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.

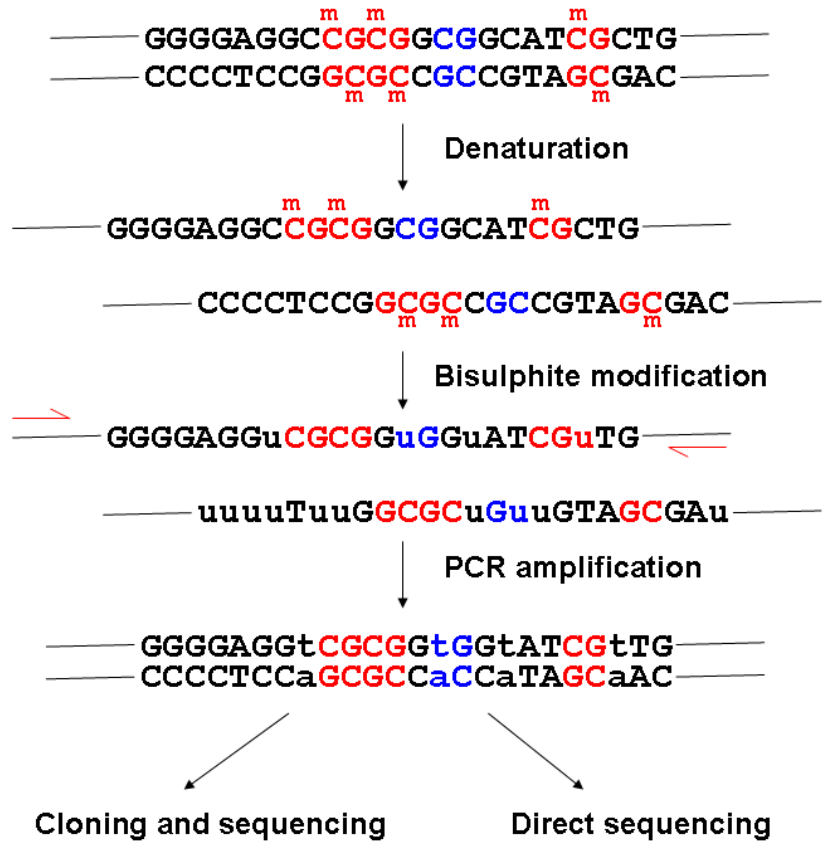

Figure 4.1: Bisulphite sequencing. Note the two strands no longer complement after bisulphite modification so different primers are needed to amplify the top and bottom strands. Shown here is the amplification and sequencing of top strand only.

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).

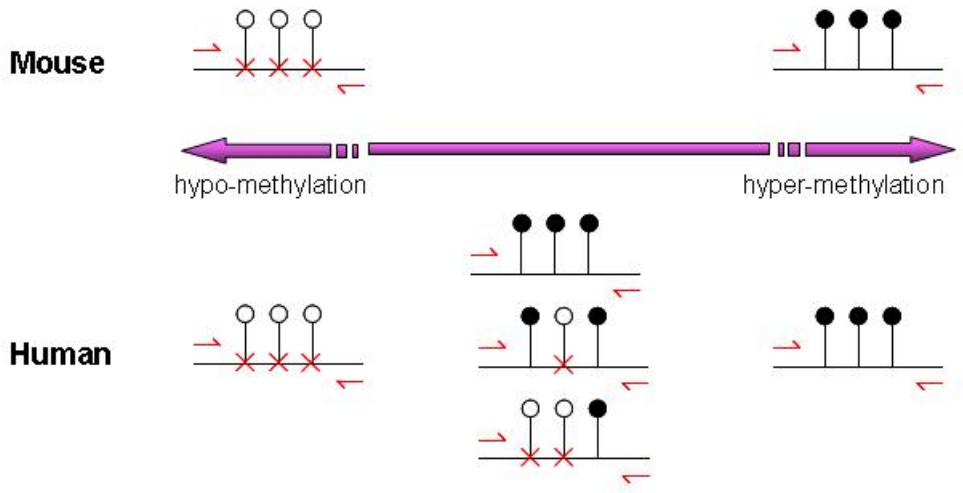

Figure 4.2: Hypothesised model of CGI methylation patterns on the human and mouse X chromosomes. Based on RPMA results, mouse CGIs are either hypomethylated or hypermethylated, but a proportion of human CGIs are intermediately methylated.

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			
	RPMA	Lit.	CCGG sites	Female Hi Hp Ms		Male Hi Hp Ms	CCGG sites	Female Hi Hp Ms		Male Hi Hp Ms		RPMA	Lit.	CCGG sites	Female Hi Hp Ms		Male Hi Hp Ms	CCGG sites	Female Hi Hp Ms		Male 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^{TM} (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.

