Chapter IV - Sequence Analysis of Vertebrate Orthologous Imprinted Regions (SAVOIR)

4.1 Introduction

4.1.1 Aims of this chapter

This chapter describes the assembly, analysis and annotation of multi-species sequences generated in chapter III with an emphasis on the human 11p15.5 orthologous regions. Where appropriate, links are made to genomic imprinting. However, the available sequences also provide an opportunity to explore the genomic landscapes of species for which little or no high-quality sequences exist.

The results begin with a description of the assembly of BAC clone sequences and the comparison of resulting finished sequence contigs with available WGS draft sequences. I then discuss the gene content of the sequences and how these compare with human before going on to investigate what comparative sequence analysis can tell us about both broad and refined features of the genomic landscapes. Next the highly variable interspersed repeat and C+G contents of the sequences are explored. Finally, a description of the SAVOIR website is provided and how the data generated here is being used by the SAVOIR consortium, established to make the most of the rich resources.

At the onset of this project (August 2003) the genome sequences of human and mouse were practically complete and comprehensive analyses of their genomes were published (International Human Genome Sequencing Consortium. 2001, Waterston et al. 2002). In addition there were draft genome sequences for the rat (*Rattus*

norvegicus) and Fugu (*Takifugu rubripes*) genomes. There was therefore a huge unexplored gap in the vertebrate phylogenetic tree between eutherians and fish (see chapter I). In the intervening years a wealth of additional draft genome sequences has become available (http://genome.ucsc.edu/ and http://www.ensembl.org/). However, some vertebrate classes remain underrepresented including marsupial and monotreme mammals. The draft genome sequence of the South American opossum (*Monodelphis domestica*) was recently analysed and gave the first broad insights into marsupial genomics (Mikkelsen et al. 2007). Draft genome sequences for a second marsupial, the Tammar wallaby (*Macropus eugenii*) and the monotreme platypus (*Ornithorhynchus anatinus*) have been generated and their analyses are eagerly awaited. However, as discussed in the previous chapter, the draft WGS sequence for platypus lacks coverage in the IC1-IC2 regions, despite a 6x genome coverage. With only a 2x genome coverage of sequence reads performed for the tammar wallaby this assembly too will have limitations of coverage and accuracy.

The finished sequences generated in chapter III therefore provide an early opportunity to study the genetic basis of biological diversity. Furthermore, as these sequences are amongst the first to be finished to the levels of accuracy of human and mouse sequences it is appropriate to compare them with WGS assemblies, where available. Coverage and quality of these 'draft' WGS assemblies is variable, both between species and regions within a species, and is dependent upon technical (i.e. depth of sequencing performed) and biological (e.g. C+G and repeat contents) factors.

The finished BAC clone sequences generated in chapter III were subject to the same high quality standards as those for the human genome

(http://www.genome.gov/10000923). Likewise, the regional assembly of multi-species sequences followed guidelines established in the human genome project. This included the recommendation to submit finished neighbouring BAC sequences with at least 2 kb overlaps to confirm that they overlap without generating too much sequence redundancy (Adam Felsenfeld [Human Genome Working Group], personal communication). AGP (A Golden Path) files containing the information required to generate a consensus from finished BAC clone sequences were created according to specifications detailed at: http://www.ncbi.nlm.nih.gov/projects/genome/guide/Assembly/AGP_Specification.html. These consensus sequences form the basis for all subsequent analyses presented here.

Sequence alone can rarely, if ever, provide biological insight without some form of annotation. Typical annotation features include genes, unusual nucleotide composition (e.g. CpG islands) and repeat elements and all of these are discussed in this chapter. Annotation is best performed on high-quality sequence to minimise ambiguities caused by incomplete or error prone sequences. Initially consensus sequences were analysed in the semi-automated 'otter' pipeline by the analysis coding (anacode) group (Searle et al. 2004). This preliminary analysis uses a combination of similarity searches against public DNA and protein databases and *ab initio* gene prediction algorithms. Annotation of gene structures based on this analysis was subsequently performed by Charles Steward in the HAVANA group (http://www.sanger.ac.uk/HGP/havana/, Ashurst et al. 2005).

Gene annotations were categorised according to available evidence; Known genes, Novel genes (coding sequence), Novel transcripts, Putative genes and Pseudogenes. Known genes are identical to human cDNA or protein sequences held within the

NCBI Entrez Gene database (Maglott et al. 2007) at the time of annotation. With the accumulation of evidence over time the numbers of genes within this category will increase. Novel genes have an open reading frame (ORF) and are identical to cDNA or protein sequences but have not yet made it into the Entrez Gene database with an official name, approved by the Human Genome Organisation (HUGO) gene nomenclature committee (Bruford et al. 2007). For some species annotated here (e.g. wallaby and platypus) there is a paucity of species-specific mRNA data in the public domain and therefore the novel gene category is commonly used. The novel transcript category is defined as for novel gene with the exception that an ORF cannot be unambiguously defined. Non-coding genes, such as *H19*, fall within this category. Putative genes match spliced ESTs but are lacking a significant ORF. Generally these are short genes or gene fragments. Finally, pseudogenes (processed and unprocessed) are annotated where there is homology to proteins but the coding sequence (CDS) is interrupted by stop codons. Importantly an active copy of the gene, with full CDS giving rise to the protein, should have been identified elsewhere in the genome.

Repeats, once thought to be the 'junk' of a genome are proving to be anything but junk with a growing list of functional and structural roles in genomes (Berg. 2006). Mammalian genomes typically comprise of about 50% of repeat sequences, which may still be active in a lineage or the relics of ancestral genes. Active repeat elements (transposons) can, and do, reshape the genome causing rearrangements as they transpose from one genomic region to another. Consequences of this can be seen in the creation of entirely new genes or rearrangement of existing genes and therefore these repeat elements are driving genome evolution (Deininger and Batzer. 1999, Lowe et al. 2007, Szabo et al. 1999). The availability of finished sequences for highly

diverged species allows us to compare repeat contents and address their role in genome expansion (the C-value paradox) and biological processes.

In contrast to the sequences of eutherian mammals, much less is known about the marsupial and monotreme genome repeat contents. Analysis of repeats in these non-eutherian mammals could provide valuable insight into the evolutionary history of the mammalian genome.

Many studies have shown the biological relevance of high and low C+G contents in multiple vertebrate genomes. For example, the correlation between C+G content and gene density, repeat densities and type, Giemsa stained chromosome bands and recombination rates. Within a given genomic region there are wide variations from the genome average of C+G content. In the human genome these deviations from the average were classified into 5 categories termed isochores (Bernardi. 1995, Bernardi. 2000). Isochores with below average C+G contents were termed low (L)1 and L2 with C+G contents <38% and 38-42%, respectively. Equally, high (H) isochores (H1, H2 and H3) contain C+G contents of 42-47%, 47-52% and >52%, respectively. The sequencing of the human genome indicated that in any given window size the C+G contents are not homogeneous and therefore the prefix 'iso' is misleading. Regardless of the term used, compartmentalising regions based on C+G content does have predictive value for genome function and therefore warrants further study between species.

The observed frequency of CpG dinucleotides (CpGs) in the human and mouse genomes (2% and 1.7%, respectively) are lower than the expected 4.4% obtained by multiplying the typical fraction of cytosine and guanine nucleotides (0.21 x 0.21). The paucity of CpGs can be explained because of the instability of methylated cytosine in the CpGs which is spontaneously deaminated resulting in thymine and

an accumulation of TpGs (CpA on the reverse strand) (Sved and Bird. 1990). By contrast, the deamination of unmethylated cytosine nucleotides results in uracil which the cell rapidly repairs. Clusters of unmethylated CpGs occurring at a higher frequency are termed CpG islands and their association with the promoter regions of many genes are one example of the functional relevance of C+G content (Bird. 1986). Indeed the differential methylation of these CpG islands (termed differentially methylated regions (DMR), Sasaki et al. 1992) are a hallmark of imprinted gene expression. It is therefore important to identify CpG islands in the genomes of species studied as a pre-requisite to the ultimate determination of methylation status.

The power of mouse genetics to further our knowledge of human gene regulation and disease is enhanced by the identification of single nucleotide polymorphisms (SNPs). Genetic crosses of inbred mouse strains have demonstrated that SNPs are readily detected (Adams et al. 2005, Lindblad-Toh et al. 2000). However, genome-wide distributions of these invaluable markers are low or non-existent in some model species and/or regions (Salcedo et al. 2007). The sequencing across IC1 and IC2 domains in the Western Mediterranean short-tailed mouse (*Mus spretus*) (chapter III) provides an opportunity to systematically identify sequence variants between this species and the finished *Mus musculus musculus* sequence (NCBI Build 37). As discussed in chapter I, sequence variants allow the discrimination between parental alleles in the congenic mouse strain SD7 that is widely used in IC1 and IC2 domain imprinting research.

4.2 Sequence assemblies

To gain a picture of the genomic landscape of orthologous sequences from imprinted gene regions, individual finished BAC sequences first require assembling into non-redundant sequence contigs. Of the 101 BAC clones selected for sequencing from orthologous imprinted gene regions, 96 have been finished, spanning 10.8 Mb of DNA sequence (chapter III). The assembly of these sequences followed by their comparison with whole genome shotgun assemblies is discussed here.

4.2.1 Assembly of BAC sequences

The overlapping BACs selected for sequencing constitute a tile path. For each species a tile path format (TPF), tab delimited, file was created which lists the mapped order of sequence accession numbers, provides the international clone identifier and name of contig in which the BAC was physically mapped. Table IV-1 shows an example of a TPF file for wallaby chromosome 2, orthologous to human chromosome 11p15.5.

Table IV-1. Example of a tile path format file.

##species=Wallaby chromosome=2

CU467493	MEKBa-205I8	Wallaby_2ctg192
CU458744	MEKBa-283A6	Wallaby_2ctg192
CR855994	MEKBa-69G12	Wallaby_2ctg192
CT008508	MEKBa-459D3	Wallaby_2ctg192
CR925759	MEKBa-346C2	Wallaby_2ctg192
CR848708	MEKBa-201B9	Wallaby_2ctg192
CU024874	MEKBa-517H11	Wallaby_2ctg192
CU024865	MEKBa-439O2	Wallaby_2ctg192
CU311200	MEKBa-183M18	Wallaby_2ctg192
CU041371	MEKBa-363O3	Wallaby_2ctg192
CU062506	MEKBa-183I7	Wallaby_2ctg192
CT990571	MEKBa-465N20	Wallaby_2ctg192

The top line of the TPF file defines the species and chromosome or region mapped. Accession numbers from sequence submissions to EMBL are provided (first column) for each mapped BAC clone (middle column). MEKBa denotes the international clone name for *Macropus eugenii* (tammar wallaby) BACs constructed at the Arizona Genomics Institute.

I uploaded TPF files into the in-house software and interface package 'ChromoView' (Ben Tubby, Darren Grafham *et al.*, unpublished), which is a chromosome viewer developed with the aim of creating an AGP file from the submitted TPF file. Within ChromoView overlaps between finished BAC sequences were confirmed using cross_match (Green,P. unpublished, www.phrap.org). Importantly, automated overlap detection was manually checked by me and, if necessary, modified within ChromoView prior to the export of sequences for analysis. Generally sequence overlaps of 2 kb were submitted to the EMBL database to support the clone overlaps whilst minimising redundancy in finished sequence. In a few cases it was necessary to submit extended sequence overlaps to validate an assembly when sequence complexities such as duplications, deletions or high polymorphism rates were present. Coordinates for assembled contiguous sequences were calculated to give the AGP file. Table IV-2 shows the AGP file created from the TPF in Table IV-1, representing the 1.5 Mb region of wallaby chromosome 2, orthologous to human chromosome 11p15.5. Each component of the AGP file

(object in Table IV-2) contributes a beginning and end coordinate to the object. These coordinates therefore define the unique sequences to be used to construct a linear sequence. AGP files for each species and each region sequenced in chapter III were exported from ChromoView into the ACeDB database 'otterlace' for sequence analysis and annotation (section 4.3).

Table IV-2. Example of 'a golden path' (AGP) format file.

Object				Component				
Name	Beginning	End	Part number	Type	ID	Beginning	End	Orient-ation
Wallaby-chr2	1	146445	1	F	CU467493.1	1	146445	+
Wallaby-chr2	146446	222894	2	F	CU458744.1	2001	78449	+
Wallaby-chr2	222895	395592	3	F	CR855994.1	2001	174698	+
Wallaby-chr2	395593	397586	4	F	CT008508.1	2001	3994	+
Wallaby-chr2	397587	542343	5	F	CR925759.7	2001	146757	+
Wallaby-chr2	542344	734871	6	F	CR848708.12	1997	194524	+
Wallaby-chr2	734872	801067	7	F	CU024874.2	2001	68196	+
Wallaby-chr2	801068	980226	8	F	CU024865.1	2001	181159	+
Wallaby-chr2	980227	1100864	9	F	CU311200.1	2001	122638	+
Wallaby-chr2	1100865	1215597	10	F	CU041371.1	2001	116733	+
Wallaby-chr2	1215598	1350537	11	F	CU062506.1	2001	136940	+
Wallaby-chr2	1350538	1528894	12	F	CT990571.1	4019	182375	+

The AGP file describes the assembly of a 1.5 Mb region of wallaby chromosome (chr) 2 from component parts; F, finished sequence. The component IDs are sequence accession numbers deposited in the EMBL database. The extent of the component sequences used to build the AGP object is defined by a beginning and end coordinate together with orientation; +, forward.

4.2.2 Comparison with whole genome shotgun sequence assemblies

The availability of considerable amounts of regional finished sequence enables the comparison of each of the 9 SAVOIR selected regions with publicly available WGS sequence assemblies and therefore an evaluation of the coverage and quality of WGS assemblies can be determined. For the species of interest in this thesis WGS

sequence assemblies are available for chicken, platypus and South American opossum but not wallaby, although this is anticipated in the near future. The start and end sequences for each finished sequence were searched against the relevant species sequence assembly at the UCSC genome browser using BLAT (Kent. 2002). Corresponding intervals are provided in Table IV-3.

Table IV-3. Comparison of finished and draft genome sequences.

Species	Region	SAVOIR accessions		Finished sequence span (bp)	WGS Location	From	To	WGS contig span (bp)
		First	Last					
Chicken	IC1-IC2	BX663531	BX640540	1162135	chr5	13991503	15164319	1172817
Platypus	IC1-IC2	CT573284	CU469422	>762175	Multiple	NA	NA	?
Platypus	DLK1-DIO3	CT573041	CR536601	795237	Ultra378	6200104	6929554	729421
Platypus	STX16-GNAS	CT573151	CT030198	484161	Ultra516	7821325	8305994	484670
Platypus	GRB10	CT978601	NA	135754	chr4	9157830	9298211	140382
Platypus	PLAGL1	CU207364	NA	152936	Multiple	NA	NA	?
Platypus	IGF2R	?	CR933560	383712	Multiple	NA	NA	?
Platypus	SLC38A2/4	BX936293	BX936288	>331739	Multiple	NA	NA	?
Platypus	UBE3A	CR938721	NA	171880	Ultra222	9157649	9265994	108346
Platypus	DNMT1	CU326346	NA	206126	Multiple	NA	NA	?
Opossum	IGF2	CU468641	NA	136229	Multiple	NA	NA	?

