Evolution and gene regulation of the genomic imprinting mechanism

Thesis

Evolution and gene regulation of the genomic imprinting mechanism

by

Andrew James Mungall B.Sc. (Hons)

A thesis submitted in partial fulfilment of the requirements of the Open University (Life and Biomolecular Sciences discipline) for the degree of Doctor of Philosophy

21st December 2007

The Wellcome Trust Sanger Institute

Open University

Abstract

Genomic imprinting describes an epigenetic mechanism by which genes are active or silent depending on their parental origin. Imprinting exists in plants and mammals, but how this monoallelic expression mechanism has evolved is not understood at the molecular level. Here I describe the mapping, sequencing and analysis of vertebrate orthologous imprinted regions spanning 11.5 Mb of genomic sequence from species with and without genomic imprinting. In eutherian (placental) mammals, imprinting can be regulated by differential DNA methylation, non-coding RNAs, enhancers and insulator elements. The systematic sequence comparison of the *IGF2-H19* imprinting cluster, in eutherians and marsupials (tammar wallaby and opossum), has revealed the presence of the enigmatic non-coding RNA *H19* in marsupials. Furthermore, we have characterised the marsupial *H19* expression status and identified key regulatory elements required for the germline imprinting of the neighbouring *IGF2* gene. All the major hallmarks of the imprinting mechanism of the *IGF2-H19* locus were found to be conserved in therian mammals. In mammals, this imprinting system is therefore the most conserved germline derived epigenetic mechanism discovered so far.

The high-quality genomic sequences have provided early glimpses of the genomic landscapes for species such as the monotreme platypus and marsupial tammar wallaby for which little was previously known. Comparative sequence analysis was used to identify candidate regulatory elements in the neighbouring imprinting centre 1 and 2 regions of human chromosome 11p15.5. Nine novel enhancer elements were identified following *in vitro* gene-reporter assays and correlation of conserved sequences with recent ENCODE data revealed probable functions for a further 24 elements.

This project has led to the formation of the Sequence Analysis of Vertebrate Orthologous Imprinted Regions (SAVOIR) consortium and resources developed here are being used by the imprinting community to further our knowledge of the evolution of the genomic imprinting mechanism.

Acknowledgements

I'm very grateful to many people for their help, advice and encouragement over the course of this project. I'll start by thanking all past members of Team62 (1993-2007). Each of you played a role in generating an environment conducive to fun, friendship and cutting edge science. It's a shame that the good things had to come to an end and I will miss you all.

I am indebted to my supervisors Ian Dunham and Wolf Reik. The productive and fun collaborative projects between Sanger and Babraham groups all stem from a Keystone meeting held at Taos, New Mexico in 2002. This goes to prove, what I've always known, that skiing and science do go well together!
I am particularly grateful to Ian for pushing me to register for the Ph.D. and his mentorship for all these years. I hope we will work (and ski) together again.

There are many groups at the Sanger Institute I thank for their expertise and helpfulness. These are: Jackie Bye, Carol Carder, Paul Hunt, Mark Madison, Klaudia Walter, Matt Jones and the sub-cloning teams, Sarah Sims, Karen Oliver and the shotgun sequencing teams, Stuart McLaren and the auto-prefinishing team, Lucy Matthews and her finishing team. I also thank Nancy Holroyd and the faculty small sequencing projects team, James Gilbert and the anacode group and Charlie Steward from the HAVANA group. For their invaluable scripts and bioinformatic know-how I thank Dave Beare and Carol Scott.

I have thoroughly enjoyed meeting and working with groups from the SAVOIR consortium, including the Ferguson-Smith, Kelsey and Reik groups. Special thanks go to Guillaume Smits (for his infectious enthusiasm and long phone calls!) and Carol Edwards (for the shared interests). I've learnt a lot from you all.

I'm very grateful to John Collins, Luc Smink and Ian Dunham for the critical reading of this thesis and Ian Sudbery for assistance with the formatting. Thanks to Christina Hedberg-Delouka, Jane Rogers and Alex Bateman for support on the Committee of Graduate Studies and my thesis committee group; Manolis Dermitzakis, Mark Ross, Wolf and Ian for their insight and Stephan Beck for his third party monitoring as well as taking an interest in my studies.

The completion of my thesis will come as a great relief to my family. I've been so fortunate to have your inspiration, love and support. It's been a testing time but we made it. Thanks especially to Karen, Susie and Caitlin for your patience and understanding. I love you all.

Finally, there's a real danger that I've neglected to thank some of you. If you were hoping for a mention here then please read on...you may not find your name but at least someone else will have read my thesis!

Table of Contents

List of Figures

List of Tables

Abbreviations used in this thesis

Abbreviation	Description
3C	Capturing Chromosome Conformation
aa	Amino Acid
ACeDB	A C. elegans DataBase
AGI	Arizona Genomics Institute
AGP	A Golden Path
AGRF	Australian Genome Research Facility
BAC	Bacterial Artificial Chromosome
BCM-HGSC	Baylor College of Medicine – Human Genome Sequencing Center
BLAST	Basic Local Alignment Search Tool
BLAT	Basic Local Alignment Tool
bp	Base pair(s)
BSA	Bovine Serum Albumin
BWS	Beckwith-Wiedemann Syndrome
C+G	Cytosine and guanine content
cDNA	Complementary DNA
CDS	CoDing Sequence
CFTR	Cystic fibrosis transmembrane conductance regulator
CTCF	CCCTC binding factor
CUGI	Clemson University Genomics Institute
DMD	Differentially methylated domain (5' of H19 gene)
DMR	Differentially methylated region
DNA	DeoxyriboNucleic Acid
EBI	European Bioinformatics Institute
ECR	Evolutionary Conserved Region
EMBL	European Molecular Biology Laboratoriums
EMBOSS	European Molecular Biology Open Software Suite

ENCODE	ENCyclopedia Of DNA Elements
ePCR	Electronic PCR
ERV	Endogenous Retrorvirus
EST	Expressed Sequence Tag
FASTA	DNA and protein file format
FISH	Fluorescent In Situ Hybridisation
FPC	FingerPrinting Contigs
Gb	Giga-basepairs
H3K9	Histone 3 lysine 9
H3K27	Histone 3 lysine 27
HAVANA	Human And Vertebrate Analysis aNd Annotation
HGP	Human Genome Project
HTML	Hypertext Markup Language
HUGO	Human Genome Organisation
IC	Imprinting Centre
Imprintace	Imprinting implementation of ACeDB
kb	kilo base pairs
kg	kilogram
LCR	Locus Control Region
LINE	Long Interspersed Nuclear Element
LTR	Long Terminal Repeat
MAR	Matrix Attachment Region
Mb	Megabase pairs
mg	Milligram
MIR	Mammalian-wide Interspersed Repeat
mm	Millimetre
MGSC	Mouse Genome Sequencing Consortium
MIR	Mammalian-wide Interspersed Repeat
MMU7	Mus Musculus chromosome 7
mRNA	Messenger RNA
miRNA	Micro RNA

Myr	Million years
NCBI	National Center for Biotechnology Information
ncRNA	Non-coding RNA
NIH	National Institutes of Health
nt	Nucleotide
OMIM	Online Mendelian Inheritance of Man
ORF	Open Reading Frame
PAC	P1-derived Artificial Chromosome
PCR	Polymerase Chain Reaction
pg	Picogram
PIP	Percentage Identity Plot
PSV	Paralogous Sequence Variants
PWS/AS	Prader-Willi/Angelman Syndrome
QH	Quantitative Hypervariable
QTL	Quantitative Trait Loci
RACE	Rapid Amplification of cDNA Ends
RFLP	Restriction Fragment Length Polymorphism
RNA	Ribonucleic acid
rRNA	Ribosomal RNA
RT-PCR	Reverse Transcription PCR
SAVOIR	Sequence Analysis of Vertebrate Orthologous Imprinted Regions
SCF	Standard Chromatograph Format
SINE	Short Interspersed Nuclear Element
snoRNA	Small nucleolar RNA
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variants
STS	Sequence Tagged Site
TF	Transcription Factor
TFBS	Transcription Factor Binding Site
TPF	Tile Path Format
TRF	Tandem Repeat Finder

tRNA	Transfer RNA
TSS	Transcription Start Site
UCSC	University of California, Santa Cruz
UMIST	University of Manchester Institute of Science and Technology
UPD	Uniparental disomies
pUPD	Paternal UPD
UTR	UnTranslated Region
VEGA	Vertebrate Genome Annotation
VISTA	Program for visualising global DNA sequence alignments
rVISTA	Regulatory VISTA
WGS	Whole Genome Shotgun
WUGSC	Washington University Genome Sequencing Center
WWW	World Wide Web
XCI	X chromosome inactivation

Publications arising from this work

Guillaume Smits*, Andrew J. Mungall*, Sam Griffiths-Jones, Paul Smith, Delphine Beury, Lucy Matthews, Jane Rogers, Andrew J. Pask, Geoff Shaw, John L. Vandeberg, John R. McCarrey, the SAVOIR consortium, Marilyn B. Renfree, Wolf Reik, Ian Dunham (2007). A conserved epigenetic mechanism controls the *H19* non-coding RNA and *IGF2* imprinting in therians. *Submitted.*

CA. Edwards, AJ. Mungall, L. Matthews, E. Ryder, DJ. Gray, AJ. Pask, G. Shaw, JAM. Graves, J. Rogers, I. Dunham, MB. Renfree, AC. Ferguson-Smith; and the SAVOIR consortium. (2007). The evolution of an imprinted domain in mammals. *Submitted.*

Carol A. Edwards*, Willem Rens*, Oliver Clarke, Andrew J. Mungall, Timothy Hore, Jennifer A. Marshall Graves, Ian Dunham, Anne C Ferguson-Smith, Malcolm A Ferguson-Smith. (2007). The evolution of imprinting: chromosomal mapping of orthologues of mammalian imprinted domains in monotreme and marsupial mammals. *BMC Evolutionary Biology* **7**:157.

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.

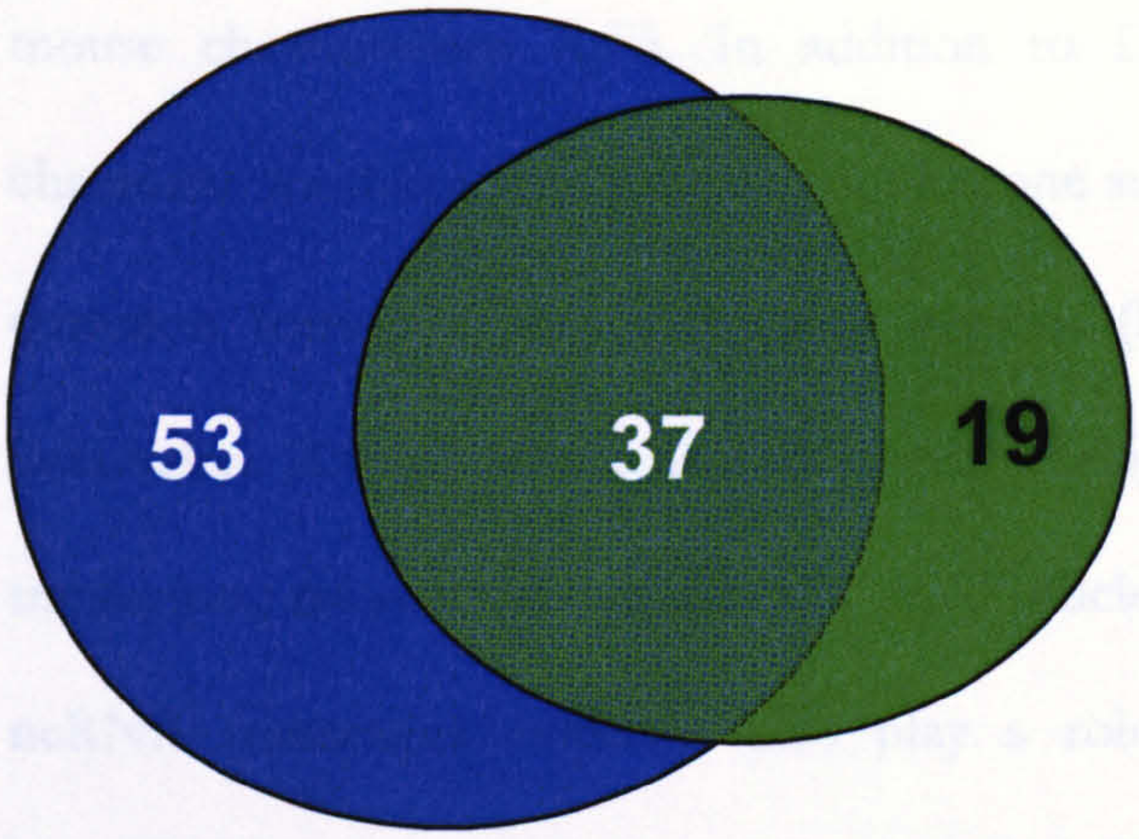


Figure I.1. Venn diagram of mouse and human imprinted genes.

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.

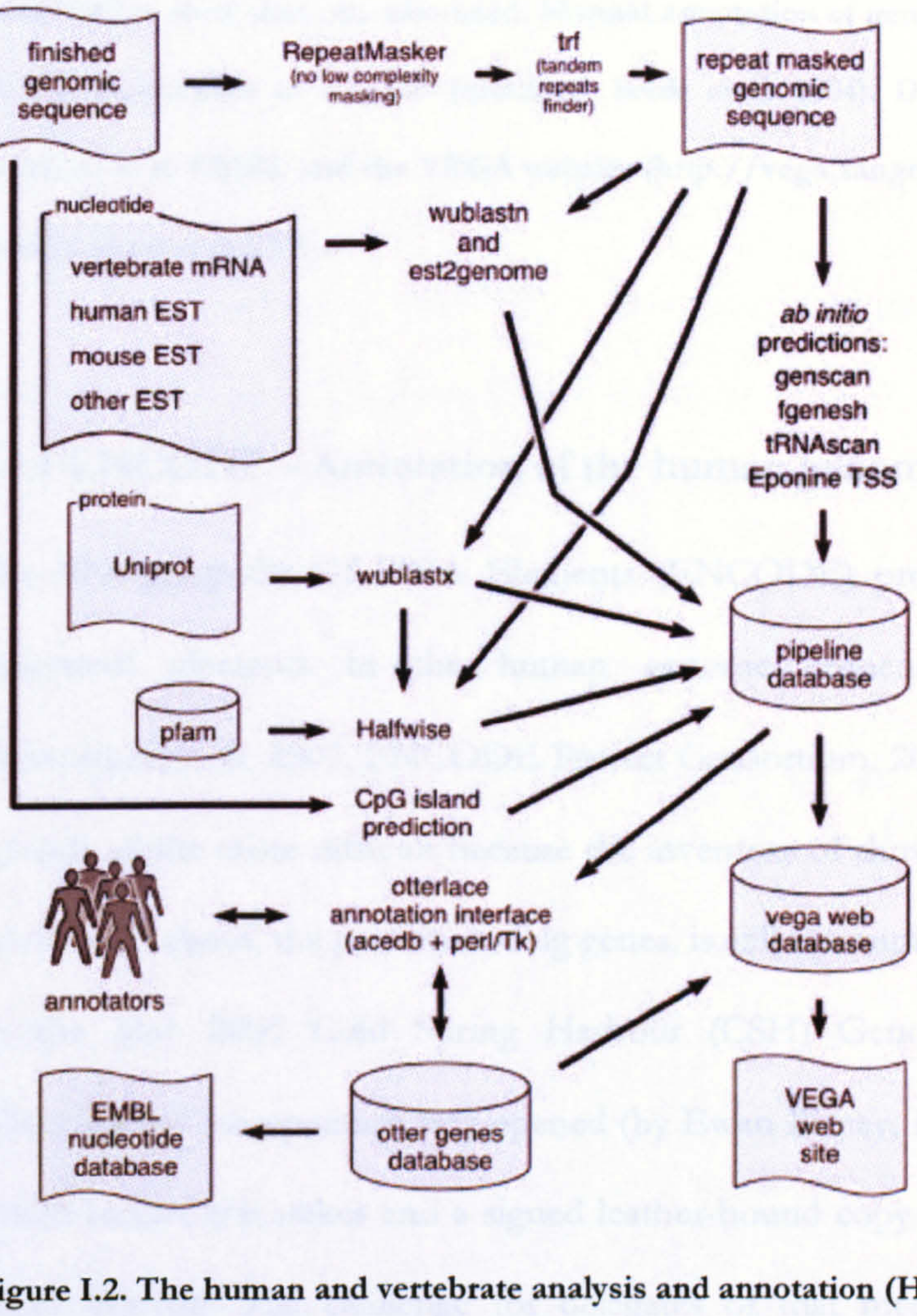


Figure I.2. The human and vertebrate analysis and annotation (HAVANA) pipeline.

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 N^{th} 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).

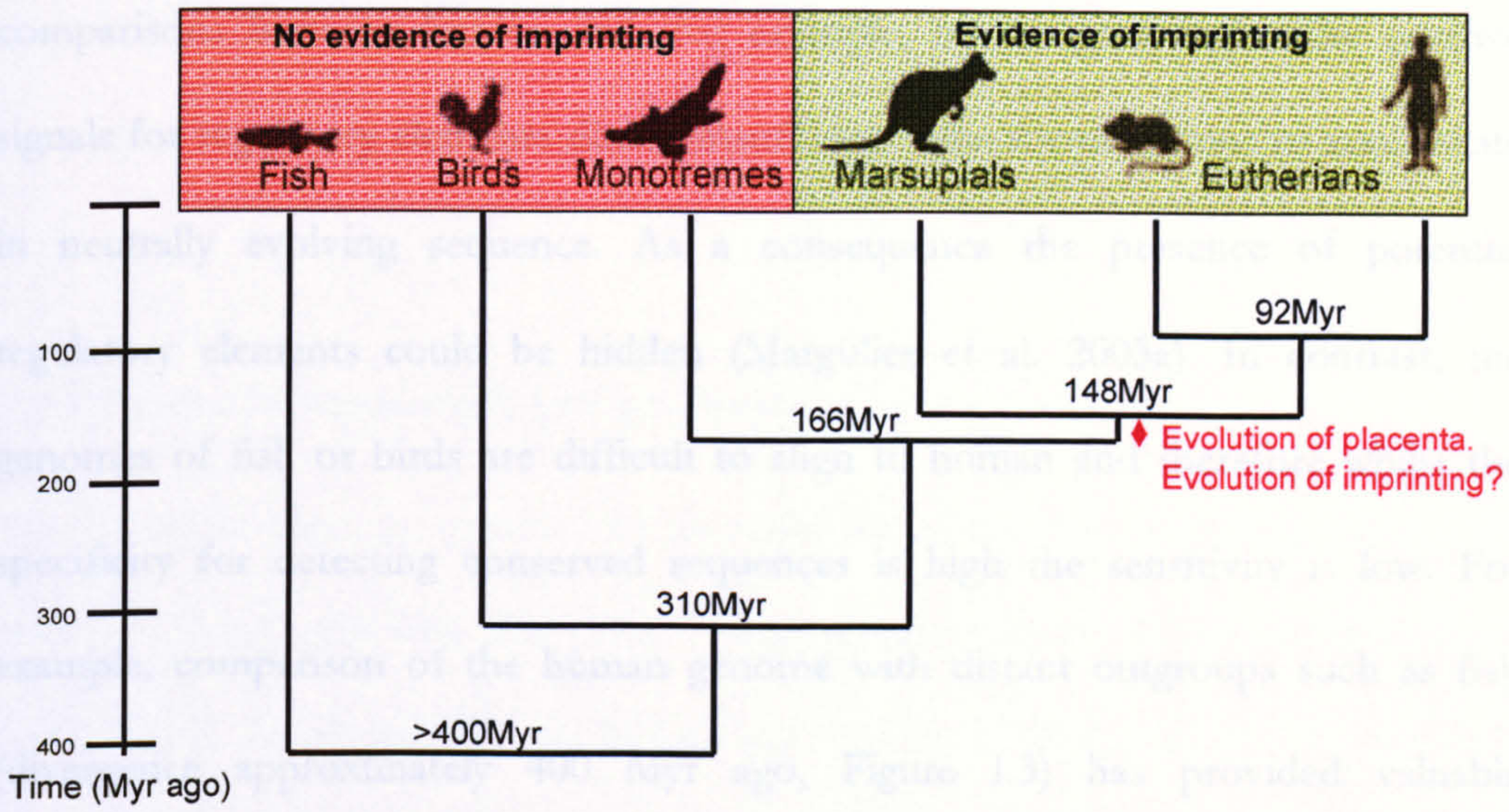


Figure I.3. Phylogeny of vertebrate species.

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).

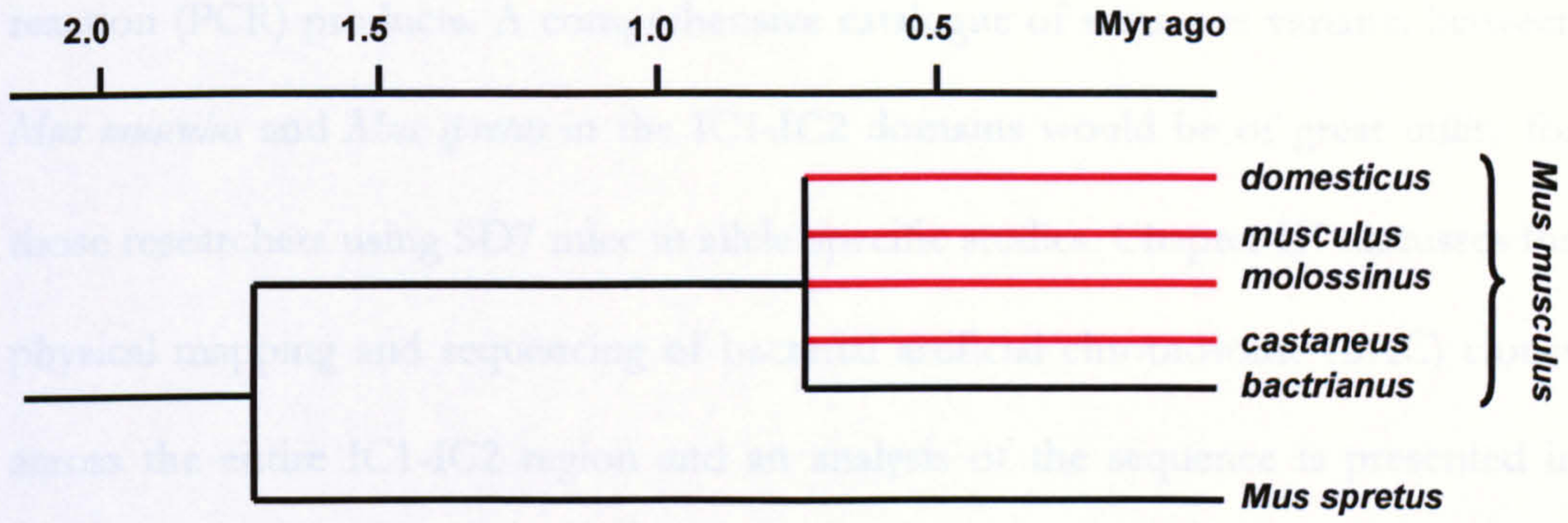


Figure I.4. Partial evolutionary tree of the genus *Mus*.