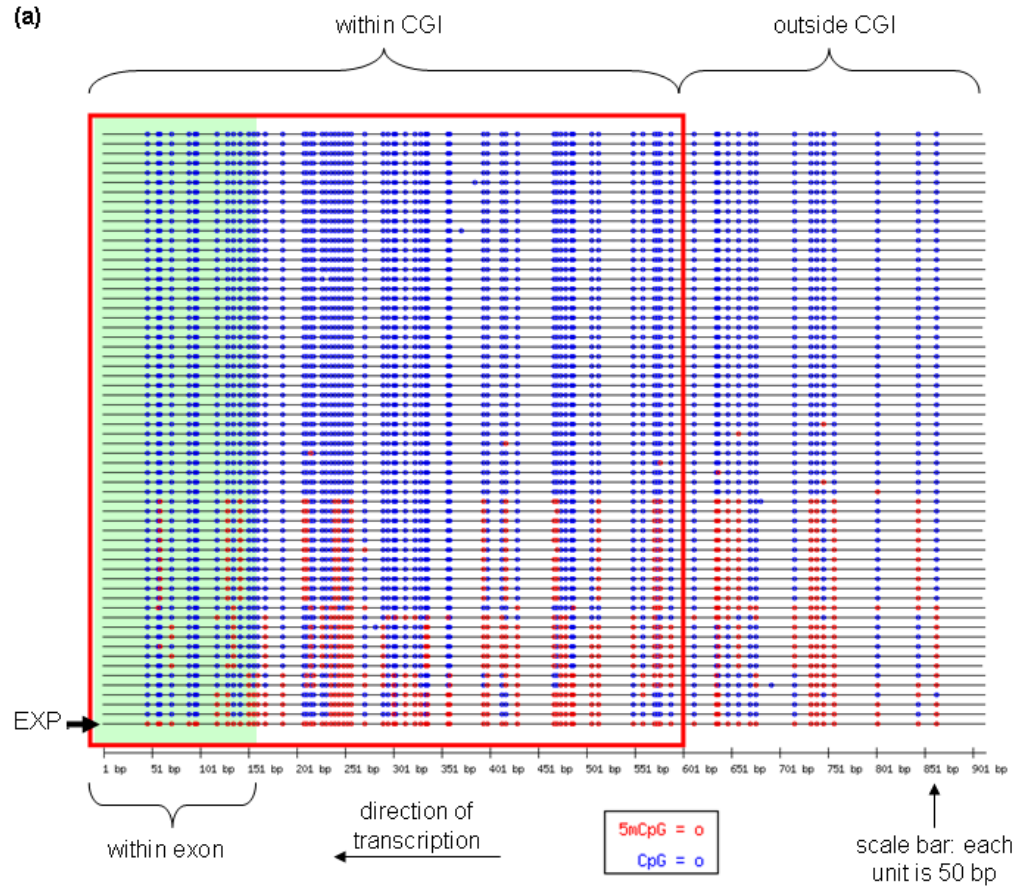

Figure 4.3: Graphical representation of CGI methylation. (a) shows CGI methylation pattern of a gene inactivated by XCI in female. Each line represents an individual molecule with unmethylated (blue) and methylated (red) CpG dinucleotides (circles). The bottom line ('EXP') represents a reference sequence 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.

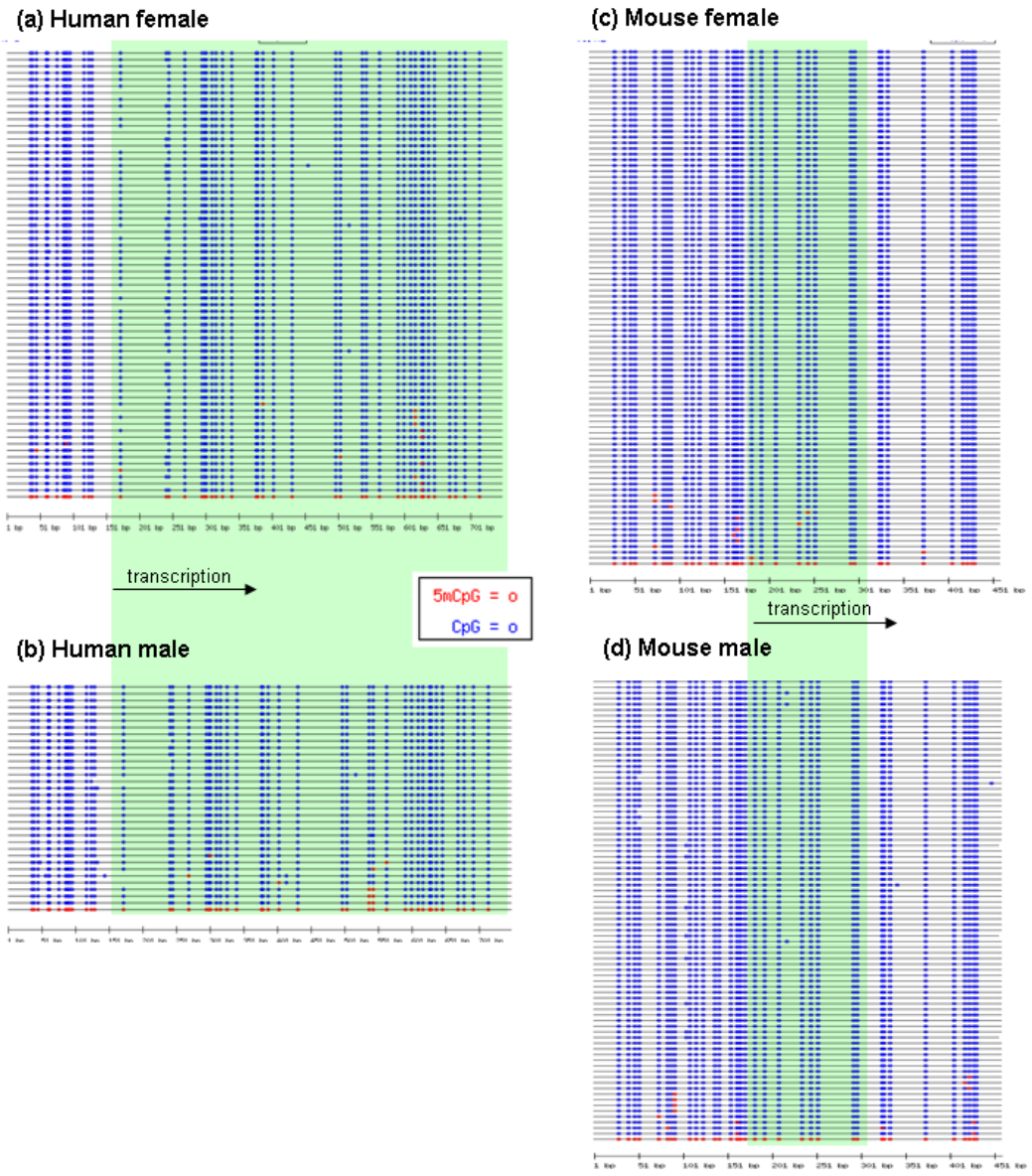

Figure 4.4: CGI methylation profiles of human *DDX3X* and mouse *Ddx3x*. Both genes escape from XCI. The human CGI had a faint signal in the *Hpa*II lane, indicating low level methylation, but the mouse CGI was shown to be hypomethylated in RPMA. The human fragment covers the middle third of the CGI 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.