WGS assemblies compared are: Chicken, galGal3 (May 2006); Platypus, ornAna1 (March 2007); Opossum, monDom4 (January 2006). NA, Not applicable. Multiple, unmapped contigs. Finished sequence in the platypus IC1-IC2 region is not contiguous. ?, Unknown.

The chicken sequence assembled from mapped BACs and corresponding region of the WGS assembly are remarkably similar in length (1,162,135 bp and 1,172,817 bp, respectively, Table IV-3). However, it should be noted that since the original chicken assembly of a 6.6X genome coverage of sequence reads (galGal2, February 2004, Hillier et al. 2004) an additional 198,000 reads from contig ends and poor quality regions have been added to create the May 2006 assembly (galGal3). The current chicken WGS assembly therefore constitutes an enhanced draft genome sequence. In contrast, of the 9 regions sequenced in platypus, only 4 lie in ultracontigs (Table IV-3). Ultracontigs are defined as ordered and orientated sequence contigs linked to the platypus physical map using BAC end-sequence and

in silico digest data. Only the *GRB10* containing sequence has been mapped to a platypus chromosome (chromosome 4) in the ornAna1 (March 2007) WGS assembly. As can be seen from dot-plots (methods to visualise sequence similarity, Maizel and Lenk. 1981) between WGS and finished sequences the order and orientation of contigs within ultracontigs is generally accurate (Figure IV.1). However, there are some evident problems even within these curated assemblies. Most notably (at the resolution shown) in the *UBE3A* region of ultracontig 222 there are two >30 kb deletions when compared to the finished BAC sequence, CR938721 (Figure IV.1D). Sequences from CR938721 which were missing from ultracontig 222 were identified by BLASTN analyses in other WGS contigs not linked to the platypus physical map (e.g. Contig13421 and Contig17578 of the ornAna1 assembly). Discrepancies between WGS and finished sequence assembly lengths can generally be explained by missing sequences in the WGS assemblies (e.g. *DLK1-DIO3* and *UBE3A* regions, Table IV-3) or expansions in the WGS assemblies caused by the addition of arbitrary padding characters between non-overlapping sequence contigs (e.g. *GRB10*, Table IV-3).

Five of the platypus regions are spanned by multiple WGS contigs that are not linked to one another or the physical map (Table IV-3). This is the case in the IC1-IC2 region where, using the SAVOIR BAC sequences mapped in this thesis, I have identified 74 contigs with an average length of only 6 kb (range 726 bp to 29,131 bp). Reasons for the fragmented assemblies in this region of the platypus genome include the high C+G and repeat sequence contents which also hampered BAC mapping and sequencing progress (see chapter III and sections 4.5 and 4.6 below).

The observation that even genomes sequenced to a depth of >6X (as is the case for platypus) can have extended regions in which many sequence contigs cannot be mapped, re-affirms the benefit of generating finished sequence. Furthermore, even

large ultracontigs have local mis-assemblies and missing sequences only revealed by the availability of mapped and finished sequences from the clone-by-clone approach. In some genomic regions, such as the platypus IC1-IC2 region, the 'reference' draft genome will be of limited use unless additional funding is made available to enhance the WGS assembly.

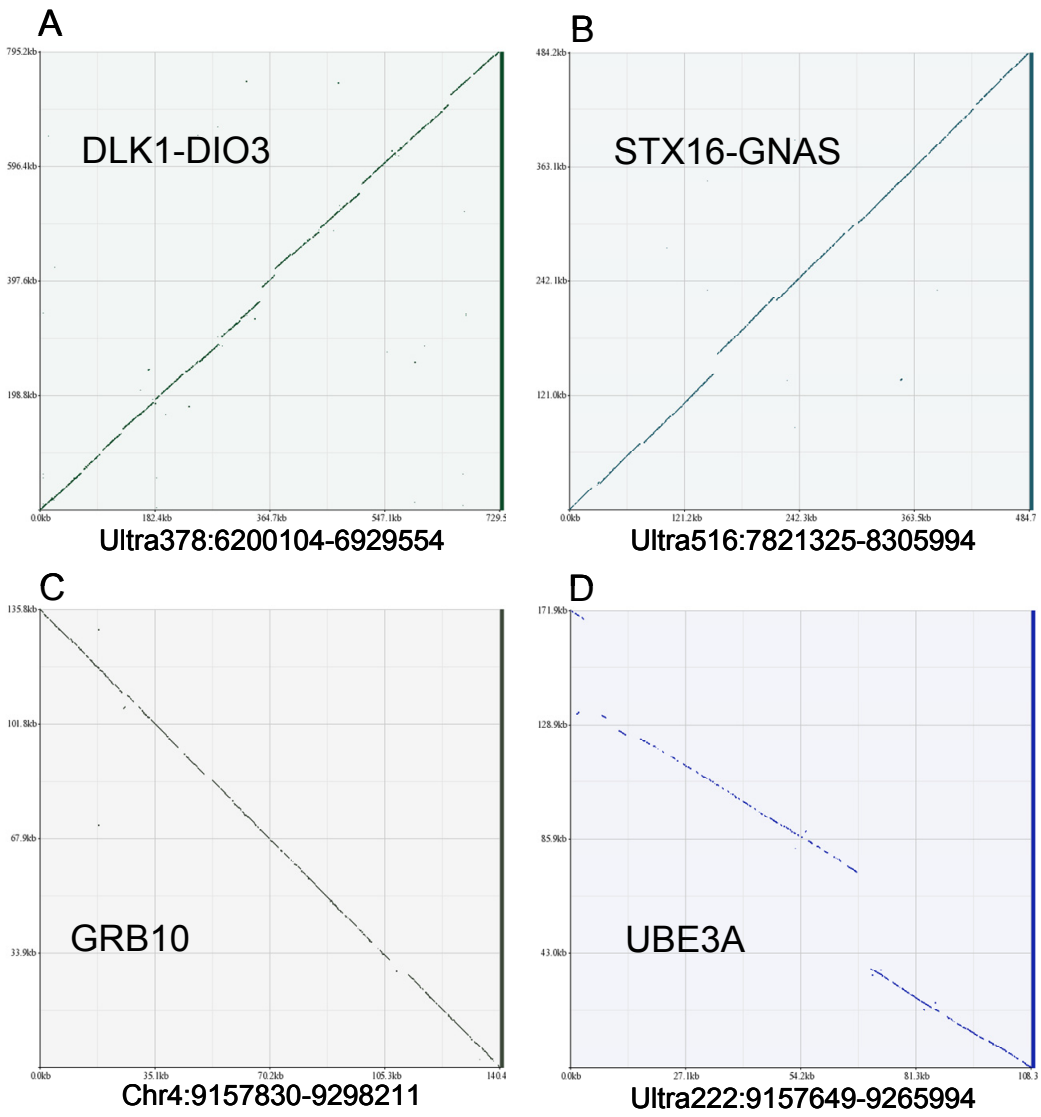

Figure IV.1. Dot-plots comparing platypus WGS contigs with finished sequences.

The platypus WGS sequences were exported from the ornAna1 (March 2007) assembly in the UCSC genome browser and compared with finished SAVOIR sequences using the dot-plot facility in the zPicture server (Ovcharenko et al. 2004a)**. For each indicated region WGS sequences are displayed on the x-axis and finished SAVOIR sequences on the y-axis. The bottom left to top right diagonals in panels A and B indicate sequences in the same orientation. The top left to bottom right diagonals in panels C and D indicate that the compared sequences are in opposite orientations.**

4.3 Multi-species and regional gene annotation

The following sections describe the gene contents of each sequenced region, orthologous to human chromosome 11p15.5, for each species (see section 4.3.5 for other SAVOIR regions). In all cases comparison is made with the human orthologous regions. Manual annotation of gene transcripts using experimental evidence and gene predictions was performed by Charles Steward (HAVANA group). A summary table of genes in the 11p15.5 region and the presence or absence of annotated orthologues is provided (Table IV-4). Only the official gene symbols are used in the text unless the gene is not listed in Table IV-4 (i.e. not present in human) in which case the full name is also provided in the text. The figures produced for each species (generated using the program genome_canvas, James Gilbert) feature a number of tracks including repeat and C+G contents of the sequences which are discussed in sections 4.5 and 4.6, respectively.

Table IV-4. Annotated human chromosome 11p15.5 genes and their orthologues.

Human gene name	Human symbol	*Mus spretus*	Wallaby	Platypus	Chicken
Keratin associated protein 5, members 1 to 6	*KRTAP5-1 to 6*	NA	Yes	NA	NA
Novel transcript	*CR626060#*	NA	Yes	NA	NA
Cathespin D	*CTSD*	NA	Yes	NA	NA
Synaptotagmin VIII	*SYT8*	NA	Yes	NA	NA
Troponin I type 2	*TNNI2*	NA	Yes	Yes	NA
Lymphocyte-specific protein 1	*LSP1*	NA	Yes	No	Yes
EnsEMBL novel	*Q8C494-like#*	NA	Yes *	No	No
Troponin T type 3	*TNNT3*	NA	Yes	Yes	Yes
Mitochondrial ribosomal protein L23	*MRPL23*	NA	Yes	NA	Yes
H19 untranslated mRNA	*H19*	Yes	Yes**	NA	No
Insulin-like growth factor 2	*IGF2*	Yes	Yes	Yes	Yes
IGF2 antisense	*IGF2AS*	Yes	No	No	No
Insulin	*INS*	Yes	Yes	Yes	Yes
Tyrosine hydroxylase	*TH*	Yes	Yes	Yes	Yes
Achaete-scute complex homolog-like 2 (Drosophila)	*ASCL2*	Yes	Yes	NA	Yes
Chromosome 11 open reading frame 21	*C11orf21*	No	No	NA	No
Tetraspanin 32	*TSPAN32*	Yes	Yes	NA	Yes
Novel transcript	*BC019904#*	No	No	NA	No
CD81 molecule	*CD81*	Yes	Yes	Yes	Yes
Tumour suppressing subtransferable candidate 4	*TSSC4*	Yes	Yes	Yes	Yes
Transient receptor potential cation channel, subfamily M, member 5	*TRPM5*	Yes	Yes	Yes	Yes
Potassium voltage-gated channel, KQT-like subfamily, member 1	*KCNQ1*	Yes	Yes	NA	Yes
KCNQ1 downstream neighbour	*KCNQ1DN*	No	No	NA	No
Cyclin-dependent kinase inhibitor 1C	*CDKN1C*	Yes	Yes	NA	Yes
SLC22A18 antisense	*SLC22A18AS*	No	No	NA	No
Solute carrier family 22, member 18	*SLC22A18*	Yes	Yes	Yes	Yes
Plekstrin homology-like domain, family A, member 2	*PHLDA2*	Yes	Yes	No	Yes
Nucleosome assembly protein 1-like 4	*NAP1L4*	Yes	Yes	Yes	Yes
Cysteinyl-tRNA synthetase	*CARS*	Yes	Yes	Yes	Yes
Oxysterol binding protein-like 5	*OSBPL5*	Yes	Yes	Yes	Yes
Chromosome 11 open reading frame 36	*C11orf36*	No	NA	No	NA
MAS-related GPR, member G	*MRGPRG*	Yes	NA	Yes	NA
MAS-related GPR, member E	*MRGPRE*	Yes	NA	Yes	NA
Zinc finger protein 195	*ZNF195*	No	NA	NA	NA

Official known gene names and symbols were taken from the NCBI Gene website except where marked with #. The presence of an orthologue is indicated in green and absence in red. NA, sequence not available. *, This novel transcript is discussed further in chapter V. **, The identification of wallaby *H19* is discussed in chapter VI.

4.3.1 Chicken (*Gallus gallus*)

Within this thesis the only orthologous imprinted gene region to be sequenced fully in chicken was a 1.2 Mb region of chromosome 5 with conserved synteny to the IC1 and IC2 domains of human and mouse. A partial analysis of much of the chicken sequences generated in this thesis has been performed by others (Paulsen et al. 2005). However, at the time of their analysis I had not completed the physical map and a gap between *KCNQ1* and *CDKN1C* genes existed in the chicken BAC contig. To enable a more comprehensive analysis of this region I subsequently closed this gap with the BAC sequences CR855369, CR855371 and CR855866. Analysis and annotation of the complete 1.2 Mb of chicken chromosome 5 sequence reveals the presence of all human orthologues for this region with the notable exceptions of *H19*, *IGF2AS*, *C11orf21*, *KCNQ1DN, SLC22A18AS* and two novel transcripts (Figure IV.2, Table IV-4). Four of these genes; *C11orf21*, *BC019904*, *KCNQ1DN* and *SLC22A18AS* appear to be specific to the human lineage and all are of unknown function. The absence of *H19* in chicken is significant because of the role of this non-coding RNA and associated regulatory elements in imprinting. The lack of imprinted gene expression of the neighbouring *IGF2* and *INS* genes in chicken would therefore appear to support a role for *H19* and/or local regulatory elements in the IC1 imprinting mechanism of mammals (Yokomine et al. 2005). In addition to orthologous human genes, 5 novel transcripts and a putative transcript were also annotated on the chicken sequence. Three of these (*AP003795.1*, *AP003795.2* and *AP003796.1*) lie between the *MRPL23* and *IGF2* genes where *H19* resides in mammals with imprinting. Was one of these the originator of the *H19* non-coding RNA? It would appear not because cross-species megaBLAST at NCBI did not reveal any matches between these chicken transcripts and the human genome. A further two novel transcripts (*AP003796.3* and

AP003796.4) lie between *IGF2* and *INS* genes and the putative transcript lies between *INS* and *TH* genes (Figure IV.2).

At the centromeric end of the cluster, between *CARS* and *OSBPL5* lies the tumour necrosis factor receptor superfamily, member 23 (*TNFRSF23*) gene (*BX640401.5*, Figure IV.2). The murine *Tnfrsf23* gene encodes a decoy receptor for tumour necrosis factor-related apoptosis-inducing ligand (TRAIL, Schneider et al. 2003). In mice, multiple copies of the *Tnfrsf* gene family exist at this locus (see below) and the solitary presence of *TNFRSF23* in chicken indicates that this was the founding member of the cluster (Bridgham and Johnson. 2004). Interestingly, no *Tnfrsf* orthologues are found at human 11p15.5 and therefore appear to have been lost in the human lineage. Other human TRAIL receptors and decoy receptors are clustered on the short arm of human chromosome 8 and therefore there may be some functional redundancy in this gene family.

