Chapter VII - Discussion

7.1 Summary

This thesis has described the physical mapping and sequencing of diverse vertebrate species, including mammals from all three orders, in 9 different genomic regions, harbouring imprinted gene orthologues or regulators of imprinting control. In total 10.8 Mb of high-quality finished sequence and a further 700 kb of near finished sequence has been generated for subsequent analyses (chapter III). A comparative analysis of the sequences, including broad genomic landscape features (such as inter-species genome expansions/contractions and evolutionary breakpoints) and fine-scale features (such as gene, repeat, C+G and polymorphism contents) was discussed. This included the finding that segmental duplications giving rise to gene families is largely responsible for the relative genomic sequence lengths in orthologous regions between species and not repeat content. This analysis also revealed the extraordinary C+G and repeat contents of the platypus genome (chapter IV). An investigation of the function of identified ECRs, conserved for at least 148 Myr of parallel evolution, in the IC1 and IC2 domains revealed the power of this approach to identify and characterise novel enhancer elements (chapter V). Indeed, almost 50% of sequences with previously unknown functions could be ascribed function. However, there are caveats with the approach including the restricted number of cell lines used and the potential for cell lines to have functionally diverged from the primary cells used to create them. Finally, a detailed analysis of the marsupial *H19* candidate region delineated by ECRs was performed to determine the ancestral mechanism of imprinting in the IC1 locus. These analyses resulted in the identification of both wallaby and opossum *H19* ncRNAs, encoding a conserved miRNA (miR-675) and a DMR that harbours predicted CTCF binding sites which demonstrate enhancer-blocking insulator function in a reporter-gene 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). The significance of these findings and how they relate to the work of others is discussed below.

7.2 Imprinting evolution

One of the principal aims of this thesis was to further our knowledge of the evolutionary origins of the genomic imprinting mechanism. Following observations that most imprinted genes occur in clusters which are co-ordinately regulated by epigenetic mechanisms (Reik and Walter. 2001), researchers turned to look for the phylogenetic distribution of imprinting. These studies demonstrated the imprinting of genes involved in resource transfer in viviparous mammals but not in oviparous taxa including monotremes and birds, thus supporting the parental conflict/kinship hypothesis (chapter I). However, very little is known of the evolution of molecular mechanisms controlling imprinting. Unlike in eutherian mammals, in which X-chromosome inactivation is random, the X-chromosome in marsupials is always paternally imprinted (inactivated, Cooper et al. 1971, Sharman. 1971). Surprisingly the *Xist* ncRNA which is essential for X-inactivation in eutherians does not perform this role in marsupials. Indeed, the homologue of *Xist* in marsupials is a protein-coding gene (Duret et al. 2006). Thus, dosage compensation of genes on the X chromosome is achieved differently between marsupial and eutherian mammals. Work stemming from this thesis (in collaboration with Guillaume Smits and Wolf

Reik at the Babraham Institute) has resulted in the identification of the *H19* gene in marsupials which is maternally expressed and paternally imprinted (using methylation) at the germline DMD upstream of *H19*. Furthermore, this thesis has shown that the wallaby DMD contains CTCF binding sites that function as an enhancer-blocking insulator (chapter VI). The boundary model of imprinting regulation at the IC1 locus (Arney. 2003) is therefore conserved between eutherians and metatherians and must have been present in the therian ancestor. This epigenetic system is therefore the most evolutionary conserved mechanism discovered to date.

This thesis can not directly answer the question: why did genomic imprinting evolve? However, the paternal expression of the marsupial foetal and post-natal growth factor *IGF2* (O'Neill et al. 2000, Suzuki et al. 2005) and maternal expression of the *H19* gene (this study) that prevents *IGF2* over-expression is entirely consistent with the parental conflict/kinship hypothesis (Wilkins and Haig. 2003).
As discussed in chapter I, several hypotheses have been suggested to explain when and how the imprinting mechanism arose; including that it was driven by X-inactivation or from an ancestrally imprinted chromosome. Using BAC resources physically mapped through the course of this thesis, and in collaboration with the Ferguson-Smith groups in Cambridge, we have demonstrated that mammalian orthologues of imprinted genes are dispersed throughout the autosomes of platypus and tammar wallaby karyotypes. These data, together with observations that the chicken orthologues of imprinted genes are also spread throughout the genome (Dunzinger et al. 2005), indicates that mammalian imprinted genes did not originate on a common imprinted autosome or the X chromosome (Edwards et al. 2007). Instead, the data suggests that a step-wise, adaptive process has evolved at each

imprinted gene cluster with the gain or loss of the imprinting mechanism as the need arose.

The hypothesis that imprinting mechanisms arose from a host defence mechanism against parasitic DNA (Barlow. 1993, McDonald et al. 2005) has been supported by the finding of suppression of parasitic elements by the deamination of methylated cytosine residues to thymine residues destroying the transposable elements (Yoder et al. 1997). It is therefore of great interest to study the repeat and C+G contents of sequences from imprinted and non-imprinted species. Analysis of platypus sequences spanning 762 kb in the orthologous IC1-IC2 region revealed an extraordinary repeat content of 70% of which 39% were SINE elements (chapter IV). Independent studies have correlated imprinted genes with SINE exclusion (Greally. 2002, Luedi et al. 2007). The high density of SINE repeats reported here in platypus, which shows no imprinting of *IGF2*, is consistent with these findings.

