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 ($\mathrm{Obs_{CpG}/Exp_{CpG}}$) 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 $\mathrm{Obs_{CpG}/Exp_{CpG}}$ ratio were calculated from the CGI sequence using the following formula:

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

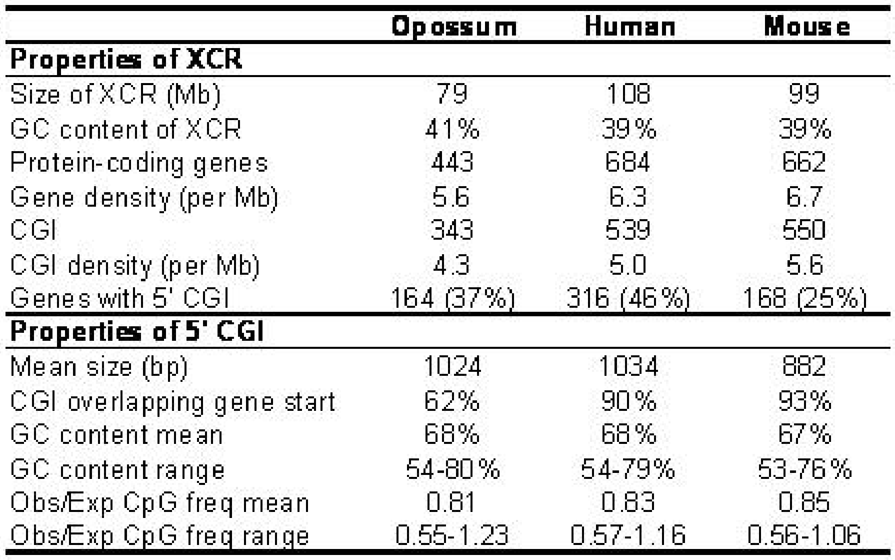
% CpG content = total(CpG) / total(ACTG) * 100

$\mathrm{Obs_{CpG}/Exp_{CpG}}$ 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 $\mathrm{Obs_{CpG}}/\mathrm{Exp_{CpG}}$ 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

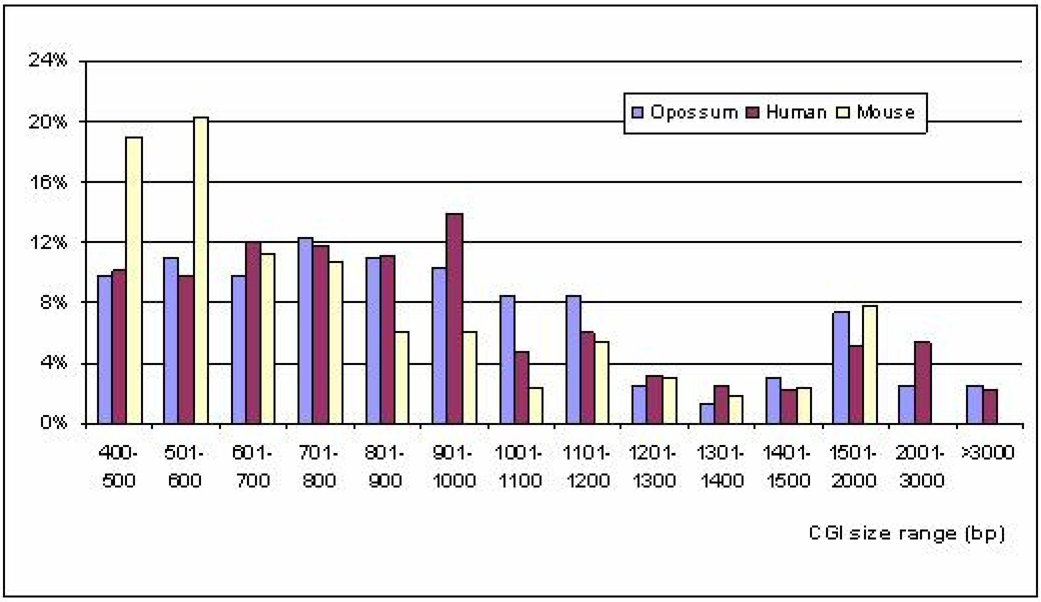


Figure 6.1: Comparison of sizes of 5' CGIs in opossum, human, and mouse XCR.