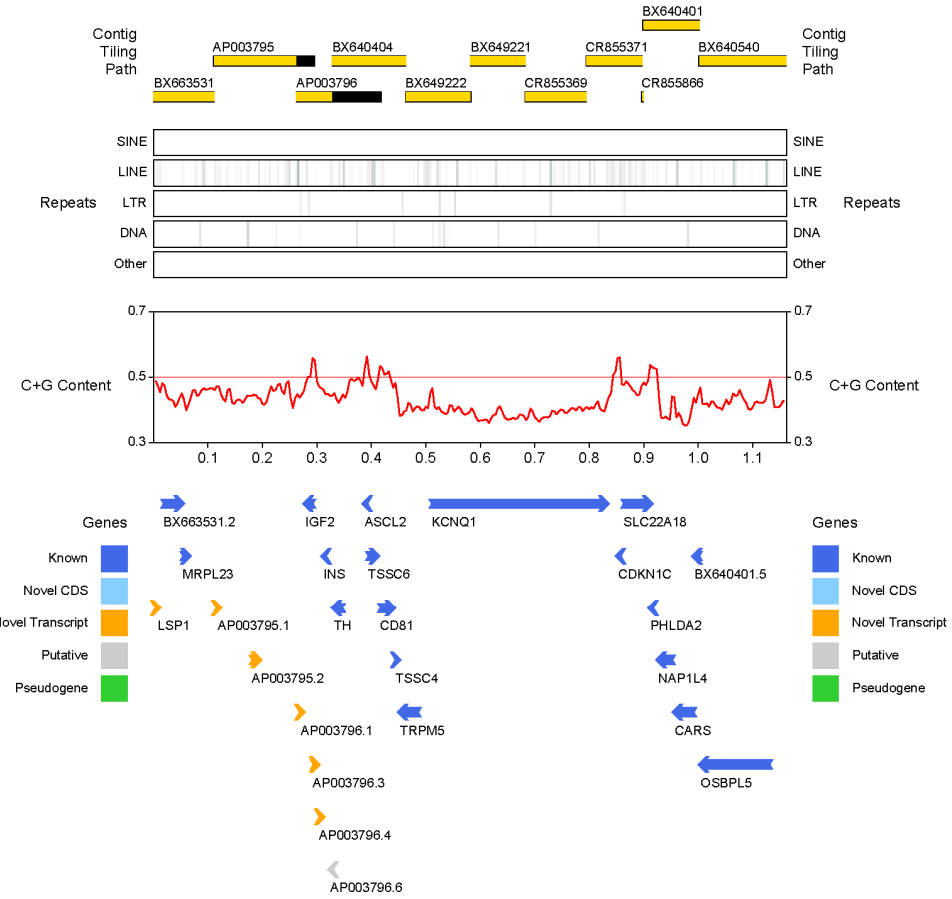

Figure IV.2. Sequence analysis and annotation of chicken chromosome 5.

Individual BAC clone sequence accession numbers in the tile path are illustrated at the top, from telomere (left) to centromere (right). The proportion of sequence contributing to the consensus is depicted in yellow. Redundant sequences not included are shown in black. Five classes of repeat elements in the sequence contig are indicated below the tile path. A plot of the C+G sequence content is shown in red. The horizontal pink line represents 50%. The scale in Mb is provided beneath the C+G plot. Known (dark blue), novel CDS (light blue), novel transcript (orange) and putative transcript (grey) annotations are indicated. Arrow heads indicate the direction of transcription. Locus names for known genes are as provided in Table IV-4 with the exception of BX663531.2 (*TNNT3*), TSSC6 (*TSPAN32*) and BC640401.5 (*TNFRSF23*).

4.3.2 Wallaby (*Macropus eugenii*)

Seven distinct regions of the tammar wallaby genome have been mapped, sequenced (chapter III) and analysed. Annotation of the 1.5 Mb of wallaby chromosome 2p sequence encompassing the entire IC1 and IC2 domains and flanking sequences reveals the following. The order and orientation of the genes between *SYT8* and *OSBPL5* are identical between human and wallaby illustrating remarkable conservation (Figure IV.4). However, the flanking regions of the wallaby sequence indicate some rearrangements. At the telomeric end of this region is a cluster of novel keratin associated protein genes and pseudogenes. *KRTAP5* family members are required for hair formation (Yahagi et al. 2004). This cluster may not be complete, since more members may exist beyond the limit of sequence generated, but the available sequence indicates more members of this family than the 6 members in the orthologous human region. Furthermore, the existence of 9 *KRTAP5* pseudogenes also suggests multiple duplication events in the wallaby lineage and a rapidly evolving locus. Perhaps these duplicated *KRTAP5* genes are undergoing neofunctionalisation (evolving to perform novel functions)? In human, two clusters of highly similar *KRTAP5* genes are found at 11p15.5 and 11q13.5. Yahagi and colleagues have proposed a model in which a segmental duplication event gave rise to the two *KRTAP5* clusters on human chromosome 11. Their model predicts that at least seven *KRTAP5* genes and one pseudogene should have existed in the primitive cluster on the ancestral species prior to the duplication event (Figure IV.3, Yahagi et al. 2004). In the human lineage progressive gene loss resulted in the 6 *KRTAP5* family members we see at 11p15.5. Intriguingly, unlike all other *KRTAP5* genes at 11p15.5, the genes *KRTAP5-1* and *KRTAP5-6* are expressed in several tissues in addition to hair root (Yahagi et al. 2004) which would support neofunctionalisation. Following the initial proposed domain duplication

event it appears that further local duplications ensued in the wallaby lineage giving rise to the increased gene and pseudogene numbers. The wallaby annotation presented here appears to support the duplication model of Yahagi and colleagues. However, the annotated wallaby *KRTAP5* genes and pseudogenes are all in the same orientation (Figure IV.4) indicating that local inversions in the human lineage may have occurred since the last common ancestor with wallaby.

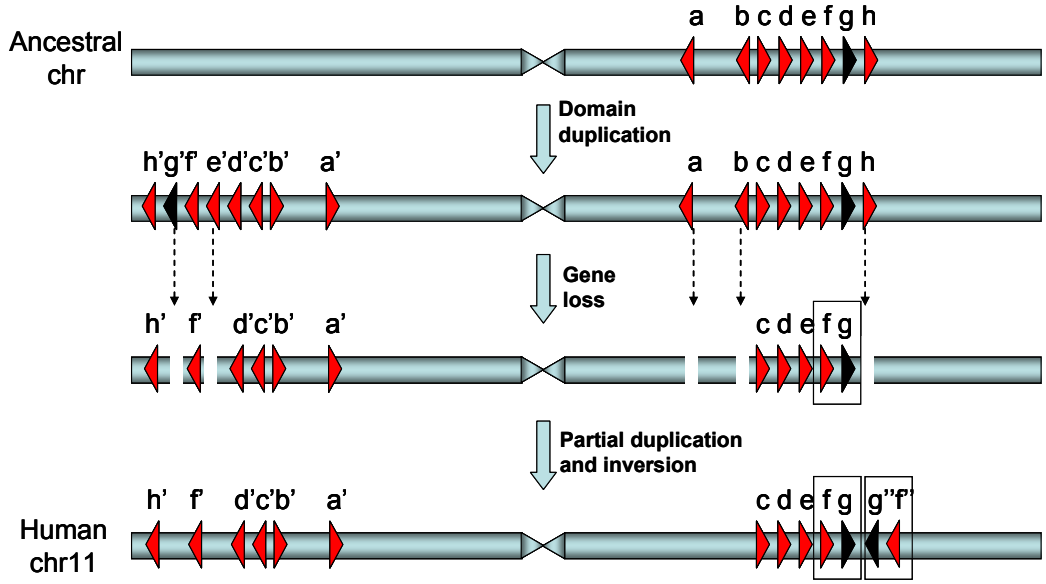

Figure IV.3. Postulated model for the evolution of *KRTAP5* family members.

Adapted from (Yahagi et al. 2004)**. Red triangles, *KRTAP5* protein-coding genes; black triangles, pseudogenes.**

Between the *CTSD* and *SYT8* genes in wallaby lies the 5' nucleotidase, cytosolic IB (*NT5C1B*) gene (Figure IV.4). The human orthologue lies on chromosome 2 band p24.2. From the sequences annotated here, and other available vertebrate genome sequences in the UCSC genome browser, it appears that *NT5C1B* has been translocated between *CTSD* and *SYT8* genes specifically in the wallaby lineage. Additional annotations in the wallaby not present in human include a ubiquitin-conjugating enzyme E2N (*UBE2N*) processed pseudogene, a regucalcin (*RGN*) processed pseudogene, a novel zinc finger C2H2 type domain containing protein

(*MEKBa-465N20.2*) and the death domain containing *TNFRSF23* gene (*MEKBa-465N20.5*) which, as noted above, is present in all species studied here except human.

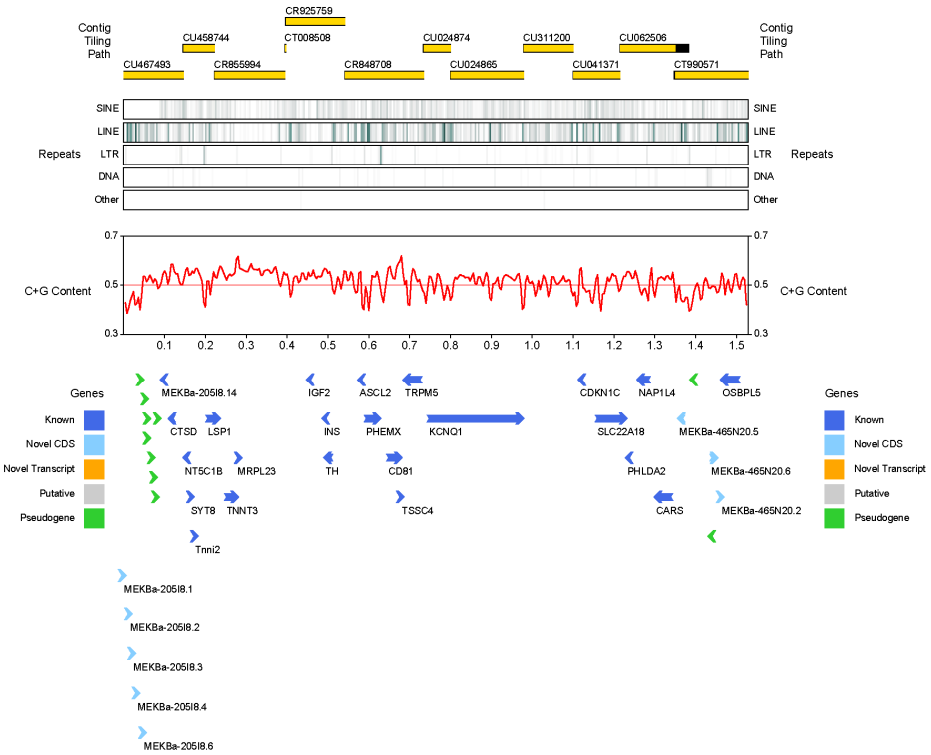

Figure IV.4. Sequence analysis and annotation of wallaby chromosome 2p.

The figure is oriented from telomere (left) to centromere (right). Features are described as in the legend to Figure IV.2. Locus names for known genes are as provided in Table IV-4 with the exception of *PHEMX* which has been re-named to *TSPAN32* and *MEKBa-205I8.14* which is an orthologue of the human novel gene (*CR626060*). The novel CDS genes MEKBa-205I8.1 to .4 and MEKBa-205I8.6 represent members of the KRTAP5 gene family. Nine KRTAP5 pseudogenes (clustered green arrow heads) lie at the telomeric end of the region.

4.3.3 Platypus (*Ornithorhynchus anatinus*)

In contrast to the 11p15.5 orthologous regions in other species, the sequencing in this region of platypus is incomplete because of difficulties in mapping BACs and

large deletions present in many of the mapped BACs (see chapter III and below for reasons why). However, the analysis and annotation of 762 kb of finished sequence from the IC1 and IC2 orthologous regions reveals the presence of known genes; *TNNI2*, *TNNT3*, *IGF2*, *INS, TH, CD81, TSSC4, TRPM5, SLC22A18, NAP1L4*, *CARS* and *OSBPL5* (Figure IV.5). The apparent absence of novel genes or transcripts annotated in platypus likely reflects the paucity of experimental evidence (e.g. ESTs and mRNAs) available for this species, together with the long branch length for platypus in mammalian phylogeny (Margulies et al. 2005a). This long branch length, as indicated by a high rate of substitutions per site, makes it difficult to use cross-species experimental evidence to annotate true platypus orthologues.

The 243,540 bp sequence contig harbouring *IGF2*, *INS* and *TH* genes generated here fully overlaps a 41 kb platypus sequence previously generated by others and used in comparative studies to investigate the mechanism of genomic imprinting at the *IGF2* locus (Weidman et al. 2004). Unfortunately, no available sequence extends across the region where *H19* and the IC1 regulatory elements might reside. From these studies it is therefore not possible to establish whether the lack of *IGF2* imprinting in monotremes (Killian et al. 2001) is a direct consequence of the absence of *H19* and/or IC1 regulatory elements.

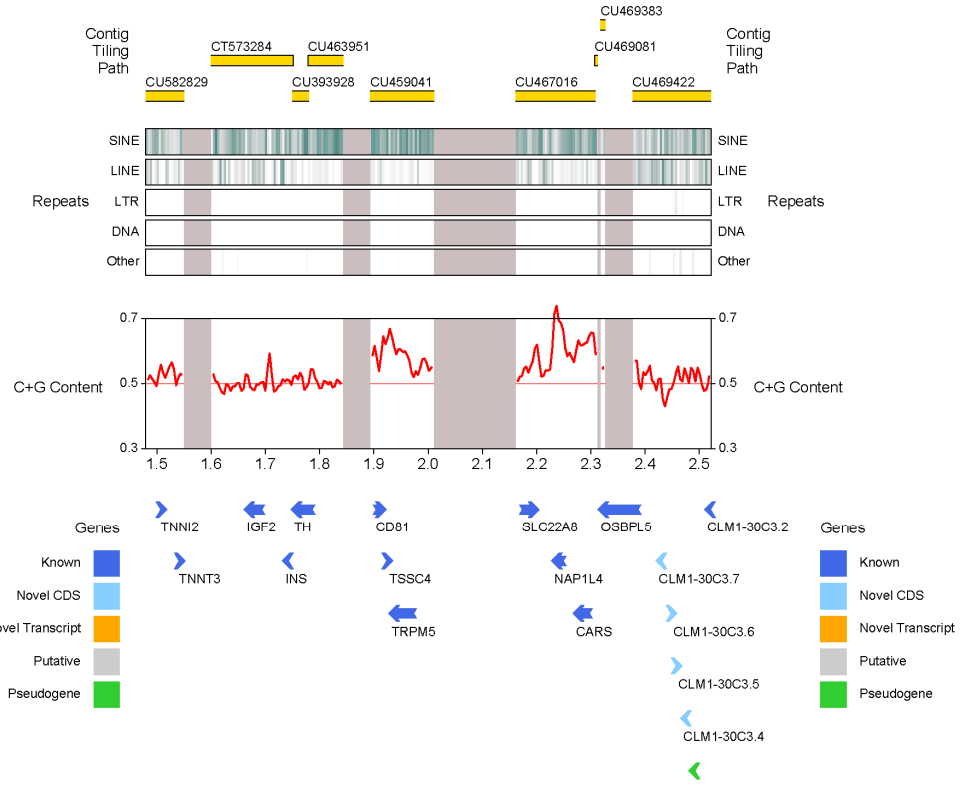

Figure IV.5. Sequence analysis and annotation of platypus chromosome 8p orthologous to human 11p15.5.

Gaps in the sequence are illustrated by grey vertical bars. The sizes of these gaps are approximations only. Genes pre-fixed CLM1-30C3.# are novel genes encoding 7 transmembrane receptor (rhodopsin family) domain containing proteins. See Figure IV.2 for legend to other features.

