Chapter I - Introduction

1.1 Opening remarks

Human biology and disease is determined by environmental, genetic and epigenetic factors in combination. To understand the causes of human disease it is therefore important to study, not only the changes in DNA sequence associated with disease pathologies, but also the regulation of gene expression and changes in chromatin structure crucial to normal human development. The sequence of the human genome (International Human Genome Sequencing Consortium. 2001, International Human Genome Sequencing Consortium. 2004) and that of a growing number of genomes (see below) provides the necessary tools and resources to study the heritable changes in cellular chromatin structure and gene expression, known as 'epigenetics'. The term epigenetics was coined by Conrad Waddington over 60 years ago to describe 'the interactions of genes with their environment, which bring the phenotype into being' (Waddington. 1942). More recently the term 'epigenomics' has been adopted to describe epigenetic changes on a genome-wide basis (Beck et al. 1999).

Key areas of study into heritable variations in gene expression include X chromosome inactivation (XCI) in mammals (reviewed in Chow et al. 2005, Heard. 2005, Huynh and Lee. 2005) and genomic imprinting (reviewed in Wilkins. 2005). Recent studies reveal mechanistic parallels between these processes (Huynh and

Lee. 2005, Reik and Lewis. 2005) but the scope of this thesis is to further our knowledge of gene regulation and evolution of the genomic imprinting mechanism.

1.2 Genomic imprinting

Genomic imprinting is a phenomenon of angiosperm (flowering) plants and placental mammals. Autosomal genomic imprinting in mammals was first described in the early 1980s following experiments to transplant haploid parental genomes into mouse zygotes thereby giving rise to embryos containing two sets of chromosomes from the same parent (McGrath and Solter. 1984, Surani et al. 1984). Following these nuclear transplantation experiments the embryos did not develop to term showing that maternal and paternal genomes are not equal. In nature, uniparental disomies (UPDs) result from the duplication of one parental allele followed by loss of the opposite parental allele. Characterised UPDs in mice revealed abnormal phenotypes (Cattanach and Kirk. 1985) that were attributed to the presence of imprinted genes transcribed from only one of the parental alleles and silenced (or imprinted) on the other. Therefore, in the regions of UPD, gene dosage is altered such that a gene may not be expressed at all (equivalent to a null allele) or over-expressed two-fold. Genomic imprinting is an epigenetic process, controlled by heritable modifications of DNA (but not the nucleotide sequence) and chromatin resulting in parent-of-origin gene expression.

The first imprinted genes to be identified were the mouse Insulin-like growth factor 2 (*Igf2*), expressed from the paternal allele only (DeChiara et al. 1991), its receptor (*Igf2r*) expressed from the maternal allele only (Barlow et al. 1991) and *H19* (cDNA clone number *19* isolated from a foetal *H*epatic library), a maternally expressed, non-coding RNA (ncRNA) gene of unknown function lying 70 kb telomeric of *Igf2* (Bartolomei et al. 1991). Ninety imprinted genes have been identified to date in mouse, 56 in human and 37 imprinted in both species (Morison et al. 2005) (http://igc.otago.ac.nz/home.html) (Figure I.1). The lower number of human imprinted genes compared with mouse in part reflects the difficulty in obtaining human biological material for imprinting studies, especially tissues and cells from very early developmental stages. The incomplete overlap between human and mouse imprinted gene sets is likely accounted for by the presence of lineage specific genes, differences in selection pressures acting on the species or simply a lack of experimental evidence.

Estimated total numbers of imprinted genes in human and mouse based on genome-scans and the proportion of mouse loci showing parental effects typically range from 100-200 (Barlow. 1995, Hayashizaki et al. 1994). A recent bioinformatic study, based on classifier programs trained with known imprinted genes, predicts 156 novel candidate imprinting genes in human (Luedi et al. 2007). Only two of these genes were experimentally tested and shown to be imprinted and therefore further work is required to validate all other candidate genes.

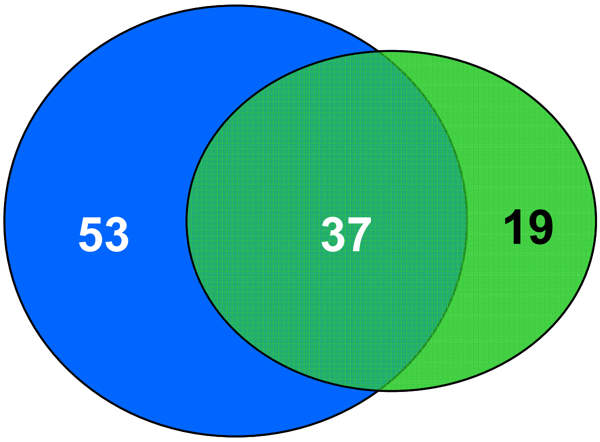

Figure I.1. Venn diagram of mouse and human imprinted genes.

90 genes are reported to be imprinted in mouse (blue), 56 in human (green) and 37 in both species. Data from http://igc.otago.ac.nz/home.html and Appendix A.

1.2.1 Common features of imprinted regions

The genetic basis for the epigenetic mechanism of imprinting is rather enigmatic. We do not fully appreciate which sequence features discriminate imprinted from non-imprinted genes and why some DNA sequences are first recognized and then differentially 'marked' in the germ-lines. Several common features of imprinted genes/regions have been reported. They generally occur in clusters with approximately 80% being physically linked within a Megabase (Mb) (Lalande. 1996, Reik and Walter. 1998) of other imprinted genes. This clustering implies that multiple genes in the region are coming under the same mechanism of control and thus there are a limited number of mechanisms. Within these clusters there are frequently parental specific differentially methylated CpG islands. These differentially methylated regions (DMRs) are often crucial to the imprinting control of the gene cluster and as such are termed imprinting centres (ICs) (Lewis and Reik. 2006). Examples include the differentially methylated domain (DMD, IC1) lying 2 kb upstream of the *H19* gene and the KvDMR (IC2) lying in intron 10 of the *Kcnq1* gene. These ICs lie in neighbouring domains of human chromosome 11p15.5 and mouse chromosome 7qF5. In addition to DNA methylation, parental-specific chromatin modifications such as core histone acetylation and methylation are also a common feature of imprinted gene regions (Margueron et al. 2005, Peters and Schubeler. 2005, Soejima and Wagstaff. 2005). Parental-specific antisense RNA transcripts, microRNAs (miRNAs), small nucleolar RNAs (snoRNAs) and longer ncRNA transcripts evidently also play a role in imprinted regulation and are identified in most imprinted regions studied (Edwards and Ferguson-Smith. 2007, Pauler et al. 2007, Yang and Kuroda. 2007, Zaratiegui et al. 2007). Examples of antisense transcripts include the paternally expressed *Kcnq1* overlapping transcript 1 (*Kcnq1ot1*) which resides within intron 10 of the maternally expressed *Kcnq1* gene (Fitzpatrick et al. 2002) and the antisense *Igf2r* RNA (*Air*), which has its promoter within intron 2 of *Igf2r* and acts in *cis* to regulate *Igf2r* expression (Sleutels et al. 2002). Increasing numbers of miRNAs are also being identified within imprinted gene clusters including the *Dlk1-Dio3* region of mouse chromosome 12 (Seitz et al. 2004). Recently miR-675 was identified within the mouse and human *H19* ncRNA although the function of both miRNA and parent transcript are currently enigmatic (Cai and Cullen. 2007). snoRNAs have been identified at the Prader-Willi/Angelman Syndrome (PWS/AS) imprinted region at human chromosome 15q11-13 (mouse chromosome 7qC) and the *DLK1-DIO3* region of human 14q32.2 (mouse chromosome 12qF1) (Cavaille et al. 2002). Finally, ncRNA transcripts such as *H19, Gtl2* and *Air* are found in imprinting regions and in the absence of experimental evidence we might speculate that they function like *Xist* on the inactive X chromosome to 'coat' the targeted chromosomal region and silence transcription of the genes. *Xist* RNA is thought to attract several histone modifying enzymes resulting in the inactive X chromosome being marked by repressive histone modifications such as histone 3 lysine 9 (H3K9) and histone 3 lysine 27 (H3K27) methylation.

Several interspersed repeat families have been reported to be enriched (e.g. long [6-8 kb]-interspersed nuclear elements, LINEs) or depleted (e.g. short [100-400bp]-interspersed nuclear elements, SINEs) within imprinted regions. This has stimulated debate as to whether these DNA elements might serve as signatures guiding the necessary epigenetic modification machineries to the imprinted regions (Reviewed in Walter et al. 2006). The increased proportion of LINE elements in imprinted regions (with >40% cytosine and guanine (C+G) content (Walter et al. 2006)) is reminiscent of the LINE enrichment observed on the X-chromosome which was proposed to assist in the spreading of epigenetic silencing along the inactive X (Lyon. 1998). Many groups have reported the relative depletion and distribution of SINE elements in imprinted gene regions (Allen et al. 2003, Greally. 2002, Ke et al. 2002), which appears to persist after normalising for C+G content (since SINE elements are C+G-rich) (Walter et al. 2006). It would appear that there is evolutionary selection pressure to maintain high levels of some interspersed repeats or low levels of others in imprinted gene regions. Reasons for such selection are not yet clear but the comparative study of repeat densities and distributions in genomes with and without imprinting should help to address this question.

1.2.2 Evolution of genomic imprinting

Given the dangers of functional haploidy, in which recessive mutations are exposed, why has imprinting evolved? Various non-mutually exclusive hypotheses have been put forward to explain the evolution of genomic imprinting. These include prevention of parthenogenesis (Kono et al. 2004, Solter. 1988), the ovarian time bomb model (Varmuza and Mann. 1994) the rheostat model (Beaudet and Jiang.

2002, McGowan and Martin. 1997) and the intralocus sexual conflict hypothesis (Day and Bonduriansky. 2004). However, possibly the most widely accepted is the parental conflict hypothesis (Moore and Haig. 1991, Wilkins and Haig. 2003), more recently termed the kinship theory (Burt and Trivers. 1998, Haig. 2004). Each of these hypotheses is discussed in turn.

1.2.2.1 Prevention of parthenogenesis

Many imprinted genes are known to be involved in development and this has stimulated many of the models which consider the selective pressures that resulted in the evolution of the function of imprinting. One of the earliest suggestions was the prevention of parthenogenesis (Solter. 1988). Parthenogenesis is the "procreation without the immediate influence of a male" as defined by Richard Owen in 1849 (Owen. 1849). Parthenogenetic embryos would be more susceptible to recessive genetic disorders because both copies of every gene are inherited from the same parent (mother). This likely explains why many genera including mammals have relinquished parthenogenesis in favour of sexual reproduction. This model is consistent with the observations that parthenogenetic embryos have a relatively normal embryo proper but poorly developed trophoblast and are unable to implant and develop fully (Barton et al. 1984). In 2004 Kono and colleagues were able to generate a viable parthenogenote mouse by deleting the *H19* transcription unit to increase *IGF2* expression levels and therefore demonstrated that genomic imprinting is the barrier to parthenogenesis (Kono et al. 2004, Kono. 2006). Some argue that this theory is unable to explain why some genes are inactivated on the paternal chromosome. But if the same reasoning is applied to prevention of molar pregnancies, which result from enucleate eggs being fertilised by one sperm that duplicates its DNA in the process of androgenesis, then this observation is understandable.