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').

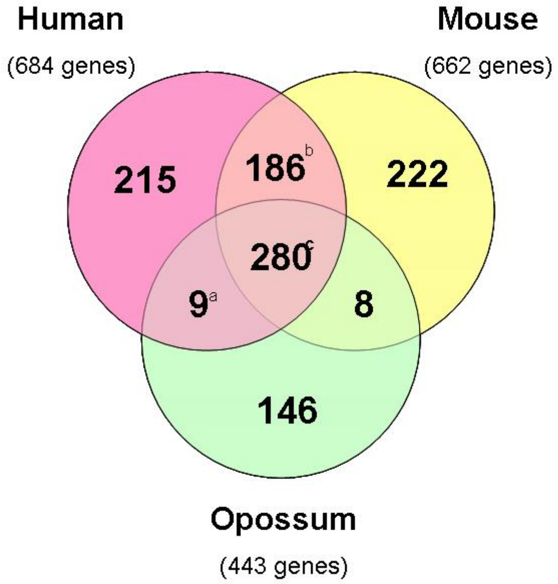
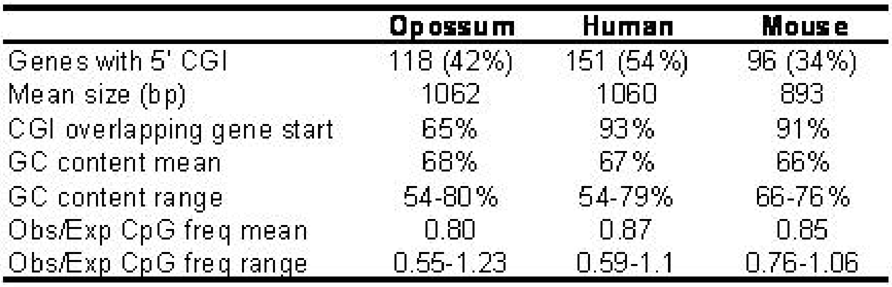


Figure 6.2: Orthologous genes on the opossum, human, and mouse X chromosomes. The number under each species name is the total number of protein-coding genes in each XCR. Genes that were identified as orthologues of the same gene (i.e. paralogues) were counted as one orthologue. Symbols: [a], two human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

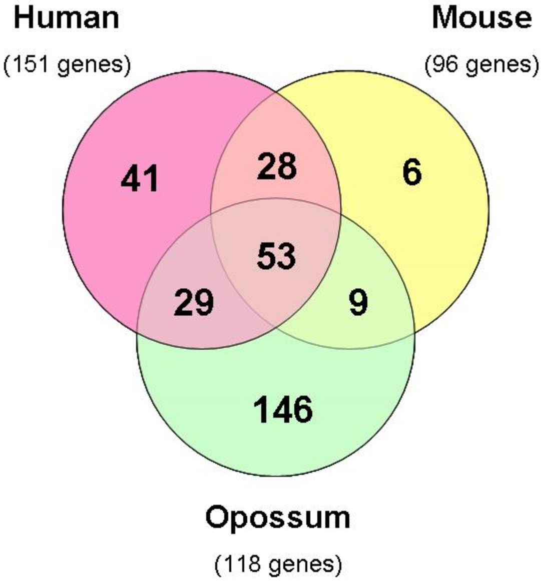


Figure 6.3: CGIs associated with the orthologous gene set. The number under each species name is the total number of genes that are associated with a 5' CGI.

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.