4.3.4 Western Mediterranean short-tailed mouse (*Mus spretus*)

The 1.4 Mb finished sequence for *Mus spretus* (chapter III) spans from non-coding transcript 1 (*Nctc1*) to 7 dehydrocholesterol reductase (*Dhcr7*) on distal chromosome 7 (qF5) and therefore encompasses both IC1 and IC2 sub-domains (Figure IV.6). As expected for species which last shared a common ancestor 1.5 Myr ago the gene order, orientation and exon-intron structure is well conserved between *Mus musculus* and *Mus spretus*. However, the sequence analysis and annotation does reveal some

surprises. The most striking of these is the apparent rapid and novel expansion of *Tnfrsf* genes between *Cars* and *Osbpl5* genes in the wild mouse (or loss in the domestic mouse; discussed further in section 4.4.3). Murine *Tnfrsf23* has been reported to be weakly imprinted (maternally expressed) in various organs with strongest expression in the trophoblast-decidua interface of placenta (Clark et al. 2002). This expression profile mirrors that of its nearest neighbour, *Osbpl5*, and suggests that the imprinting signal from IC2 (also known as KvDMR1, lying 387 kb away in intron 10 of *Kcnq1*) is sufficiently strong to silence the paternal allele in placental tissue.

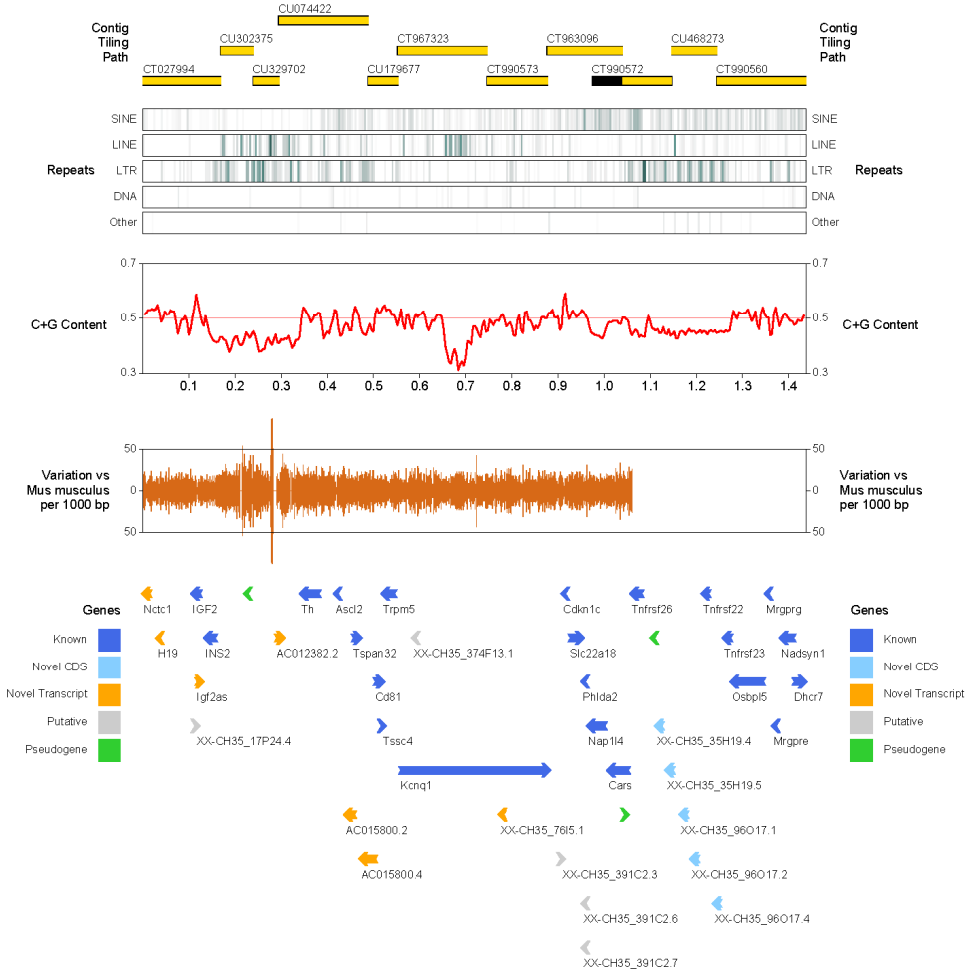

Figure IV.6. Sequence analysis and annotation of *Mus spretus* distal chromosome 7.

The number of sequence variants (both single nucleotide substitutions and insertion/deletion events) identified between a ssahaSNP (http://www.sanger.ac.uk/Software/analysis/ssahaSNP/) alignment of *Mus spretus* and *Mus musculus* sequences is plotted in 1000 bp windows. This plot ends at 1.06 Mb coincident with a tandem duplication event in *Mus spretus* (see text). Other features are as described in the legend to Figure IV.2.

4.3.5 Analysis and annotation of other SAVOIR regions

In addition to the orthologous 11p15.5 regions described above, 7 other regions were sequenced in both platypus and wallaby and an additional two regions in

platypus alone (chapter III and SAVOIR website [see below]). These sequences were assembled, analysed and annotated as described above. Figures summarising the gene, repeat and C+G contents for these regions can be found in Appendix D.

4.4 Multi-species comparative sequence analysis

Implicit in the study design to generate sequences for analysis of vertebrate orthologous imprinted regions is conserved synteny. Orthologous genes, by definition, originate from a single gene in the last common ancestor between two species. When two or more orthologous genes are physically linked (by sequence) in different species the common segment (or block) can be considered to define conserved synteny. As the annotation above has shown, the gene order and orientation is well conserved across vertebrate sequences and therefore large blocks of conserved synteny are easily defined and evolutionary breakpoints highlighted. This is exemplified by a dot-plot showing the conserved synteny between human chromosome 11p15.5 and mouse chromosome 7qF5 (Figure IV.7).

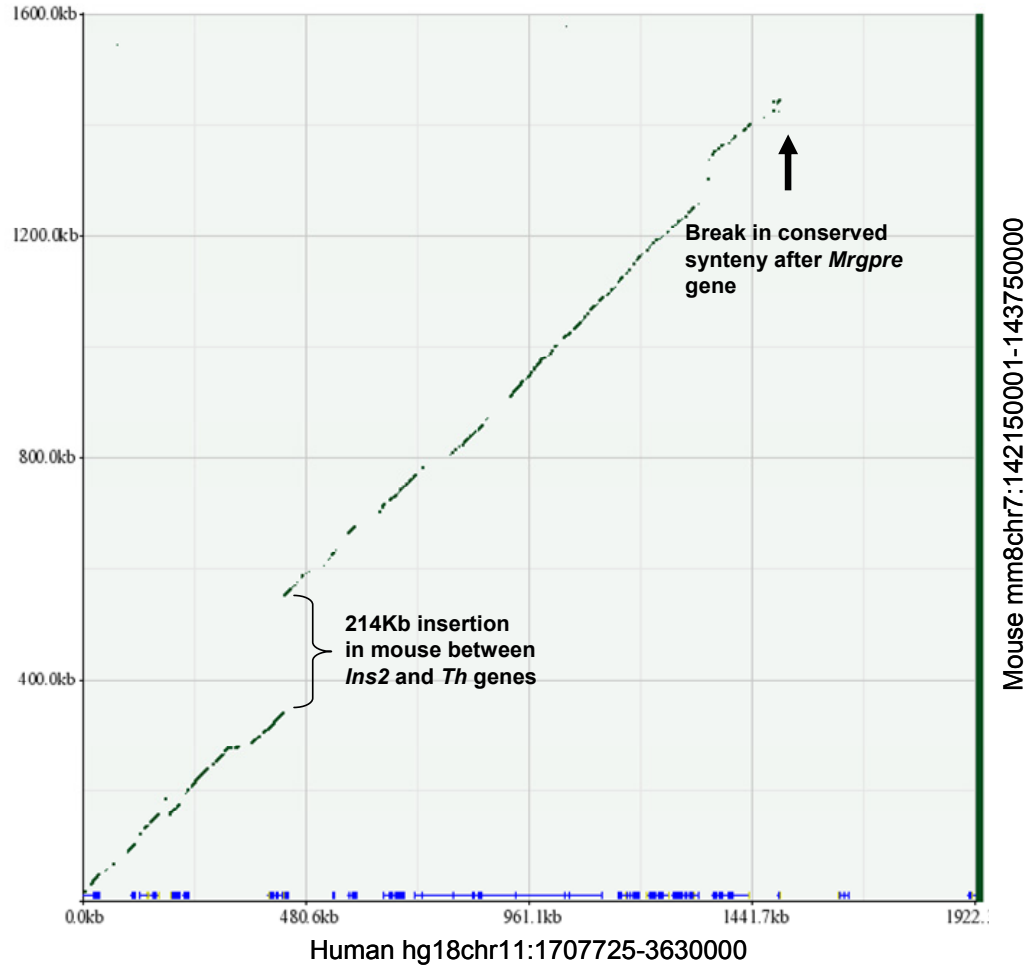

Figure IV.7. An extended block of conserved synteny between human chromosome 11p15.5 and mouse chromosome 7qF5.

A 1,922,276 bp human chromosome 11 (chr11) sequence (NCBI build 36, hg18) was aligned against a 1.6 Mb mouse chromosome 7 (chr7) sequence (NCBI build 36, mm8) using BLASTZ (Schwartz et al. 2003b) **from within the zPicture server** (Ovcharenko et al. 2004a)**. Within the dot-plot green diagonal lines denote ungapped alignments. On the x-axis human genes (exons, blue rectangles; introns, blue lines) are illustrated. The location of a 214 kb sequence insertion between mouse *Ins2* and *Th* genes is indicated. The location of an evolutionary breakpoint is marked by a black arrow.**

4.4.1 Localisation of an evolutionary breakpoint at 11p15.5

The multi-species sequence analysis and annotation described above confirms the extended block of conserved synteny, encompassing both IC1 and IC2 imprinted sub-domains, in human, mice, wallaby, platypus and chicken. The *OSBPL5* gene at the centromeric end of the IC2 domain (in human) is conserved in all species. Furthermore the *MRGPRG* and *MRGPRE* genes are conserved between human and mouse (Table IV-4). However, the human transcript *C11orf36* (NM_173590), telomeric of *MRGPRG*, may be specific to the human lineage as it is not found in the *Mus musculus* genome or the *Mus spretus* sequence reported here by BLAT or BLASTN analyses. The mouse gene *Nadsyn1* is positioned 11.3 kb telomeric of the *Mrgpre* gene (Figure IV.6) and yet in human this gene lies on the long arm of chromosome 11 band q13.4. Therefore, there is an evolutionary breakpoint mapping within 11.3 kb of sequence between *Mrgpre* and *Nadsyn1*. Centromeric of this breakpoint in human is the *ZNF195* gene. The chimp chromosome 11 sequence (March 2006) and rhesus macaque chromosome 14 sequence (January 2006) assemblies reveal the same gene order and orientation as human. The sequence of all other therian (placental and marsupial) mammals studied at the UCSC genome browser (rat, cat, dog, horse, cow and opossum) reveal the same gene arrangement as mouse. Therefore, the evolutionary breakpoint appears to have occurred early in the primate lineage. In chicken, lizards and fish the protein tyrosine phosphatase, receptor type, J (*PTPRJ*) gene lies adjacent to the *OSBPL5* gene indicating that an ancient and distinct evolutionary breakpoint exists in this region. This observation supports the notion that there are common regions harbouring chromosome breakpoints in vertebrate evolution which predicts that there are sites of genomic fragility (Pevzner and Tesler. 2003).

The mechanism by which chromosome breaks are initiated and subsequently fixed in evolution is currently unknown. However, it has been proposed that breakpoints may be associated with DNA containing repetitive sequences such as tandem repeats, gene clusters or segmental duplications (Bailey et al. 2004, Choi et al. 2006, Murphy et al. 2005, Puttagunta et al. 2000, Ruiz-Herrera et al. 2006, Stankiewicz et al. 2001). Analysis of the mouse 11.3 kb sequence containing the breakpoint reveals a C+G content of 49% and a relatively high proportion of SINE (16.79%) and simple repeat (5.76%) elements, most of which are TC-rich sequences clustered at both ends of a 1.4 kb region (Figure IV.8). It is interesting to speculate that these repeats may be underlying the breakpoint mechanism. Analysis of a 20 kb human sequence centromeric of *MRGPRE* indicates a high density of interspersed repeats (66.58%) but very few simple repeats and therefore the breakpoint mechanism is perhaps more likely to be mediated through interspersed repeats. Improvements in the ability to accurately align inter-species non-coding sequences and/or the identification of a primate species in which the gene order is the same as that in non-primate therians will be required to refine the breakpoint further and elucidate the underlying molecular mechanism.

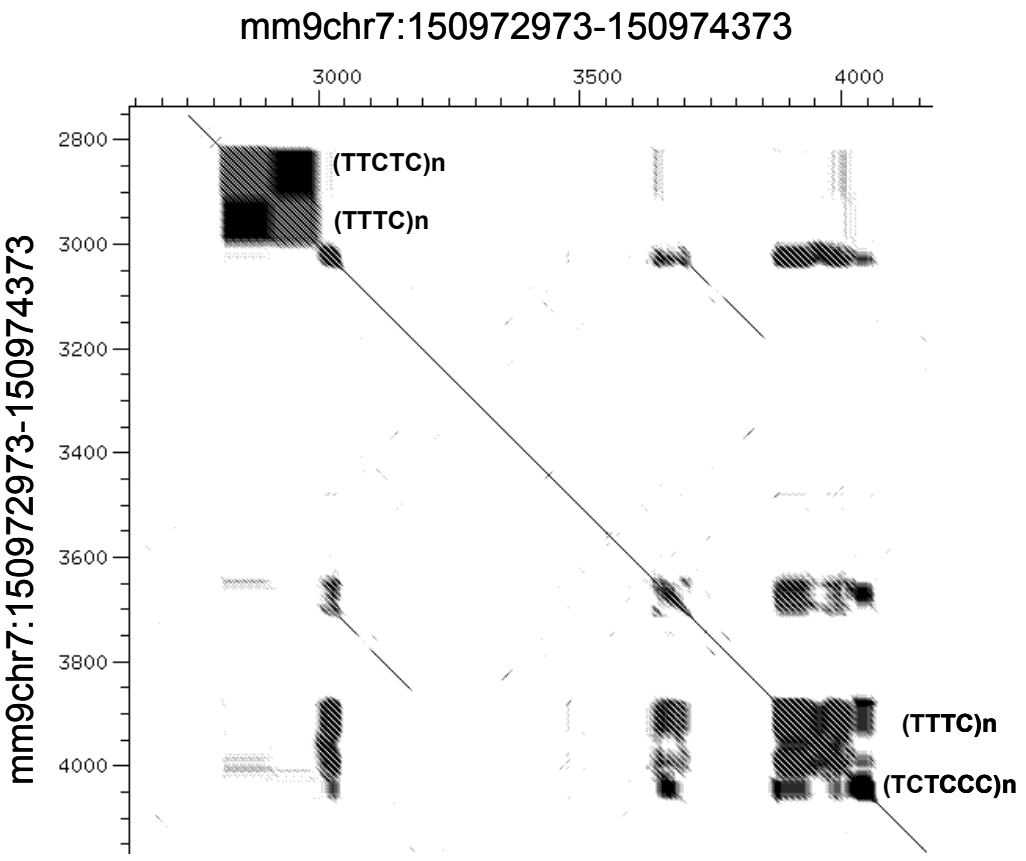

Figure IV.8. *Mus musculus* self–self dot-plot in the distal chromosome 7 evolutionary breakpoint region.