1.2.2.2 The Ovarian Time Bomb model

Varmuza and Mann in 1994 further developed the prevention of parthenogenesis model and proposed the 'Ovarian Time Bomb' hypothesis (Varmuza and Mann. 1994). They suggested that genomic imprinting evolved to protect the female from invasive trophoblastic disease. Gestational trophoblastic disease is caused by normal and ectopic pregnancies, but the risk is 1000 fold greater in molar (no maternal genome) pregnancies. The highly invasive and metastatic nature of these tumours is believed to reflect the role of the trophoblast in development. The occurrence of benign ovarian tumours is relatively high in humans with 4-7% of women affected at some point in their life (Morrow et al. 1993). Varmuza and Mann suggested that these tumours would become malignant trophoblastic disease without imprinting. Oocytes can be parthenogenetically activated throughout life but fail to form functional trophoblast because the genes required for this are inactivated in the female germ-line. Functional copies of the genes are only provided by the male germ-line after fertilisation thus protecting the female from trophoblastic disease. They suggest active copies of these genes are tolerated in the spermatocytes because the frequency of male germ-line tumours is 1000-fold lower than that of ovarian tumours. The authors postulated that most imprinted genes were 'innocent by-standers' which only became imprinted because they were recognised by the imprinting machinery regulating trophoblast-specific genes. Arguments raised against this model include the low occurrence of trophoblast disease in non human species and the concept that only one gene would need to be inactivated in the female germ-line to prevent trophoblast development (Haig. 1994, Moore. 1994, Solter. 1994). Problems with the ovarian time-bomb model are that it fails to explain why some imprinted genes are switched off in the male germ-line or the imprinting of genes involved in post-natal maternal care (e.g. *MEST* or *PEG3*, Constancia et al. 2004). It has been demonstrated that marsupials, which have non-invasive placentas, also display genomic imprinting. Marsupials and eutherian lineages diverged from each other approximately 148 million years (Myr) ago (see below) so it is highly unlikely that imprinting evolved to protect the female from invasive trophoblast but perhaps this is a useful by-product.

1.2.2.3 The Rheostat model

The Rheostat model predicts that imprinting evolved to increase the evolvability of a region through functional haploidy (Beaudet and Jiang. 2002). The model suggests that most imprinted genes are 'quantitative hypervariable' (QH) loci which exhibit great variation in both the levels of gene expression and phenotype; thus producing phenotypic variance along a continuum. Targeted genes would be those involved in continuous phenotypes such as growth and/or behaviour. As imprinted genes display functional haploidy, alleles can remain hidden from natural selection for a number of generations. The model predicts that when the selective advantage for haploidy acts on a QH locus, a rapid and reversible form of 'imprinting–dependent evolution' is created. Such a mechanism would allow a population to rapidly adjust to a changing environment (Beaudet and Jiang. 2002). This model fits with the type of genes that tend to be imprinted but fails to explain why imprinting is not seen in other vertebrates. Imprinted genes have also been shown in a previous study to be evolving no more rapidly than non-imprinted genes (Hurst and McVean. 1998).

1.2.2.4 Intralocus sexual conflict model

The intralocus sexual conflict hypothesis (Day and Bonduriansky. 2004) provides a potential explanation for much of the currently available empirical data, and it also makes new predictions about patterns of genomic imprinting that are expected to evolve but that have not, as of yet, been looked for in nature. The basis of this theory is the assumption of intralocus sexual conflict occurring when selection at a locus favours different alleles in males versus females (Anderson and Spencer. 1999, Rice and Chippindale. 2001). Reproductive success ensures the transmission of high fitness alleles from parent to offspring. It follows then that males will more likely pass on high male-fitness alleles and females will more likely pass on high female-fitness alleles to their offspring. Intralocus sexual conflict then ensues because the inherited alleles are expressed differently in the sexes. Natural selection should, therefore, favour modifier loci that silence maternally inherited alleles in males and conversely, paternally inherited alleles in females. In this system genomic imprinting would be selected for because this form of epigenetic inheritance would mitigate the severity of intralocus sexual conflict. Unlike other theories, this theory focuses on the evolution of the locus causing the imprinting (modifier) and not the imprinted locus itself.

1.2.2.5 Parental conflict/kinship hypothesis

The parental conflict hypothesis was first formalised in 1989 by Haig and Westoby (Haig and Westoby. 1989) then later refined by Haig and Moore in 1991 (Moore and Haig. 1991). However, it was recently recognised that Willson and Burley had the earlier insight in their 1983 book "Mate Choice in Plants: Tactics, Mechanisms and Consequences" (Haig and Westoby. 2006, Willson and Burley. 1983). Collectively they proposed that imprinting arose through a 'tug of war' between maternal and paternal alleles over resource allocation to offspring. The more nutrients an embryo can absorb from its mother *in utero* the more likely it is to survive to reproduce. This may have detrimental affects on the mother's health and ability to provide for future offspring. This model suggests that paternal genes within the embryo would be selected for extracting more resources from the mother, whereas maternal genes would be selected for moderation of nutrient acquisition by the current offspring in favour of future ones which may be by different fathers. The conflict hypothesis therefore predicts that imprinted genes will be involved in resource acquisition by an offspring from its mother. In most mammals resource transfer from mother to foetus occurs across the placenta for neonates and following birth in the process of lactation. Therefore imprinted loci might be expected to be involved in placental and embryonic growth, suckling and neonatal behaviour. Experimental support for this theory has grown (Haig. 2004) and includes the imprinting of the mouse *Igf2* and *Igf2r* genes (Barlow et al. 1991, DeChiara et al. 1991). Foetally expressed *Igf2*, which is involved in nutrient transfer, is expressed only from the paternal allele whereas *Igf2r*, which binds to *Igf2* and sequesters it for degradation, is maternally expressed. More recent studies suggest that imprinted genes play a vital role in regulating the supply and demand of maternal nutrients *in utero* (Reik et al. 2003). Even recent studies in Arabidopsis provide evidence that a transcription factor *MEA (MEDEA)*, a critical gene responsible for endosperm formation (embryo nourishing tissue), has rapidly evolved a new function, supporting the conflict hypothesis (Spillane et al. 2007). A number of arguments have been raised against the conflict hypothesis. For example Hurst argues that with this model imprinting should not persist in a monogamous species, such as the mouse *Peromyscus polionotus,* because the maternal and paternal genomes have identical interests (Hurst. 1998). However, it should be noted that the conflict model does not predict that there would be a rapid loss of imprinting if a species switches to a monogamous lifestyle. Further results that seem to contradict the conflict model were presented in a review of uniparental disomies (UPDs) by Hurst and McVean (Hurst and McVean. 1997). The conflict model predicts that paternal UPDs (pUPDs) should be growth enhancing whereas this review claimed that most pUPDs are generally growth restricted or show no phenotype. This could be seen as misleading because for those pUPDs that show phenotypes most are severe resulting in prenatal lethality and growth enhancement is often evident in early development but lost as the phenotypes progress. This inconsistency can also be explained by an over-allocation of resources to the placenta by paternally derived genes. This is supported by evidence of enlarged placentas in androgenotes and some pUPDs. A review of the physiological functions of imprinted genes (Tycko and Morison. 2002), showed that the majority of imprinted genes with *in vivo* data conform to the conflict hypothesis. Although this model is far from proven it is the one that currently best fits empirical data.

1.2.3 The mechanism of genomic imprinting

The question of when and how the genomic imprinting mechanism evolved has been the subject of much research and debate and forms the core motivation for this thesis. Hypotheses include the host defence theory (Barlow. 1993, McDonald et al. 2005), X-inactivation driven evolution (Huynh and Lee. 2005, Lee. 2003) and chromosomal duplication (Walter and Paulsen. 2003).

1.2.3.1 Host defence mechanism

One of the earliest theories is that imprinting mechanisms arose from the host defence mechanism against foreign DNA (Barlow. 1993). This was based on the link between imprinting and methylation, and is supported by the fact that retroviral, repetitive and transposable elements are usually methylated within the genome (Yoder et al. 1997). Yoder and colleagues suggest that the primary role of cytosine methylation is the suppression of parasitic elements but it does have a secondary function in allele specific gene expression. Further evidence for imprinting co-opting this mechanism comes from the imprinted regions themselves. The paternally expressed retrotransposon-like 1 (*Rtl1*) and paternally expressed 10 (*Peg10*) genes are both derived from long terminal repeat (LTR) retrotransposons of the Ty3/Gypsy family (Ono et al. 2001, Seitz et al. 2003, Suzuki et al. 2007, Youngson et al. 2005), whereas *Znf127* and *U2afbp-rs* are both intronless retrotransposed X-linked genes. Many imprinted domains also have tandem repeat sequences within them which have been suggested to play a role in the regulation of imprinting. Most of these repeats are not conserved between mammalian sequences but in the *Dlk1/Gtl2* cluster there is a conserved C+G-rich region which contains tandem repeats in human, mouse and sheep. The differential methylation of this region plays a crucial role in the imprinting of the cluster. Further analysis of imprinted domains in distantly related species is needed to see if there is a correlation between DNA methylation, repeat type/composition and imprinting. A key question is whether there is involvement of differential methylation in imprinted loci of marsupials? A recent study has revealed differential methylation in the 5' region of the Tammar wallaby orthologue of *PEG10* (Suzuki et al. 2007), supporting the idea that imprinting mechanisms may have a common origin.

1.2.3.2 X-inactivation driven evolution of imprinting

Similarities between features of XCI and genomic imprinting have been long recognised (Huynh and Lee. 2005, Lee. 2003, Reik and Lewis. 2005) and have led to the hypothesis of X-inactivation being the 'driving force' behind imprinting evolution (Huynh and Lee. 2005, Lee. 2003). This theory is based on the common features of XCI and imprinting e.g. differential methylation, non coding RNAs, antisense transcripts and presence of CCCTC binding factor (CTCF). Lee proposes that such mechanisms existed in mammalian ancestors for different purposes but were co-opted onto the X chromosome to silence the paternal X. Once they had been fixed on the X chromosome the mechanisms were adopted by autosomes in order to overcome various biological obstacles. In marsupials the paternal X chromosome is always inactivated as is the case in extraembryonic tissue in mice. Both XCI and imprinting have yet to be observed in monotremes, the most ancient extant mammalian relatives, so currently it is not known which process appeared first (Rens et al. 2007).

The *XIST* ncRNA is key in the initiation of X inactivation in eutherian mammals and is derived from a protein-coding gene in marsupials with no role in X inactivation (Duret et al. 2006). This change in function occurred after the divergence between eutherian and marsupial lineages and therefore X chromosome dosage compensation mechanisms must have evolved independently. This would appear to refute the hypothesis that X inactivation was the driving force behind genomic imprinting. However, it seems likely that the same molecular tools have been re-used for different evolutionary adaptations.

1.2.3.3 Chromosomal duplication

Along similar lines is the theory that imprinting mechanisms evolved through chromosomal duplications and imprinted genes were originally found on one (or a few) ancestral pre-imprinted chromosome region(s) (Walter and Paulsen. 2003). Walter and Paulsen noted that genes associated with imprinted regions frequently had paralogues which were linked to other imprinted regions and were often imprinted themselves. This observation is not surprising because paralogous genes are likely to have similar functions as they derive from a common ancestral gene. Thus, they may harbour the same imprint signals and be under the same functional selective pressure to be imprinted. Examples include the murine imprinted genes *Ins1* and *Ins2* (orthologue of human *INS*) which both code for insulin.

