	638	2	+ + vf	+ - -	5	+ f -	+ - -	Y	i	0 / 9
REPS2	868	2	+ + +	+ vf vf	2	+ + vf	+ - -	Y	i	2 / 9
NHS	2657	3	+ + vf	+ - -	2	+ + vf	+ - -	Y	i	3 / 9
SCML1	1594	2	+ + vf	+ - -	3	+ + vf	+ - -	Y	i	0 / 9
PHKA2	1240	2	+ + f	+ vf vf	4	+ f -	+ - -	Y	i	2 / 9
MAP3K15	848	2	+ + f	+ - -	3	+ f vf	+ - -	Y	i	N/A
SH3KBP1	1529	1	+ + vf	+ - -	2	+ + vf	+ - -	Y	i	4 / 9
SAT	668	5	+ + f	+ - -	2	+ + vf	+ - -	Y	i	1 / 9
NR0B1	1228	1	+ + vf	+ f -	2	+ vf vf	+ - -	Y	i	N/A
GK	1129	2	+ + f	+ - -	4	+ + vf	+ - -	Y	i	0 / 9
SRPX	458	2	+ + f	+ - -	2	+ - -	+ - -	Y	i	0 / 9

(Table 3.1 continued)

Gene	CGI size (bp)	RPMA primer pair 1 CCGG sites	Female Hi Hp Ms	Male Hi Hp Ms	RPMA primer pair 2 CCGG sites	Female Hi Hp Ms	Male Hi Hp Ms	Female-specific methylation?	XCI state prediction	Lit.
Hypomethylated in both female and male										
GPM6B	910	3	+ - -	+ - -	3	+ - -	+ - -	N	e	8 / 9
CA5B	584	4	+ - -	+ - -	3	+ - -	+ - -	N	e	9 / 9
AP1S2	1546	3	+ - -	+ - -	5	+ - -	+ - -	N	e	9 / 9
CTPS2	1012	3	+ - -	+ - -	2	+ - -	+ - -	N	e	9 / 9
EIF1AX	1536	3	+ - -	+ - -	3	+ - -	+ - -	N	e	9 / 9
PRDX4	639	4	+ - -	+ - -	4	+ - -	+ - -	N	e	0 / 9
ZFX	1482	2	+ - -	+ - -	3	+ - -	+ - -	N	e	9 / 9
Cxorf38	0	3	+ - -	+ - -	3	+ - -	+ - -	N	e	N/A
CRSP2	941	3	+ - -	+ - -	5	+ - -	+ - -	N	e	6 / 6
USP9X	2246	5	+ - -	+ - -	5	+ - -	+ - -	N	e	9 / 9
UTX	1750	2	+ - -	+ - -	2	+ - -	+ - -	N	e	9 / 9
RP11-30A20.3	519	4	+ - -	+ - -	2	+ vf -	+ - -	N	e	9 / 9
U2AF1L2	633	1	+ vf -	+ - -	1	+ - -	+ - -	N	e	N/A
Methylation state difficult to define										
SMS	1348	2	+ f -	+ - -	6	+ - -	+ - -	U	U	0 / 6
RS1	446	5	+ f -	+ - -	4	+ vf -	+ - -	U	U	2 / 9
DDX3X	2423	4	+ f -	+ - -	2	+ f -	+ f f	U	U	9 / 9
PIR	440	3	+ - -	+ - -	2	+ f vf	+ f -	U	U	9 / 9
CXorf15	1549	1	+ vf vf	+ - -	3	+ - -	+ - -	U	U	9 / 9
RBBP7	1507	1	+ + +	+ f f	4	+ - -	+ - -	U	U	9 / 9
GPR64	1156	7	+ - -	+ - -	6	fail	fail	U	U	N/A
MAOB	527	4	+ - -	+ - -	2	+ f f	+ f f	U	U	N/A
EGFL6	572	2	+ f vf	+ - -	3	+ f f	+ - -	U	U	N/A
RAB9A	499	4	+ f f	+ f f	1	+ f -	+ - -	U	U	9 / 9
OFD1	894	2	+ f -	+ - -	3	+ f vf	+ - -	U	U	6 / 6
CA5BL	968	1	+ - f	+ f vf	1	+ - f	+ f f	U	U	9 / 9
SYAP1	480	1	+ + +	+ f -	2	+ f f	+ f -	U	U	9 / 9
CDKL5	1305	2	+ vf vf	+ - -	3	+ vf -	+ - -	U	U	2 / 9
EIF2S3	617	2	+ vf -	+ vf -	1	+ f -	+ f -	U	U	9 / 9
TMEM47	1219	3	+ f vf	+ - -	3	fail	fail	U	U	0 / 9
CLCN4	561	1	+ + -	+ vf -	3	+ f -	+ vf -	U	U	5 / 9
CNKSR2	1572	2	+ + vf	+ vf -	2	+ + vf	+ vf -	U	U	N/A
RP2	933	2	+ + -	+ f -	1	+ + -	+ f -	U	U	0 / 9
Hypermethylated in both female and male										
CXorf20	837	3	+ + -	+ + -	3	+ + -	+ + -	N/A	N/A	N/A
SIAH1L	446	2	+ + f	+ + f	1	+ + f	+ + f	N/A	N/A	N/A
YY2	563	1	+ + f	+ + f	1	+ + f	+ + f	N/A	N/A	N/A
FTHL17	548	1	+ + +	+ + +	3	+ + f	+ + f	N/A	N/A	N/A
RP1-154K9.1	662	1	+ + f	+ + f	2	+ + vf	+ + vf	N/A	N/A	N/A
MKRN4	482	4	+ + vf	+ + -	6	+ + -	+ + -	N/A	N/A	4 / 4

3.3.4 Analysis of CGI methylation on the mouse X chromosome

All 53 mouse CGIs were successfully assayed by RPMA in female and male mouse embryonic fibroblasts (MEF). Of these, 49 CGIs had data for both amplicons and four others had results from a single amplicon.

Similar to, but more extreme than the situation in human samples, almost all CGIs were hypomethylated in male samples (Table 3.2, Figure 3.6). For 44 (83%) CGIs, no PCR products were amplified using the *Hpa*II digested material. Another seven (13%) CGIs had a complete lack of signal in the *Hpa*II lane for one amplicon, and a faint or very faint signal in the *Hpa*II lane for the other amplicon. Only two CGIs had a signal in the *Hpa*II lane that is comparable with the signal from the *Hind*III control.

However, the methylation patterns in mouse female samples are very different from that observed in human samples. Only three CGIs, covering the 5' regions of the genes *Utx*, *Eif2s3x*, and *Ddx3x*, were clearly hypomethylated in both female and male samples (Figure 3.6, a-c). The CGIs of the orthologous genes in human appeared to be either hypomethylated (*UTX*) or having low level methylation (*EIF2S3*, *DDX3X*) in female.

In contrast, the vast majority of mouse CGIs (48 out of 53) exhibited female-specific methylation in at least one amplicon (Figure 3.6, d-f), and most (35 out of 48) had both amplicons methylated in female samples. Eight CGIs had one amplicon clearly hypermethylated, and the other amplicon was either absent (1 case), had a strong signal in the *Hpa*II lane but also a very faint signal from the *Msp*I control (2 cases), or had a faint signal in the *Hpa*II lane (2 cases). For

the other five CGIs, all amplicons had a strong signal in the *Hpa*II lane but also a faint or very faint signal from the *Msp*I control. Interestingly, the patterns suggesting intermediate levels of methylation, observed in a considerable fraction of human CGIs, were never seen in mouse CGIs.

In the remaining two cases, the CGIs showed some degrees of methylation in both female and male samples.

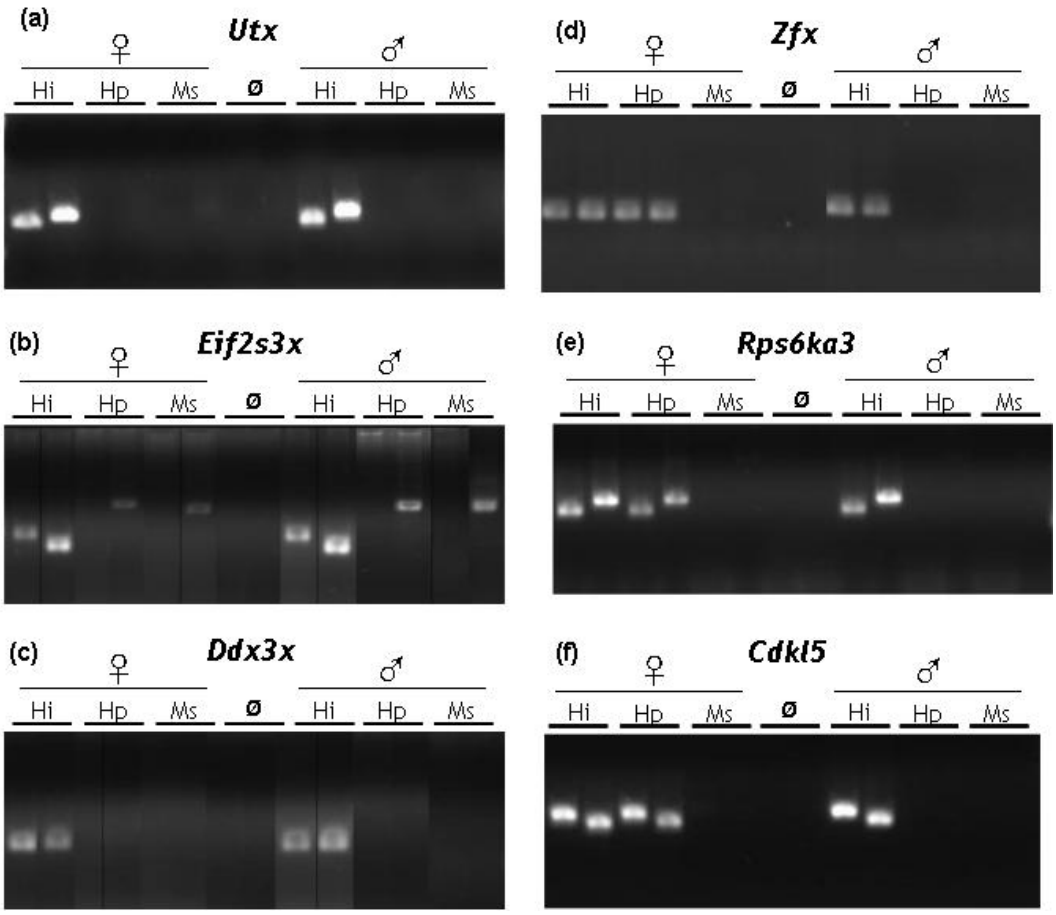

Figure 3.6: RPMA results of mouse CGIs with hypomethylation or female-specific methylation of CGI. (a-c) are examples of CGIs with hypomethylation in both female and male samples. (d-f) are examples of CGIs with female-specific methylation. Lanes in (b) and (c) have been re-organised as the original loading pattern had the two amplicons on separate gels.

Table 3.2: RPMA results of mouse CGIs. '+' = strong band; 'f' = faint band; 'vf' = very faint band; '-' = no visible band. The 'XCI state prediction' is based on RPMA results. For most mouse CGIs, no record of XCI status can be found in literature. XCI status is colour-coded: pink for X-inactivated and violet for escapee.

Gene	Human orthologue	CGI size (bp)	RPMA primer pair 1			RPMA primer pair 2			Female-specific methylation?	XCI state prediction
			CCGG sites	Female Hi Hp Ms	Male Hi Hp Ms	CCGG sites	Female Hi Hp Ms	Male Hi Hp Ms		
Female-specific methylation in at least one amplicon										
Hccs	HCCS	729	3	+ + -	+ - -	3	+ + -	+ f -	Y	i
Msl31	MSL3L1	1029	2	+ + -	+ vf -	3	+ + -	+ - -	Y	i
A630018P17Rik	RP11-30A20.3	661	3	+ + -	+ - -	2	+ + -	+ - -	Y	i
Rab9	RAB9A	504	2	+ + -	+ - -	2	+ + -	+ - -	Y	i
Ofd1	OFD1	846	2	+ + -	+ vf -	3	+ + -	+ - -	Y	i
Gpm6b	GPM6B	641	4	+ + -	+ - -	4	+ + -	+ vf -	Y	i
Piga	PIGA	527	2	+ + -	+ - -	6	+ + -	+ - -	Y	i
U2af1-rs2	U2AF1L2	1089	2	+ + -	+ - -	3	+ + -	+ - -	Y	i
Ap1s2	AP1S2	1112	3	+ + -	+ - -	3	+ + -	+ - -	Y	i
4932441K18Rik	CXorf15	1152	2	+ + -	+ - -	2	+ + -	+ - -	Y	i
Rbbp7	RBBP7	1362	3	+ + -	+ - -	2	+ + -	+ - -	Y	i
Reps2	REPS2	516	3	+ + -	+ - -	2	+ + -	+ - -	Y	i
ENSMUSG00000067171	NHS	1354	2	+ + -	+ - -	3	+ + -	+ - -	Y	i
Rai2	RAI2	1016	2	+ + -	+ - -	2	+ + -	+ - -	Y	i
Cdk15	CDKL5	899	2	+ + -	+ - -	2	+ + -	+ - -	Y	i
Phka2	PHKA2	769	2	+ + -	+ - -	3	+ + -	+ - -	Y	i
Gpr64	GPR64	463	3	+ + -	+ - -	3	+ + -	+ - -	Y	i
Pdha1	PDHA1	489	6	+ + -	+ - -	2	+ + -	+ - -	Y	i
ENSMUSG00000031303	MAP3K15	728	3	+ + -	+ - -	6	+ + -	+ - -	Y	i
Sh3kbp1	SH3KBP1	999	3	+ + -	+ - -	2	+ + -	+ - -	Y	i
Eif1ay	EIF1AX	783	4	+ + -	+ - -	3	+ + -	+ - -	Y	i
Rps6ka3	RPS6KA3	1155	3	+ + -	+ - -	4	+ + -	+ - -	Y	i
Cnksr2	CNKSR2	1447	2	+ + -	+ - -	5	+ + -	+ - -	Y	i
Mbtps2	MBTPS2	730	3	+ + -	+ - -	7	+ + -	+ - -	Y	i
Sms	SMS	1001	2	+ + -	+ - -	2	+ + -	+ - -	Y	i
Ptchd1	PTCHD1	1174	3	+ + -	+ - -	2	+ + -	+ - -	Y	i
Zfx	ZFX	1590	2	+ + -	+ - -	3	+ + -	+ - -	Y	i
Siah1b	SIAH1L	686	3	+ + -	+ - -	8	+ + -	+ - -	Y	i
Mid1ip1	MID1IP1	1532	2	+ + -	+ - -	2	+ + -	+ - -	Y	i
Atp6ap2	ATP6AP2	744	3	+ + -	+ - -	1	+ + -	+ - -	Y	i
Usp9x	USP9X	1660	1	+ + -	+ - -	4	+ + -	+ - -	Y	i
Cask	CASK	1174	2	+ + -	+ - -	7	+ + -	+ - -	Y	i
Maoa	MAOA	504	2	+ + -	+ - -	3	+ + -	+ - -	Y	i
Slc9a7	SLC9A7	550	3	+ + -	+ - -	2	+ + -	+ - -	Y	i
Phf16	PHF16	1017	2	+ + -	+ - -	5	+ + -	+ - -	Y	i
Arhgap6	ARHGAP6	463	3	+ + -	+ - -	2	+ + -	+ - -	Y	i
2900002G04Rik	RP11-393H10.2	1286	2	+ + -	+ - -	4	+ f -	+ - -	Y	i
Prdx4	PRDX4	409	3	+ + -	+ - -				Y	i
Tmem47	TMEM47	793	5	+ f -	+ - -	2	+ + -	+ - -	Y	i
Crsp2	CRSP2	974	2	+ + -	+ - -	2	+ f f	+ f f	Y	i
Gyk	GK	659	1	+ + vf	+ - -	2	+ + -	+ - -	Y	i
Lancl3	LANCL3	1078	2	+ + -	+ - -	1	+ + vf	+ f -	Y	i
Prps2	PRPS2	679	2	+ + vf	+ - -	2	+ + vf	+ - -	Y	i
Tmsb4x	TMSB4X	2054	1	+ + f	+ - -	1	+ + f	+ - -	Y	i
Syap1	SYAP1	594	1	+ + vf	+ - -				Y	i
Pdk3	PDK3	497	4	+ + vf	+ - -				Y	i
Pola1	POLA	485	3	+ + vf	+ - -				Y	i
Hypomethylated in both female and male										
Eif2s3x	EIF2S3	651	1	+ - -	+ - -	2	+ - -	+ - -	N	e
Ddx3x	DDX3X	1274	2	+ - -	+ - -	3	+ - -	+ - -	N	e
Utx	UTX	1739	3	+ - -	+ - -	2	+ - -	+ - -	N	e
Hypermethylated in both female and male										
Xkh	XK	652	1	+ + vf	+ f vf	1	+ + vf	+ + -	N/A	N/A
Bcor	BCOR	2118	3	+ + -	+ + -	2	+ + -	+ + -	N/A	N/A

3.4 Discussion

In this chapter, I wish to use the methylation state of the 5' CGI as an indicator to predict the genes' XCI status. To this end, CGI methylation of 90 human X-linked genes, known to include both inactivated and escapee genes, were examined in cultured fibroblast cells using a rapid method, RPMA. An attempt was made to predict the gene's XCI status based on the methylation profile of its CGI: hypomethylated CGIs were assumed to associate with escapee genes and hypermethylated CGIs were associated with inactivated genes. Since hypomethylation of a single CCGG site, thus cleavage by *Hpa*II is sufficient to prevent PCR amplification, a prediction of escapee was only made when both amplicons in one CGI gave a pattern consistent with hypomethylation. In total, a prediction was made for 64 (71%) genes.

Over half of the human CGIs were found to have female-specific methylation (Table 3.1, Figure 3.4), and the 51 associated genes (57% of all assayed genes) were predicted to be subject to XCI. Previously, Carrel and Willard (2005) created an inactivation profile for the whole human X chromosome by analysing gene expression from human X_a and X_i in a panel of rodent/human somatic cell hybrids. From the S3 region, 128 genes were assayed in their study and 83 (65%) were found to be effectively silenced (expression detected from no X_i hybrid or up to three X_i hybrids). This proportion of silenced genes based on expression is similar to the estimation based on CGI methylation in this study. Moreover, 42 out of the 51 genes with a prediction of inactivation in this thesis were examined in the study by Carrel and Willard (2005). Thirty-nine out of the 42 were shown to be inactivated. *MAOA* (which encodes monoamine oxidase type

A) (Figure 3.4, f), one of those whose inactivation state disagreed, was found to escape XCI in the previous study but should be subject to XCI according to the methylation test. In Carrel and Willard's study, the XCI status of *MAOA* was also tested using a quantitative SNP assay in primary human cells, and interestingly, the expression from X_i never exceeded 15% in any cell line tested, so *MAOA* may still be relatively effectively silenced. In addition, several other studies, using diverse methods including CGI methylation analysis, expression analysis by RT-PCR, and allele-specific expression analysis by SNP genotyping, all found evidence supporting inactivation of *MAOA* (Benjamin *et al.*, 2000; Hendriks *et al.*, 1992; Nordquist and Oreland, 2006).

Notably, a considerable fraction of human CGIs appeared to be hypomethylated. Thirteen CGIs were clearly hypomethylated at both amplicons in both female and male samples (Table 3.1, Figure 3.4). If absence of CGI methylation is a reliable indicator of poor inactivation, these genes are predicted to escape from XCI in the fibroblast cells used. Alternatively, these genes may not rely on methylation for maintaining stable inactivation. Comparison to published data favours the former possibility: of these 13 genes with hypomethylated CGIs, 11 were also assayed by Carrel and Willard (2005), and ten were found to escape from XCI (expression detected in all, or in one case, 8 out 9, X_i hybrids). The only disagreement was *PRDX4*, for which no other XCI data are available in literature.

In total, 64 genes had a prediction of their XCI status made based on clear CGI methylation patterns as assayed by RPMA. Of these, 53 genes were assayed in cell hybrids in the previous study, and for 49 genes (92%), observations from the previous study strongly support predictions made in current study. Therefore,

CGI methylation as measured by RPMA can serve as a good indicator of the gene's XCI status.

Previous researchers have suggested that XCI in mouse is more complete than XCI in human (Disteche, 1995). To investigate whether this difference in XCI is reflected in difference in CGI methylation on the X chromosome, I extended the methylation analysis to the mouse orthologues of human S3 genes. Of the 53 mouse genes assayed, only the three known escapees, *Utx* (Greenfield *et al.*, 1998), *Eif2s3x* (Ehrmann *et al.*, 1998), and *Ddx3x* (Disteche *et al.*, 2002), gave a pattern consistent with escape from XCI (Figure 3.6, a-c). Forty-eight genes were shown to have a profile consistent with inactivation (Table 3.2, Figure 3.6, d-f), including *Zfx*, which is known to be silenced (Adler *et al.*, 1991; Ashworth *et al.*, 1991). XCI statuses for most of the 48 genes have not been reported in literature previously, but available evidence suggests that the vast majority of genes on the mouse X chromosome undergo normal XCI (Disteche, 1995), which is consistent with my finding.

Comparing the CGI methylation profiles of the orthologous human and mouse genes reveals a much higher proportion of escapees in human. The same trend has also been observed by other researchers. Whereas more than 15% of genes on the human X chromosome were found to escape XCI (Carrel and Willard, 2005), less than ten escapees have been identified on the mouse X chromosome so far (Disteche *et al.*, 2002), so it might not be surprising that I failed to discover any novel escapees in mouse. Although an extensive XCI profiling has not been done for the entire mouse X chromosome, many of the genes that escape from XCI in human were found subject to inactivation in mouse, suggesting more complete XCI in mouse (Disteche, 1995). This observation is also supported by

comparative analysis of a conserved gene cluster around *SMCX/Smcx*, a known escapee in both human and mouse (Tsuchiya *et al.*, 2004). Whereas there is a large escapee domain in human, *Smcx* is the only escapee in the region in mouse. Similarly, I found the genes surrounding the three mouse escapees to all have a CGI methylation profile consistent with inactivation, in contrast to the situation in human where the escape domains are larger.

Using the rapid RPMA approach, I confirmed that a much higher percentage of genes are silenced on the mouse X_i than on the human X_i, and identified a large number of mouse genes that undergo XCI, most of which is novel finding.

It is well known that 5' CGIs are generally unmethylated in somatic tissues in healthy individuals. Even in the case of XCI or imprinting, the CGI is only methylated at one of the two alleles. Therefore it is curious that in my study, six human CGIs (Figure 3.5d) and one mouse CGI displayed identical hypermethylation in both female and male samples. This methylation pattern is more commonly observed in non-5' CGIs. A closer examination of the exact locations of the assayed sites revealed that most sites are either intronic (as the case of Figure 3.5d) or in the exon of a single-exon gene. Interestingly, the only other human CGI associated with a single-exon gene had one amplicon hypermethylated in both sexes while the other amplicon had normal female-specific methylation. Moreover, in an additional case in mouse, one assayed site is intronic and was hypermethylated in male, while the other site is located towards the end of exon1 and showed low level methylation. It is very possible that these CGIs, or the assayed portions of them, are not critical in transcriptional control.

Another interesting difference between human and mouse comes from the unclassifiable cases. The remaining 19 CGIs in human mostly had faint signals

in female, and sometimes in male, *Hpa*II lane, suggesting intermediate levels of methylation (Figure 3.5, a-c). These ambiguous methylation patterns make interpretation of XCI status difficult, but interestingly, were never seen in any mouse CGIs. Unfortunately, the RPMA approach is limited in its ability to examine details of methylation. At this stage, it was not possible to distinguish between patterns formed by a small proportion of hypermethylated CGIs and that by a large proportion of intermediately methylated CGIs. To test which situation gave rise to these intermediate methylation patterns, more detailed methylation information, ideally at single CpG resolution, needs to be obtained.

In this chapter, I have demonstrated that RPMA is a rapid method of predicting XCI status for a large fraction of genes with a CGI. RPMA results are consistent with much more complete gene silencing in mouse XCI than in human. In addition, there is a class of intermediately methylated CGIs unique to the human species and I shall carry out further study of the detailed methylation patterns for a representative set of genes in the next chapter.

Chapter 4

Comparison of detailed CGI methylation patterns on the human and mouse X chromosomes

4.1 Introduction

In the early days of DNA methylation research, methods of analysing CGI methylation mainly relied on methylation-sensitive restriction enzymes. Studies using these techniques helped to build our initial understanding of the phenomenon, but as discussed in Chapter 3, were limited by the availability of enzyme cleavage sites and resolution. In 1992, Frommer and colleagues described a completely new approach that 'fixes' the methylation pattern in the DNA sequence and produces a single-basepair resolution readout using the dideoxy sequencing method. The key to the novel method is the chemical sodium bisulphite, which preferentially deaminates cytosine residues to uracil in single-stranded DNA, but rarely reacts

with 5-methylcytosine. Bisulphite-converted DNA is used as a template for PCR amplification of regions under investigation. All uracils and thymines are amplified as thymines in the PCR products; only 5-methylcytosines are amplified as cytosines. The PCR products can be cloned and individual clones can be sequenced to give methylation maps of single DNA molecules in the original genomic DNA sample (Figure 4.1). Using DNA sequences with known methylation patterns, Frommer and colleagues (1992) optimised the conditions to completely convert cytosines, but leave 5-methylcytosines essentially non-converted. Then they analysed the methylation state of two CpG dinucleotides in a number of human tissue samples. The methylation state of one of these CpGs was confirmed using restriction-enzyme digestion, and the other CpG was impossible to assay using the restriction method. Bisulphite sequencing was found to be a reliable method for methylation analysis and it offers the major advantage of the ability to analyse methylation state of single cytosines on single DNA molecules.

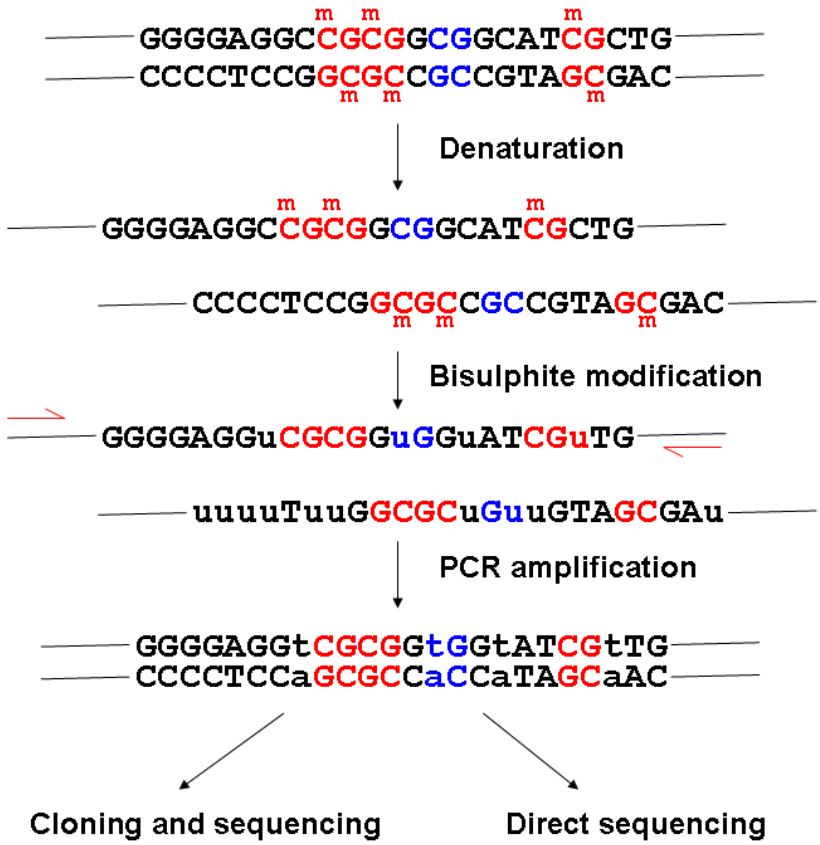

Figure 4.1: Bisulphite sequencing. Note the two strands no longer complement after bisulphite modification so different primers are needed to amplify the top and bottom strands. Shown here is the amplification and sequencing of top strand only.

Bisulphite sequencing provides an ideal method to study cell populations with mixed methylation profiles. One such example is X chromosome inactivation. Using restriction enzymes, it has been shown that the 5' region of the human *HPRT* and mouse *Hprt* genes is methylated on the inactive X, but unmethylated on the active X (Lock *et al.*, 1986; Wolf *et al.*, 1984; Yen *et al.*, 1984), but it was not clear whether all CpGs in the region were uniformly methylated. Nor was it known whether the methylation patterns are faithfully reproduced after cell division.

To address these questions, Park and colleagues (1994) analysed methylation patterns of 32 CpGs in a 371 bp region 5' of the mouse *Hprt* gene. They confirmed the X_i-specific methylation suggested by previous studies, but found that not all CpGs were methylated on all molecules, and levels of methylation at individual CpGs varied greatly. Moreover, the heterogeneous methylation patterns were not only observed within a tissue, but even in clonal cell populations, suggesting a dynamic nature of maintenance of methylation.

The fine resolution of bisulphite sequencing is also important to reveal methylation changes in early development, when the material is scarce and states of individual CpGs vary. A differentially methylated region (DMR) was identified 5' to the imprinted gene *H19* in mouse, which is methylated in the non-expressing paternal allele (Ferguson-Smith *et al.*, 1993). Using a restriction-PCR method, it was shown that this region is methylated in sperms but not eggs and the pattern is carried over to pre-implantation embryos (Tremblay *et al.*, 1995). Bisulphite sequencing of the 5' end of this DMR helped to identify the 5' differentially methylated boundary and provided much insight into the establishment of this boundary in early embryos (Olek *et al.*, 1996).

Methylation at the 5' promoter region of the tumour suppresser gene *Rb* has been studied using restriction enzymes, and suggested CGI methylation as a potential mechanism of oncogenesis (Greger *et al.*, 1989; Sakai *et al.*, 1991). However, only a limited number of sites were studied in this way and it was not clear if the whole island was methylated, because in theory, it's enough to methylate just a few critical sites, e.g. transcription factor binding sites. Also the dynamics of methylation was not clear. Stirzaker and colleagues (1997) re-analysed this area (27 CpGs in 161 bp containing the core promoter sequence) using bisulphite

sequencing, and found that every single CpG in the whole island was extensively methylated in tumours, but completely unmethylated in normal cells from the same patients. Moreover, examination of individual clones, derived from individual DNA molecules in the tumour sample, revealed a mosaicism of methylation pattern in one patient, although the whole tumour was presumably originated from a single clone. This demonstrated a continuing dynamics of the maintenance of methylation patterns.

The precision of bisulphite sequencing makes it the gold standard of methylation analysis. Even studies employing other high throughput methods normally check a selection of samples using bisulphite sequencing as confirmation. When Weber and colleagues (2005) found aberrant methylation at CGIs in tumour cells in the large scale MeDIP-chip study, they carried out bisulphite sequencing for four genes to comfirm the MeDIP data. Bisulphite sequencing can also be applied to large-scale studies. Instead of cloning and sequencing, the PCR products can be directly sequenced to give an averaged methylation profile. The Human Epigenome Project aims to sequence all human genes in all major tissues, in order to characterise the genome-wide DNA methylation profile for all human genes. An analysis of chromosomes 6, 20, and 22 provided high resolution methylation information from 2524 amplicons, comprising coding, non-coding, and evolutionarily conserved sequences associated with 873 genes in 12 different tissues (Eckhardt *et al.*, 2006).

In recent years, the development of next generation sequencing technologies makes it possible to carry out whole genome bisulphite sequencing. A first such study has just been published for the Arabidopsis epigenome (Lister *et al.*, 2008). Using sequencing-by-synthesis technology, highly integrated epigenome

maps were generated, revealing the interplay between DNA methylation, transcription and small RNA transcription. Currently, the application of this technology is still limited in its scope, because the short sequence length generated by the new sequencing methods presents a major obstacle for conventional sequence assembly algorithms, but with continuing advances in sequencing and bioinformatics, it will become more widely applied and provide unprecedented depth of understanding of DNA methylation.

4.1.1 Aims of this chapter

In the RPMA study, I identified hyper- and hypomethylated CGIs that are associated with X-inactivated and escapee genes. A much higher proportion of mouse genes were found to be associated with hypermethylated CGIs. I also found a third group of CGIs with intermediate levels of methylation. This methylation profile is not informative of the associated gene's XCI status, and interestingly, is specific to the human samples. I thus have the hypotheses that a) mouse CGIs are more densely methylated than human CGIs; and b) human CGIs are more variably methylated than mouse CGIs, as illustrated in Figure 4.2. To test these hypotheses, the methylation patterns need to be examined in much higher resolution, for which bisulphite sequencing provides the ideal method. Sodium bisulphite treatment 'fixes' the methylation pattern of genomic DNA by converting all unmethylated cytosine bases to uracil. The target region can then be PCR amplified, subcloned and sequenced. Each clone represents an individual DNA molecule in the original sample, so the proportion of methylated molecules, as well as single-base resolution CpG methylation state of each molecule, can be

examined (Figure 4.1).

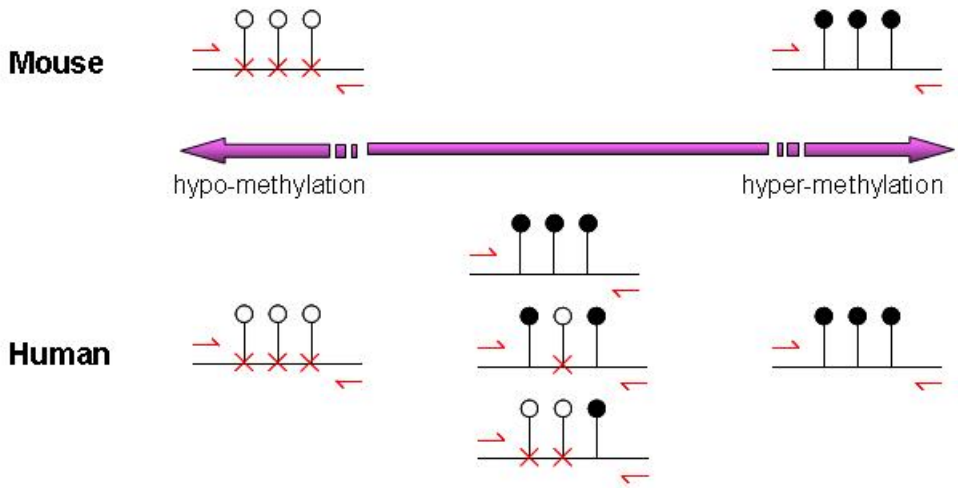

Figure 4.2: Hypothesised model of CGI methylation patterns on the human and mouse X chromosomes. Based on RPMA results, mouse CGIs are either hypomethylated or hypermethylated, but a proportion of human CGIs are intermediately methylated.

In this chapter, bisulphite sequencing is employed to study the detailed methylation profiles of a selected group of human and mouse CGIs. The aims include:

1. To verify methylation patterns interpreted from RPMA profiles.

2. To confirm differences in CGI methylation between the human and mouse X chromosomes.

3. To investigate whether the intermediate methylation pattern as revealed by RPMA was formed by a small portion of hypermethylated CGIs or by a large portion of intermediately methylated CGIs.