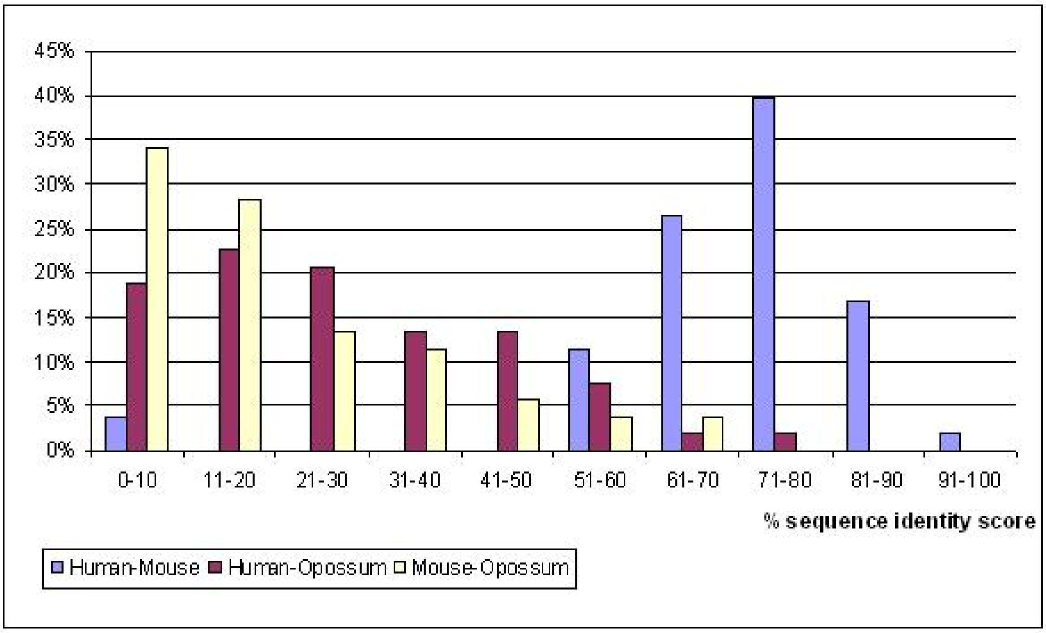


Figure 6.4: Homology between 5' CGIs of the orthologous genes.

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 differential 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.