This model claims that regulatory elements would have existed before duplication and were transmitted to both duplicated domains. Rearrangements of these domains and further duplications brought about the clusters we see today. Many of these duplication events can be observed in *Fugu* (*Takifugu rubripes*) and therefore the authors argue that random monoallelic expression may have existed in other vertebrate clades prior to the onset of imprinting. How this random mechanism might have persisted for over 200 Myr between the divergence of *Fugu* and the emergence of mammals is unclear. However, it is of interest to know whether imprinted genes are clustered on one (or a few) chromosomes in organisms such as birds and monotremes in which imprinting has not been demonstrated. Resources for these species are now available to address this issue and are investigated in this thesis.

1.3 Genomic sequencing

The true value of genomic sequence lies in its annotation, but the quality of annotation is intrinsically linked to the quality of sequence. Genome sequencing projects operate on a continuum from low sequence coverage (e.g. 2x) through 'draft' (typically 6x) to 'finished' sequence which may have 8-12x coverage and significant efforts to resolve ambiguities. The definition of finished sequence is somewhat arbitrary but following a meeting of Sequencing Centres held in Bermuda in 1997 there was broad agreement that finished sequences should contain no more than 1 error in 10,000 bp and have no gaps (http://www.genome.gov/10000923). In practice the error rate is significantly lower for the human genome project (1 in 651,000 bp), at least for sequencing centres such as the Sanger Institute (Schmutz et al. 2004).

The draft assemblies of the human genome announced in 2000 had important short-comings with less than 90% euchromatic DNA represented and approximately 150,000 gaps (International Human Genome Sequencing Consortium. 2001, Venter et al. 2001). Despite underlying physical maps for much of the genome the correct order and orientation of many local segments were unknown. It took hundreds of scientists another 3 years before the human genome was finally declared complete and yet even today curation of the genome continues. I think it is fair to say that no other vertebrate genome will have this level of curation. Since the completion of human and mouse genomes there has been much debate about the trade-off between numbers of genomes sequenced, depth of shotgun sequence provided and degree of finishing for those genomes (Blakesley et al. 2004, Margulies et al. 2005b). With current sequencing capabilities this has become an issue of cost. The value of draft sequence is undeniable, however, issues of sequence coverage and accuracy must be accounted for in any studies planning to utilise these sequences. This thesis will illustrate examples of missing functional elements as a direct result of low sequence coverage (chapters IV and VI). As this thesis is concerned with identifying conserved regulatory elements in local regions, whilst generating lasting resources for the imprinting community, it was deemed important to generate comprehensive physical clone maps and finished sequences across each region in each species. Since the start of this study whole genome shotgun (WGS) sequencing for the species selected have been generated and in all except wallaby draft assemblies produced (Table I-1). In some cases it was therefore possible to make use of the WGS sequences, for example, to design probes for BAC library screening. It is also informative to compare WGS sequences with the finished sequences generated here (chapter IV).

Table I-1. Details of genome sequences for species used in this thesis.

Species	Common name	Current assembly version	UCSC code	Assembled Sequence length (Gb)	Status (coverage)	Reference
Homo sapiens	Human	NCBI build 36.1 (March 2006)	hg18	2.86	Finished	(International Human Genome Sequencing Consortium. 2004)
Mus musculus	House mouse	NCBI build 37 (July 2007)	mm9	2.56	Finished	(Waterston et al. 2002)
Macropus eugenii	Tammar wallaby	NA	NA	NA	Low (2x)	NP
Monodelphis domestica	South American short-tailed, grey opossum	Broad Institute (Jan 2006)	monDom4	3.50	Draft (6.5x)	(Mikkelsen et al. 2007)
Ornithorhynchus anatinus	Platypus	WUGSC v5.0.1 (March 2007)	ornAna1	1.84	Draft (6.0x)	NP
Gallus gallus	Red jungle fowl chicken	WUGSC v2.1 (May 2006)	galGal3	1.05	Draft (6.6x)	(Hillier et al. 2004)

Mus spretus is not shown here because there is no plan to sequence this genome. UCSC code, assembly description at the UCSC genome browser; WUGSC, Genome Sequencing Center, Washington University, St Louis; NA, Not yet assembled; NP, Not yet published.

1.4 Genome annotation

As noted above the true value of genome sequence lies in its annotation. At the Sanger Institute all finished sequences undergo a preliminary analysis and annotation. Sequences are entered into a highly automated and refined analysis pipeline developed by the 'anacode' (analysis coding) group before highly skilled computer biologists in the human and vertebrate analysis and annotation (HAVANA) group annotate gene structures based on experimental and gene prediction evidence. Figure I.2 provides a detailed description of the analysis and annotation processes performed.

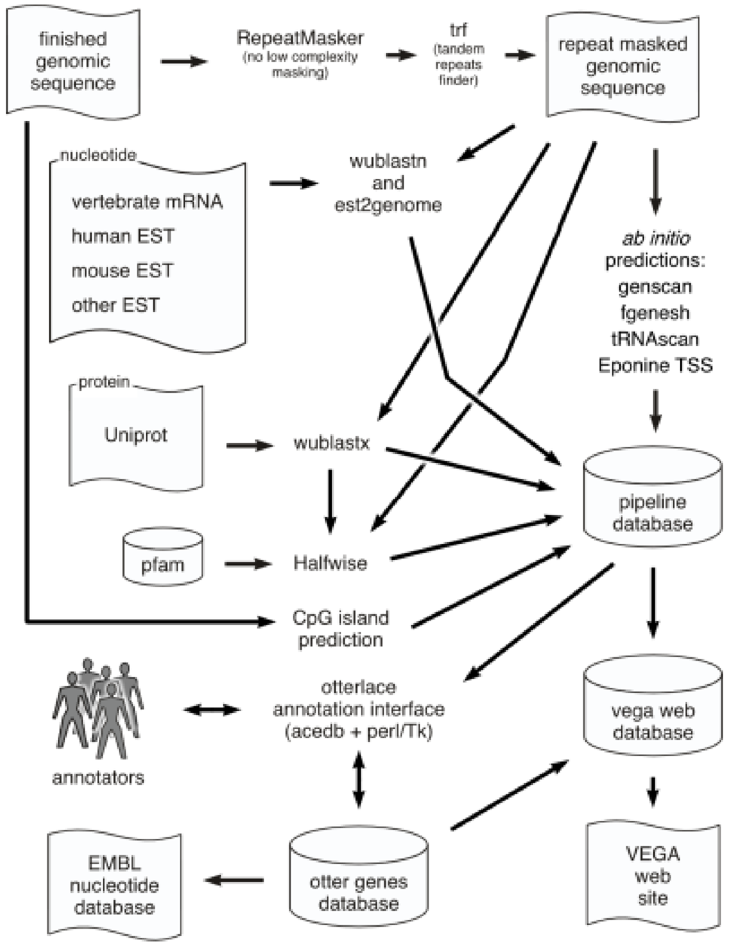

Figure I.2. The human and vertebrate analysis and annotation (HAVANA) pipeline.

The schema provides an organisation of the data flow from finished genomic sequence through to release of annotations in the public databases. With the exception of CpG island prediction using EMBOSS scripts (Rice et al. 2000) **all other analyses were performed with sequences masked for interspersed (RepeatMasker, http://ftp.genome.washington.edu/RM/RepeatMasker.html) and tandem (TRF,** Benson. 1999) **repeats. Nucleotide sequence databases are searched using wuBLASTN (http://blast.wustl.edu) and significant matches aligned to the repeat-free sequences using est2genome** (Mott. 1997)**. Translated nucleotide sequences are searched against the Uniprot (http://www.uniprot.org) protein databases. Annotation of protein domains within the Pfam database** (Bateman et al. 2004) **is performed using Genewise (Halfwise) software** (Birney et al. 2004)**. The *Ab initio* gene prediction algorithms genscan** (Burge and Karlin. 1997) **and fgenesh** (Salamov and Solovyev. 2000) **are run. Additionally tRNAscan** (Lowe and Eddy. 1997)**, to identify tRNA genes, and Eponine TSS** (Down and Hubbard. 2002)**, to identify predicted**

transcription start sites, are also used. Manual annotation of gene structures is performed in an implementation of ACeDB (otterlace, Searle et al. 2004)**. Data release is achieved via submission to EMBL and the VEGA website (http://vega.sanger.ac.uk). Figure reproduced from** (Ashurst et al. 2005)**.**

1.4.1 ENCODE – Annotation of the human genome

The ENCyclopedia Of DNA Elements (ENCODE) project aims to identify all functional elements in the human genome sequence (ENCODE Project Consortium et al. 2007, ENCODE Project Consortium. 2004). This Herculean task is made all the more difficult because the inventory of those functional elements we know most about, the protein-coding genes, is still incomplete.

In the May 2000 Cold Spring Harbour (CSH) Genome Biology meeting a "GeneSweep" competition was opened (by Ewan Birney, Ensembl) and the winner would receive the stakes and a signed leather-bound copy of "the double helix" by James Watson. The challenge for delegates of that meeting was to predict the number of (protein-coding) genes in the human genome (http://web.archive.org/web/20050428090317/www.ensembl.org/Genesweep/).

A strict definition of a gene was imposed such that "alternatively spliced transcripts all belong to the same gene, even if the proteins that are produced are different". The draft publications of the human genome sequences in 2001 (International Human Genome Sequencing Consortium. 2001, Venter et al. 2001) revealed an estimated 30,000 to 40,000 genes. At the Genome of *Homo Sapiens* CSH symposium in 2003 the winner was announced (Lee Rowen, Institute for Systems Biology) who had predicted 25,947 genes (Pennisi. 2003). Lee's educated guess was the lowest in a range extending to 153,478 genes with a mean of 61,710 (I guessed 51,000!). At that time the Ensembl Gene build 33 number was 24,500, noticeably lower than the estimates from the draft sequences and a far cry from the long-held estimate of 100,000 genes.

Ensembl currently lists 21,858 annotated known protein-coding genes in human, a further 1828 novel genes, 4150 RNA genes and 2136 pseudogenes. Although the numbers of protein-coding genes have remained relatively stable in recent years it is clear from the wealth of new data recently obtained from 1% of our genome (approximately 30 Mb), as a result of the ENCODE pilot project (ENCODE Project Consortium et al. 2007), that the numbers of identified ncRNA genes will increase. Furthermore there is clear evidence for multiple splice forms for protein-coding genes. Manual annotation by the HAVANA group of protein-coding genes in the ENCODE regions, termed GENCODE annotations (Harrow et al. 2006), reveal 2,608 transcripts clustered into 487 loci i.e. on average 5.4 transcripts per locus. Experimental validation of predicted protein-coding transcripts indicates that GENCODE annotations are at least 98% complete (Guigo et al. 2006).

The observation that much of the genome (approximately 90%) is transcribed and approximately 50% is spliced confirms that we know less about the human gene complement than we thought. Furthermore, genes frequently overlap both on the same and opposite DNA strands (ENCODE Project Consortium et al. 2007).