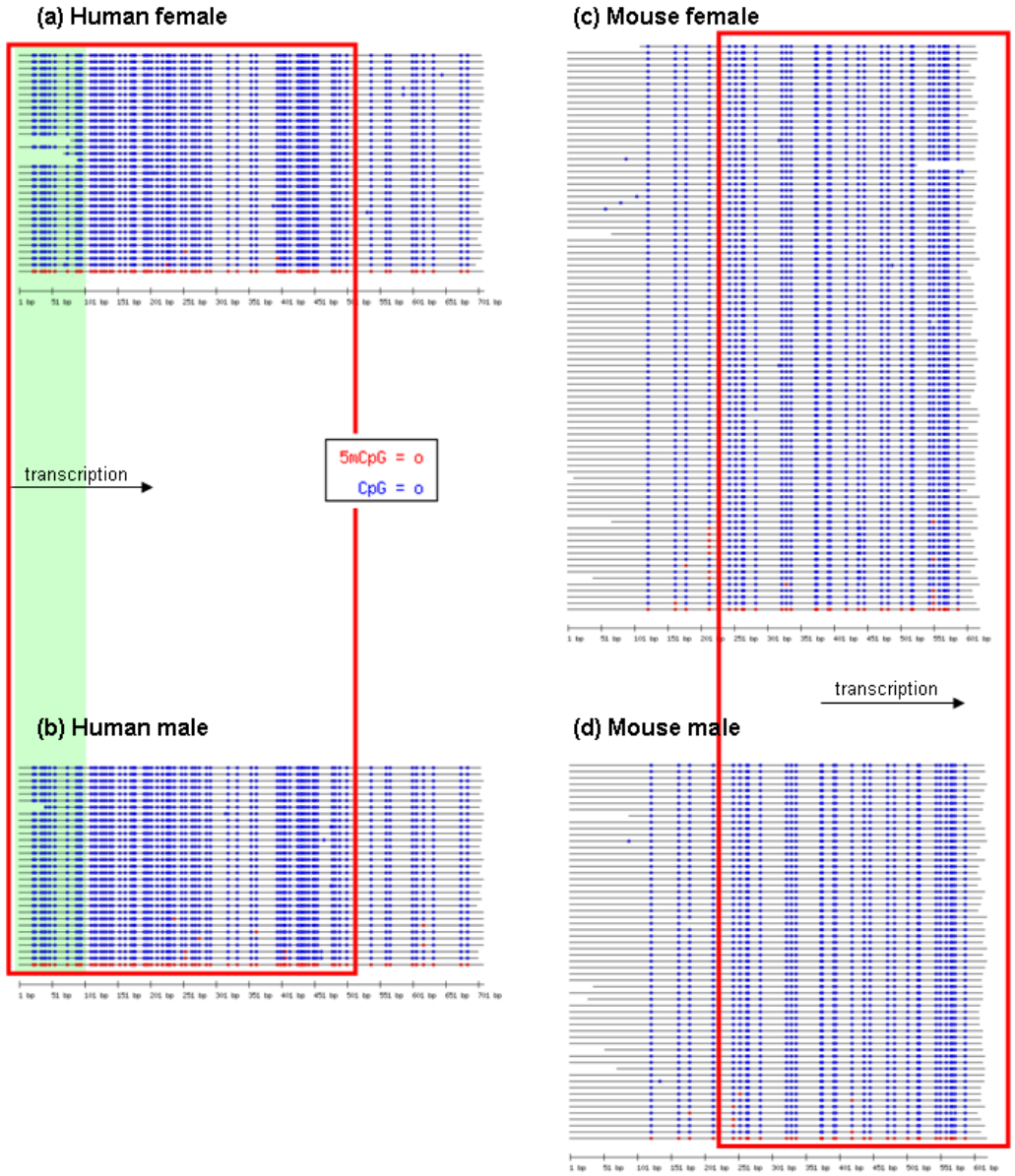

Figure 4.5: CGI methylation profiles of human *UTX* and mouse *Utx*. Both genes escape from XCI and both CGIs were shown to be hypomethylated in RPMA. The human fragment covers the end third of the CGI and the mouse fragment covers the first quarter of the CGI. The island regions are indicated by the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

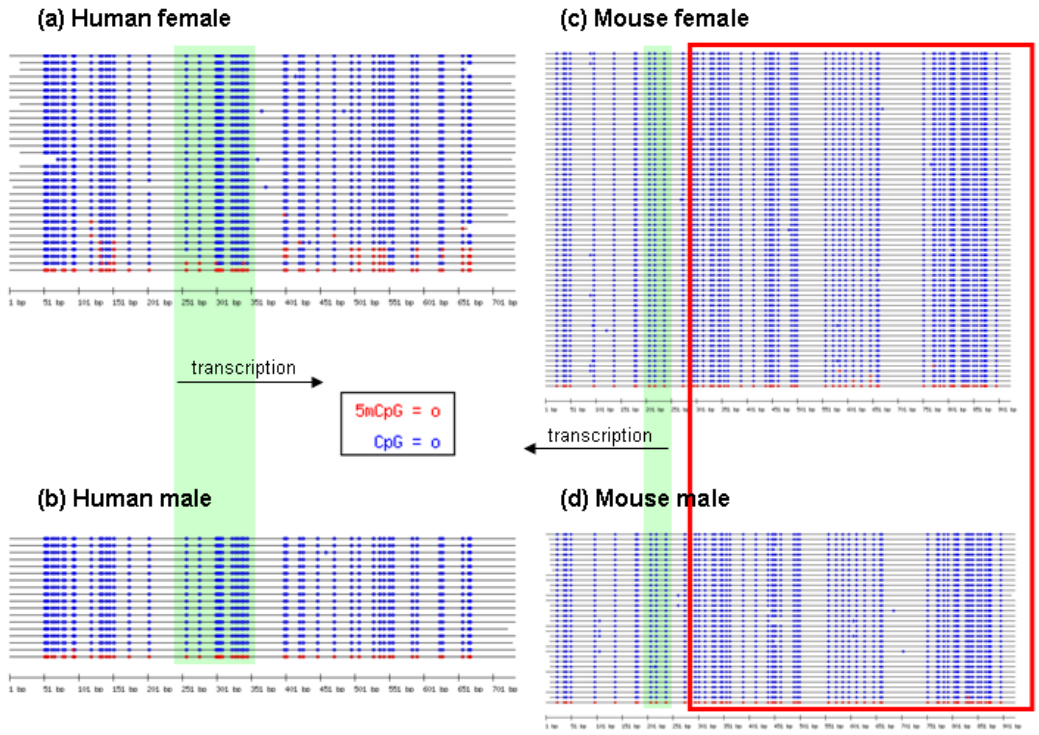

Figure 4.6: CGI methylation profiles of human *EIF2S3* and mouse *Eif2s3x*. Both genes escape from XCI. The human CGI had a faint signal in the *Hpa*II lane, indicating low level methylation, but the mouse CGI was shown to be hypomethylated in RPMA. The human fragment covers the entire CGI and the mouse 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.