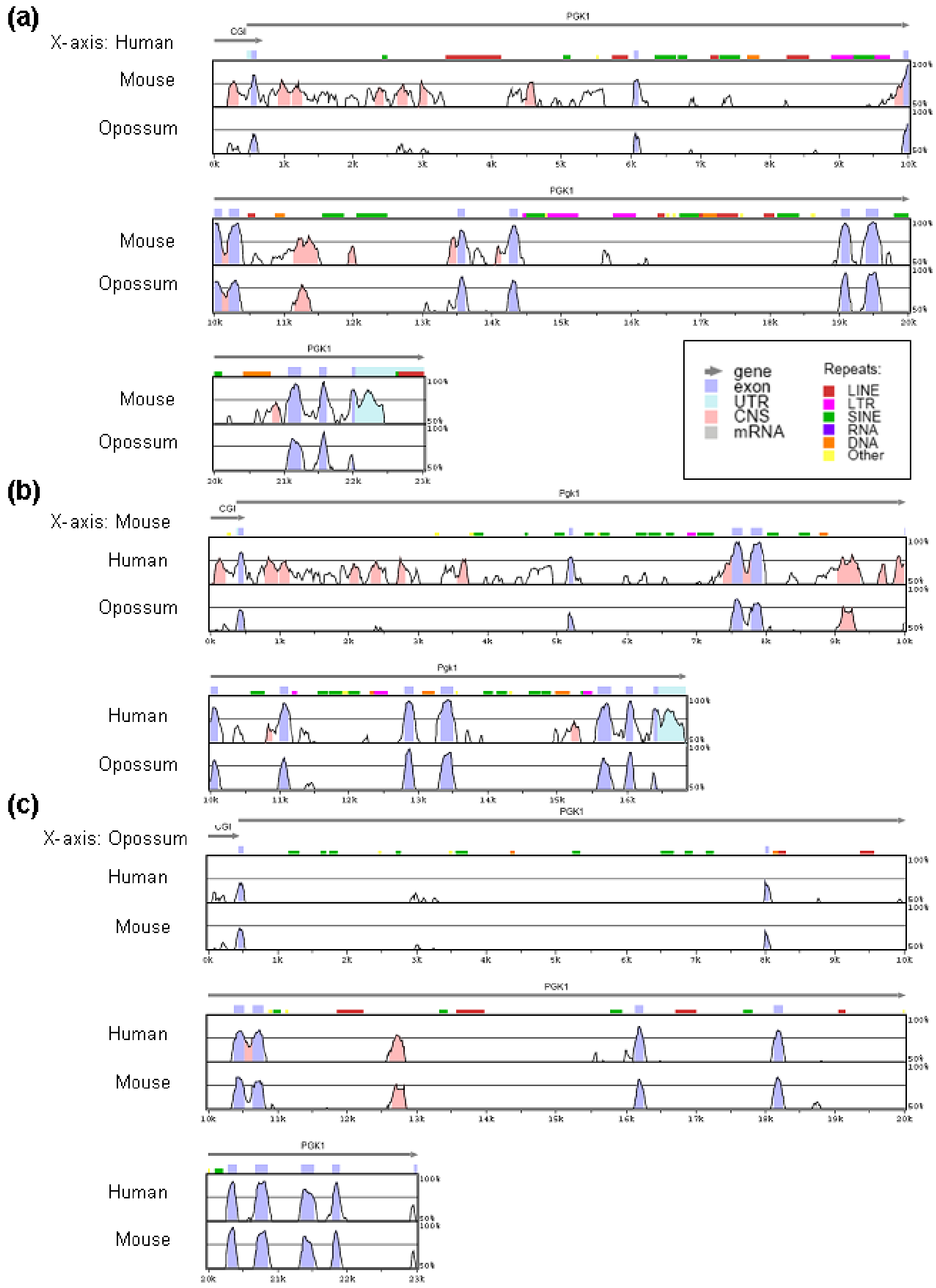


Figure 6.5: VISTA plot of human, mouse and opossum *PGK1* and its CGI. Sequences for human, mouse, and opossum were compared using the MLAGAN alignment program. (a) shows the alignment of the human locus with the mouse and opossum loci; (b) shows the alignment of the mouse locus with the human and opossum 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.

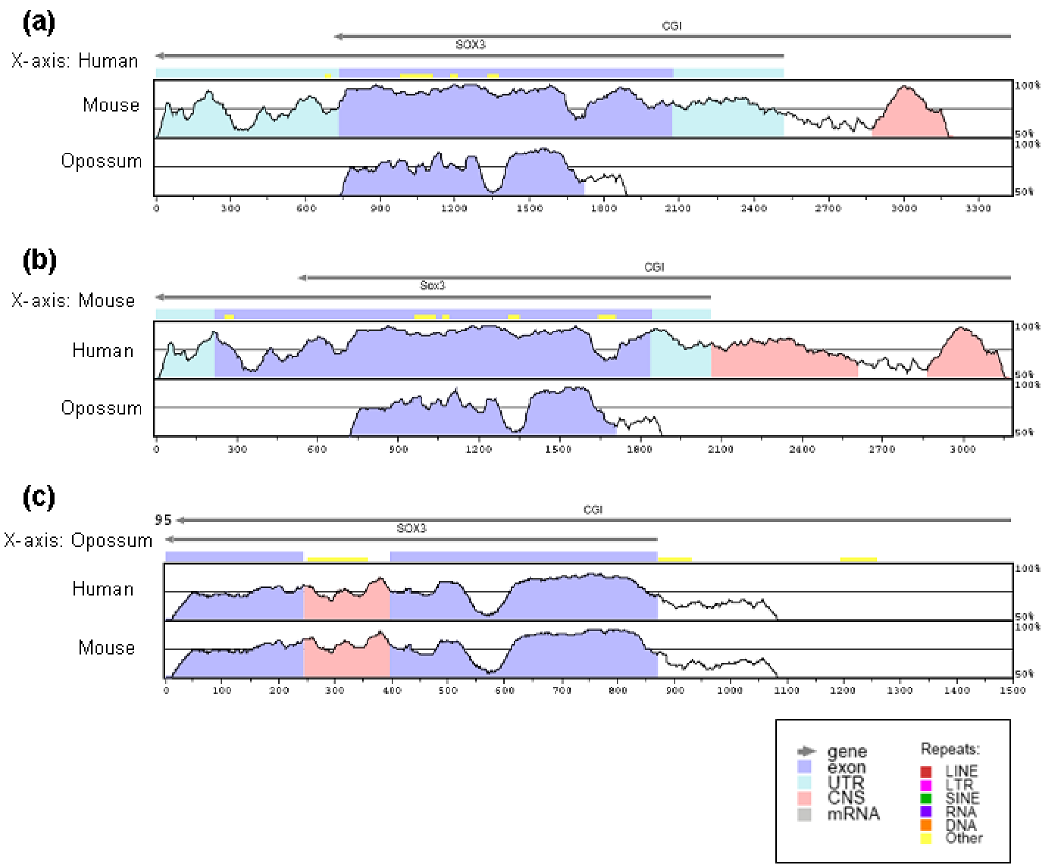


Figure 6.6: VISTA plot of human, mouse and opossum *SOX3* and its CGI. Sequences for human, mouse, and opossum were compared using the MLAGAN alignment program. (a) shows the alignment of the human locus with the mouse and opossum loci; (b) shows the alignment of the mouse locus with the human and opossum 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.