Following the ENCODE pilot project it is becoming clear that the previous definition of a gene is no longer adequate (Gerstein et al. 2007). The historical definition of a gene, as a unit of hereditary resulting in a specific characteristic of an organism, dates back to the concepts of Gregor Mendel (1866) and later Thomas Hunt Morgan (1915)(MENDEL. 1950, Morgan. 1915). The concept of a gene as a discrete locus in the genome no longer applies because it is well established that regulatory elements responsible for the correct expression of a gene product can lie great distances along the DNA sequence from the promoter, exons and so on. Of course, in 3-D space as a consequence of chromatin structure even apparently distant enhancers can lie in close proximity to the promoter (discussed further below).

By definition, ncRNA transcripts do not have a long and unambiguous open reading frame (ORF) and are therefore more difficult to predict computationally than protein-coding genes. The diverse functions of ncRNA transcripts include a role in protein synthesis by transfer RNAs (tRNAs) and ribosomal RNAs (rRNAs), transcriptional and translational regulation by miRNAs and RNA processing by snoRNAs. Large ncRNAs also exist with demonstrable function (e.g. *XIST*) or with as yet unknown function (e.g. *H19*, see chapter VI). The observation of high genomic transcription renders it quite likely that the number of ncRNAs in the human genome (including those yet to be categorised) will significantly increase over time (Bertone et al. 2004, Carninci. 2006, Cheng et al. 2005). Our increased knowledge of human genome transcription and regulation has led Gerstein and colleagues to propose a new gene definition: "A gene is a union of genomic sequences encoding a coherent set of potentially overlapping functional products" (Gerstein et al. 2007).

1.4.2 Enhancing human genome annotation

Comparison of human and rodent genomes indicated that approximately 5% of the human sequence is evolutionarily constrained despite only 2% being protein-coding (Gibbs et al. 2004, Waterston et al. 2002). The ENCODE consortium have subsequently confirmed that only 40% of the constrained bases lie in protein-coding exons or their untranslated regions (UTRs) (ENCODE Project Consortium et al. 2007). A further 20% of constrained bases overlap with known regulatory elements and therefore the remaining 40% of constrained bases are of no known function. Thus it remains a great challenge in biology to identify and characterise *cis*-regulatory elements in our genome. In contrast to the relative success of computational prediction of protein-coding exons, *in silico* identification of *cis*-regulatory elements is difficult because their syntax or grammar is largely unknown. Comparing DNA sequences between diverse species is a well documented means of identifying functionally important sequences (Boffelli et al. 2003, Hardison et al. 1997, Kellis et al. 2003, Loots et al. 2000, Margulies et al. 2003b, Visel et al. 2007, Woolfe et al. 2005). So what are these functional elements? There will surely be elements in our genome for which function has yet to be described and methods to identify them are introduced below. The functions that have been ascribed generally fit the category of transcriptional regulation.

1.4.3 Transcriptional regulation

Efforts are now underway to identify the function of the 20,000-25,000 genes or their products in our genome. We know relatively little about the identity and function of transcriptional regulatory elements and how they act to coordinate complex spatial and temporal gene expression. Simplistically *cis*-acting regulatory elements can be categorised into promoters, enhancers, silencers, insulators and locus control regions (LCR, including imprinting control regions) (Maston et al. 2006). These *cis*-acting sequences contain DNA binding sites for *trans*-acting factors that determine whether transcription is enhanced or repressed. Additionally, matrix attachment regions (MARs) are believed to have a structural role in the formation of active and silent domains of transcription.

Imprinted gene regulation involves complex interactions between all categories of regulatory elements listed above. The role of LCRs (or ICs) in the regions studied here have been well documented. Equally, the regions have been extensively studied for transcription and CpG islands and therefore most, if not all, promoters are known. By contrast relatively little is known about tissue specific enhancers and insulators and yet those characterised in the IC1 domain have been shown to play crucial roles in the regulation of imprinted gene expression (Bell and Felsenfeld. 2000, Hark et al. 2000, Leighton et al. 1995).

1.4.3.1 Enhancers

Cis-acting regulatory sequences that markedly increase the expression of a neighbouring gene are called enhancers. Enhancer activity was first identified following experiments in which segments of the SV40 tumour virus were observed to significantly increase transcription of a heterologous human gene with promoter (Banerji et al. 1981). The first mammalian enhancer identified was lymphocyte-specific, residing in the immunoglobulin heavy-chain locus (Banerji et al. 1983). Each enhancer element typically consists of multiple TFBSs each of which may only occupy 6 to 10 bases of DNA.

As of December 2007 the VISTA Enhancer Browser (http://enhancer.lbl.gov/) lists 309 elements with experimentally confirmed enhancer activities in transgenic mice. The number of enhancers identified through other reporter assays, both *in vivo* and *in vitro*, may be somewhat higher. Never-the-less, given the fact that many genes have multiple isoforms, each of which might be spatially or temporally expressed differently, there are likely many more enhancers to be found.

Enhancer (or indeed other regulatory) effects in humans have been identified in patients carrying translocations which result in the separation of regulatory elements from their target genes (Abbasi et al. 2007, Lettice et al. 2003). Data thus far shows that enhancer function is largely independent of orientation and distance from the gene promoter with which they interact, at least in 2-D space. Enhancers have been characterised which lie great distances from their target genes. For example, 7 enhancers of the human *DACH* gene reside in an 870 kb gene desert flanking this gene (Nobrega et al. 2003) and the enhancer ZPA regulatory sequence (ZRS) regulates mouse Sonic hedgehog (*Shh*) expression from over 1 Mb away (Lettice et al. 2003). Examples of enhancers lying within the introns of target genes, between neighbouring genes and even beyond neighbouring genes exist. So how do enhancers exert their function on gene promoters? Evidence is mounting for a model in which DNA looping brings distal enhancers to their site of action at a promoter (Celniker and Drewell. 2007). Chromosome conformation capture (3C, Dekker et al. 2002) studies within the mouse IC1 domain have demonstrated that DNA looping brings distal endodermal enhancers in contact with *IGF2* or *H19* promoters in an allele-specific manner (Murrell et al. 2004).

It is laborious to test extended DNA regions for enhancer activity by serial deletion analysis in reporter gene assays. However, evolutionary conserved regions (ECRs) are more likely to be enriched for functionally constrained elements and the technology now exists enabling us to clone and test relatively large numbers of ECRs in gene reporter assays (see chapter V). The success of this approach relies on the identification of ECRs which is largely determined by the ability to accurately align two or more sequences (see below).

1.4.3.2 Insulator elements

Insulator elements function to block genes from the inappropriate transcriptional affects of nearby genes, thus compartmentalising the genome into discrete functional domains. Insulator activity can be classified as enhancer-blocking or barrier (Gaszner and Felsenfeld. 2006). Barrier insulators are cis-acting regulatory sequences preventing the spreading of heterochromatin into euchromatic regions. Enhancer-blocker insulators block the long-range interactions of a distal enhancer with a proximal promoter when located between these elements and therefore compartmentalise the genome, preventing the inappropriate transcription of neighbouring genes or alleles. One of the best known examples of enhancer-blocker activity acting in an allele-specific manner is the *IGF2/H19* locus of human and mouse (Bell and Felsenfeld. 2000, Hark et al. 2000). Like enhancers, insulator elements contain DNA binding sites for transcription factors including the well characterized and highly conserved CTCF protein which contains 11 zinc finger domains. CTCF binding at the unmethylated *H19* DMD prevents access of downstream endodermal enhancers from acting on the *IGF2* promoter on the maternal allele. Methylation of the *H19* DMD on the paternal allele prevents binding of CTCF allowing the enhancers to activate the *IGF2* promoter. Mutation of the CTCF zinc fingers or DNA binding sites within the *H19* DMD prevent insulator function leading to biallelic expression of *Igf2* and demonstrate the critical importance of enhancer-blocking activity in the imprinting mechanism (Renda et al. 2007).

1.5 Sequence alignment

The comparison of genome sequences is an important tool with which to identify functional elements in the human genome (Cooper and Sidow. 2003, Loots et al. 2000, Nardone et al. 2004, Woolfe et al. 2005). The fundamental premise is that functionally important regions will tend to evolve more slowly than non-functional sequences because mutations in functional DNA are likely to be deleterious and are therefore selected against (Kimura. 1983). Thus, between any two sequences the degree of conservation is a function of their evolutionary distance and differences in local mutation rates (Hardison et al. 2003). The comparison of orthologous sequences to identify *cis*-regulatory elements has been termed phylogenetic footprinting (Tagle et al. 1988).

Tools used for the large-scale alignments of sequences are now readily available from the World Wide Web (WWW, Table I-2) and their underlying algorithms generally fall into two groups; global (sometimes referred to as hierarchical) and local.

1.5.1 Global alignments

Global alignments require co-linearity, i.e. those sequences from regions of conserved synteny are first defined, because the alignment will be applied across the entire lengths of all query sequences. Global alignment algorithms therefore work best with sequences of similar length and nucleotide composition. Typically global alignment programs take as input pairwise local alignments and output regions containing significant alignment in the same order and orientation between species. Examples of global alignment programs are given in Table I-2. These methods rely on 'chaining' which describes an ordered set of locally aligned segments such that the Nth local alignment coordinates are less than those of the (N+1)th. This increases the likelihood that the aligned sequences are truly orthologous and not paralogous.

1.5.2 Local alignment

Local alignments identify regions of similarity within long sequences that are often widely divergent overall. Specifically, the strategy adopted is to identify all similarities between two sequences and then combine these pairwise alignments into multiple alignments (Batzoglou. 2005). Typically local alignments begin with a short exact or imperfect match ("word") which is then used to initiate potentially larger alignments. The best known example is the basic local alignment search tool (BLAST, Altschul et al. 1990). However, more recent adaptations such as BLASTZ are optimized for large sequence alignments (Schwartz et al. 2003b). BLASTZ computes local alignments for sequences of any length based on the assumption that the input sequences are related and share blocks of high conservation, separated by regions lacking homology and varying in length. BLASTZ is therefore an appropriate tool for use in this thesis because the orthology of the sequences is known (from the prior mapping) but may be highly divergent (e.g. human-chicken comparisons). For whole genome sequence comparisons BLASTZ alignments are provided to MULTIZ which builds a multiple alignment from local pairwise alignments of a designated reference sequence with other sequences of interest. Alignments viewed in the UCSC genome browser have typically been generated using the MULTIZ program (Blanchette et al. 2004). Examples of local alignment programs are provided in Table I-2.

Table I-2. Features of local and global sequence alignment algorithms.

Alignment strategy	Local	Global
Example programs	BLASTZ (Schwartz et al. 2003b) TBA/MULTIZ (Blanchette et al. 2004) PatternHunter (Ma et al. 2002) MUMmer (Kurtz et al. 2004) DIALIGN (Morgenstern. 1999)	MLAGAN (Brudno et al. 2003) MAVID (Bray and Pachter. 2004) PARAGON ClustalW (Larkin et al. 2007) GRIMM-Synteny (Bourque et al. 2004) MAP2 (Ye and Huang. 2005) Mauve (Darling et al. 2004)
Frequently used algorithm	Smith-Waterman (Smith and Waterman. 1981)	Needleman-Wunsch (Needleman and Wunsch. 1970)
Genome-wide Sensitivity	High	Lower
Local sensitivity	Lower	High
Speed	Slow	Fast
Consequence of insertions or deletions (indels)	Short indels explicitly gapped. Long indels implicitly gapped or interpreted as missing data.	Short and long indels explicitly gapped
Examples of visualisation tools	zPicture (Multi-zPicture) (Ovcharenko et al. 2004a) PipMaker (Multi-PipMaker) (Schwartz et al. 2000) Mulan (Ovcharenko et al. 2005)	VISTA (Multi-VISTA) (Mayor et al. 2000)

Adapted from (Dewey and Pachter. 2006)

1.5.3 zPicture

With so many sequence alignment tools available and new ones emerging on a regular basis it can be difficult to choose the correct one. Alignment tools that incorporate highly detailed dynamic visualisation modules which facilitate the identification of potentially functional regions, for subsequent experimental testing, are required. In this thesis the alignment tool of choice has been BLASTZ (Schwartz et al. 2003b) run from within the zPicture server (http://zpicture.dcode.org/, Ovcharenko et al. 2004a). zPicture (for two species) and Multi-zPicture (for >2 species) are extensions to the widely used predecessors PipMaker and MultiPipMaker (Schwartz et al. 2000, Schwartz et al. 2003a). One advantage of multiple pairwise alignments over simple pairwise analysis is that loss of a given element within a single lineage does not hamper the identification of the element if it remains functional in the genomes of other species.