4.2 High resolution analysis of methylation levels on the mouse and human X chromosomes by bisulphite sequencing

4.2.1 Identification of target CGIs

For the purpose of confirming the exact methylation profiles that give rise to specific RPMA patterns observed in Chapter 3, I selected CGIs that displayed a range of methylation levels by RPMA in both human and mouse. To investigate the details of the intermediate methylation patterns in human samples, most islands were chosen from this category, giving a good representation of different kinds of intermediate PRMA patterns. The final selection consisted of CGIs associated with 16 human-mouse orthologous gene pairs. The human CGIs included three clearly hypomethylated CGIs, four clearly hypermethylated CGIs, and nine intermediately methylated CGIs that were associated with either inactivated or escapee genes according to Carrel and Willard (2005). The mouse CGIs included three clearly hypomethylated CGIs, nine clearly hypermethylated CGIs, had four hypermethylated CGIs with a faint *Hpa*II signal in one female or male sample (the closest to an intermediate pattern). RPMA results and XCI statuses (predicted from methylation status and recorded in literature) of the 16 human-mouse CGI pairs are summarised in Table 4.1.

Table 4.1: Human and mouse genes chosen for bisulphite sequencing analysis of CGI methylation. Information presented in this table is reproduced from Tables 3.1 and 3.2. XCI status is colour-coded: pink for X-inactivated, violet for escapee, and yellow for undetermined. The 'RPMA' columns contain XCI status prediction based on RPMA results described in Chapter 3, and the 'Lit' columns contain XCI status recorded in literature. Information about human genes is based on Carrel and Willard (2005), where the numbers of somatic cell hybrids retaining X_i that showed gene expression is presented. For most mouse genes, no record of XCI state can be found in the literature.

HUMAN	XCI status		RPMA primer pair 1			RPMA primer pair 2			MOUSE	XCI status		RPMA primer pair 1			RPMA primer pair 2		
			CCGG	Female	Male	CCGG	Female	Male				CCGG	Female	Male	CCGG	Female	Male
	RPMA	Lit.	sites	Hi Hp Ms	Hi Hp Ms	sites	Hi Hp Ms	Hi Hp Ms		RPMA	Lit.	sites	Hi Hp Ms	Hi Hp Ms	sites	Hi Hp Ms	Hi Hp Ms
Polarised methylation in both human and mouse																	
POLA	i	0 / 9	2	+ + -	+ - -	3	+ + -	+ - -	Pola1	i	i	3	+ + vf	+ - -			
UTX	e	0 / 9	2	+ - -	+ - -	2	+ - -	+ - -	Utx	e	e	3	+ - -	+ - -	2	+ - -	+ - -
ZFX	e	9 / 9	2	+ - -	+ - -	3	+ - -	+ - -	Zfx	i	i	2	+ + -	+ - -	3	+ + -	+ - -
Intermediate methylation in human; polarised methylation in mouse																	
PRPS2	i	0 / 5	2	+ f -	+ - -	3	+ + -	+ - -	Prps2	i		2	+ + vf	+ - -	2	+ + vf	+ - -
ATP6AP2	i	1 / 9	1	+ + -	+ - -	2	+ f -	+ - -	Atp6ap2	i		3	+ + -	+ - -	1	+ + -	+ - -
SMS	U	0 / 6	2	+ f -	+ - -	6	+ - -	+ - -	Sms	i		2	+ + -	+ - -	2	+ + -	+ - -
CDKL5	U	2 / 9	2	+ vf vf	+ - -	3	+ vf -	+ - -	Cdkl5	i		2	+ + -	+ - -	2	+ + -	+ - -
OFD1	U	6 / 6	2	+ f -	+ - -	3	+ f vf	+ - -	Ofd1	i		2	+ + -	+ vf -	3	+ + -	+ - -
RAB9A	U	9 / 9	4	+ f f	+ f f	1	+ f -	+ - -	Rab9	i		2	+ + -	+ - -	2	+ + -	+ - -
SYAP1	U	9 / 9	1	+ + +	+ f -	2	+ f f	+ f -	Syap1	i		1	+ + vf	+ - -			
DDX3X	U	9 / 9	4	+ f -	+ - -	2	+ f -	+ f f	Ddx3x	e	e	2	+ - -	+ - -	3	+ - -	+ - -
EIF2S3	U	9 / 9	2	+ - -	+ - -	1	+ f -	+ - -	Eif2s3x	e	e	1	+ - -	+ - -	2	+ - -	+ - -
Polarised methylation in human; intermediate methylation in mouse																	
TMEM47	i	0 / 9	3	+ + vf	+ - -	3	Failed	Failed	Tmem47	i		5	+ f -	+ - -	2	+ + -	+ - -
HCCS	i	0 / 9	5	+ + -	+ - -	2	+ + -	+ - -	Hccs	i		3	+ + -	+ - -	3	+ + -	+ f -
MSL3L1	i	3 / 9	3	+ + -	+ - -	6	+ + -	+ - -	Msl3l1	i		2	+ + -	+ vf -	3	+ + -	+ - -
CRSP2	e	6 / 6	3	+ - -	+ - -	5	+ - -	+ - -	Crsp2	i		2	+ + -	+ - -	2	+ f f	+ f f

4.2.2 Primer design and PCR

CGI sequences, together with 500 bp upstream and downstream sequences, were repeatmasked and bisulphite converted *in silico*, assuming that all CpGs were methylated. This converted sequence was used to design primers to amplify 400-800 bp overlapping fragments covering the entire CGI. The CpG dinucleotides were masked with 'XX' for primer design purpose so that primers would contain no CpGs, thus ensuring that hypo- and hypermethylated sequences would be amplified with similar efficiencies. Primers were designed using Primer3 with relaxed parameters to accommodate the very unusual sequence composition: melting temperature was between 55 and 65 °C, and primer length was between 18 and 27 bases. For CGIs of human genes *POLA*, *ZFX*, *ATP6AP2*, *RAB9A*, *EIF2S3*, and *CRSP2*, primers were provided by Christine Burrows (personal communication). All primer combinations and their predicted amplicon sizes are listed in Appendix II.

Human and mouse genomic DNA were extracted from female and male cultured fibroblast cells and treated with sodium bisulphite. The bisulphite modification was carried out using the EZ DNA methylation kit™ (Zymo Research) following manufacturer's recommendations (section 2.10.5). Because of the heterogeneous nature of CpG methylation, the bisulphite-converted DNA contains a variety of sequences with different cytosines converted or unconverted, so it was not possible to check primer specificity through BLAST against a reference sequence database. Therefore primers were used directly in PCR with female and male bisulphite-converted DNA using previously optimised PCR conditions (Vardhman Rakyan, personal communication, and section 2.11.3). A fraction of

each PCR product was visualised using gel electrophoresis. PCR products providing the best coverage for each CGI were used for subcloning and sequencing. At least one pair of primers was successful for each CGI.

4.2.3 Cloning and sequencing

The selected PCR products were column-purified using a PCR purification kit (section 2.10.1) and ligated with the plasmid pGEM®-T Easy (section 2.12.1). Plasmid DNA was introduced into Mach1 cells by chemical transformation (section 2.12.2), and seven white colonies from each ligation were checked for successful subcloning of the correctly-sized PCR product using colony PCR (section 2.11.4). If more than two colonies contained the right product, the clone was progressed into sequencing. For each successful ligation, up to 384 clones were sequenced (by the Faculty Small Sequencing Projects team at the Wellcome Trust Sanger Institute, section 2.13). The resulting sequences were processed using Gap4 (section 2.14.4) prior to methylation analysis, as follows: Sequences were aligned to the *in silico*bisulphite-converted sequence (the 'reference'), the flanking vector sequences were removed, and low quality sequences were discarded. Forward and reverse reads of the same sequence were joined and sequences missing more than 10% of the expected full length were discarded. All sequences from the same ligation, each representing an individual CGI molecule, were exported into a single text file, together with the genomic and the reference sequences.

4.2.4 Methylation analysis

Further sequence processing was carried out to fulfil the strict input format requirement of the subsequent methylation analysis. The sequences of each analysis region needed to be correctly aligned, of equal length (padded with '-' wherever necessary), and to contain only the characters a, t, c, g, and n. After experimenting with a number of alignment programs, MUSCLE (Labarga *et al.*, 2007) was found to give best alignment for my purpose and was used in all subsequent analysis. To ensure accurate methylation analysis, each CpG dinucleotide has to be located at exactly the same position in each sequence, and the C and G should not be separated by one or more '-', so the alignments were manually checked in Genedoc (section 2.14.5) for correct alignment of all CpGs. Methylation analysis was carried out using a modified version of MethTools, a collection of open source Perl scripts that makes graphical representation of bisulphite sequencing results and calculates methylation densities (Grunau *et al.*, 2000). Instead of using the web interface, the scripts were downloaded and modified as follows. The original program can analyse several forms of cytosine methylation. For this study, only CpG methylation was considered. For better visualisation of the methylation distribution, clones were ordered by methylation level in the graphical representation. A shell script was also created to automate the analysis. The source code of all scripts used in this study is included in Appendix IV and all modifications are logged in the code. The reference sequence was included in every output graph (named 'EXP' for expected) to demonstrate the expected positions of CpG dinucleotides. An example of the graphical representation of bisulphite sequencing results is shown in Figure 4.3.

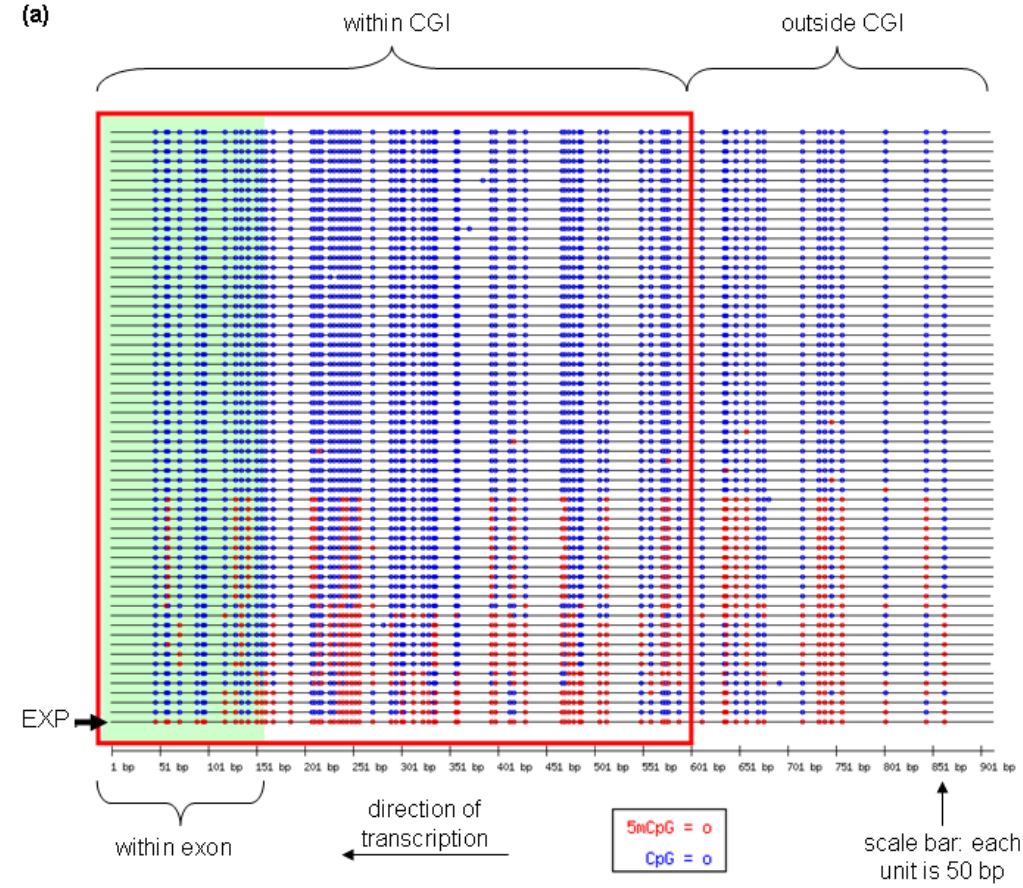

Figure 4.3: Graphical representation of CGI methylation. (a) shows CGI methylation pattern of a gene inactivated by XCI in female. Each line represents an individual molecule with unmethylated (blue) and methylated (red) CpG dinucleotides (circles). The bottom line ('EXP') represents a reference sequence where every expected CpG is methylated in silico (thus all red). The island region is framed by the red box and exon is shadowed in green. Direction of transcription is indicated by arrow. (b) shows CGI methylation pattern of the same gene, but in a male sample, where CGIs on the only X chromosome are expected to be unmethylated. Note that all molecules are comprised of mostly unmethylated CpGs (blue), except for the artificial reference sequence at the bottom, which demonstrates a fully methylated molecule.

4.2.5 Comparison between human and mouse methylation profiles

Bisulphite sequencing was successful for ten human CGIs and ten mouse CGIs in both female and male samples (Figures 4.4-4.15). Eight human CGIs and eight mouse CGIs form orthologous pairs (Figures 4.4-4.11). One or two amplicons were sequenced for each CGI, and at least 40 individual molecules, each representing the island of an individual chromosome, were sequenced for most amplicons.

The single X chromosome in male remains active, so CGIs on the male X chromosome are expected to be free from methylation. As expected, in both human and mouse, all male CGIs are clearly hypomethylated, regardless of the methylation situation in the corresponding female sample (Figures 4.4-4.15). For each male CGI, the vast majority of island molecules are completely free from methylation, and the remaining cases only have one or a few CpGs methylated.

The female CGIs of the eight orthologous pairs can be divided into three groups according to their behaviour in human and mouse.

The CGIs of *DDX3X/Ddx3x*, *UTX/Utx* and *EIF2S3/Eif2s3x* were also hypomethylated in female samples in both human and mouse (Figures 4.4-4.6). Like their male companions, the female CGIs of the three known escapees had the majority of island molecules completely unmethylated, apart from a small number with one or two methylated CpGs. Only three molecule of the human *EIF2S3* CGI (n=31) had more than two CpGs methylated in females.

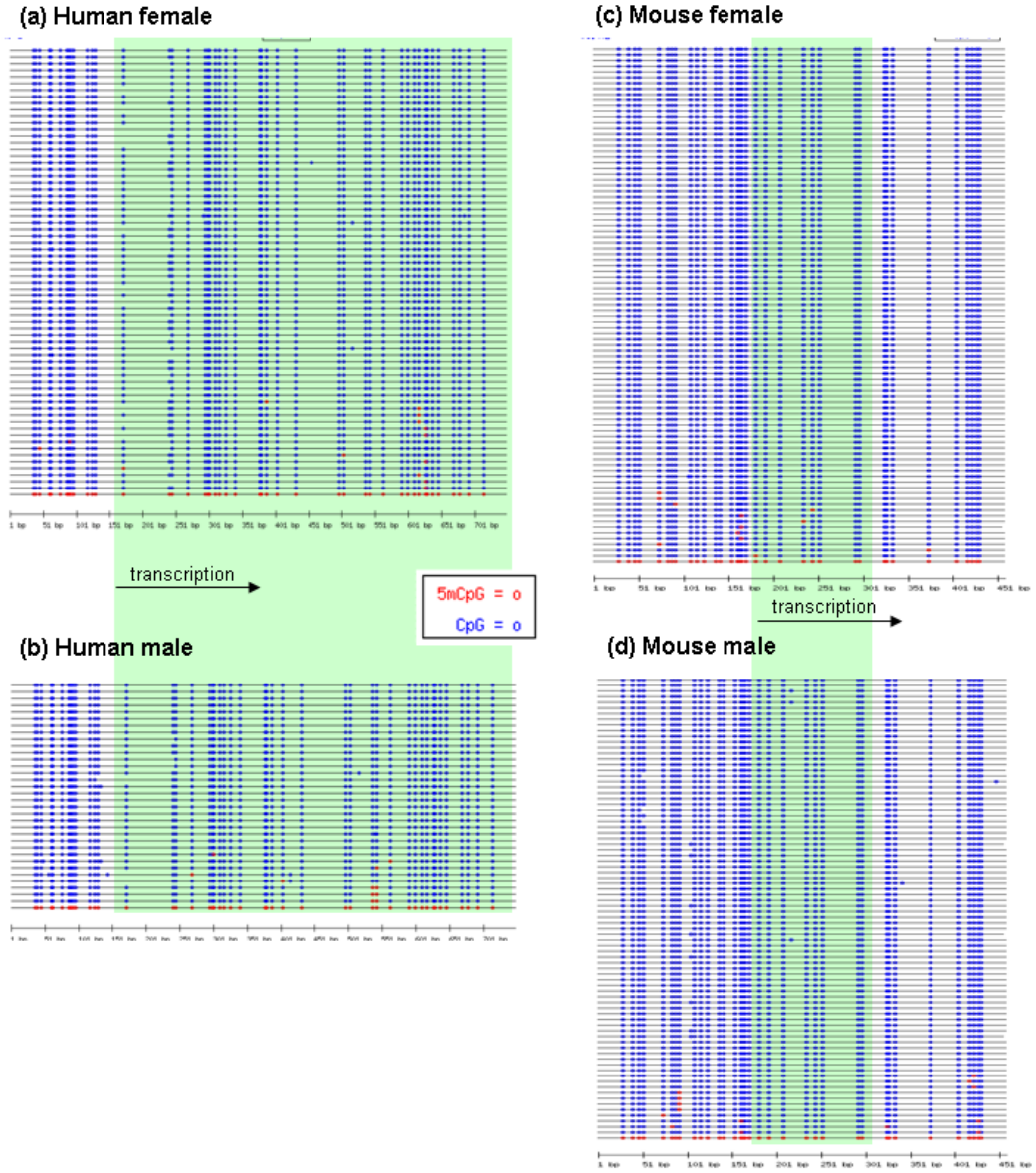

Figure 4.4: CGI methylation profiles of human *DDX3X* and mouse *Ddx3x*. Both genes escape from XCI. The human CGI had a faint signal in the *Hpa*II lane, indicating low level methylation, but the mouse CGI was shown to be hypomethylated in RPMA. The human fragment covers the middle third of the CGI and the mouse fragment covers the first half of the CGI. Exons are shadowed in green. Directions of transcription are indicated by arrows. For detailed annotation see Figure 4.3.

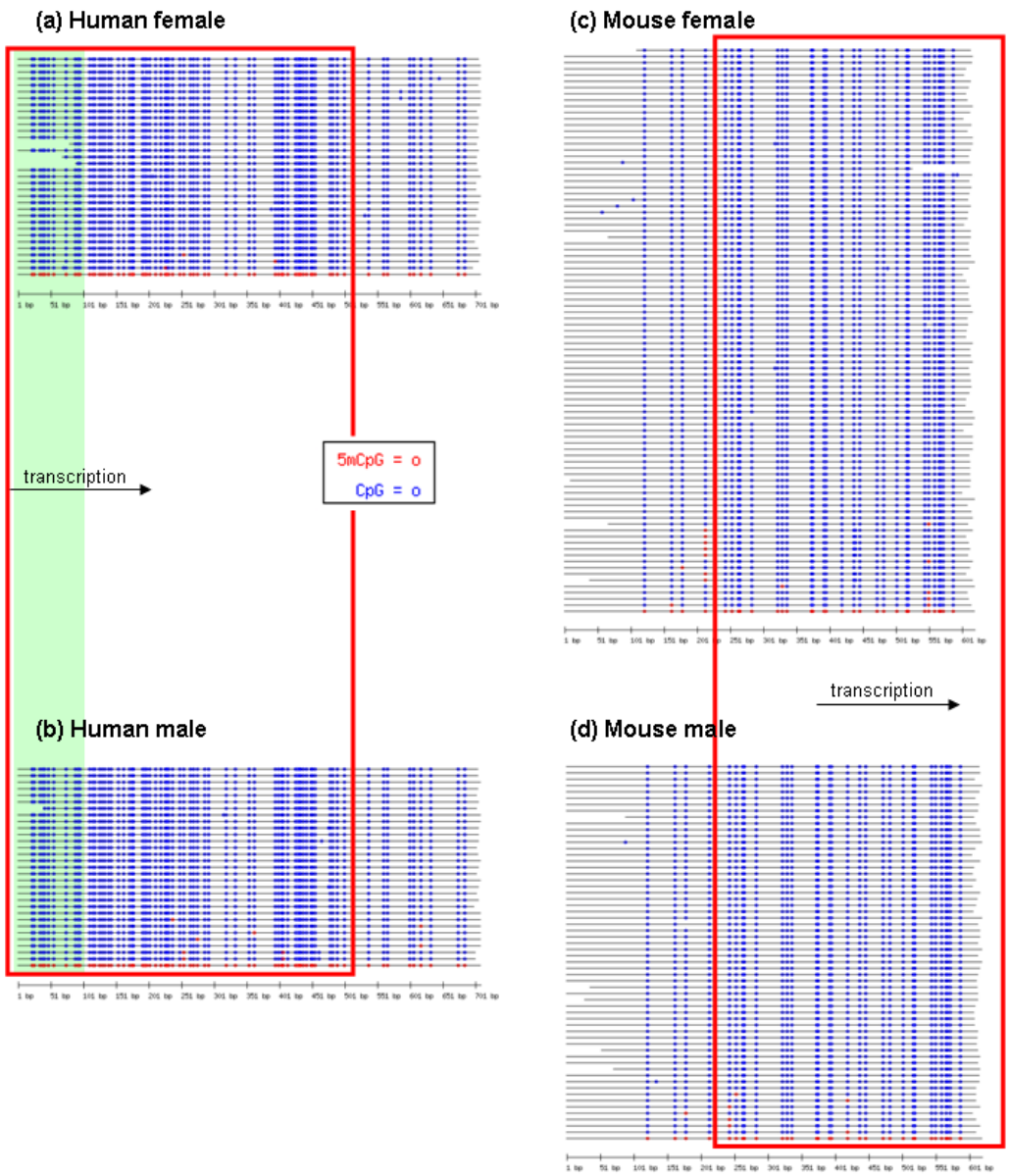

Figure 4.5: CGI methylation profiles of human *UTX* and mouse *Utx*. Both genes escape from XCI and both CGIs were shown to be hypomethylated in RPMA. The human fragment covers the end third of the CGI and the mouse fragment covers the first quarter of the CGI. The island regions are indicated by the red boxes and exons are shadowed in green. Directions of transcription are indicated by arrows. For detailed annotation see Figure 4.3.

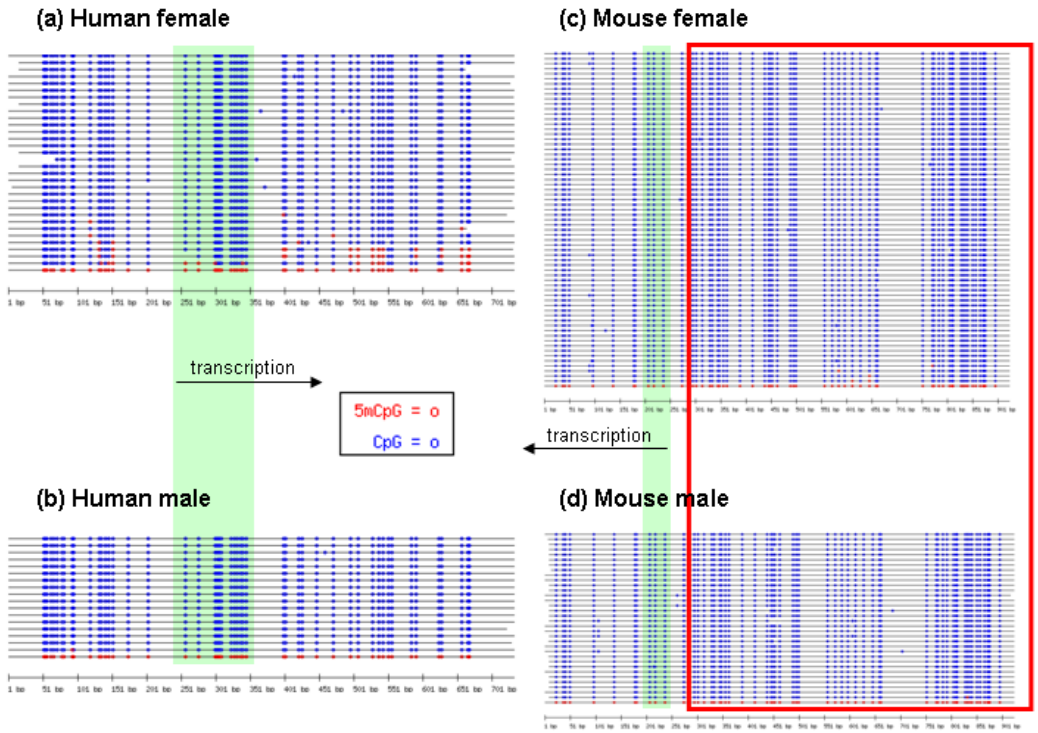

Figure 4.6: CGI methylation profiles of human *EIF2S3* and mouse *Eif2s3x*. Both genes escape from XCI. The human CGI had a faint signal in the *Hpa*II lane, indicating low level methylation, but the mouse CGI was shown to be hypomethylated in RPMA. The human fragment covers the entire CGI and the mouse fragment covers almost the entire CGI (missing the first three CpGs). The island region is indicated by the red box and exons are shadowed in green. Directions of transcription are indicated by arrows. For detailed annotation see Figure 4.3.

The CGI of the human gene *OFD1* was also hypomethylated in both sexes, but the CGI of the mouse orthologue *Ofd1* showed striking methylation difference between female and male samples (Figure 4.7). Around half of the *Ofd1* island molecules were hypomethylated like their male companions, but the remaining island molecules had more than 30% of all CpGs methylated. This is the picture that would be expected if half of the clones were derived from methylated CGIs on the inactive X.

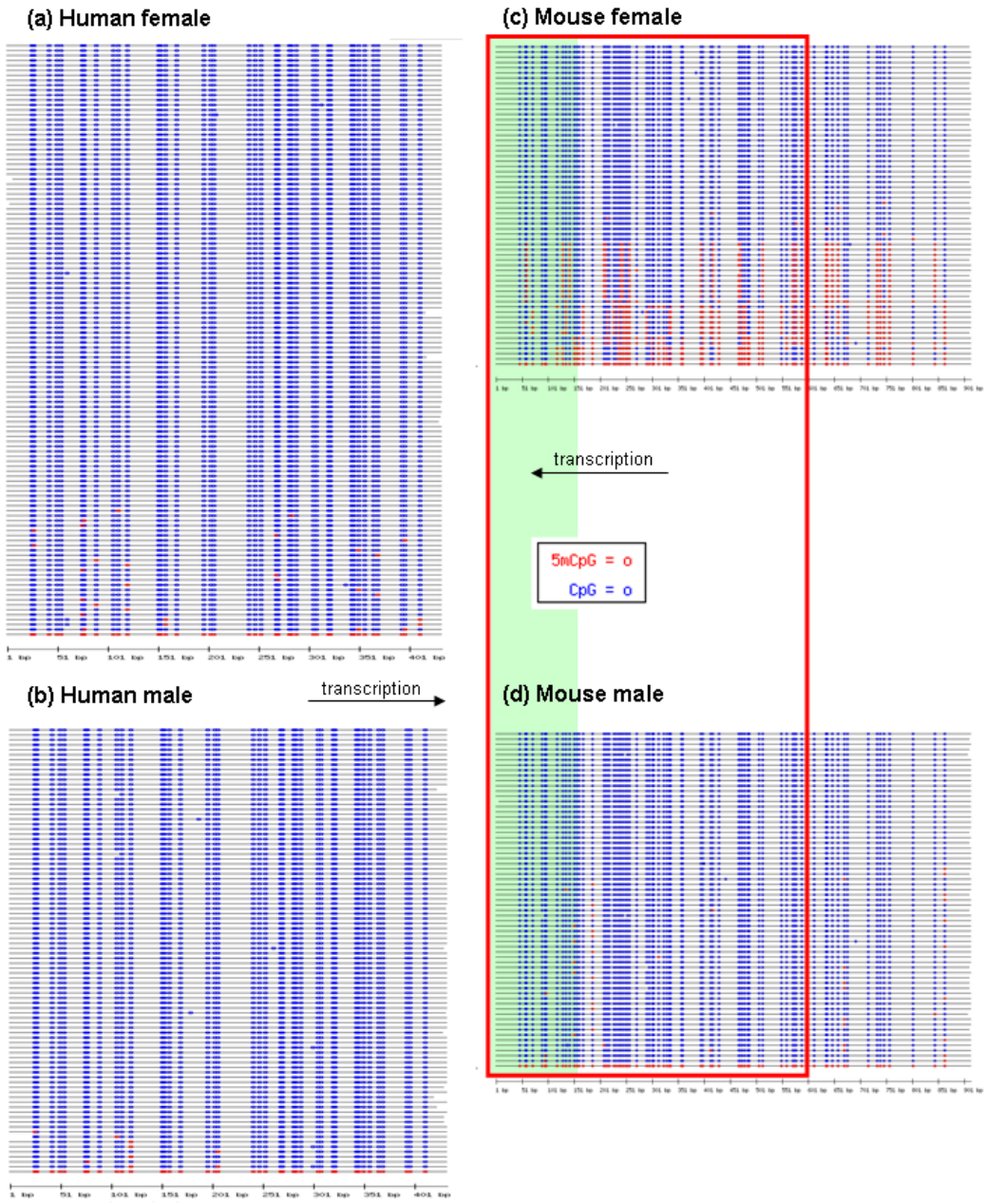

Figure 4.7: CGI methylation profiles of human *OFD1* and mouse *Ofd1*. The human gene escapes from XCI. The human CGI had a faint signal in the *Hpa*II lane, indicating low level methylation, while the mouse CGI was shown to be hypermethylated in RPMA. The human CGI also overlaps the 5' region of *TRAPPC2*, the mouse orthologue for which is not X-linked. The human fragment covers the first half of the CGI and the mouse fragment covers most of the CGI. The island regions is indicated by the red box and exon is shadowed in green. Directions of transcription are indicated by arrows. For detailed annotation see Figure 4.3.

For the remaining four orthologous pairs, the CGIs displayed female-specific methylation in both human and mouse, but to different extents (Figures 4.8-4.11). In each case, the female island molecules can be easily divided into two groups based on their methylation levels: one group resembling the male pattern, possibly derived from the inactive X, and the other with considerably higher levels of methylation, presumably derived from the active X. The CGI of *POLA/Pola1* showed very similar methylation patterns in human and mouse (Figure 4.8). The human CGI had 65% of island molecules methylated, where 41-77% of all CpGs were methylated. The mouse CGI had a slightly higher proportion (72%) of methylated molecules, but the methylation levels of individual molecules (41-78%) were similar to those of human. Difference in methylation levels between human and mouse is more obvious for the CGI of *MSL3L1/Msl3l1* (Figure 4.9). The human CGI had 39% of the female island molecules methylated, where 48-65% of all CpGs were methylated. A higher proportion of the female island molecules were methylated for the mouse CGI, where the two amplicons had 60% and 44% of female island molecules methylated respectively. Heavier methylation was also noted for the mouse island, where up to 100% of all CpGs were methylated. Similar pattern was seen for the CGI of *ATP6AP2/Atp6ap2* and *PRPS2/Prps2* (Figures 4.10 and 4.11). Whereas female-specific methylation was seen in both species, a higher proportion of the female island molecules were methylated in mouse, and the mouse islands were methylated to much greater extents.

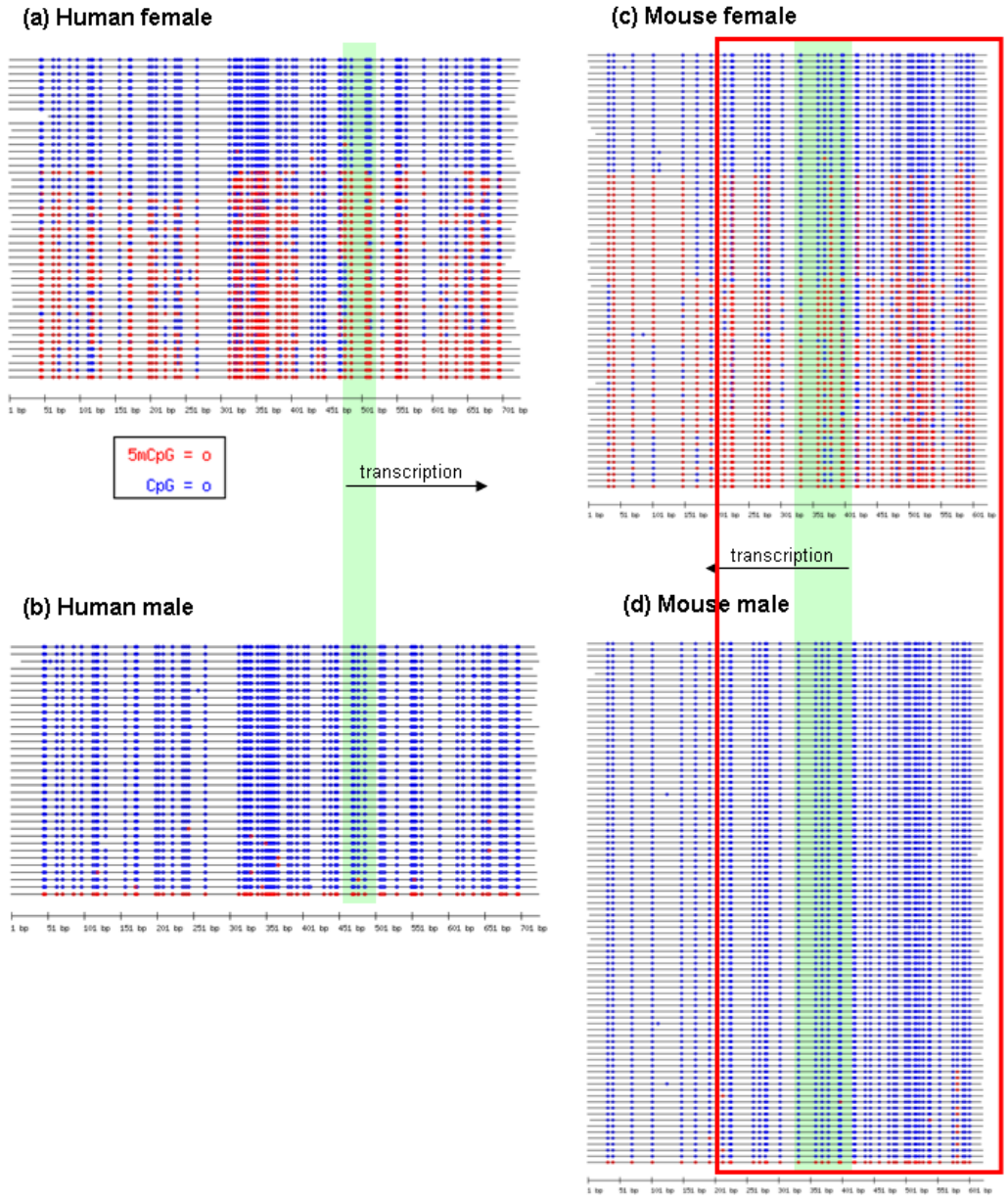

Figure 4.8: CGI methylation profiles of human *POLA* and mouse *Pola1*. Both genes are X inactivated and both CGIs were shown to be hypermethylated in RPMA. The human fragment covers the entire CGI and the mouse fragment covers most of the CGI. The island region is indicated by the red box and exons are shadowed in green. Directions of transcription are indicated by arrows. For detailed annotation see Figure 4.3.