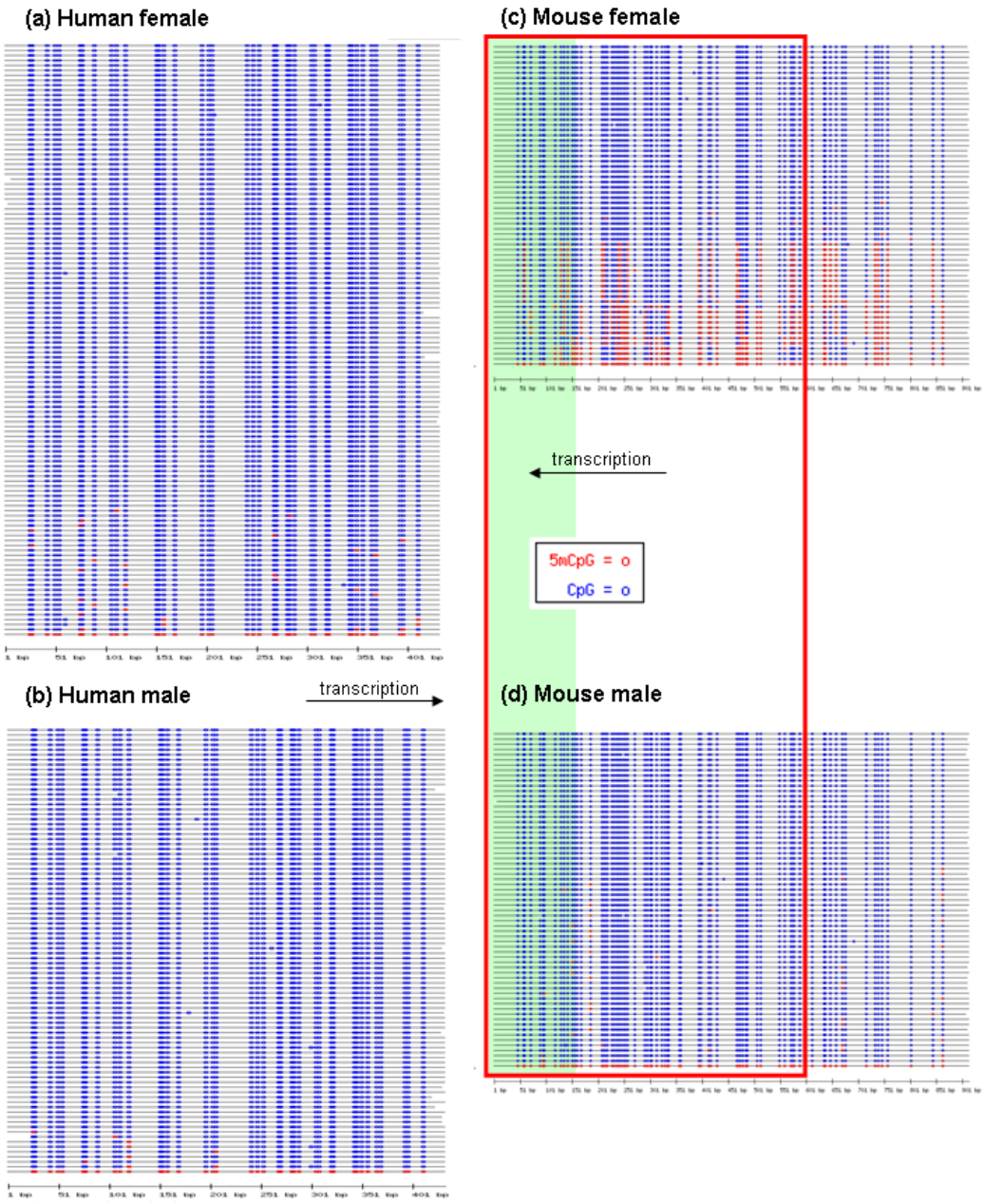

Figure 4.7: CGI methylation profiles of human *OFD1* and mouse *Ofd1*. The human gene escapes from XCI. The human CGI had a faint signal in the *Hpa*II lane, indicating low level methylation, while the mouse CGI was shown to be hypermethylated in RPMA. The human CGI also overlaps the 5' region of *TRAPPC2*, 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.