The zPicture server is highly intuitive to the molecular biologist and importantly offers numerous features of value. These include the customised real-time processing of alignment data and the ability to export ECRs to portals such as rVISTA (Loots and Ovcharenko. 2004) for analysis of conserved transcription factor binding sites (TFBS). Input sequences can either be imported from the UCSC genome browser, together with annotation of e.g. genes and repeats, or uploaded from a user's computer. User-defined visualisation parameters include: the choice of input sequences to be used as the reference, with which all others are compared; the output style for sequence homology i.e. percentage identity plots (PIP) in the form of unconnected dots or smooth trace (VISTA) conservation plots; ability to modify annotation such as the location of exons, introns and UTRs; the percent identity and length parameters for ECRs. This thesis makes use of an additional feature of the zPicture server, which is the facility to generate large dot-plots (chapter IV).

1.6 Informative species

If the parental conflict/kinship hypothesis is valid then oviparous (egg-laying) animals such as birds, reptiles and non-placental mammals e.g. the duck-billed platypus (monotreme) would not be expected to show genomic imprinting in early development because the mother lays down all of the nutrients they use in development before fertilisation. Also there is no influence of the paternal genome on the maternal energy contribution. Viviparous (live-bearing) animals would be expected to imprint because the embryo plays a role in determining the amount of resources it receives from its mother. Limited imprinting might also be expected in marsupial mammalian lineages such as the tammar wallaby (in which the developing foetus leaves the uterus early and migrates to the pouch for subsequent development). Supporting evidence of this has been found for both *Igf2* and *Igf2r* genes which have been shown to be biallelically expressed in the monotremes (platypus and echidna) and imprinted in the marsupials *Monodelphis domestica* and *Didelphis virginiana* (South and North American opossums, respectively) (Killian et al. 2000, Killian et al. 2001, O'Neill et al. 2000). The chicken genome which diverged from the human lineage over 310 Myr ago also does not show imprinting in the genes studied to date (Nolan et al. 2001, O'Neill et al. 2000, Yokomine et al. 2001, Yokomine et al. 2005). The mechanism of genomic imprinting in animals is, therefore, thought to have arisen during the radiation of marsupial mammals from monotreme mammals, some 148-166 Myr ago (Bininda-Emonds et al. 2007, Kumar and Hedges. 1998) (Figure I.3).

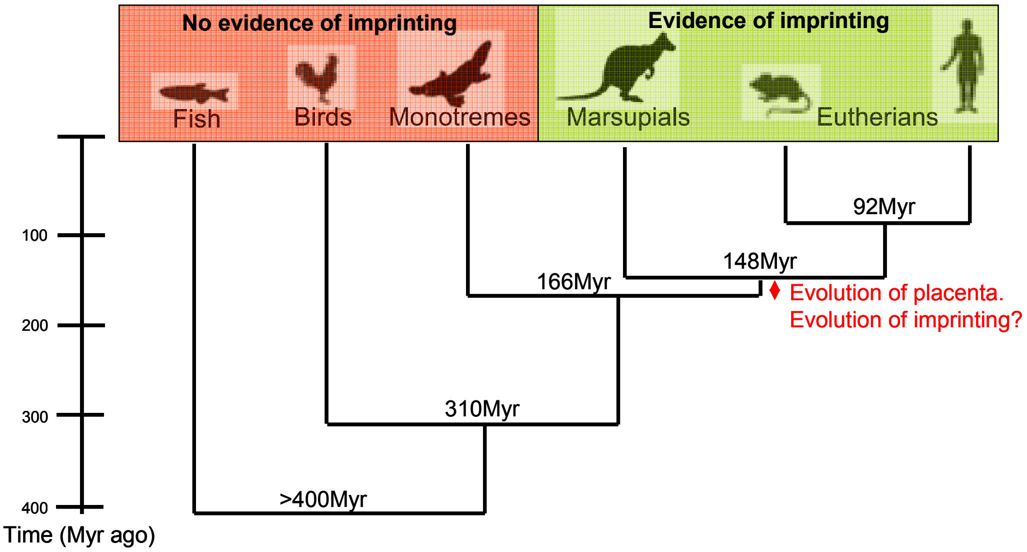

Figure I.3. Phylogeny of vertebrate species.

Mammals are divided into three groups; Monotremes, Marsupials and Eutherians. Eutherian mammals diverged approximately 148 Myr ago from the marsupial mammals, which in turn diverged from the egg-laying monotremes approximately 166 Myr ago. The apparent absence of genomic imprinting in monotremes and presence in marsupials suggests that the function and mechanism of imprinting evolved between 148 and 166 Myr ago and may have co-evolved with placentation. Divergence times taken from Bininda-Emonds et al. 2007, Kumar and Hedges. 1998**.**

At the onset of this project (August 2003) only four WGS sequence assemblies for vertebrates (human, mouse, rat and pufferfish) existed and were publicly available in genome browsers such as Ensembl or University of California, Santa Cruz (UCSC). As a consequence large evolutionary niches, spanning some 300 million years, were unrepresented. To date, the genomes of 28 vertebrates, including 21 mammals, have been sequenced to varying degrees of completion (see below) and many more are on the way (http://www.genome.gov/10002154). These include the first WGS sequences of a bird (chicken), monotreme (duck-billed platypus) and marsupial (South American grey short-tailed opossum). These are important genome additions to the vertebrate phylogenetic tree (Figure I.3) because previous sequence comparisons between human and, for example, mouse gave high false positive signals for regulatory elements due to insufficient time for mutations to accumulate in neutrally evolving sequence. As a consequence the presence of potential regulatory elements could be hidden (Margulies et al. 2003a). In contrast, the genomes of fish or birds are difficult to align to human and therefore whilst the specificity for detecting conserved sequences is high the sensitivity is low. For example, comparison of the human genome with distant outgroups such as fish (divergence approximately 400 Myr ago, Figure I.3) has provided valuable information on both protein-coding genes and regulatory elements, at least for regulators of development including many transcription factors such as SOX21, PAX6, HLXB9 and SHH (Woolfe et al. 2005). The chicken genome (divergence more than 310 Myr ago, Figure I.3) has also been shown to be a good predictor of coding genes but may be too distantly related in order to identify many non-coding regulatory regions in the human genome (Hillier et al. 2004). The monotreme and marsupial sequences therefore fill an important evolutionary niche.

The utility of marsupial and monotreme genomic sequences to characterise mammalian regulatory elements through comparative analyses has been demonstrated for a few small regions, including the lymphoblastic leukaemia derived sequence 1 (*LYL1*) locus (Chapman et al. 2003) and a region of chromosome 7q13.3 encompassing the cystic fibrosis transmembrane conductance regulator (*CFTR*) gene (Margulies et al. 2005a). In the context of this study, sequence comparison between species in which genomic imprinting has been demonstrated and those in which it has not (at least for the few loci studied) is expected to reveal functional elements involved in the mechanism of imprinting

(Figure I.3). Such elements could be crucial to the 'reading' of imprinting signals (marks) and might include enhancer, silencer or insulator elements.

1.6.1 Placental mammals (eutherians)

The infraclass eutheria contains all placental mammals that are nourished via the placenta during gestation. The placenta of eutherian mammals is considered much more highly developed than in metatherians (section 1.6.2), because it invasively implants into the uterine wall and nourishes the foetus during prolonged intrauterine gestation. For the purpose of this thesis the terms eutherian or placental mammals are used interchangeably.

1.6.1.1 Human

The earliest fossils of anatomically modern humans (*Homo sapiens*) were found in Africa and dated to 130,000 years ago. The closest living relatives of *Homo sapiens* are the Bonobo (*Pan paniscus*) and Common (*Pan troglodytes*) chimpanzees that diverged from the human lineage some 6.5 Myr ago. Comparison between the genome sequences of chimpanzee and human indicate that 98.4% of the DNA sequence is identical (Chimpanzee Sequencing and Analysis Consortium. 2005). This level of sequence conservation is of utility in phylogenetic shadowing i.e. the identification of functional DNA elements unique to one lineage (Boffelli et al. 2003). However, this level of sequence identity is of lesser use in identifying regulatory elements common to both species (phylogenetic footprinting). The choice of species to compare to human is therefore of fundamental importance and should reflect the biological question(s) being asked. In the context of this thesis all other model organisms discussed below are studied to further our knowledge of human gene regulation in regions harbouring imprinted gene orthologues. Consequently all genome comparisons are made using human as the reference.

1.6.1.2 Mouse

The mouse and human lineages diverged from one another approximately 92 Myr ago (Figure I.3). Mice are the most commonly utilized animal research model and therefore the *Mus musculus musculus* genome (strain C57BL/6J) was sequenced in parallel with that of the human and continues to provide a key experimental tool with which to interpret the human genome (Waterston et al. 2002). Much of what we know about genomic imprinting was elucidated in the laboratory mouse. Indeed, nuclear transplantation studies in mice revealed that both paternal and maternal genomes were required for successful embryogenesis (McGrath and Solter. 1984, Surani et al. 1984). Furthermore many of the first phenotypes ascribed to imprinting anomalies were observed in mouse genetic studies in which Robertsonian and reciprocal translocations were employed to generate uniparental disomies (parental isodisomies) or uniparental duplications of whole chromosomes or regions (Cattanach and Kirk. 1985).

Genomic resources available from wild mice such as *Mus spretus* (the Western Mediterranean short-tailed mouse) are complementing those of the laboratory mice strains derived from *Mus musculus domesticus*, *Mus musculus musculus* and *Mus musculus castaneus* sub-species (Guenet and Bonhomme. 2003). Evolutionarily the *Mus spretus* and *Mus musculus* species diverged from one another approximately 1.5-2.0 Myr ago and the *Mus musculus* sub-species last shared a common ancestor 0.5-1.0 Myr ago (Figure I.4, Guenet and Bonhomme. 2003).