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.

PAGE/PAGES EXCLUDED UNDER INSTRUCTION FROM UNIVERSITY

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.

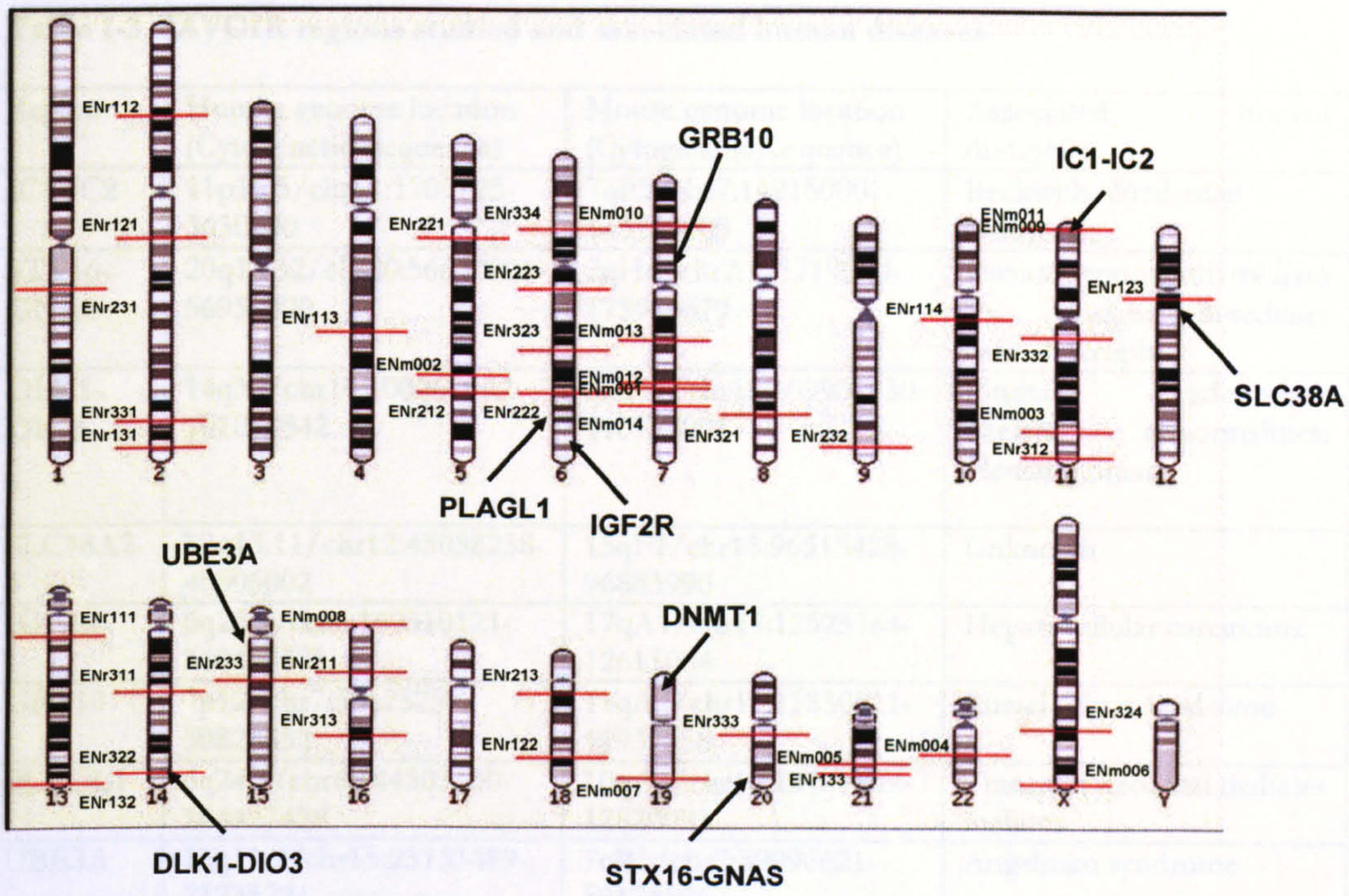


Figure I.9. SAVOIR regions studied.

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).

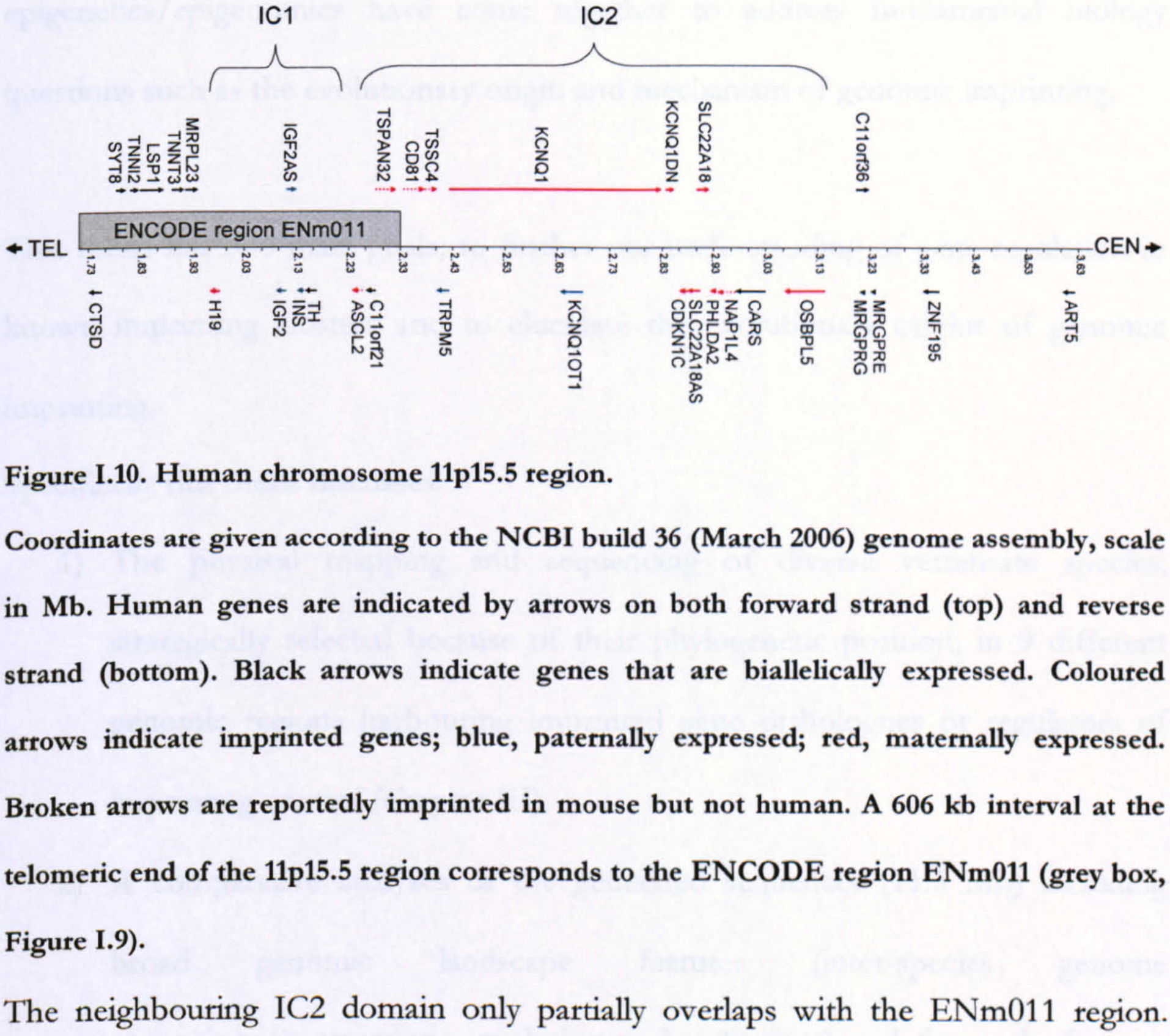


Figure I.10. Human chromosome 11p15.5 region.

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).

Chapter II - Materials and Methods

All solutions and media used are listed at the end of this chapter.

2.1 DNA manipulation methods

2.1.1 Polymerase Chain Reaction (PCR)

PCR was performed in a 96-well microtitre plate (ABgene) in a PTC-225 (MJ Research) thermal cycler or in 0.5 ml microcentrifuge tubes in a DNA Thermal Cycler (Perkin Elmer). 15 µl reactions were prepared except where noted.

1. A premix sufficient for the number of planned reactions was prepared, allowing for a 1X reaction mix once the DNA template was added (usually 10 µl of mix and 5 µl of template).
2. The final reaction contained 1X buffer (Buffer 1 unless otherwise specified), 200 µM of each of the four nucleotides (Pharmacia), 40 ng of each primer, and 0.5 units/µl of DNA polymerase (*Taq* (Applied Biosystems Amplitaq) or *KOD* Hot-start (Novagen)).
3. PCR amplifications using *Taq* polymerase were performed under the same cycling profile (except where specified): 94°C for 5 minutes, followed by 35 cycles at 94°C for 30 seconds, annealing temperature (specific to each primer pair) for 30 seconds and 72°C for 30 seconds, and finally followed by 1 extension cycle at 72°C for 5 minutes.
4. PCR amplifications using KOD polymerase were performed with the cycling profile: 94°C for 2 minutes (to activate the Hot-start KOD polymerase), followed by 35 cycles at 94°C for 15 seconds, 60°C for 30

seconds and 68°C for 3 minutes, and finally followed by 1 extension cycle at 68°C for 5 minutes.

5. Reaction products were visualised by agarose gel electrophoresis and staining with ethidium bromide (see below).

2.1.2 DNA templates

The templates used were:

1. Bacterial colonies picked into 100 µl of sterile water and 5 µl used directly.
2. 2 µl of overnight bacterial culture inoculated into 100 µl of sterile water and 5 µl used directly.
3. cDNA or bacterial pools.
4. DNA excised from an agarose gel into 100 µl sterile water, left overnight at 4°C and 5 µl used directly.
5. Human (Sigma D-3035), mouse (Coriell and Babraham Institute), wallaby (kind gifts of Marilyn Renfree and Jenny Graves), opossum (kind gift of Guillaume Smits), platypus (kind gift of Marilyn Renfree) or chicken (Novagen 69233) genomic DNA at 10 ng/µl.

2.1.3 Agarose gel electrophoresis

1. An agarose gel was prepared (2.5% for most PCR amplified products and 1% for fragments over 1 kb) in 1X TBE and ethidium bromide (250 ng/µl).
2. DNA was added to the appropriate amount of 6X loading buffer (e.g. 5 µl of PCR product and 1 µl of 6X loading buffer) and loaded. In the case of Buffer 2, the samples were loaded directly.
3. Size markers (100 bp or 1 kb ladders, Invitrogen) were also loaded.

4. Mini-gels (50 ml) were run at 80 Volts for 10-15 minutes and larger gels (250 ml) were run at 190 Volts for the time required to obtain satisfactory separation, typically 45 minutes.
5. DNA was visualised under UV light on a transilluminator and photographed with a Polaroid camera (Kodak) or digital system (UVP).

2.1.4 Size markers

100 bp and 1 kb DNA ladders for agarose gel electrophoresis were supplied from Invitrogen (15628-019 and 15615-024 respectively).

A wide-range analytical marker DNA was used in fingerprinting (Promega DG1931). This marker provides an evenly spaced distribution of 32 DNA fragments ranging from 702 bp to 29,950 bp and 4 smaller fragments (498, 525, 536 and 645 bp) resulting from a mixture of restriction enzyme digests of Lambda and θX174 DNAs.

2.1.5 Restriction enzyme digests

2.1.5.1 Liquid DNA

1. Up to 10 μg of bacterial clone, or plasmid, DNA was used in a reaction containing the appropriate 1X buffer, 1 mM spermidine, 100 μg/ml BSA and 20-50 units of the appropriate enzyme.
2. The DNA was digested for 2 hours or overnight at the temperature recommended by the supplier of the enzyme (typically NEB).
3. The DNA was subjected to agarose gel electrophoresis and visualised.

2.1.5.2 PCR products

1. After PCR amplification, the required amount (usually 5-10 µl) was transferred to a new 0.5 ml microcentrifuge tube, and 5 units of the restriction enzyme added.
2. DNA was digested for 1 hour at the recommended temperature (obtained from the NEB catalogue) and visualised by gel electrophoresis.

2.2 DNA extraction

2.2.1 Phenol/chloroform extraction of plasmids

1. To 100-700 µl of sample was added an equal volume of Tris-buffered phenol. Sample volumes smaller than 100 µl were made up to 100 µl with water to ease subsequent steps.
2. The samples were mixed thoroughly by vortex then the two phases separated by centrifugation for 1 minute at 13,200 rpm.
3. The aqueous (upper) layer was carefully removed to a fresh 1.5 ml eppendorf tube by pipetting.
4. An equal volume of phenol/chloroform/isoamyl alcohol (25:24:1) was added, mixed and spun.
5. To the carefully removed aqueous layer was added an equal volume of chloroform/isoamyl alcohol (24:1). Following mixing and centrifugation the aqueous layer containing DNA was removed for ethanol precipitation (see below).

2.2.2 Bacterial clone DNA micro-preparations

1. A single colony was inoculated into 500 µl of 2xTY broth containing the appropriate antibiotic and grown overnight at 37°C with shaking at 300 rpm.
2. 250 µl aliquots of culture were dispensed into 96-well round-bottom plates (Greiner).
3. The cells were pelleted at 2500 rpm for 4 minutes. The plate was inverted to drain the supernatant.
4. The pellet was resuspended in 25 µl of solution I (GTE).
5. 25 µl of fresh 0.2M NaOH and 1% SDS was added and mixed by gently tapping.
6. 25 µl of 5M Acetate, 3M K^+ was added and mixed by tapping gently.
7. The well contents were transferred to a 96-well 0.2 µm filter-bottom plate (Corning #3504).
8. The filter plate was taped on top of a 96-well round-bottom plate (Greiner) containing 100 µl of isopropanol.
9. The plates were centrifuged at 2500 rpm for 2 minutes and the filter plate, containing precipitated protein and bacterial genomic DNA, was discarded.
10. The round-bottom plate was incubated at room temperature for 30 minutes and then spun at 3200 rpm for 20 minutes at room temperature.
11. Supernatant was removed by inverting the plate and dabbing onto tissue.
12. 100 µl of 70% ethanol was added to each well and the plate tapped gently. The plate was then spun at 3200 rpm, for 20 minutes at room temperature.
13. The supernatant was removed by inversion and the pellet dried at room temperature.
14. The DNA was resuspended in 5 µl of $T_{0.1}E$ with RNase (10 µl of 1 mg/ml RNase per 1 ml of $T_{0.1}E$).

2.2.3 Bacterial clone DNA mini-preparations

DNA mini-preparations from 3-5 ml bacterial cultures were performed using the QIAprep spin mini-prep kit (Qiagen) as per manufacturer's instructions. Quantification of DNA was performed using a NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies).

2.2.4 Bacterial clone DNA midi-preparations

DNA midi-preparations from 50 ml bacterial cultures were performed using the Qiagen plasmid midi prep kit as per manufacturer's instructions. Quantification of DNA was performed using a NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies).

2.3 DNA purification

2.3.1 Ethanol precipitation

1. In a 1.5 ml microcentrifuge tube, 0.1 volumes of 3M sodium acetate (pH 5.2) and either 1 volume of isopropanol or two and a half volumes of ethanol were added to the DNA.
2. The samples were mixed well by vortexing and incubated for 20 minutes at –20°C.
3. The DNA was then pelleted in a benchtop microcentrifuge at 13,200 rpm and washed once with 70% ethanol.
4. The pellet was left to air dry and then re-suspended in an appropriate amount of $T_{0.1}E$.

5. The DNA recovery was tested by agarose gel electrophoresis and/or NanoDrop spectrophotometry.

2.3.2 Gel purification

1. The DNA fragment was excised from the agarose gel with a clean scalpel.
2. The gel slice was weighed in a 1.0 ml (1 g) eppendorf tube.
3. The gel slice was then purified using a Qiaquick Gel Extraction Kit™ (Qiagen) according to the manufacturer's instructions.
4. The DNA recovery was tested by agarose gel electrophoresis and/or NanoDrop spectrophotometry.

2.3.3 Exonuclease/Shrimp Alkaline Phosphatase (ExoSAP) purification of PCR products

1. A premix sufficient for the number of planned reactions was prepared allowing for a 1X reaction mix once the PCR reaction was added (usually a 15 µl PCR reaction volume).
2. The final reaction contained 1X reaction buffer (SAP Buffer), 1X Dilution buffer, 1 unit/µl of Shrimp Alkaline Phosphatase (USB) and 1 unit/µl exonuclease I (USB).
3. The mixture was incubated at 37°C for 30 minutes, followed by denaturation of the enzymes at 80°C for 15 minutes.

2.4 Clone resources

2.4.1 Bacterial artificial chromosome (BAC) libraries

Table II-1. Whole genome BAC library details.

Species	*Ornithorhynchus anatinus*	*Macropus eugenii*	*Monodelphis domestica*	*Gallus gallus*	*Mus spretus*
Common name	Duck-billed platypus	Tammar wallaby	Grey short-tailed opossum	White leghorn chicken	Algerian mouse
Library source	Clemson University Genomics Institute (CUGI)	Arizona Genomics Institute (AGI)	Benaroya Research Institute at Virginia Mason	Wageningen	Children's Hospital Oakland Research Institute (CHORI)
Constructed by	Jeff Tomkins	M. Luo	Andrew Stuart	Richard Crooijmans	Baoli Zhu
Supplied by	CUGI	AGI	BACPAC Resources	Geneservices Ltd.	BACPAC Resources
Number of clones in library	230,400	239,616	364,800	49,920	165,888
Number of gridded filters	12	13	21	4	11
Average insert size (kb)	143	166	175	134	181
Genome coverage	10.8x	11.36x	15x	5.5x	11.1x
Cloning site	*Hin*dIII	*Hin*dIII	*Eco*RI	*Hin*dIII	*Eco*RI
Cloning vector	pCUGIBAC1	pCUGIBAC1	pCC1BAC	pECBAC1	pTARBAC2.1
Library code	OA_Bb	ME_KBa	VMRC-18	WAG	CHORI-35

2.5 Cloning

2.5.1 pGEM T-Easy cloning

1. 3 µl of purified PCR products were A-tailed using 1mM dATP and *Taq* polymerase (Applied Biosystems Amplitaq) in a 10 µl reaction buffered with NEB.
2. A-tailing reactions were incubated on a thermal cycler block at 70°C for 30 minutes.

3. 10 µl ligation reactions were performed containing 5 units of T4 DNA ligase (Roche), 5 ng pGEM T-Easy vector, 1X ligation buffer (Roche) and 2 µl A-tailed PCR product or appropriate controls. Reactions were mixed and incubated overnight at 4°C.

2.5.2 Gateway® cloning

Gateway® cloning (by recombination) was performed as described by the manufacturer (Invitrogen) except for the following modifications.

2.5.2.1 Creating Gateway® Entry (pENTR) clones

The donor vector (pDONR^{TM}223, kind gift from James Hartley) was used to generate all pENTR^{TM} clones. Primers used to amplify sequence for cloning were synthesised with the following 5' adaptor (*att*B) sequences to increase the specificity of cloning orientation:

Oligo_1 5'-AAAGTTGGCATG<specific primer sequence>-3'

Oligo_2 5'-GAAAGTTGGGTA< specific primer sequence>-3'

To introduce full-length *att*B recombination sites into the PCR product a second round of PCR was performed using the stpDONR223.att.E primers:

Oligo_1 5'-GGGGACAACTTTGTACAAAAAAGTTGGCATG-3'

Oligo_2 5'-GGGGACAACTTTGTACAAGAAAGTTGGGTA-3'

1. The pDONR^{TM}223 plasmid in host *E. coli* DB3.1 cells was cultured in 5 ml LB broth containing Specintomycin (50 µg/ml).
2. Following overnight growth at 37°C with shaking at 280 rpm plasmid DNA was extracted (see section 2.2.3).
3. The sequences to be cloned were amplified in a 25 µl PCR reaction containing 1x KOD buffer (Novagen), 200 µM dNTPs, 1mM $MgSO_4$, 0.25

units KOD Hot-Start DNA polymerase, 200 nM sequence specific primers with *att*B adaptors (see above) and 50 ng DNA.

4. The amplifications were performed with the cycling profile: 94°C for 2 minutes (to activate the Hot-start KOD polymerase), followed by 35 cycles at 94°C for 15 seconds, 60°C for 30 seconds and 68°C for 3 minutes, and finally followed by 1 extension cycle at 68°C for 5 minutes.
5. One microlitre of the above reaction product was diluted in 100 µl DDW and 5 µl used in a second-round KOD PCR reaction, set-up as above except the use of stpDONR223.att.E primers (see above).
6. Amplification of this second-round PCR utilised the cycling profile: 94°C for 2 minutes, followed by 5 cycles at 94°C for 15 seconds, 45°C for 30 seconds and 68°C for 3 minutes. This was followed by 20 cycles at 94°C for 15 seconds, 55°C for 30 seconds and 68°C for 3 minutes and finally followed by 1 extension cycle at 68°C for 5 minutes.
7. PCR products were loaded into wide-wells of a 1% agarose gel. Following electrophoresis (section 2.1.3) bands were excised and gel purified (QIAquick, Qiagen).
8. 10 µl BP reactions contained 1X BP3 buffer (see solutions), 1 µl BP ClonaseII enzyme, 50 ng *att*B PCR product and 150 ng pDONR223 vector DNA.
9. BP reactions were incubated overnight at 25°C in a thermal cycler (MJ Research) and terminated by the addition of Proteinase K (2 µg) and incubation at 37°C for 10 minutes.
10. Microtitre plate transformation of *E. coli* Mach1™ cells (Invitrogen) were performed as described in section 2.7.1.
11. pENTR clones were verified by sequencing (see section 2.13).