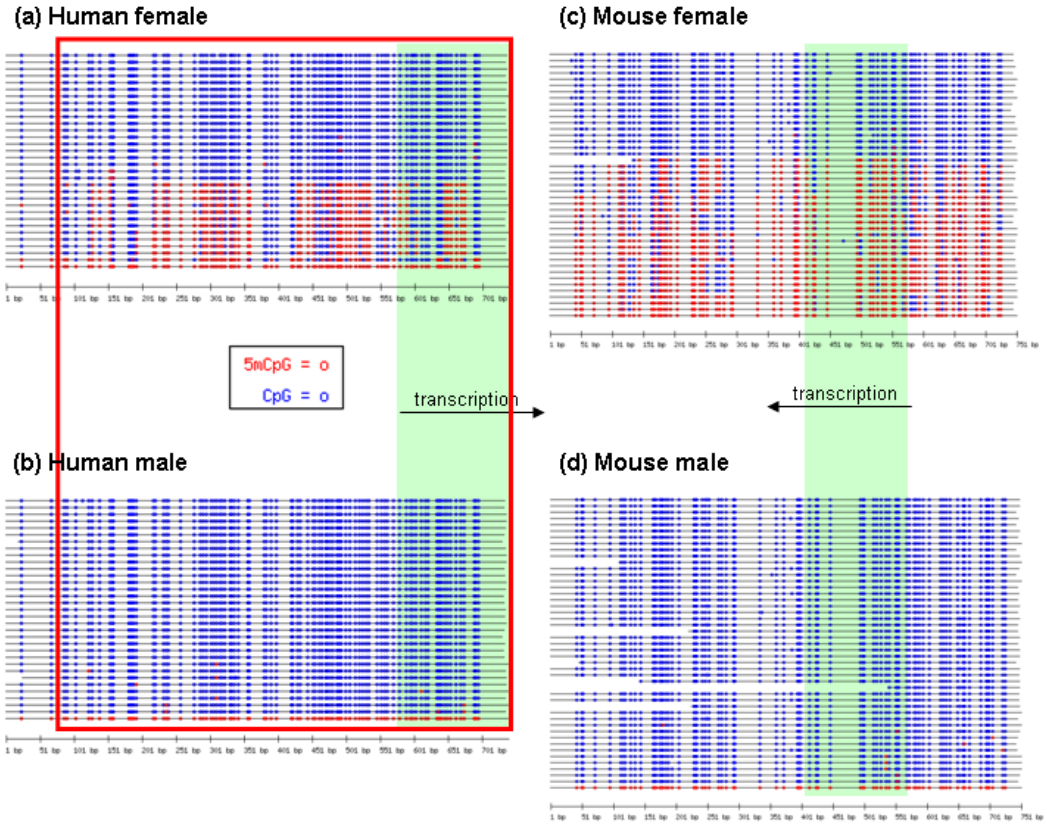

Figure 4.9: CGI methylation profiles of human *MSL3L1* and mouse *Msl3l1*. The human gene is X inactivated and both CGIs were shown to be hypermethylated in RPMA. The human fragment covers the first half of the CGI and the mouse fragment covers the first two thirds of the CGI. The island region is indicated by the red box and exons are shadowed in green. Directions of transcription are indicated by arrows. For detailed annotation see Figure 4.3.

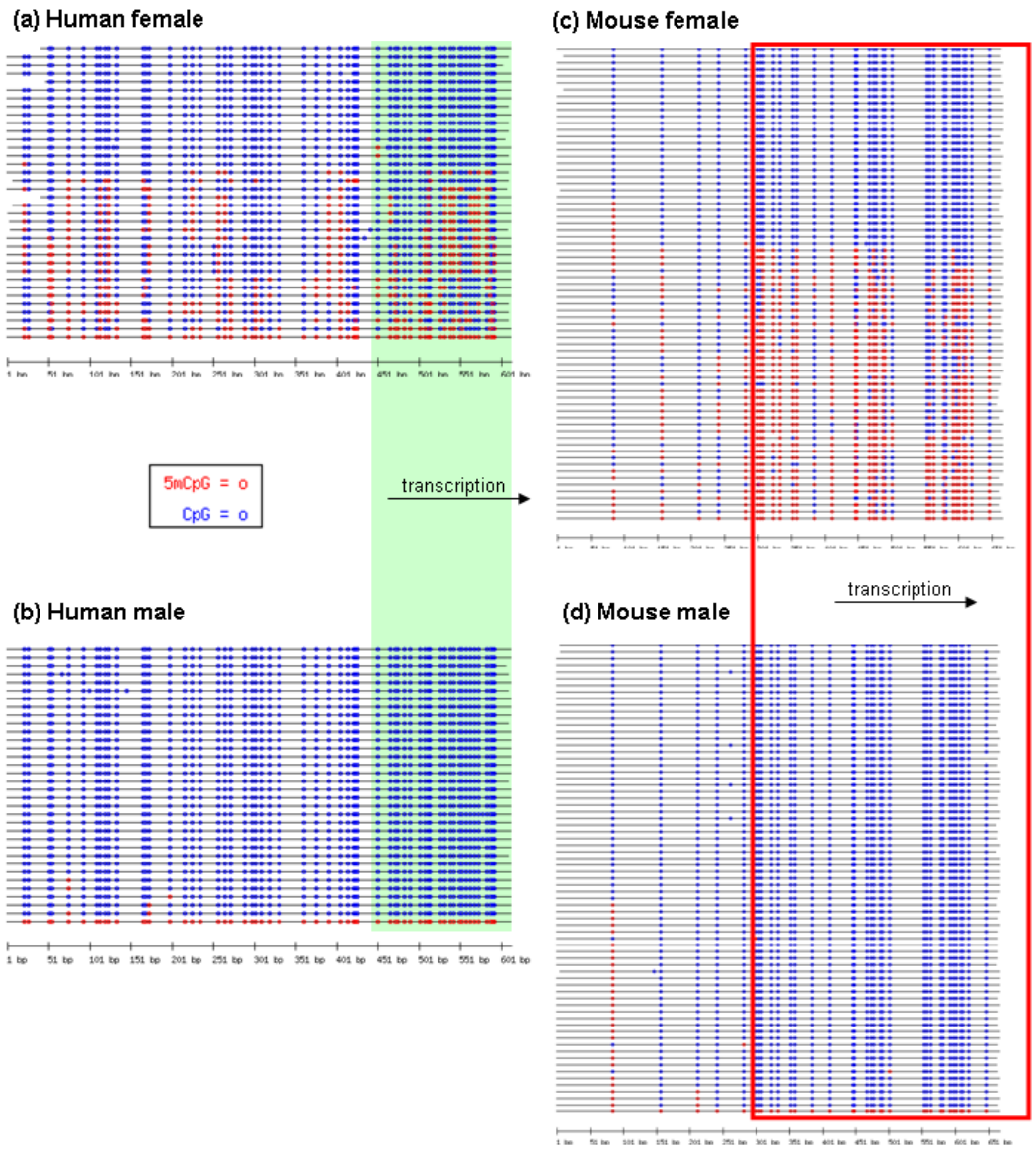

Figure 4.10: CGI methylation profiles of human *ATP6AP2* and mouse *Atp6ap2*. The human gene is X inactivated. Both CGIs were shown to be hypermethylated in RPMA but the human CGI had a faint signal in the *Hpa*II lane of one amplicon. Both fragments cover the first half of the CGI. The island region is indicated by the red box and exon is shadowed in green. Directions of transcription are indicated by arrows. For detailed annotation see Figure 4.3.

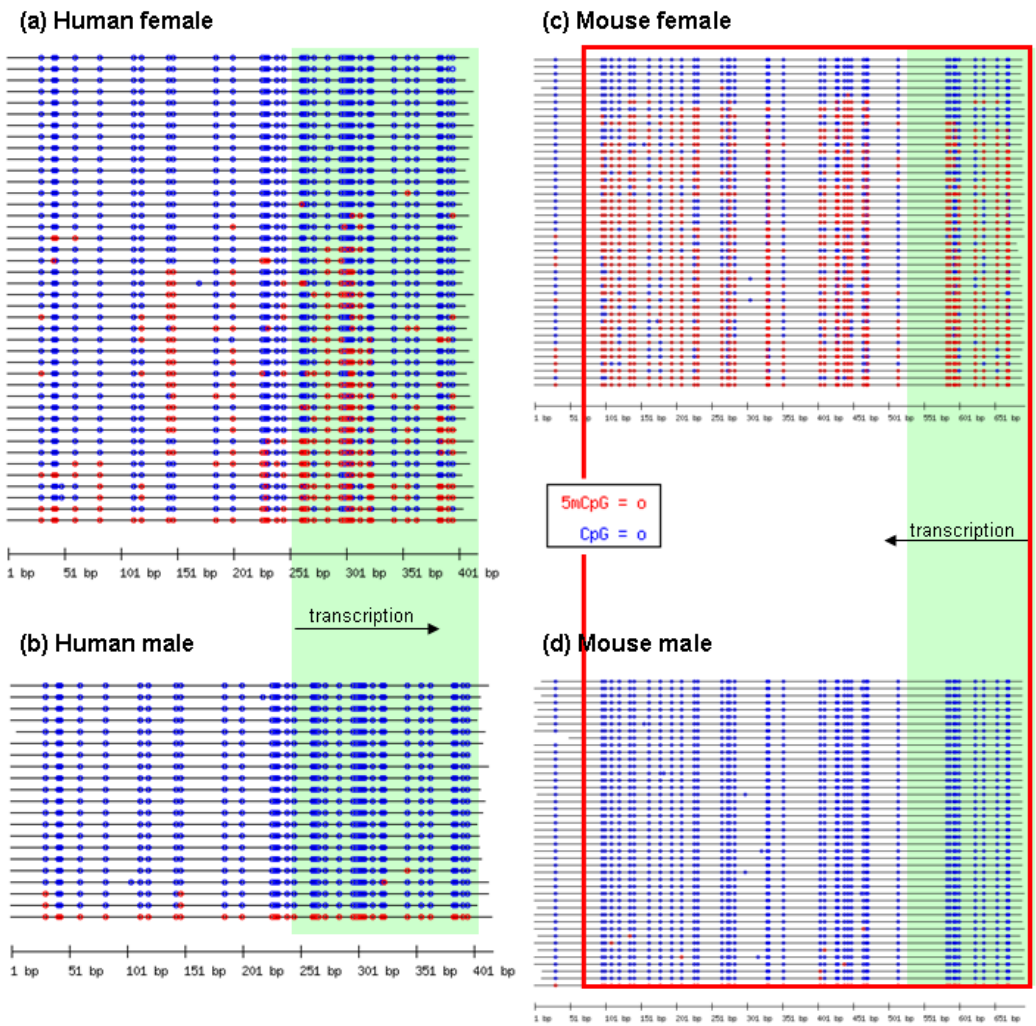

Figure 4.11: CGI methylation profiles of human *PRPS2* and mouse *Prps2*. The human gene is X inactivated. Both CGIs were shown to be hypermethylated in RPMA but the human CGI had a faint signal in the *Hpa*II lane of one amplicon. The human fragment covers the first half of the CGI and the mouse fragment covers most of the CGI. The island region is indicated by the red box and exons are shadowed in green. Directions of transcription are indicated by arrows. For detailed annotation see Figure 4.3.

Two CGIs in human and two CGIs in mouse had no information available from the orthologous genes in the other species. The CGI of the human escapee *RAB9A* was clearly hypomethylated in both sexes (Figure 4.12), but the CGI of

the other human escapee *CDKL5* had a small number of female island molecules with low level methylation (Figure 4.13). The CGIs of the mouse genes *Syap1* and *Hccs* (Figures 4.14 and 4.15) were like the ones shown in Figures 4.7-4.11. A good proportion of the female island molecules were methylated to great extents.

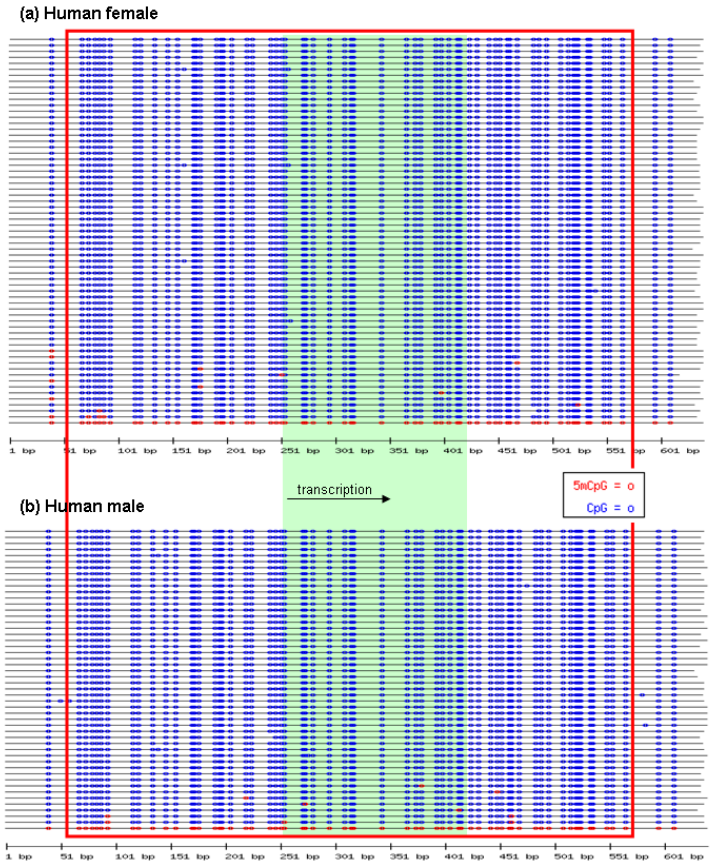

Figure 4.12: CGI methylation profiles of human *RAB9A*. This gene escapes from XCI and its CGI had a very faint signal in the *Hpa*II lane in RPMA. The bisulphite sequenced fragment covers the entire CGI. The island region is indicated by the red box and exon is shadowed in green. Direction of transcription is indicated by arrow. For detailed annotation see Figure 4.3.

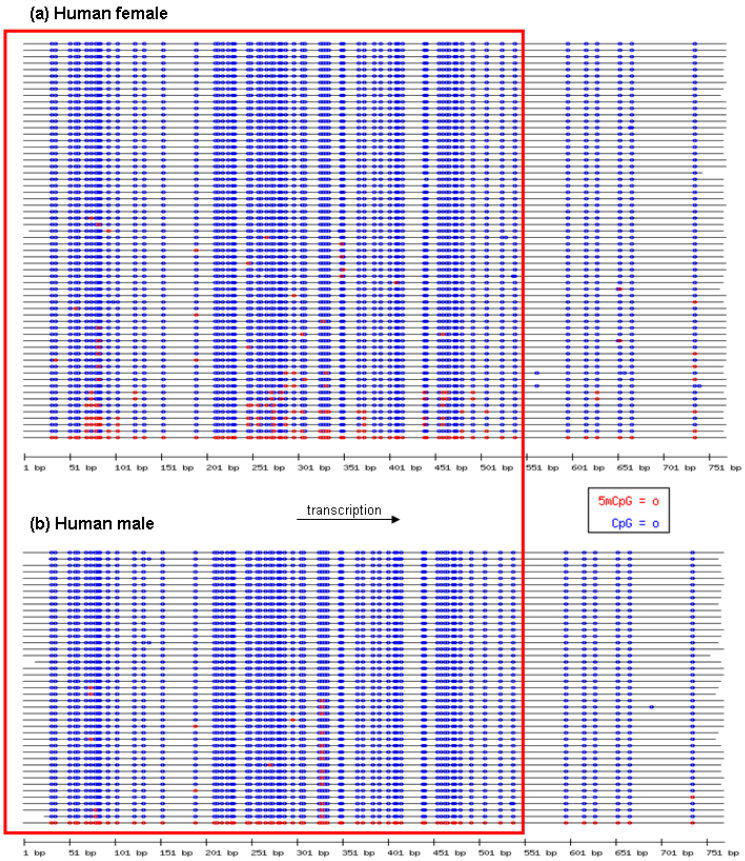

Figure 4.13: CGI methylation profiles of human *CDKL5*. This gene escapes from XCI. Its CGI was shown to be hypomethylated in RPMA, but had had a very faint signal in the *Hpa*II lane. The bisulphite sequenced fragment covers the second half of the CGI and is in intron1. The island region is indicated by the red box . Direction of transcription is indicated by arrow. For detailed annotation see Figure 4.3.

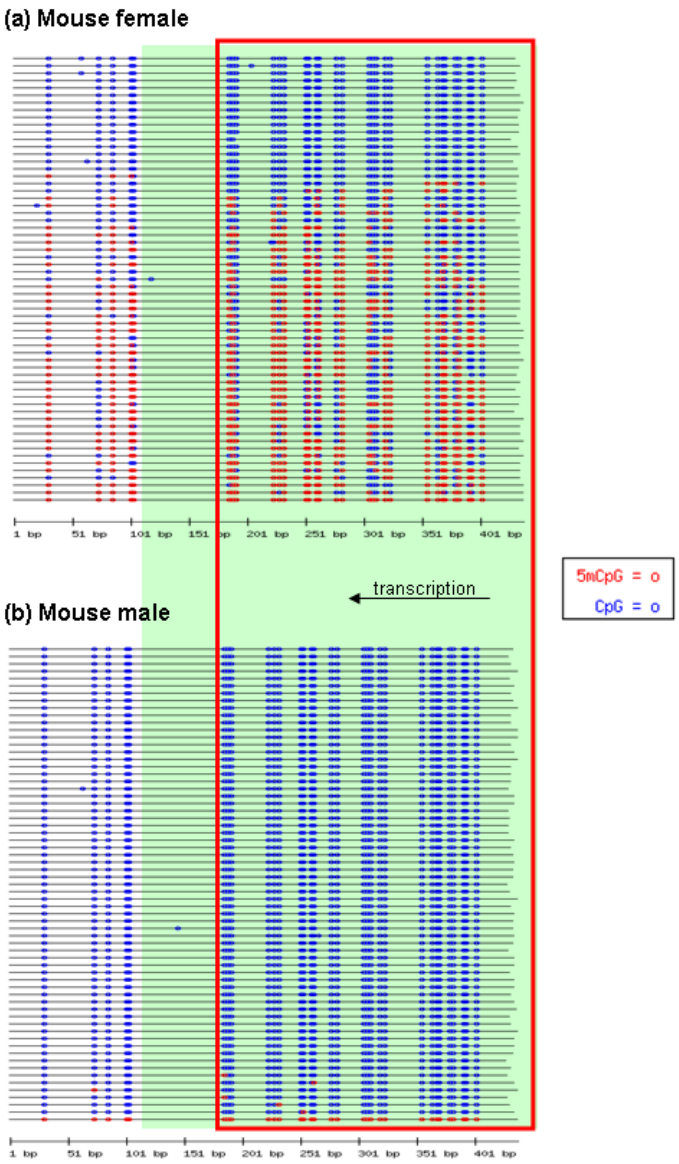

Figure 4.14: CGI methylation profiles of mouse *Syap1*. The CGI was shown to be hypermethylated in RPMA. The bisulphite sequenced fragment covers the first half of the CGI. The island region is indicated by the red box and the exons is shadowed in green. Direction of transcription is indicated by arrow. For detailed annotation see Figure 4.3.

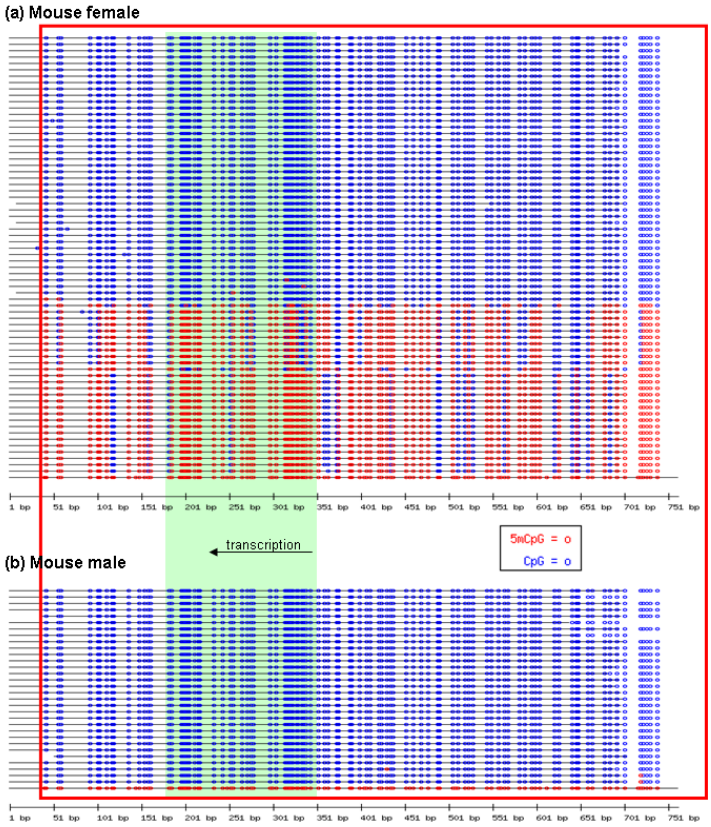

Figure 4.15: CGI methylation profiles of mouse *Hccs*. The CGI was shown to be hypermethylated in RPMA. The bisulphite sequenced fragment covers almost the entire CGI. The island region is indicated by the red box and the exons is shadowed in green. Direction of transcription is indicated by arrow. For detailed annotation see Figure 4.3.

4.2.6 Comparison between bisulphite sequencing and RPMA results

In total, results were compared from bisulphite sequencing and RPMA analysis for 12 RPMA targets in six human CGIs and 11 RPMA targets in eight mouse CGIs (summarised in Table 4.2, note that some RPMA fragments are not contained in available bisulphite sequences). For each RPMA target, the assayed CCGG sites

were identified in the bisulphite sequencing results, and a template availability value was calculated, which is the percentage of clones in which these CCGG sites were fully methylated (thus available as PCR template).

This comparison has shown the limitations of RPMA. When the *Hpa*II band is absent in the RPMA assay, the template availability according to bisulphite sequencing is zero. The same is also true in the two (male) cases where the *Hpa*II band is classified as very faint (vf). In cases where the *Hpa*II band was comparable with the *Hind*III band (+), the template is always present, but the template availability ranges from a fairly low 8.6% to as high as 61.7%, encompassing a wide range of methylation levels. The RPMA results are least informative of the true methylation levels when the *Hpa*II band was classified as faint (f). Most such cases are characterised by a zero template availability, though there are three examples where the figure lies between 2.4% and 8.6%.

Table 4.2: Comparison between bisulphite sequencing and RPMA results. Sites assayed in RPMA were identified in bisulphite sequencing results, and the percentage of clones that would support a *Hpa*II (Hp) signal in RPMA was calculated (Template). RPMA information is reproduced from Tables 3.1 and 3.2, where '+' = strong band; 'f' = faint band; 'vf' = very faint band; '-' = no visible band.

	RPMA primer pair 1				RPMA primer pair 2			
Gene	Female		Male		Female		Male	
	Hp signal	Template	Hp signal	Template	Hp signal	Template	Hp signal	Template
Human								
POLA	+	11.1%	-	0.0%	+	42.2%	-	0.0%
PRPS2	f	2.4%	-	0.0%	+	N/A	-	N/A
ATP6AP2	+	11.4%	-	0.0%	f	8.6%	-	0.0%
OFD1	f	N/A	-	N/A	f	0.0%	-	0.0%
RAB9A	f	0.0%	f	0.0%	f	0.0%	-	0.0%
EIF2S3	-	0.0%	-	0.0%	f	6.5%	-	0.0%
Mouse								
Pola1	+	45.0%	-	0.0%				
Prps2	+	67.4%	-	0.0%	+	19.6%	-	0.0%
Atp6ap2	+	8.6%	-	0.0%	+	61.3%	-	0.0%
Ofd1	+	58.8%	vf	0.0%	+	41.2%	-	0.0%
Syap1	+	61.7%	-	0.0%				
Eif2s3x	-	0.0%	-	0.0%	-	0.0%	-	0.0%
Hccs	+	37.7%	-	0.0%	+	37.7%	f	0.0%
Msl3l1	+	40.0%	vf	0.0%	+	45.0%	-	0.0%

4.3 Discussion

The initial RPMA study, described in Chapter 3, indicated that the majority of CpG islands on the mouse X_i are heavily methylated, with a small percentage associated with genes escaping from XCI being hypomethylated. The data from analysis of human CGIs, by contrast, indicated both a larger percentage of genes with unmethylated CGIs and a greater variation in the methylation levels of the methylated CGIs. To investigate this variation in greater detail, bisulphite sequencing was carried out on 32 CGIs (16 human-mouse orthologous pairs) with varied methylation profiles in human and mouse. Ten human and ten mouse CGIs, including eight orthologous pairs, were successfully bisulphite sequenced in both female and male samples.

Before analysing the bisulphite sequence data for methylation status, it is first important to confirm the efficiency of bisulphite conversion. In almost all sequences, cytosines were almost completely absent outside the CpG dinucleotides. Therefore the conversion was very efficient and the resulting sequence data are trustworthy. As a further precaution to ensure any CpG dinucleotide seen in the final sequence was a true representation of methylation, it was investigated whether or not bisulphite conversion is less efficient when an unmethylated C is part of a CpG. CGIs on the only X chromosome in males are expected to be unmethylated, thus provide a perfect substrate to test the efficiency of bisulphite conversion and the ability to clone and sequence DNA molecules with such unusual sequence composition. Indeed the bisulphite sequences of all the male samples are almost completely devoid of cytosines. Given the extremely low frequency of cytosine appearance in general (data not shown), it is highly

likely that the CG dinucleotides observed in male samples are indicative of very occasional CpG methylation.

In using the bisulphite sequencing approach to study CGIs on the female X chromosomes, there is an assumption that approximately equal numbers of molecules studied in a large sample will be derived from the active and the inactive X chromosome (Stöger *et al.*, 1997). However, the possibility of PCR bias because of differences in sequence composition should be considered. If the test region contains molecules with vastly different methylation states, the cytosine content will differ substantially between molecules in the bisulphite-converted DNA. In the case of CGIs associated with X-inactivated genes, where half of the molecules are expected to be unmethylated but the other half hypermethylated, we would expect half of the molecules to be extremely T-rich, and the other half relatively C-rich. It is possible that one group of sequences will be amplified preferentially. Warnecke and colleagues (1997) studied such PCR bias in a number of regions and found that the bias to be sequence-dependent, but was mostly towards amplification of unmethylated DNA. This finding is consistent with PCR bias observed in a number of bisulphite sequencing studies, both on autosomal and X-linked genes (Stöger *et al.*, 1997; Stirzaker *et al.*, 1997). In the results of this study, the methylated clones in female samples tend to make up around half of the total number, consistent with expectation, so it is unlikely that such PCR bias occurred. Additional support comes from the RPMA data. Wherever methylation is suggested by RPMA, it was also observed in the bisulphite sequencing results, so complete bias against amplification of the methylated molecules could be ruled out.

The first aim of this thesis was to confirm whether the RPMA data represent

true patterns of methylation. As seen in Figures 4.4-4.15, in both the human and the mouse samples, the bisulphite sequencing results are mostly consistent with methylation patterns interpreted from the RPMA results. A negative *Hpa*II signal in RPMA is always associated with absence of methylation in the bisulphite sequencing results, whereas a strong *Hpa*II signal in RPMA is in most cases backed up by considerable methylation. However, the limitations of the RPMA approach are also obvious, as revealed by Table 4.2. Intermediate RPMA patterns, characterised by presence of faint bands in the *Hpa*II lane, do not correlate with a consistent methylation profile revealed by bisulphite sequencing. In addition, details of methylation levels are lost in the RPMA results as a wide range of methylation densities can give rise to the same strong RPMA signal. In general, bisulphite sequencing results confirm that RPMA is a useful method of studying methylation, most suitable for an initial screen.

The major purpose of this detailed methylation analysis was to confirm that the CGI methylation differences suggested by RPMA between human and mouse are real and not artefactual. In support to the RPMA results, all but three mouse genes investigated showed female-specific CGI methylation, while a much larger number of human genes were hypomethylated at their CGIs in female samples. In addition, variation of methylation densities were compared between the human and mouse CGIs. In all samples that showed female-specific methylation, the female island molecules consisted of two distinct groups, one group unmethylated, and the other methylated. This bimodal distribution of methylation densities is consistent with the samples containing active and inactive alleles and was observed also in a previous study of the X-linked gene *FMR1* (Stöger *et al.*, 1997). Another noticeable feature is the heterogeneity of methylation: within the

same island, the methylated molecules always presented a diversity of methylation densities. Such methylation mosaicism was also recorded in the previous study (Stöger *et al.*, 1997) and has been proposed to manifest the dynamic nature in maintaining an overall stable methylation level (Genereux *et al.*, 2005).

A wide range of methylation densities were found in the female CGIs of both human and mouse, but the methylation densities of the mouse islands tend to occupy the denser end of the spectrum, with most medians above 60%. One molecule of the *Msl3l1* CGI was even fully methylated. In contrast, no a single one of the hundreds of human island molecules sequenced had more than 80% CpGs methylated, and most human CGIs had methylation density medians below 50%. When the homologous human-mouse CGIs are compared, it is also clear that, when the island was methylated in both species, the mouse samples always showed heavier methylation than the human samples, both in terms of the proportions of methylated clones and the methylation densities of individual clones (Figures 4.8-4.11).

Two human CGIs had low density methylation that is not observed in any mouse samples (Figures 4.6 and 4.13). Interestingly, the two CGIs with similar methylation states are associated with genes with apparently divergent XCI states. One of these CGIs is associated with a well-established escapee, *EIF2S3* (Ehrmann *et al.*, 1998), and the other, *CDKL5*, is shown to be inactivated (expression from X_i only observed in two out of nine cell hybrids) according to Carrel and Willard (2005). It is possible that such low level methylation is not enough to have an impact on the XCI state of a gene.

From the 20 CGIs for which bisulphite data were produced, a prediction of the XCI state of the associated gene was previously made based on the RPMA results

for five human and ten mouse genes. All these predictions were strongly supported by the bisulphite sequencing results. For five human CGIs, no predictions could be made based on RPMA results because of ambiguous methylation status. Based on bisulphite sequencing results, three of these CGIs were clearly hypomethylated, and two CGIs both had low proportions of methylated clones, which all had only low level methylation, leading to a prediction of escape (4/5 agree with Carrel and Willard (2005)).

In this study, it has been confirmed that CGI methylation on the human X chromosome does exhibit a different density variation from that in mouse, at least in fibroblasts. Both species display a wide range of methylation densities, but the mouse CGIs tend to be either completely unmethylated or densely methylated, whereas most human CGIs are only moderately methylated. A revised model of CGI methylation on the X chromosome in these two species is illustrated in Figure 4.16.

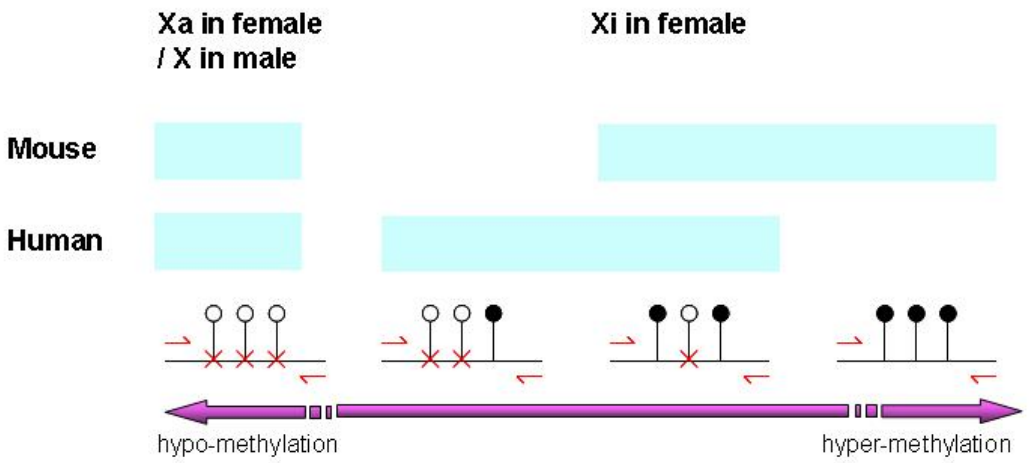

Figure 4.16: Variation of CGI methylation on the human and mouse X chromosomes - revised model (see Figure 4.2 for the original model).

Chapter 5

Investigation of CpG island methylation on the opossum X chromosome

5.1 Introduction

Marsupial and eutherian mammals shared a common ancestor approximately 180 MYA (Murphy *et al.*, 2004). The separation of the marsupial lineage is half way between the split between birds and mammals $\sim$ 310 MYA (Bininda-Emonds *et al.*, 2007; Hedges and Kumar, 2004) and the radiation of the eutherian clades $\sim$ 100 MYA (Bininda-Emonds *et al.*, 2007; Woodburne *et al.*, 2003), making them an important group for studying the biology and evolution of mammals. Although the marsupials share some basic mammalian characteristics with the eutherians, for example giving birth to live young, presence of mammary glands and fur, they also display a great number of differences, the most distinct being

the early developmental stage at birth and the long and complicated lactation period. Such divergence is also apparent in the genetics of the sex chromosomes.

Marsupial mammals share the same XY sex determination system with the eutherians, and also achieve dosage compensation through XCI, but with distinct differences. In both groups, the inactive X chromosome is late replicating and turned into heterochromatin. Key histone modification on the eutherian inactive X chromosome, like hypoacetylation of histone H4, was identified on the marsupial X. However, the initiation and maintenance of XCI appear to be very different in the marsupials. Most noticeably, the master switch gene for random XCI in eutherians, *XIST*, is not present in marsupials, and XCI in marsupials is imprinted, where the paternal X is always silenced. It has also been suggested that methylation of the CGIs on the inactive X chromosome, contributing towards maintenance of XCI in the eutherians mammals, may be absent from the marsupial mammals (Kaslow and Migeon, 1987).

To investigate whether the apparently poor maintenance of XCI states in marsupials (Cooper *et al.*, 1977; Samollow *et al.*, 1987) is associated with lack of CGI methylation, Kaslow and Migeon (1987) studied the 5' CGI of *G6PD* in a North American marsupial, the Virginian opossum (*Didelphis virginiana*). *G6PD* is stably inactivated in human and mouse, and its 5' CGI was found to be methylated on the inactive X in both species. Its XCI appears to be less stable in several marsupial species, where expression of the paternal (inactive) allele of *G6PD* was detected in tissue samples or cultured fibroblast cells (Samollow *et al.*, 1987). Kaslow and Migeon (1987) found this CGI to be hypomethylated in both male and female samples, possibly leading to its incomplete XCI. However, despite lack of methylation, expression of the paternal allele was at best partial, so they

proposed that methylation may not play a role in marsupial XCI, but that it serves to stabilise XCI in eutherians. Later, bisulphite sequencing analysis of the 5' CGI of *G6PD* in an Australian marsupial, the common wallaroo (*Macropus robustus*), confirmed hypomethylation (Loebel and Johnston, 1996). The paternal allele of *G6PD* was previously found to be completely repressed in this species, so the authors speculated that CGI methylation might not be required for maintenance of XCI in marsupials.

At the time of commencement of this study, the notion that 5' CGIs are not methylated on the marsupial X chromosome is based on methylation analysis of only a single gene, *G6PD*. During the course of the study, a new study was described where Hornecker and colleagues (2007) examined CGI methylation of *G6PD* and PGK1 in adult females of the grey short-tailed opossum (*Monodelphis domestica*). Using bisulphite sequencing, they concluded that both CGIs were hypomethylated on the paternal X (also the inactive X), so the previous perception about CGI methylation in marsupial XCI remained unchallenged.

However, there is now evidence for differential methylation at the 5' CGIs of imprinted autosomal genes in marsupials (Smits *et al.*, 2008; Suzuki *et al.*, 2007). Suzuki and colleagues (2007) reported that *PEG10* is imprinted in tammar wallaby, with almost exclusive expression from the paternal allele. Its 5' promoter region, covered by a CGI, was heavily methylated on the maternal allele but unmethylated on the paternal allele. The same CGI also covers the 5' promoter region of an adjacent gene, *SGCE*, but the *SGCE* part of the CGI was completely unmethylated, correlating with the biallelic expression pattern of *SGCE*. A second differentially methylated region associated with an imprinted gene in marsupials was identified at the *H19* promoter and exon 1 in tammar wallaby in a

recent study of the *H19-IGF2* imprinting locus in mammals (Smits *et al.*, 2008). This association between differential methylation and imprinting is similar to the situation in eutherians. On the basis of these observations, and considering the dearth of information on X-linked genes, it remains a possibility that differential methylation of X chromosome genes may be involved in marsupial XCI.