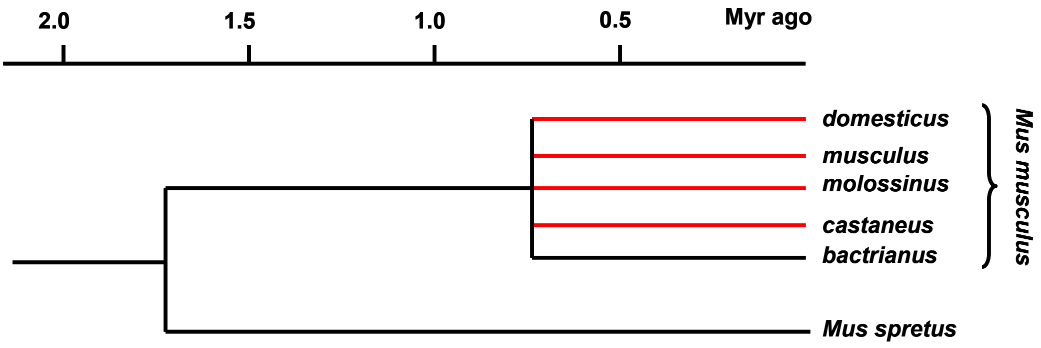

Figure I.4. Partial evolutionary tree of the genus *Mus*.

The Western Mediterranean short-tailed mouse (*Mus spretus*) last shared a common ancestor with the *Mus musculus* complex species around 1.5-2.0 Myr ago. The 5 *Mus musculus* species diverged from one another some 0.5-1.0 Myr ago and those at the origin of classical laboratory strains are highlighted with red branches. Adapted from Guenet and Bonhomme. 2003.

The geographical ranges of *Mus musculus* and *Mus spretus* overlap but there is little evidence of natural hybrids reflecting their different species. Laboratory mice strains can produce viable offspring with *Mus spretus*. However, male offspring are sterile. Congenic strains (containing a region of interest that is selectively transferred by sexual reproduction from its original background into another strain) offer a powerful tool to dissect the genetic (or epigenetic) control of complex traits (Burgio et al. 2007). The SD7 congenic strain contains the distal portion of *Mus spretus* chromosome 7 (including both IC1 and IC2 imprinted sub-domains) on an otherwise C57BL/6J genetic background (Dean et al. 1998). The SD7 strain has been widely used to study the parental origin of alleles to establish or confirm the imprinting status of genes on distal chromosome 7 (Fitzpatrick et al. 2007, Weber et al. 2003). Such studies have relied on the prior identification of polymorphisms (restriction fragment length polymorphisms (RFLPs) or single nucleotide polymorphisms (SNPs)) to distinguish parental alleles. In the majority of cases polymorphisms are identified *ad hoc*, usually by sequencing polymerase chain reaction (PCR) products. A comprehensive catalogue of sequence variants between *Mus musculus* and *Mus spretus* in the IC1-IC2 domains would be of great utility for those researchers using SD7 mice in allele-specific studies. Chapter III discusses the physical mapping and sequencing of bacterial artificial chromosome (BAC) clones across the entire IC1-IC2 region and an analysis of the sequence is presented in chapter IV.

1.6.2 Marsupial mammals (metatherians)

Marsupial (metatherian) mammals occupy one of three main groups in extant mammalian phylogeny (Figure I.3). There are 270 species of marsupials mainly found in Australasia and South America with only one, the Virginia opossum (*Didelphis virginiana*), found in North America today. Marsupials are thought to have originated in North America before colonizing the supercontinent of Gondwana which separated 38-84 Myr ago to form South America, Australia and Antarctica (Veevers. 1991). American and Australian marsupials are thought to have diverged from one another approximately 60 Myr ago (Nilsson et al. 2004), equivalent to the time when human and, for example, lemur lineages split. Metatherian species diverged approximately 148 Myr ago from the eutherian mammals and are collectively termed therians. Divergence between therian and monotreme (prototherian) mammals occurred approximately 166 Myr ago (Figure I.3, Bininda-Emonds et al. 2007).

The position of marsupial species on the phylogenetic tree renders them of particular interest in evolutionary biology. With the exception of some opossums (see below) females have a pouch or marsupium (from which the name ‘marsupial’ is derived) in which their young are reared from weeks 4-5 in their development. The developing embryo crawls up its mother’s belly and into the pouch where it locates a teat and feeds there for many weeks. A consequence of the early birth in marsupials is that the placenta is more primitive (and less invasive) than its eutherian counterpart. However, the placenta is fully functional and expresses hormones essential to mammalian pregnancy and parturition (the process of giving birth) (Renfree and Blanden. 2000, Renfree and Shaw. 2000). Leaving the relative safety of the marsupial yolk sac in the womb exposes the embryo to greater risk of infection. However, a powerful anti-microbial agent, possibly beta-lactoglobulin, is believed to be present in the complex milk provided to the tiny embryo in the pouch (Ambatipudi et al. 2007, Lefevre et al. 2007). The same anti-microbial molecule has not been found in the milk of placental mammals whose young develop their own immune systems within the safety of the mother's womb. Milk composition changes with the development of the pouch young from birth to the joey stage. Indeed the same mother can produce different milk compositions to her offspring, of different developmental ages, suckling on different teats. This marsupial adaptation to reproduction and nutrition may be particularly important during difficult seasons when carrying a large foetus to term is dangerous to the mother and her offspring.

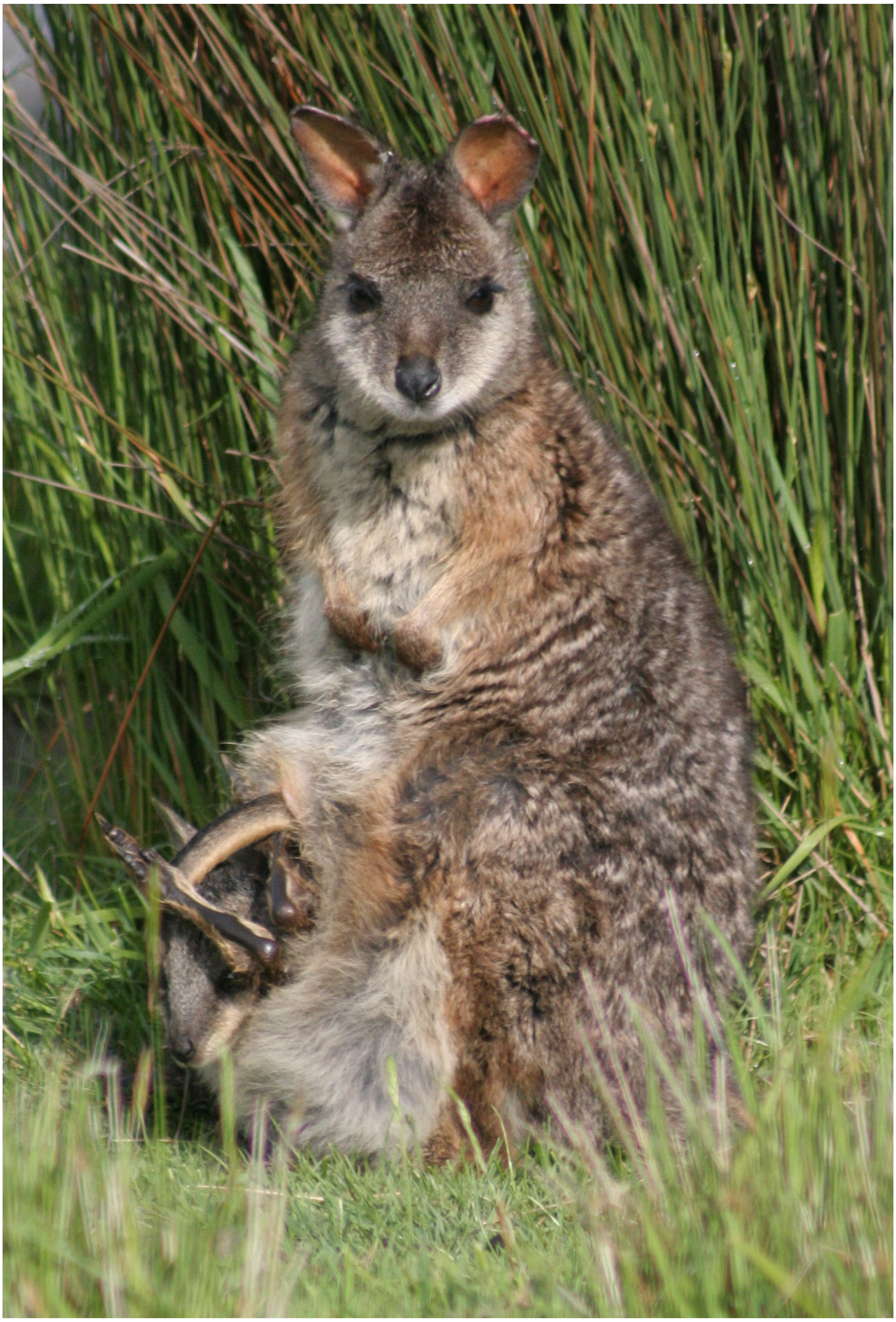

Figure I.5. Tammar wallaby (*Macropus eugenii*) with large pouch young ('joey').

Photo courtesy of Geoff Shaw (Department of Zoology, The University of Melbourne).

Analysis of genomic imprinting in marsupials in which there is minimal maternal investment in the embryo promises to provide answers to the question - *why did* *genomic imprinting evolve?* A literature search reveals that thus far 6 imprinted genes, in four regions, have been identified in marsupial species (*IGF2*, *IGF2R*, *MEST(PEG1)*, *PEG10*, *SGCE* and *INS*) (Ager et al. 2007, Killian et al. 2000, O'Neill et al. 2000, Suzuki et al. 2005, Suzuki et al. 2007). The generation of maps and sequence in some of these and other regions will provide the necessary resources with which to identify additional imprinted genes and regulatory elements in marsupials.

1.6.2.1 Tammar wallaby

The tammar wallaby (*Macropus eugenii,* Figure I.5) is a small (approximately 8 kg) member of the kangaroo family and due to its availability and ease of handling is the species of choice for much marsupial research. This research includes the study of immunogenetics (Deakin et al. 2007, Siddle et al. 2006), neurobiology, neoplasia, developmental and reproductive biology and genomic imprinting (Ager et al. 2007, Graves and Westerman. 2002, Wakefield and Graves. 2005). The tammar wallaby is found in large numbers on Kangaroo Island off the South coast of Australia. The wallaby genome comprises about 3.6 Giga-basepairs (Gb) and is cytogenetically arranged as 8 large chromosomes (2n=16) (http://www.agrf.org.au/Default.aspx?tabid=89). The embryo develops *in utero* for only 26 days before birth. At birth the embryo measures approximately 16 mm and weighs only 400 mg, the size of a broad bean. This stage of development is equivalent to a 40-day human embryo or 15-day mouse embryo (Tyndale-Biscoe and Renfree. 1987). The neonate uses forelimbs (hindlimbs are not yet developed) to climb up to the mother’s pouch to locate a teat where it will gain all the nutrients it requires for further development.

The utility of marsupial sequences to aid in the annotation of the human genome has been shown in the *LYL1* gene region in which the conserved sequences were shown to be exonic or regulatory in nature including transcription factor binding sites (TFBSs) (Chapman et al. 2003). Furthermore, new human genes have been identified as a direct consequence of human-wallaby genomic sequence comparisons. Examples include the RNA binding motif protein, X-linked (*RBMX*) and related genes (Delbridge et al. 1999, Lingenfelter et al. 2001). It is therefore anticipated that wallaby sequences generated in this thesis will serve to identify potentially novel genes and regulatory elements in the human genome. Inclusion of a highly divergent marsupial species, the South American short-tailed grey opossum (*Monodelphis domestica*), in the study will also improve the significance of any findings relating to an ancestral imprinting mechanism.