2.5.2.2 Creating Gateway expression clones

For optimal LR recombination between pENTR clones and destination vectors the pENTR clones were linearised with an appropriate restriction endonuclease (see section 2.1.5.1).

1. LR reactions containing 75 ng destination plasmid DNA, 150 ng linear pENTR plasmid DNA and 1 µl LR ClonaseII enzyme were made up to 10 µl with TE (pH 8.0). Positive (pENTR™-gus) and negative (DDW) controls were also performed.
2. LR reactions were performed in the wells of an ABgene 96-well plate and incubated overnight at 25°C in a thermal cycler (MJ Research). Reactions were terminated by the addition of Proteinase K (2 µg) and incubation at 37°C for 10 minutes.
3. Microtitre plate transformation of *E. coli* Mach1™ cells (Invitrogen) were performed as described in section 2.7.1.

2.6 Making chemically competent cells

1. A single colony of the *E. coli*, T1 phage-resistant, Mach1™ strain (Invitrogen) was picked into 5 ml LB broth for overnight growth at 37°C with shaking at 280rpm.
2. 1 ml of overnight culture was used to inoculate 110 ml LB broth in a 500ml conical flask and incubated at 37°C with shaking at 280 rpm.
3. After 1hr 45min a 1 ml aliquot of the culture was taken and an OD_{550nm} measured. The desired OD_{550nm} of 0.48 was generally achieved in less than 2 hours with Mach1™ cells (doubling time of approximately 50min).
4. 25 ml of culture was added to pre-chilled 50 ml Falcon tubes and further chilled on ice for 15 minutes.

5. Cells were pelleted by centrifugation at 2500rpm in a Beckman J6-MC centrifuge cooled to 4°C.
6. Supernatant was discarded and excess media dabbed onto absorbent tissue paper.
7. 4 ml of cold (4°C) TfbI was added to the cell pellets on ice and the pellets re-suspended by vigorous tapping of the tubes.
8. Following a 15 minute incubation on ice the cells were re-pelleted at 2500rpm, 4°C for 10 minutes.
9. Supernatant was discarded and the pellets re-suspended in 1 ml cold TfbII solution added using pre-chilled 1 ml pipette tips.
10. Cells were again incubated on ice for 15 minutes before taking 250 µl and 100 µl aliquots into pre-frozen 1.5 ml eppendorf tubes.
11. Tubes were rapidly frozen in an ethanol/dry ice bath for storage at -70°C.

2.7 Transformation

1. 0.5 µl of ligation reactions were added to the wells of an ABgene microtitre plate, frozen at -20°C and transferred to a cold block (Stratagene).
2. A 250 µl or 100 µl aliquot of competent Mach1™ cells was transferred from a -70°C freezer to a 1.5ml cold block (Stratagene).
3. After 2 minutes the cells were flick mixed and 10 µl added to each of the ligation reactions and controls.
4. Cells were kept cold for 20 minutes before heat shocking at 42°C for 50 seconds in a pre-heated thermal cycling block (MJ Research).
5. Heat shocked cells were replaced on the cold block for 2 minutes before adding 90 µl of LB broth.

6. Cells were incubated at 37°C for 1.5 hours without shaking.
7. The entire transformation was plated onto blue/white selective agar plates and incubated overnight at 37°C. Colonies generally containing inserts (white) and those generally without (blue) were scored.

2.7.1 Microtitre plate transformation

Steps 1 to 6 above were performed.

1. Large square 48-well Bioassay plates (Genetix) were poured with 250 ml LB agar containing appropriate antibiotic selection.
2. 90 µl and 10 µl aliquots of transformed Mach1™ cells were plated and spread onto the plates using 2-4 roll-and-grow plating beads (Q-biogene) per well.
3. Plates were left to dry then inverted and incubated at 37°C overnight.

2.8 Tissue Culture

2.8.1 Resuscitating frozen human Caucasian hepatocyte carcinoma (HepG2) cells

1. A Nunc vial containing a frozen aliquot of HepG2 cells (ECACC Number 85011430; CB number 04A014) was taken from liquid nitrogen and defrosted at 37°C.
2. A few drops of pre-warmed (37°C) EMEM plus (see solutions and media) was added to the Nunc vial and the re-suspended cells transferred to a 15 ml Falcon tube. An additional 12 ml EMEM plus was added to this tube.
3. Centrifugation was performed at 1200 rpm for 5 minutes to pellet the cells.

4. The supernatant (containing freeze media) was poured off into 1% Virkon solution.
5. Cells were resuspended by pipetting in 5 ml EMEM plus media and the cells transferred to a 25 cm^2 flask for overnight growth at 37°C and 5% CO2.

2.8.2 Splitting adherent HepG2 cells

When a monolayer of cells was observed covering more than 80% of a 75cm^2 flask (approximately 1x10^7 cells) cells were split according to the following procedure. Note that HepG2 cells grow relatively slowly in islands. Cells in a 75cm^2 flask typically require splitting every 5 or 6 days.

1. EMEM plus media was aspirated off into 1% Virkon and the cells washed twice with 10 ml PBS.
2. 3 ml of pre-warmed trypsin, 0.5% EDTA solution was added to the cell monolayer, poured off and repeated.
3. The flask was incubated at 37°C for two minutes until the cells could be seen flowing down the flask bottom on tapping.
4. 10 ml of pre-warmed EMEM plus media was added to the cells with repeated pipetting to resuspend the cells.
5. To 13 ml of EMEM plus in one flask and 14 ml in a second was added 2 ml and 1 ml of cell suspension respectively. These flasks were labelled with cell type (HepG2), passage number, cell dilution (e.g. 1 in 5 and 1 in 10) and date.
6. The flasks were incubated in a humidified 37°C incubator with 5% CO_2.

2.8.3 Freezing cells for storage

1. Steps 1 to 4 of the splitting procedure above were performed and the resuspended cells transferred to a 50 ml Falcon tube.

2. Cells were pelleted by centrifugation at 1200 rpm for 5 minutes and the media poured off into 1% Virkon solution.
3. Cells were washed with 10 ml PBS, re-pelleted and the solution poured off.
4. Cells were resuspended in 1 ml freeze media (FBS and 10% DMSO).
5. 550 µl of cells were added to labelled cryo vials (Nunc), placed in a cryo 1°C freezing container (Nalgene, Cat. No. 5100-0001) and frozen at -70°C.
6. Once frozen the cells were transferred into liquid nitrogen for long-term storage.

2.9 Transient transfection of HepG2 cells

Steps 1 to 4 of the splitting procedure (2.8.2) were performed.

Cells were counted in the 25 x 25 gridded area (0.0025mm^2) of a haemacytometer to determine cell concentration. For example:

Cell count	Cell concentration (cells/ml)	Required dilution
66	$66x10^4$	1 in 10 (2 ml in 20 ml)
100	$100x10^4$	1 in 15 (1.32 ml in 20 ml)
130	$130x10^4$	1 in 20 (1 ml in 20 ml)

1. Per 96-well plate to transfect a pre-mix sufficient for 140 wells was prepared containing 560 µl EMEM (without additives) and 31.5 µl GeneJuice transfection reagent (Novagen 70967). This pre-mix was incubated for 5 minutes at room temperature.
2. 21.1 µl of the pre-mix above was added to each of the 24 wells containing 400 ng pGL3 test-construct and 8 ng pRL-CMV co-reporter plasmid DNA, mixed and incubated for 5 minutes.

3. An appropriate volume (see table above) of pre-warmed EMEM plus media was added to a 50 ml tube (Falcon). A volume of cells to give a final concentration of 6.6 x 10^4 cells/ml was added to the tube and mixed.
4. Cells were poured into a sterile multi-channel chamber and 150 µl (1 x 10^4 cells) pipetted into each well of a 96-well plate (Falcon #3072).
5. 6.2 µl of the DNA/GeneJuice complex (from step 1) was added to the cells and mixed by pipetting.
6. The 96-well plate was placed into a plastic box (containing water in circular Petri dishes to maintain humidity) and incubated at 37°C with 5% CO_2 for 48 hours. Cells should not be more than 95% confluent prior to dual luciferase reporter assays.

2.10 Dual luciferase reporter assays

All assays were carried out using the Dual-Luciferase Reporter Assay Kit (Promega #E1960). Approximately 6 ml reagents (LARII and Stop & Glo) were required for each 96-well plate to be assayed, 3 ml for the plate and 3ml for the luminometer (Berthold LB96V) injection lines.

1. Media from each well of a 96-well plate was aspirated into 1% Virkon solution.
2. Cells were washed with 250 µl PBS.
3. 23 µl 1 X passive lysis buffer (PLB – Promega #E1941) was added to the cells followed by incubation at room temperature for 1 hour with shaking (The Belly Dancer, Stovall Life Science, Inc).
4. 20 µl of the cell lysates were transferred to a white 96-well PE Optiplate (Perkin Elmer #6005290).

5. Firefly and renilla luciferase levels were assayed using a luminometer (Berthold LB96V) equipped with dual injectors, one for each of the two luciferase substrates. The injectors were programmed to dispense 30 µl of luciferase assay reagent II (LARII) and 30 µl Stop & Glo reagent. Each injection was followed by a 1.8 second delay and a 10 second measurement time.
6. Data was collected using the WinGlow (Berthold Technologies) software package and analysed in Microsoft® Office excel.

2.11 Library screening

2.11.1 PCR radiolabelling of STSs

DNA probes were radiolabelled by PCR (Feinberg and Vogelstein. 1983, Hodgson and Fisk. 1987).

1. The required fragment was amplified from either genomic DNA or cDNA as appropriate.
2. The fragments were separated on a 2.5% agarose gel, cut out and transferred to a 0.5 ml microcentrifuge tube containing 100 µl of sterile water. The DNA was allowed to diffuse out of the gel slice (at least one hour).
3. Using Guy's buffer, 9.5 µl reactions were set up containing the required primers, 2.5 µl of the liquid surrounding the gel slice, nucleotides (except dCTP) and *Taq* (Applied Biosystems) DNA polymerase. A single drop of mineral oil was placed on top of the reaction mixture.
4. 0.5 µl of ^{32}P-dCTP (3000 Ci/mol, Amersham Pharmacia Biotech) was added.
5. PCR was performed in a DNA Thermal Cycler (Perkin Elmer) under the following cycling profile, 94°C for 5 minutes, 25 cycles of 93°C for 30

seconds, 55°C for 30 seconds, 72°C for 30 seconds and 1 cycle of 72°C for 5 minutes.

6. After completion, the probes were denatured in the thermal cycler at 99°C for 5 minutes and then snap chilled in ice water.
7. The probes were then added to the hybridisation mix.

2.11.2 Screening of library filters by hybridisation of PCR-labelled probes

High-density gridded library filters of mouse, wallaby, opossum, platypus and chicken BAC clones were imported (Table II-1). Filters were screened as follows:

1. STSs were radiolabelled as described above.
2. Up to 13 (22x22 cm) filters were sequentially placed in a 22x22x5 cm sandwich box with sufficient hybridisation buffer to cover the filters. A plastic sheet, cut to size, was placed on top to reduce evaporation. The filters were pre-hybridised at 65°C for 2 hours with shaking at 50 rpm (Innova 4000 orbital shaking incubator, New Brunswick Scientific).
3. The filters were removed and the denatured probe was added to hybridisation solution in the box and mixed.
4. The filters were added back into the box and carefully submerged under the hybridisation mix. The plastic sheet was replaced on top.
5. After hybridisation overnight at 65°C and 50 rpm shaking, the filters were washed by rinsing twice in 2X SSC at room temperature for 5 minutes. The filters were then washed twice in 0.5X SSC and 1% N-lauroyl-sarcosine at 65°C for 30 minutes, before rinsing twice in 0.2X SSC at room temperature for 5 minutes.

6. The washed filters were wrapped in Saran wrap (Dow Chemical Co.) and exposed to pre-flashed Fuji Medical X-ray film (036010) (or equivalent) overnight. If longer exposure was required the wrapped filters were sandwiched between intensifying screens and stored at –70°C.
7. Occasionally filters were re-washed to 0.2X SSC with 1% N-lauroyl-sarcosine at 65°C for 30 minutes if required (e.g. to remove high background signals).
8. The autoradiographs were developed and labelled with the name of the filter and the data entered into implementations of ACeDB.

Positive BACs resulting from this screening were received from distribution centres (Table II-1) in agar stabs and subsequently inoculated into LB broth containing 7.5% glycerol and chloramphenicol. Four identical aliquots of each BAC were stored at -70°C in 96-well microtitre plates, representing Archive, Backup, Working and Gridding copies. Filters corresponding to region-specific subsets of bacterial clones were gridded by the Sanger Institute clone resource group. These "polygrid" filters were screened as above except that (pre-)hybridisations were performed in 15 ml tubes (Falcon) and only one probe was screened against one filter to establish the STS content of arrayed BACs. For all other BAC clone experiments the Working copy was used.

2.12 Landmark production

2.12.1 Primer design

Primers were designed using the web-based Primer3 program (Rozen and Skaletsky. 2000) (http://www.genome.wi.mit.edu/genome_software/other/primer3.html). Additional primers for amplification and subsequent cloning of full-length cDNA were designed using the perl script Expresso (Dave Beare). For cloning of PCR

products *att*B adaptor sequences (Gateway®) or restriction endonuclease recognition sites were added to the 5' end of oligonucleotide primers.

2.12.2 Primer synthesis

1. Primers were synthesised at the Sanger Institute by David Frazer and Di Gibson or externally by Sigma-Genosys (Haverhill, UK). Primer concentrations were supplied in both cases.
2. Primers were stored at –20°C and working dilutions for PCR prepared at 100ng/µl for each primer in pairs.
3. The primers were tested at three different annealing temperatures, 55°C, 60°C and 65°C, using the standard cycling on thermal cyclers to establish optimal PCR conditions.

2.12.3 Primer sequences

All primer sequences used are available in the Appendix B (CD at the back of this thesis).

2.13 Plasmid and PCR product sequencing

All plasmid end-sequencing or PCR product sequencing was performed by the Sanger Institute Faculty Small Sequencing Projects (FSSP) team. Sequence-ready reactions were supplied containing 5 µM primer, 5-100 ng purified DNA (depending on size of PCR product or plasmid) in a total volume of 7 µl.

2.14 Bacterial clone fingerprinting

*Hin*dIII fingerprinting of bacterial clones was performed using the standard protocol below (Marra et al. 1997).

2.14.1 Restriction endonuclease digestion

1. Bacterial clones were micro-prepped as described in section 2.2.2 above.
2. 2.6 µl of water, 0.9 µl of NEB buffer 2 and 20 units of *Hin*dIII (NEB) were added to each well, mixed by gentle tapping and then the plate centrifuged up to 1000 rpm to collect the contents.
3. The plate was incubated at 37°C for 2 hours.
4. The reaction was terminated by addition of 2 µl of 6X Dye Buffer II and the plate centrifuged up to 1000 rpm to collect the contents.

2.14.2 Gel preparation and loading

1. A 1% gel mix was prepared using 450 ml of 1X TAE and 4.5 g agarose and poured at 4°C. A 121-well comb was placed in the gel and allowed to set for 45 minutes. The comb was then removed.
2. 3-4 litres of 1X TAE was added to the gel tank.
3. 0.8 µl of the marker (Promega, DG1931) was loaded in the first well and then in every subsequent fifth well.
4. 2.0 µl of each sample was then loaded into the empty wells between markers.
5. The gel was run in a coldroom (4°C) at 85 V for 16 hours.

2.14.3 Gel staining

1. The gel was trimmed to ~19 cm and stained with vistra green stain for 45 minutes. The gel was covered whilst staining to prevent light degradation of the vistra green.
2. The gel was then rinsed with 0.5 litres of deionised water. The gel was visualised and the image recorded using a Molecular Dynamics scanner.

2.15 Computational analysis

2.15.1 ACeDB

For each species an implementation of ACeDB (http://www.acedb.org) was used to track mapping, sequencing and analysis data:

Chicken	'gallusace'
Platypus	'platypace'
Wallaby	'wallabase'

These databases are curated by me but the underlying code maintained by Carol Scott.

For other species and/or test data my own 'Imprintace' database was used.

2.15.2 Sequence analysis and annotation

Clones selected for sequencing are uploaded into ChromoView using the perl script tpf2oracle. The ChromoView web interface links to gull to track sequencing

progress and displays pre-computed sequence overlaps. Finished sequence assemblies, from ChromoView, are exported as an AGP file for sequence analysis. Finished BAC sequences were analysed in the standard HAVANA (Human And Vertebrate Analysis aNd Annotation) pipeline (http://www.sanger.ac.uk/HGP/havana/). The sequences were analysed for repeats using RepeatMasker (Smit, AFA., Hubley, R. and Green, P., "RepeatMasker" at http://www.repeatmasker.org) and tandem repeats finder (trf, Benson. 1999). The repeat masked sequences were subsequently screened in similarity searches against the public domain DNA and protein databases using the BLAST suite of programs. The exon and gene prediction programs, Genscan (Burge and Karlin. 1997), Fgenesh (Solovyev et al. 1994), tRNAscan (Fichant and Burks. 1991, Lowe and Eddy. 1997) and Eponine TSS (Down and Hubbard. 2002) were used to predict possible gene structures. The unmasked sequence was used in C+G content analysis and prediction of CpG islands. The completed sequences were visualised in the DNA map display of ACeDB ('otterlace', Searle et al. 2004) and this database used for the manual annotation of gene structures.

2.15.3 Multi-species comparative sequence analysis

Web based multi-species comparative sequence analyses were performed using the zPicture server (Ovcharenko et al. 2004a) (http://zpicture.dcode.org).

With assistance from Carol Scott and Paul Bevan, a SAVOIR consortium webpage showing comparative sequence-ready maps for each region and species being studied was implemented (http://www.sanger.ac.uk/PostGenomics/epicomp). Clone map and sequence data was regularly updated using MySQLMan tables (v1.09, Gossamer Threads Inc).

2.15.4 BLAST and BLAT

Web based BLAST analyses were performed at NCBI (http://www.ncbi.nlm.nih.gov/BLAST), EnsEMBL (http://www.ensembl.org/) or Sanger Institute (http://www.sanger.ac.uk/cgi-bin/blast). A local UNIX installation of BLASTN was run on the command line as follows. A BLASTN database of FASTA format sequences was created using the 'pressdb <seq.fa>' command. The command 'blastn <database> <query.fa> > <output>' was used to redirect the BLASTN output to a file.

BLAT (Kent. 2002) was performed at the UCSC Genome browser (http://genome.ucsc.edu/).

2.15.5 Electronic polymerase chain reaction (ePCR)

To establish the STS content of a sequence an in-house implementation of the program electronic PCR (ePCR, Schuler. 1997) was performed. Usage:

e-PCR STS_database Sequence [options]

Primer sequences used to create the STS database were exported from ACeDB and formatted in the text editor 'emacs' (http://www.gnu.org/software/emacs). The database is of the format:

STSname Oligo_1(5' ->3') Oligo_2(5'->3') Amplicon_size(bp)

The interrogated sequence is in FASTA format and ePCR options were:

M=50. The margin of discrepancy between the observed and expected amplicon sizes.

N=0. The number of mismatches allowed.

W=7. The word size

In addition the option –mid was used when the mid-point of the STS match was required and not the range (default).

2.15.6 Perl and EMBOSS scripts

The following perl scripts were generated by Dave Beare (unless stated otherwise) and can be found in the directory /nfs/chr22/gid22/perl/

scf2fa Converts sequence SCF files to FASTA format sequences

Expresso Used for ORF oligo design

MatchReport Batch BLAST and output processing tool (original software written by Luc Smink).

tpf2oracle Submits species specific tile path format to Oracle tracking database (written by James Gilbert)

CpGcount Calculates the frequency of CpG dinucleotides in a sequence

cpgplot Plots CpG rich regions (EMBOSS script)

newcpgreport Predicts CpG islands (EMBOSS script)

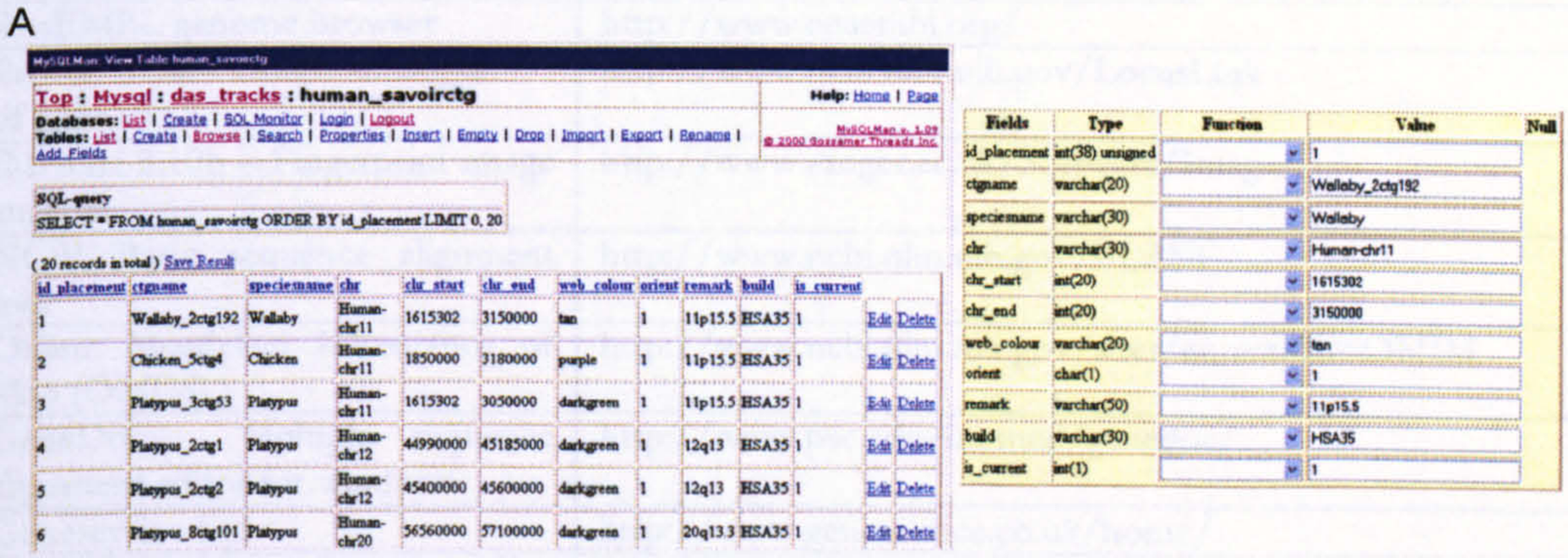
tracey Tool to retrieve data from the internal trace archive

2.15.7 MySQL tables

Updates to the SAVOIR website mapping displays (chapter IV) were made using MySQL tables that were regularly updated using a web-based database manager tool (MySQLMan v1.09, Gossamer Threads Inc). Four tables exist; savoirctg, savoirclone, savoireg and savoirer. The first two describe the mapped contigs and clones, respectively (Figure II.1). The third table (savoireg) contains details of the known EnsEMBL genes and the final table lists ENCODE pilot project regions (ENCODE Project Consortium. 2004). As new mapped clone contigs were generated the savoirctg table was updated with the contig name in the format

<species>_<chr>ctg<FPCid> (e.g. Wallaby_2ctg192 for the wallaby chromosome 2 region orthologous to human 11p15.5). Approximate coordinates of the contigs in the human genome assembly (NCBI Build35) were provided, using the gene content of the vertebrate sequences as a guide. Individual clones within the maps were entered into the savoirclone table with their international clone name, species name, sequencing status, sequence accession number (if available) and approximate position relative to human. Like the contigs, the position of clones was determined by gene content and/or sequence length. In total the MySQL database contains information on 99 BAC clones in 20 contigs across the 9 SAVOIR regions and 5 species.