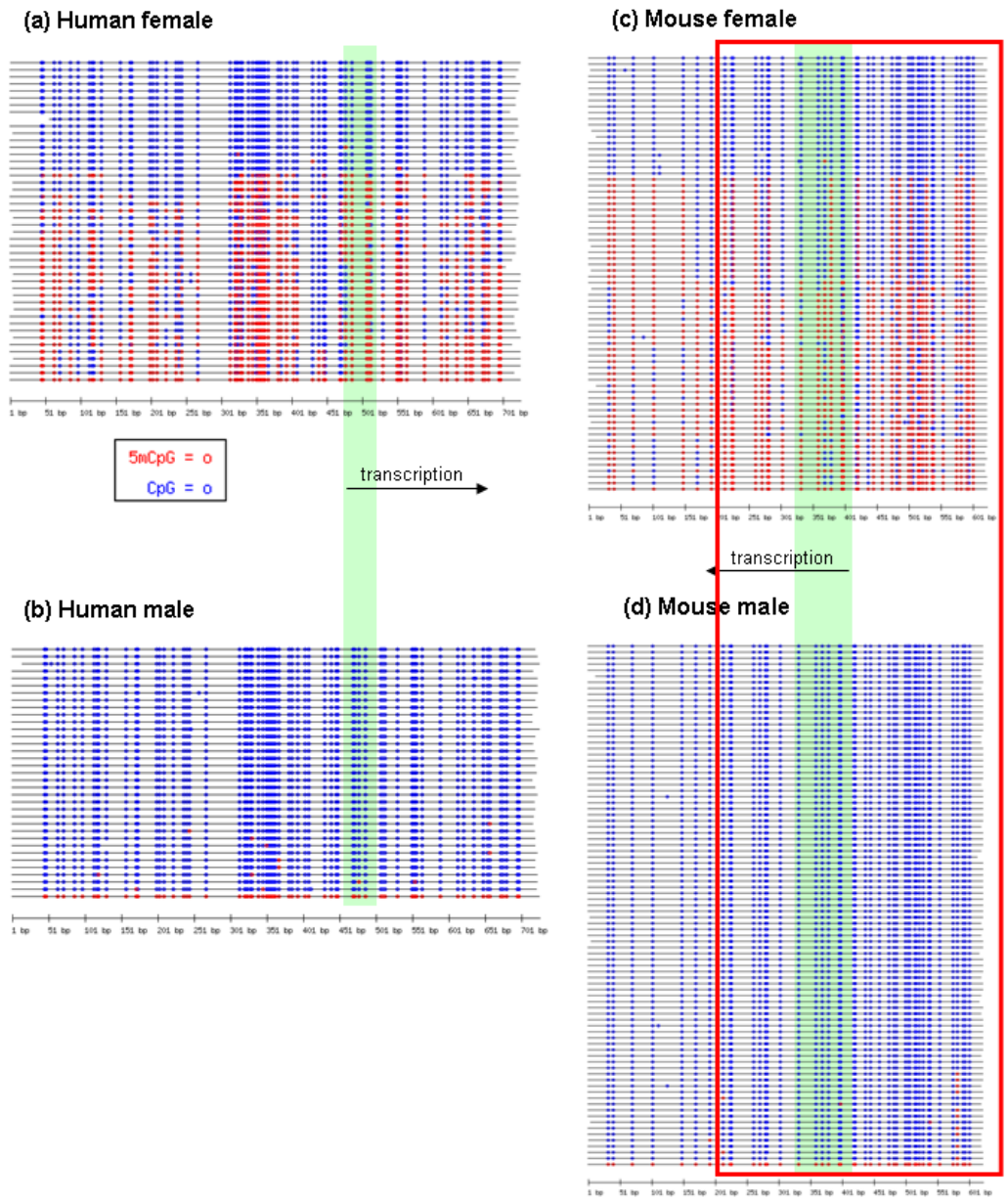

Figure 4.8: CGI methylation profiles of human *POLA* and mouse *Pola1*. Both genes are X inactivated and both CGIs were shown to be hypermethylated in RPMA. The human fragment covers the entire CGI and the mouse fragment covers most of the CGI. The island region is indicated by the red box and exons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

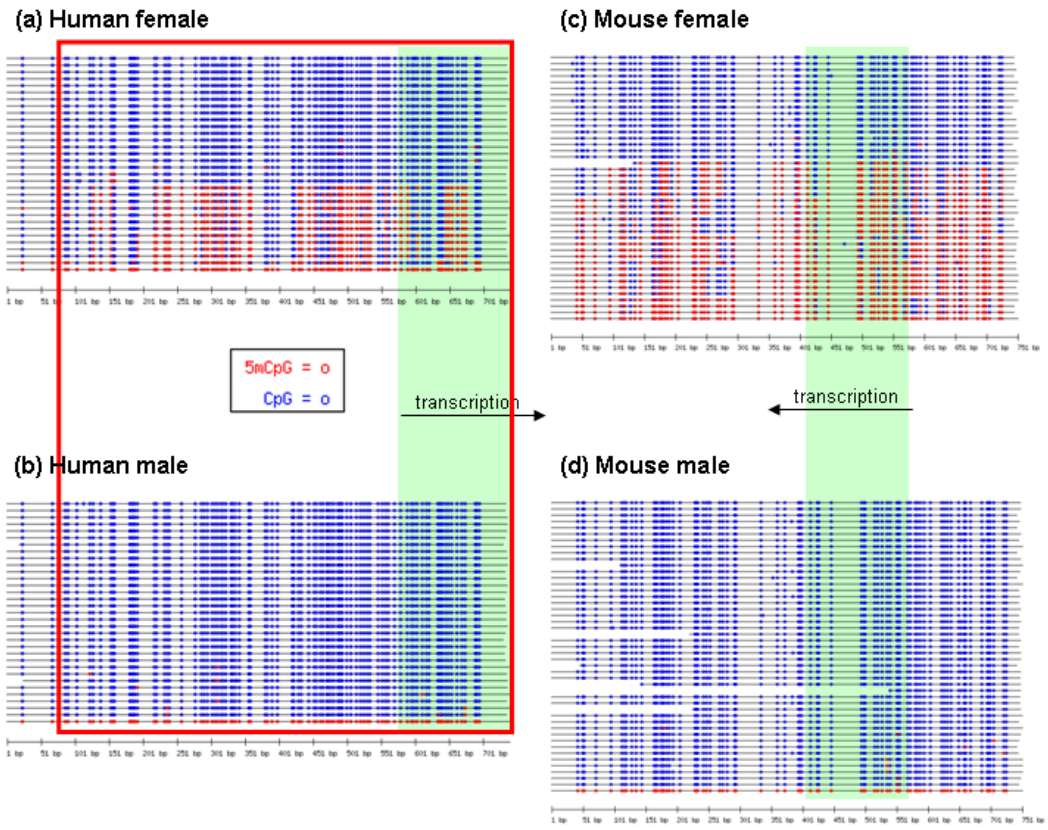

Figure 4.9: CGI methylation profiles of human *MSL3L1* and mouse *Msl3l1*. The human gene is X inactivated and both CGIs were shown to be hypermethylated in RPMA. The human fragment covers the first half of the CGI and the mouse fragment covers the first two thirds of the CGI. The island region is indicated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