The availability of the first whole genomic sequence of a marsupial (Mikkelsen *et al.*, 2007) provides a great opportunity to examine the methylation states of a greater number of X-linked genes in marsupials and to ascertain the importance of methylation in marsupial XCI.

5.1.1 The grey short-tailed opossum (*M. domestica*)

There are more than 300 species of marsupial mammals currently identified, mainly in Australia, where they are the dominant mammalian group (more than 200 species, including kangaroos, possums and koalas), and in South America (more than 90 species of opossums and shrew opossums) (Wilson and Reeder, 2005). Only one species is found in North America, the Virginian opossum (*D. virginiana*).

The chosen species for the first marsupial genome sequencing project was the grey short-tailed opossum, *M. domestica* (Figure 5.1), a South American marsupial that is used as a model organism in a wide range of biological studies (reviewed in Samollow, 2006). It is small, fast-growing, rapid-breeding, and can easily adapt to standard rodent facilities, so it is probably not surprising that *M. domestica* has been raised in pedigreed lab colonies for over 25 years (VandeBerg, 1989). Some earlier work on marsupial XCI was carried out using the Virginian

opossum, the original 'opossum', but as *M. domestica* gains popularity as an experimental model, *D. virginiana* is used much less. Both *D. virginiana* and *M. domestica* belong to the Didelphidae family. For the purpose of this thesis, the name 'opossum' refers to *M. domestica*.

Figure 5.1: South American, grey short-tailed opossum (*Monodelphis domestica*). Courtesy of Paul Samollow, Southwest Foundation for Biomedical Research, San Antonio. Freely distributed image obtained from `http://www.genome.gov/`

5.1.2 The opossum X chromosome

A high quality draft of the genomic sequence of opossum was described in May 2007 (Mikkelsen *et al.*, 2007). The genome of a partially inbred female opossum was sequenced using the whole genome shotgun approach to 6.8x depth. The

current assembly, monDom5, covers 99% of the estimated euchromatic sequence, and has excellent accuracy and contiguity.

The opossum X chromosome is very small compared to the autosomes. It is the smallest chromosome in opossum, representing only 2% of the total opossum genome, and is approximately 76 Mb in length (Mikkelsen *et al.*, 2007). In comparison, the sizes of X chromosomes in eutherian mammals tend to be comparable with that of the autosomes. The human, mouse, and dog X chromosomes each represent approximately 5% of the genome and are about twice the size of the opossum X (Lindblad-Toh *et al.*, 2005; Ross *et al.*, 2005; Waterston *et al.*, 2002). The size difference reflects the addition of the XAR to the X chromosome after the split between marsupial and eutherian mammals. Comparative mapping confirmed the conservation of synteny between the opossum X chromosome and the XCR of the human X chromosome (Mikkelsen *et al.*, 2007).

The opossum X chromosome is also unique in its sequence composition in relation to its autosomes. In eutherian mammals, the X chromosome has lower (G+C) content and CpG density than the autosomes. The situation in the opossum genome is just the opposite. The (G+C) content of the opossum X is not only higher than that of the autosomes, which have the lowest mean autosomal (G+C) content among all mammals sequenced, but also higher than any other X chromosome sequenced (Mikkelsen *et al.*, 2007). Similarly, despite the extremely low CpG frequency in opossum autosomes (two-fold lower than in other mammals), the CpG density of the opossum X chromosome is comparable with that of the eutherian X chromosomes (Mikkelsen *et al.*, 2007). One hypothesis proposes that the decline of (G+C) content, especially CpG dinucleotides, is balanced by a recombination-mediated (G+C)-biased gene conversion process (Duret 2006).

Consistent with this model, the opossum autosomes have lower recombination rates than those of other sequenced mammals, while the recombination rate of the opossum X was estimated to be higher than that of the autosomes, possibly as a result of their dramatic size difference (Mikkelsen *et al.*, 2007). If at least one recombination event is required for each chromosome pair at meiosis to ensure proper disjunction, then the small size of the X would result in very high recombination rate relative to the autosomes. Interestingly, this is also the opposite of the situation in eutherian mammals, where the X chromosome has a lower recombination rate than the autosomes.

5.1.3 Aims of this chapter

The aim of the work described in this chapter is to explore a large number of CGIs on the opossum X chromosome to investigate the hypothesis that lack of differential methylation between the female X chromosomes is a common feature of marsupial X-linked genes.

5.2 Identification of CGIs on the opossum X chromosome

Information about the opossum genes used in this study was obtained from the Ensembl database (v47) based on the sequence assembly monDom5. Opossum CGIs were predicted (by Val Curwen) using the cpg program with the same parameters that were applied in identifying human and mouse CGIs (section 3.2.2).

There were 448 protein coding genes found on the opossum X chromosome, 144 (32.1%) of which had a predicted CGI within 2 kb of the 5' end. CGI sizes range between 400 bp and 5 kb, with an average size of about 1 kb. Of the 130 CGIs that are each associated with a single gene, 94 overlap with exon1, the vast majority extending into intron1, and the remaining 36 are upstream from the 5' end of the genes. Seven pairs of genes, transcribed in opposite directions, had a CGI covering the 5' regions of both genes. Four of these CGIs overlap with exon1 of one gene and are a short distance upstream from the 5' end of the other gene, two CGIs are upstream from both genes, and one CGI overlaps with exon1 of both genes.

Out of these 144 genes with a predicted 5' CGI on the opossum X chromosome, 110 have putative human orthologues according to Ensembl annotation; 107 of these are also X-linked and are of particular interest for this study. For convenience, in this thesis all opossum genes are referred to using the gene ID of their putative human orthologues (predicted in Ensembl).

5.3 Assessing CGI methylation on the opossum X chromosomes by RPMA

In my previous studies on CGI methylation in human and mouse, RPMA was shown to be a reliable method to assess rapidly the methylation states of a large number of CGIs. Therefore in the initial stage of the project, RPMA was used to examine methylation of a number of CGIs on the opossum X chromosome.

Predicted CGIs on the opossum X chromosome, with 500 bp flanking se-

quences, were extracted from Ensembl, and primers were designed to amplify 150-250 bp products containing multiple *Hpa*II/*Msp*I cleavage sites lying within the predicted CGI (section 3.3.2). Since bisulphite sequencing in human revealed that RPMA may produce a signal in case of low level methylation, three pairs of primers were designed for each CGI wherever possible to increase the predictive power of the method. Primers were tested in PCR amplification using male genomic DNA. For the 41 CGIs targeted (see below for details on selection of candidate CGIs), the expected products were successfully amplified for 37 CGIs: nine of these had three pairs of successful primers, 18 had two, and 10 had only one. All primer combinations, their predicted amplicon sizes, and the number of enclosed CCGG sites are listed in Appendix I.

Genomic DNA was extracted from the liver of one female and one male opossum (section 2.9.1). Restriction enzyme digestion was performed as described in Chapter 3, but a simpler PCR protocol (optimised by Frances Lovell, personal communication, section 2.11.2) was applied.

5.3.1 Analysis of CGI methylation on the opossum X chromosome

Initially, I selected nine candidate CGIs to assay: *G6PD*, *HPRT1*, *PGK1*, *SMCX*, *RBM10*, *UBE1*, *USP11*, *SOX3*, and *DIAPH2*. Three genes, *G6PD*, *HPRT1*, and *PGK1*, are housekeeping genes that have been extensively studied in human, mouse, and marsupials. All three are from evolutionary stratum S1, so are among the first genes recruited into XCI and have had a long time to achieve methylation. Allele-specific expression of all three genes and the CGI methylation state of

G6PD have been studied in a number of marsupial species, making them obvious candidates for further methylation studies. Two additional genes, *SOX3* and *DIAPH2*, were chosen from S1. *SOX3* is particularly interesting as it has been suggested to be the X-borne homologue of the mammalian male-sex-determining gene SRY (Foster and Graves, 1994). Therefore in theory it should be the first excluded from recombination and has had the longest time of all X-linked genes to acquire methylation. I assayed CGI methylation for both genes by RPMA in a pilot study preceding work described in Chapter 3. *SOX3* had female-specific CGI methylation in both species, consistent with the gene being X inactivated. *DIAPH2*, a partial escapee in human (Carrel and Willard, 2005), had female-specific CGI methylation in mouse but not in human. To provide a contrast to the above five genes, another four genes were chosen from the younger evolutionary stratum S2; genes on which have been recruited into XCI for a shorter time and a greater proportion escape from XCI. *SMCX* escapes XCI in both human and mouse (Agulnik *et al.*, 1994), and the CGI was found to be hypomethylated in 6 of 11 eutherian orders (12 of 18 species) tested by Jegalian and Page (1999). The *RBM10* (inactivated) - *UBE1* (escapee) - *USP11* (partial escapee) gene cluster in human contains a potential boundary separating inactivated and escapee domains (XCI status from Carrel and Willard (2005)). The orthologous genes on the opossum X form a cluster in the same order, so it is interesting to investigate whether such domains and boundary are also present in opossum. In addition, the marsupial *UBE1* has a homologue on the Y chromosome (Mitchell *et al.*, 1992), so is expected to escape from XCI and free from CGI methylation, thus providing a negative control for methylation analysis.

RPMA results were obtained for seven of these genes, but no PCR primers

were successfully designed for the CGIs of *HPRT1* or *RBM10*. Consistent with expectation, and as in human and mouse, no CGIs showed a pattern indicating hypermethylation in male samples (Table 5.1). However, not all male results showed clear hypomethylation. In some cases, a very faint or faint signal was present in the *Hpa*II lane, which may be caused by low level methylation at the enzyme cleavage sites or incomplete digestion. These faint signals in the *Hpa*II lane were often accompanied by a similar signal in the *Msp*I lane, where cleavage is not affected by methylation, so it is more likely that the *Hpa*II signal has resulted from partial digestion rather than methylation. The opossum DNA was extracted from tissue samples, so may still contain substances that inhibit digestion, for example glycogen. This problem was not seen in human an mouse, possibly due to the different source of the test DNA (from cultured fibroblast cells). Attempts were made to remove this problem, including use of fresh batch of restriction enzymes, digestion with higher concentration of enzymes, re-purification of the DNA samples, and use of spermidine to facilitate digestion, but without success.

In spite of the problems with digestion, giving rise to faint signals in male samples, speculations about methylation states are still possible through comparison between male and female results. In all but one CGIs, the female pattern was identical (4 CGIs) or very similar (2 CGIs) to the male pattern (Table 5.1). The only exception is *PGK1*, which showed a greater level of methylation in the female sample in both amplicons (Table 5.1, Figure 5.2). RPMA results for these seven CGIs are consistent with female-specific methylation being rare on the opossum X chromosome.

In order to explore this apparent lack of sex-specific differential methylation of CGIs on the opossum X chromosome further, I extended the study to another 15

CGIs, selected from the centromeric end of the chromosome. As expected, most CGIs were not hypermethylated in male samples (Table 5.1). As before, faint or very faint signals were frequently found in the *Hpa*II and *Msp*I lanes, indicative of partial digestion. Therefore, no attempt was made to define the methylation state of a CGI, as described in Chapter 3, but the island was simply scored for presence or absence of female-specific methylation.

Four amplicons, three of which contain only one cleavage site, had strong PCR products for all enzymes in both sexes, making them not informative for estimating presence of female-specific methylation. Eleven CGIs exhibited identical (5) or very similar (6) RPMA patterns in both sexes (Table 5.1). One of them, the CGI of *ATP6AP2* had a strong signal in the *Hpa*II lane for both sexes, the *Msp*I lane was clean in the female sample and had only a faint signal in the male sample, suggesting hypermethylation at both female and male islands. Heavier methylation in female was detected for the CGIs associated with *MPP1*, *FAAH2*, and *F8A*∼ (Table 5.1, Figure 5.2). For all three CGIs, the male samples had a very faint signal in the *Hpa*II and *Msp*I lanes, while the female samples had a strong signal in the *Hpa*II lane, but only a very faint signal in the *Msp*I lane. Repeated experiments showed identical or very similar patterns. For the remaining CGI, the only amplicon available was not informative so no conclusion could be made. In general, these results are still consistent with lack of female-specific methylation, but uncovered a small cluster of CGIs with heavier methylation in females.

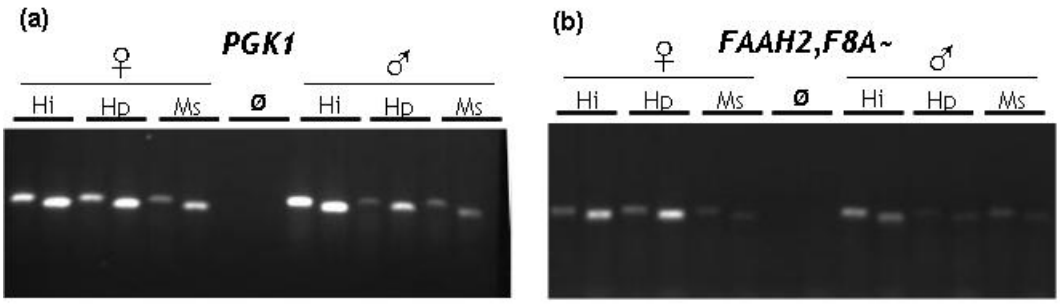

Figure 5.2: RPMA results of three opossum genes with female-specific methylation of CGI. *FAAH2* and *F8A*∼ only had one amplicon each.

Interestingly the *MPP1*, *FAAH2* and *F8A*∼ genes reside in the same neighbourhood of a 1.5 Mb region on the opossum X chromosome, only separated by a few non-CGI-associated genes. Having previously observed higher female methylation associated with the CGI of *PGK1*, I hypothesised that I might have success in detecting additional examples of methylation among its neighbouring genes. I assayed 15 CGIs surrounding *PGK1*. Disregarding the three uninformative amplicons, which had strong signals in the *Msp*I lane, 12 CGIs exhibited identical (6) or very similar (6) methylation patterns in both sexes (Table 5.1). Interestingly, the three CGIs immediately surrounding *PGK1*, *FGF16*, *IAG2*, *ITM2A*, all showed some degree of female-specific methylation (Table 5.1). The CGIs of *FGF16* and *ITM2A* each had two amplicons, one showed no difference between sexes, but the other with strong signal in the *Hpa*II lane in female and no signal in male. The only amplicon of the *IAG2* CGI had a strong signal in the *Hpa*II lane in female and a faint signal in male.

In summary, 37 CGIs were successfully assayed. The majority (73%) showed no methylation differences between the female and male samples (Table 5.1, Figure 5.3). Only seven CGIs, in two distinct clusters, demonstrated greater

level of methylation in the female samples (Table 5.1, Figure 5.3). However, a weak signal was frequently detected in the *Msp*I lanes in both female and male RPMA results, suggesting incomplete digestion and making it difficult to estimate the exact extent of methylation.

Table 5.1: RPMA results of opossum CGIs. '+' = strong band; 'f' = faint band; 'vf' = very faint band; '-' = no visible band. One CGI is shared by each pair of the following: *BCAP31* and *ABCD1*, *IDH3G* and *SSR4*

	RPMA primer pair 1			RPMA primer pair 2			RPMA primer pair 3			Female-
Gene	CCGG	Female	Male	CCGG	Female	Male	CCGG	Female	Male	specific
	sites	Hi Hp Ms	Hi Hp Ms	sites	Hi Hp Ms	Hi Hp Ms	sites	Hi Hp Ms	Hi Hp Ms	methylation?
USP11	3	+ - -	+ - vf	3	+ - -	+ - -				N
UBE1	3	+ vf vf	+ vf vf	3	+ vf -	+ vf vf				N
SMCX	3	+ vf -	+ vf -	2	+ - -	+ - -				N
G6PD	3	+ vf -	+ vf -	2	+ f f	+ f f				N
SOX3	2	+ vf vf	+ vf vf	2	+ vf vf	+ vf vf				N
DIAPH2	5	+ - -	+ - -	4	+ - -	+ - -				N
PGK1	2	+ + f	+ vf vf	1	+ + f	+ f vf				Y
GDI1	4	+ - -	+ - -	2	+ f -	+ f -				N
ATP6AP1	1	+ + -	+ + f	1	+ + +	+ + +				N
PLXNA3	3	+ - -	+ - -	2	+ f vf	+ vf vf				N
SEPHS2	5	+ - -	+ - -							N
DKC1	2	+ f f	+ f f	2	+ + +	+ + +				N
MPP1	2	+ + vf	+ vf vf							Y
FAAH2	2	+ + vf	+ vf vf							Y
F8A~	1	+ + vf	+ vf vf							Y
TSC22D3	2	+ f -	+ vf vf	5	+ f -	+ f -				N
TMEM164	4	+ - -	+ - -	1	+ + +	+ + +	1	+ + f	+ f f	N
AMMECR1	3	+ vf vf	+ vf vf							N
AR	3	+ - -	+ - -	3	+ vf -	+ vf -	1	+ - -	+ - -	N
MSN	1	+ + +	+ + +							U
BCAP31	2	+ + f	+ + f	3	+ f -	+ f -	2	+ - -	+ - -	N
IDH3G	2	+ vf -	+ - vf	3	+ vf -	+ vf -				N
HTATSF1	1	+ f f	+ f vf	3	+ vf vf	+ vf -	3	+ f -	+ vf -	N
RBMX	2	+ vf -	+ vf -	2	+ - -	+ - -				N
GPR101	2	+ - -	+ - -							N
ZIC3	1	+ + +	+ + +	1	+ + f	+ f f	2	+ vf vf	+ vf vf	N
CHM	2	+ - -	+ - -	2	+ f f	+ f vf	3	+ f -	+ f -	N
DACH2	3	+ - -	+ - -	2	+ f -	+ f -	3	+ f -	+ vf -	N
FGF16	3	+ + vf	+ - -	7	+ - -	+ - -				Y
IAG2	1	+ + f	+ f f							Y
ITM2A	1	+ f f	+ f f	1	+ + -	+ - -				Y
BRWD3	2	+ f f	+ f f	2	+ f f	+ f f	2	+ f f	+ f f	N
ALG13	1	+ vf vf	+ - -	2	+ f vf	+ f -				N
OGT	3	+ - -	+ - -	1	+ f vf	+ f vf				N
TAF1	4	+ - -	+ - -	2	+ - -	+ - -	1	+ + +	+ + +	N
ZMYM3	2	+ - vf	+ - vf							N
SNX12	1	+ f vf	+ vf vf							N

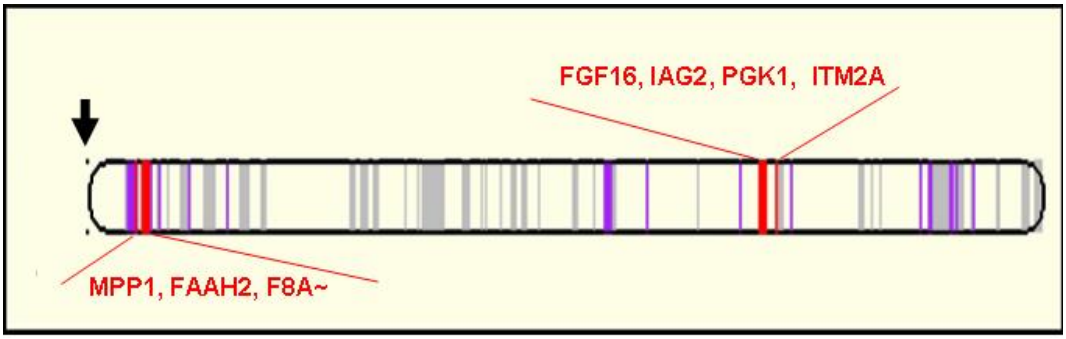

Figure 5.3: Distribution of assayed CGIs on the opossum X chromosome. Locations of all assayable CGIs are indicated by grey bars. The assayed CGIs are coloured: those with female-specific methylation are in red, and the ones without female-specific methylation are in purple. Location of centromere is indicated by arrow.

5.4 Assessing CGI methylation on the opossum X chromosomes by bisulphite sequencing

The data described indicate that most CGIs on the female opossum X chromosomes are hypomethylated. This is consistent with the previous picture for other marsupial species but extends the basis for this conclusion well beyond the study of a single gene. Most significantly, the RPMA analysis has uncovered a number of apparent examples of higher female methylation. If confirmed, this would be the first demonstration of this phenomenon in marsupial and would open up the possibility of a role for methylation in marsupial XCI that was not previously appreciated. Owing to the possibility of incomplete digestion, as well as the limited resolution of RPMA, it was necessary to confirm these findings using a more accurate method. To this end, six CGIs that showed differential methylation by RPMA, together with two CGIs that are associated with well-studied genes in

marsupials (*HPRT1* and *G6PD*), were examined by bisulphite sequencing. It was not possible to design primers for the *MPP1* CGI, which also showed differential methylation by RPMA. Bisulphite sequencing was performed as described in Chapter 4. All primer combinations and their predicted amplicon sizes are listed in Appendix II.

The CGIs of *G6PD*, *HPRT1*, *FAAH2*, *FGF16*, *IAG2*, *PGK1*, and *ITM2A* were successfully bisulphite sequenced in both female and male samples (Figures 5.4-5.10). *F8A*∼ results were only available for the female sample (Figure 5.11). One or two amplicons were sequenced for each CGI, and at least 40 individual molecules, each representing the island of an individual chromosome, were sequenced for each amplicon. Some amplicons contain CpG dinucleotides outside the actual island, these CpGs were not counted in the methylation statistics presented below.

Most CGI molecules were clearly hypomethylated in the male samples. For five CGIs, the majority of island molecules were completely free from methylation, and the remaining ones had only one or two CpGs methylated (Figures 5.4-5.8). The CGI of *ITM2A* had a small proportion of island molecules with three out of 22 CpGs methylated (Figure 5.10). For the CGI of *PGK1*, a small proportion of island molecules had nearly identical low level methylation, but most other island molecules had no methylated CpG dinucleotides (Figure 5.9).

The CGIs of *G6PD* and *HPRT1* were hypomethylated in both sexes (Figures 5.4, 5.5). Like their male companions, the female CGIs had the majority of island molecules completely unmethylated, apart from a small number with one or two methylated CpGs. Only one molecule of the *HPRT1* CGI (n=77) and six molecules of the *G6PD* CGI (n=155) had more than two CpGs methylated in

females.

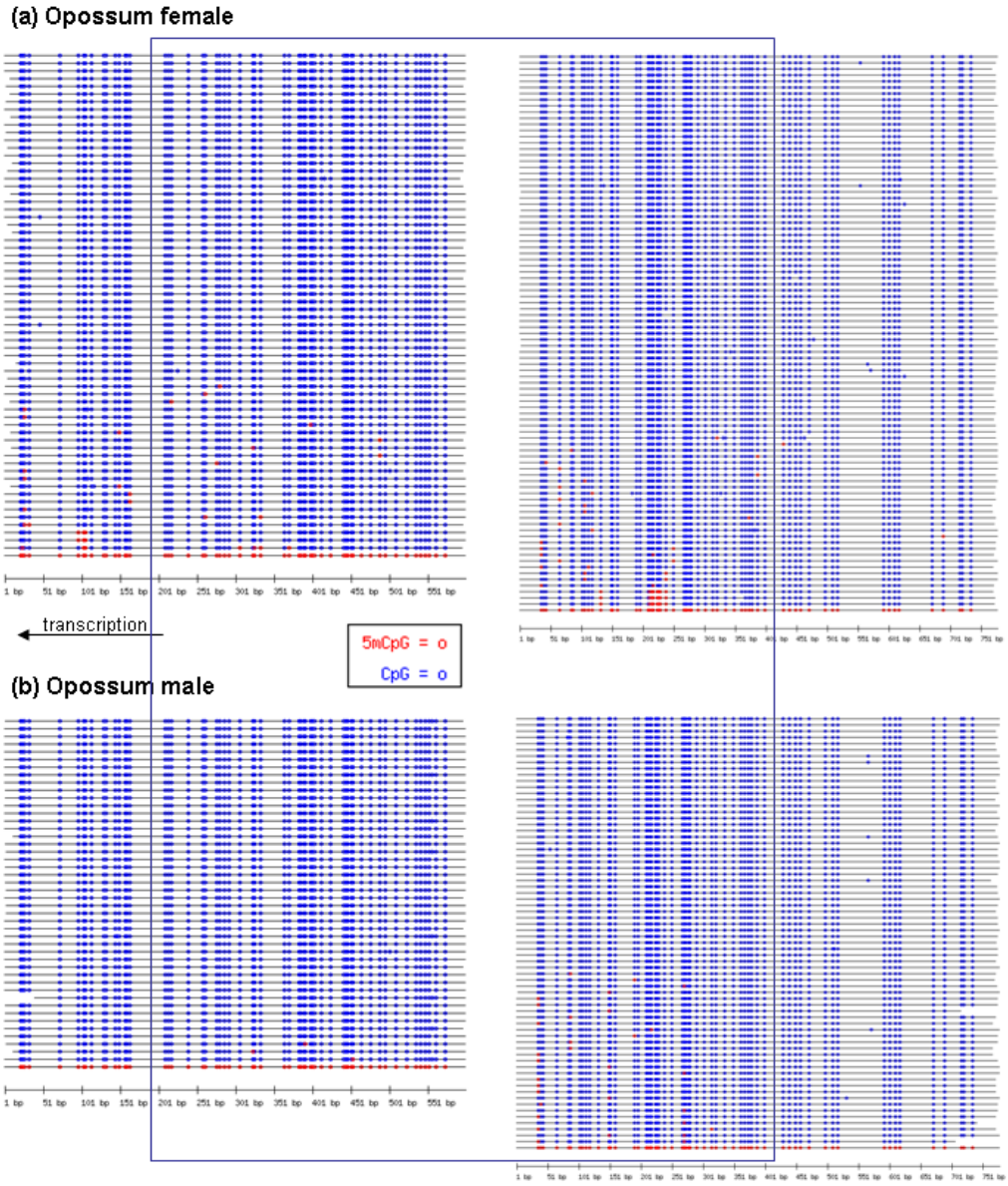

Figure 5.4: CGI methylation profiles of opossum *G6PD*. This CGI was shown to lack female-specific methylation in RPMA. The bisulphite sequenced fragments are overlapping and cover the first two thirds of the CGI. The overlapped region is indicated by the blue box. The CGI is upstream of the gene and direction of transcription is indicated by arrow. For detailed annotation see Figure 4.3.

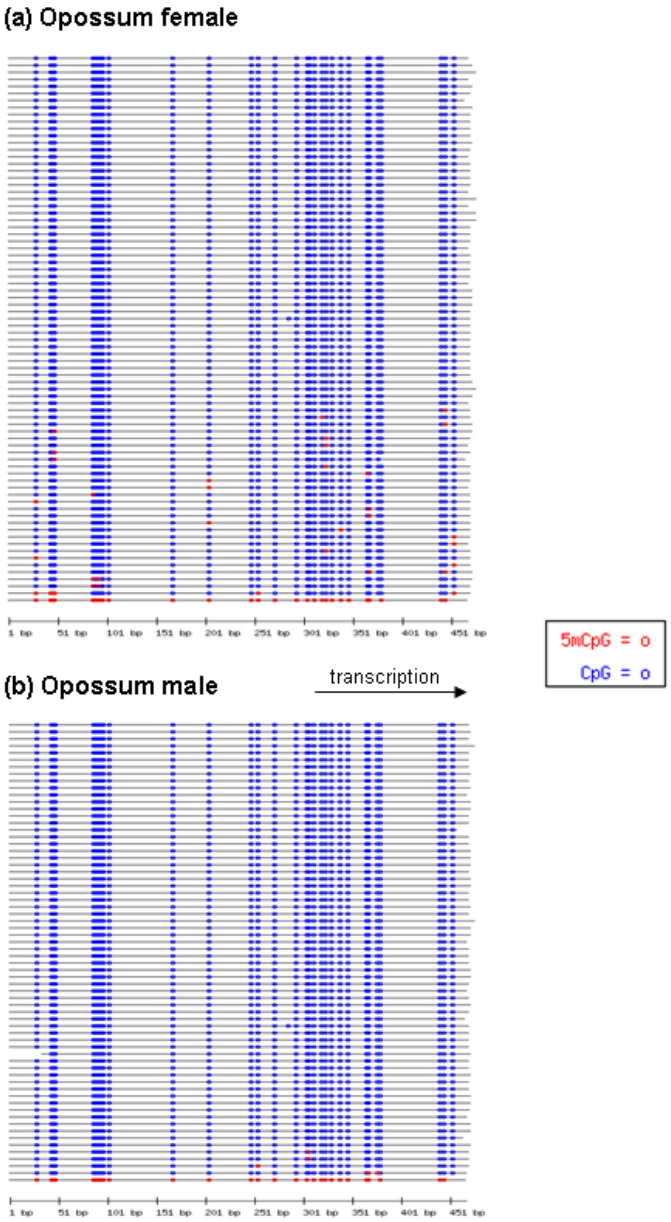

Figure 5.5: CGI methylation profiles of opossum *HPRT1*. It was not possible to assay this CGI by RPMA. The bisulphite sequenced fragment covers the first two thirds of the CGI and is upstream of the transcription start site. Direction of transcription is indicated by arrow. For detailed annotation see Figure 4.3.

As predicted by RPMA, the CGIs of *FAAH2*, *FGF16*, *IAG2*, *PGK1*, and

ITM2A all showed clearly heavier methylation in the female samples (Figures 5.6-5.10). However, unlike the mouse and human female CGIs which contained distinct hypo- and hypermethylated portions, there was a continuum of methylation densities for all opossum female CGIs and the methylation densities were generally low. Therefore it was not easy to decide which molecules might have come from the active X, and which from the inactive X. Since the single X in male can be regarded as an equivalent as the active X in female, the methylation pattern of each male CGI was used as benchmark for methylation level of X_a. The female molecules with higher level of methylation than the most heavily methylated molecule from male are presumably derived from the X_i.

Using this approach, *FAAH2* and *FGF16* exhibited a CGI methylation pattern most typical of more heavily methylated female X_i and unmethylated male X / female X_a (Figures 5.6, 5.7). Methylation was scarce in the male samples, but around half of the island molecules in the female samples had greater than male level of methylation. A similar CGI methylation pattern was observed for *PGK1* and *ITM2A*, but their male samples had a very small number of island molecules with more elevated methylation, so a smaller portion of female island molecules had greater than male level of methylation (Figures 5.9, 5.10). In the case of *IAG2*, the female CGI was still more methylated than the male CGI, but had lower proportion of methylated molecules and lower levels of methylation in the methylated molecule compared to the above islands (Figure 5.8). Around half of the island molecules in *F8A*∼ CGI female samples were also methylated (Figure 5.11). Although the male results were not available for comparison, given the methylation patterns in other male CGIs it is most likely that the *F8A*∼ CGI was also hypomethylated in males.

It should be noted that, overall, the extent of methylation is still low in females, even for the most highly methylated molecules. On the other hand, a clear difference between the sexes was observed, not just for the level of methylation, but also in the pattern of methylation. There are regions that were clearly methylated in female island molecules that were classified as being from the X_i, but were devoid of methylation in the male (see Figures 5.6 and 5.9 for example).

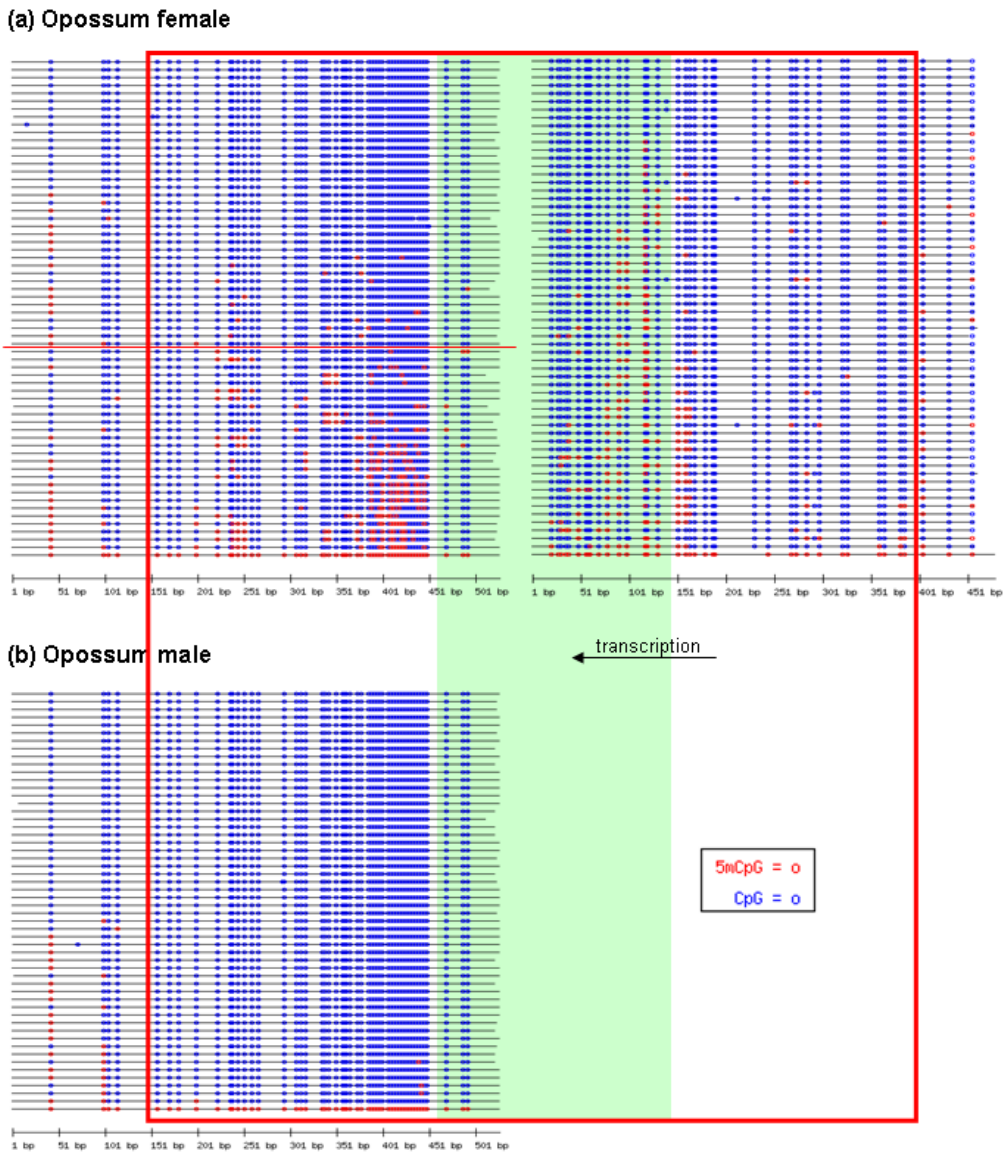

Figure 5.6: CGI methylation profiles of opossum *FAAH2*. This CGI was shown to have heavier methylation in females in RPMA . The bisulphite sequenced fragments are overlapping (but with no overlapping CpGs) and cover the entire CGI, but one amplicon only had data from female samples. The island region is indicated by the red box and the exon is shadowed in green. Female molecules potentially derived from the X_a or X_i are divided by the horizontal red line based on methylation levels in male. Direction of transcription is indicated by arrow. For detailed annotation see Figure 4.3.