A



B

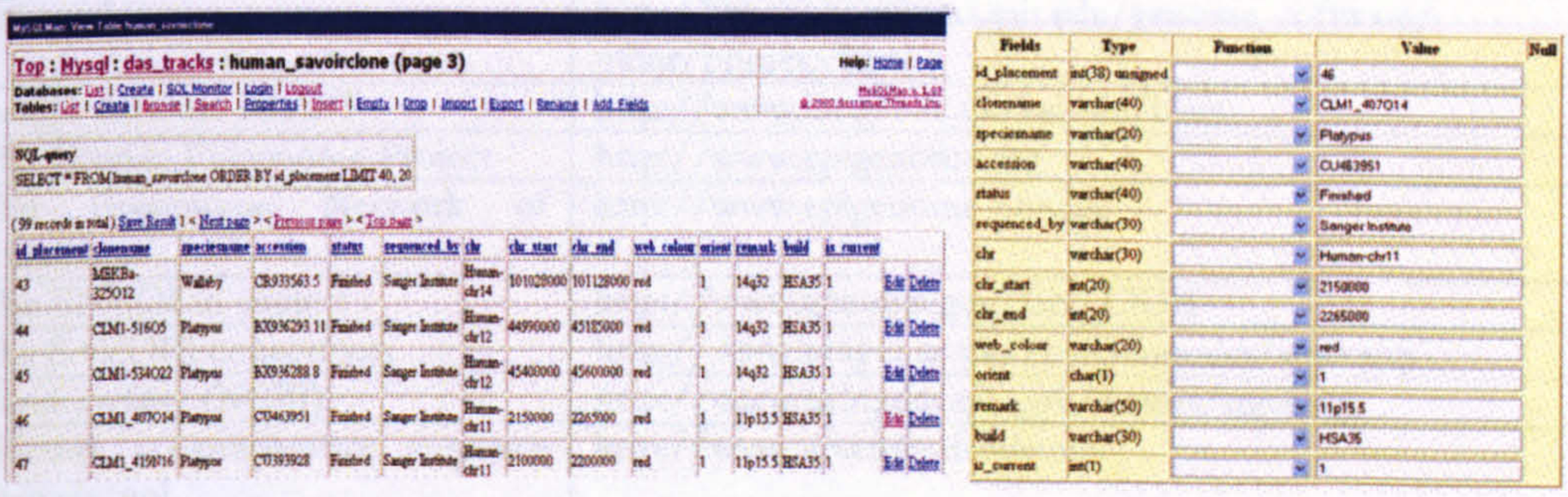


Figure II.1. SAVOIR contig and clone MySQL tables.

Tables containing data for the mapped contigs (A) and sequenced clones (B) are shown. Data is entered manually into the web forms on the right and include features such as contig or clone name, species of origin, orthologous human genomic region, feature coordinate (in the NCBI build 35 assembly of the human genome) and sequencing status of the clone.

2.16 URLs

Table II-2. URLs visited.

Arizona Genomics Institute – BAC orders	http://www.genome.arizona.edu/orders/
Harwell Mouse Imprinting	http://www.mgu.har.mrc.ac.uk
Catalogue of imprinted genes	http://cancer.otago.ac.nz/IGC/Web/home.html
Zebrafish EST BLAST site	http://134.174.23.160/zfBlast/PublicBlast.htm
Fugu genome project	http://fugu.hgmp.mrc.ac.uk
University of California, Santa Cruz (UCSC) genome browser	http://genome.ucsc.edu
Chicken EST project	http://www.chick.umist.ac.uk
CpG island identification	http://www.ebi.ac.uk/cpgplot/
Clemson University Genomics Institute – BAC orders	https://www.genome.clemson.edu/groups/bac/
ClustalW multiple sequence alignment tool	http://www.ebi.ac.uk/clustalw/index.html
EMBOSS	http://emboss.sourceforge.net/
ENCODE	http://www.genome.gov/10005107
EnsEMBL human blastview	http://www.ensembl.org/Homo_sapiens/blastview
EnsEMBL genome browser	http://www.ensembl.org/
Entrez Gene - Online catalogue of gene loci	http://www.ncbi.nlm.nih.gov/LocusLink
IMAGE 3.10b – Fingerprint image analysis	http://www.sanger.ac.uk/Software/Image/
NCBI Basic sequence alignment tool	http://www.ncbi.nlm.nih.gov/BLAST
Online Mendelian Inheritance of Man (OMIM)	http://www.ncbi.nlm.nih.gov/sites/entrez?db=OMIM
GeneDoc – Multiple sequence alignment editor for windows	http://www.psc.edu/biomed/genedoc/
Geneservice Ltd	http://www.geneservice.co.uk/home/
Pipmaker	http://bio.cse.pse.edu/pipmaker
Primer 3.0	http://www.genome.wi.mit.edu/genome_software/other/primer3.html
Sanger Institute BLAST	http://www.sanger.ac.uk/cgi-bin/blast
The Human Epigenome Project	http://www.epigenome.org
The Epigenome Network of Excellence	http://www.epigenome-noe.net
The ENCODE project	http://www.genome.gov/ENCODE
The SAVOIR consortium	http://www.sanger.ac.uk/PostGenomics/epicomp
Trace archive (NCBI)	http://www.ncbi.nlm.nih.gov/Traces/trace.cgi
zPicture - Comparative sequence analysis tool	http://www.zpicture.dcode.org

2.17 Solutions and media

Table II-3. Solutions and media used

1X TE 10 mM Tris-HCl (pH 8.0) 1 mM EDTA	1X $T_{0.1}E$ 10 mM Tris-HCl (pH 8.0) 0.1 mM EDTA
10X NEB PCR buffer 670 mM Tris-HCl (pH 8.8) 166 mM $(NH_4)_2SO_4$ (enzyme grade) 67 mM MgCl (pH 8.8)	Guys Buffer 500 mM KCl 100 mM Tris pH 8.3 15mM $MgCl_2$
28% Sucrose/cresol red solution 1X $T_{0.1}E$ 28% w/v sucrose 0.008% w/v cresol red	6X Glycerol loading dyes 30% v/v glycerol 0.1% w/v bromophenol blue 0.1% w/v xylene cyanol 5 mM EDTA (pH 7.5)
6X Dye Buffer II 0.25% bromophenol blue 0.25% xylene cyanol 15% Ficoll (Type 400: Pharmacia)	Vistra Green stain For 1 gel: 0.01M Tris HCl 0.0001M EDTA (pH 7.4) 50 µl Vistra green (Amersham)
10X TAE 890mM Tris base 0.05M EDTA 5.71% glacial acetic acid	10X TBE 890 mM Tris base 890 mM Borate 20mM EDTA (pH 8.0)
	LB broth 10 mg/ml bacto-tryptone 5 mg/ml yeast extract 10 mg/ml NaCl (pH 7.4)
Hybridisation buffer 6X SSC Denhardt's solution 1% N-lauroyl-sarcosine 50mM Tris-HCl (pH 7.4) 10% w/v dextran sulphate	2 x TY 16 mg/ml bacto-tryptone 10 mg/ml yeast extract 5 mg/ml NaCl
20X SSC 3M NaCl 0.3M Trisodium citrate	SAP buffer 200 mM Tris HCl (pH 8.0) 100 mM $MgCl_2$

ExoSAP dilution buffer 50mM Tris HCl (pH 8.0)	100X Denhardt's solution 20 mg/ml Ficoll 400-DL 20 mg/ml polyvinylpyrrolidine 40 20 mg/ml BSA (pentax fraction V)
Solution I (GTE) 50 mM glucose 10 mM EDTA 25 mM Tris-HCl (pH 8.0)	Solution II 5M NaOH (0.2M final conc.) 10% SDS (1% final conc.)
Solution III 3M KOaC (pH5.5)	EMEM plus Minimal Essential Medium Eagles including 1% non-essential amino acids (Sigma M5650) 10% Foetal bovine serum 2mM L-Glutamine 100 units/ml penicillin 100µg/ml streptomycin
Antibiotics: Ampicillin 50 mg/ml (final conc. 100 µg/ml) Gentamycin 50 mg/ml (final conc. 7 µg/ml) Kanamycin 10 mg/ml (final conc. 50 µg/ml) Spectinomycin 10 mg/ml (final conc. 100 µg/ml)	1X PBS 137mM NaCl 10mM phosphate 2.7mM KCl pH 7.4
1X BP3 buffer 20mM Tris-Cl (pH7.5) 4mM EDTA 6mM spermidine-HCl 5% glycerol 45mM NaCl	

Chapter III - Mapping and sequencing of vertebrate orthologous imprinted regions

3.1 Introduction

3.1.1 Aims of this chapter

The aim of the work covered in this chapter was to generate regional BAC maps in species with and without genomic imprinting and derive sequences from minimally overlapping tile paths of BACs. The map and sequence resources presented in this chapter are not only critical to addressing the overall aims of the thesis but were released to the scientific community according to the Fort Lauderdale agreement (http://www.wellcome.ac.uk/assets/wtd003207.pdf) to serve as a significant and lasting resource to be used by the imprinting community as well as groups studying vertebrate genome biology. Using a clone-by-clone sequencing approach not only facilitates the finishing of sequences but also creates a resource of BACs for further studies. Indeed, in collaboration with research groups in Cambridge, the mapped BAC clones have been used as fluorescence *in situ* hybridisation (FISH) probes to identify chromosomal locations of genes in both tammar wallaby and platypus to address hypotheses concerning the evolutionary origins of genomic imprinting (Edwards et al. 2007). Other potential applications for mapped BAC clones include the generation of transgenic lines for functional genetics studies and micro-array comparative genomic hybridisation (Ylstra et al. 2006). The prompt release of sequences generated during this project has allowed others to perform their own analyses. For example, the chicken chromosome 5 sequence has been analysed for

its conserved synteny with imprinted mammalian species to address how structural features of imprinting evolved in the IC1 and IC2 domains (Paulsen et al. 2005). Whilst the study of these different regions is expected to give insight into the imprinting mechanisms acting in each region (and potentially unique to that region) the main focus of this thesis is the 11p15.5 orthologous region.

3.1.2 Different methods of genome sequencing

A full understanding of gene regulation within orthologous imprinted gene regions will require complete and accurate sequence. The genomes of multi-cellular organisms which have had their sequence 'finished' to agreed levels of accuracy, according to the Bermuda principles (chapter I), include human, mouse, the nematode worm, *Caenorhabditis elegans* and flowering plant *Arabidopsis thaliana* (International Human Genome Sequencing Consortium. 2004, The Arabidopsis Genome Initiative. 2000, The C. elegans Sequencing Consortium. 1998, Waterston et al. 2002). The strategies for generating finished sequence were essentially the same for these genome projects. They all made use of a hierarchical approach in which physical maps of bacterial clones were first assembled and anchored to chromosomes using genetic, radiation hybrid or yeast artificial chromosome clone maps (Bentley et al. 2001, Dunham et al. 1999, Mungall et al. 1997). Minimally overlapping bacterial clones (e.g. cosmids or bacterial artificial chromosomes (BACs)), that collectively represent a minimal tiling path across a genomic region, were subsequently chosen for shotgun sequencing. This approach is known as a clone-by-clone approach.

The alternative to a clone-by-clone sequencing strategy is a whole genome shotgun (WGS) sequencing strategy in which the DNA of an entire genome is sub-cloned

into plasmids, sequenced and computer assembled. For small genomes such as those of viruses or bacteria this approach is effective, however, for much larger genomes this approach alone is problematic. For illustration consider a typical mammalian genome of 3 Gbp (3,000,000,000 bp) as a Constable Landscape painting and each sequence read (say 500 bp) as a pixel in that painting. With a sufficient number of sequence reads and assuming a random distribution it should be possible to represent every pixel of the painting. Assembling the pixels into clusters (contigs) would be a challenging enough prospect. Of course, in a Constable painting there is a considerable amount of sky, much of which looks the same. This is also the case in mammalian genomes in which approximately 50% of the sequence is repetitive, posing a real problem to its correct assembly. The clone-by-clone mapping approach alleviates many of the problems associated with genomic repeats since it greatly reduces (approximately 20,000 fold) the complexity of an assembly to a region of about 150 kb, the average length of a BAC clone.

The WGS approach, by definition, cannot be targeted to regions of interest and therefore is not cost effective when wishing to address localised questions. WGS approaches are also not conducive for generating high-quality 'finished' sequence because the plasmids sequenced are too numerous for storage. The majority of vertebrate genome sequences now held in the public databases have been sequenced to draft quality only and assembled using a hybrid strategy in which sequence read pairs from smaller plasmids are assembled within a scaffold of larger insert clones (e.g. BACs). Whilst there can be no doubt of the utility of having multiple draft genome sequences in the public databases it is important to acknowledge their limitations.

3.1.2.1 Limitations of draft sequence assemblies

Possibly the greatest problem faced with WGS sequence assemblies is one of coverage. The randomness of the WGS approach means that for biological, technical or simply statistical reasons some regions of a genome will be very well represented whereas others will be under-represented or not represented at all. The same can be said for the shotgun of an individual clone. However, at this scale it is more straightforward to supplement the shotgun sequence reads with directed approaches used in the finishing phase. An issue linked with coverage is that of accuracy. Regions of low-coverage are inherently more prone to low accuracy because there is no confirmation of the base sequence provided by increased depth of coverage.

Why is the method of sequencing important? To identify all genes and their regulatory elements in a given region it is crucial to have good coverage across that region. As discussed in chapter I, comparative sequence analysis has been widely used to identify functional elements. The initial and most crucial step in sequence comparison is to align those sequences and the results of any such methods will only be as good as the underlying alignment. We can be much more confident about the presence or absence of a gene or regulatory element when we know that there are no gaps in the sequences being compared. For the reasons outlined above I therefore decided to use the clone-by-clone mapping and sequencing strategy in this thesis. The strategy is similar to that used in the human genome project (HGP, (Bentley et al. 2001, International Human Genome Sequencing Consortium. 2001) but with some significant differences in the early mapping stages.

3.1.2.2 Modification of the HGP strategy

With the complete sequence of the human genome (and others) came the opportunity for comparative mapping and sequencing of any other vertebrate. This is made possible by local conserved synteny as observed by the co-location of genes along chromosome arms of different species. Human genomic sequences from the orthologous region(s) of interest can be used to search (using, for example, BLASTN) a plethora of sequence databases containing known genes, expressed sequence tags (ESTs), BAC-end sequences or WGS sequence reads from other species. Since identified sequences must share homology with the query human sequence, the regions of homology can be used to design DNA probes.

There is a selection of methods available for DNA probe design. Perhaps the most widely used is the 'overgo' approach (Ross et al. 1999, Vollrath. 1999). Overgos are pairs of oligonucleotide primers (typically 24-mers) that overlap one another by 8 bp designed within regions of high similarity between the orthologous sequences of two species. These partially complementary primers are then used to generate 40 bp double-stranded radiolabelled DNA probes in a fill-in reaction using ^{32}P-dATP and ^{32}P-dCTP (McPherson et al. 2001). An alternative to these short overgo probes are the longer STS probes. STS primers are typically shorter (20-mers) and therefore less costly than the overgo primers but the resulting PCR radiolabelled amplicons are longer and thus provide greater hybridisation specificity. Furthermore, the same STS primers can be used in regular PCR to rapidly test BACs for landmark content data. The proven success of STS probes in previous large-scale mapping projects led me to use this methodology here.

3.1.3 Species and regions studied

The informative species used in this thesis were selected because of their unique phylogenetic positions with respect to hypotheses of the evolution of genomic imprinting and have been introduced in chapter I. Comparative mapping and sequencing in wallaby (with evidence of imprinting) and platypus (without evidence of imprinting) for 8 genomic regions orthologous to known imprinting gene clusters of human and mouse has been performed. Additionally, the IC1-IC2 orthologous regions in chicken, *Mus spretus* and South American grey, short-tailed opossum (IC1 region only) were mapped and sequenced. The selection of the IC1-IC2 and other regions for mapping and sequencing reflects those most intensively studied, many of which are of interest to imprinting groups in the Cambridge region. 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, chapter IV). Finally, the wallaby and platypus orthologues of the DNA methyltransferase 1 (*DNMT1*) gene, responsible for *de novo* and maintenance DNA methylation and therefore of fundamental importance to the imprinting mechanism were mapped and sequenced.

3.2 Bacterial clone contig construction

The process of clone-by-clone sequencing can be conceptually divided into the sequential steps (Figure III.1); map construction, clone selection for sequencing, sub-clone library construction, shotgun sequencing, sequence assembly, directed finishing and sequence verification. For this strategy to be successful three criteria need to be met:

1) Bacterial clone libraries for the species of choice should be available. The sources of libraries used in this thesis are listed in chapter II, Table II-1.

2) The second requirement is the availability of genomic resources (e.g. genomic DNA and orthologous sequences) to generate unique DNA landmarks in a given species. Frequently these orthologous sequences are derived from gene sequences (e.g. ESTs or mRNAs) deposited in public databases, which can be used to generate inter- or intra-species-specific markers for library screening (see below).

3) Finally the infrastructure for large-scale genomic sequencing needs to be in place. At the Wellcome Trust Sanger Institute a streamlined sequencing pipeline has been implemented with each of the sequential steps above, from sub-clone library construction to finished sequence quality assessment, being performed by specialist teams and technology. Supporting this infrastructure are a series of databases, each interacting with one another, to track the clones through the sequencing pipeline.

The following sections describe all stages of the mapping process (performed by me) from marker generation through to sequence clone selection.

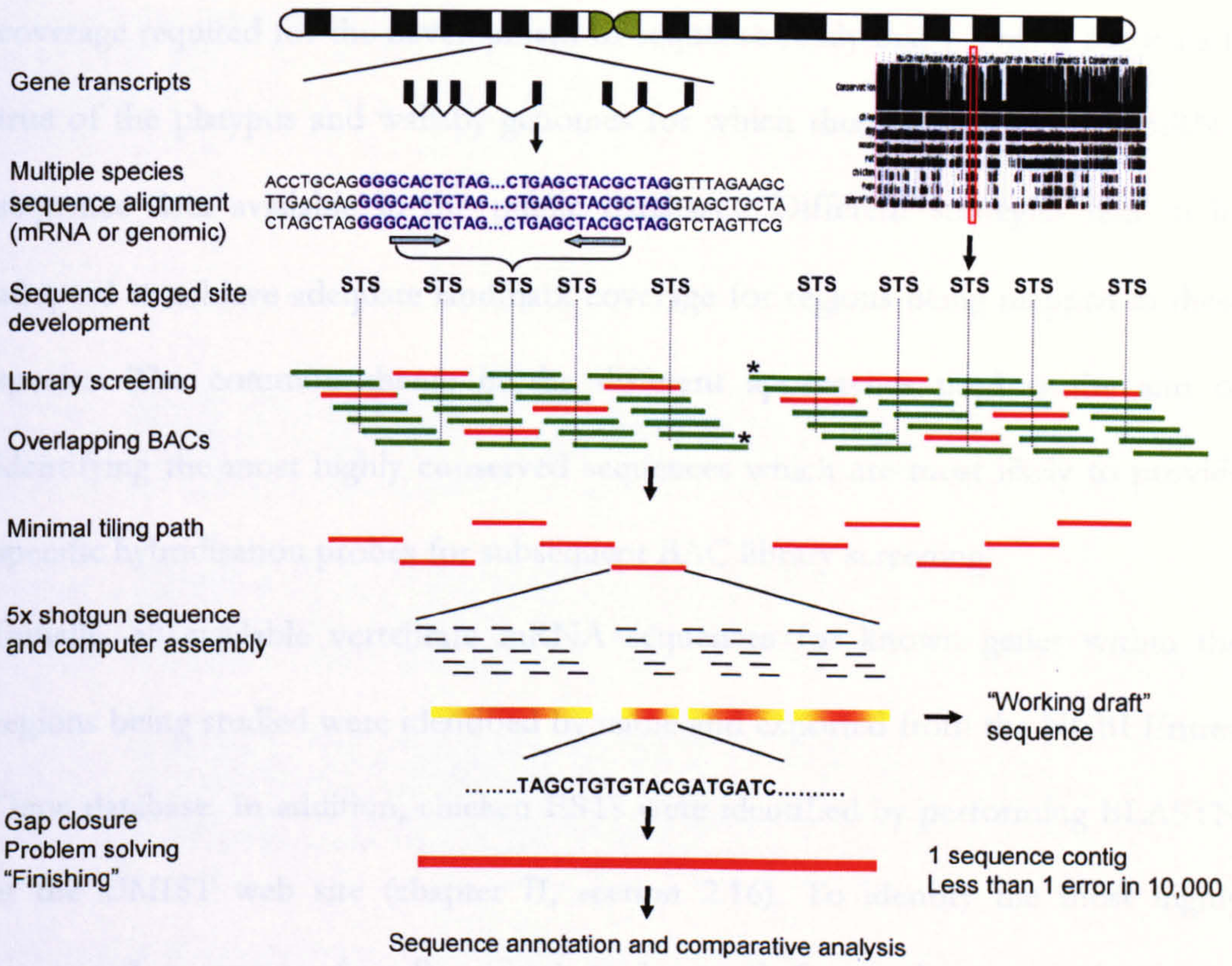


Figure III.1. Mapping and sequencing strategy.