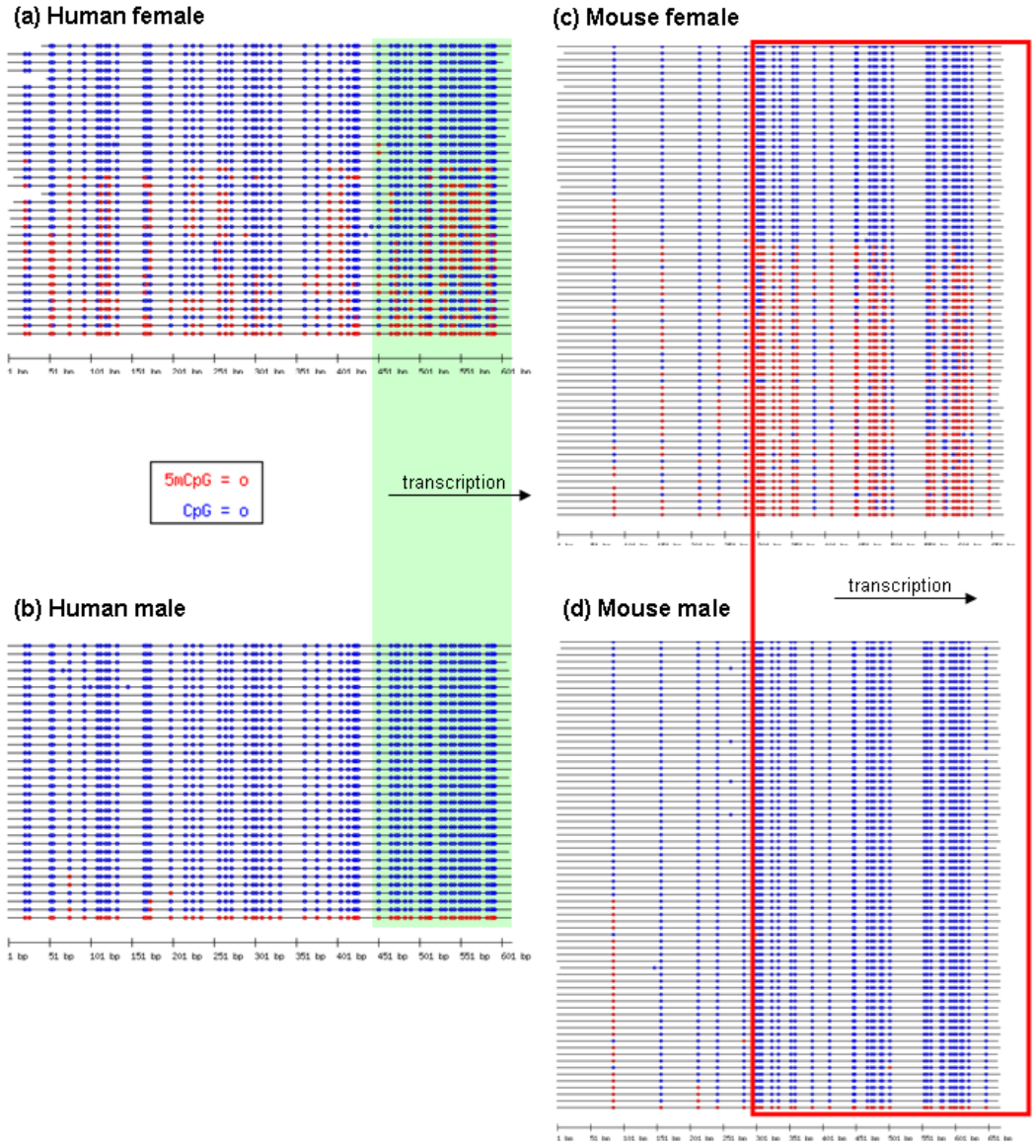

Figure 4.10: CGI methylation profiles of human *ATP6AP2* and mouse *Atp6ap2*. The human gene is X inactivated. Both CGIs were shown to be hypermethylated in RPMA but the human CGI had a faint signal in the *Hpa*II lane of one amplicon. Both fragments cover the first half of the CGI. The island region 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.