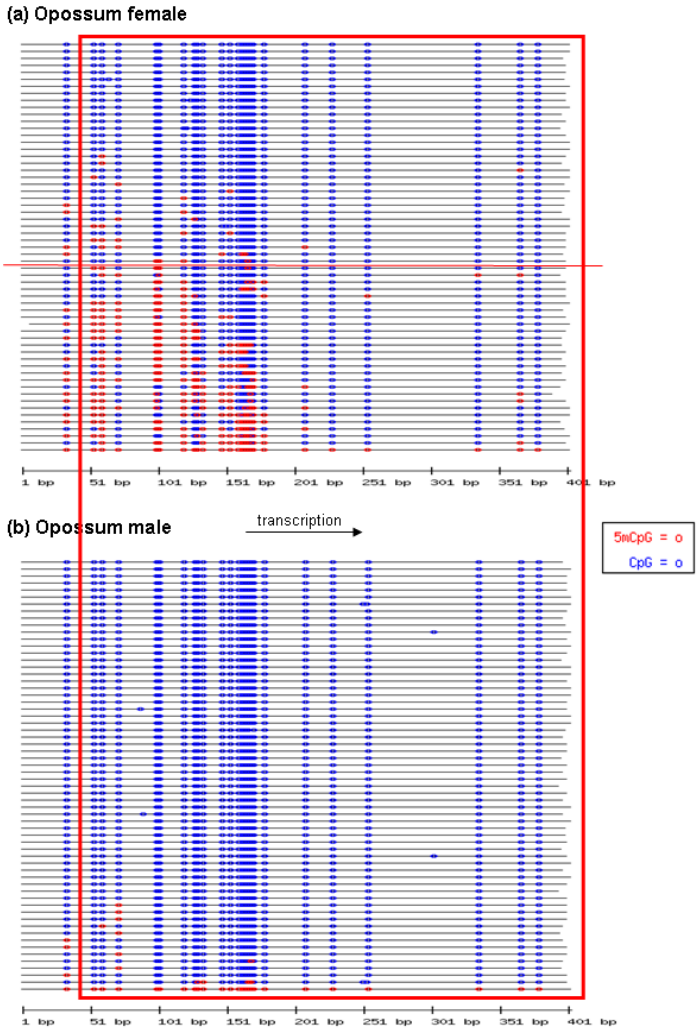

Figure 5.7: CGI methylation profiles of opossum *FGF16*. This CGI was shown to have heavier methylation in females in RPMA. The bisulphite sequenced fragment covers the first quarter of the CGI, and is upstream of the transcription start site. The island region is indicated by the red box . Female molecules potentially derived from the X_a or X_i are divided by the horizontal red line based on methylation levels in male. Direction of transcription is indicated by arrow. For detailed annotation see Figure 4.3.

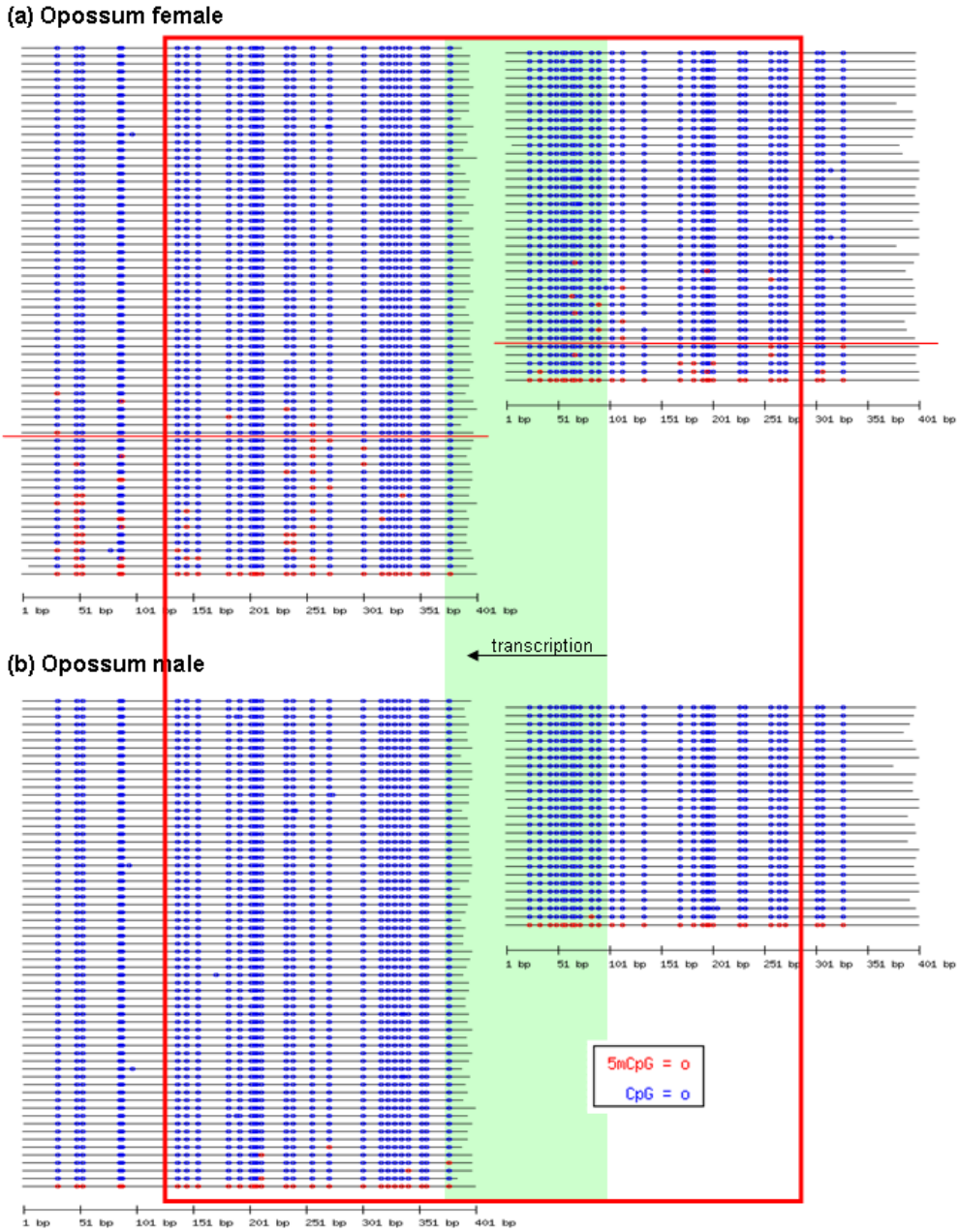

Figure 5.8: CGI methylation profiles of opossum *IAG2*. This CGI was shown to have heavier methylation in females in RPMA . The bisulphite sequenced fragments are overlapping (but with no overlapping CpGs) and cover the entire CGI. The island region is indicated by the red box and the exon is shadowed in green. Female molecules potentially derived from the X_a or X_i are divided by the horizontal red line based on methylation levels in male. Direction of transcription is indicated by arrow. For detailed annotation see Figure 4.3.

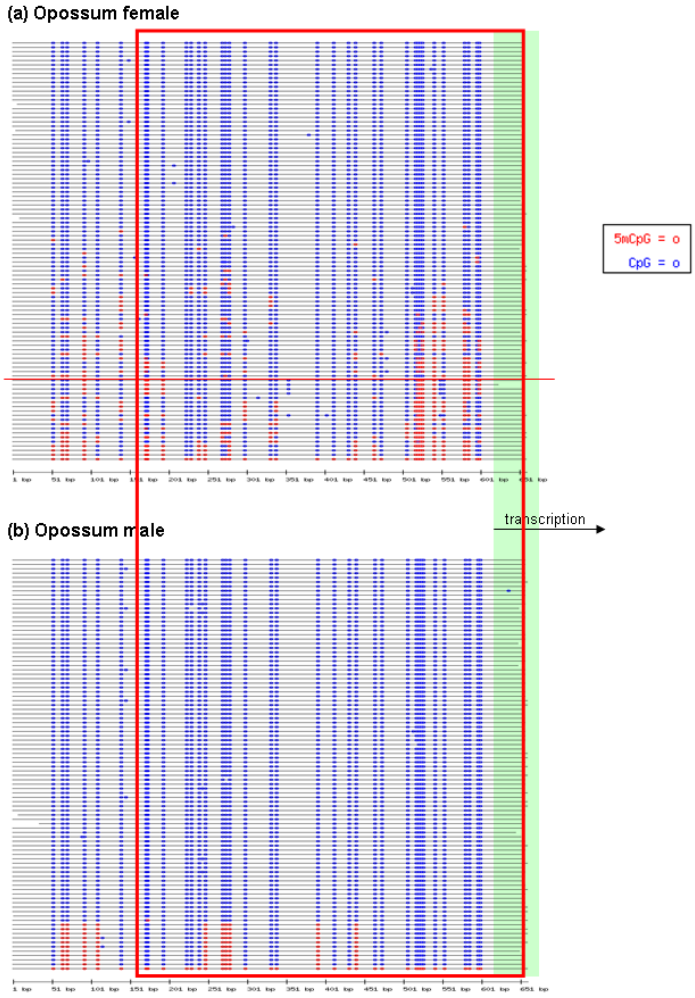

Figure 5.9: CGI methylation profiles of opossum *PGK1*. This CGI was shown to have heavier methylation in females in RPMA . The bisulphite sequenced fragment covers the entire CGI. The island region is indicated by the red box and the exon is shadowed in green. Female molecules potentially derived from the X_a or X_i are divided by the horizontal red line based on methylation levels in male. Direction of transcription is indicated by arrow. For detailed annotation see Figure 4.3.

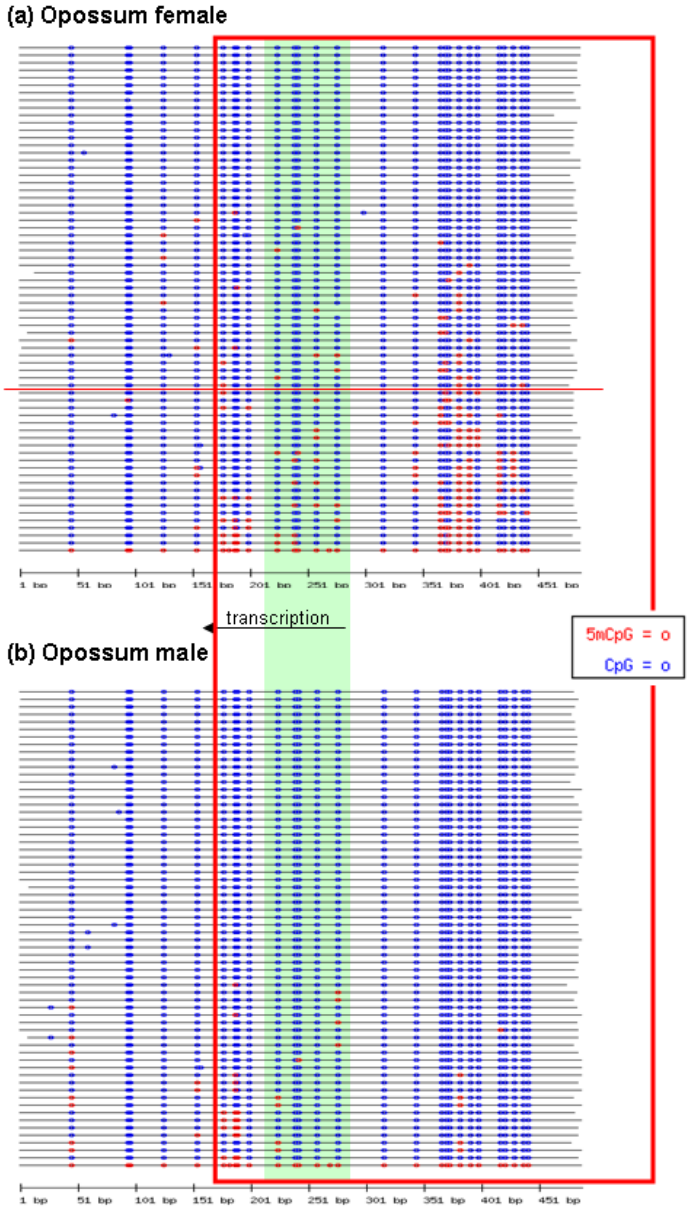

Figure 5.10: CGI methylation profiles of opossum *ITM2A*. This CGI was shown to have heavier methylation in females in RPMA . The bisulphite sequenced fragment covers the first half of the CGI. The island region is indicated by the red box and the exon is shadowed in green. Female molecules potentially derived from the X_a or X_i are divided by the horizontal red line based on methylation levels in male. Direction of transcription is indicated by arrow. For detailed annotation see Figure 4.3.

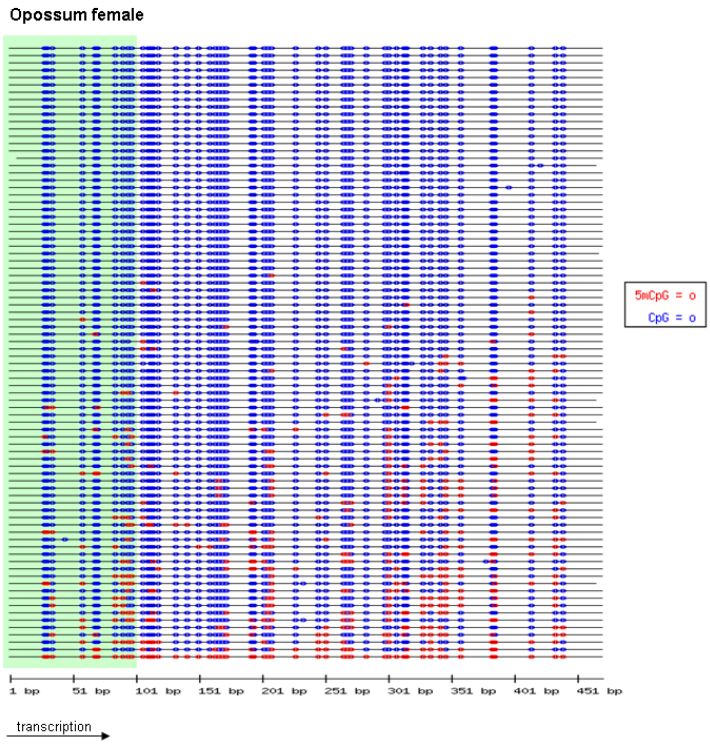

Figure 5.11: CGI methylation profiles of opossum *F8A*∼. This CGI was shown to have heavier methylation in females in RPMA. Bisulphite sequencing was only successful using the female samples. The bisulphite sequenced fragment covers a quarter of the island in the second half of the CGI. The exon is shadowed in green. Direction of transcription is indicated by arrow. For detailed annotation see Figure 4.3.

5.5 Comparison of methylation status across multiple tissues

Following confirmation of the elevation in CGI methylation levels on some female opossum X chromosomes, I decided to investigate whether or not such methylation is limited to the liver. Therefore, all eight CGIs successfully assayed using liver samples were examined by bisulphite sequencing in three additional tissue

samples from the same individuals: heart, kidney, and spleen. The CGIs of *FGF16*, *IAG2*, *PGK1*, and *ITM2A* were successfully assayed in all three tissues (except *PGK1* in male kidney); the CGI of *HPRT1* was successfully assayed in kidney and spleen; but the CGI of *FAAH2* failed in all tissues. For *FA8*∼ results were produced from kidney and spleen, and, as previously seen for liver, only female samples were successfully assayed. For convenience of comparison, methylation density was calculated for each CpG to produce an average methylation level plot for each CGI (Figure 5.12).

Methylation patterns of all test CGIs in the other tissues were similar to those found in liver. The CGIs of *G6PD* and *HPRT1* were hypomethylated in both sexes in all tissues (Figures 5.13 and 5.14). The CGIs of *FGF16*, *PGK1*, and *ITM2A* showed differential methylation consistent with the female X_i being more heavily methylated than the male X / female X_a (Figures 5.15-5.17). The CGI of *F8A*∼ shared a similar methylation pattern in female samples with these three, and is likely to have sex-specific differential methylation although no firm conclusion can be made due to the lack of male results (Figure 5.19). Heavier methylation in female was also observed in the *IAG2* CGI across all tissues, but the methylation was always at very low levels (Figure 5.18).

It is noted from Figures 5.15-5.19 that the distribution of methylation densities along a CGI is not uniform. Interestingly, in all methylated female samples, the methylation density peaks were located around the same positions in all tissues. In general, the methylation levels were slightly higher in liver and spleen than in heart and kidney, but this difference in methylation densities does not affect the distribution of methylation densities in any single tissue, only altering the height of density peaks.

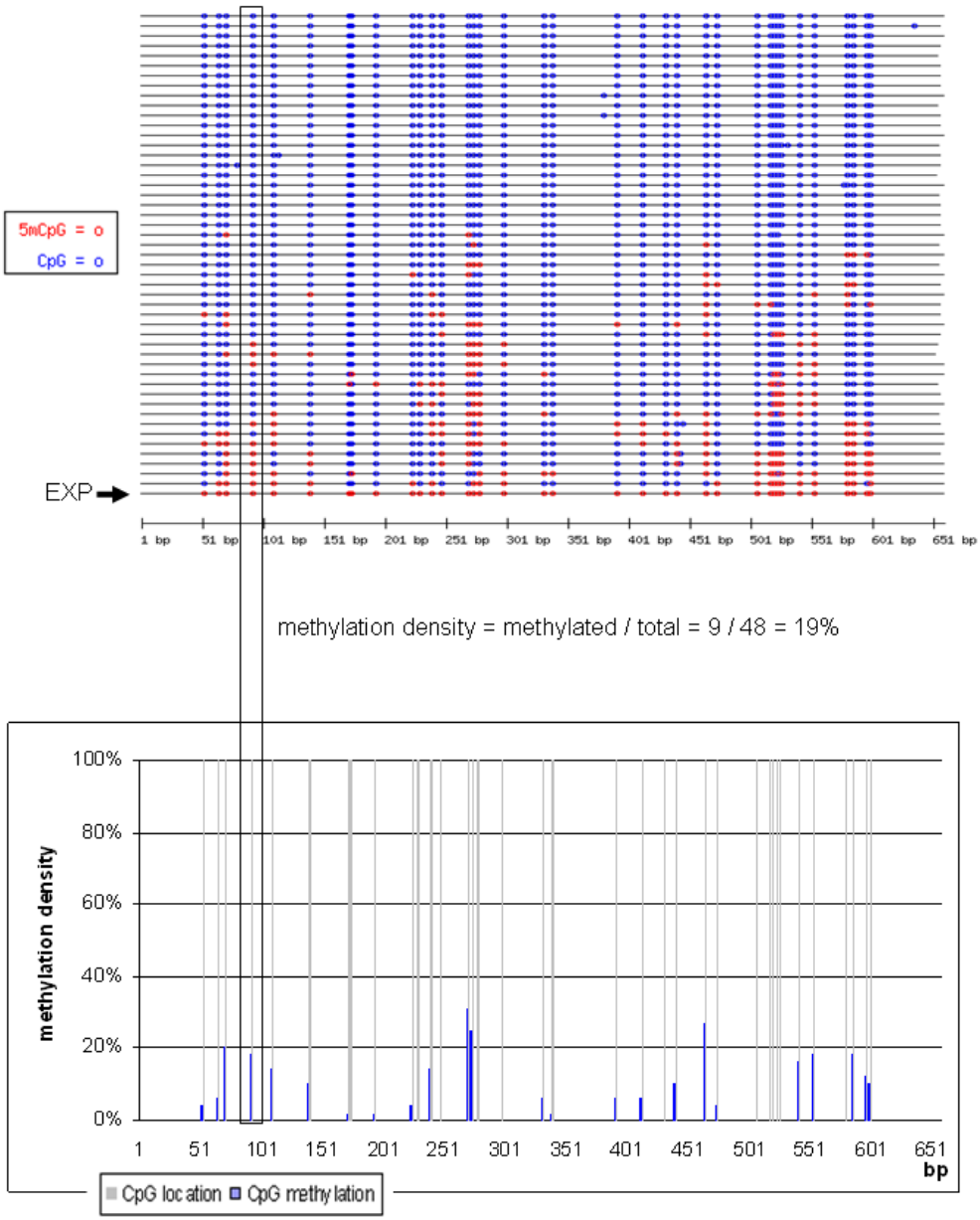

Figure 5.12: Graphical representation of average methylation levels in individual CGI. Methylation density was calculated for each CpG as the percentage of molecules methylated at this CpG out of the total number of molecules where this CpG is present. Methylation densities are shown in a bar chart where the grey bars indicate locations of expected CpGs according to the reference sequence ('EXP') and blue bars represent methylation levels. Methylation density calculation for an example CpG is demonstrated in the figure. Note that the 'EXP' sequence was excluded in the calculation.

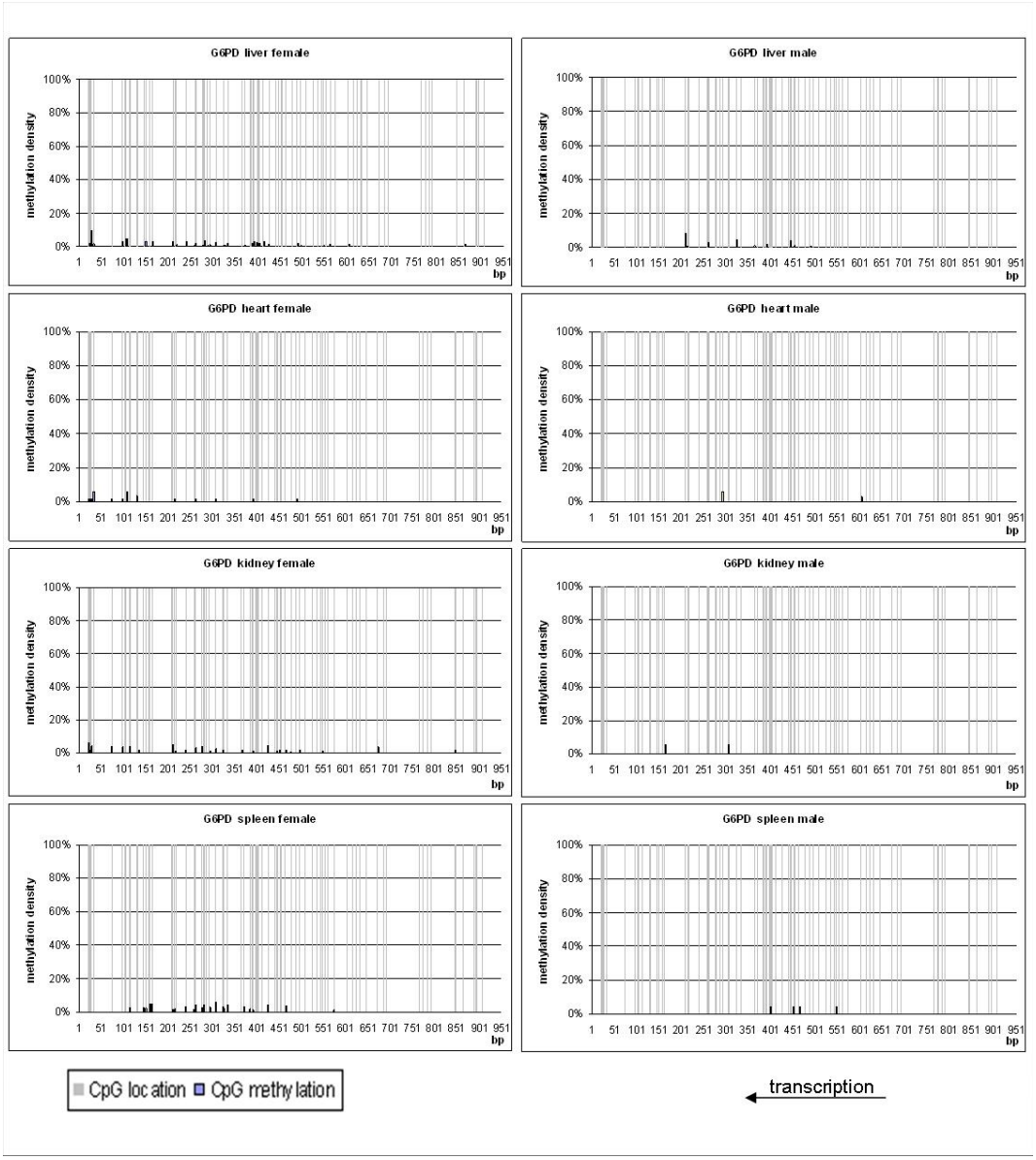

Figure 5.13: CGI methylation profiles of opossum *G6PD* in four tissues. Methylation levels are average of two overlapping sequences, the first covering 1-599 bp, and the second covering 175-951 bp. For female heart and male kidney samples, only the first sequence was available. For male heart and spleen samples, only the second sequence was available. Direction of transcription is indicated by arrow. For detailed annotation see Figure 5.12.

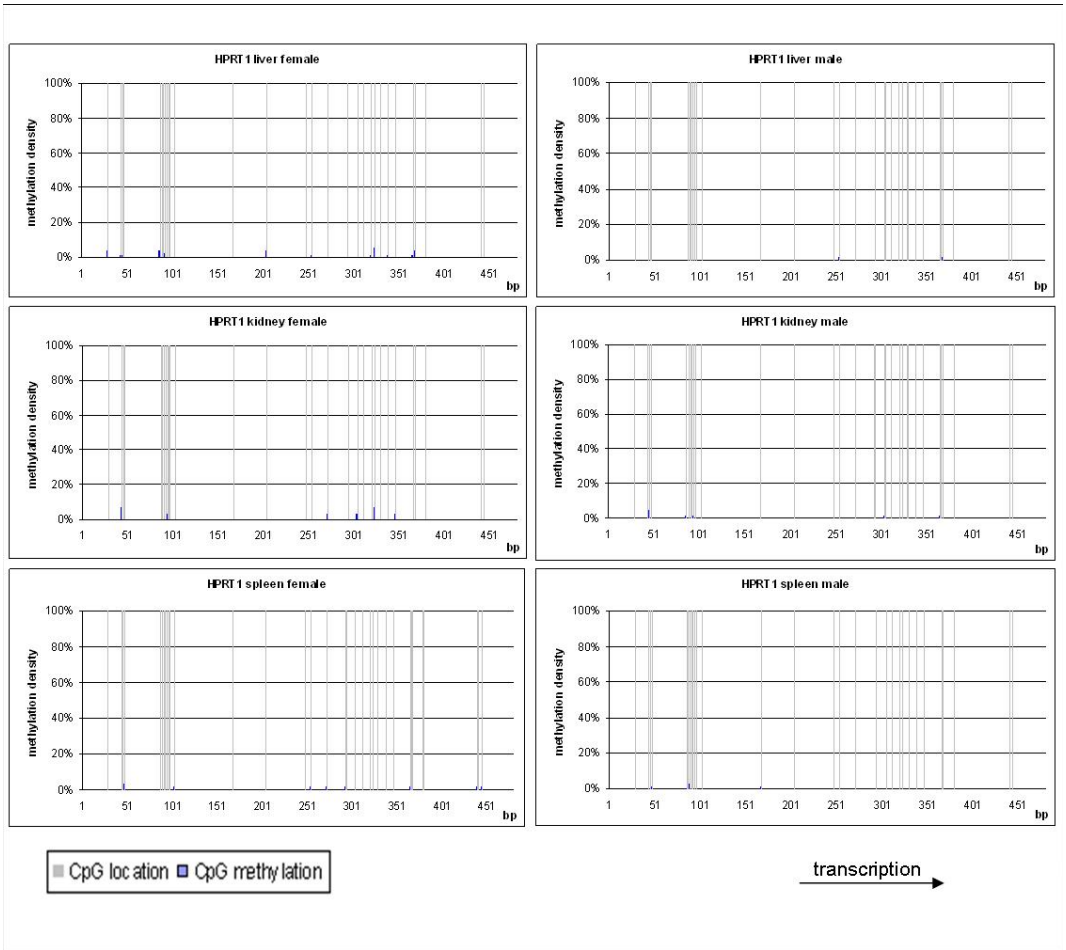

Figure 5.14: CGI methylation profiles of opossum *HPRT1* in three tissues. Direction of transcription is indicated by arrow. For detailed annotation see Figure 5.12.

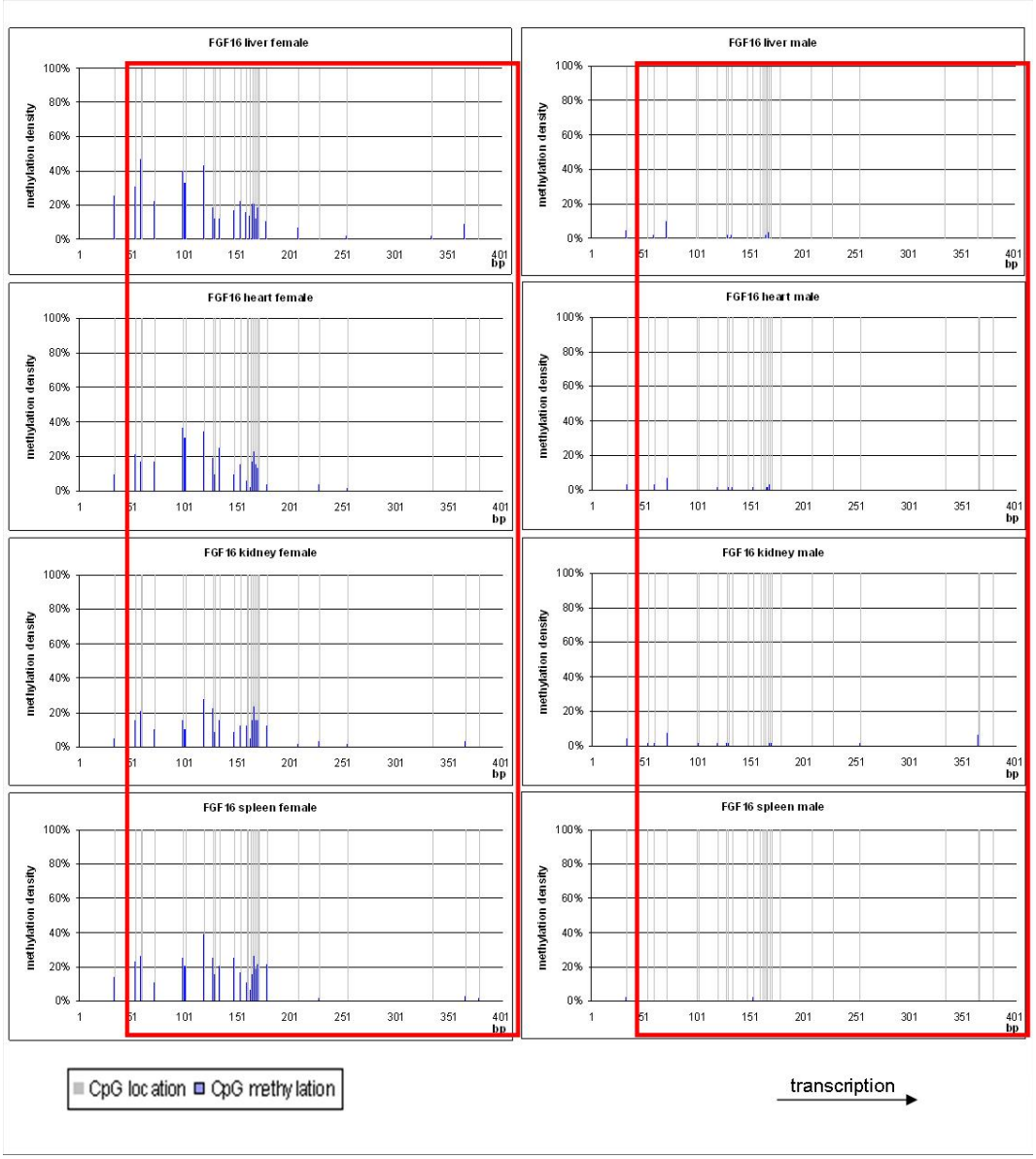

Figure 5.15: CGI methylation profiles of opossum *FGF16* in four tissues. The island region is indicated by the red box. Direction of transcription is indicated by arrow. For detailed annotation see Figure 5.12.

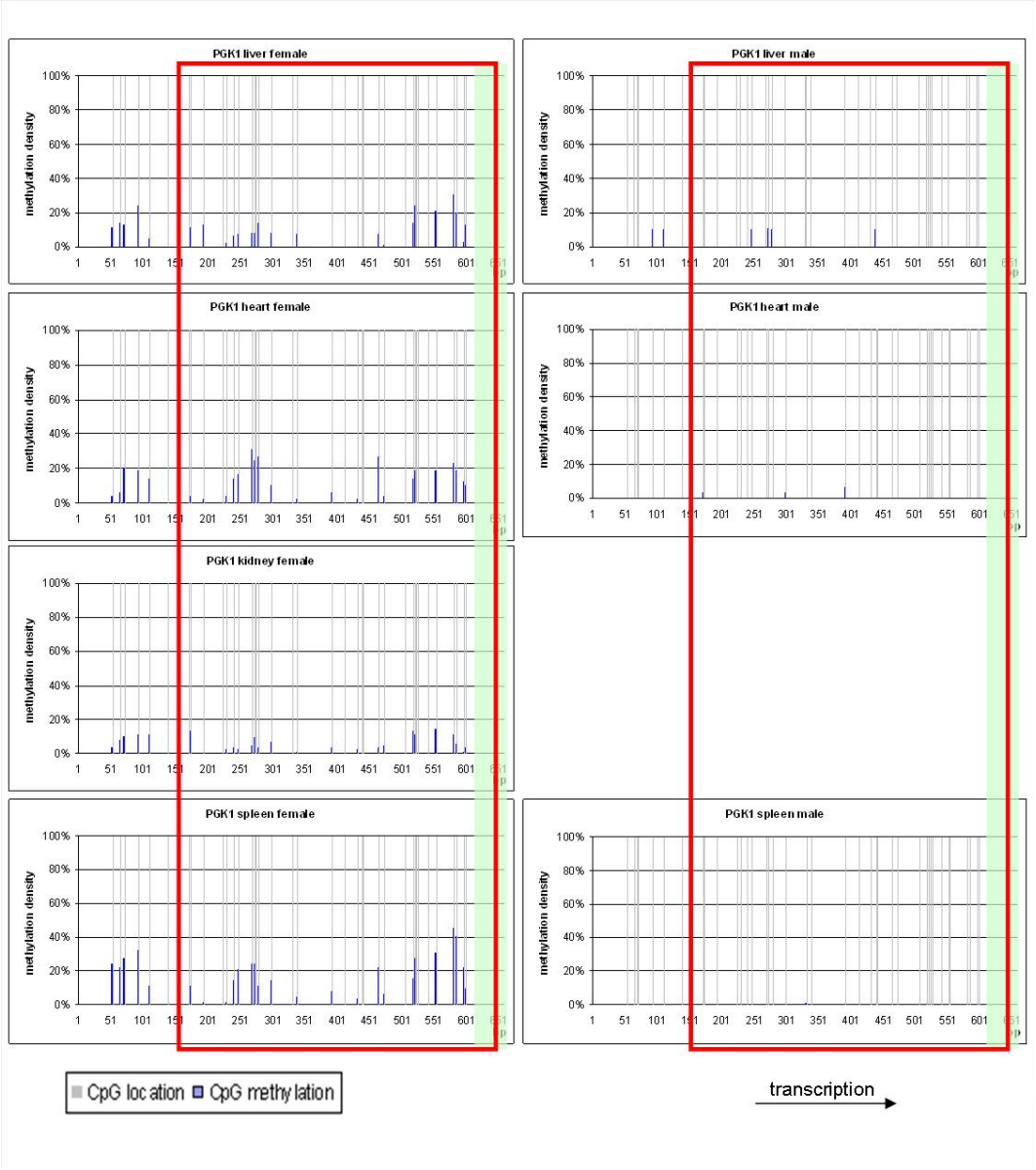

Figure 5.16: CGI methylation profiles of opossum *PGK1* in four tissues. Bisulphite sequencing using the male kidney sample was unsuccessful. The island region is indicated by the red box and the exon is shadowed in green. Direction of transcription is indicated by arrow. For detailed annotation see Figure 5.12.