Highly conserved sequences between distantly related species were used to generate hybridisation probes for species-specific BAC library screening. Identified BACs were imported and assembled into contigs from which minimally overlapping BACs were selected for shotgun sequencing. Computational assembly of sub-clone sequences from each BAC was performed to generate a "working draft". A high quality "finished" sequence is generated following rounds of manual editing and problem solving by skilled biologists.

3.2.1 Marker development

The mapping and sequencing strategy employed in this project is depicted in Figure III.1 and is broadly the same as that used in the HGP (Bentley et al. 2001, Mungall and Humphray. 2003). There are, however, some significant strategic differences, especially in the initial marker development and these are described as follows. Although BAC libraries are becoming available for a growing number of species (chapter II, Table II-1), few of these genomes have adequate marker (landmark)

coverage required for the development of sequence-ready maps. This is particularly true of the platypus and wallaby genomes for which there is little EST or mRNA sequence data available in the public databases. Different strategies had to be adopted to achieve adequate landmark coverage for regions being mapped in these species. The common theme in the different approaches used is the aim of identifying the most highly conserved sequences which are most likely to provide specific hybridisation probes for subsequent BAC library screening.

Initially, all available vertebrate mRNA sequences for known genes within the regions being studied were identified by name and exported from the NCBI Entrez Gene database. In addition, chicken ESTs were identified by performing BLASTN at the UMIST web site (chapter II, section 2.16). To identify the most highly conserved sequences (usually exons) to be used for marker generation, gene sequences were aligned using ClustalW (EBI) and the alignments read into the GeneDoc program for manual inspection and annotation (http://www.nrbsc.org/downloads, Figure III.2).

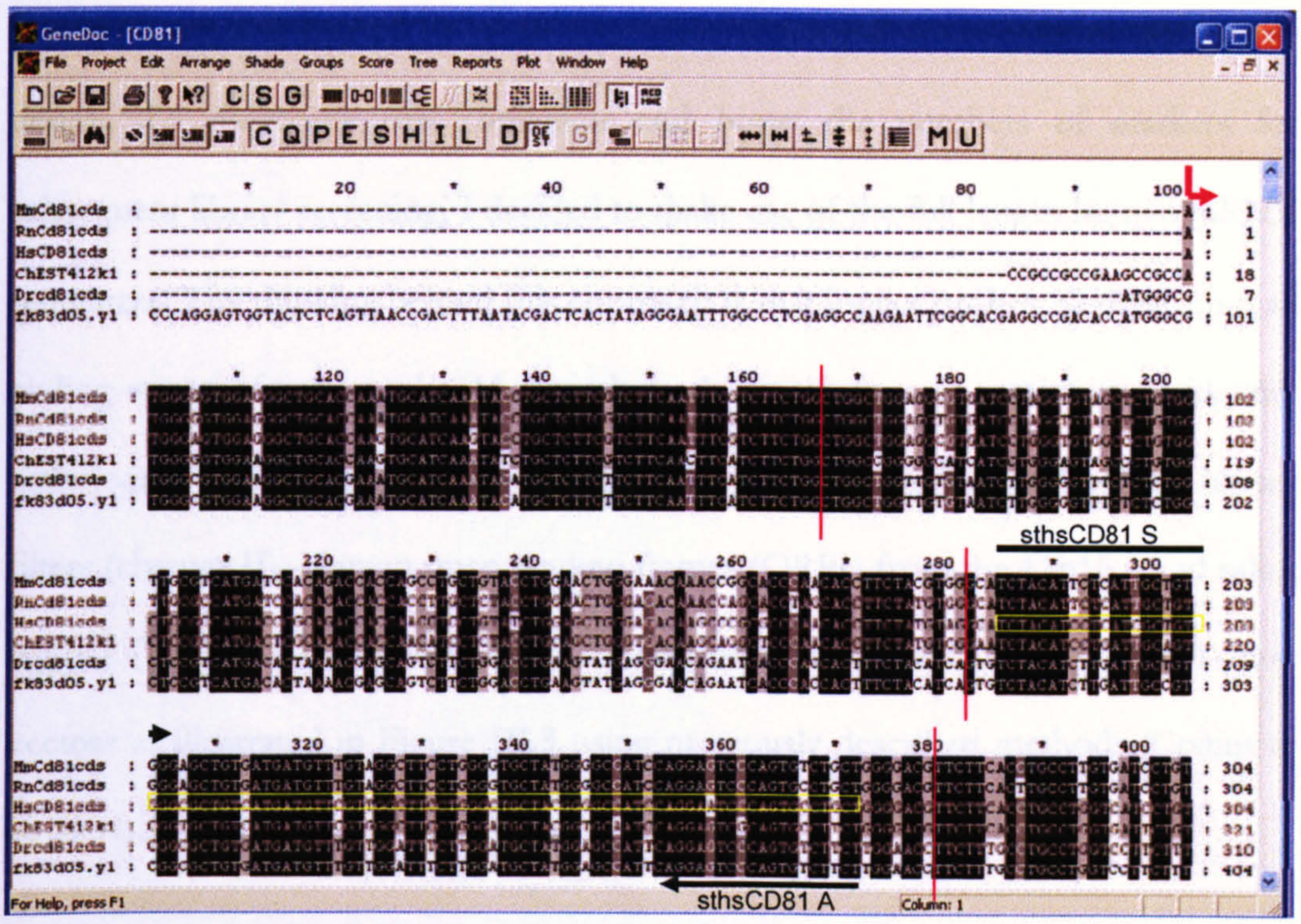


Figure III.2. Multi-species sequence alignment of *CD81* gene sequences.

From top to bottom the sequences are *Mus musculus Cd81* coding sequence, *Rattus norvegicus Cd81* coding sequence, *Homo sapiens CD81* coding sequence, *Gallus gallus CD81* EST, *Danio rerio Cd81* coding sequence and *Fugu rubripes Cd81* EST. The translation start is labelled with a red arrow. Exon-intron boundaries are demarked by red vertical lines. An 88 bp STS (sthsCD81) outlined in yellow is amplified with sense (S) and antisense (A) oligonucleotides (black arrows).

The start codon (ATG) and exon/intron boundaries for each gene were annotated by comparing the mRNA sequences with the finished human genome sequence in the UCSC genome browser. Human sequences within the most highly conserved exonic sequences, with a minimum length of 80 bp, were then submitted to PRIMER 3.0 for primer design. Sequence tagged sites (STSs, one form of landmark) were tested at three different annealing temperatures (typically 55°C, 60°C and 65°C) on genomic DNA from chicken, wallaby and platypus. This approach generated useful cross-species probes for approximately 30% of genes. However, the utility of

this approach was limited by small exon sizes which put constraints on the primer design. To overcome this constraint and boost the numbers of markers for subsequent library screening, I decided to make use of the full length human mRNA sequences. The thinking behind this approach is that longer probes, derived from all coding exons of a gene, should provide both specificity and sensitivity to identify orthologues of that gene upon reduced stringency hybridisation to the BAC library filters (chapter II). Human open reading frames (ORFs) from the 11p15.5 and other regions (Table III-1 and Table III-2, respectively) were cloned into pGEM T-Easy vectors as illustrated in Figure III.3 using previously described methods (Collins et al. 2004).

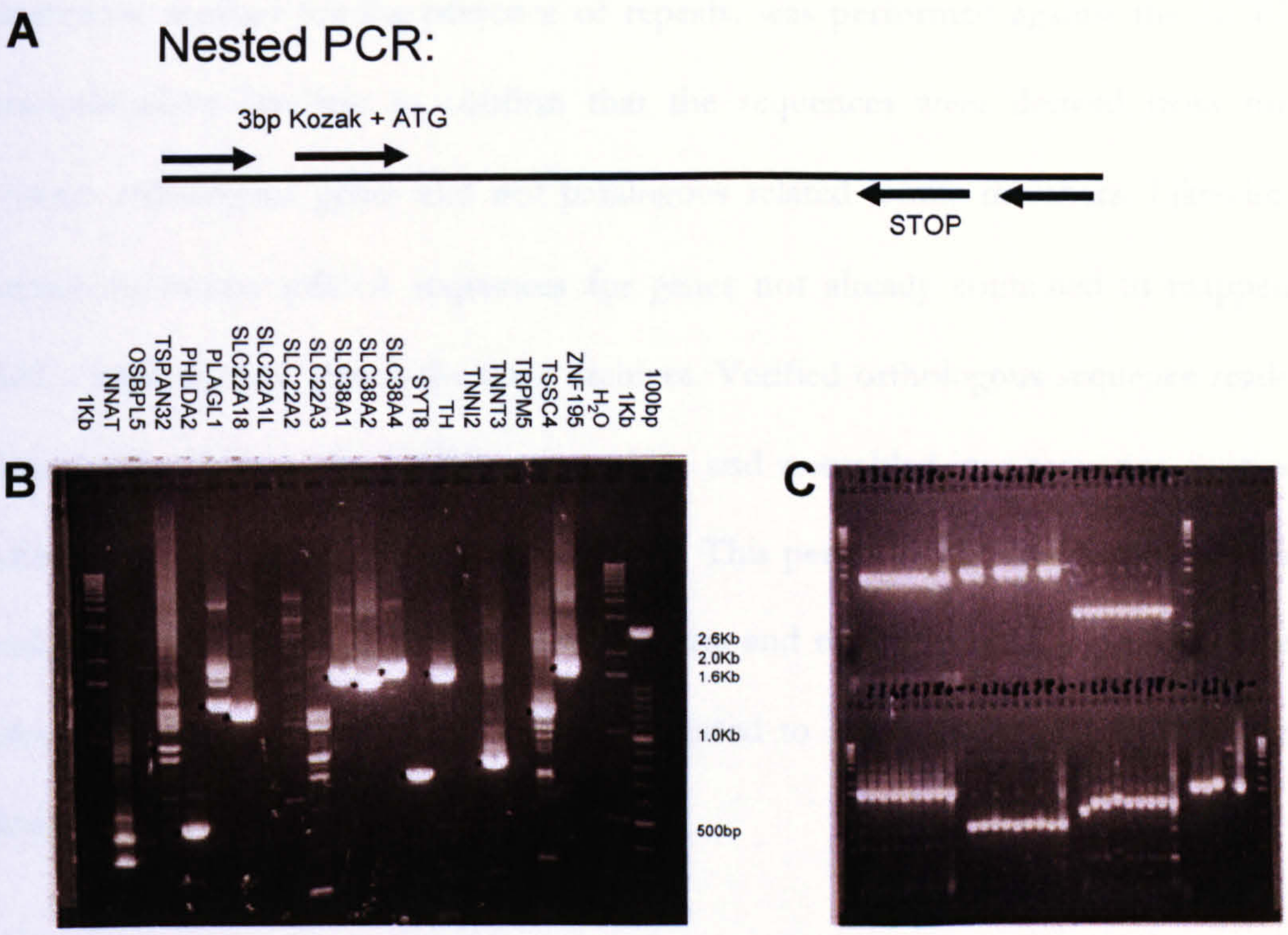


Figure III.3. Strategy for cloning human open reading frames.

Nested PCR was used to amplify full length human ORFs (A) and PCR products were separated by electrophoresis in a 1% agarose gel (B). Products of the expected size were excised, A-tailed, cloned into pGEM T-Easy vectors and transformed into *E. coli* JM109 cells. Bacterial colonies were tested by PCR and reaction products visualised in a 1% agarose gel (C). Three representative PCR products were sequenced to verify the gene inserts.

Sixty four percent (25/39) of human ORFs were successfully cloned and sequence verified from all regions (Table III-1 and Table III-2). This provided another useful set of hybridisation probes for wallaby and platypus BAC library screening at reduced stringency and increased the overall marker density. Between the two approaches probes for 75% (21/28) of genes in the 11p15.5 region were developed.

To further increase the number of markers, recently available platypus and wallaby whole genome shotgun sequence reads were identified using discontiguous megaBLAST at the NCBI trace archive. Reciprocal BLAST analysis of the identified

sequences, masked for the presence of repeats, was performed against the NCBI non-redundant database to confirm that the sequences were derived from the correct orthologous genes and not paralogous related family members. Likewise, human or mouse mRNA sequences for genes not already contained in mapped BACs were used to search the trace archives. Verified orthologous sequence reads were imported from the NCBI trace archive and assembled into sequence contigs within a GAP4 database (Staden et al. 2000). This permitted the manual inspection and editing of sequences to resolve ambiguities and remove vector sequences. The edited consensus sequences were then submitted to PRIMER 3.0 for STS primer design.

Lastly, during the gap closure phase of mapping (section 3.2.3) STSs were designed within BAC end-sequences. The total numbers of designed and tested markers, available for library screening, for each species are given in Table III-3.

Table III-1. Cloning of human ORFs from the 11p15.5 region.

Gene	Expression	mRNA Accession	Length (bp)	CDS/ORF coordinates	Translation frame	Cloned
CTSD	B	NM_001909.3	2205	134-1372	2	Yes
SYT8	B	NM_138567.2	1672	97-615	1	Yes
TNNI2	B	NM_003282.1	701	27-575	3	No
LSP1	B	NM_002339.1	1631	109-1128	1	Yes
TNNT3	B	NM_006757.1	1000	13-789	1	Yes
MRPL23	B	NM_021134.2	701	56-517	2	Yes
H19#	M	XR_000200	1072	N/A	N/A	No
IGF2	P	NM_000612.2	1356	553-1095	1	Yes
IGF2AS	P	NM_016412.1	2056	128-841	2	No
INS	P	NM_000207.1	450	45-377	3	Yes
TH	M	NM_000360.1	1816	20-1513	2	Yes
ASCL2	M	NM_005170.2	1864	621-1202	3	No
C11orf21	B	NM_014144.1	2967	259-657	1	No
TSPAN32	B	NM_005705.3	1309	161-1033	2	Yes
CD81	M	NM_004356.3	1497	234-944	3	Yes
TSSC4	M	NM_005706.2	1443	182-1171	2	Yes
TRPM5	P	NM_014555.2	3913	10-3507	1	No
KCNQ1	M	NM_000218.2	3262	109-2139	1	No
KCNQ1DN#	M	NM_018722.1	1067	635-841	2	Yes
CDKN1C	M	NM_000076.1	1511	261-1211	3	Yes
SLC22A18AS#	M	NM_007105.1	1342	499-1260	1	No
SLC22A18	M	NM_002555.3	1549	205-1479	1	Yes
PHLDA2	M	NM_003311.3	937	57-515	3	Yes
NAP1L4	M	NM_005969.3	2564	142-1269	1	Yes
CARS	B	NM_001751.3	2524	71-2317	2	Yes
OSBPL5	M	NM_020896.2	3873	117-2756	3	No
FLJ36102#	B	NM_173590.1	1817	405-881	3	No
ZNF195	B	NM_007152.1	2394	46-1935	1	Yes
ART5	B	NM_053017.2	1477	341-1216	2	No

#There is no current evidence for a protein product for these genes. B, biallelic expression; M, maternal expression in at least one tissue of mouse and/or human; P, paternal expression.

Table III-2. Cloning of human ORFs from non-11p15 imprinted domains.

Gene	Expression	Chromosome Location	mRNA Accession	Length (bp)	CDS/ORF coordinates	Translation frame	Cloned
SLC38A1	B	12q13.11	NM_030674.2	3105	455-2038	2	Yes
SLC38A2	B	12q13.11	NM_018976.3	4861	343-1863	1	Yes
SLC38A4	P	12q13.11	NM_018018.2	3965	365-2008	2	Yes
GNAS	M	20q13.2	NM_000516.4	1926	357-1541	3	Yes
DLK1	P	14q32	NM_003836.3	1547	154-1305	1	Yes
DIO3	P	14q32	NM_001362.1	2066	221-1057	2	No
MEG3#	M	14q32	XR_000167	1236	N/A	N/A	No
PLAGL1	P	6q24	NM_002656.2	4354	1936-3171	1	Yes
IGF2R	M	6q26	NM_000876.1	9090	148-7623	1	No
SLC22A2	M	6q26	NM_003058.2	2512	171-1838	3	No
SLC22A3	M	6q26-27	NM_021977.2	5624	28-1698	1	No
NNAT	P	20q11.2-q12	NM_005386.2	1338	128-373	2	Yes

#There is no current evidence for a protein product for these genes. B, biallelic expression; M, maternal expression in at least one tissue of mouse and/or human; P, paternal expression.

3.2.2 Library screening

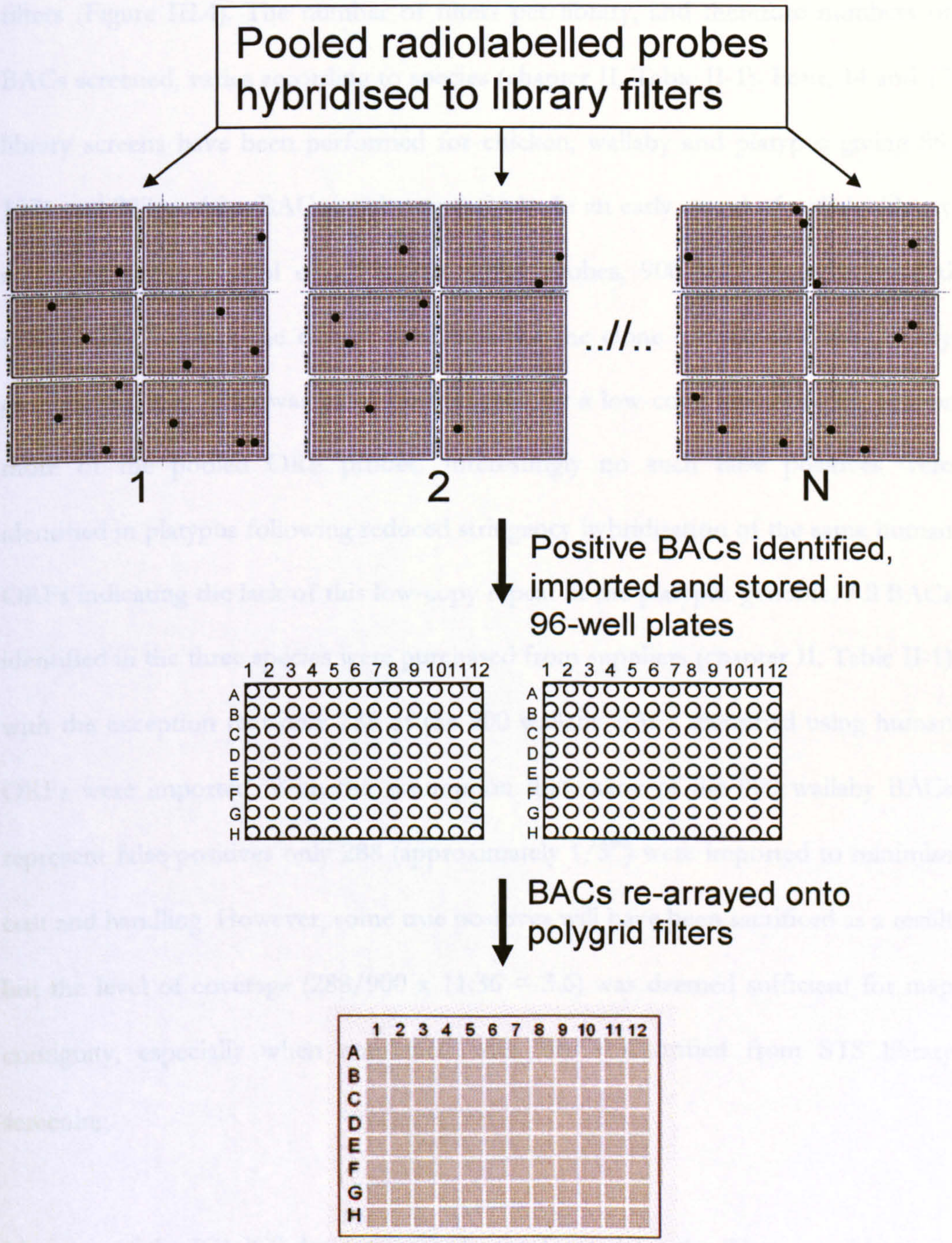


Figure III.4. Library screening strategy.

See text for details.

In order to identify BACs for all 9 regions in wallaby, platypus and chicken (IC1-IC2 region only), successfully developed STSs or ORFs were PCR radiolabelled and pooled for hybridisation to high density gridded arrays of BAC clones on 22x22 cm filters (Figure III.4). The number of filters per library, and therefore numbers of BACs screened, varies according to species (chapter II, Table II-1). Four, 14 and 17 library screens have been performed for chicken, wallaby and platypus giving 86, 1571 and 927 positive BAC signals respectively. In an early round of wallaby library screening, using a pool of 17 human ORF probes, 900 BACs were identified representing 5 times the expected number for the clone coverage of the wallaby genome (11.36x). This was most likely caused by a low-copy repeat within one or more of the pooled ORF probes. Interestingly no such false positives were identified in platypus following reduced stringency hybridisation of the same human ORFs indicating the lack of this low-copy repeat in the platypus genome. All BACs identified in the three species were purchased from suppliers (chapter II, Table II-1) with the exception that only 288 of the 900 wallaby BACs identified using human ORFs were imported. With the assumption that many of the 900 wallaby BACs represent false positives only 288 (approximately 1/3rd) were imported to minimise cost and handling. However, some true positives will have been sacrificed as a result but the level of coverage (288/900 x 11.36 = 3.6) was deemed sufficient for map contiguity, especially when combined with BACs identified from STS library screening.

Mapping of the IC1-IC2 domains was also performed for the Western wild mouse (*Mus spretus*) to provide a resource of sequence variants (chapter IV). Forty four evenly spaced STSs (approximately 1 STS per 35 kb), were designed to the

contiguous high-quality *Mus musculus* sequence exported from the UCSC genome browser. Of these STSs, 38 (86%) were successfully amplified from genomic DNA of a congenic mouse strain (SD7) containing *Mus spretus* distal chromosome 7 on a *Mus musculus* background. Thirty three STSs were radiolabelled and combined in 3 pools for hybridisation to the *Mus spretus* (SPRET/Ei) BAC library identifying a total of 103 BACs.