The dot-plot shows a 1.4 kb mouse sequence; Assembly mm9 (NCBI build 37, July 2007), position chr7:150972973-150974373. TC-rich simple repeats at both ends of the sequence are indicated. The dot-plot was created using Dotter software (Sonnhammer and Durbin. 1995)**.**

4.4.2 Broad scale finished sequence comparisons

The availability of finished sequence for mouse (*Mus musculus*), *Mus spretus*, wallaby and chicken in the region of conserved synteny with human 11p15.5 enables a comparison of genome expansion or contraction relative to human. As the platypus and opossum sequences in this region are incomplete these species are not included in this analysis. For all other species sequences spanning *IGF2* to *OSBPL5* genes were compared with human (Table IV-5). For both murine species and wallaby the *IGF2* to *OSBPL5* interval is larger than that of human. The corresponding chicken

interval is 23% smaller. However, when taking the genome sizes into account (using data from genomesize.com) all species, with the possible exception of tammar wallaby for which the genome size is currently unknown, have a proportional increase in the size of this region. Indeed the chicken genome is approximately one third the size of the human genome and yet the chicken *IGF2-OSBPL5* region is 77% that of the human (Table IV-5). The conserved synteny and apparent resistance to sequence loss raise the questions; is there strong evolutionary selection pressure to maintain the IC1-IC2 domain structure? And if so what might those pressures be? In mice and wallaby imprinting of genes in the region may provide the selection, but, there is no evidence of imprinting in chicken (O'Neill et al. 2000, Yokomine et al. 2005). Of course, across the whole 11p15.5 orthologous regions there may be different selection pressures at work. To gain a better understanding of this I refined the analysis between genes by plotting intergenic distances (Figure IV.9).

Table IV-5. Relative genomic sizes between *IGF2* to *OSBPL5* genes.

Species	From	To	Interval Size (bp)	Fraction Of human interval	C-value (pg) [ratio to human]
Human (hg18chr11)	2,113,329	3,106,954	993,626	1.00	3.50 [1.00]
Mouse (mm9chr7)	149,841,715	150,927,628	1,085,914	1.09	3.28 [0.94]
Mus spretus (chr7)	113,814	1,330,144	1,216,331	1.22	3.68 [1.05]
Wallaby (chr2p)	458,282	1,504,068	1,045,787	1.05	3.13-5.58* [0.89-1.59]
Chicken (chr5)	293,183	1,063,032	769,850	0.77	1.25# [0.36]

Coordinates from base 1 of the *IGF2* CDS to base 1 of the *OSBPL5* CDS are provided. For human and mouse these coordinates were extracted from the UCSC genome browser assemblies as indicated; hg18 (NCBI Build 36), mm9 (NCBI build 37). For *Mus spretus*, wallaby and chicken HAVANA annotations were used. Haploid genome sizes (C-values) were obtained from http://www.genomesize.com where available. pg, picograms. NA, not available. #, C-value for *Gallus domesticus* was used. *, a range of wallaby C-values is provided, however, the value for tammar wallaby is unknown.

The observed expansions are not uniformly distributed (Figure IV.9). Murine sequences indicate a large (214 kb) insertion between *Ins2* and *Th* genes (orthologues of human *INS* and *TH* genes, respectively, Figure IV.7). Interestingly, deletion of this region in mouse models has no discernible effect on the imprinting of neighbouring genes or other detectable phenotypic consequence (Shirohzu et al. 2004). Sequence analysis of the murine 214 kb expansion is largely the result of interspersed and tandem repeats which occupy 77% of this sequence. Of the interspersed repeats 23.5% are LINE elements and 22% are LTR elements, representing double the average for the full 1.6 Mb IC1 and IC2 containing region. The repetitive nature of the *Ins2-Th* intergenic region is demonstrated by the dot-plot comparison of *Mus musculus* and *Mus spretus* sequences (Figure IV.11). Whether this repeat-rich region serves any function, such as isolating the neighbouring IC1 and IC2 domains, is unclear. However, it has been noted that the human interval between *TH* and *ASCL2* genes also contains retroelements leading to the suggestion that the absence of retroelements in the non-imprinted chicken region may support a functional role in imprinting (Yokomine et al. 2005).

Whilst the finished sequence for platypus is not contiguous between *IGF2* and *OSBPL5* genes there are three finished sequence contigs spanning the genes: *IGF2* to *TH*, *CD81* to *TRPM5* and *NAP1L4* to *CARS*. The *INS* to *TH* interval in the (non-imprinted) platypus sequence is 46 kb and therefore, whilst smaller than the murine expansion, is larger than the human and wallaby intervals of 11 kb and 17 kb, respectively. The correlation of this sequence expansion with imprinting therefore seems doubtful and rather, perhaps, reflects the tolerance of mutation in this region of decreased functional constraint.

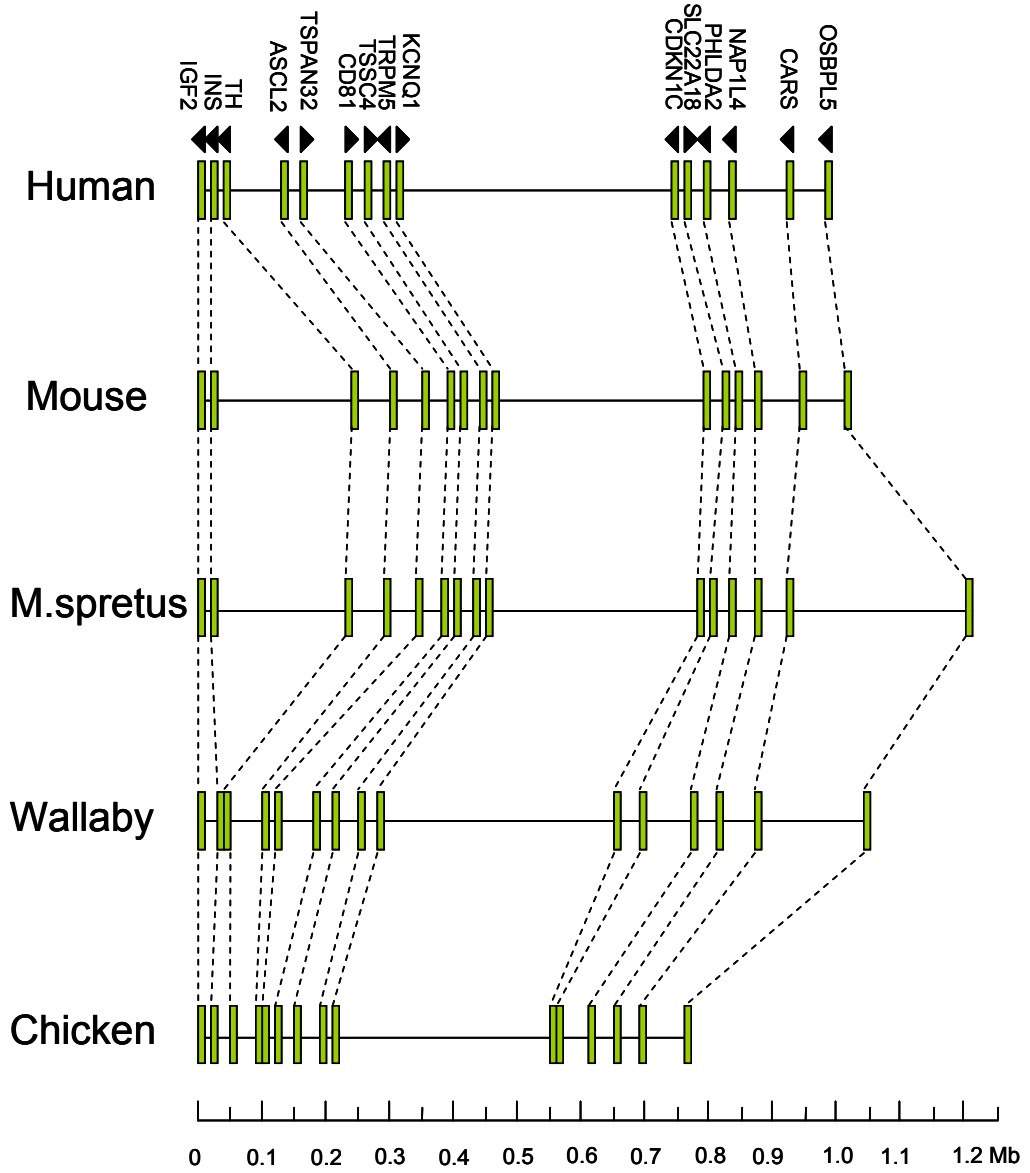

Figure IV.9. Comparison of the genomic structures between *IGF2* and *OSBPL5* genes.

For each indicated species the CDS start position for each gene is illustrated by green vertical bars. The arrow heads indicate direction of transcription. The scale is provided along the bottom in Mb.

The murine species and wallaby reveal an additional expansion between *CARS* and *OSBPL5* genes relative to human. This is the result of tumour necrosis factor receptor superfamily (*Tnfrsf*) genes, members of which are present in mice, wallaby and chicken but have been lost in the human lineage. In chicken only a single copy

of *Tnfrsf23* exists and is therefore thought to be the ancestral gene (Bridgham and Johnson. 2004). Orthologues of *TNFRSF23* are quite divergent across vertebrates as revealed by ClustalW alignment of amino acid sequences. However, they do share a conserved arrangement of multiple cysteine residues (Figure IV.10). These cysteine residues are characteristic of the extracellular ligand-binding domains of other TNF receptor family members (Marsters et al. 1992). The high level of divergence, elsewhere in the alignment, may be related to species-specific selective pressures on the adaptive immune system.

Tandem duplication of the *Tnfrsf23* locus in *Mus musculus* has given rise to the *Tnfrsf22* gene and a third family member, *Tnfrsf26* (Engemann et al. 2000, Schneider et al. 2003). All three family members are protein coding, and two (*Tnfrsf22* and *Tnfrsf23*) encode decoy receptors for TRAIL (Schneider et al. 2003). In *Mus spretus* further segmental duplication of the *Tnfrsf22* and *Tnfrsf23* genes has greatly expanded the region (Figure IV.11). An additional 6 transcripts, including 5 potentially protein coding genes (Novel CDS) and 1 pseudogene, have been annotated (Figure IV.6). It is possible that this region is polymorphic and that the single *Mus spretus* individual sequenced represents just one haplotype of a copy number variable region. Additional individuals would need to be sequenced to establish if this is the case.

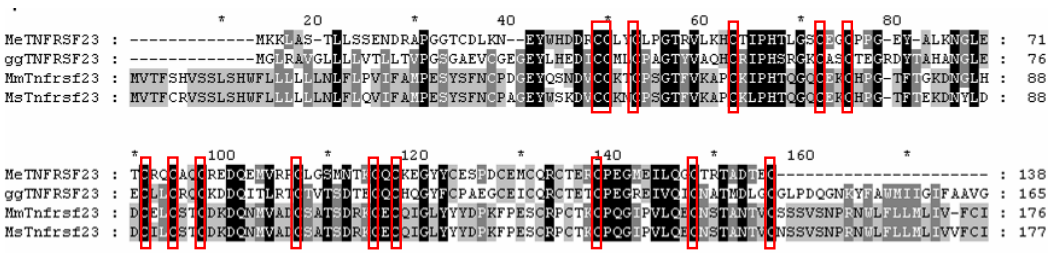

Figure IV.10. Amino acid sequence alignment of TNFRSF23 orthologues.

Wallaby (MeTNFRSF23), chicken (ggTNFRSF23), *Mus musculus* (MmTnfrsf23) and *Mus spretus* (MsTnfrsf23) protein sequences were aligned in ClustalW 1.83 (http://www.ebi.ac.uk/Tools/clustalw/index.html) and viewed using GeneDoc software (http://www.nrbsc.org/gfx/genedoc/). Amino acids are shaded according to their level of conservation; black, conserved in all 4 species; dark grey, conserved in 3 species and light grey, conserved in 2 species. Conserved cysteine residues are boxed in red.

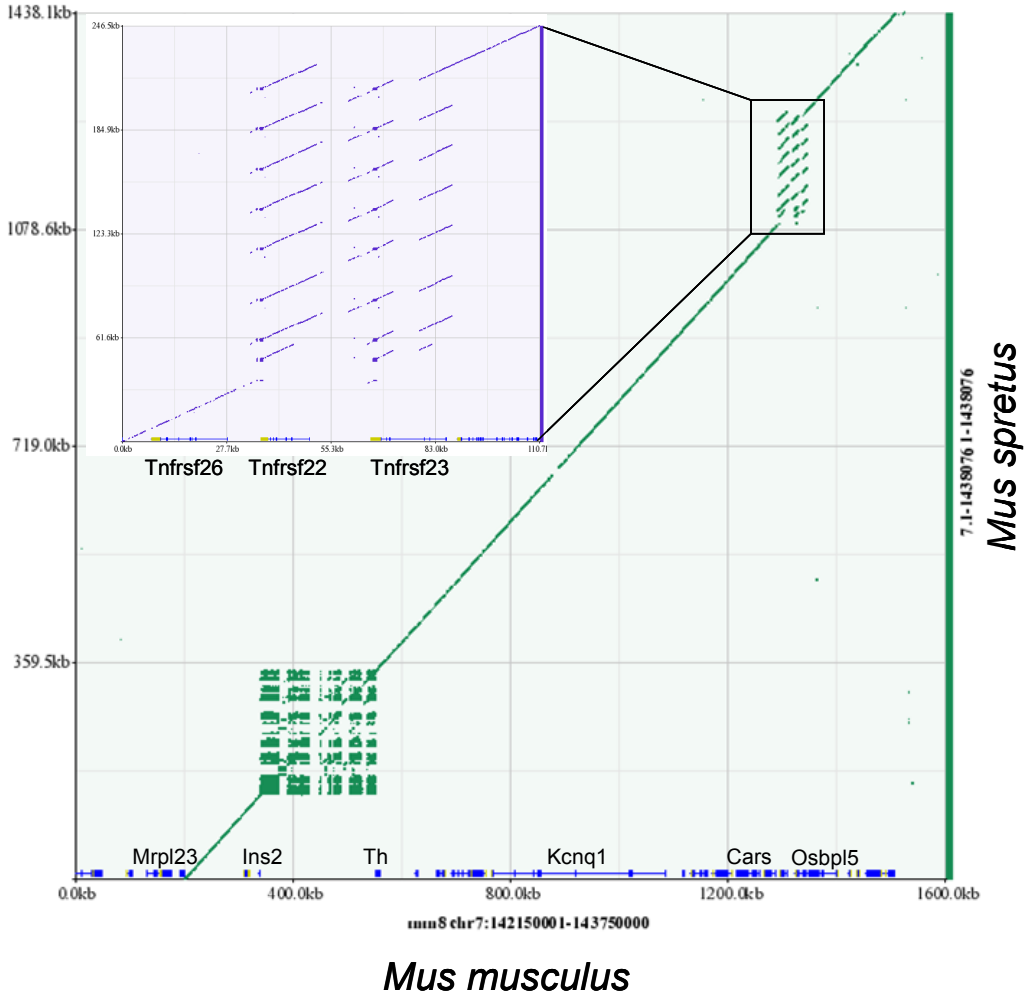

Figure IV.11. Dot-plot of *Mus musculus* and *Mus spretus* sequences in the IC1 and IC2 domains. **The green diagonal line represents the extent of pairwise sequence alignment (i.e. one-to-one orthology between the two species). Both sequences were masked for interspersed repeats but not tandem repeats from within the zPicture server, hence the observed repeats between *Ins2* and *Th* genes. The boxed region has been expanded to reveal the tandem duplication of *Tnfrsf* family members. On the x-axis *Mus musculus* genes (exons, blue rectangles; introns, blue lines) are illustrated. The y-axis label is on the right.**

4.4.3 Fine scale sequence comparisons - Sequence variant discovery between *Mus spretus* and *Mus musculus* species

To establish the imprinting status of genes generally relies upon the prior identification of polymorphisms (restriction fragment length polymorphisms (RFLPs) or SNPs) to distinguish parental alleles. Genome-wide polymorphism discovery efforts in mice suffer from low density because of the inbred strains typically used. To my knowledge there has not been a systematic effort to identify sequence variants (polymorphisms) between domestic and wild mice including the entire IC1 and IC2 domains and as a consequence the imprinting status of some murine genes is unknown (Engemann et al. 2000). This, and the extensive use of the experimental *Mus musculus-spretus* hybrid (SD7) to follow allele specific effects, provided the motivation for sequencing the entire IC1 and IC2 domains of *Mus spretus*. Alignment of 1,064,899 bp of *Mus musculus* (mm8chr7:142352177-143417075) and 1,060,554 bp *Mus spretus* finished sequence spanning from *Nctc1* to *Cars* was performed using ssahaSNP (http://www.sanger.ac.uk/Software/analysis/ssahaSNP). The sequences aligned deliberately excluded the duplicated *Tnfrsf23* genes and pseudogenes which were very likely to give mis-alignments between paralogous matches and not true orthologous alignments. As a result, paralogous sequence variants (PSVs) would be mistaken for true sequence variants (Estivill et al. 2002). Post-processing of the ssahaSNP output file, containing identified single nucleotide substitutions and insertion or deletion (indel) events, was performed using a perl script (ssahaParse, Dave Beare). A total of 16,831 sequence variants (14,156 single nucleotide variants (SNVs) and 2675 indels) were obtained equating to a sequence variant rate of 1 per 63 bp on average. The number of variants per 1000 bp is plotted on Figure IV.6. From this plot it is evident that the variants are not evenly distributed. In particular

there is a region at approximately 280 kb which is apparently rich in variants between the murine species. However, this region lies within the previously discussed 214 kb murine-specific expansion between *Ins2* and *Th* genes. Indeed the dot-plot of murine sequences indicates the repetitive nature of this interval (Figure IV.11) and suggests that many of the variants identified in this interval may be PSVs. Other than this region the level of variants elsewhere are within reasonable bounds.