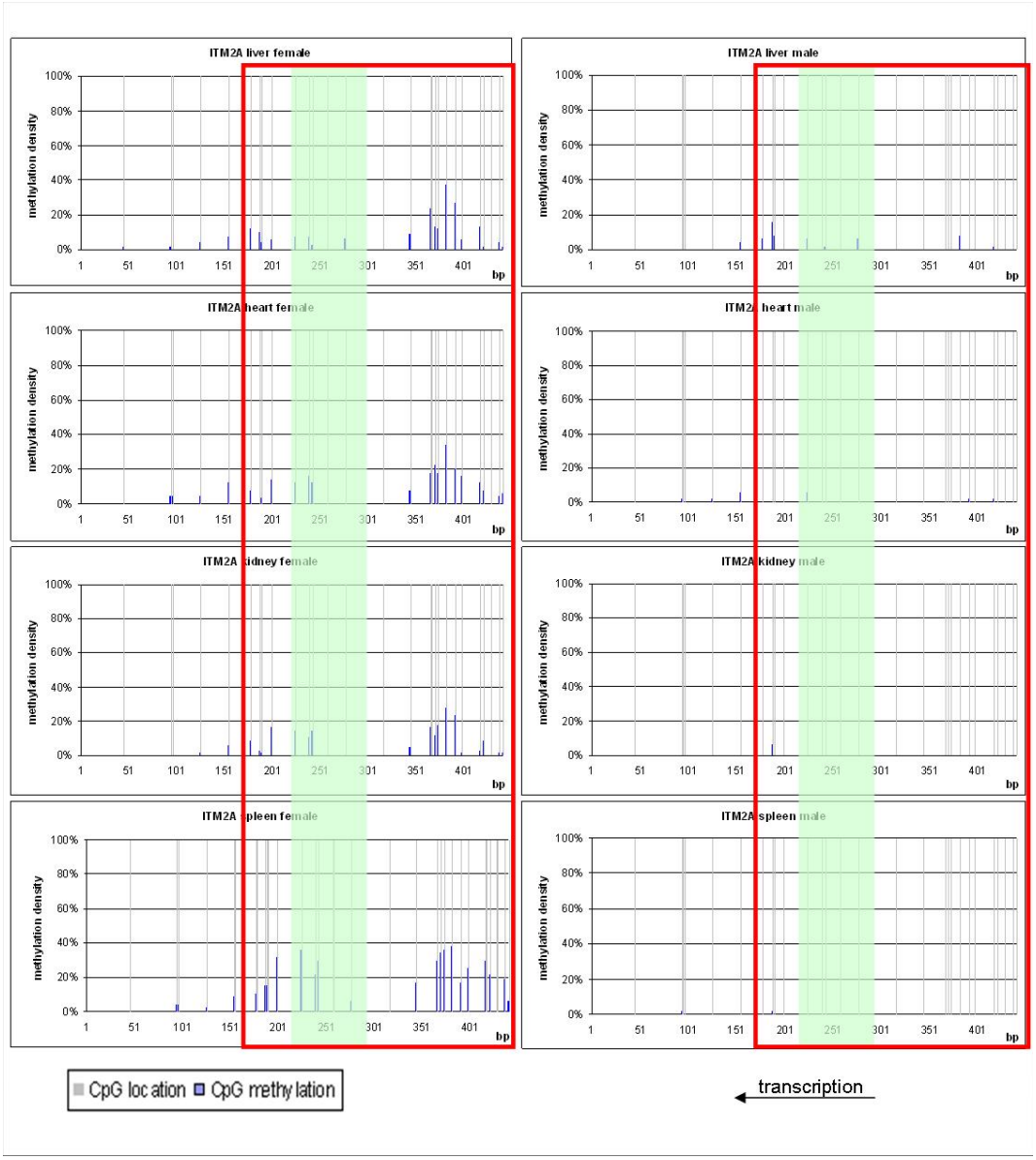

Figure 5.17: CGI methylation profiles of opossum *ITM2A* in four tissues. The island region is indicated by the red box and the exon is shadowed in green. Direction of transcription is indicated by arrow. For detailed annotation see Figure 5.12.

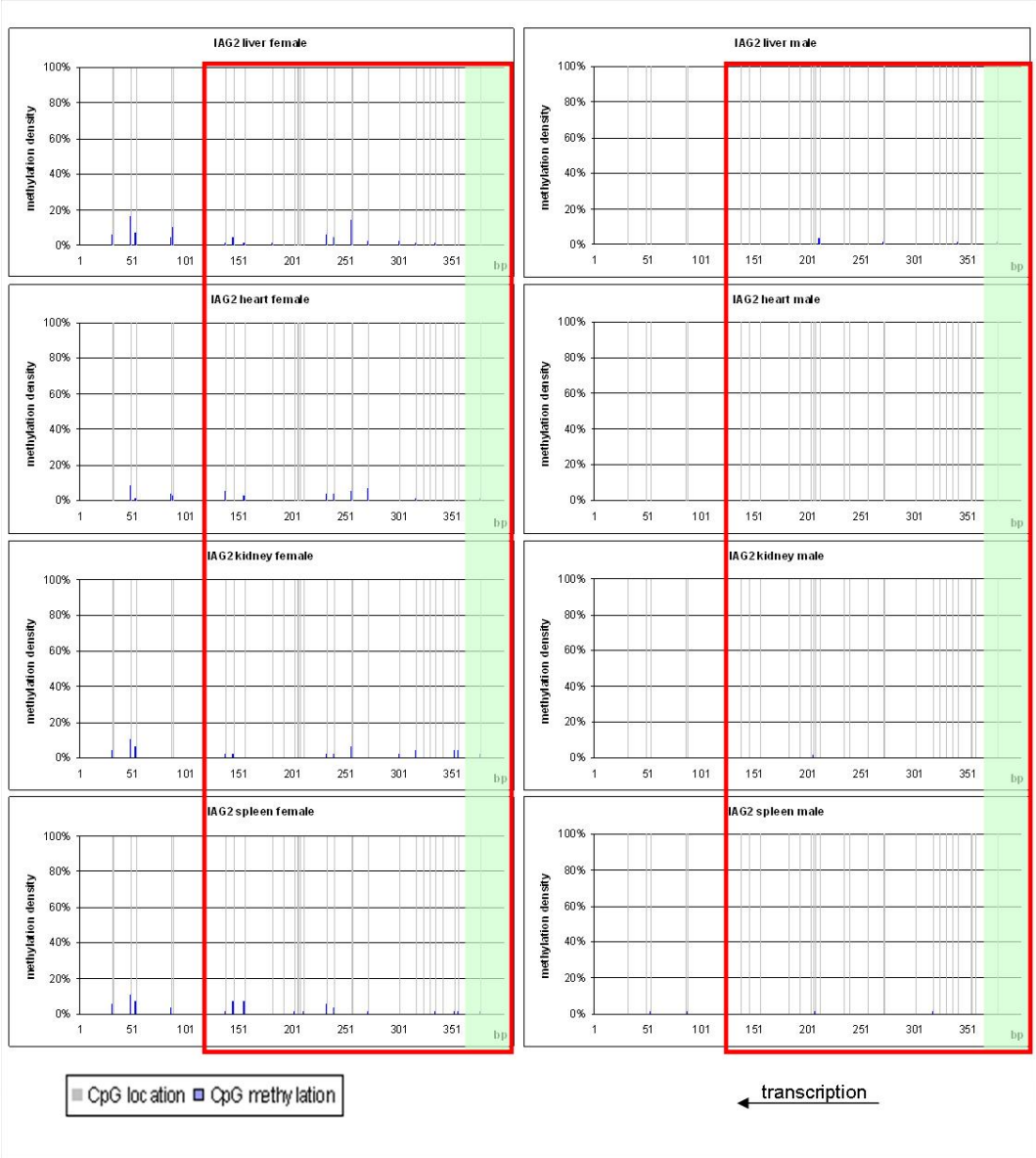

Figure 5.18: CGI methylation profiles of opossum *IAG2* in four tissues. The island region is indicated by the red box and the exon is shadowed in green. Direction of transcription is indicated by arrow. For detailed annotation see Figure 5.12.

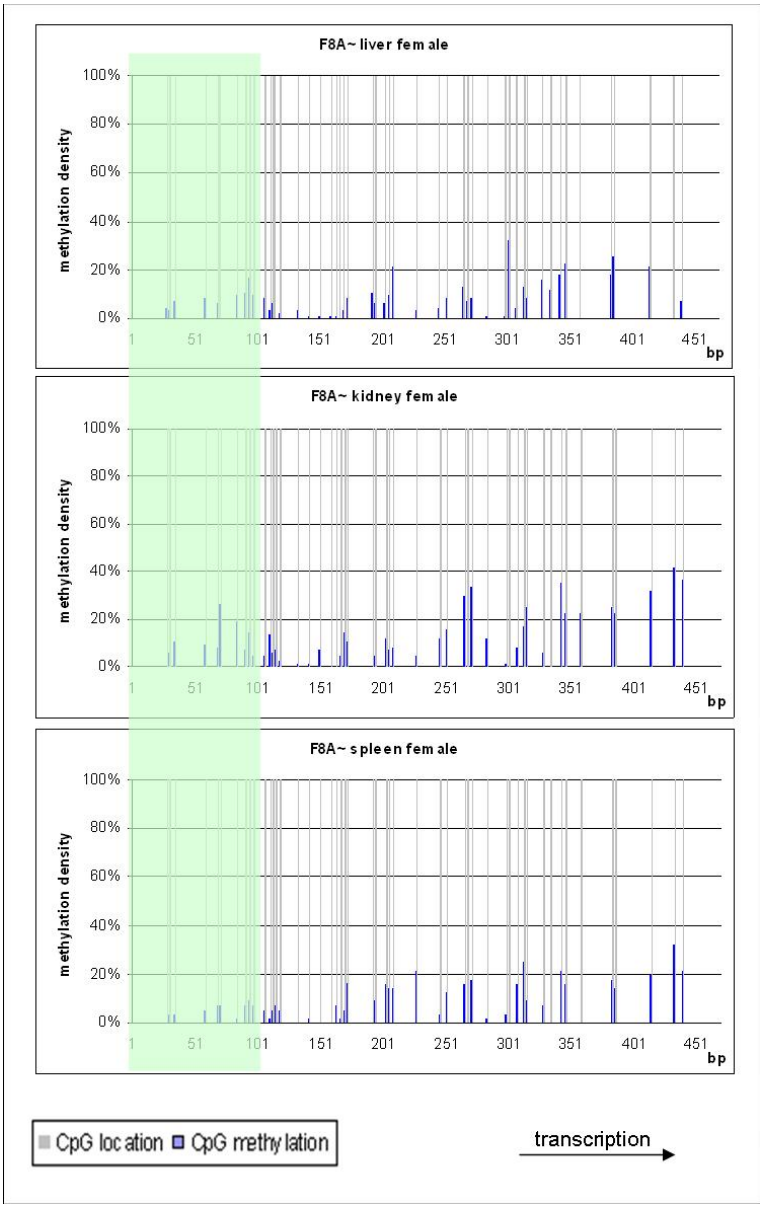

Figure 5.19: CGI methylation profiles of opossum *F8A*∼ in three tissues. Bisulphite sequencing was only successful using the female samples. The exon is shadowed in green. Direction of transcription is indicated by arrow. For detailed annotation see Figure 5.12.

5.6 Discussion

Before this study, the only two methylation studies of marsupial X-linked 5' CGIs were both in *G6PD*. Using methylation-sensitive restriction enzymes, Kaslow and Migeon (1987) found no sex difference in methylation of the CGI at the 5' end of *G6PD* in the Virginian opossum. Loebel and Johnston (1993) made the same observation in a distantly related marsupial species, the common wallaroo, using both methylation-sensitive restriction enzymes and bisulphite sequencing analysis. The study described in this chapter greatly extends our knowledge of this area by including almost a third of all CGIs of X-linked protein coding genes in the grey short-tailed opossum. Methylation analysis is described of DNA from female and male opossum using RPMA analysis for 37 CGIs. For eight of these, bisulphite sequencing analysis was performed to gain methylation profiles of single base pair resolution.

In this study, methylation of the 5' CGI of *G6PD* was examined in another marsupial species, the grey short-tailed opossum. Bisulphite sequencing analysis confirmed hypomethylation in both female and male samples in four different tissues, consistent with the previous findings in other marsupial species (Kaslow and Migeon, 1987; Loebel and Johnston, 1993). Similar to the situation in wallaroo, the majority of CpGs were completely unmethylated in all clones sequenced (Loebel and Johnston, 1993). Therefore, if the analysis in this study was also limited to this gene, I would have come to the same conclusion as previous researchers and suggest that methylation may not be a feature of CGIs on the marsupial X chromosome. However, the wealth of other information generated from the current study has presented a different picture.

For a set of seven genes, comprising nearly a fifth of all CGIs examined in this study, clear differential methylation was observed between females and males. Bisulphite sequencing analysis of the CGIs of *FAAH2*, *FGF16*, *IAG2*, *PGK1*, and *ITM2A* confirmed the sex-specific difference. While the male samples were notably hypomethylated, up to half of the clones in each female sample, presumably representing the inactive X chromosomes, were methylated to various extents (Figures 5.6-5.10), just like the CGIs of X-inactivated genes in human and mouse. This is the first observation of female-specific methylation on any marsupial X chromosome. It has demonstrated the risk of relying on only a single observation to derive a general rule.

The presence of higher CGI methylation for X-linked genes in opossum females opens up the possibility that methylation may play a role in maintaining XCI in some marsupial genes. It is noted that the methylation densities in opossum CGIs were much lower than the ones found in human and mouse, but it is not known what level of methylation is needed to silence any gene. Maybe some of the methylation densities observed in this study is already enough to have an effect. It is also possible that a lower level of methylation can stabilise silencing in conjunction with other epigenetic modifications. Having found heavier methylation in females for not one, but a group of genes in opossum, it is highly likely that female-specific methylation will be identified for additional genes in opossum and in other marsupial species. Some of these genes might have high enough methylation to repress gene expression. If methylation does play a role in marsupial XCI, it probably has not become as widespread in marsupials as in eutherians, so that many CGIs still remain unmethylated even on the inactive X.

It has been noted that XCI in marsupials is not always stable and expression

of paternal alleles have been observed in at least one species for all genes studied (reviewed in Cooper *et al.*, 1993). Previous researchers thus proposed that DNA methylation 'locks in' the inactivation state and that lack of CGI methylation on the marsupial X chromosome might be responsible for this leaky XCI (Kaslow and Migeon, 1987). Later evidence confirmed that DNA methylation is a late event in the XCI process (Lock *et al.*, 1987), but it has not been possible to test whether CGI methylation protects a marsupial gene from 'leaky' XCI, since no methylated CGI was recognised before this study. It would therefore be of interest to compare the newly uncovered methylation profiles with the effectiveness of XCI of the relevant genes. Allele-specific expression of three genes, *G6PD*, *HPRT*, and *PGK1*, has been assessed in a related species, *D. virginiana*. Interestingly, the CGI methylation profiles of these genes in *M. domestica* correlates with their paternal allele expression in *D. virginiana*. The CGI of *G6PD* was hypomethylated in both sexes in *M. domestica*; and expression of the paternal allele was detected in tissues as well as in cultured fibroblasts of *D. virginiana* females (Migeon *et al.*, 1989; Samollow *et al.*, 1989). In contrast, CGI methylation was observed here in *M. domestica* female samples for *PGK1*; and no paternal allele expression was found in either tissues or cultured fibroblasts of *D. virginiana* (Samollow 1986, Samollow 1989). *HPRT* has a similar CGI methylation profile to that of *G6PD*. If the correlation observed for *G6PD* and *PGK1* extends to *HPRT1*, expression of the paternal allele in *D. virginiana* females is expected. Although the situation *in vivo* is unknown, it has been demonstrated that the paternal allele of *HPRT* from *D. virginiana* becomes reactivated in cultured fibroblasts, which is indicative of unstable XCI (Migeon *et al.*, 1989).

During the course of this study, Hornecker and colleagues (2007) described the

first study of CGI methylation of *G6PD* and *PGK1* in *M. domestica*. Bisulphite sequencing analysis was performed on both islands in a number of tissues, including liver, which is used in my study. Similar to my observation, they also found the CGI of *G6PD* to be hypomethylated in both sexes. However, their found hypomethylation of the *PGK1* CGI in both sexes whereas my study showed clear female-specific methylation for the same island. There are a number of possible explanations for the discrepancy between the two studies. Firstly, they only assayed a small number of CpGs in a small number of molecules (for *PGK1* in liver, 17 CpGs in 6 molecules, compared with 36 CpGs in 95 molecules in the current study), so might have missed low level methylation by chance. However, their accompanying expression study, in which biallelic expression was found to be common for *PGK1*, does support lack of methylation at this CGI. It is also possible that there is big inter-individual variation in CGI methylation on the opossum X chromosome, and methylation in some individuals is too low to affect expression. The fact that *G6PD* is stably inactivated in some marsupial species despite lack of CGI methylation indicates operation of other maintenance mechanism. Therefore, if methylation plays a role in maintenance of XCI in marsupials, it might form a dispensable layer of 'extra insurance', and allows for greater extents of plasticity.

In recent years there has been increased interest in the role DNA methylation plays in regulating tissue-specific gene expression. In his initial report of CGI, Bird (1985) noted that CGIs remain unmethylated in tissue types and developmental stages where a gene is not expressed, and concluded that it is unlikely for DNA methylation to have a regulatory role in tissue-specific gene expression. More recently, an increasing number of tissue-specific differentially methylated

regions (tDMRs) have been identified and the question has been raised again. A large-scale systematic study by Eckhardt and colleagues (2006), contributing towards the International Epigenome Project, found that only 17% of the analysed genes had tDMRs located in their 5' UTRs and that two-thirds of these tDMRs did not correlate with gene expression. I found the general levels of methylation vary among different tissues for all seven CGIs examined, but not substantially (Figures 5.15-5.19). In contrast, the methylation density distribution patterns remain remarkably stable across different tissue types (Figures 5.15-5.19). Based on these results, methylation of X-linked CGIs in opossum may not play a role in tissue-specific expression control.

This study has provided the first evidence that methylation may have a role in XCI in marsupials, at least for a subset of genes. Greater understanding of the significance of CGI methylation in marsupial XCI will come from further study of expression and methylation patterns of X-linked genes in several different species.

Chapter 6

Comparative analysis of features of 5' CGIs in human, mouse and opossum

6.1 Introduction

CpG islands were initially identified when genomic DNA was digested with rare cutter restriction enzymes, whose recognition sequences are CpG rich and methylation sensitive (Antequera and Bird, 1993; Cooper *et al.*, 1983). These islands were found to have the common characteristics of high (G+C) content, and close to random expectation of CpG dinucleotide frequency (Bird, 1986). CGIs are often located at the 5' region of genes and have been implicated in transcriptional control, so are of particular interest both in gene expression studies and as markers in search for novel gene sequences. Ideally, they should be defined by directly testing for hypomethylation, but this is not always practical, especially

in chromosome or even genome scale studies, so computational methods to identify putative CGIs are very useful. Knowing the characteristics of CGIs allowed development of *in silico* methods to search for CGIs. Using the sequence data available in public databases, Gardiner-Garden and Frommer (1987) proposed a first CGI-search algorithm, in which a CGI was defined as a sequence greater than 200 bp in length, having a (G+C) content greater than 50%, and having an observed CpG / expected CpG ratio (Obs_{CpG}/Exp_{CpG}) of greater than or equal to 0.6.

The Gardiner-Garden and Frommer definition has been widely applied in analysis of CGIs, and was used in CGI identification in the International Human Genome Project (Lander *et al.*, 2001). A total of 29890 CGIs were found in the non-repetitive portion of the genomic sequence, comparable with the previous prediction of 35000 (Lander *et al.*, 2001). The CGIs are highly GC-rich. Most islands have a (G+C) content between 60% and 70%, compared to the genome average of 41%. A similar analysis of CGI was performed for the International Mouse Genome Project (Waterston *et al.*, 2002). The mouse genome was found to contain considerably less CGIs than the human genome (15500 compared with 27000, note that the number of human islands are slightly different from the previous paper, due to different computer algorithms used in the two studies), despite the slightly higher (G+C) content of the mouse genome (42% compared with 41%). In both genomes, a good correlation was found between gene density and CGI density on a chromosome, although there are a few outliers in human (Lander *et al.*, 2001; Waterston *et al.*, 2002).

The full sequence of the opossum genome has also become available recently, but analysis of opossum CGIs was not described in the report (Mikkelsen *et al.*,

2007). Therefore it is of interest to compare the features of CGIs on the X chromosomes of human, mouse, and opossum.

6.1.1 Aims of this chapter

The aims of the work described in this chapter are:

1. To explore the density and gene-association of predicted CGIs in the homologous region of the human, mouse, and opossum X chromosomes.
2. To compare characteristics of the subset of CGIs present in the 5' region of genes.
3. To compare characteristic of CGIs associated with orthologous genes in all three species.

6.2 CGIs in the conserved region of human, mouse and opossum X chromosome

The opossum X chromosome is believed to represent the ancestral X chromosome in therian mammals (Graves, 1995). Its homologous region in human, the XCR, has undergone few arrangements and includes the entire long arm and the proximal short arm (Graves, 1995). The X chromosome in mouse is more rearranged, but still well conserved in several homology blocks (Ross *et al.*, 2005); seven of them make up almost the entirety of the XCR. For the purpose of this study, the homologous regions in the three species (all referred to as the 'XCR' for convenience) were defined as the complete X chromosome in opossum, the

XCR in human, and the homology blocks corresponding to the human XCR in mouse. The boundary of the human XCR was mapped at 46.85 Mb on the human X chromosome according to the sequence of the opossum X chromosome (Mikkelsen *et al.*, 2007), in agreement with the boundary between evolutionary strata S2 and S3 as defined from analysis of the human X chromosome (Ross *et al.*, 2005). The locations of the mouse homology blocks that make up the mouse XCR were obtained from the SyntengyView of the Ensembl browser (v50).

For each XCR, all protein-coding genes and CGIs were extracted from the Ensembl database (v50) using a Perl script via Ensembl API. For each gene, 5' CGI was searched for within 5000 bp of the gene start. When a 5' CGI was found, its size in base pair, DNA sequence, and the distance from the CGI end to the gene start (a negative value indicating overlapping with gene 5' end) were recorded. The % GC content, CpG frequency, and $\text{Obs}_{\text{CpG}}/\text{Exp}_{\text{CpG}}$ ratio were calculated from the CGI sequence using the following formula:

% GC content = total(G+C) / total(ACTG) * 100

% CpG content = total(CpG) / total(ACTG) * 100

$\text{Obs}_{\text{CpG}}/\text{Exp}_{\text{CpG}}$ ratio = total(CpG) / (freq(C) * freq(G) * total(ACTG))

The features of XCR-linked protein-coding genes and 5' CGIs associated with such genes in these three species are summarised in Table 6.1. All three XCR have similar (G+C) content and abundance of predicted CGIs, but a significantly lower proportion of mouse genes have CGIs present in their 5' regions. In addition, the mouse 5' CGIs are smaller than the opossum and human ones. The distribution of 5' CGI sizes were plotted for all three species in Figure 6.1. Whereas the opossum and human CGIs show a more even distribution between 400 and 1000 bp, small CGIs with sizes between 400 and 600 bp make up a good proportion

of the mouse CGIs. However, the mouse CGIs are no 'weaker' than the opossum and human ones in terms of (G+C) content and Obs_{CpG}/Exp_{CpG} ratios.

Over 90% of human and mouse 5' CGIs overlap with the gene start, but this is only the case for two-thirds of opossum 5' CGIs. One possible cause of this difference is higher number of incomplete annotation of 5' gene start in opossum. To investigate this possibility, annotation of several opossum genes with non-overlapping 5' CGIs were manually checked in the Ensembl browser, and in all cases the 5' end of predicted transcript was incomplete.

Table 6.1: Comparison of XCR gene and CGI features in opossum, human, and mouse.

	Opossum	Human	Mouse
Properties of XCR			
Size of XCR (Mb)	79	108	99
GC content of XCR	41%	39%	39%
Protein-coding genes	443	684	662
Gene density (per Mb)	5.6	6.3	6.7
CGI	343	539	550
CGI density (per Mb)	4.3	5.0	5.6
Genes with 5' CGI	164 (37%)	316 (46%)	168 (25%)
Properties of 5' CGI			
Mean size (bp)	1024	1034	882
CGI overlapping gene start	62%	90%	93%
GC content mean	68%	68%	67%
GC content range	54-80%	54-79%	53-76%
Obs/Exp CpG freq mean	0.81	0.83	0.85
Obs/Exp CpG freq range	0.55-1.23	0.57-1.16	0.56-1.06

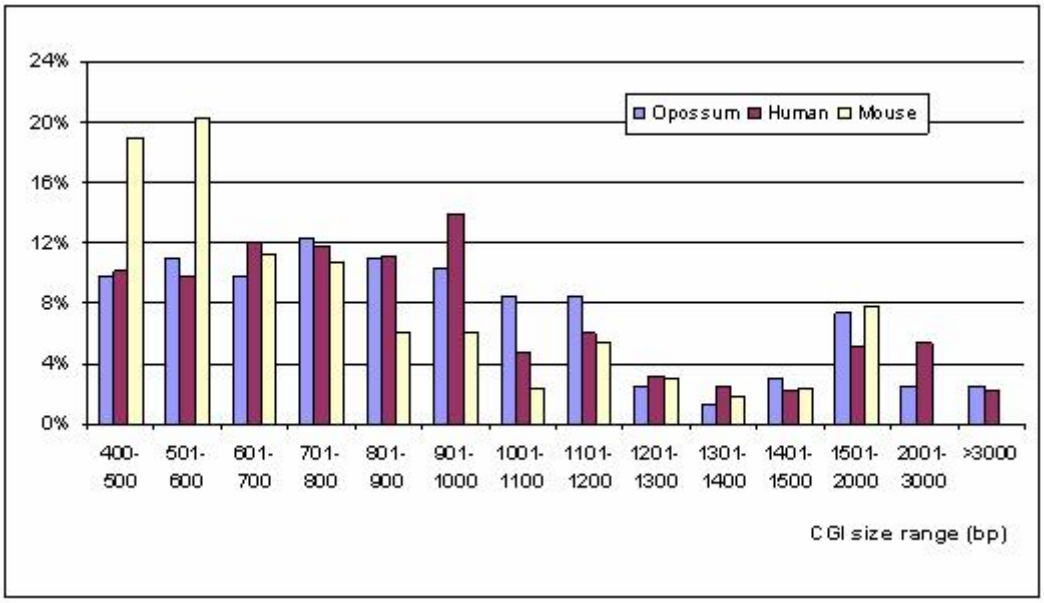

Figure 6.1: Comparison of sizes of 5' CGIs in opossum, human, and mouse XCR.

6.3 CGIs associated with orthologous genes on human, mouse and opossum XCR

For each XCR-linked protein-coding gene in each species, X-linked putative orthologues in the other two species were identified, again using a Perl script via Ensembl API. In the cases of presence of paralogues, all paralogous genes were recorded. The resulting lists of orthologous genes for each species were compared manually for consistency. Three human orthologues of opossum X-linked genes are outside the human XCR. Sixteen mouse orthologues of human XCR genes are outside of the mouse XCR homology blocks. The numbers of orthologous genes are summarised in Figure 6.2. A set of 280 genes are preserved on the X chromosomes of all three (thereafter referred to as the 'shared gene set').

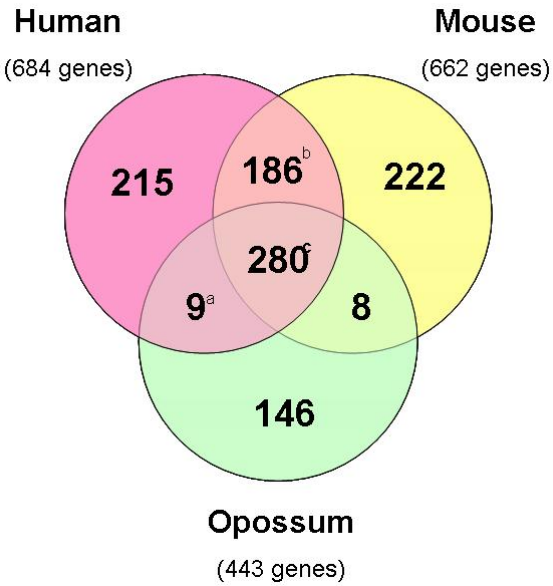

Figure 6.2: Orthologous genes on the opossum, human, and mouse X chromosomes. The number under each species name is the total number of protein-coding genes in each XCR. Genes that were identified as orthologues of the same gene (i.e. paralogues) were counted as one orthologue. Symbols: [a], two human genes are outside of XCR; [b], eight mouse genes are outside of XCR; [c], one human gene and eight mouse genes are outside of XCR.

Features of 5' CGIs associated with these 280 genes in opossum, human, and mouse are compared in Table 6.2. In all three species but especially in mouse, the shared gene set is enriched with 5' CGI-associated genes (compare Table 6.2 with Table 6.1). The average sizes of 5' CGIs associated with this set of genes are also slightly larger than the XCR average. The numbers of genes with a 5' CGI in the shared gene set are summarised in Figure 6.3. A set of 53 genes have a 5' CGI in all three species.

Table 6.2: Comparison of 5' CGIs associated with the orthologous gene set.

	Opossum	Human	Mouse
Genes with 5' CGI	118 (42%)	151 (54%)	96 (34%)
Mean size (bp)	1062	1060	893
CGI overlapping gene start	65%	93%	91%
GC content mean	68%	67%	66%
GC content range	54-80%	54-79%	66-76%
Obs/Exp CpG freq mean	0.80	0.87	0.85
Obs/Exp CpG freq range	0.55-1.23	0.59-1.1	0.76-1.06

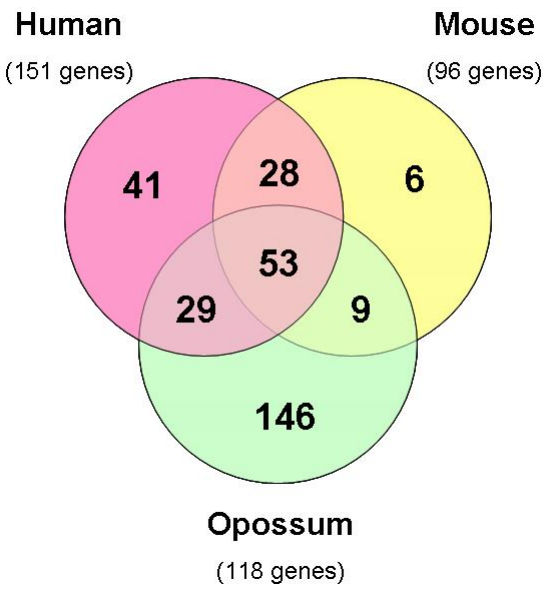

Figure 6.3: CGIs associated with the orthologous gene set. The number under each species name is the total number of genes that are associated with a 5' CGI.

6.4 Conservation of 5' CGIs associated with orthologous genes

In the shared gene set, 53 genes have a 5' CGI identified in all three species. It is of interest to investigate whether sequence homology is conserved in CGIs

associated with orthologous genes in different species. For each of the 53 sets of orthologous genes, CGI sequences from all three species were aligned using the ClustalW2 multiple sequence alignment programme (Larkin *et al.*, 2007) and the distribution of percentage identity scores are plotted in Figure 6.4. The homology between human and mouse islands is on average 70%. There is generally a lack of sequence homology between human and opossum islands or between mouse and opossum islands. Details of the 53 genes and their CGIs in the three species are recorded in Appendix III.

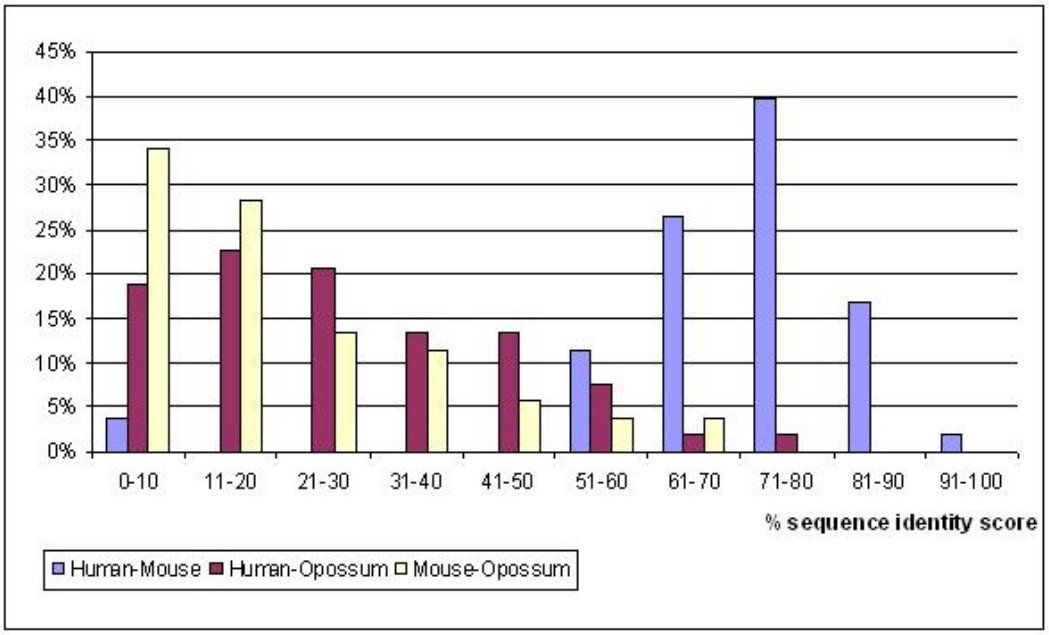

Figure 6.4: Homology between 5' CGIs of the orthologous genes.

To explore the relationship of island homology and gene homology, the full gene and island sequences were compared for several sets of orthologous genes using the VISTA server (Frazer *et al.*, 2004).

The CGI of *PGK1* is differentially methylated in human and mouse, and dif-

ferential methylation was also detected in opossum in the current study. Human-mouse homology and human-opossum homology for this island are both close to average, but opossum-mouse homology is low. Alignments of gene and island sequences in the three species showed that homology is limited to exons (Figure 6.5). The portion of CGIs outside gene start showed poor homology.

SOX3, presumably the oldest gene recruited into XCI, has a very high sequence identity between CGIs in the three species. However, for the gene in all three species, most of the island overlaps with the exons, and the island homology is mainly accounted for by gene homology (Figure 6.6). Like the case of *PGK1*, the portions of CGIs outside exons showed poor homology.

The CGI of *PGRMC1* showed the highest sequence homology between human and opossum, as well as high sequence homology between mouse and opossum. Again, the homology can largely be explained by the great extent of overlapping between the island and exon of genes (Figure 6.7). The CGI covers the entire exon1 in all three species, and is almost entirely made up of exonic sequence in opossum.

PIM2 showed among the lowest island homology between human and opossum or between mouse and opossum. The CGI overlaps with exons 1 and 2 in all three species, but a high sequence divergence between opossum and human or mouse was revealed for both exons (Figure 6.8).

In all cases, the sequence homology between CGIs in different species reflects the extent of overlap between islands and exons and the sequence homology between the exons.