Finally, to support the research findings in the IC1 region of tammar wallaby (chapter VI), the South American grey, short-tailed opossum was mapped in this region alone. Thirteen STSs were designed, 5 from the published 10,837 bp *Monodelphis domestica IGF2* sequence (DQ519591.1, Lawton et al. 2007) and 8 from the end sequences of 4 WGS sequence contigs (section 3.4, Figure III.10). Six of these STSs were radiolabelled and pooled to screen the opossum BAC library filters, identifying 30 BACs.

Table III-3. Mapping resources developed.

Common name	Western Wild mouse	Tammar wallaby	Grey Short-tailed opossum	Duck-billed Platypus	Chicken	TOTALS
Species name	*Mus spretus*	*Macropus eugenii*	*Monodelphis domestica*	*Ornithorhynchus anatinus*	*Gallus gallus*	
Number of STSs designed	44	233	13	265	179	734
Number of STSs passing primer testing (%)	38 (86)	200 (86)	11 (85)	177 (67)	168 (94)	594 (81)
Number of STSs in pooled hybridisation	33	109	6	92	42	282
Number of BACs identified	103	1571	30	927	86	2717
Number of BACs in FPC	97	1369	30	732	73	2301
Number of BAC contigs$	1	173	1	69	5	249
Number of BACs with landmark content established*	83	358	3	397	80	921
Number of BACs selected for sequencing	11	34	3	32	12	92

$, at least two clones. *, by PCR and/or hybridisation.

3.2.3 Landmark content mapping

Landmark content mapping is an approach that determines the presence or absence of a unique genomic feature (landmark) in a clone, and uses that information to establish overlaps between clones. Landmarks serve to anchor contigs (overlapping DNA fragments, Staden. 1979) to a framework map such as the human transcript map of a region and are therefore of great utility in establishing the order and orientation of contigs across the region. Often, but not exclusively, landmarks are STSs. The landmark content of clones can be achieved by hybridisation, PCR or electronic PCR (ePCR – chapter II). To establish which of the radiolabelled probes in a library screening pool have identified which BACs and therefore where the BACs map relative to each other on the framework map, individual STSs were used to screen the BAC collections by 96 or 384-well PCR and/or hybridisation to the 'polygrid' filters arrayed by the in-house clone resource group. Landmark content analysis by PCR (Figure III.5) has the advantage of being very rapid; clones received from external providers (chapter II, Table II-1) can be tested the following day using colony PCR (chapter II). However, PCR is not practical when testing thousands of clones with individual STSs. In contrast, screening of the polygrid filters by DNA hybridisation (Figure III.6) enables thousands of clones to be tested simultaneously for the presence of an STS but generating the polygrids is time-consuming. Since the polygrids are arrayed with chicken, wallaby and platypus BACs, highly conserved probes frequently gave orthologous signals from more than one species. Where an STS developed in one species cross-hybridises to the BACs of another species this gave independent confirmation of conserved synteny.

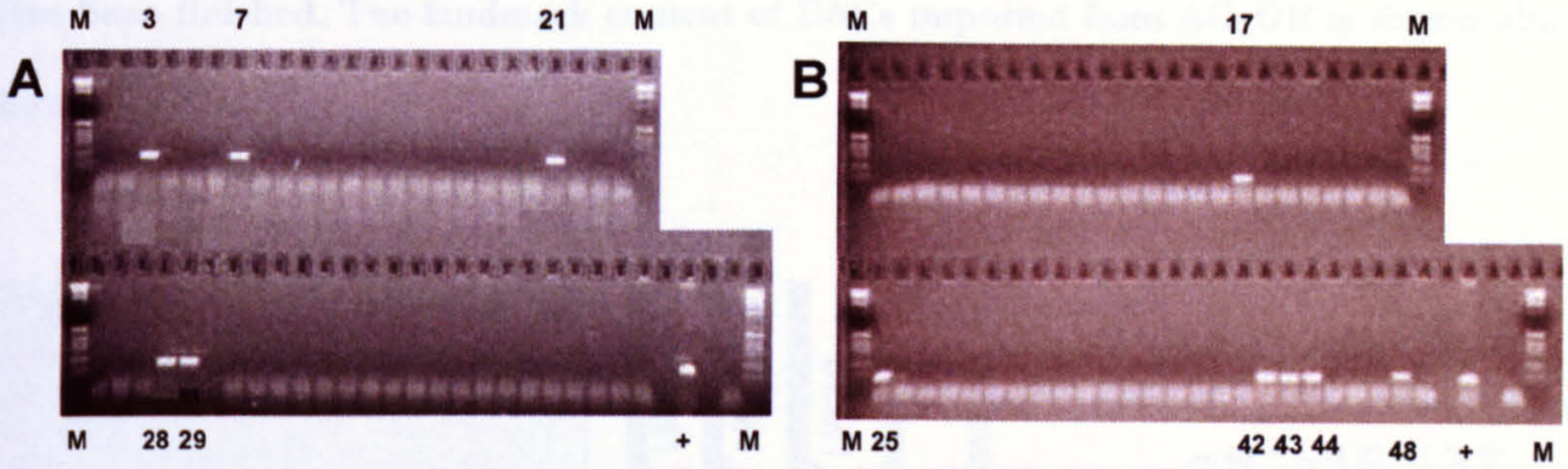


Figure III.5. Landmark content analysis of chicken BACs by colony PCR.

STSs for Chicken ESTs stggMRPL23 (A) and stggKCNQ1 (B) were tested against 49 different BAC colonies (lanes 1-49), chicken genomic DNA (+) and water (lane 53). PCR products were separated by 2.5% agarose gel electrophoresis. M, size marker. Numbered lanes indicate BACs that are positive for the STS tested. BAC names are provided in the boxes. Landmark content data was imported into FPC (via ACeDB) to complement fingerprinting data and assist with contig building. In the chicken BAC contig (C) horizontal lines represent individual BAC clones the length of which is proportional to the number of bands they contain. Red highlighted clones were selected for sequencing and

have been finished. The landmark content of BACs imported from ACeDB is shown above the clones.

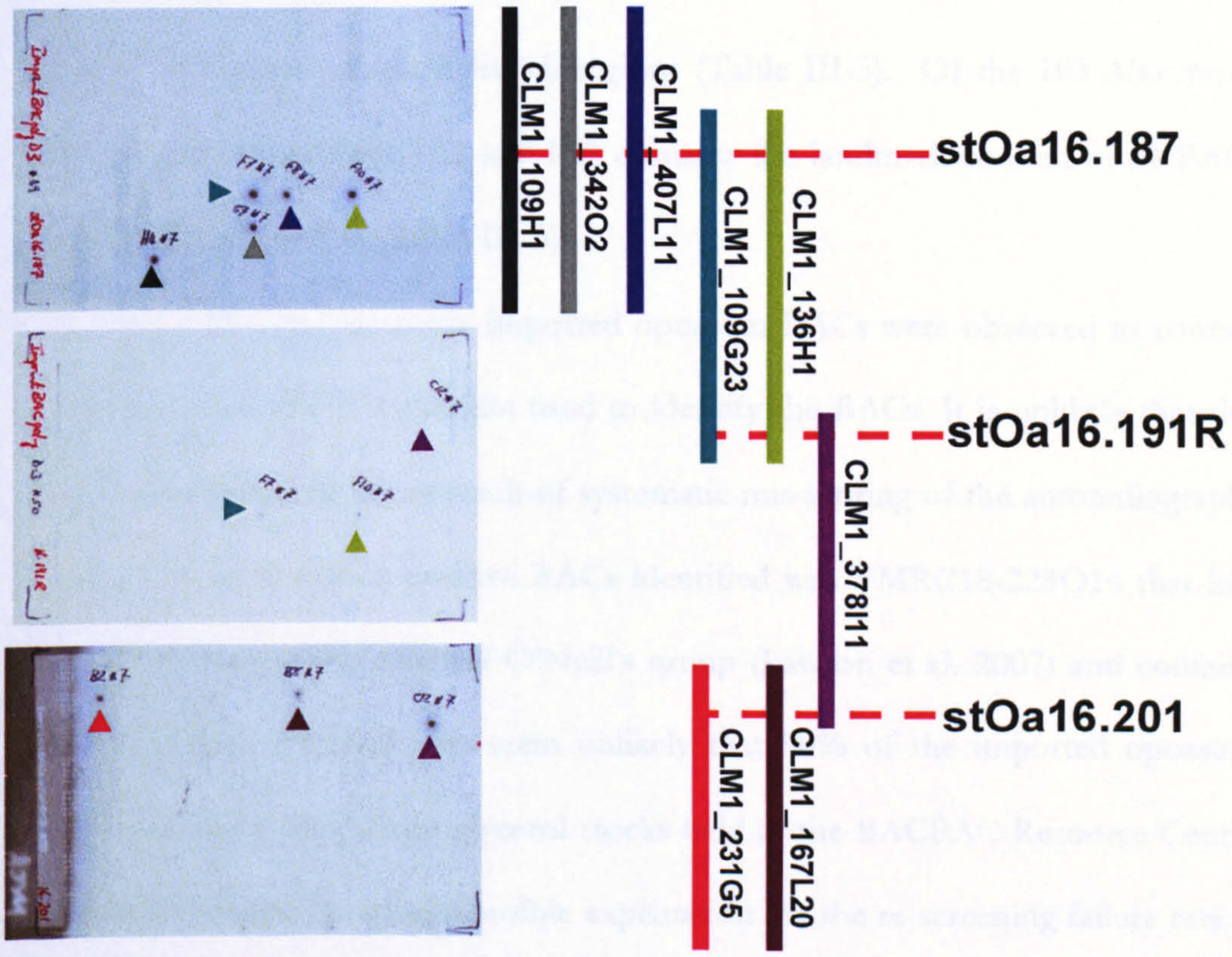


Figure III.6. Landmark content mapping through polygrid screening.

Three platypus STSs developed from platypus WGS ultracontig16 were screened by hybridisation against identical copies of the ImprintBACpoly_D_3 (generation 3) polygrid. STS stOa16.187 is contained within 5 BACs, 2 of which were also identified by stOa16.191R. A third BAC (CLM1_378I11) identified by stOa16.191R is also positive for stOa16.201. The contig constructed is part of a larger contig mapping to platypus chromosome 8p orthologous to the *STX16-GNAS* complex region. Clones CLM1_109G23 and CLM1_378I11 were two of the four selected for sequencing from this region.

Four generations of polygrid filters were created during this study with each generation arrayed with more clones than the previous one. Polygrids were screened by hybridisation with 12 STSs from chicken, 127 STSs from wallaby and 94 STSs from platypus resulting in the landmark content of 30, 284 and 298 BACs respectively. In addition 93 STSs from chicken, 45 STSs from wallaby and 73 STSs

from platypus were amplified by PCR from 78 chicken, 170 wallaby and 309 platypus BACs respectively. Accounting for redundancy between mapping methods (PCR and hybridisation) the landmark content of 80 chicken, 358 wallaby and 397 platypus BACs was obtained for all regions (Table III-3). Of the 103 *Mus spretus* BACs identified from the IC1 and IC2 domains the landmark content of 83 BACs was established by PCR (Table III-3).

Only three (10%) of the thirty imported opossum BACs were observed to contain the 6 STSs from the IC1 domain used to identify the BACs. It is unlikely that this poor re-screening rate is the result of systematic mis-scoring of the autoradiographs because one of the three positive BACs identified was VMRC18-223O16 that had been FISH mapped by Michael O'Neill's group (Lawton et al. 2007) and contains the *IGF2* gene. It would also seem unlikely that 90% of the imported opossum BACs were mis-picked from glycerol stocks held at the BACPAC Resource Center (CHORI). Perhaps the most plausible explanation for the re-screening failure rate is that one or more of the six clustered STSs used to screen the library cross-hybridises to BACs from other genomic regions as a result of a low copy repeat not present in the RepeatMasker library. The specificity of PCR revealed that the 3 BACs mapping to the IC1 domain, including VMRC18-223O16, partially overlap and encompass the entire IC1 domain and flanking regions (Figure III.10).

3.2.4 Restriction endonuclease fingerprinting

The single enzyme digestion (fingerprinting) of BACs (Marra et al. 1997) provides a high-throughput means to assemble bacterial clone contigs and has been widely used to physically map genomes (Gregory et al. 2002, Humphray et al. 2007, McPherson et al. 2001). Identified and imported BACs following library screening

for all species were digested with the restriction endonuclease *Hin*dIII enzyme to generate a clone fingerprint (chapter II). Gel images from the *Hin*dIII fingerprinting experiments were processed using IMAGE software (http://www.sanger.ac.uk/Software/Image). Briefly, each lane of a 121-well 1% agarose gel needs to be tracked and individual bands called. This enables the conversion in IMAGE of raw data into a set of normalised integers corresponding to individual fingerprint bands. In practice a great deal of manual editing is required to produce the final data set. The extent of overlap between two clones is established by statistical comparison of their shared fingerprint bands within the program FingerPrinting Contig (FPC, Soderlund et al. 1997). When combined with landmark content data, imported into FPC from ACeDB, the order and orientation of these contigs is readily determined (Figure III.7). Resulting species-specific FPC databases therefore offer a powerful tool with which to build contigs and select optimal tile paths for sequencing. Table III-3 details the numbers of BACs fingerprinted and contigs assembled for each of the species. For each species the average *Hin*dIII fragment (band) size multiplied by the number of bands in any given contig is used to determine its approximate length prior to the availability of sequence.

In the IC1/IC2 region large, uninterrupted fingerprinting contigs have been generated for *Mus spretus*, wallaby and chicken. The largest contig generated has 73 BACs spanning almost 1.6 Mb from the *CTSD* gene to the *OSBPL5* gene of wallaby chromosome 2p (Figure III.7D and Figure III.8). In the orthologous region of platypus chromosome 3p, 8 contigs were assembled.

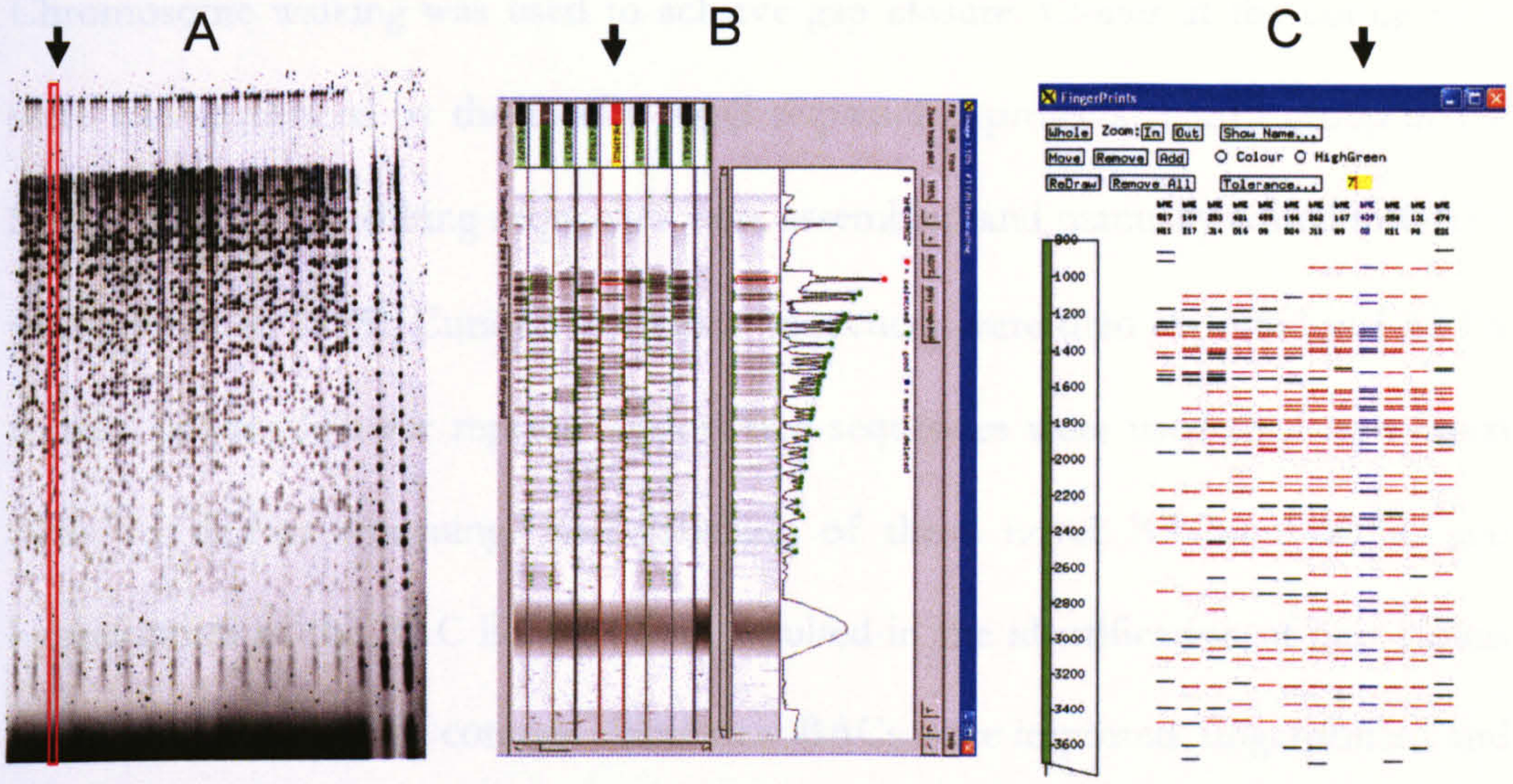


Figure III.7. The process of fingerprint mapping.

Micro-prepped BAC DNAs in a 96-well plate are digested with *Hin*dIII and loaded in a 121-well 1% agarose gel with size markers in every 5th well (A). After scanning the gel, band-calling is performed in the IMAGE 3.10b software (B). The fingerprints are digitised (C) for reading into FPC. The figure illustrates fingerprint mapping of wallaby BAC clones from the IC1-IC2 domains. The black arrow follows the BAC bME439O2 through this process.

3.2.5 Gap closure

Bacterial clone contigs assembled through the combined use of fingerprinting and landmark content analysis will inevitably contain gaps, either as a result of inadequate marker density used to screen the BAC libraries or an under representation of certain genomic regions within the libraries screened. This under representation could simply be due to statistics, for example, even with estimated genome coverage of 99%, 1 in 100 regions might be missing. Alternatively bias may have been introduced into the library because of the restriction enzyme chosen to digest genomic DNA for cloning and/or may be due to the instability of some sequences (e.g. tandem repeats) in the vector host (*E. coli*). Some foreign DNA may even be harmful to *E. coli* when transcribed or translated.

Chromosome walking was used to achieve gap closure. Clones at the contig ends were end-sequenced by the faculty small sequencing projects (FSSP) group at the Sanger Institute. Resulting sequences were assembled and manually edited in GAP4 (Bonfield et al. 1995). Curated consensus sequences were then exported in FASTA format and masked for repeats. Repeat-free sequences were used to design novel STSs for library screening. Radiolabelling of these novel BAC-end STSs, and hybridisation to the BAC library filters, resulted in the identification of new clones mapping to the ends of contigs. These new BACs were imported, fingerprinted and merged with existing contigs in FPC for analysis against other contig ends. The landmark content of gap closure clones was also established to identify clones overlapping only slightly and therefore not detected by fingerprinting. Iterative chromosome walking was performed until all contigs in an orthologous region were joined or the density of repeats flanking the gaps precluded further walking.

3.3 FISH mapping of BACs to wallaby and platypus chromosomes

The BAC resources developed in this thesis permitted us to begin to address questions of the origins of the genomic imprinting mechanism. There are several theories to account for how the mechanism may have evolved which include the hypothesis that it was driven by the evolution of X chromosome inactivation (XCI, Lee. 2003), or that it arose from an ancestrally imprinted chromosome (Walter and Paulsen. 2003) (details in chapter I). If BACs containing orthologues of therian imprinted genes were found to map to sex chromosomes in ancestral species, in which imprinting has not been identified, this would support the hypothesis that the mechanism of genomic imprinting evolved from selection pressures acting on XCI. Likewise if BACs containing orthologous genes imprinted in therians but not in

monotremes or birds were found to lie on one or a few autosomal chromosomes in the platypus or chicken this would lend some support to the hypothesis of an ancestrally imprinted chromosome.

In collaboration with Carol Edwards (Department of Physiology, Development and Neuroscience, Cambridge) and Willem Rens (Department of Veterinary Medicine, Cambridge) FISH mapping of platypus and wallaby BACs to platypus and wallaby metaphase chromosome spreads from cells in culture was performed to establish the genomic location of imprinted gene orthologues (Edwards et al. 2007). Eight orthologues of imprinted genes, representing seven clusters, were localised to platypus chromosomes. A further eight tammar wallaby BACs containing imprinted gene orthologues were mapped to wallaby chromosomes (Table III-4).

Table III-4. Summary of chromosomal locations of genes studied in human, mouse, wallaby, platypus and chicken genomes.

FISH mapped genes	Human location	Mouse location	Wallaby location	Platypus location	Chicken Location
MRPL23/IGF2/CD81	11p15.5	7F5	2p	3p	5
DLK1/DIO3	14q32	12E-F1	1q	1q	5
GNAS	20q13	2E1-H3	1q	8p	20
GRB10	7p12	11A1	3p	4p	2
IGF2R	6q26	17A-C	2q	Centric 2	3
SLC38A4	12q13	15F1	3p	2q	1
UBE3A	15q12	7C	5#	18p	1

#(Rapkins et al. 2006)

The finding of imprinted gene orthologues distributed throughout the autosomal chromosomes of platypus and chicken indicates that both 'XCI driven evolution' and 'single ancestrally imprinted chromosome' hypotheses are unlikely to be true. Although the mechanism of platypus dosage compensation is unknown it seems probable that this mechanism preceded that of genomic imprinting. However, the biological mechanisms required to silence autosomal genes (imprinting) may have been reused from those of XCI in a process called exaptation (Gould and Vrba.

1982). Taken together with observations that some imprinted loci in placental mammals are not imprinted in marsupial mammals (for example *CDKN1C*, Suzuki et al. 2005) this suggests that the mechanism of genomic imprinting arose gradually after the evolution of viviparity and continued to evolve convergently in therian lineages.