The density of sequence variants identified here should be of great use to those wishing to discriminate between parental alleles in studies using the congenic SD7 mouse line (described in chapter I). The imprinting community working on either IC1 or IC2 domains in mice commonly work with mice crossed between C57BL/6J and SD7 strains. Therefore for all regions in which a SNP has been identified between *Mus musculus* and *Mus spretus* the parental origin of that region can be determined. Applications making use of allelic discrimination include embryonic allelic expression, to address, for example, the question: 'when does imprinted expression start'? When combined with ChIP these studies can also determine which histone modifications are present on each allele with a given expression? Allele-specific PCR can also be used to distinguish differential methylation and higher order chromatin structure. All applications have a different requirement for SNPs. For allele specific expression the SNP must be located within an exon, for methylation analysis the SNP should be near a CpG island and in chromosome conformation capture (3C) the SNP should be positioned near the restriction site used. Having a catalogue of all sequence variants in the IC1 and IC2 domains therefore enables the most appropriate choice of SNP for a given application. All 16,831 sequence variants identified have been provided to collaborating groups and will be made publicly available following publication.

4.5 Repeat contents of sequences

As described above, despite overall conservation of gene content, order and orientation, there is variation between the size of genomic regions (Table IV-5, Figure IV.9). The genomic content of coding sequences does not account for the observed variation in size. I therefore looked at non-coding regions for evidence of molecular events that may have contributed to the observed genome compression or expansion. The most likely explanation for size variation comes from the evolutionary insertion of transposable elements giving rise to interspersed repeats.

To test this assumption RepeatMasker (RepBase update 24th September 2007) was run on each finished sequence using the appropriate species option (see chapter II). Repeat and C+G contents of the sequences were obtained from the table output file from RepeatMasker and are shown in Table IV-6.

Table IV-6. Repeat and C+G contents of multi-species sequences generated here.

Species name (Common name)	Orthologous human region (Genes spanned)	Sequence (bp)	Bases masked (bp) [%]	C+G content (%)	CpG number [%]	Predicted number of CpG islands
Mus spretus (Algerian mouse)	11p15.5 (Ctsd-Osbpl5)	1438076	458755 [31.90]	47.26	12583 [1.7]	33
Macropus eugenii (Tammar wallaby)	11p15.5 (CTSD-OSBPL5)	1528894	473278 [30.96]	51.21	16451 [2.2]	40
Monodelphis domestica (South American opossum)	11p15.5 (IGF2-INS)	136229	32134 [23.59]	61.52	4784 [7.0]	39
Ornithorhynchus anatinus (Platypus)	11p15.5 (CTSD-OSBPL5#)	762175	536742 [70.34]	53.77	30272 [7.9]	319
Gallus gallus (Chicken)	11p15.5 (CTSD-OSBPL5)	1162135	58415 [5.03]	43.18	13359 [2.3]	28
Macropus eugenii (Tammar wallaby)	20q13.3 (STX16-GNAS)	529114	230368 [43.54]	37.34	2593 [1.0]	11
Ornithorhynchus anatinus (Platypus)	20q13.3 (STX16-GNAS)	484161	158570 [32.75]	41.62	5034 [2.1]	16
Macropus eugenii (Tammar wallaby)	12q13 (SLC38A2/4)	319152	174824 [54.78]	36.32	1338 [0.8]	5
Ornithorhynchus anatinus (Platypus)	12q13 (SLC38A2/4)	331739	213804 [64.45]	43.86	4558 [2.7]	22
Macropus eugenii (Tammar wallaby)	14q32 (DLK1-DIO3)	1674705	1030874 [61.56]	38.11	6218 [0.7]	10
Ornithorhynchus anatinus (Platypus)	14q32 (DLK1-DIO3)	795237	393524 [49.49]	44.62	8217 [2.1]	30
Macropus eugenii (Tammar wallaby)	7p11.2-p12 (GRB10)	164317	64640 [39.34]	38.65	881 [1.1]	1
Ornithorhynchus anatinus (Platypus)	7p11.2-p12 (GRB10)	135754	54128 [39.87]	43.51	2048 [3.0]	2
Macropus eugenii (Tammar wallaby)	6q26 (IGF2R)	159825	62821 [39.31]	37.3	911 [1.1]	3
Ornithorhynchus anatinus (Platypus)	6q26 (IGF2R)	383712	217394 [56.66]	50.02	14266 [7.4]	135
Macropus eugenii (Tammar wallaby)	19p13.2 (DNMT1)	159867	54979 [34.39]	42.69	1201 [1.5]	2
Ornithorhynchus anatinus (Platypus)	19p13.2 (DNMT1)	206126	78958 [38.31]	54.26	6376 [6.2]	54

Repeat and C+G contents were obtained from RepeatMasker (version open-3.1.8, RepBase update 20070924). The number of CpG dinucleotides in a given sequence was determined using the perl script CpGcount (Dave Beare). CpG islands were predicted using the program newcpgreport (EMBOSS).

Recent or remnant copies of repeats resulting from insertion events vary widely between the vertebrates studied. For example, in the 11p15.5 orthologous region, 70% of available platypus sequences and only 5% of the chicken region are repetitive (Table IV-6). The chicken repeat content is consistent with the findings of

others (Hillier et al. 2004, Thomas et al. 2003). The two-fold relative expansion of this chicken chromosome 5 region can therefore not be accounted for by repeat elements. This contrasts with the observations of Thomas and colleagues who sequenced the greater cystic fibrosis transmembrane (*CFTR*) gene region in multiple vertebrate species and showed a correlation between repeat content and size of region (Thomas et al. 2003). In those regions for which only wallaby and platypus orthologous sequences were generated, there are regions with similar repeat contents (e.g. *GRB10* and *DNMT1* regions). However, all other regions have very different repeat contents (Table IV-6). The repeat distributions in human, mouse, wallaby and platypus were investigated further to see whether mammalian repeat contents and genome sizes are correlated.

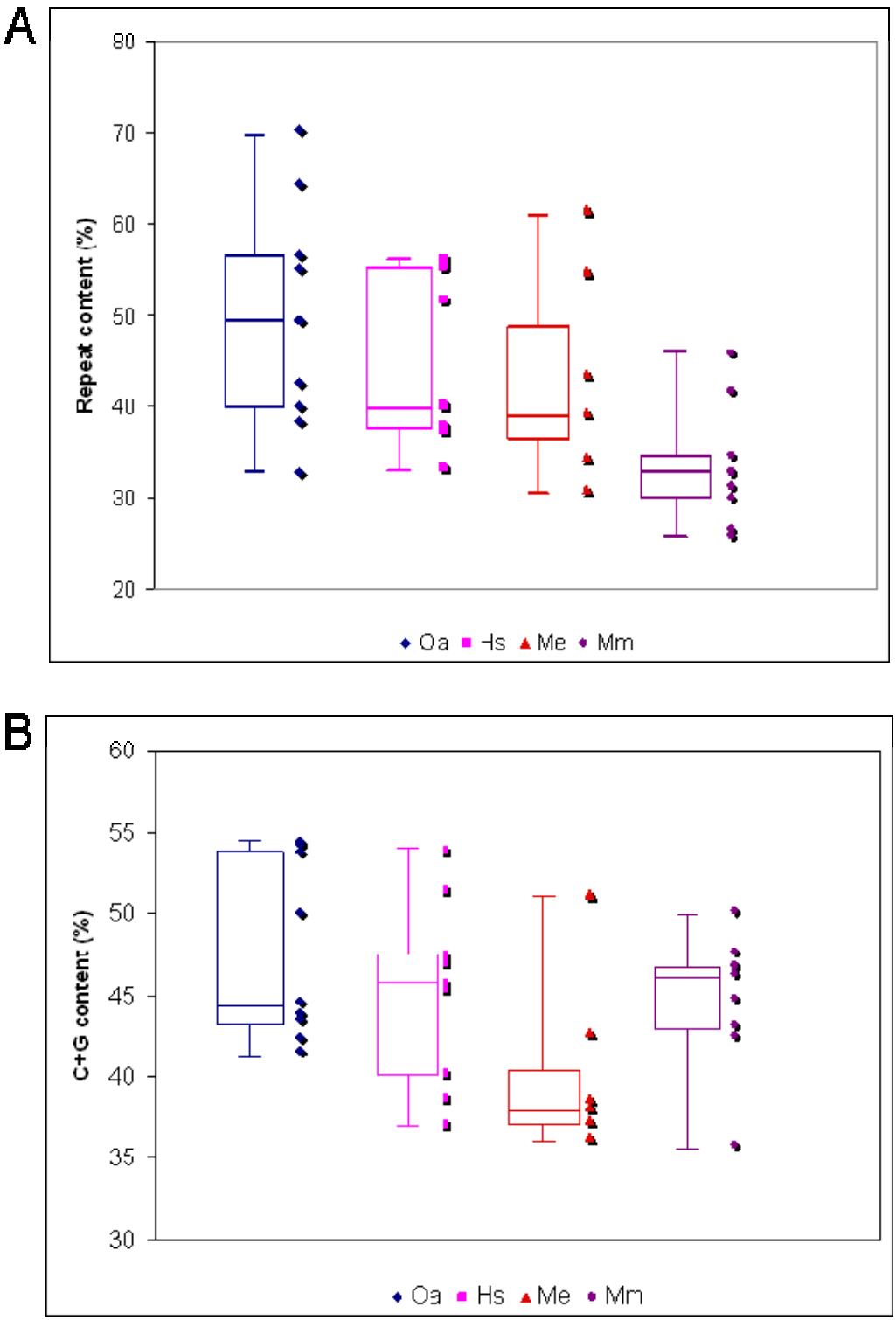

Figure IV.12. Box-and-Whisker plots of repeat and C+G contents in the SAVOIR regions.

The repeat contents (A) and C+G contents (B) for each indicated species is shown; Oa, platypus; Hs, human; Me, wallaby; Mm, mouse . Each plot corresponds to data from 9 distinct genomic regions per indicated species, except for wallaby which has finished sequence for 7 regions only. Each box-and-whisker plot reveals (from bottom to top) the minimum, 25th percentile, median, 75th percentile and maximum statistics for the data. The colour matched raw data is plotted to the right of the box-and-whisker plots.

The box-and-whisker plots (Figure IV.12) provide two common measures of variation, the range and the interquartile range. It is immediately apparent that the mouse repeat contents are much more uniform across the 9 SAVOIR regions with 50% of the data (interquartile range, within the box) having a repeat content of 30-35% (Figure IV.12A). The median for all mouse data is 32.8% which is significantly lower than the reported 41.2% for the whole mouse genome (Waterston et al. 2002). This indicates that the imprinting domains, in mouse, are either selecting against the acquisition of repeat elements or are actively deleting these elements. In contrast, there is a much wider distribution of repeats for human, platypus and wallaby regions and the repeat contents for these species are significantly higher with median values of 40.1%, 49.5% and 39.3%, respectively. As for the mouse, the human repeat content is lower than the genome average (47.6%, International Human Genome Sequencing Consortium. 2001). For both human and wallaby the repeat distributions are heavily skewed upwards indicating that there are regions of unusually high repeat content. The opposite appears to be true for the platypus which has the greatest range, highest median and, as noted above, the most extreme repetitive region sequenced. The reported genome size for platypus (3.06pg) is smaller than that of human (3.50pg, http://www.genomesize.com) and yet the repeat content is higher, at least across the regions sequenced here. Overall the data would not appear to support the link between repeat content and genome size.

4.5.1 Orthologous 11p15.5 region repeats

So what do the repeat contents of amniotes look like by class? To determine this, relative contents of different interspersed repeat types were plotted for each species for the orthologous 11p15.5 region (Figure IV.13).

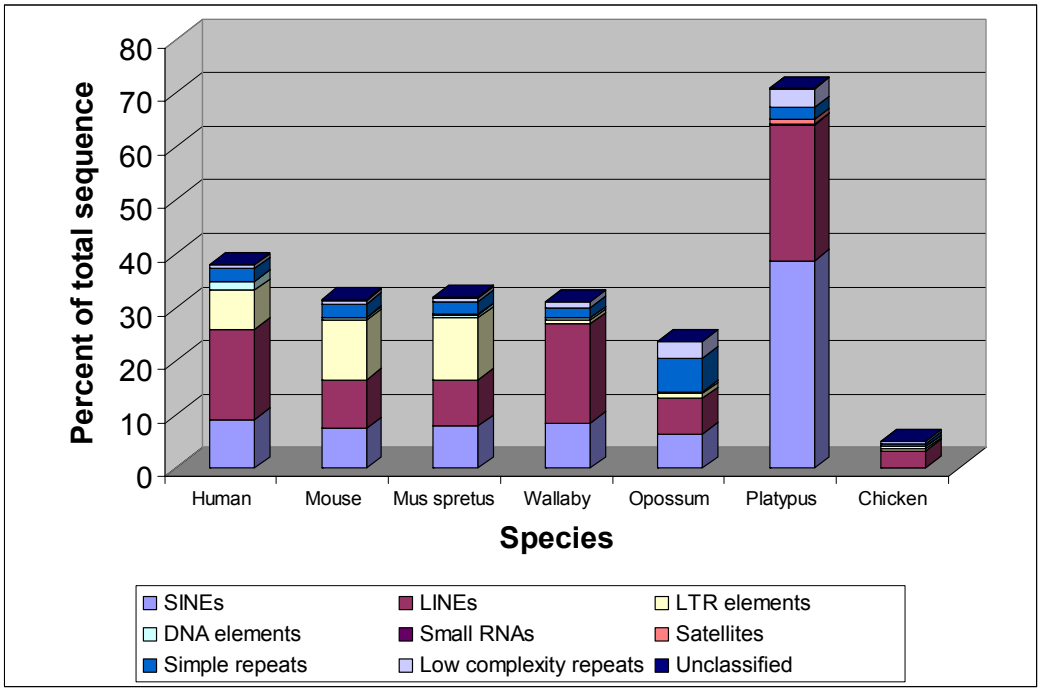

Figure IV.13. Relative content of repeat types within the 11p15.5 orthologous regions.