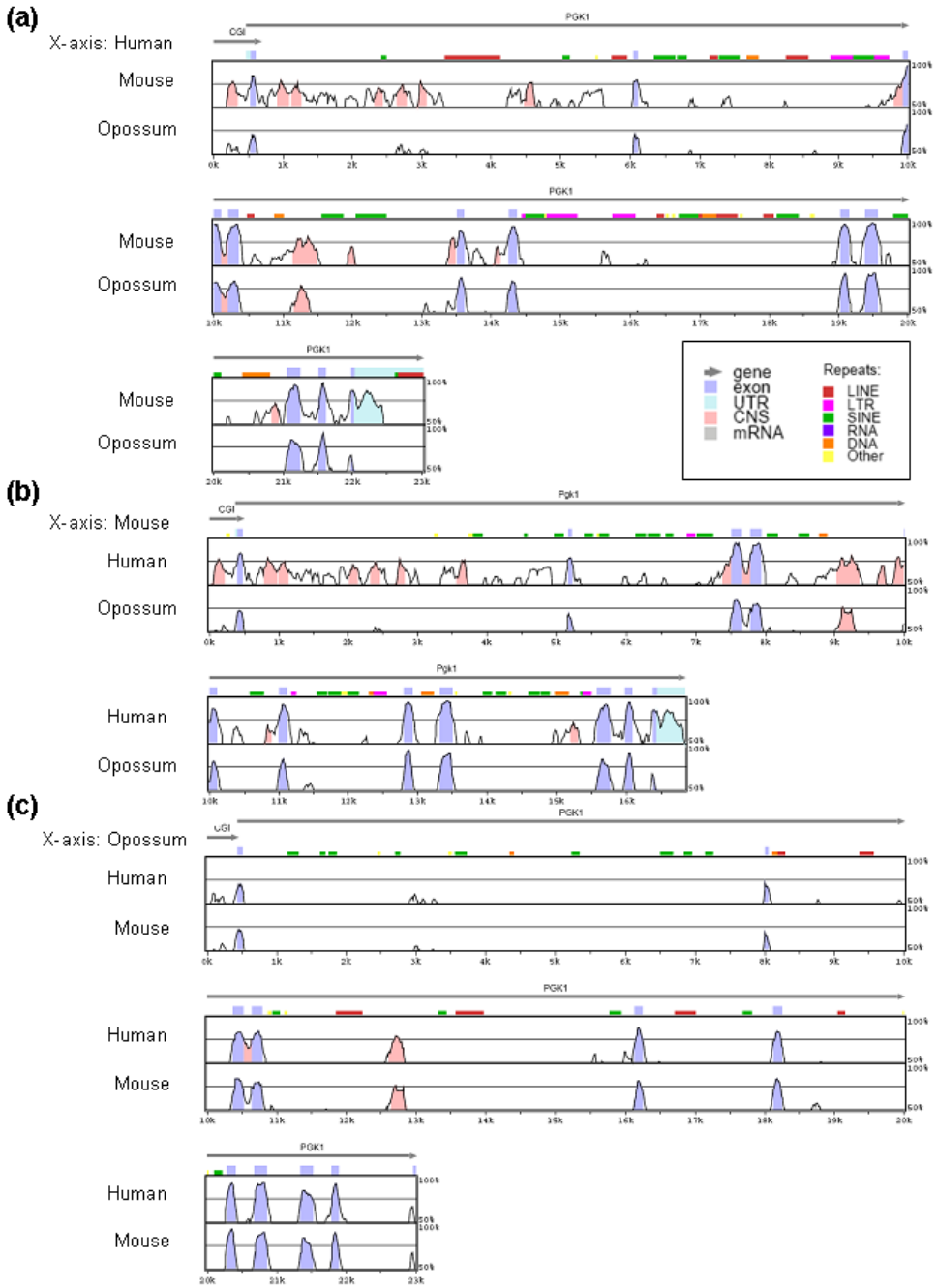

Figure 6.5: VISTA plot of human, mouse and opossum *PGK1* and its CGI. Sequences for human, mouse, and opossum were compared using the MLAGAN alignment program. (a) shows the alignment of the human locus with the mouse and opossum loci; (b) shows the alignment of the mouse locus with the human and opossum loci; and (c) shows the alignment of the opossum locus with the human and mouse loci. Conserved regions with more than 70% sequence similarity (Y-axis) over a 100 base pair window are coloured: conserved non-coding sequences in pink, exons in violet, and UTRs in light-blue.

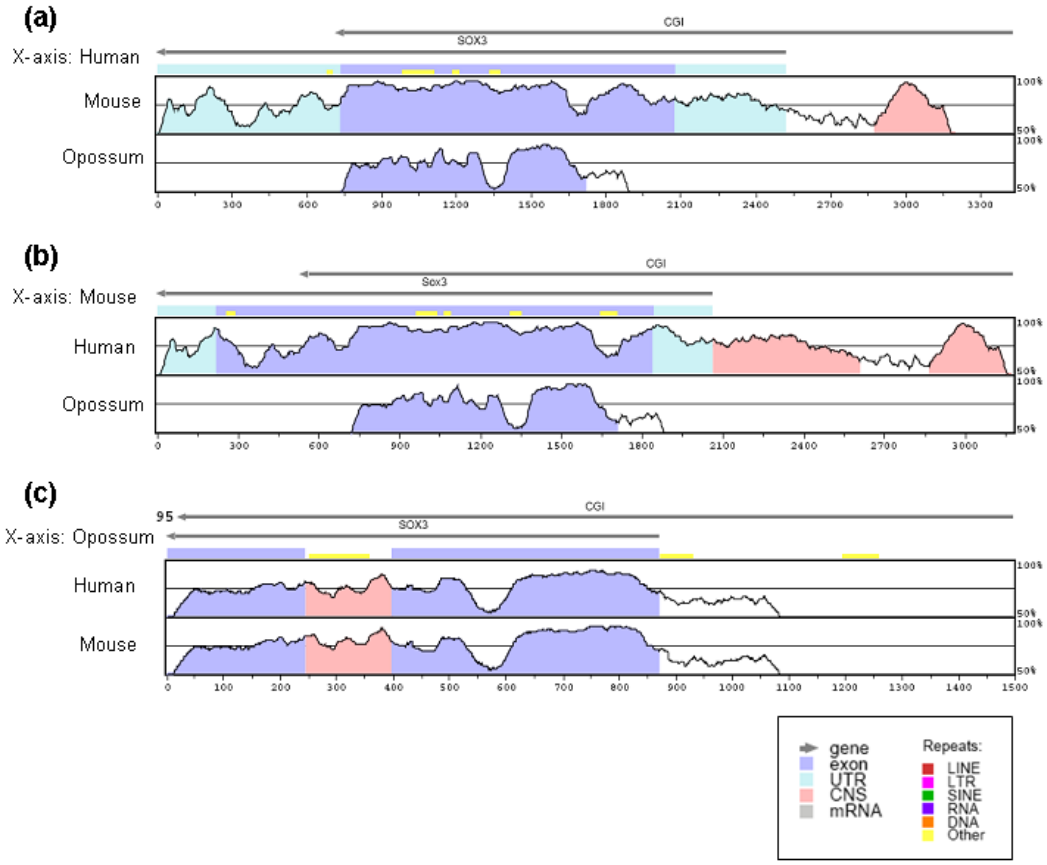

Figure 6.6: VISTA plot of human, mouse and opossum *SOX3* and its CGI. Sequences for human, mouse, and opossum were compared using the MLAGAN alignment program. (a) shows the alignment of the human locus with the mouse and opossum loci; (b) shows the alignment of the mouse locus with the human and opossum loci; and (c) shows the alignment of the opossum locus with the human and mouse loci. Conserved regions with more than 70% sequence similarity (Y-axis) over a 100 base pair window are coloured: conserved non-coding sequences in pink, exons in violet, and UTRs in light-blue.

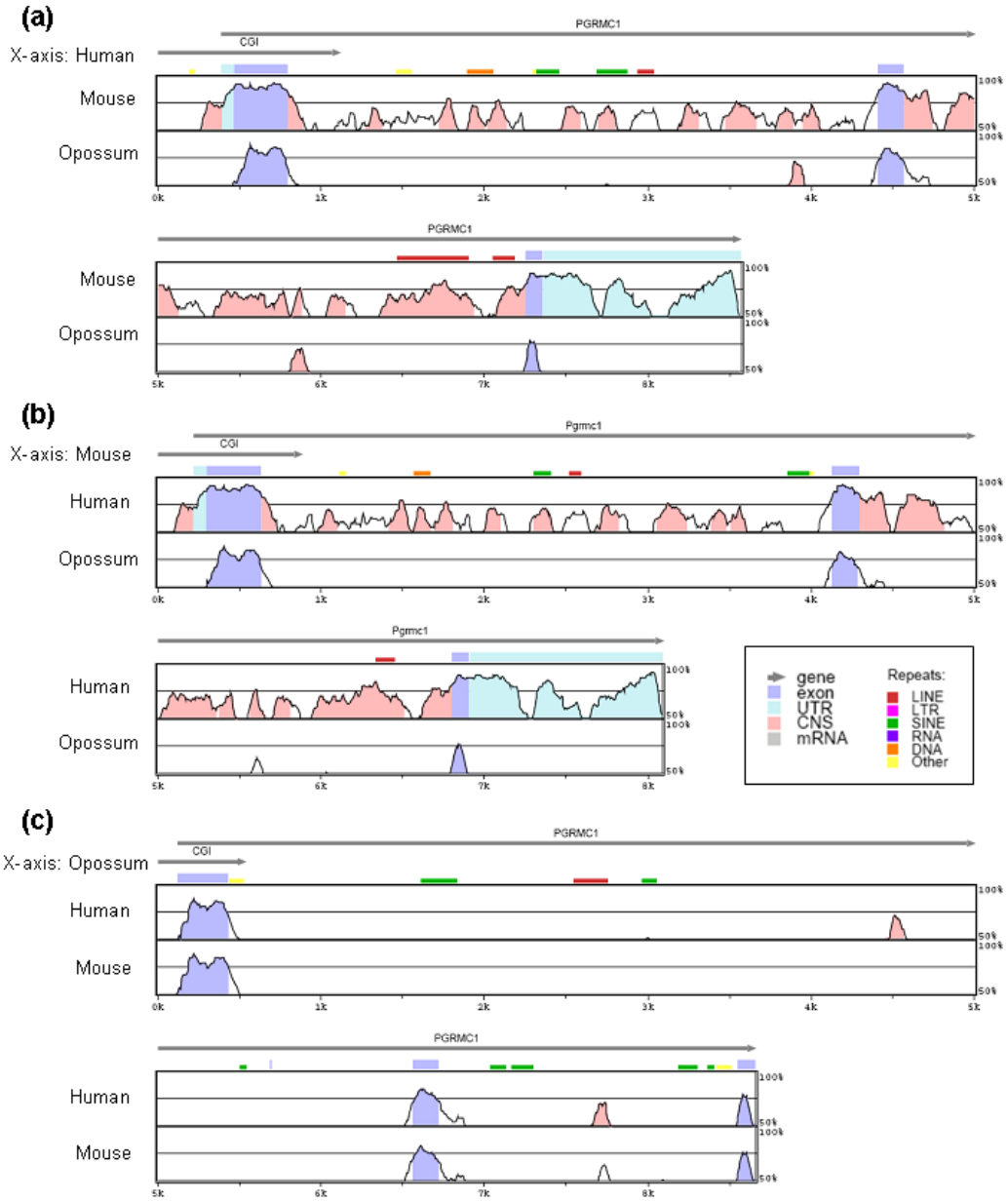

Figure 6.7: VISTA plot of human, mouse and opossum *PGRMC1* and its CGI. Sequences for human, mouse, and opossum were compared using the MLAGAN alignment program. (a) shows the alignment of the human locus with the mouse and opossum loci; (b) shows the alignment of the mouse locus with the human and opossum loci; and (c) shows the alignment of the opossum locus with the human and mouse loci. Conserved regions with more than 70% sequence similarity (Y-axis) over a 100 base pair window are coloured: conserved non-coding sequences in pink, exons in violet, and UTRs in light-blue.

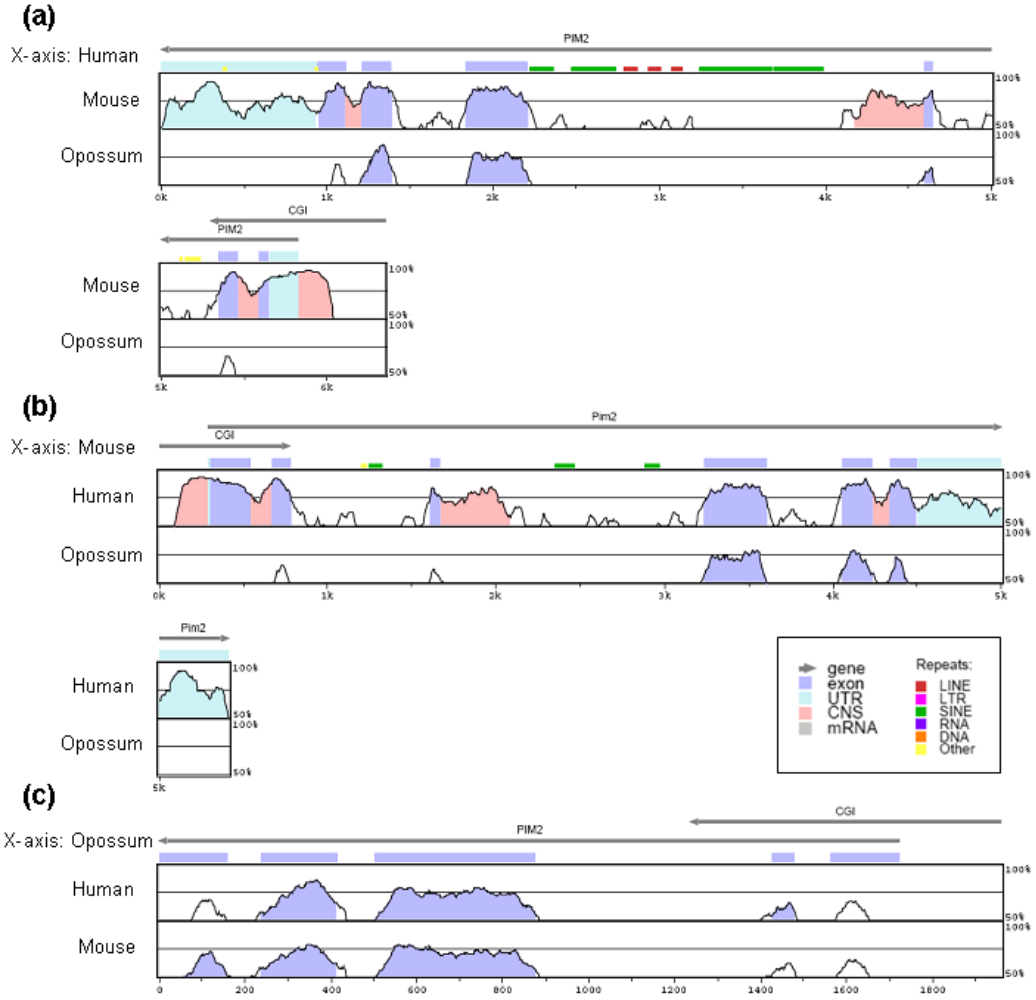

Figure 6.8: VISTA plot of human, mouse and opossum *PIM2* and its CGI. Sequences for human, mouse, and opossum were compared using the MLAGAN alignment program. (a) shows the alignment of the human locus with the mouse and opossum loci; (b) shows the alignment of the mouse locus with the human and opossum loci; and (c) shows the alignment of the opossum locus with the human and mouse loci. Conserved regions with more than 70% sequence similarity (Y-axis) over a 100 base pair window are coloured: conserved non-coding sequences in pink, exons in violet, and UTRs in light-blue.

6.5 Discussion

In this chapter a comparative analysis of 5' CGIs in the conserved region of the human, mouse and opossum X chromosomes was carried out.

The (G+C) content and CGI density of the human XCR are very similar to the whole human X chromosome (Ross *et al.*, 2005), but much lower than the autosomes (Lander *et al.*, 2001). It has been noted that the distribution of CGIs in the human genome was found to vary greatly among some of the chromosomes. While most chromosomes have around 10 CGIs per Mb, the Y chromosome has only 2.9 CGIs per Mb, and chromosome 19, on the other extreme, has 43 CGIs per Mb. The density of CGIs on individual chromosomes correspond relatively well with the density of genes, although there are a few outliers. Apart from the Y chromosome, consistent with its low gene density, the X chromosome has the lowest predicted CGI frequency, at only 5.25 CGI per Mb, which is half of the estimated genome average (Lander *et al.*, 2001; Ross *et al.*, 2005).

Similar to the situation in human, the mouse X chromosome is also among the mouse chromosomes with lowest (G+C) content and CGI density (Waterston *et al.*, 2002). Interestingly, the mouse XCR has slightly higher frequency of CGI than the human XCR, despite the generally lower (G+C) content and CGI frequency of the mouse genome. This is possibly due to the more homogeneous distribution of CGIs among the mouse chromosomes, where no extreme CGI densities like that of human chromosome 19 can be found (Waterston *et al.*, 2002).

The opossum XCR, which is the entire X chromosome, has a slightly higher (G+C) content than that of human and mouse XCR. The CGI frequency is

slightly lower than that of human and mouse, but the CGI to gene ratios are very similar in these three species. However, this is not the general pattern of the opossum genome. As described in Chapter 5, the opossum genome has low overall (G+C) content and CpG density, but curiously the (G+C) content of the opossum X is similar to that of human and mouse, and actually highest in all amniotes. The CpG density of the opossum X is also much higher than the genome average (1.4 compared with 0.9), but similar to that of human and mouse. There has been speculation that this is due to the short length and high recombination frequencies of the opossum X (Duret *et al.*, 2006).

A set of 280 genes were found to be shared by the X chromosome of all three species. These make up a third of all XCR genes in human and mouse, and half of all XCR genes in opossum. The good gene conservation is consistent with Ohno's law, which predicted a very conserved mammalian X chromosome as a result of the XCI (Ohno, 1967). The evolutionary distance between two species is reflected by the number of orthologous genes. Apart from the 280 genes shared by all three species, more than 200 genes are shared by human and mouse, but only less than ten genes are shared by opossum and human or opossum and mouse.

The set of genes preserved on all three X chromosomes were found to be enriched with CGI-association. One possible explanation is that most housekeeping genes are associated with a 5' CGI (Larsen *et al.*, 1992) and the vital housekeeping genes are less likely to be lost from the X chromosome. It is also possible that rearrangement and duplication of genes might disrupt the association of genes and CGIs.

A set of 53 genes from the preserved gene set are associated with a 5' CGI in all three species. Sequence alignment of the homologous CGIs showed good

conservation between human and mouse but poor conservation between human and opossum or mouse and opossum. When the alignment was extended to the full length of genes, it became clear that the sequence homology is mainly limited to exons, suggesting that sequence conservation is not necessary for island function.

Chapter 7

Discussion

This thesis has described the analysis of 5' CGI methylation states in a large number of X-linked genes in three species, (human, mouse, and opossum) belonging to the two extant mammalian groups that use X chromosome inactivation for dosage compensation.

A dynamic landscape of X chromosome CGI methylation patterns has been revealed in the females of these three species. On one extreme, the vast majority of CGIs assayed in mouse were found to be methylated in mouse females following the RPMA analysis (Chapter 3). Single base pair resolution methylation maps generated by bisulphite sequencing revealed a heterogeneous pattern of methylation within individual islands, but most of the methylated island molecules tend to have more than half of the CpG dinucleotides methylated (Chapter 4). Only three CGIs were hypomethylated in females, where the lack of methylation was almost complete, indistinguishable from the methylation pattern seen in males. A different pattern was seen in the human CGIs. In the RPMA analysis, a much greater proportion of human islands were found to be hypomethylated

in females as well as in males (Chapter 3). The difference between human and mouse is also obvious in the levels of methylation in individual CGIs. As shown by bisulphite sequencing, the human islands were never methylated to the same extent as that of their mouse homologues (Chapter 4). In addition, intermediate levels of methylation were found in a good proportion of human CGIs, but never in mouse (Chapters 3 and 4). At the other end of the spectrum, the first large scale methylation analysis of marsupial CGIs to date revealed an overall hypomethylation of 5' CGIs on the opossum X chromosome.

Methylation of 5' CGIs is rarely seen on the autosomes, but very common on eutherian X chromosomes, presumably playing a role in maintaining the silencing state of genes on the inactive X (Kaslow and Migeon, 1987). One aim of this thesis was to test whether CGI methylation serves as a good indicator for a gene's XCI status, as previous research has established a correlation between CGI methylation and gene silencing from XCI (Tribioli *et al.*, 1992). The work presented in this thesis has greatly extended this correlation (Chapter 3). For the human genes, a comparison of their CGI methylation states with their XCI status determined using the somatic cell hybrid system (Carrel and Willard, 2005) demonstrated a very strong correlation between CGI methylation and gene silencing (49/53 in agreement). For the mouse genes with known XCI status, their CGI methylation states enabled perfect prediction of their XCI status.

The findings in this thesis have also provided novel XCI information for a great number of genes. For eleven human genes that could not be assayed in the previously published human XCI profile (Carrel and Willard, 2005), an XCI state prediction was made based on their 5' methylation profile in this study. In the case of mouse, XCI information uncovered for most genes in this thesis is novel.

Chapter 7

Discussion

This thesis has described the analysis of 5' CGI methylation states in a large number of X-linked genes in three species, (human, mouse, and opossum) belonging to the two extant mammalian groups that use X chromosome inactivation for dosage compensation.

A dynamic landscape of X chromosome CGI methylation patterns has been revealed in the females of these three species. On one extreme, the vast majority of CGIs assayed in mouse were found to be methylated in mouse females following the RPMA analysis (Chapter 3). Single base pair resolution methylation maps generated by bisulphite sequencing revealed a heterogeneous pattern of methylation within individual islands, but most of the methylated island molecules tend to have more than half of the CpG dinucleotides methylated (Chapter 4). Only three CGIs were hypomethylated in females, where the lack of methylation was almost complete, indistinguishable from the methylation pattern seen in males. A different pattern was seen in the human CGIs. In the RPMA analysis, a much greater proportion of human islands were found to be hypomethylated

in females as well as in males (Chapter 3). The difference between human and mouse is also obvious in the levels of methylation in individual CGIs. As shown by bisulphite sequencing, the human islands were never methylated to the same extent as that of their mouse homologues (Chapter 4). In addition, intermediate levels of methylation were found in a good proportion of human CGIs, but never in mouse (Chapters 3 and 4). At the other end of the spectrum, the first large scale methylation analysis of marsupial CGIs to date revealed an overall hypomethylation of 5' CGIs on the opossum X chromosome.

Methylation of 5' CGIs is rarely seen on the autosomes, but very common on eutherian X chromosomes, presumably playing a role in maintaining the silencing state of genes on the inactive X (Kaslow and Migeon, 1987). One aim of this thesis was to test whether CGI methylation serves as a good indicator for a gene's XCI status, as previous research has established a correlation between CGI methylation and gene silencing from XCI (Tribioli *et al.*, 1992). The work presented in this thesis has greatly extended this correlation (Chapter 3). For the human genes, a comparison of their CGI methylation states with their XCI status determined using the somatic cell hybrid system (Carrel and Willard, 2005) demonstrated a very strong correlation between CGI methylation and gene silencing (49/53 in agreement). For the mouse genes with known XCI status, their CGI methylation states enabled perfect prediction of their XCI status.

The findings in this thesis have also provided novel XCI information for a great number of genes. For eleven human genes that could not be assayed in the previously published human XCI profile (Carrel and Willard, 2005), an XCI state prediction was made based on their 5' methylation profile in this study. In the case of mouse, XCI information uncovered for most genes in this thesis is novel.

The differences in CGI methylation between human and mouse correlate with the extent of escape from XCI seen in these species, providing some clue to the mechanism of escape. It is currently not clear if escapees occur from failure to establish or failure to maintain inactivation. The latter scenario is supported by experimental demonstration that a mouse escapee gene, *Smcx*, is inactivated in early embryonic development followed by re-activation (Lingenfelter *et al.*, 1998). The progressive loss of CGI methylation has been proposed as a possible reason for re-activation (Disteche *et al.*, 2002). The findings in this thesis are also in support of this hypothesis. In mouse, the vast majority of CGIs studied were hypermethylated, possibly contributing towards more complete XCI than in human. In human, a good proportion of CGIs studied showed low level or hypomethylation, in agreement with the extent of escape from XCI. The lower level of methylation in human than mouse is consistent with the suggestion that escape from XCI is associated with a failure to maintain the inactive state.

Very little is known about the inactivation status of X-linked genes in marsupials. Prior to this study, allozyme studies of a small number of genes showed that these genes are less stably inactivated in marsupials than in human and mouse (reviewed in Cooper *et al.*, 1993). The less stable XCI in marsupials has been linked to lack of CGI methylation (Kaslow and Migeon, 1987), which was confirmed for *G6PD* in several marsupial species (Kaslow and Migeon, 1987; Loebel and Johnston, 1996). Findings presented in this thesis have greatly expanded our knowledge about CGI methylation on the marsupial X chromosome. CGI methylation analysis was performed for 36 novel genes, comprising almost a third of all assayable genes on the X chromosome of a model marsupial, the grey short-tailed opossum (*M. domestica*). Four-fifths of them showed indistinguishable patterns

of hypomethylation between female and male. If escaping from XCI is a result of failure to maintain the inactive state, as speculated above, a high level of escape from XCI might be anticipated in marsupials. However, this may not be the case as there might be other maintenance mechanisms in place. For example, the CGI of *G6PD* is not methylated in the common wallaroo (Loebel and Johnston, 1996), but the paternal allele is completely repressed in this species, and re-activation was only seen in culture cell lines (Johnston *et al.*, 1978; Richardson *et al.*, 1971).

This thesis has also demonstrated for the first time existence of female-specific methylation on a marsupial X chromosome. Based on limited evidence, it is believed that methylation is not part of the molecular mechanism of marsupial XCI (Graves, 1996, 2006), but a fifth of all genes assayed in this study showed some degree of differential methylation between the two sexes (Chapter 5). Bisulphite sequencing analysis of seven CGIs in four different tissues confirmed the presence of female-specific methylation, and also revealed conservation of methylation patterns across different tissues. It is not clear whether such degree of methylation may bear any functional significance in the absence of expression data, but interestingly, the CGI methylation states of three genes uncovered in this study correlate with their expression from the inactive X in a related species, the Virginian opossum, in the same fashion that CGI methylation profiles of human and mouse genes correlate with their XCI status. It leaves open the possibility that DNA methylation may play a role in maintaining XCI for a subset of marsupial genes.

7.1 Future directions

Many of the features associated with XCI in the eutherian mammals are absent from the XCI system in the marsupial mammals, which is thought to represent a more ancestral mechanism. Studying the details of XCI in marsupials is therefore of key importance to help our understanding of the evolutionary origin of this unique form of dosage compensation in therian mammals. Contrary to the common view that CGI methylation is absent from the marsupial X chromosome, this thesis has described the first evidence of female-specific CGI methylation on a marsupial X (Chapter 5). It is of immediate interest to investigate whether this extent of methylation may have any functional significance in regulation of marsupial XCI. Expression studies could be carried out in parallel with methylation studies to see if level of methylation is correlated with expression from the inactive X. The imprinted XCI in marsupial enables easy identification of gene expression from the active or inactive X chromosome, which can only be achieved in human and mouse using experimental systems that don't necessarily represent situations in normal cells. This thesis has also shown a stable methylation pattern across different tissues in the same individual. This could be extended to comparison of methylation profiles of the same CGIs among different individuals. If such methylation bears functional significance, it is more likely to be preserved in multiple individuals. In addition to DNA methylation, it is worthy to search for other factors that may affect stability of XCI in marsupials, for example histone modifications.

In the eutherian domain, little is known about the XCI status of X-linked genes in most eutherian mammals, as the expression-based techniques used to

assess XCI in human and mouse are mostly difficult to extend to other species. In order to understand how the landscape of gene silencing has evolved, it is necessary to study a group of genes' XCI status in a number of species on different branches of the mammalian evolutionary tree. Moreover, multi-species comparison will give an opportunity to define more precisely the characteristics that influence the XCI status of a gene or region. This thesis has demonstrated an indirect method to predict rapidly the XCI status of a large number of genes based on their 5' CpG island methylation status (Chapter 3), and therefore provided a platform for constructing a comparative XCI profile in mammals. Recent advances in sequencing technology have greatly promoted the growth of genomic sequencing projects. During the course of this project, full genomic sequence has become available for an increasing number of mammals, covering a great total branch length of the evolutionary tree, and providing a good representation of the diversity of mammalian species (Figure 7.1). In this thesis, the S3 region, containing both inactivated and escapee genes, where different XCI patterns are known between different species, was studied in human and mouse (Chapters 3 and 4). This work can be extended to other mammalian species to make predictions of the ancestral XCI status for a number of genes. As demonstrated by Jegalian and Page's study of the XCI status of four genes in 18 species (1999), generation of XCI profiles for even a small number of genes in these species will greatly expand our understanding of the evolution of XCI.

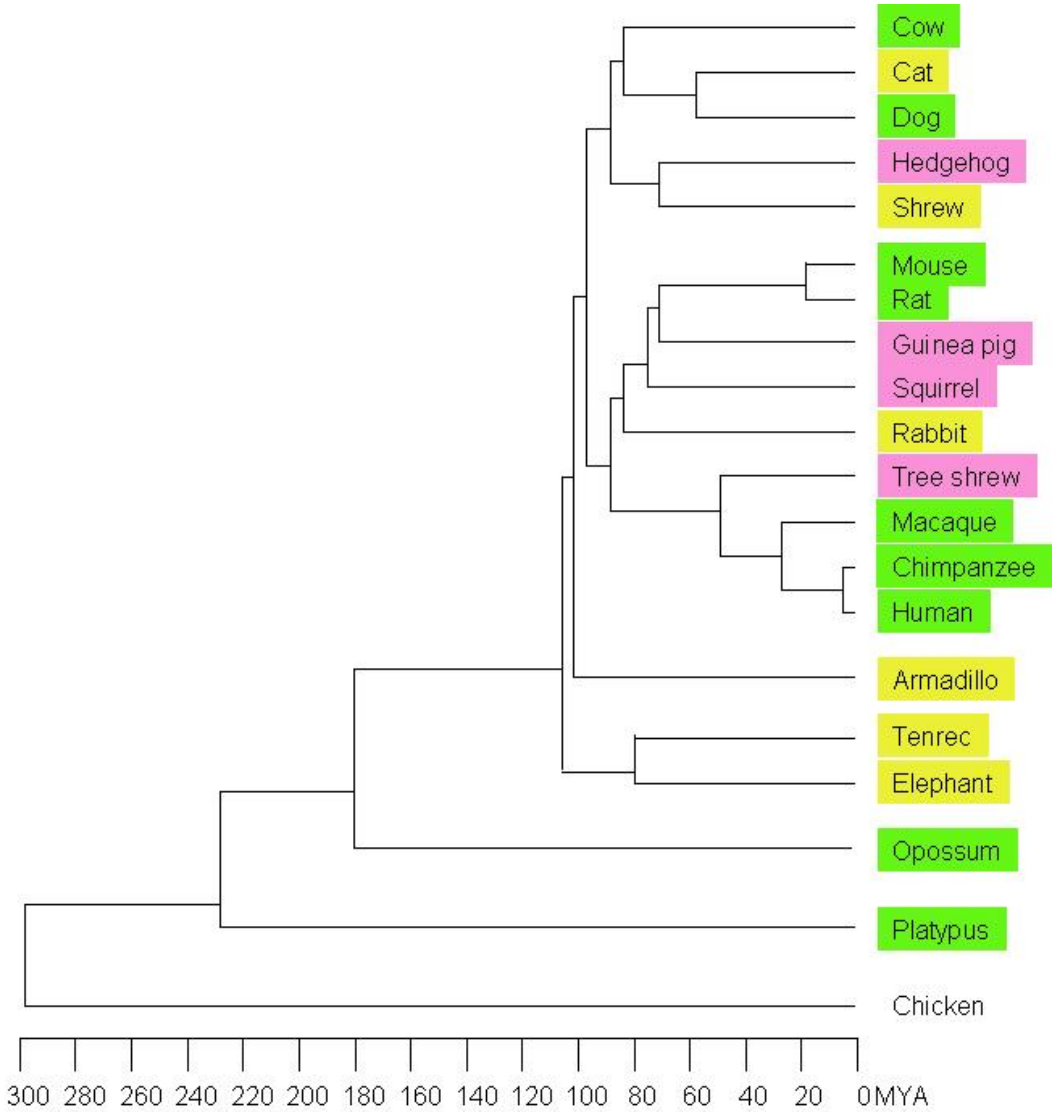

Figure 7.1: Phylogeny and time scale of mammalian species with available sequence information. Chicken is outgroup. Species highlighted in green have completed medium to deep coverage sequences; species in yellow have completed low coverage sequences; and sequencing are in process towards low-coverage for those in pink.

On the other dimension, more in-depth methylation analysis could be carried out on the basis of current findings. As many X-linked genes are not associated with a predicted CGI yet undergo stable XCI, it is of great interest to know whether DNA methylation still plays a role in maintaining their silencing. To this end, oligo-array-based technology could be applied to the entire human X chromosome. It may reveal CpG-rich domains at 5' region of genes with methylation patterns typically associated with inactivated or escapee genes as established in this study. Most recently, bisulphite sequencing has been combined with advanced sequencing technology to yield a single base resolution map of the Arabidopsis methylome (Lister *et al.*, 2008). With further development of technology, similar details may be revealed for the human and other X chromosomes and enable more thorough understanding of the dynamics of DNA methylation in X chromosome inactivation.

7.2 Conclusions

This thesis has shown the potential of DNA methylation analysis to further our understanding of X chromosome inactivation. Described here is the most extensive methylation analysis to date of 5' CGIs associated with genes subject to and escaping from XCI. The generation of fine methylation maps has revealed a great extent of variation in DNA methylation, on the levels of molecules, islands, and species. Comparative study of three species with different extent of escape has provided support to the nature of escape being a lack of maintenance of XCI. It has also provided the first evidence of female-specific methylation on any marsupial X chromosome. This study has formed a basis of more extensive multi-species studies to explore the evolutionary landscape of X chromosome inactivation.

References

Adler DA, Bressler SL, Chapman VM, Page DC and Disteche CM. Inactivation of the Zfx gene on the mouse X chromosome. *Proc Natl Acad Sci U S A*, 88(11):4592–4595, 1991.

Adler DA, Rugarli EI, Lingenfelter PA, Tsuchiya K, Poslinski D *et al.* Evidence of evolutionary up-regulation of the single active X chromosome in mammals based on Clc4 expression levels in *mus spretus* and *mus musculus*. *Proc Natl Acad Sci U S A*, 94(17):9244–9248, 1997.

Adorján P, Distler J, Lipscher E, Model F, Müller J *et al.* Tumour class prediction and discovery by microarray-based DNA methylation analysis. *Nucleic Acids Res*, 30(5):e21, 2002.

Agulnik AI, Mitchell MJ, Mattei MG, Borsani G, Avner PA, Lerner JL and Bishop CE. A novel X gene with a widely transcribed Y-linked homologue escapes X-inactivation in mouse and human. *Hum Mol Genet*, 3(6):879–884, 1994.

Antequera F and Bird A. Number of CpG islands and genes in human and mouse. *Proc Natl Acad Sci U S A*, 90(24):11995–11999, 1993.

Ashworth A, Rastan S, Lovell-Badge R and Kay G. X-chromosome inactivation may explain the difference in viability of XO humans and mice. *Nature*, 351(6325):406–408, 1991.

Bailey JA, Carrel L, Chakravarti A and Eichler EE. Molecular evidence for a relationship between LINE-1 elements and X chromosome inactivation: the Lyon repeat hypothesis. *Proc Natl Acad Sci U S A*, 97(12):6634–6639, 2000.