Not only do the chromosomal locations of representative BAC clones in an imprinting gene cluster region allow us to address specific evolutionary questions they also give important new genomic markers in otherwise poorly characterised genomes. To illustrate this with an example, one of the earliest identified and possibly best characterised imprinted genes, *IGF2*, was not mapped to a chromosome following the recent sequence and assembly of the platypus genome. A platypus *IGF2* cDNA sequence (AF225876) lies within a 12,590 bp unanchored WGS sequence contig (Contig27935 of the March 2007, ornAna1 assembly). The FISH mapping of platypus BAC CLM1_349H20 in this study now reveals that *IGF2* and flanking genes reside on platypus chromosome 3p.

3.4 Sequence clone selection

Having two independent data sets, landmark content and fingerprinting, within FPC is invaluable in selecting a minimally overlapping set of BACs (tile path) for sequencing. Since BAC fingerprints are overlapped according to the probability of sharing restriction fragments (bands), very small overlaps between BACs, corresponding to only one or two bands, will be missed in FPC. Empirically, BACs overlapping by approximately 20% of their lengths can be assembled into contigs based on fingerprint data alone. However, two BACs overlapping by as little as 100 bp can be identified if they contain a common STS. Marker content data is therefore imported into FPC from ACeDB before manual sequence clone selection. When

choosing BACs for sequencing in FPC it is important to select clones which approximate to the average insert size reported for the BAC library (chapter II, Table II-1), thus avoiding deleted or chimaeric BACs. This is also aided by the selection of clones in which all fingerprint bands are accounted for in neighbouring clones. This precludes the selection of clones at the very extremes of the mapped contig. Tags are manually assigned to the clones selected for sequencing in FPC which are then automatically entered into the Oracle database ('gull') for tracking through subsequent stages in the sequencing pipeline.

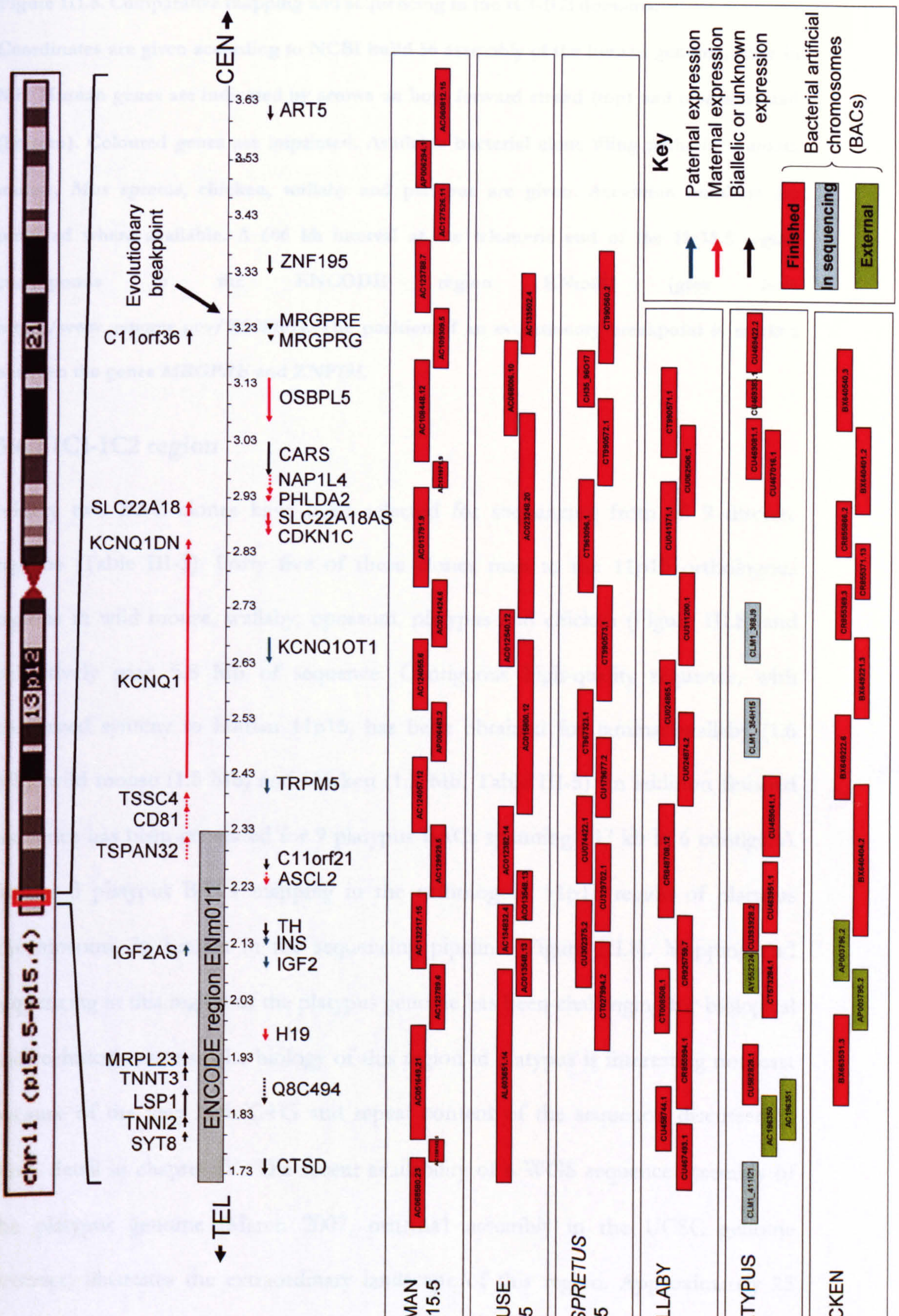

chr11 (p15.5–p15.4)
CEN
Evolutionary breakpoint
3.63
ART5
3.53
3.43
ZNF195
3.33
3.23
MRGPRE
MRGPRG
C11orf36
3.13
OSBPL5
3.03
CARS
NAP1L4
PHLDA2
SLC22A18AS
CDKN1C
SLC22A18
KCNQ1DN
2.93
2.83
2.73
KCNQ1OT1
KCNQ1
2.63
2.53
2.43
TRPM5
TSSC4
CD81
TSPAN32
2.33
C11orf21
ASCL2
2.23
TH
INS
IGF2
IGF2AS
2.13
2.03
H19
MRPL23
TNNT3
LSP1
TNNI2
SYT8
1.93
Q8C494
1.83
ENCODE region ENm011
1.73
CTSD
TEL
Key
Paternal expression
Maternal expression
Biallelic or unknown expression
Bacterial artificial chromosomes (BACs)
Finished
In sequencing
External
HUMAN 11p15.5
MOUSE 7qF5
M. SPRETUS 7qF5
WALLABY 2p
PLATYPUS 3p
CHICKEN 5

Figure III.8. Comparative mapping and sequencing in the IC1-IC2 domains.

Coordinates are given according to NCBI build 36 assembly of the human genome, scale in Mb. Human genes are indicated by arrows on both forward strand (top) and reverse strand (bottom). Coloured genes are imprinted. Available bacterial clone tiling paths for human, mouse, *Mus spretus*, chicken, wallaby and platypus are given. Accession numbers are provided where available. A 606 kb interval at the telomeric end of the 11p15.5 region corresponds to the ENCODE region ENm011 (grey box, http://www.genome.gov/10005107). The position of an evolutionary breakpoint is marked between the genes *MRGPRE* and *ZNF195*.

3.4.1 IC1-IC2 region

Ninety two BAC clones have been selected for sequencing from all 9 discrete regions (Table III-3). Forty five of these clones map to the 11p15 orthologous regions in wild mouse, wallaby, opossum, platypus and chicken (Figure III.8) and collectively span 5.8 Mb of sequence. Contiguous high-quality sequence, with conserved synteny to human 11p15, has been obtained for tammar wallaby (1.6 Mb), wild mouse (1.5 Mb) and chicken (1.3 Mb, Table III-5). In addition finished sequence has been generated for 9 platypus BACs spanning 737 kb in 6 contigs. A further 3 platypus BACs mapping in the orthologous 11p15 region of platypus chromosome 3p remain in the sequencing pipeline (Figure III.8). Mapping and sequencing in this region of the platypus genome has been challenging for biological and technical reasons. The biology of this region in platypus is interesting not least because of the very high C+G and repeat content of the sequence, discussed in more detail in chapter IV. The recent availability of a WGS sequence assembly of the platypus genome (March 2007, ornAna1 assembly in the UCSC genome browser) illustrates the extraordinary landscape of this region. Approximately 25 million sequence reads representing a 6-fold coverage of the platypus genome were assembled into 205,536 contigs. The N50 length of all contigs in the genome is 967

kb where N50 length is defined as the largest value of *n* for which 50% of the basepairs in the genome lie in contigs with a length greater or equal to *n*. WGS contigs identified by BLASTN using mRNA sequences from the 11p15 region and available BAC sequences from the BAC resource developed here span a total of 450 kb and have an average length of only 6 kb (Appendix C). Whether the absence of large contigs in this region is due to the lack of coverage, problems with the assembly or a combination of the two is unclear. In stark contrast, all other regions of interest are contained within single multi-Megabase platypus WGS ultracontigs (chapter IV).

Deletions of BAC inserts are generally rare owing to their low-copy number (Osoegawa et al. 2000, Shizuya et al. 1992). Despite the general stability of BAC clones, 6 clones selected for sequencing from the platypus orthologous 11p15 region were observed to be significantly smaller (27 to 103 kb) than the 143 kb average cloned insert size for the library. So what is special about this region of the platypus genome? The most likely explanation for the observations is that high local repeat and C+G contents are responsible, discussed in detail in chapter IV. An indication of the unusual base composition of the region came with the observation of very few *Hin*dIII (AAGCTT) or *Eco*RI sites (GAATTC) within platypus BACs mapping to the region in contrast to *Bam*HI (GGATCC) sites (Figure III.9).

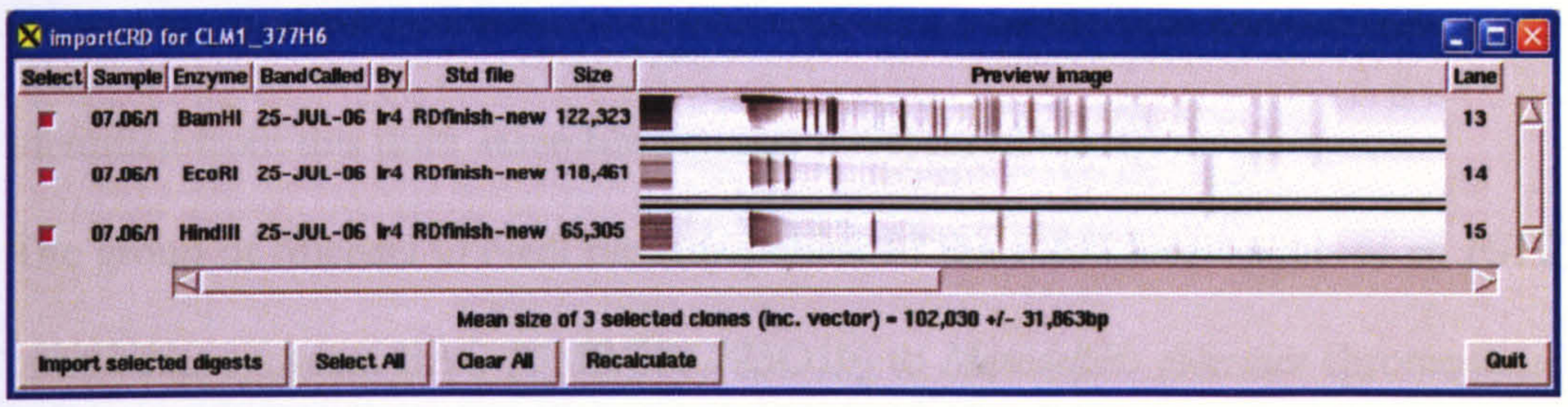


Figure III.9. Restriction endonuclease digests for platypus BAC CLM1_377H6.

Three digests for CLM1_377H6 were performed and viewed using importCRD in-house software. From top to bottom the digests are *Bam*HI, *Eco*RI and *Hind*III.

In addition to the biological challenges faced with mapping in this region of the platypus genome a technical error compounded these difficulties. Platypus BAC library filters were purchased from Clemson University Genomics Institute (CUGI) for screening with radiolabelled probes. In May 2007 CUGI sent correspondence stating that when clone arraying robots (Q-Bots) were updated, two duplication patterns were transposed on one of the machines. The result of this being that identified positive signals on 9 of the 12 library filters were incorrectly addressed and the imported clones were not the clones containing the markers of interest. The filters at fault had been screened with 26 STSs, 18 of which map to the 11p15 orthologous region, and identified 251 BACs. Of these BACs 25% were found to be mis-identified. In an 'average' genomic region such a loss in clone coverage would have little effect on map construction. However, taken together with the biological observations noted above mapping and sequencing progress in this region has been hampered.

Having a second marsupial sequence that diverged from tammar wallaby some 60 Myr ago would add significance to any findings in any one marsupial species (see chapter VI). The recent draft genome sequence of the grey, short-tailed opossum (*Monodelphis domestica*) (Mikkelsen et al. 2007) had many gaps between small scaffold

contigs in the IC1 orthologous region. The genes *IGF2* and *H19* could not be identified from this draft assembly (January 2006, monDom4).

The group of Michael O'Neill recently published the FISH localization of an *IGF2* containing opossum BAC (VMRC18-223O16) to *Monodelphis domestica* chromosome (MDO) 5q3 (Lawton et al. 2007). Therefore MDO 5q3 shares conserved synteny with tammar wallaby chromosome 2p (Table III-4). The BAC VMRC18-223O16 was kindly provided by Michael O'Neill and entered into the Sanger Institute sequencing pipeline. To extend the opossum IC1 map, draft genome scaffolds were identified by BLASTN using wallaby evolutionary conserved regions (ECRs, chapter IV). Scaffolds 1397 (12,123 bp), 1389 (12,182 bp), 792 (18,709 bp) and 777 (23,475 bp) were identified, masked for repeats and unique STSs designed. Six STSs were radiolabelled and pooled for hybridisation to the 11 VMRC-18 library filters. As discussed above only three BACs were identified as mapping to the IC1 orthologous region. Based on landmark content mapping all three BACs were assembled into a single contig. In addition to VMRC18-223O16 the BACs VMRC18-490C6 and VMRC18-151I14 were selected for sequencing and together span the entire IC1 locus from *MRPL23* to *INS* genes (Figure III.10).

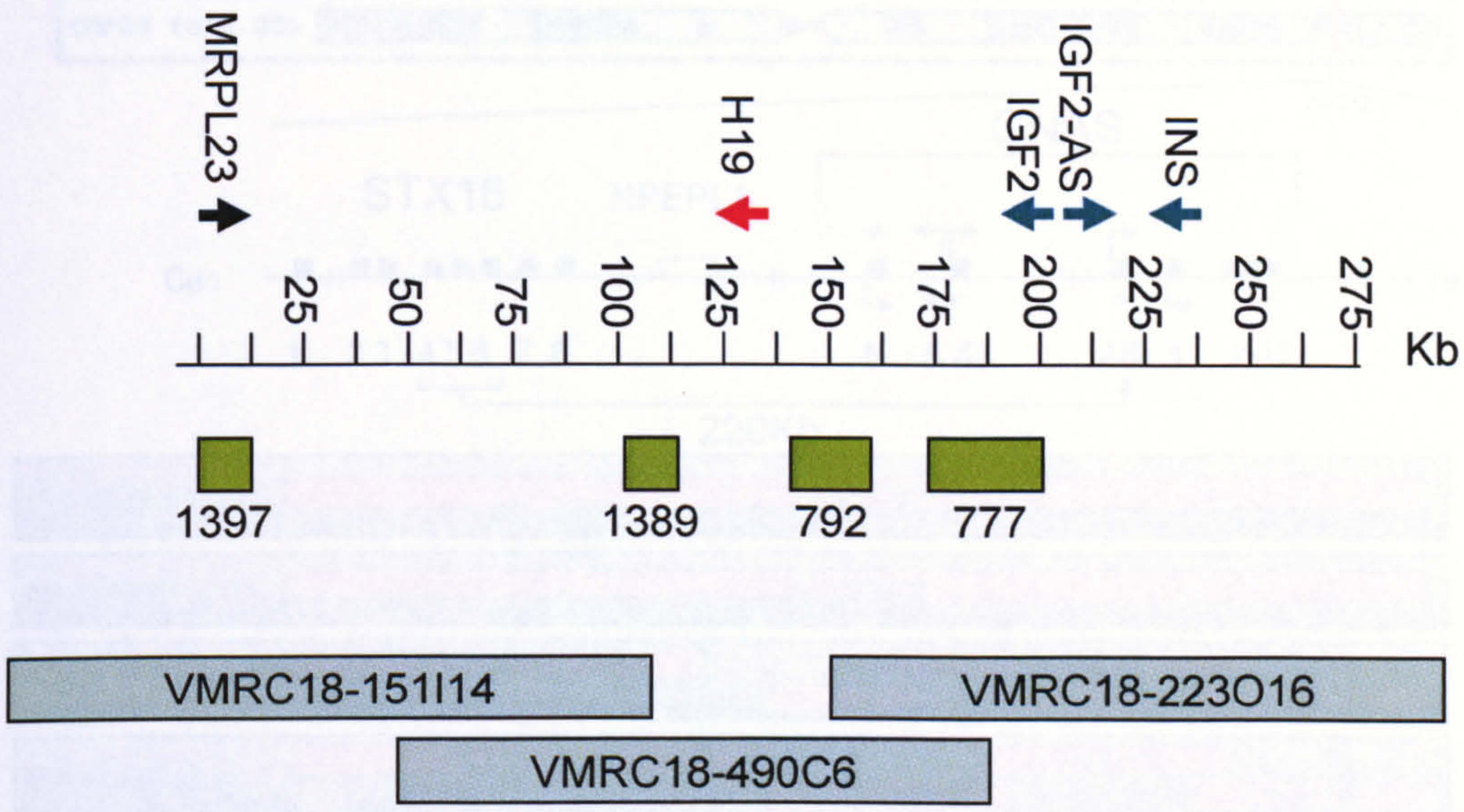


Figure III.10. Schematic of opossum mapping in orthologous IC1 region.

Green rectangles, opossum draft genome assembly scaffolds; Blue rectangles, opossum BACs.

3.4.2 *STX16-GNAS* region

Four BACs have been mapped and sequenced from the *STX16-GNAS* orthologous regions of wallaby and platypus, mapping to chromosomes 1q and 8p respectively (Figure III.11). The four, overlapping, platypus BACs span 490 kb of sequence. In wallaby, one gap remains between the BACs MEKBa-420A22 and MEKBa-266M20. Gene annotation of platypus and wallaby BAC sequences (chapter IV) reveal the presence of the family with sequence similarity 38, member A (*FAM38A*) gene positioned between *NPEPL1* and *GNAS* loci. The 3' end of *FAM38A* in wallaby BAC MEKBa-420A22 and extra-large exon of *GNAS* (*GNAS-XL*) in BAC MEKBa-266M20 indicate that this gap is no more than 100 kb in size. In human and mouse the *FAM38A* gene resides on chromosomes 16q24.3 and 8qE1, respectively.

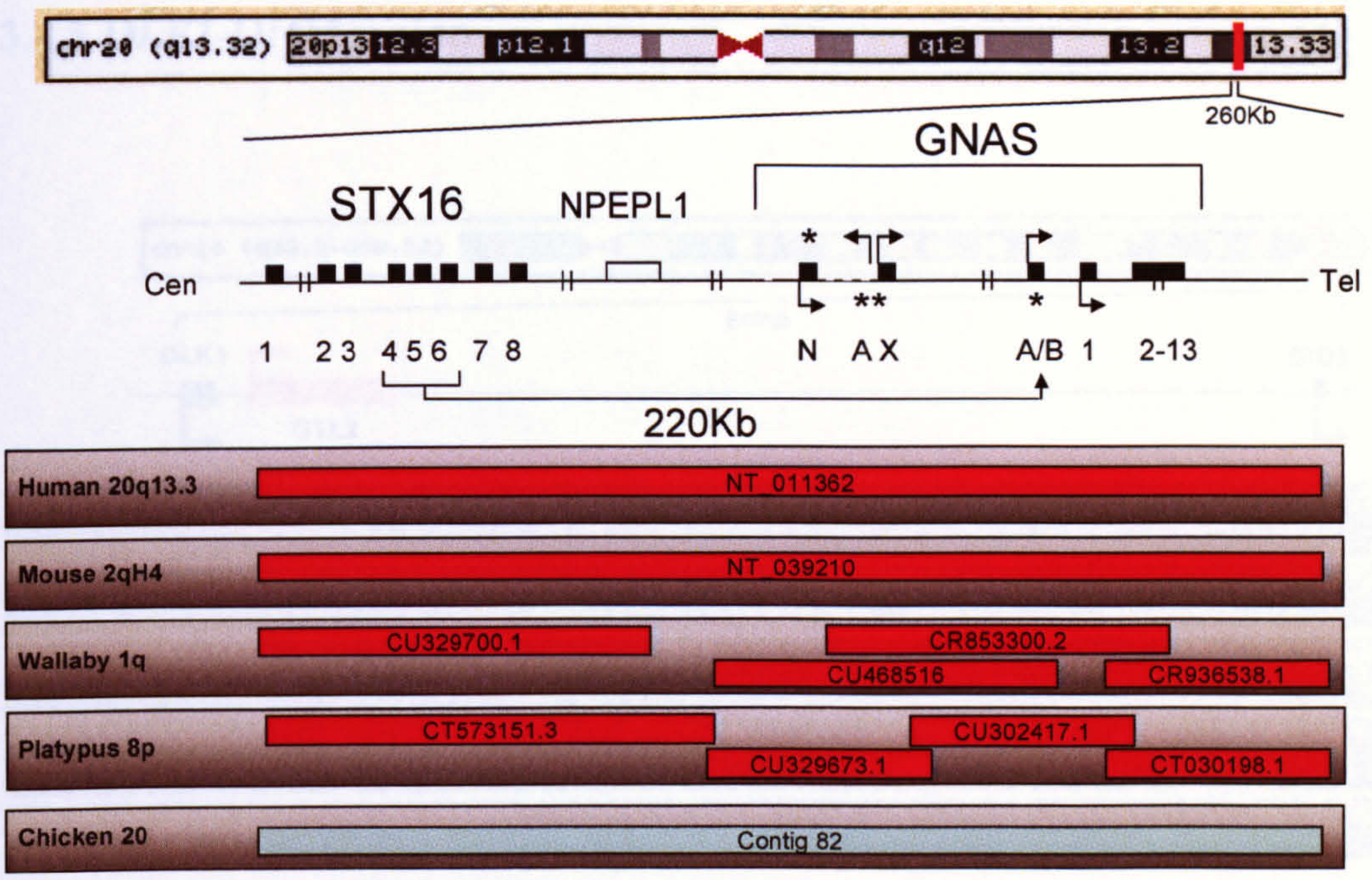


Figure III.11. Schematic of the GNAS complex region.