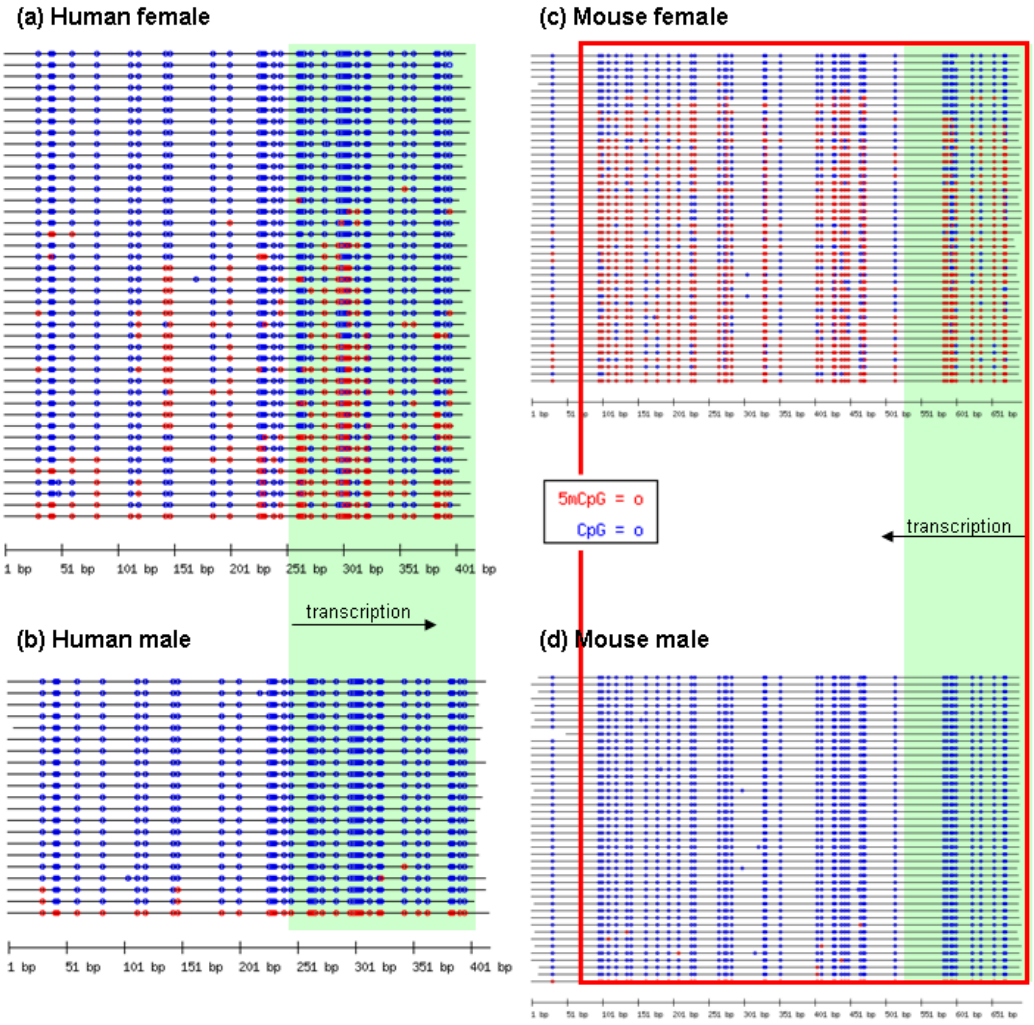

Figure 4.11: CGI methylation profiles of human *PRPS2* and mouse *Prps2*. The human gene is X inactivated. Both CGIs were shown to be hypermethylated in RPMA but the human CGI had a faint signal in the *Hpa*II lane of one amplicon. The human fragment covers the first half of the CGI and the mouse fragment 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.

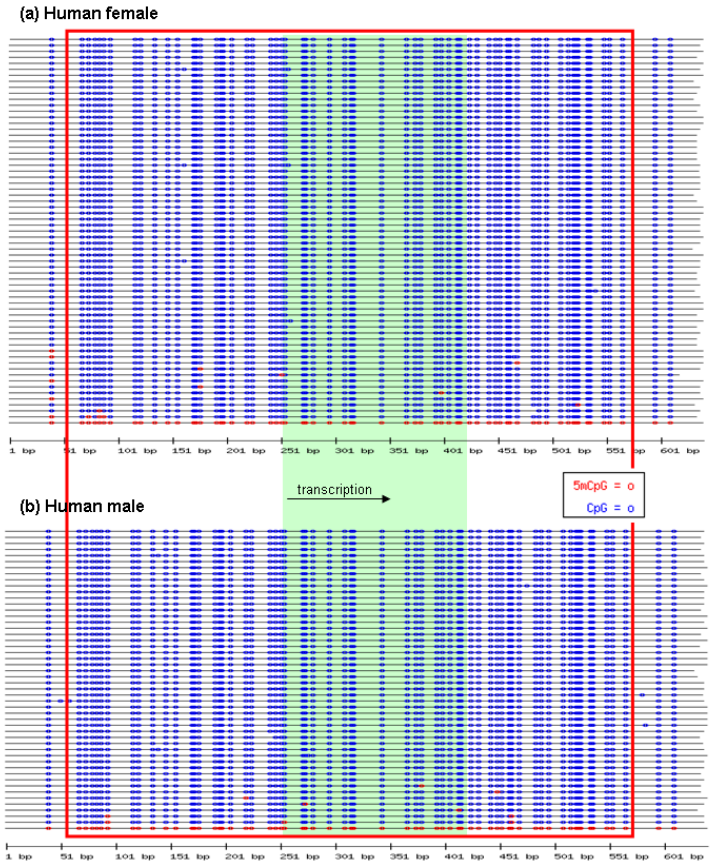

Figure 4.12: CGI methylation profiles of human *RAB9A*. This gene escapes from XCI and its CGI had a very faint signal in the *Hpa*II lane in RPMA. The bisulphite sequenced fragment covers the entire CGI. The island region is indicated by the red box and exon is shadowed in green. Direction of transcription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

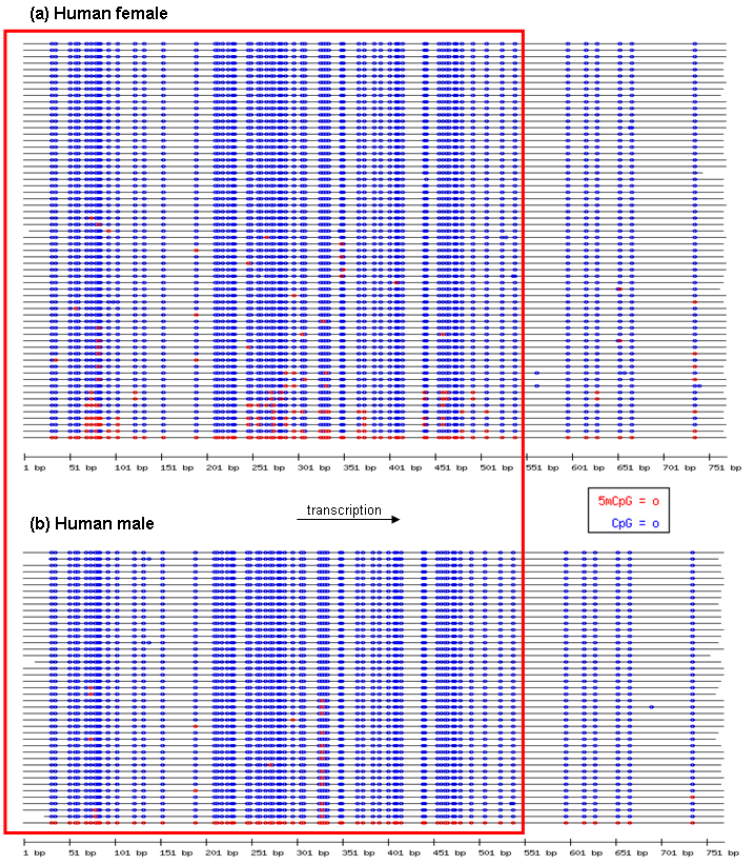

Figure 4.13: CGI methylation profiles of human *CDKL5*. This gene escapes from XCI. Its CGI was shown to be hypomethylated in RPMA, but had had a very faint signal in the *Hpa*II lane. The bisulphite sequenced fragment covers the second half of the CGI and is in intron1. The island region is indicated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