Baker BS, Gorman M and Marín I. Dosage compensation in Drosophila. *Annu Rev Genet*, 28:491–521, 1994.

Barr ML and Bertram EG. A morphological distinction between neurones of the male and female, and the behaviour of the nucleolar satellite during accelerated nucleoprotein synthesis. *Nature*, 163(4148):676, 1949.

Bell AC, West AG and Felsenfeld G. The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. *Cell*, 98(3):387–396, 1999.

Benjamin D, Bakel IV and Craig IW. A novel expression based approach for assessing the inactivation status of human X-linked genes. *Eur J Hum Genet*, 8(2):103–108, 2000.

Bininda-Emonds ORP, Cardillo M, Jones KE, MacPhee RDE, Beck RMD *et al.* The delayed rise of present-day mammals. *Nature*, 446(7135):507–512, 2007.

Bird A, Taggart M, Frommer M, Miller OJ and Macleod D. A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. *Cell*, 40(1):91–99, 1985.

Bird AP. CpG-rich islands and the function of DNA methylation. *Nature*, 321(6067):209–213, 1986.

Bird AP and Taggart MH. Variable patterns of total DNA and rDNA methylation in animals. *Nucleic Acids Res*, 8(7):1485–1497, 1980.

Boggs BA, Connors B, Sobel RE, Chinault AC and Allis CD. Reduced levels of histone H3 acetylation on the inactive X chromosome in human females. *Chromosoma*, 105(5):303–309, 1996.

Borsani G, Tonlorenzi R, Simmler MC, Dandolo L, Arnaud D *et al.* Characterization of a murine gene expressed from the inactive X chromosome. *Nature*, 351(6324):325–329, 1991.

Brockdorff N, Ashworth A, Kay GF, McCabe VM, Norris DP, Cooper PJ, Swift S and Rastan S. The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. *Cell*, 71(3):515–526, 1992.

Brown CJ, Ballabio A, Rupert JL, Lafreniere RG, Grompe M, Tonlorenzi R and Willard HF. A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. *Nature*, 349(6304):38–44, 1991.

Brown CJ, Hendrich BD, Rupert JL, Lafrenière RG, Xing Y, Lawrence J and Willard HF. The human Xist gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. *Cell*, 71(3):527–542, 1992.

Brown CJ and Willard HF. The human X-inactivation centre is not required for maintenance of X-chromosome inactivation. *Nature*, 368(6467):154–156, 1994.

Bull J. *Evolution of sex determining mechanisms.* Benjamin Cummings, Menlo Park, CA, 1983.

Carrel L, Cottle AA, Goglin KC and Willard HF. A first-generation X-inactivation profile of the human X chromosome. *Proc Natl Acad Sci U S A*, 96(25):14440–14444, 1999.

Carrel L and Willard HF. X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature*, 434(7031):400–404, 2005.

Cattanach BM and Isaacson JH. Controlling elements in the mouse X chromosome. *Genetics*, 57(2):331–346, 1967.

Cattanach BM and Williams CE. Evidence of non-random X chromosome activity in the mouse. *Genet Res*, 19(3):229–240, 1972.

Chadwick LH, Pertz LM, Broman KW, Bartolomei MS and Willard HF. Genetic control of X chromosome inactivation in mice: definition of the Xce candidate interval. *Genetics*, 173(4):2103–2110, 2006.

Clemson CM, McNeil JA, Willard HF and Lawrence JB. XIST RNA paints the inactive X chromosome at interphase: evidence for a novel RNA involved in nuclear/chromosome structure. *J Cell Biol*, 132(3):259–275, 1996.

Clerc P and Avner P. Role of the region 3' to Xist exon 6 in the counting process of X-chromosome inactivation. *Nat Genet*, 19(3):249–253, 1998.

Cline TW and Meyer BJ. Vive la différence: males vs females in flies vs worms. *Annu Rev Genet*, 30:637–702, 1996.

Comb M and Goodman HM. CpG methylation inhibits proenkephalin gene expression and binding of the transcription factor AP-2. *Nucleic Acids Res*, 18(13):3975–3982, 1990.

Cooper DN, Taggart MH and Bird AP. Unmethylated domains in vertebrate DNA. *Nucleic Acids Res*, 11(3):647–658, 1983.

Cooper DW, Edwards C, James E, Sharman GB, VandeBerg JL and Graves JAM. Studies on metathcrian sex chromosome. VI. A third state of an X-linked gene: partial activity for the paternally derived *pgk-a* allele in cultured fibroblasts of *macropus giganteus* and *m. pangi*. *Aust J Biol Sci*, 30:431–433, 1977.

Cooper DW, Johnston PG, Watson JM and Graves JAM. X-inactivation in marsupials and monotremes. *Semin Dev Biol*, 4:117–128, 1993.

Cooper DW, VandeBerg JL, Sharman GB and Poole WE. Phosphoglycerate kinase polymorphism in kangaroos provides further evidence for paternal X inactivation. *Nat New Biol*, 230(13):155–157, 1971.

Costanzi C and Pehrson JR. Histone macroH2A1 is concentrated in the inactive X chromosome of female mammals. *Nature*, 393(6685):599–601, 1998.

Costello JF, Frühwald MC, Smiraglia DJ, Rush LJ, Robertson GP *et al.* Aberrant CpG-island methylation has non-random and tumour-type-specific patterns. *Nat Genet*, 24(2):132–138, 2000.

Coulondre C, Miller JH, Farabaugh PJ and Gilbert W. Molecular basis of base substitution hotspots in *escherichia coli*. *Nature*, 274(5673):775–780, 1978.

Csankovszki G, Nagy A and Jaenisch R. Synergism of Xist RNA, DNA methylation, and histone hypoacetylation in maintaining X chromosome inactivation. *J Cell Biol*, 153(4):773–784, 2001.

Csankovszki G, Panning B, Bates B, Pehrson JR and Jaenisch R. Conditional deletion of Xist disrupts histone macroH2A localization but not maintenance of X inactivation. *Nat Genet*, 22(4):323–324, 1999.

Davidson RG, Nitowsky HM and Childs B. Demonstration of two populations of cells in the human female heterozygous for glucose-6-phosphate dehydrogenase variants. *Proc Natl Acad Sci U S A*, 50:481–485, 1963.

Disteche CM. Escape from X inactivation in human and mouse. *Trends Genet*, 11(1):17–22, 1995.

Disteche CM, Filippova GN and Tsuchiya KD. Escape from X inactivation. *Cytogenet Genome Res*, 99(1-4):36–43, 2002.

Duret L, Eyre-Walker A and Galtier N. A new perspective on isochore evolution. *Gene*, 385:71–74, 2006.

Eckhardt F, Lewin J, Cortese R, Rakyan VK, Attwood J *et al.* DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat Genet*, 38(12):1378–1385, 2006.

Ehrlich M, Gama-Sosa MA, Huang LH, Midgett RM, Kuo KC, McCune RA and Gehrke C. Amount and distribution of 5-methylcytosine in human DNA from different types of tissues of cells. *Nucleic Acids Res*, 10(8):2709–2721, 1982.

Ehrmann IE, Ellis PS, Mazeyrat S, Duthie S, Brockdorff N *et al.* Characterization of genes encoding translation initiation factor eif-2gamma in mouse and human: sex chromosome localization, escape from X-inactivation and evolution. *Hum Mol Genet*, 7(11):1725–1737, 1998.

Ellison J, Passage M, Yu LC, Yen P, Mohandas TK and Shapiro L. Directed isolation of human genes that escape X inactivation. *Somat Cell Mol Genet*, 18(3):259–268, 1992.

Esteller M, Corn PG, Baylin SB and Herman JG. A gene hypermethylation profile of human cancer. *Cancer Res*, 61(8):3225–3229, 2001.

Ferguson-Smith AC, Sasaki H, Cattanach BM and Surani MA. Parental-origin-specific epigenetic modification of the mouse H19 gene. *Nature*, 362(6422):751–755, 1993.

Fialkow PJ. X-chromosome inactivation and the *xg* locus. *Am J Hum Genet*, 22(4):460–463, 1970.

Filippova GN, Cheng MK, Moore JM, Truong JP, Hu YJ, Nguyen DK, Tsuchiya KD and Disteche CM. Boundaries between chromosomal domains of X inactivation and escape bind CTCF and lack CpG methylation during early development. *Dev Cell*, 8(1):31–42, 2005.

Fisher EM, Beer-Romero P, Brown LG, Ridley A, McNeil JA *et al.* Homologous ribosomal protein genes on the human X and Y chromosomes: escape from X inactivation and possible implications for Turner syndrome. *Cell*, 63(6):1205–1218, 1990.

Foster JW and Graves JA. An *sry*-related sequence on the marsupial X chromosome: implications for the evolution of the mammalian testis-determining gene. *Proc Natl Acad Sci U S A*, 91(5):1927–1931, 1994.

Frazer KA, Pachter L, Poliakov A, Rubin EM and Dubchak I. VISTA: computational tools for comparative genomics. *Nucleic Acids Res*, 32(Web Server issue):W273–W279, 2004.

Frigola J, Ribas M, Risques RA and Peinado MA. Methylome profiling of cancer cells by amplification of inter-methylated sites (AIMS). *Nucleic Acids Res*, 30(7):e28, 2002.

Frommer M, McDonald LE, Millar DS, Collis CM, Watt F, Grigg GW, Molloy PL and Paul CL. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc Natl Acad Sci U S A*, 89(5):1827–1831, 1992.

Gardiner-Garden M and Frommer M. CpG islands in vertebrate genomes. *J Mol Biol*, 196(2):261–282, 1987.

Gartler SM and Riggs AD. Mammalian X-chromosome inactivation. *Annu Rev Genet*, 17:155–190, 1983.

Genereux DP, Miner BE, Bergstrom CT and Laird CD. A population-epigenetic model to infer site-specific methylation rates from double-stranded DNA methylation patterns. *Proc Natl Acad Sci U S A*, 102(16):5802–5807, 2005.

Gitan RS, Shi H, Chen CM, Yan PS and Huang THM. Methylation-specific oligonucleotide microarray: a new potential for high-throughput methylation analysis. *Genome Res*, 12(1):158–164, 2002.

Gonzalgo ML and Jones PA. Rapid quantitation of methylation differences at specific sites using methylation-sensitive single nucleotide primer extension (Ms-SNuPE). *Nucleic Acids Res*, 25(12):2529–2531, 1997.

Gonzalgo ML, Liang G, Spruck CH, Zingg JM, Rideout WM and Jones PA. Identification and characterization of differentially methylated regions of genomic DNA by methylation-sensitive arbitrarily primed pcr. *Cancer Res*, 57(4):594–599, 1997.

Goodfellow P, Pym B, Mohandas T and Shapiro LJ. The cell surface antigen locus, *mic2x*, escapes X-inactivation. *Am J Hum Genet*, 36(4):777–782, 1984.

Gorman JG, di Re J, Treacy AM and Cahan A. The application of -Xga antiserum to the question of red cell mosaicism in female heterozygotes. *J Lab Clin Med*, 61:642–649, 1963.

Goto Y, Gomez M, Brockdorff N and Feil R. Differential patterns of histone methylation and acetylation distinguish active and repressed alleles at X-linked genes. *Cytogenet Genome Res*, 99(1-4):66–74, 2002.

Graves JA. 5-azacytidine-induced re-expression of alleles on the inactive X chromosome in a hybrid mouse cell line. *Exp Cell Res*, 141(1):99–105, 1982.

Graves JA. The evolution of mammalian sex chromosomes and the origin of sex determining genes. *Philos Trans R Soc Lond B Biol Sci*, 350(1333):305–11; discussion 311–2, 1995.

Graves JA. Mammals that break the rules: genetics of marsupials and monotremes. *Annu Rev Genet*, 30:233–260, 1996.

Graves JA and Watson JM. Mammalian sex chromosomes: evolution of organization and function. *Chromosoma*, 101(2):63–68, 1991.

Graves JAM. Sex chromosome specialization and degeneration in mammals. *Cell*, 124(5):901–914, 2006.

Greenfield A, Carrel L, Pennisi D, Philippe C, Quaderi N *et al.* The *utx* gene escapes X inactivation in mice and humans. *Hum Mol Genet*, 7(4):737–742, 1998.

Greger V, Passarge E, Höpping W, Messmer E and Horsthemke B. Epigenetic changes may contribute to the formation and spontaneous regression of retinoblastoma. *Hum Genet*, 83(2):155–158, 1989.

Grunau C, Schattevoy R, Mache N and Rosenthal A. MethTools–a toolbox to visualize and analyze DNA methylation data. *Nucleic Acids Res*, 28(5):1053–1058, 2000.

Grunstein M. Histone acetylation in chromatin structure and transcription. *Nature*, 389(6649):349–352, 1997.

Heard E, Rougeulle C, Arnaud D, Avner P, Allis CD and Spector DL. Methylation of histone H3 at Lys-9 is an early mark on the X chromosome during X inactivation. *Cell*, 107(6):727–738, 2001.

Hedges SB and Kumar S. Precision of molecular time estimates. *Trends Genet*, 20(5):242–247, 2004.

Hendriks RW, Chen ZY, Hinds H, Schuurman RK and Craig IW. An X chromosome inactivation assay based on differential methylation of a CpG island coupled to a VNTR polymorphism at the 5' end of the monoamine oxidase A gene. *Hum Mol Genet*, 1(3):187–194, 1992.

Herman JG, Graff JR, Myöhänen S, Nelkin BD and Baylin SB. Methylation-specific PCR: a novel PCR assay for methylation status of CpG islands. *Proc Natl Acad Sci U S A*, 93(18):9821–9826, 1996.

Herman JG, Latif F, Weng Y, Lerman MI, Zbar B *et al.* Silencing of the VHL tumor-suppressor gene by DNA methylation in renal carcinoma. *Proc Natl Acad Sci U S A*, 91(21):9700–9704, 1994.

Herzing LB, Romer JT, Horn JM and Ashworth A. Xist has properties of the X-chromosome inactivation centre. *Nature*, 386(6622):272–275, 1997.

Hore TA, Koina E, Wakefield MJ and Graves JAM. The region homologous to the X-chromosome inactivation centre has been disrupted in marsupial and monotreme mammals. *Chromosome Res*, 15(2):147–161, 2007.

Hornecker JL, Samollow PB, Robinson ES, Vandeberg JL and McCarrey JR. Meiotic sex chromosome inactivation in the marsupial *monodelphis domestica*. *Genesis*, 45(11):696–708, 2007.

Huang TH, Laux DE, Hamlin BC, Tran P, Tran H and Lubahn DB. Identification of DNA methylation markers for human breast carcinomas using the methylation-sensitive restriction fingerprinting technique. *Cancer Res*, 57(6):1030–1034, 1997.

Huang TH, Perry MR and Laux DE. Methylation profiling of CpG islands in human breast cancer cells. *Hum Mol Genet*, 8(3):459–470, 1999.

Huynh KD and Lee JT. X-chromosome inactivation: a hypothesis linking ontogeny and phylogeny. *Nat Rev Genet*, 6(5):410–418, 2005.

Jegalian K and Page DC. A proposed path by which genes common to mammalian X and Y chromosomes evolve to become X inactivated. *Nature*, 394(6695):776–780, 1998.

Jeppesen P and Turner BM. The inactive X chromosome in female mammals is distinguished by a lack of histone H4 acetylation, a cytogenetic marker for gene expression. *Cell*, 74(2):281–289, 1993.

Johnston PG, Sharman GB, James EA and Cooper DW. Studies on metatherian sex chromosomes. VII. Glucose-6-phosphate dehydrogenase expression in tissues and cultured fibroblasts of kangaroos. *Aust J Biol Sci*, 31:415–424, 1978.

Jones PA, Taylor SM, Mohandas T and Shapiro LJ. Cell cycle-specific reactivation of an inactive X-chromosome locus by 5-azadeoxycytidine. *Proc Natl Acad Sci U S A*, 79(4):1215–1219, 1982.

Kaslow DC and Migeon BR. DNA methylation stabilizes X chromosome inactivation in eutherians but not in marsupials: evidence for multistep maintenance of mammalian X dosage compensation. *Proc Natl Acad Sci U S A*, 84(17):6210–6214, 1987.

Kawakami T, Okamoto K, Ogawa O and Okada Y. XIST unmethylated DNA fragments in male-derived plasma as a tumour marker for testicular cancer. *Lancet*, 363(9402):40–42, 2004.

Kay GF, Ashworth A, Penny GD, Dunlop M, Swift S, Brockdorff N and Rastan S. A candidate spermatogenesis gene on the mouse Y chromosome is homologous to ubiquitin-activating enzyme E1. *Nature*, 354(6353):486–489, 1991.

Keohane AM, O'neill LP, Belyaev ND, Lavender JS and Turner BM. X-inactivation and histone H4 acetylation in embryonic stem cells. *Dev Biol*, 180(2):618–630, 1996.

Keshet I, Yisraeli J and Cedar H. Effect of regional DNA methylation on gene expression. *Proc Natl Acad Sci U S A*, 82(9):2560–2564, 1985.

Labarga A, Valentin F, Anderson M and Lopez R. Web services at the European bioinformatics institute. *Nucleic Acids Res*, 35(Web Server issue):W6–11, 2007.

Lahn BT and Page DC. Four evolutionary strata on the human X chromosome. *Science*, 286(5441):964–967, 1999.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC *et al.* Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics*, 23(21):2947–2948, 2007.

Larsen F, Gundersen G, Lopez R and Prydz H. CpG islands as gene markers in the human genome. *Genomics*, 13(4):1095–1107, 1992.

Lee JT and Jaenisch R. Long-range *cis* effects of ectopic X-inactivation centres on a mouse autosome. *Nature*, 386(6622):275–279, 1997.

Lee JT and Lu N. Targeted mutagenesis of Tsix leads to nonrandom X inactivation. *Cell*, 99(1):47–57, 1999.

Lewis JA. Structure and expression of the Chinese hamster thymidine kinase gene. *Mol Cell Biol*, 6(6):1998–2010, 1986.

Lin H, Gupta V, Vermilyea MD, Falciani F, Lee JT, O'Neill LP and Turner BM. Dosage compensation in the mouse balances up-regulation and silencing of X-linked genes. *PLoS Biol*, 5(12):e326, 2007.

Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB *et al.* Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature*, 438(7069):803–819, 2005.

Lingenfelter PA, Adler DA, Poslinski D, Thomas S, Elliott RW, Chapman VM and Disteche CM. Escape from X inactivation of *smcx* is preceded by silencing during mouse development. *Nat Genet*, 18(3):212–213, 1998.

Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH and Ecker JR. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell*, 133(3):523–536, 2008.

Lobanenkov VV, Nicolas RH, Adler VV, Paterson H, Klenova EM, Polotskaja AV and Goodwin GH. A novel sequence-specific DNA binding protein which interacts with three regularly spaced direct repeats of the CCCTC-motif in the 5'-flanking sequence of the chicken *c-myc* gene. *Oncogene*, 5(12):1743–1753, 1990.

Lock LF, Melton DW, Caskey CT and Martin GR. Methylation of the mouse *hprt* gene differs on the active and inactive X chromosomes. *Mol Cell Biol*, 6(3):914–924, 1986.

Lock LF, Takagi N and Martin GR. Methylation of the *hprt* gene on the inactive X occurs after chromosome inactivation. *Cell*, 48(1):39–46, 1987.

Loebel DA and Johnston PG. Analysis of DNase 1 sensitivity and methylation of active and inactive X chromosomes of kangaroos (*macropus robustus*) by *in situ* nick translation. *Chromosoma*, 102(2):81–87, 1993.

Loebel DA and Johnston PG. Methylation analysis of a marsupial X-linked CpG island by bisulfite genomic sequencing. *Genome Res*, 6(2):114–123, 1996.

Lyon MF. Gene action in the X-chromosome of the mouse (*mus musculus l.*). *Nature*, 190:372–373, 1961.

Lyon MF. X-chromosome inactivation: a repeat hypothesis. *Cytogenet Cell Genet*, 80(1-4):133–137, 1998.

Lyon MF, Searle AG, Ford CE and Ohno S. A mouse translocation suppressing sex-linked variegation. *Cytogenetics*, 15:306–323, 1964.

McKeon C, Ohkubo H, Pastan I and de Crombrugghe B. Unusual methylation pattern of the alpha 2 (l) collagen gene. *Cell*, 29(1):203–210, 1982.

Merlo A, Herman JG, Mao L, Lee DJ, Gabrielson E, Burger PC, Baylin SB and Sidransky D. 5' CpG island methylation is associated with transcriptional silencing of the tumour suppressor *p16/cdkn2/mts1* in human cancers. *Nat Med*, 1(7):686–692, 1995.

Mermoud JE, Costanzi C, Pehrson JR and Brockdorff N. Histone macroH2A1.2 relocates to the inactive X chromosome after initiation and propagation of X-inactivation. *J Cell Biol*, 147(7):1399–1408, 1999.

Meyer BJ. Sex in the wormcounting and compensating X-chromosome dose. *Trends Genet*, 16(6):247–253, 2000.

Migeon BR, de Beur SJ and Axelman J. Frequent derepression of *g6pd* and *hprt* on the marsupial inactive X chromosome associated with cell proliferation *in vitro*. *Exp Cell Res*, 182(2):597–609, 1989.

Migeon BR, Chowdhury AK, Dunston JA and McIntosh I. Identification of TSIX, encoding an RNA antisense to human XIST, reveals differences from its murine counterpart: implications for X inactivation. *Am J Hum Genet*, 69(5):951–960, 2001.

Migeon BR, Lee CH, Chowdhury AK and Carpenter H. Species differences in TSIX/Tsix reveal the roles of these genes in X-chromosome inactivation. *Am J Hum Genet*, 71(2):286–293, 2002.

Mikkelsen TS, Wakefield MJ, Aken B, Amemiya CT, Chang JL *et al.* Genome of the marsupial *monodelphis domestica* reveals innovation in noncoding sequences. *Nature*, 447(7141):167–177, 2007.

Mitchell MJ, Woods DR, Wilcox SA, Graves JA and Bishop CE. Marsupial Y chromosome encodes a homologue of the mouse Y-linked candidate spermatogenesis gene *ube1y*. *Nature*, 359(6395):528–531, 1992.

Mohandas T, Sparkes RS, Hellkuhl B, Grzeschik KH and Shapiro LJ. Expression of an X-linked gene from an inactive human X chromosome in mouse-human hybrid cells: further evidence for the noninactivation of the steroid sulfatase locus in man. *Proc Natl Acad Sci U S A*, 77(11):6759–6763, 1980.

Morey C, Arnaud D, Avner P and Clerc P. Tsix-mediated repression of Xist accumulation is not sufficient for normal random X inactivation. *Hum Mol Genet*, 10(13):1403–1411, 2001.

Mukhopadhyay R, Yu W, Whitehead J, Xu J, Lezcano M *et al.* The binding sites for the chromatin insulator protein CTCF map to DNA methylation-free domains genome-wide. *Genome Res*, 14(8):1594–1602, 2004.

Murphy WJ, Pevzner PA and O'Brien SJ. Mammalian phylogenomics comes of age. *Trends Genet*, 20(12):631–639, 2004.

Namekawa SH, VandeBerg JL, McCarrey JR and Lee JT. Sex chromosome silencing in the marsupial male germ line. *Proc Natl Acad Sci U S A*, 104(23):9730–9735, 2007.

Nguyen DK and Disteche CM. Dosage compensation of the active X chromosome in mammals. *Nat Genet*, 38(1):47–53, 2006.

Nordquist N and Oreland L. Monoallelic expression of *maoa* in skin fibroblasts. *Biochem Biophys Res Commun*, 348(2):763–767, 2006.

Ogawa Y and Lee JT. Xite, X-inactivation intergenic transcription elements that regulate the probability of choice. *Mol Cell*, 11(3):731–743, 2003.

Ohno S. *Sex chromosomes and sex-linked genes.* Springer-Verlag Berlin, 1967.

Ohno S and Hauschka TS. Allocycly of the X-chromosome in tumors and normal tissues. *Cancer Res*, 20:541–545, 1960.

Ohno S, Kaplan WD and Kinosita R. Formation of the sex chromatin by a single X-chromosome in liver cells of *rattus norvegicus*. *Exp Cell Res*, 18:415–418, 1959.

Olek A, Oswald J and Walter J. A modified and improved method for bisulphite based cytosine methylation analysis. *Nucleic Acids Res*, 24(24):5064–5066, 1996.

Park JG and Chapman VM. CpG island promoter region methylation patterns of the inactive-X-chromosome hypoxanthine phosphoribosyltransferase (*hprt*) gene. *Mol Cell Biol*, 14(12):7975–7983, 1994.

Penny GD, Kay GF, Sheardown SA, Rastan S and Brockdorff N. Requirement for Xist in X chromosome inactivation. *Nature*, 379(6561):131–137, 1996.

Plath K, Mlynarczyk-Evans S, Nusinow DA and Panning B. Xist RNA and the mechanism of X chromosome inactivation. *Annu Rev Genet*, 36:233–278, 2002.

Rastan S. Non-random X-chromosome inactivation in mouse X-autosome translocation embryos–location of the inactivation centre. *J Embryol Exp Morphol*, 78:1–22, 1983.

Rastan S and Robertson EJ. X-chromosome deletions in embryo-derived (EK) cell lines associated with lack of X-chromosome inactivation. *J Embryol Exp Morphol*, 90:379–388, 1985.

Rens W, Grützner F, O'brien PCM, Fairclough H, Graves JAM and Ferguson-Smith MA. Resolution and evolution of the duck-billed platypus karyotype with an X1Y1X2Y2X3Y3X4Y4X5Y5 male sex chromosome constitution. *Proc Natl Acad Sci U S A*, 101(46):16257–16261, 2004.

Richardson BJ, Czuppon AB and Sharman GB. Inheritance of glucose-6-phosphate dehydrogenase variation in kangaroos. *Nat New Biol*, 230(13):154–155, 1971.

Riggs AD. X inactivation, differentiation, and DNA methylation. *Cytogenet Cell Genet*, 14(1):9–25, 1975.

Ross MT, Grafham DV, Coffey AJ, Scherer S, McLay K *et al.* The DNA sequence of the human X chromosome. *Nature*, 434(7031):325–337, 2005.

Russell GJ, Walker PM, Elton RA and Subak-Sharpe JH. Doublet frequency analysis of fractionated vertebrate nuclear DNA. *J Mol Biol*, 108(1):1–23, 1976.

Russell L and Montgomery C. The use of X-autosome translocations in locating the X-chromosome inactivation center. *Genetics*, 52:470–471, 1965.

Russell LB. Genetics of mammalian sex chromosomes. *Science*, 133:1795–1803, 1961.

Russell LB. Mammalian X-chromosome action: inactivation limited in spread and region of origin. *Science*, 140:976–978, 1963.

Sakai T, Toguchida J, Ohtani N, Yandell DW, Rapaport JM and Dryja TP. Allele-specific hypermethylation of the retinoblastoma tumor-suppressor gene. *Am J Hum Genet*, 48(5):880–888, 1991.

Samollow PB. Status and applications of genomic resources for the gray, short-tailed opossum, monodelphis domestica, an american marsupial model for comparative biology. *Australian Journal of Zoology*, 54:173–196, 2006.

Samollow PB, Ford AL and VandeBerg JL. X-linked gene expression in the Virginia opossum: differences between the paternally derived *gpd* and *pgk-a* loci. *Genetics*, 115(1):185–195, 1987.

Samollow PB, Johnston PG, Ford AL and VandeBerg JL. X-linked gene expression in metatherian fibroblasts: evidence from the *gpd* and *pgk-a* loci of the virginia opossum and the red-necked wallaby. *Biochem Genet*, 27(5-6):313–320, 1989.

Schiebel K, Weiss B, Wöhrle D and Rappold G. A human pseudoautosomal gene, ADP/ATP translocase, escapes X-inactivation whereas a homologue on Xq is subject to X-inactivation. *Nat Genet*, 3(1):82–87, 1993.

Schmid CW. Does SINE evolution preclude Alu function? *Nucleic Acids Res*, 26(20):4541–4550, 1998.

Schneider-Gädicke A, Beer-Romero P, Brown LG, Nussbaum R and Page DC. *ZFX* has a gene structure similar to *zfy*, the putative human sex determinant, and escapes X inactivation. *Cell*, 57(7):1247–1258, 1989.

Schumacher A, Kapranov P, Kaminsky Z, Flanagan J, Assadzadeh A *et al.* Microarray-based DNA methylation profiling: technology and applications. *Nucleic Acids Res*, 34(2):528–542, 2006.

Shapiro LJ, Mohandas T, Weiss R and Romeo G. Non-inactivation of an X-chromosome locus in man. *Science*, 204(4398):1224–1226, 1979.

Sharman GB. Late DNA replication in the paternally derived X chromosome of female kangaroos. *Nature*, 230(5291):231–232, 1971.

Simmler MC, Cattanach BM, Rasberry C, Rougeulle C and Avner P. Mapping the murine Xce locus with (CA)n repeats. *Mamm Genome*, 4(9):523–530, 1993.

Singer-Sam J, Grant M, LeBon JM, Okuyama K, Chapman V, Monk M and Riggs AD. Use of a HpaII-polymerase chain reaction assay to study DNA methylation in the *pgk-1* CpG island of mouse embryos at the time of X-chromosome inactivation. *Mol Cell Biol*, 10(9):4987–4989, 1990a.

Singer-Sam J, LeBon JM, Tanguay RL and Riggs AD. A quantitative HpaII-PCR assay to measure methylation of DNA from a small number of cells. *Nucleic Acids Res*, 18(3):687, 1990b.

Smits G, Mungall AJ, Griffiths-Jones S, Smith P, Beury D *et al.* Conservation of the H19 noncoding RNA and H19-IGF2 imprinting mechanism in therians. *Nat Genet*, 40(8):971–976, 2008.

Spencer JA, Sinclair AH, Watson JM and Graves JA. Genes on the short arm of the human X chromosome are not shared with the marsupial X. *Genomics*, 11(2):339–345, 1991a.

Spencer JA, Watson JM and Graves JA. The X chromosome of marsupials shares a highly conserved region with eutherians. *Genomics*, 9(4):598–604, 1991b.

Stein R, Sciaky-Gallili N, Razin A and Cedar H. Pattern of methylation of two genes coding for housekeeping functions. *Proc Natl Acad Sci U S A*, 80(9):2422–2426, 1983.

Stöger R, Kajimura TM, Brown WT and Laird CD. Epigenetic variation illustrated by DNA methylation patterns of the fragile-X gene *fmr1*. *Hum Mol Genet*, 6(11):1791–1801, 1997.

Stöger R, Kubicka P, Liu CG, Kafri T, Razin A, Cedar H and Barlow DP. Maternal-specific methylation of the imprinted mouse *igf2r* locus identifies the expressed locus as carrying the imprinting signal. *Cell*, 73(1):61–71, 1993.

Stirzaker C, Millar DS, Paul CL, Warnecke PM, Harrison J, Vincent PC, Frommer M and Clark SJ. Extensive DNA methylation spanning the *rb* promoter in retinoblastoma tumors. *Cancer Res*, 57(11):2229–2237, 1997.

Suzuki S, Ono R, Narita T, Pask AJ, Shaw G *et al.* Retrotransposon silencing by DNA methylation can drive mammalian genomic imprinting. *PLoS Genet*, 3(4):e55, 2007.

Swartz MN, Trautner TA and Kornberg A. Enzymatic synthesis of deoxyribonucleic acid. XI. Further studies on nearest neighbor base sequences in deoxyribonucleic acids. *J Biol Chem*, 237:1961–1967, 1962.

Takagi N and Sasaki M. Preferential inactivation of the paternally derived X chromosome in the extraembryonic membranes of the mouse. *Nature*, 256(5519):640–642, 1975.

Takai D and Jones PA. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc Natl Acad Sci U S A*, 99(6):3740–3745, 2002.

Tremblay KD, Saam JR, Ingram RS, Tilghman SM and Bartolomei MS. A paternal-specific methylation imprint marks the alleles of the mouse *h19* gene. *Nat Genet*, 9(4):407–413, 1995.

Tribioli C, Tamanini F, Patrosso C, Milanesi L, Villa A *et al.* Methylation and sequence analysis around Eagi sites: identification of 28 new CpG islands in Xq24-Xq28. *Nucleic Acids Res*, 20(4):727–733, 1992.

Tsuchiya KD, Greally JM, Yi Y, Noel KP, Truong JP and Disteche CM. Comparative sequence and X-inactivation analyses of a domain of escape in human Xp11.2 and the conserved segment in mouse. *Genome Res*, 14(7):1275–1284, 2004.

Tykocinski ML and Max EE. CG dinucleotide clusters in *mhc* genes and in 5' demethylated genes. *Nucleic Acids Res*, 12(10):4385–4396, 1984.

VandeBerg J. The gray short-tailed opossum (monodelphis-domestica) as a model didelphid species for genetic research. *Australian Journal of Zoology*, 37:235–247, 1989.

Veyrunes F, Waters PD, Miethke P, Rens W, McMillan D *et al.* Bird-like sex chromosomes of platypus imply recent origin of mammal sex chromosomes. *Genome Res*, 18(6):965–973, 2008.

Wakefield MJ, Keohane AM, Turner BM and Graves JA. Histone underacetylation is an ancient component of mammalian X chromosome inactivation. *Proc Natl Acad Sci U S A*, 94(18):9665–9668, 1997.

Wallis MC, Waters PD, Delbridge ML, Kirby PJ, Pask AJ *et al.* Sex determination in platypus and echidna: autosomal location of *sox3* confirms the absence of *sry* from monotremes. *Chromosome Res*, 15(8):949–959, 2007.

Warnecke PM, Stirzaker C, Melki JR, Millar DS, Paul CL and Clark SJ. Detection and measurement of pcr bias in quantitative methylation analysis of bisulphite-treated DNA. *Nucleic Acids Res*, 25(21):4422–4426, 1997.

Warren WC, Hillier LW, Graves JAM, Birney E, Ponting CP *et al.* Genome analysis of the platypus reveals unique signatures of evolution. *Nature*, 453(7192):175–183, 2008.

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915):520–562, 2002.

Weber M, Davies JJ, Wittig D, Oakeley EJ, Haase M, Lam WL and Schübeler D. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat Genet*, 37(8):853–862, 2005.

Welshons WJ and Russell LB. The y-chromosome as the bearer of male determining factors in the mouse. *Proc Natl Acad Sci U S A*, 45(4):560–566, 1959.

Wilson DE and Reeder DM, editors. *Mammal Species of the World: A Taxonomic and Geographic Reference*. JHU Press, 2005.

Wolf SF, Jolly DJ, Lunnen KD, Friedmann T and Migeon BR. Methylation of the hypoxanthine phosphoribosyltransferase locus on the human X chromosome: implications for X-chromosome inactivation. *Proc Natl Acad Sci U S A*, 81(9):2806–2810, 1984.

Woodburne MO, Rich TH and Springer MS. The evolution of tribospheny and the antiquity of mammalian clades. *Mol Phylogenet Evol*, 28(2):360–385, 2003.

Wu J, Salido EC, Yen PH, Mohandas TK, Heng HH *et al.* The murine *xe169* gene escapes X-inactivation like its human homologue. *Nat Genet*, 7(4):491–496, 1994.

Xue F, Tian XC, Du F, Kubota C, Taneja M *et al.* Aberrant patterns of X chromosome inactivation in bovine clones. *Nat Genet*, 31(2):216–220, 2002.

Yen PH, Patel P, Chinault AC, Mohandas T and Shapiro LJ. Differential methylation of hypoxanthine phosphoribosyltransferase genes on active and inactive human X chromosomes. *Proc Natl Acad Sci U S A*, 81(6):1759–1763, 1984.

Yen ZC, Meyer IM, Karalic S and Brown CJ. A cross-species comparison of X-chromosome inactivation in Eutheria. *Genomics*, 90(4):453–463, 2007.