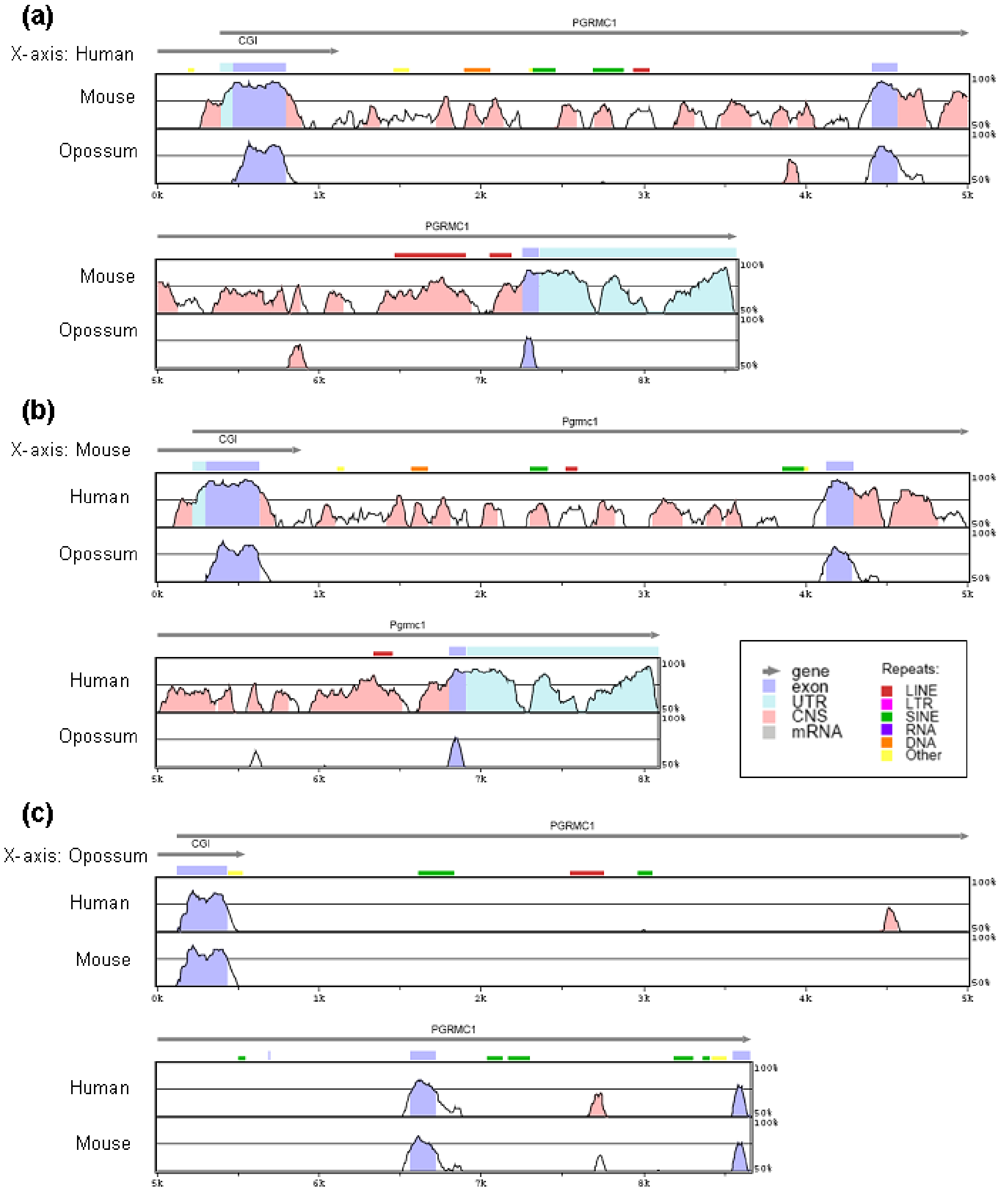


Figure 6.7: VISTA plot of human, mouse and opossum *PGRMC1* and its CGI. Sequences for human, mouse, and opossum were compared using the MLAGAN alignment program. (a) shows the alignment of the human locus with the mouse and opossum loci; (b) shows the alignment of the mouse locus with the human and 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.