1.6.2.2 South American, grey short-tailed opossum

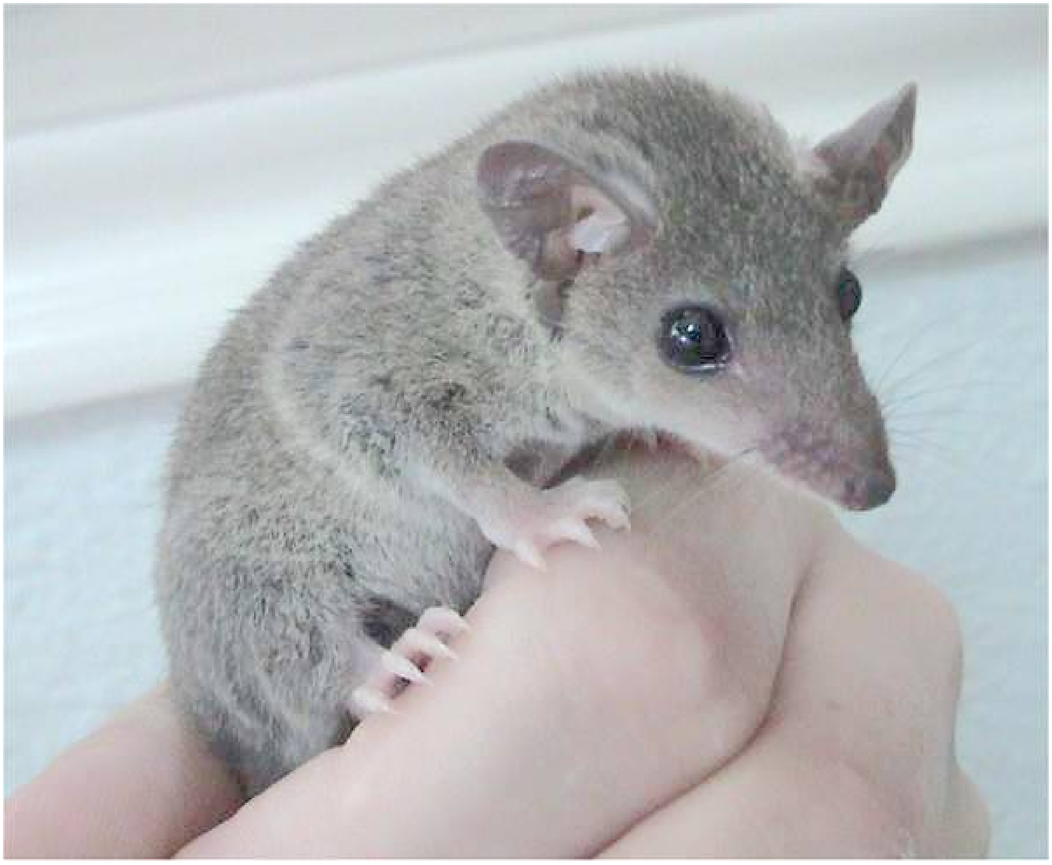

Figure I.6. South American, grey short-tailed opossum (*Monodelphis domestica*). Photo reproduced from Wikimedia Commons (http://en.wikipedia.org/wiki/Image:Monodelphis_domestica.jpg) under the Creative Commons Attribution ShareAlike 2.5 licence.

The grey, short-tailed opossum (*Monodelphis domestica*, Figure I.6) is found in South America and is one of about 100 opossum species found world-wide. It is an ideal marsupial research model due to its small size, ability to breed in captivity and the ease in which neonates can be studied. Female opossums, unlike most other marsupial species, lack a pouch and therefore the young cling to the mother's teats and can simply be removed for study. *Monodelphis domestica* was the first marsupial genome sequenced (to draft status) and provides a unique perspective on the organization and evolution of mammalian genomes (Mikkelsen et al. 2007).

This opossum has been the subject of much research into the mechanisms of genomic imprinting (Killian et al. 2000, Lawton et al. 2007, Murphy and Jirtle. 2003, O'Neill et al. 2000, Rapkins et al. 2006, Vu et al. 2006, Weidman et al. 2004, Weidman et al. 2006, Yokomine et al. 2006) and complements research on the distantly related tammar wallaby.

1.6.3 Monotreme mammals (prototheria)

Only five species of monotremes (prototheria) are known to exist today including the duck-billed platypus (*Ornithorhynchus anatinus*), a short-beaked echidna (*Tachyglossus aculeatus*) and three species of long-beaked echidna (genus *Zaglossus*). Monotremes are the most ancient of extant mammals having last shared a common ancestor with humans approximately 166 Myr ago (Figure I.3 and Bininda-Emonds et al. 2007). As with all mammals monotremes are warm-blooded, have hair on their bodies, a single lower jaw bone, three middle ear bones and produce milk. The young suckle milk not from defined nipples but from mammary glands secreting milk through skin patches on the mother's abdomen. Unlike viviparous eutherians, monotremes are oviparous. The young develop within a small leathery egg (like that of a reptile) which is retained *in utero* for approximately 28 days before being laid and incubated in a burrow for a further 10 days.

Genomic imprinting has not been observed in monotremes, although only a few loci known to be imprinted in therians have been studied to date (see below). This limited data would appear to support the kinship theory which predicts that imprinting exists in species in which there is at least some contribution of maternal resources to the embryo and polyandry (mating with multiple males) is common. The principle organ responsible for nutrient exchange between mother and offspring is the placenta which is noticeably absent from oviparous animals.

Imprinting has therefore been hypothesised to have co-evolved with placentation (Figure I.3, Reik and Lewis. 2005). Whether imprinting exists at a very early developmental stage *in utero*, when the monotreme egg is rapidly increasing in volume through the acquisition of nutrients, remains to be determined. Of course, nutrient transfer from mother to offspring is not limited to exchange across the placenta or membranes of an egg but also in lactation. It will be of great interest to see whether monotreme genes involved in lactation and/or the behaviour of suckling are imprinted. To test these hypotheses high-quality genomic sequences are required for many more genes in regions of known therian imprinting.

1.6.3.1 Platypus

The duck-billed platypus (*Ornithorhynchus anatinus*, Figure I.7) is the only extant member of the family Ornithorhynchidae. The platypus is a very timid creature and captive breeding programmes have generally not been successful. The low numbers of extant monotreme species, early divergence from therian mammals and mix of mammalian, reptilian and unique morphological and physiological features have resulted in the platypus being a popular research subject in evolutionary biology.

Figure I.7. The monotreme platypus (*Ornithorhynchus anatinus*).

Photo reproduced from Wikimedia Commons (http://en.wikipedia.org/wiki/Image:Platypus.jpg) under the GNU Free Documentation Licence.

In molecular biology the platypus genome has already revealed important information about the sex chromosome systems of amniotes. Unlike all other mammals platypus have ten sex chromosomes which form a multivalent chain in meiosis (Grutzner et al. 2004). No sex-determining orthologue of the SRY gene has been identified on any of the platypus X chromosomes. Some of the X chromosomes have homology with the human X chromosome and others have homology with the Z chromosome of birds. Therefore, in some regards the platypus genome shares features with mammals or birds as befits its phylogenetic position (Figure I.3).

Until recently very little genomic sequence data existed for platypus. However, Margulies and colleagues have reported the sequence analysis of a 1.26 Mb region of the platypus genome orthologous to human 7q31.3 (Margulies et al. 2005a). This region contains the cystic fibrosis transmembrane conductance regulator (*CFTR*) and neighbouring genes and was thus referred to as the 'greater *CFTR* region'. Only

14% of the platypus sequence could be aligned with human, considerably lower than the 45-70% alignments observed between human and non-primate eutherians (Thomas et al. 2003). The genomic landscape of the platypus greater *CFTR* region would appear to contain some unique mammalian features. Despite containing all orthologous genes in the same order and orientation as in human, the greater *CFTR* region of platypus is 24% smaller. Contrary to expectations that larger genomes are the result of increased repeat contents, the repeat content of platypus was 44.9%, somewhat higher than the 40.3% repeat content in the corresponding human region. Furthermore over half of the platypus repeats were SINE elements, a level not observed in any other vertebrate sequenced to date. The platypus C+G content in the greater *CFTR* region (49.5%) is also high when compared with other mammalian genomes. To establish whether these features are specific to the greater CFTR region or representative of the platypus genome as a whole requires further sequencing and analysis.

Genomic imprinting has not been observed in the platypus genome, at least for the foetal growth regulator genes *IGF2*, *IGF2R* and *UBE3A* which are imprinted in therians (Killian et al. 2000, Killian et al. 2001, Rapkins et al. 2006, Weidman et al. 2004). As a member of the monotremes at the base of mammalian phylogeny the platypus is an important addition to any study of imprinted gene regulation.

1.6.4 Birds

Birds (class Aves) are the most diverse tetrapod vertebrates in existence with approximately 10,000 species. All birds are warm-blooded, feathered and are oviparous. However, the eggs, unlike those of reptiles or monotremes, are hard-shelled, made largely of calcium carbonate. Birds last shared a common ancestor with mammals approximately 310 Myr ago (Figure I.3). Evolutionary biologists had long contested the origin of birds, however, the 19th Century discovery in Germany of the fossilised bird *Archaeopteryx lithographica* with its reptilian teeth, clawed forelimbs and long bony tail indicated its evolutionary descent from the therapod dinosaurs.

1.6.4.1 Chicken

Figure I.8. Red Jungle Fowl and White Leghorn chickens (*Gallus gallus*).

Photo of Red Jungle Fowl pair courtesy of ARKive.org under the Creative Commons Attribution-Non-commercial-Share Alike 3.0 licence. Photo of White Leghorn chicken courtesy of http://www.clipartguide.com.

Domestic chickens are of immense commercial importance world-wide for their meat and egg production. Chickens are also the most widely used bird models in biological research. Genomic resources (e.g. BAC libraries, ESTs) for the Red Jungle Fowl (thought to be the ancestral breed) and White Leghorn chicken breeds are available. Furthermore, the draft genome sequence of the Red Jungle Fowl (*Gallus gallus*) is proving to be a very useful distant outgroup for comparison with the human genome (Hillier et al. 2004). The chicken genome is separated from the human genome by approximately 1.7 substitutions per site in orthologous, neutrally evolving sequences (Hillier et al. 2004). Aligning orthologous sequences whose mean genetic distance exceeds one substitution per site is reported to be problematic (Margulies et al. 2005b). However, any similarity observed between human and chicken sequences are very likely due to constrained sequences of functional importance.

Being oviparous chickens make no post-fertilization contribution of maternal resources to their offspring and therefore in keeping with the kinship hypothesis it is of no surprise that genes imprinted in some therians, *IGF2*, *IGF2R*, *INS* and *ASCL2* are all biallelically expressed in chickens (Nolan et al. 2001, O'Neill et al. 2000, Yokomine et al. 2005). Interestingly, birds do show some characteristics that are in accordance with the kinship hypothesis such as polygamy and post-hatching parental care. The theory predicts that imprinting will evolve only if paternal alleles can influence maternal investment in offspring. Since maternal investment in offspring continues after fertilization in birds as well as mammals it is, in theory at least, plausible for imprinting to exist in birds.