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 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 they do not wholly account for such extreme levels. Since transposable elements tend to attract DNA methylation in order to suppress their transcription, the relative reduction in C+G content in these sequences may reflect the process of 5mC to T mutation by deamination.

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°C, low for a warm-blooded mammal (usually about 37°C) and intriguingly its genome has a level of CpGs between those of eutherian mammals and cold-blooded fish (Jabbari and Bernardi. 2004 and this study). This raises the possibility that the ancestral vertebrate genome had high CpG and methylation levels and that over the course of evolution there has been 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). We might speculate that in the therian ancestor methylation silencing of DNA provided a major selective advantage to the mother in viviparous species which in turn was counteracted by paternal suppression of maternal alleles, hence parental conflict. Once this system was finely balanced and evolutionarily fixed any retrotransposition of repeats into the region upsetting the methylation balance would be selected against.

7.3 Improving human genome annotation

The map and sequence resources presented in this thesis are not only critical to addressing the overall aims of the thesis but have been used to improve human genome annotation through comparative sequence analysis. This was illustrated by the identification of a novel gene that is conserved in all mammals studied and partially overlaps the *LSP1* gene at 11p15.5 (chapter IV). Conserved novel endodermal enhancers within the IC1-IC2 regions have also been identified and characterised (chapter V). The function of additional constrained sequences (ECRs) were predicted based on correlation with epigenetic data from the ENCODE project (ENCODE Project Consortium et al. 2007). The success of comparative sequence analysis is largely dependent upon the quality of the underlying sequences and sequence alignment methods.

7.3.1 Benefits of finished sequence

There have been huge technical advances in DNA sequencing in recent years (Bentley. 2006, Fredlake et al. 2006, Ryan et al. 2007, Shendure et al. 2004). For 30 years the sequencing method of choice has been the traditional Sanger chain termination chemical sequencing reaction. (Sanger et al. 1977). The new methodologies include micro-fluidic devices, sequencing by hybridisation and sequencing by synthesis. All new technologies strive for greatly reduced reagent volumes and cost whilst delivering extremely high-throughput (for review see Bentley. 2006). These sequencing technologies are proving to be extremely useful for re-sequencing applications where short reads are compared to a reference sequence. However, accurate *de novo* sequencing (techniques that do not depend on any prior knowledge of the sequence) of large (e.g. non-viral or -bacterial) genomes or genomic regions has yet to be proven using these novel technologies. With efforts focusing on ultra high-throughput (re-)sequencing (see Archon X-PRIZE for genomics: www.xprize.org) we should be cautious not to lose sight of the importance of high-quality genome sequence. Typically this entails the generation of a highly automated shotgun sequence followed by a directed, less automated and more costly 'finishing' phase.

The extent to which new genome sequences should be finished is the subject of debate (Blakesley et al. 2004, Green. 2007). Looking in the genome browsers today it is clear that a strategy to generate multiple incomplete sequences has been chosen over fewer complete sequences. This approach represents a compromise between phylogenetic breadth and depth of sequence redundancy (linked to coverage and accuracy) in a given species. This is not to say that analyses of increasing numbers of 2x coverage genome sequences do not reveal some interesting biology. Broad orthology between related species, lineage-specific and ancient repeat contents, partial gene and other evolutionary constrained sequences and polymorphisms can be identified from low coverage genome sequences. There are, of course, limitations of this approach. Analyses requiring complete features (e.g. genes or repeats) or their relative order and orientation cannot be determined from incomplete sequence. Furthermore, low coverage sequencing struggles to resolve segmental duplications (She et al. 2004). Thus, in this study, the identification and characterisation of duplicated genes and pseudogenes (e.g. *TNFRSF* and *KRTAP5* gene families) derived from segmental duplications probably would not have been possible without finished sequence. Also, without complete sequence coverage in the wallaby or opossum IC1 regions it would not have been possible to identify the *H19* gene and regulatory elements in these marsupials. Indeed, the opossum *H19* gene is not represented in the draft WGS assembly of the opossum genome (Mikkelsen et al. 2007).

Consideration should be given to generations of future experimenters using the sequences as foundations on which their research builds. Whilst the limitations of reference sequences may be evident to those of us who have been involved in their generation, to the thousands of scientists looking at consensus sequences in the genome browsers it may not be so clear. The mapping and sequencing strategy adopted in this thesis can be used to improve the sequence quality for targeted regions of draft quality genomes by using these WGS assemblies for probe generation to screen BAC libraries. This does assume the availability of a BAC library for the species of interest but these are now widely available for diverse species (http://www.genome.gov/10001844). The additional cost of finishing sequences in the short term will pay dividends in both cost and time to all those that use them in the long term.