The repeat contents of species on the x-axis were obtained from the table output file from RepeatMasker. The colour-coded key to repeat class is provided at the bottom.

The percentage of total sequence occupied by repeat element is similar for this region between human, mice and wallaby. Some caution should be placed in interpreting the opossum repeat content because only 136 kb of finished sequence (accession number CU468641.1) was available for analysis. Within the therian species (eutherian/placental and metatherian/marsupial) the SINEs are similar in proportion (6 to 9%). In contrast, LINEs are twice as abundant in human and wallaby as the mice (or limited opossum sequence) and LTRs are almost entirely absent from non-eutherian species.

In sharp contrast to the therian species the monotreme (platypus) has a striking amount of repetitive sequence in this region (>70%) of which 64% is split between SINE (39%) and LINE (25%) elements. Indeed, there are more SINE elements in

this region of the platypus genome than the combined repeat content of any other single species (Figure IV.13). Although not so dramatic, Margulies and colleagues also reported high levels of platypus SINEs (approximately 25%) in the greater *CFTR* region (Margulies et al. 2005a) and this would therefore appear to be a feature of the platypus genome.

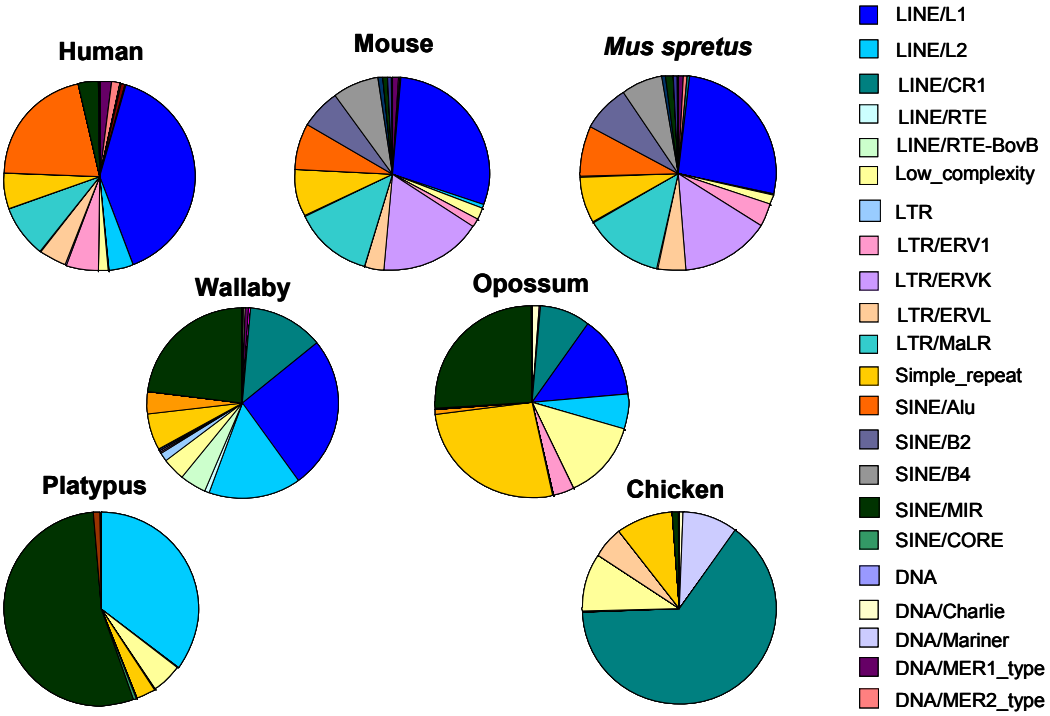

Figure IV.14. Repeat composition of multi-species sequences in the orthologous 11p15.5 region.

The pie charts show the distribution, based on contribution by total sequence length, of repeat families per species as indicated. The eutherian (placental) mammals are shown at the top, the metatherian (marsupial) species in the centre, the monotreme (platypus) at the bottom left and bird (chicken) at the bottom right of the figure. The key to repeat class/family is provided on the right.

An in-depth analysis of the repeat classes and families reveals considerable difference between species clades. The human and murine repeat compositions are broadly similar. However, both mice contain approximately 15% of LTR/ERVK

repeats which are almost entirely absent from human (0.005%). The main other difference between human and mice is the composition of SINE elements. In human, most SINE repeats are *Alu* elements (21%) with the remainder from the mammalian-wide interspersed repeat (MIR) family (3%). In mice, SINE repeats can be separated into *Alu* (8%), B2 (7%), B4 (8%) and MIR (1%) families (Figure IV.14).

The repeat composition of marsupial repeats is quite distinct, with features of eutherian and monotreme sequences, befitting its position in mammalian phylogeny (chapter I). By comparison with the eutherian species there are a much greater proportion of MIR elements (23-26%) in the marsupials but less than half the proportion found in the platypus (54%).

By far the majority of repetitive elements identified within chicken are of the LINE/CR1 family (64%) and together with the SINE/MIR family (1%) are the only interspersed repeat elements common to all 7 species studied here. It therefore would appear that relatively few insertions of MIR elements in the ancestral chicken genome was followed by a huge expansion of this repeat element in the monotreme followed by a progressive loss in marsupial and eutherian species (Figure IV.14).

Does the high-density presence of SINE repeats in platypus prevent imprinting of the region? John Greally has reported a strong association of imprinted regions with lower than average SINE contents (Greally. 2002). Since SINE elements tend to attract DNA methylation in order to suppress their transcription this could upset the balance of allelic methylation in imprinted domains and would therefore be actively selected against. The C+G-rich nature of SINEs may explain the extreme C+G content observed in platypus. However, if SINEs are silenced by methylation and methylated cytosine is prone to deamination resulting in thymine (Coulondre et

al. 1978, Duncan and Miller. 1980) wouldn't the C+G and CpG content of platypus be decreasing? This paradox is discussed further below.

MIRs are somewhat of a misnomer since they are present in non-mammalian sequences such as the chicken sequence reported here. MIRs are thought to have derived from an ancestral CORE-SINE element (Gilbert and Labuda. 1999, Gilbert and Labuda. 2000). Only the platypus sequence studied here has the archetypal CORE-SINE element (1%). CORE-SINEs are non-autonomous retrotransposons that require the enzymatic machinery of active LINE partners to spread within the genome (Ohshima and Okada. 2005). CORE-SINEs have been of great recent interest, providing important insights into mammalian evolution and function. This was very recently illustrated by Santangelo and colleagues who demonstrated the exaptation (a biological adaptation where the biological function currently performed by the adaptation was not the function performed while the adaptation evolved under earlier natural selection pressures) of an ancient CORE-SINE element into a mammalian neuronal enhancer (Santangelo et al. 2007). Another example in which an ancient relic of a transposable element has acquired function was demonstrated by Bejerano and co-workers who showed that a SINE element positioned 0.5 Mb from the neuro-developmental transcription factor gene (*Isl1*) in mouse behaves as an enhancer for this gene. Intriguingly the originator SINE element appears still to be active in the 'living fossil' coelacanth (Bejerano et al. 2006). A growing number of imprinted genes appear to have arisen through exaptation of retrotransposons (Suzuki et al. 2007, Wood et al. 2007, Youngson et al. 2005) and confirm the importance of interspersed repeats once considered to be 'junk' (Berg. 2006).

4.6 C+G content and CpG islands

The differential methylation of alleles is a hallmark of imprinted genes. DNA methylation occurs almost exclusively on the cytosine in CpG dinucleotides in vertebrates. It is therefore of interest to compare the C+G and CpG contents between species with and without imprinted gene regulation.

The C+G content of sequences was obtained from the output of the RepeatMasker program (Table IV-6). The level of CpG dinucleotides that constitute a CpG island is arbitrary, however, CpG islands are typically defined as having a length greater than 200 bp, a C+G content greater than 50% and observed over expected ratio greater than 0.60 (Gardiner-Garden and Frommer. 1987). The EMBOSS script newcpgreport was used, with above parameters, to identify CpG islands in each of the species and each of the regions sequenced. The C+G, CpG contents and number of CpG islands predicted within the regional sequences are shown in Table IV-6. As expected there is considerable variability both between regions and species. This is investigated further below. If we consider the sequences for human, mouse, wallaby and platypus in all SAVOIR regions it is apparent that the C+G content for all species is higher than the reported genome averages (or other genomic regions for wallaby and platypus, Table IV-7). This likely reflects the fact that 7 of the 9 regions sequenced correspond to cytogenetically light bands upon Giemsa staining of human chromosomes. These light bands are known to be C+G and gene-rich regions of the genome. The two regions corresponding to a dark band in human and mouse are the *STX16-GNAS* locus at human 20q13.32 (mouse 2qH4) and the DNA cytosine-methyltransferase 1 (*DNMT1)* locus at human 19p13.2 (mouse 9qA3). The *DNMT1* gene is itself not an imprinted gene, but its protein product is a maintenance methyltransferase enzyme critical for the correct imprinting of some

genes (Li et al. 1993a, Li et al. 1993b). Therefore 7 out of 8 regions containing at least one imprinted gene in human and/or mouse lie within light bands.

Table IV-7. Comparison of repeat and C+G contents between SAVOIR and other reported regions.

	Average repeat content (%)		Average C+G content (%)	
Species	SAVOIR Regions	Genome-wide	SAVOIR	Genome-wide
Human	45.24	47.61	45.2	41
Mouse	33.49	41.18	44.88	41.8
Wallaby	43.41	37#	40.23	37.3#
Platypus	49.97	44.9#	47.61	45.9#

#, Genome-wide figures are not currently available for wallaby and platypus. Figures were therefore taken from other sequenced regions in wallaby and platypus (Margulies et al. 2005a).

As is the case with the distribution of repeats, there are also clear species differences in C+G contents. Whilst the C+G ranges in mouse and wallaby are very similar, their medians are not (Figure IV.12B). In wallaby the distribution is skewed towards lower C+G content (median of 38.1%) whereas in mouse the data are skewed towards higher C+G content (median of 46.2%). For both mouse and wallaby the interquartile range reveals a more limited range of C+G contents when compared with human and platypus. This relatively tight distribution of C+G content in the mouse genome has been reported before (Waterston et al. 2002). The overall C+G content in the platypus sequences are higher than those for other species. What could account for such varied C+G contents? To investigate whether the high proportion of SINE elements in platypus are responsible for the high platypus C+G content, the orthologous 11p15.5 sequences were divided into unique and repeat containing fractions. Interestingly, the repeats which comprise 70% of the sequence in this region have a C+G content of 51%. By comparison, the unique fraction has a C+G content of 61%. So although the repeat content in platypus contributes in raising the C+G content above other mammalian levels it is the unique sequence

that is important (see below). In human, mouse and wallaby the reasons for varied C+G content are less clear but may include altered mutational or repair mechanisms (Sueoka. 1988, Wolfe et al. 1989) and/or differences in selection for C+G (Bernardi et al. 1988).

It has been hypothesised that there is an inverse correlation between C+G content and body temperature following investigations into the frequency of CpGs and methylated cytosine residues (5mC) between fish, amphibians, birds and mammals (Jabbari and Bernardi. 2004, Jabbari et al. 1997). The platypus body temperature is 30-32^{o}C, low for a warm-blooded mammal. Fish and amphibians are cold-blooded, so their body temperature changes with that of their environment. The CpG and 5mC levels for fish and amphibians are two-fold higher than the warm-blooded eutherians and birds and not correlated with repeat content (Jabbari et al. 1997). Intriguingly the platypus genome has a level of CpGs between those of eutherian mammals and fish (Figure IV.15 and Jabbari and Bernardi. 2004). This raises the possibility that the ancestral vertebrate genome had high CpG and methylation levels and that over the course of evolution there was a progressive depletion of CpGs and corresponding methylation. This depletion may have been brought about by deamination of 5mC which has been shown to occur at higher rates in warmer body temperatures (Shen et al. 1994). Whether the characteristics of the platypus genome regions sequenced here are typical of the genome should soon be known with the pending analysis of the WGS assembly. Additional monotreme and reptile sequences would also help to address issues of CpG dynamics in evolution.

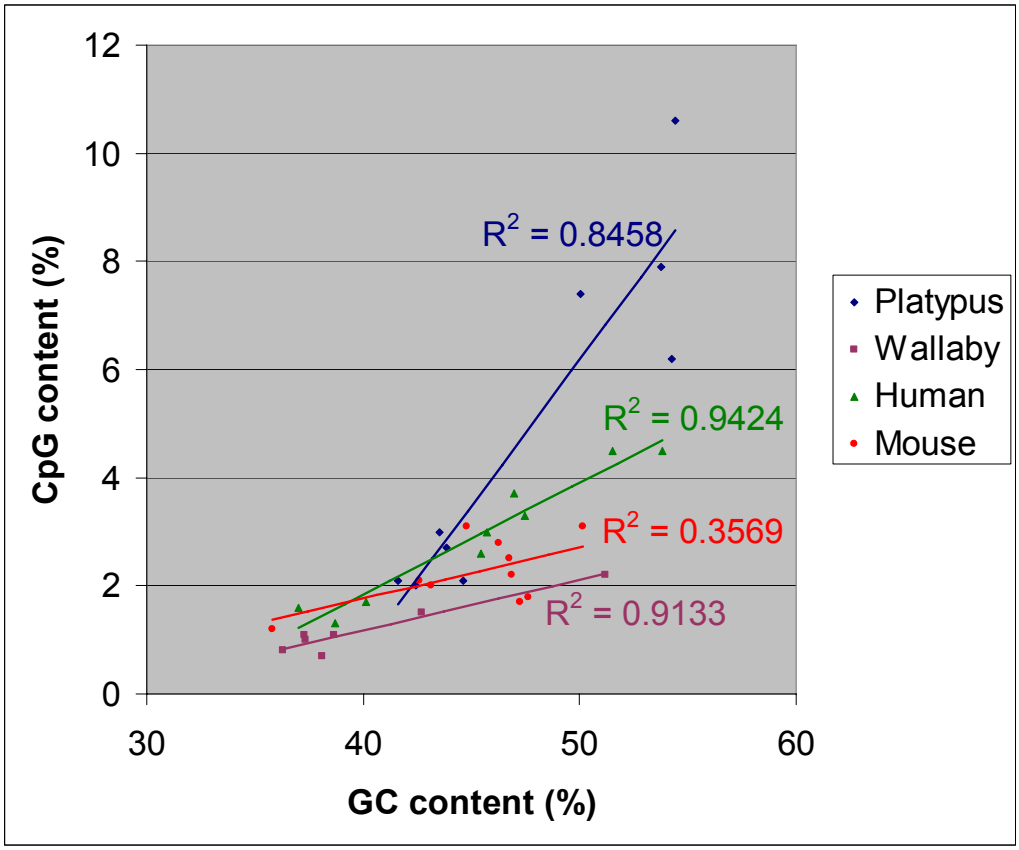

Figure IV.15. Plot of CpG and C+G contents for multi-species regional sequences.