The mapping of quantitative traits such as those responsible for egg production, quality and viability has revealed autosomal regions with parent-of-origin specific effects (Tuiskula-Haavisto and Vilkki. 2007). Many of the mapped quantitative trait loci (QTL) reside on chicken macrochromosomes and lie in or close to regions of conserved synteny with therian imprinted gene clusters. These regions also exhibit asynchronous DNA replication, an epigenetic feature associated with imprinted gene clusters (Dunzinger et al. 2005). Although parent-of-origin monoallelic gene expression has not been observed in chicken perhaps the chromatin environments of the macro-chromosomes were conducive for the subsequent evolution of the genomic imprinting mechanism. As is the case for monotremes more experiments are required to verify the phylogenetic distribution of genomic imprinting in birds.

1.7 Genomic regions studied

Eight distinct genomic regions orthologous to known imprinting gene clusters of human and mouse have been selected for study here. Additionally the DNA methyltransferase 1 (*DNMT1*) gene, responsible for *de novo* and maintenance DNA methylation, and therefore of fundamental importance to the imprinting mechanism, was included (Figure I.9). The selection of regions for vertebrate mapping and sequencing reflects those most intensively studied, many of which are of interest to imprinting groups in the Cambridge region. These include the Reik and Kelsey groups in the laboratory of developmental genetics and imprinting at the Babraham Institute, and the Ferguson-Smith group in the department of physiology, development and neuroscience at the University of Cambridge. A strong collaborative relationship between the groups led to the formation of the SAVOIR (Sequence Analysis of Vertebrate Orthologous Imprinted Regions) consortium (http://www.sanger.ac.uk/PostGenomics/epicomp). The human and mouse cytogenetic and sequence locations of the 9 regions studied together with human diseases associated with epigenetic anomalies are provided in

Table I-3.

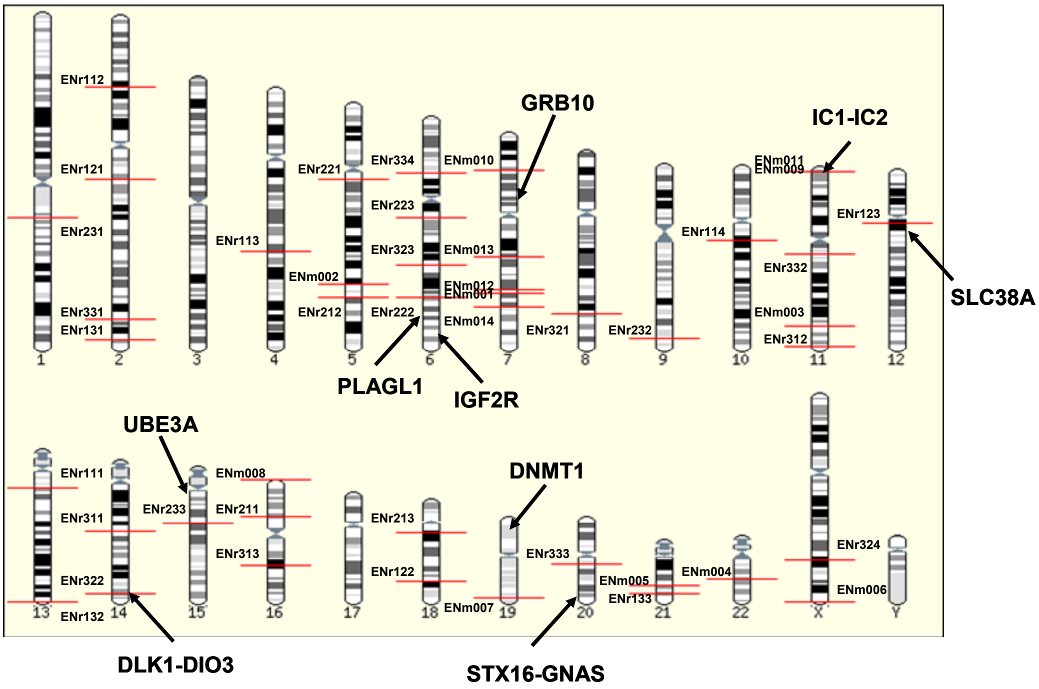

Figure I.9. SAVOIR regions studied.

The regions mapped and sequenced in this thesis are indicated by arrows on the human chromosome ideograms. Red horizontal lines indicate regions selected for study by the ENCODE pilot project (ENCODE Project Consortium. 2004)**. Only the ENm011 ENCODE region overlaps with a selected SAVOIR region (IC1-IC2, near the short-arm telomere of chromosome 11).**

Table I-3. SAVOIR regions studied and associated human diseases.

Region	Human genome location (Cytogenetic/sequence)	Mouse genome location (Cytogenetic/sequence)	Associated human disease(s)	OMIM ID
IC1-IC2	11p15.5/chr11:1707725-3630000	7qF5/chr7:142150001-143750000	Beckwith-Wiedeman syndrome	130650
STX16-GNAS	20q13.32/chr20:56650001-56950000	2qH4/chr2:173719260-173989679	Pseudohypoparathyroidism Ib; Albright hereditary osteodystrophy	603233; 103580

DLK1-DIO3	14q32/chr14:100262982-101099542	12qF1/chr12:109901030-110728904	Pituitary adenomas; Skeletal abnormalities; Hemangiomas	176290;
				605636;
				608149;
				601038

SLC38A2-4	12q13.11/chr12:45038238-45506002	15qF1/chr15:96515428-96883990	Unknown	608065
IGF2R	6q25.3/chr6:160310121-160447573	17qA1/chr17:12525764-12613064	Hepatocellular carcinoma	147280
GRB10	7p12/chr7:50625259-50828652	11qA1/chr11:11830511-11937358	Russel-Silver syndrome	601523
PLAGL1	6q24.2/chr6:144303130-144427428	10qA2/chr10:12781107-12820083	Transient neonatal diabetes mellitus	601410
UBE3A	15q11.2/chr15:23133489-23235221	7qB5/chr7:59096621-59174596	Angelman syndrome	105830
DNMT1	19p13.2/chr19:10105022-10166811	9qA3/chr9:20657612-20703275	Neoplasia	126375

Human sequence locations are from the NCBI build 36 (March 2006, hg18) genome assembly. Mouse sequence locations are from the NCBI build 36 (Feb 2006, mm8) genome assembly. OMIM ID, Online Mendelian Inheritance of Man identifier.

1.7.1 IC1-IC2 domains

The largest region being studied, and the focus for this thesis, spans 1.9Mb of human 11p15.5, mouse distal chromosome 7(qF5), and harbours the imprinted genes dysregulated in the overgrowth disorder Beckwith-Wiedemann Syndrome (BWS,

Table I-3). The telomeric boundary of the region studied lies at the Cathespin D (*CTSD*) gene and the centromeric boundary extends to the ADP-ribosyltransferase 5 gene (*ART5*) in human (Figure I.10). The telomeric boundary was deliberately selected to coincide with the distal end of a manually selected ENCODE region (ENm011, Figure I.9). ENm011 occupies 606 kb (31.5%) of the whole region and was selected by the ENCODE consortium (ENCODE Project Consortium et al. 2007, ENCODE Project Consortium. 2004) because of the interest in imprinted gene regulation. Imprinted genes in this IC1 domain include the maternally expressed *H19* gene, and paternally expressed *IGF2*, *IGF2* antisense (*IGF2AS*) and insulin (*INS*) genes. Mouse *Igf2* was the first imprinted gene discovered following the observation that phenotypes were different in mice carrying targeted mutations of this gene and dependent upon the parental allele transmitting the mutation (DeChiara et al. 1991).

.

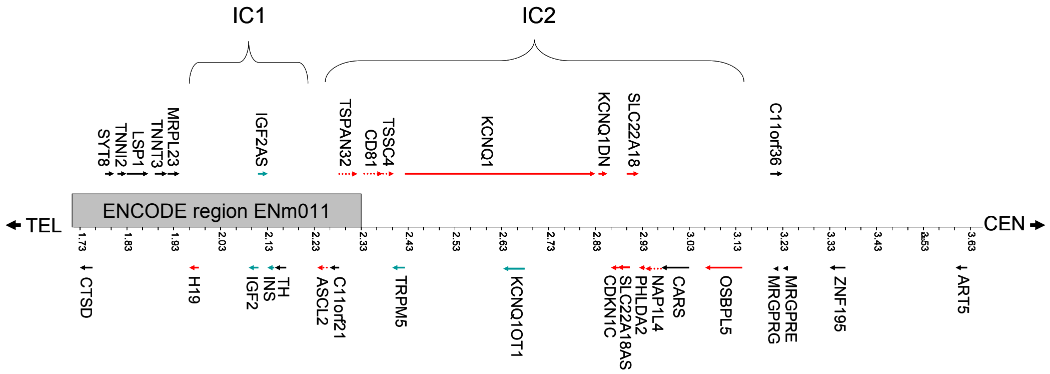

Figure I.10. Human chromosome 11p15.5 region.

Coordinates are given according to the NCBI build 36 (March 2006) genome assembly, scale in Mb. Human genes are indicated by arrows on both forward strand (top) and reverse strand (bottom). Black arrows indicate genes that are biallelically expressed. Coloured arrows indicate imprinted genes; blue, paternally expressed; red, maternally expressed. Broken arrows are reportedly imprinted in mouse but not human. A 606 kb interval at the telomeric end of the 11p15.5 region corresponds to the ENCODE region ENm011 (grey box, Figure I.9).

The neighbouring IC2 domain only partially overlaps with the ENm011 region. This 1 Mb imprinted domain contains mainly maternally expressed genes regulated by the KvDMR located within intron 10 of the potassium voltage-gated channel, KQT-like subfamily, member 1 (*KCNQ1*) gene and harbours the promoter for the paternally expressed *KCNQ1OT1* antisense ncRNA. In mice the *Kcnq1ot1* transcript is required for paternal repression of imprinted genes in the IC2 domain (Mancini-Dinardo et al. 2006).

1.8 Aims of the thesis

In this era of genomics the availability of large-scale biological resources and novel technologies enable unprecedented investigations of gene function and regulation. In recent years there has also been a growing appreciation for the role of epigenetics in health and disease and the fields of genetics/genomics and epigenetics/epigenomics have come together to address fundamental biology questions such as the evolutionary origin and mechanism of genomic imprinting.

This thesis has two main goals; to further our understanding of gene regulation in known imprinting clusters and to elucidate the evolutionary origins of genomic imprinting.

Specifically this thesis discusses:

1) The physical mapping and sequencing of diverse vertebrate species, strategically selected because of their phylogenetic position, in 9 different genomic regions harbouring imprinted gene orthologues or regulators of imprinting control (chapter III).
2) A comparative analyses of the generated sequences (11.5 Mb) including broad genomic landscape features (inter-species genome expansions/contractions, evolutionary breakpoints) and fine-scale features (gene, repeat, C+G and polymorphism contents) (chapter IV).
3) An investigation of the function of identified ECRs, conserved for at least 148 Myr, in the IC1 and IC2 domains with the specific aim of identifying and characterising novel enhancer elements (chapter V).

4) A detailed analysis of the marsupial *H19* candidate region delineated by ECRs to determine the ancestral mechanism of imprinting in the IC1 locus. This includes the identification of both wallaby and opossum *H19* ncRNAs, encoding a conserved miRNA (miR-675) and a DMR that harbours predicted CTCF binding sites and which demonstrates insulator function in an experimental assay. Thus all the major hallmarks of the eutherian *IGF2-H19* imprinting system are present in the marsupials making it the most conserved epigenetic mechanism discovered so far (chapter VI).