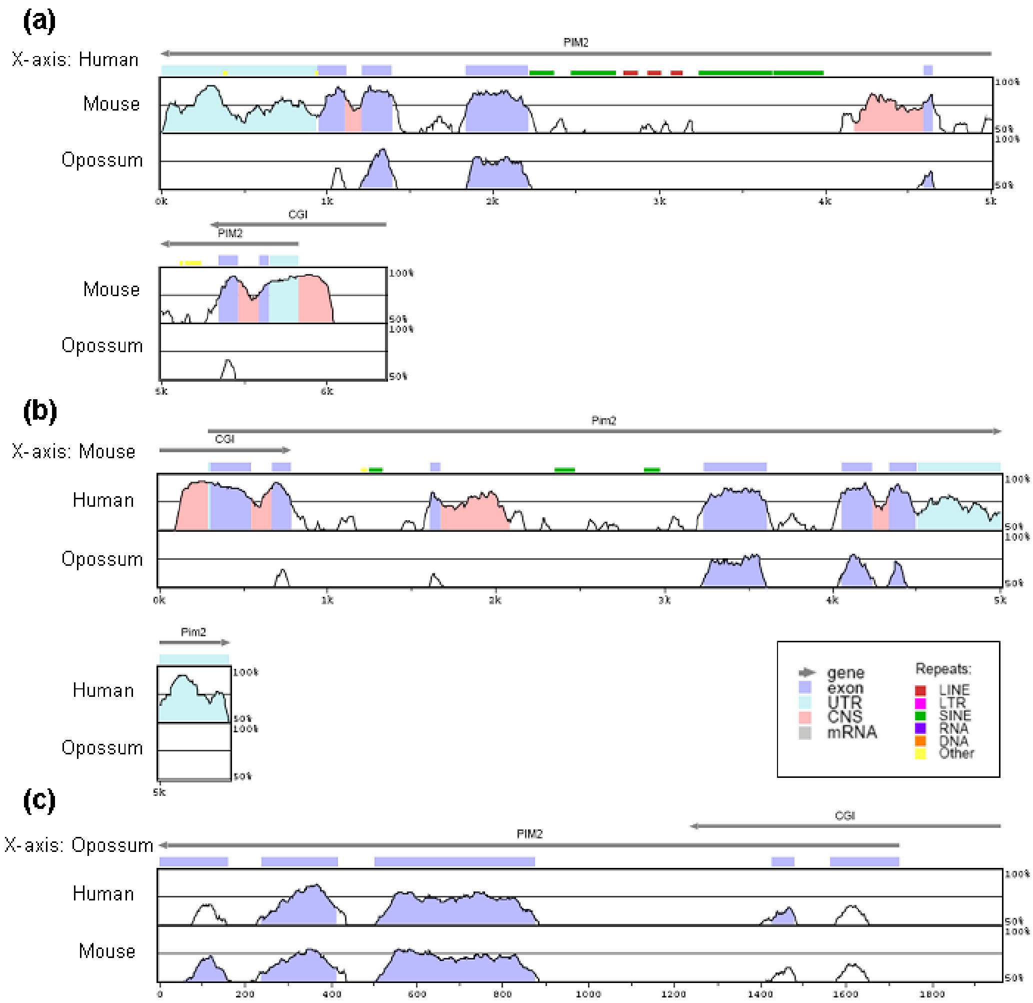


Figure 6.8: VISTA plot of human, mouse and opossum *PIM2* and its CGI. Sequences for human, mouse, and opossum were compared using the MLAGAN alignment program. (a) shows the alignment of the human locus with the mouse and opossum loci; (b) shows the alignment of the mouse locus with the human and opossum 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.