Each data point represents a distinct genomic region in the species indicated. Regression lines and coefficients (R^2) are colour-coded for each species in the key.

How do the C+G and CpG contents compare between species in the SAVOIR regions? With the exception of mouse (R^2=0.3569) there is a good positive correlation between CpG and C+G levels within species, as we might predict (Figure IV.15). The lack of a statistical positive correlation between CpG and C+G levels in mouse is likely a function of the small sample size.

The most striking observation is the high C+G content and CpG density in the platypus genome when compared with placental or marsupial genomes (Figure IV.15). Indeed, with the exception of the wallaby orthologous 11p15.5 region (51.21% C+G, 2.2% CpG), there is no overlap between platypus and wallaby CpG contents across the spectrum of C+G content. The high levels of C+G and

corresponding CpG frequency in the platypus regions illustrates that the CpG island parameters commonly used for annotation of the eutherian genomes may not be adequate for annotation of the platypus genome. Using those parameters 578 CpG islands were identified in platypus compared with 229, 126 and 72 identified in the corresponding human, mouse and wallaby regions, respectively (Table IV-6). The limited opossum sequence in the IC1 region also has high numbers of predicted CpG islands. Furthermore an 83 kb CpG island was predicted in the platypus *UBE3A* region. By comparison the largest CpG island found in the entire human genome spans 36.6 kb on chromosome 10 and 95% of human CpG islands are less than 1.8 kb (International Human Genome Sequencing Consortium. 2001). If we look again at the 11p15.5 region, the C+G contents are almost identical between human and platypus (53.84% and 53.77%, respectively, Table IV-6). However, the CpG density in human is 1.8 times lower (4.5% and 7.9% in human and platypus, respectively) and yet there is no evidence for a greater gene density in this region of the platypus genome (Table IV-8). It therefore seems likely that the increased CpG frequency in platypus is not functionally correlated. I therefore conclude that the parameters appropriate for CpG island identification in the platypus genome should be adjusted to account for higher CpG frequencies at a given regional C+G content.

Table IV-8. Predicted CpG islands in the human 11p15.5 orthologous sequences.

Species	Sequence length (bp)	Number of CpG islands	Average CpG island density	Number of annotated genes (pseudogenes)
Human	1648045	102	1 per 16.2 kb	49 (7)
Mouse	1609784	41	1 per 39.3 kb	30 (3)
Mus spretus	1438076	33	1 per 43.6 kb	33 (3)
Wallaby	1528894	40	1 per 38.2 kb	29 (8)
Opossum	136229	39	1 per 3.5 kb	ND
Platypus	763115	319	1 per 2.4 kb	8
Chicken	1162135	28	1 per 41.5 kb	25

The same sequences used in the repeat analyses were used here. CpG island prediction was performed using the EMBOSS newcpgreport program with parameters: Window size, 100bp; Shift, 1; Minimum length, 200 bp; Minimum average observed/expected ratio, 0.6; Minimum percentage, 50%. ND, not yet determined.

4.7 SAVOIR consortium website

In order to make the most of the resources developed during this thesis we established collaborative research programmes with local and international groups with a common interest in elucidating the ancestral mechanisms of imprinting. Collectively this group is known as the SAVOIR consortium and currently comprises 6 research groups and supporting teams in 5 establishments. Links to each of these groups, and their specific research interests, can be found at: http://www.sanger.ac.uk/PostGenomics/epicomp/participants.shtml.

Consistent with the Wellcome Trust data release policies for large-scale community resource projects ((Bentley. 1996) and the Fort Lauderdale agreement, http://www.wellcome.ac.uk/doc_wtd003208.html) we have submitted all BAC sequences to the EMBL DNA database as they were being generated. Furthermore a SAVOIR website was created (with assistance from Paul Bevan and Carol Scott). This website (Figure IV.16) provides an overview of the project, a full list of consortium participants and importantly links to the contig mapping. These maps

display the tiling paths of BACs sequenced for each species and region. The maps are anchored to an Ensembl style 'gene view' webpage with hyperlinks into the Ensembl human database (Figure IV.17). The mouse-over facility enables the user to identify BAC clone names, the status of sequencing of that clone and where available a link through to the sequence accession file in EMBL. The data displayed on the website is curated by me using hypertext markup language (html) files for text based pages (e.g. overview and participants pages) or underlying MySQL tables to display the mapping and sequencing data. The MySQL tables accessed by the website are described in chapter II.

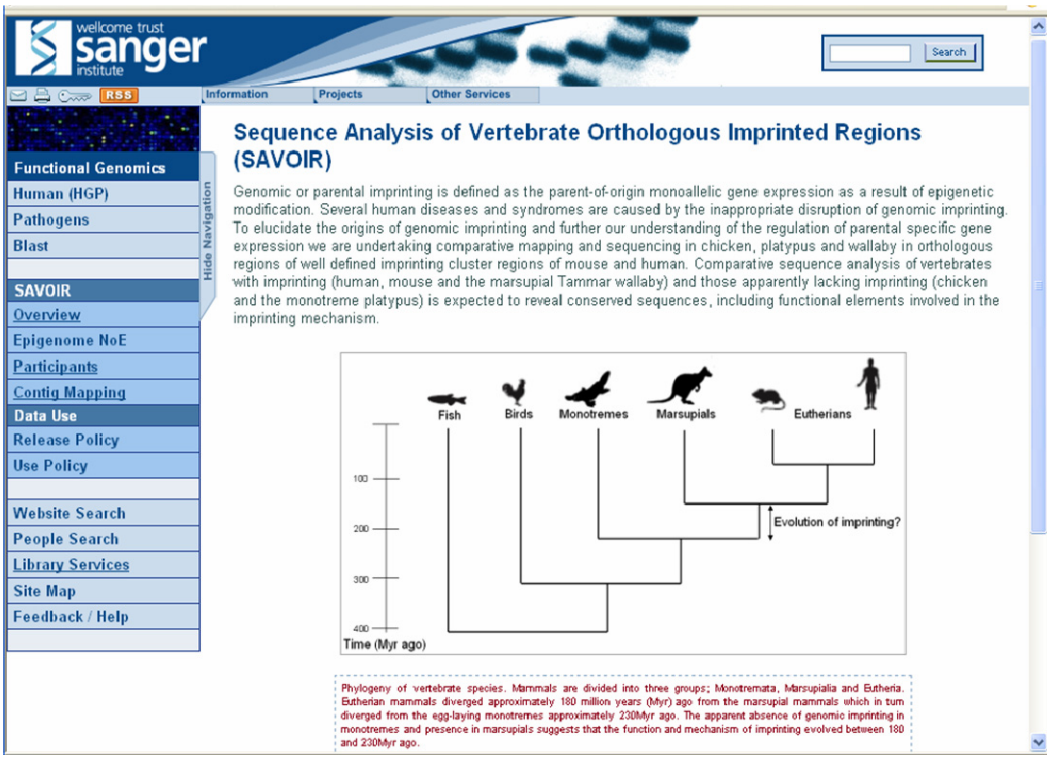

Figure IV.16. The SAVOIR website (http://www.sanger.ac.uk/PostGenomics/epicomp).

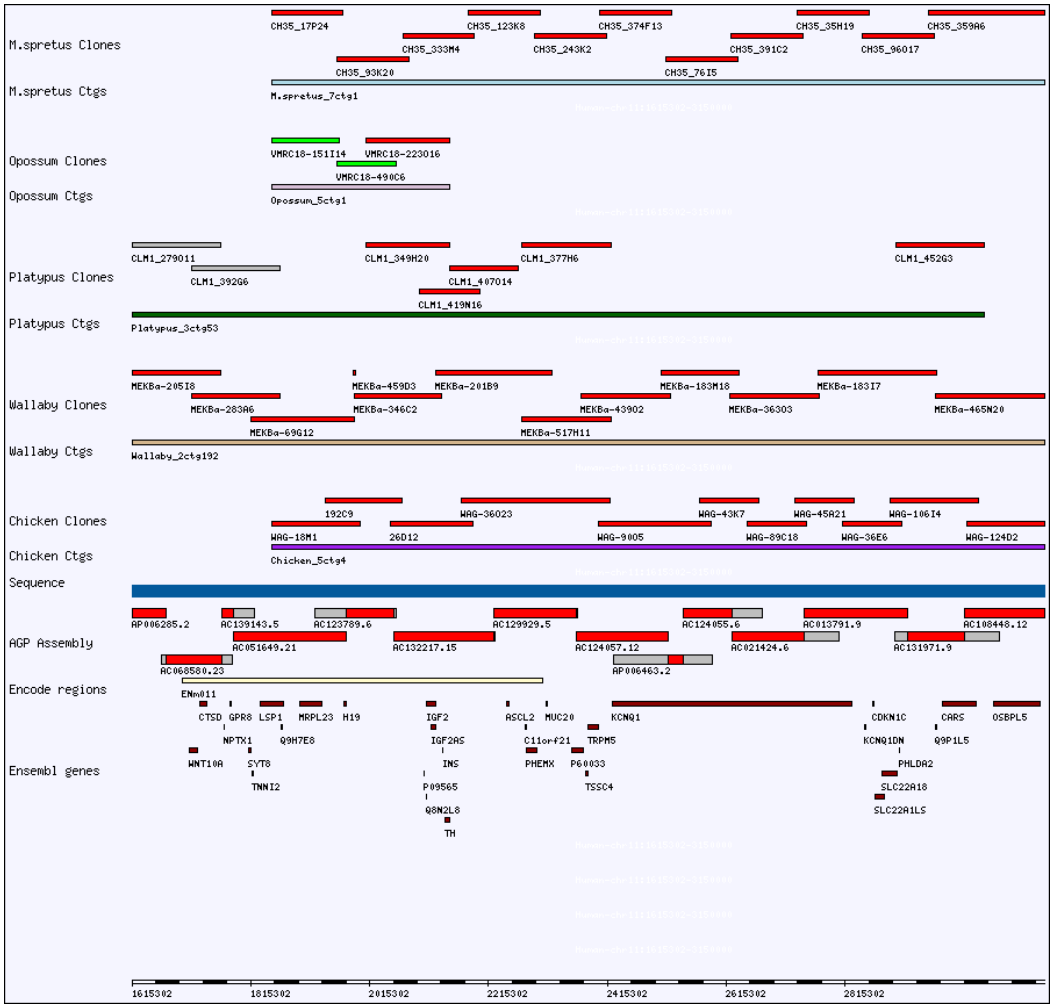

Figure IV.17. SAVOIR contig view.

The Ensembl style view shown was obtained by following the link to 'Contig Mapping' from the SAVOIR homepage and selecting the 'Wallaby_2ctg192' fingerprint contig name. The resulting display shows the clones selected for sequencing in each species; M.spretus, *Mus spretus*; Opossum, *Monodelphis domestica*; Wallaby, *Macropus eugenii*; Platypus, *Ornithorhynchus anatinus*; Chicken, *Gallus gallus*. The clone name is shown beneath each bar. Colour of the bars indicates sequence status; red, finished; grey, finishing in progress; green, in shotgun sequencing. The extent of the mapped fingerprint contigs are shown under the BACs for each species. As a point of reference the human BAC tiling path is shown beneath the blue bar. The red portions of each sequence accession contribute to the human consensus sequence. ENCODE regions and EnsEMBL known genes are depicted. A scale-bar is shown at the bottom.

4.8 Discussion

From the sequence analysis and annotation presented here the differences in relative expansions/contractions cannot be wholly accounted for by interspersed repeat contents and are significantly due to segmental duplications. The *KRTAP5* and *Tnfrsf* gene families flanking the IC1 and IC2 domains illustrate that these duplications have given rise to gene families, members of which have gained mutations and lost function (pseudogenisation) whilst others are presumably evolving novel functions (neofunctionalisation).

The contribution of interspersed repeats to genome expansion may be underrepresented because different lineages are expected to have different mutation rates and therefore very ancient interspersed repeats may not be detected computationally. The overall repeat contents may therefore be considerably higher than current estimates. This appears to be true in the mouse genome which has seen high transposable element insertion followed by a higher nucleotide substitution rate as determined by the study of lineage specific insertion events (Waterston et al. 2002). The availability of increasing amounts of sequence should enable similar studies in diverse genomes and help to explain the wide variation in vertebrate genome sizes (the C-value paradox).

The utility of CpG island prediction to reveal unmethylated (often functional) sites of the genome is hampered by the general methylation of C+G-rich retroelements (Yoder et al. 1997). Furthermore, CpG island prediction currently relies upon sequence compositional thresholds which were established in 1987 with available vertebrate sequences (Gardiner-Garden and Frommer. 1987). The sequence databases have since exponentially increased in size and diversity and for some species, at least, the commonly used parameters (≥200bp, ≥50%, O/E ≥0.6) will be either too conservative or too liberal. False negatives would result in missing

unusual CpG dinucleotide densities in otherwise AT-rich sequences of known function. Alternatively CpG islands may be over-predicted because of the high C+G and corresponding CpG density of sequences as shown here in the platypus.

CG cluster annotation was recently shown to improve annotation of known promoters compared with the CpG island prediction method (Glass et al. 2007). The observed clustering is a result of the genome-wide decay of CG dinucleotide content, with preservation of CG density at certain regions. In the human genome the authors demonstrated that optimal CG clusters contain at least 27 CpGs in a sequence length of no more than 531 bp (in mouse, 24 CpGs in no more than 585 bp). Using these parameters 44,165 CG clusters were identified in the human genome, with repeats not masked. These CG clusters were shown to identify more 5' ends of genes and known hypomethylated sites than the CpG island prediction method. Importantly, the CG cluster definition is not influenced by *a priori* assumptions of sequence composition and should, therefore, be widely transferable between different species with highly variable C+G contents.

CpG island or CG cluster predictions do not directly test the methylation status of the DNA and yet the methylation is intrinsic to the function. Therefore ultimately the methylation status of genomic DNA should be targeted directly. A variety of methods exist to fractionate methylated and unmethylated DNA regions. These include methylation sensitive restriction enzyme (MSRE) fractionation, methylcytosine immunoprecipitation and bisulphite sequencing. The fractions can then be discriminated by hybridisation to micro-arrays or sequenced directly to establish methylation enriched sequences (Bernstein et al. 2007 and references within). The application of these technologies to the full human genome is now underway in Europe (http://www.epigenome.org/index.php) and U.S.A

(http://nihroadmap.nih.gov/2008initiatives.asp) and will identify truly functional CpG islands which can then be used to refine computational predictors for application to cells, tissues and even species not yet tested.

To conclude, this chapter has demonstrated that comparative sequence analysis is a powerful tool with which to investigate genome evolution. Genome expansion/contractions cannot be fully explained by interspersed repeat content but are significantly due to segmental duplication events. The platypus sequence analysis reveals high C+G and corresponding CpG content which likely reflects extraordinary SINE/MIR content of the regions studied. Consequently new parameters are required to discern unusual and functional CpG densities from background. Notably absent from this chapter is a discussion of multiple-species sequence alignment and its power to identify conserved sequences of likely functional importance. This is important in the context of establishing a comprehensive catalogue of functional elements within orthologous imprinted regions and is therefore the subject of the next chapter.