7.3.2 Improving sequence alignment and functional element prediction

Changes in genome sequencing strategies have created an urgent requirement for user friendly methods for comparing them. Currently there are large numbers of tools and servers for users to align their sequences or imported sequences from public databases. Some of these tools were introduced in chapter I. The problem facing the molecular biologist is which tool or tools to use and the decision can be an important one because subsequent research will rely upon it. The BLASTZ alignment method, used in this study to identify ECRs in the zPicture server, has been used to indirectly compare species sequences using multiple pair-wise alignments (e.g. A vs B, A vs C,...). However, alignment tools which enable direct multi-species sequence comparisons and take into account phylogenetic branch lengths and local neutral background substitution rates have been recently developed (Cooper et al. 2005, Prabhakar et al. 2006, Siepel et al. 2005). These include Gumby (Prabhakar et al. 2006) and PhastCons (Siepel et al. 2005). The Gumby conservation analysis (http://pga.jgi-psf.org/gumby/) is automatically performed when DNA sequences are submitted to the mVISTA server and conserved sequences are displayed using RankVISTA (Frazer et al. 2004). PhastCons is the analytical engine behind the conservation tracks displayed at the UCSC genome browser and is part of the PHylogenetic Analysis with Space/Time models (PHAST) package (Siepel et al. 2005). PhastCons is based on the statistical model of sequence evolution called a phylogenetic hidden Markov model (phylo-HMM). Although similar to VISTA, PhastCons considers more than two species and considers the phylogenetic relationship between sequences to be aligned. Like Gumby, this model also goes beyond percent identity, allowing for multiple substitutions per site and a higher frequency of transitions than transversions that are commonly observed. Preliminary experimental tests of these statistical methods reveal that they outperform percent identity plots in their ability to detect functional enhancer elements from global alignments of human-mouse-rat sequences (Visel et al. 2007). It would be of interest to see how these new tools perform on the more evolutionary diverse sequences generated in this thesis, in particular whether they can identify additional ECRs for functional evaluation.

7.4 Future perspectives

Imprinting and X-inactivation mechanisms are present in therian mammals but absent from birds. Resource allocation from mother to offspring occurs via the placenta in therians but also in lactation. It is therefore of key importance to study these mechanisms in the monotremes platypus and echidna. Monotreme mothers should manifest the same need as therians to conserve energy during lactation but do not show imprinting of *IGF2*. However, it has not been shown that monotreme *IGF2* is biallelically expressed in all developmental stages. Since we now know that the IC1 imprinting mechanism existed in the ancestor of therians it is also important to determine which regulatory elements (e.g. the *H19* gene, miR-675, DMD or CTCF binding sites) were required for the emergence of monoallelic expression of *IGF2*. To address these questions it is imperative to complete the clone map and sequencing in this region of platypus and/or echidna. Echidna BAC library filters are available for screening and may prove to be more amenable for mapping and sequencing than platypus has been (chapters III and IV).

Although analyses of gene regulation in this thesis have focussed on the IC1-IC2 imprinted domains, high quality contiguous sequences have been generated for 8 other regions. It will be informative to perform similar analyses across these different regions to assess how genomic imprinting mechanisms have regionally adapted. The availability of generated sequences in public databases means that the imprinting community are already beginning to address these issues.

As shown in chapter V, almost half of the ECRs conserved at least since our last common ancestor with wallaby have no known function. In addition to the enhancer and insulator functions studied in this thesis it is possible to experimentally test ECRs for promoter and silencer functions (reviewed in Maston et al. 2006). Additionally, ECRs may represent structural features of the genome such as matrix attachment regions (MARs) which are AT-rich sequences capable of associating with the nuclear matrix or scaffold (Laemmli et al. 1992). MARs are believed to be important in the formation of active and silent domains of transcription. Indeed, MARs positioned adjacent to the *IGF2* DMRs have been shown to associate with the nuclear matrix in a parental-specific manner and are therefore likely important in the regulation of genomic imprinting (Weber et al. 2003). The sequences of experimentally defined MARs are not highly conserved but genomic MAR assays will be required to exclude a structural role for ECRs.

To identify the function of all evolutionarily constrained sequences and further our understanding of transcriptional regulation, it will be necessary to study many more cell-lines, or better still primary cells, from different developmental stages and environmental conditions. Only then will we be able to unravel the complexities of spatial and temporal gene expression.

Finally, selected functional elements identified using strategies adopted in this thesis should be studied in knock-out mice models for further analysis of genetic and epigenetic imprinting diseases.

7.5 Conclusion

This thesis has shown the potential of genomics to further our understanding of epigenetic phenomena. The availability of high-quality clone maps and sequence has enabled a deeper understanding of the evolutionary origins of genomic imprinting and its regulatory mechanisms. Comparative sequence analysis is a valuable tool with which to enhance human genome annotation and will increasingly be used, in combination with other approaches, to unravel the functional and evolutionary histories of our genome. Finally, the resources generated in the course of this study have been publicly released to serve as a significant and lasting resource to be used by the imprinting community as well as groups studying vertebrate genome biology.