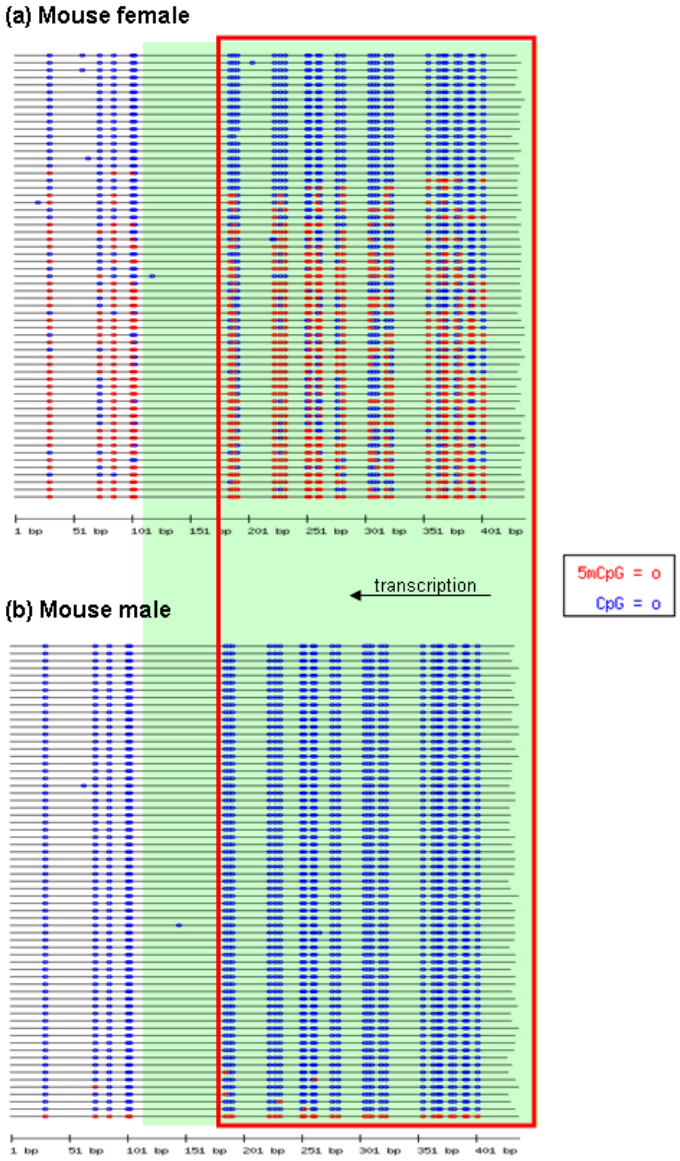

Figure 4.14: CGI methylation profiles of mouse *Syap1*. The CGI was shown to be hypermethylated in RPMA. The bisulphite sequenced fragment covers the first half of the CGI. The island region is indicated by the red box and the exons is shadowed in green. Direction of transcription is indicated by arrow.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

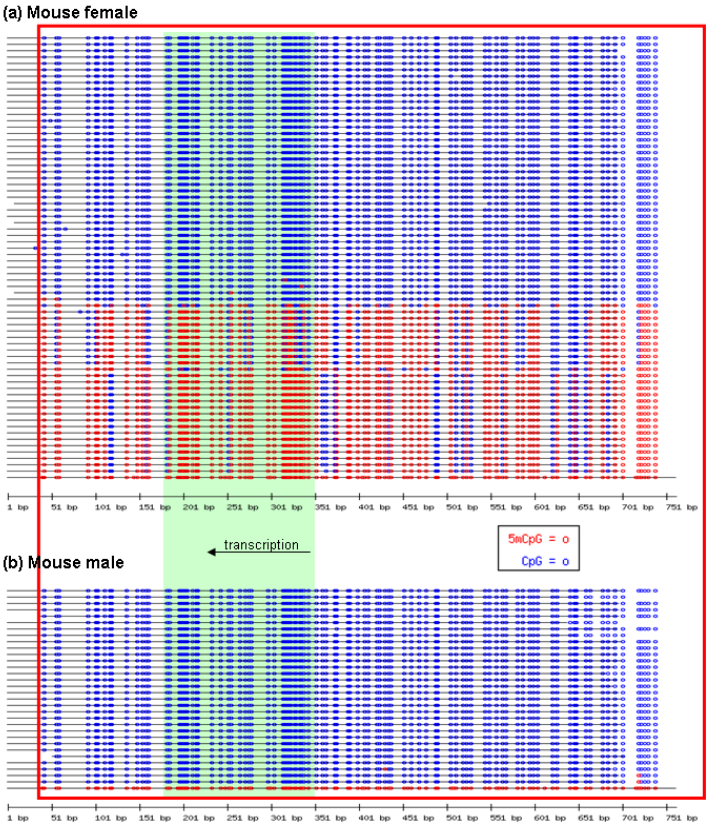

Figure 4.15: CGI methylation profiles of mouse *Hccs*. The CGI was shown to be hypermethylated in RPMA. The bisulphite sequenced fragment covers almost the entire CGI. The island region is indicated by the red box and the exons is shadowed in green. Direction of transcription is indicated by arrow. For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

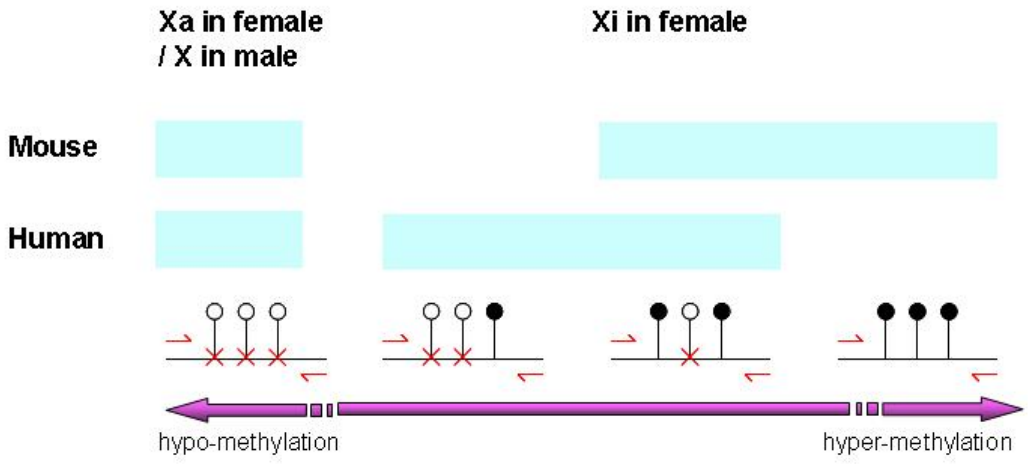

Figure 4.16: Variation of CGI methylation on the human and mouse X chromosomes - revised model (see Figure 4.2 for the original model).