The human chromosome 20 ideogram indicating the 260 kb interval at 20q13.32 containing *STX16*, *NPEPL1* and *GNAS* complex genes. A micro-deletion of *STX16* exons 4-6 affecting the *GNAS* exon A/B methylation (*) status is depicted. Further details can be found in the text. Sequence coverage of the region is illustrated by rectangular boxes; red, finished sequence; blue, unfinished sequence. Human, mouse and chicken sequences for the region were obtained from the UCSC genome browser.

3.4.3 *DLK1-DIO3* region

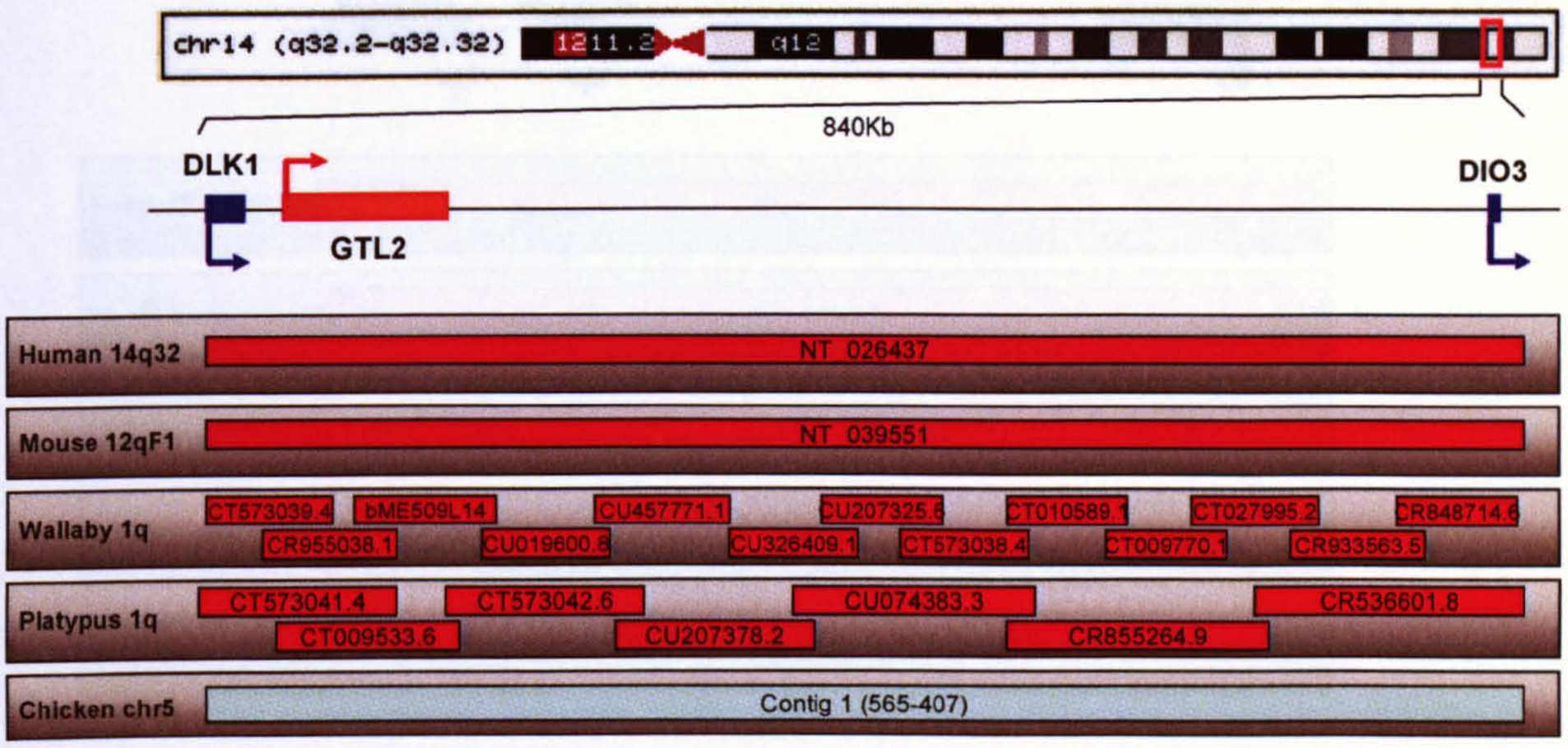


Figure III.12. Schematic of the DLK1-DIO3 region.

Key as in Figure III.11.

This region was mapped and sequenced in wallaby and platypus in collaboration with Carol Edwards and Anne Ferguson-Smith at the Department of Physiology, Development and Neuroscience, University of Cambridge. This region forms the basis for Carol's PhD thesis and is therefore not discussed further here.

3.4.4 *SLC38A2* and *SLC38A4* gene region

Orthologues of the *Slc38a2* and *Slc38a4* genes (not imprinted and imprinted in mouse, respectively) have been sequenced in both wallaby and platypus (Figure III.13). The platypus sequence in this region spans approximately 332 kb and wallaby sequence spans 319 kb (Table III-5).

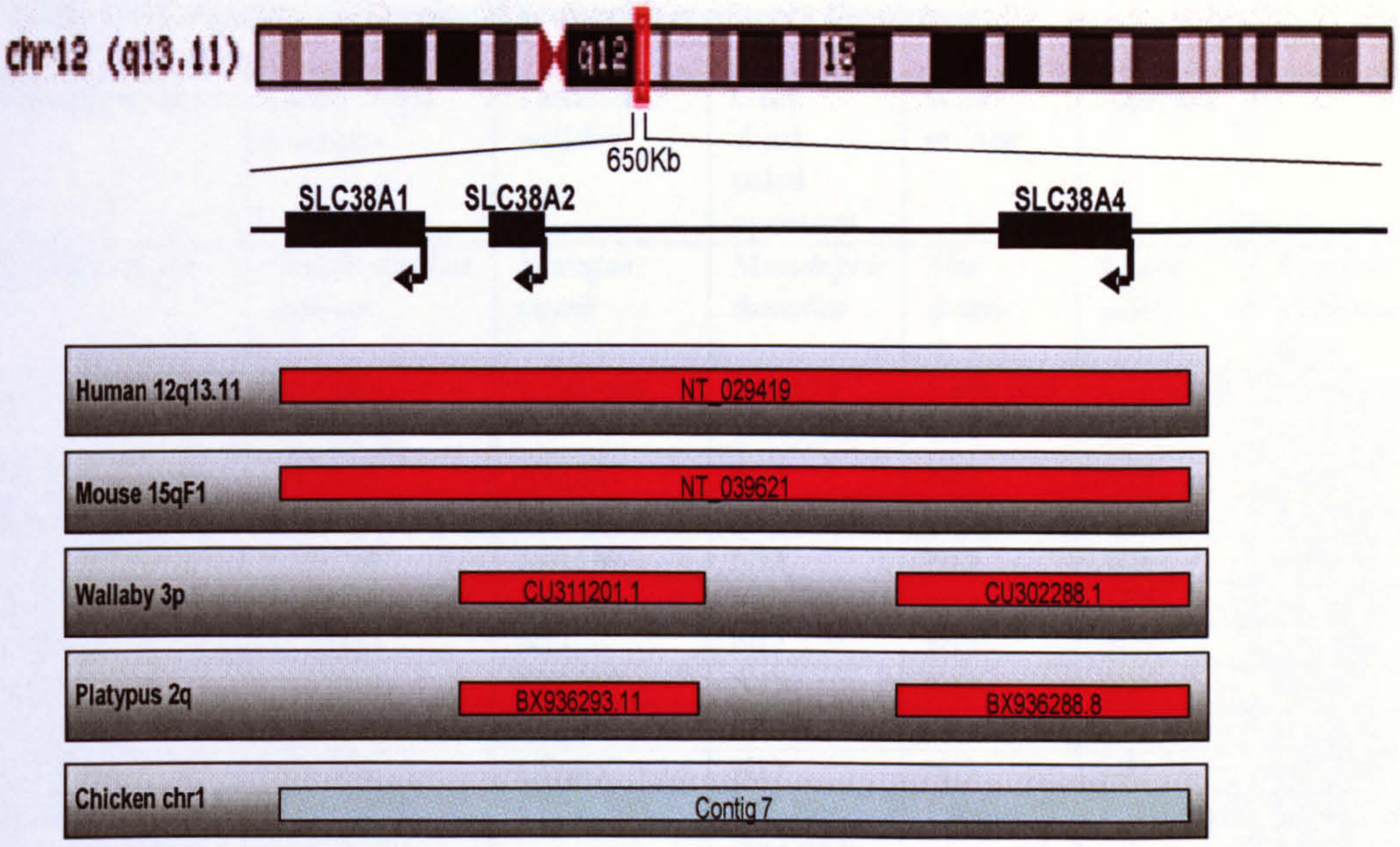


Figure III.13. Schematic of the solute carrier gene family 38 region.

Finished sequence for orthologues of two members of this gene family has been obtained in wallaby and platypus. Key as in Figure III.11.

3.4.5 *IGF2R* region

Three platypus BACs encompassing the large *IGF2R* gene were selected for sequencing. All three BACs have been finished generating a consensus sequence of 383,801 bp. Only one wallaby BAC (CU469286.1) from this region has been sequenced, spans 159,825 bp, and contains the entire *IGF2R* gene.

3.4.6 Other regions

Solitary BACs have been mapped and finished for wallaby and platypus orthologues of *GRB10* and *DNMT1* and finally single platypus BACs containing the orthologues *UBE3A* and *PLAGL1* have been sequenced to completion. In summary, a total of 10.8 Mb of finished sequence and a further 0.7 Mb of unfinished sequence from 5 amniote species and 9 different genomic regions has been obtained for subsequent analysis (Table III-5).

Table III-5. Species and regional sequence resources developed

Common name	Duck-billed platypus	Tammar wallaby	Grey short-tailed opossum	Wild mouse	Chicken	
Species name	*Ornithorhynchus anatinus*	*Macropus eugenii*	*Monodelphis domestica*	*Mus spretus*	*Gallus gallus*	Regional TOTALS (bp)
IC1-IC2	737,285 (328,330)	1,585,968	136,229 (344,833)	1,438,076	1,301,664	5,199,222 (673,163)
STX16-GNAS	490,161	450,517	ND	ND	ND	940,678
DLK1-DIO3	807,237	1,698,634	ND	ND	ND	2,505,871
SLC38A2-A4	331,739	319,152	ND	ND	ND	650,891
IGF2R	383,801	159,825	ND	ND	ND	543,626
UBE3A	171,880	ND	ND	ND	ND	171,880
GRB10	135,754	164,317	ND	ND	ND	300,071
PLAGL1	152,936	ND	ND	ND	ND	152,936
DNMT1	206,126	159,867	ND	ND	ND	365,993
Species TOTALS (bp)	3,416,919 (328,330)	4,538,280	136,229 (344,833)	1,438,076	1,301,664	10,831,168 (673,163)

Columns represent species sequenced, rows represent regions sequenced. Numbers in brackets correspond to unfinished sequence. ND, not done.

3.5 Discussion

This chapter has described the physical mapping and sequencing of distinct regions of conserved synteny in species occupying unique positions in vertebrate phylogeny. The informative species used in this study were selected to help address the questions: why, when and how did the phenomenon of genomic imprinting arise. The choice of regions to be studied was largely determined by local interest from groups at the Babraham Institute and Cambridge University which together with our group constitutes the SAVOIR consortium. Of the 17 clusters of imprinted genes known to exist in therian genomes, 8 are represented in this study.

A total of 10.8 Mb of high-quality sequence has been generated from across the regions and an additional 700 kb of sequence is in draft form with individual sequence contigs ordered and oriented where possible. Work is continuing by the Sanger Institute finishing team, led by Lucy Matthews, to bring these draft

sequences to finished form. Over 50% of the total sequence generated lies within the orthologous IC1 and IC2 regions, however, in addition to wallaby and platypus sequences, sequence was also generated for Western wild mouse, the South American opossum and chicken in this region alone.

Because of the paucity of markers available for marsupial and monotreme species novel marker generation was required. Since the exons of genes are highly evolutionarily constrained, to preserve their function, they offer a good starting point for marker development. In order to achieve contiguity of BAC clones across a region of interest, with minimal requirement for chromosome walking to close gaps, a marker density approaching 1 per 67 kb is required (Mungall and Humphray. 2003). For gene-rich regions of a genome this target marker density may be achievable, however, in gene poor regions the density of markers derived from gene sequences is likely to be inadequate. In the IC2 domain this is true for the *KCNQ1* gene which spans 404 kb in human (hg18 chr11:2422797-2826915) with distances between exons of up to 107 kb. As a consequence novel markers were required and were initially derived from the end-sequences of BAC clones mapping to the regions of interest. Iterative rounds of clone walking were then performed until map contiguity was achieved. With the exception of 5 gaps in the platypus IC1-IC2 region and a single gap in the tammar wallaby *GNAS* complex region (discussed above) all other regions in all other species were contiguated. The BACs in these contigs were the substrate for sequencing and subsequent analyses.

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

4.1 Introduction

4.1.1 Aims of this chapter

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

4.2 Sequence assemblies

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

4.2.1 Assembly of BAC sequences

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

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

```
##species=Wallaby chromosome=2
CU467493    MEKBa-205I8      Wallaby_2ctg192
CU458744    MEKBa-283A6      Wallaby_2ctg192
CR855994    MEKBa-69G12      Wallaby_2ctg192
CT008508    MEKBa-459D3      Wallaby_2ctg192
CR925759    MEKBa-346C2      Wallaby_2ctg192
CR848708    MEKBa-201B9      Wallaby_2ctg192
CU024874    MEKBa-517H11     Wallaby_2ctg192
CU024865    MEKBa-439O2      Wallaby_2ctg192
CU311200    MEKBa-183M18     Wallaby_2ctg192
CU041371    MEKBa-363O3      Wallaby_2ctg192
CU062506    MEKBa-183I7      Wallaby_2ctg192
CT990571    MEKBa-465N20     Wallaby_2ctg192
```

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

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

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

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

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

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

4.2.2 Comparison with whole genome shotgun sequence assemblies

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

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

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

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

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

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

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

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

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

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

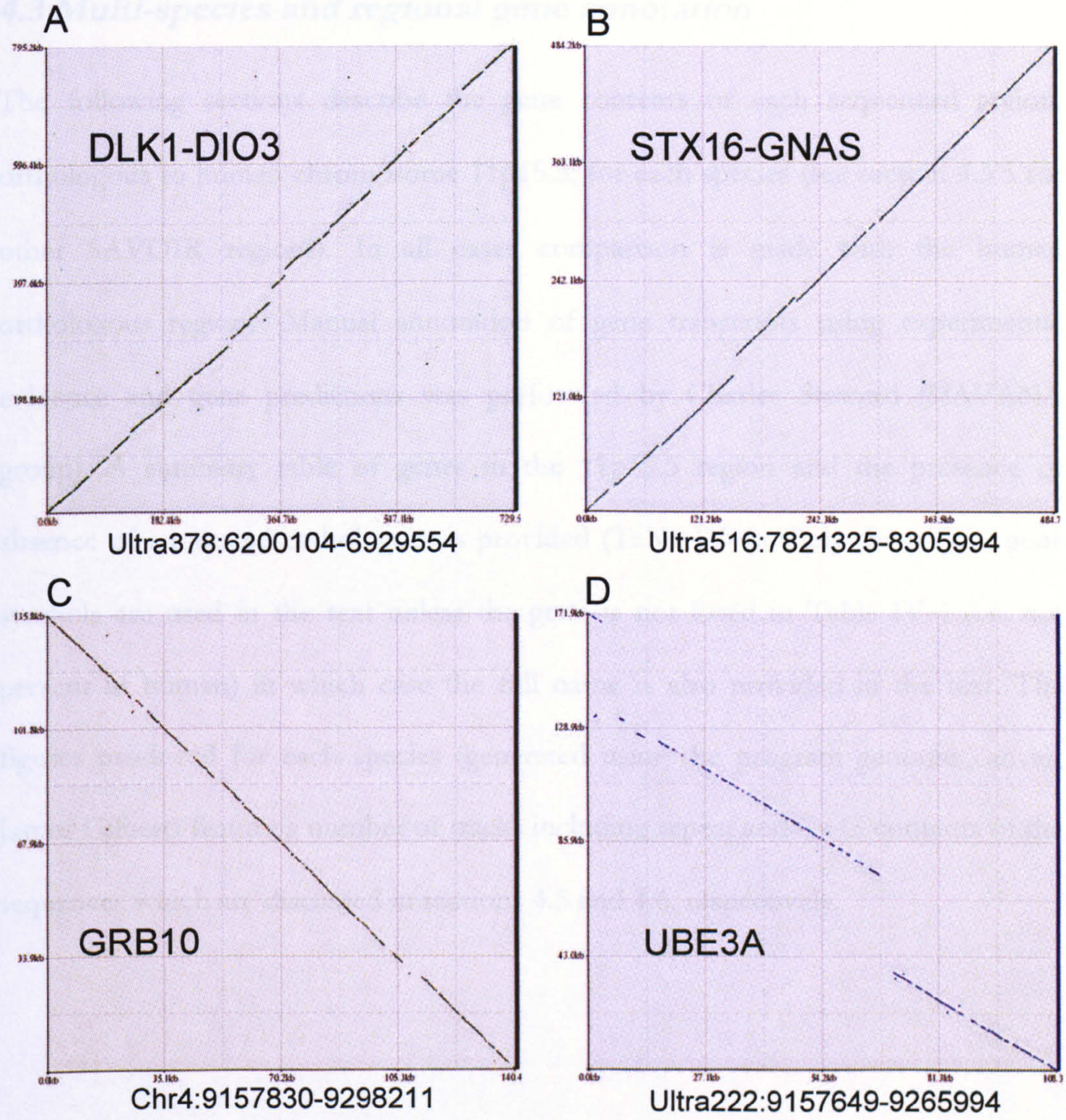


Figure IV.1. Dot-plots comparing platypus WGS contigs with finished sequences.

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

4.3 Multi-species and regional gene annotation

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

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

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

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

4.3.1 Chicken (*Gallus gallus*)

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

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

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

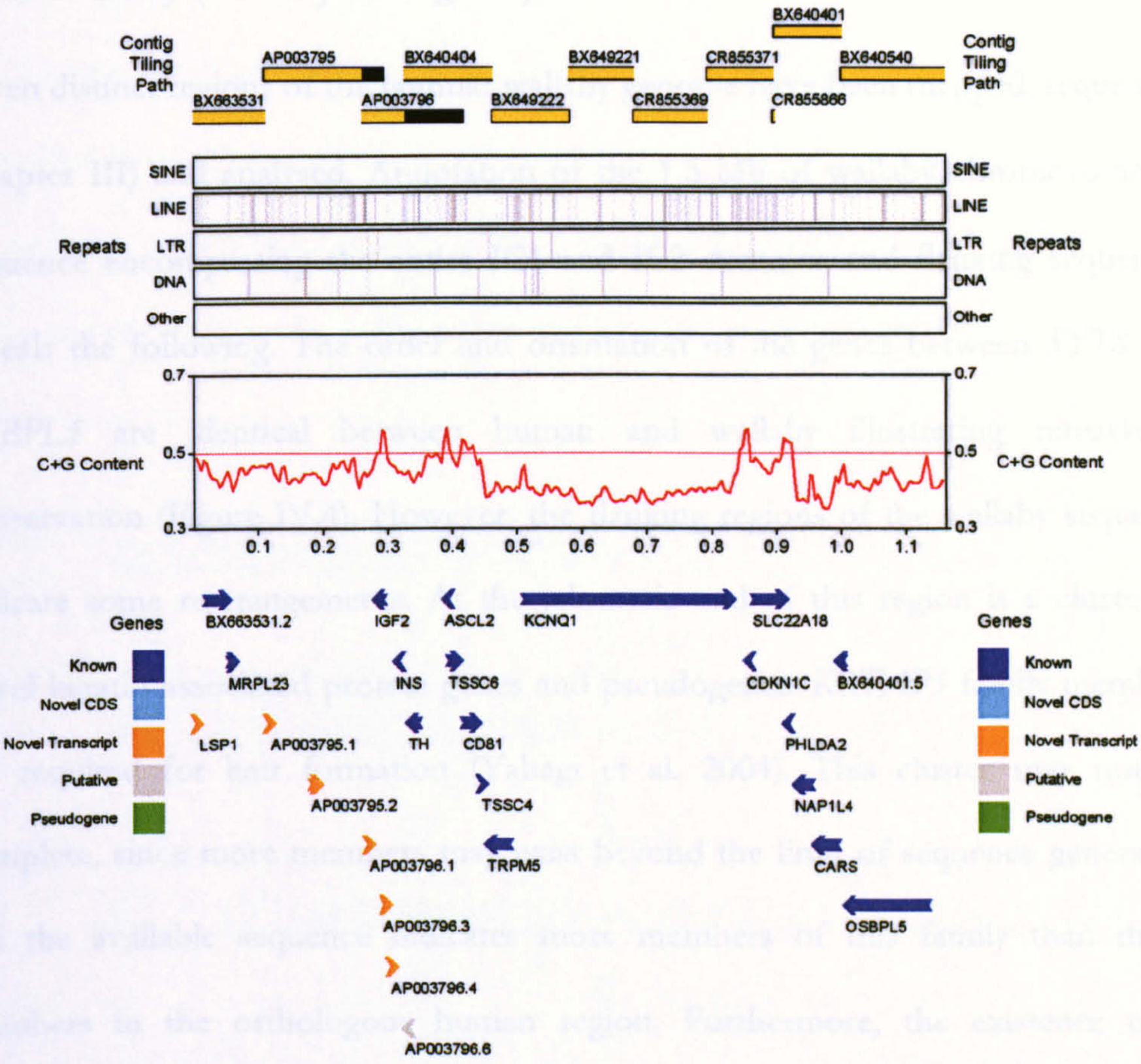


Figure IV.2. Sequence analysis and annotation of chicken chromosome 5.

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

4.3.2 Wallaby (*Macropus eugenii*)

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

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

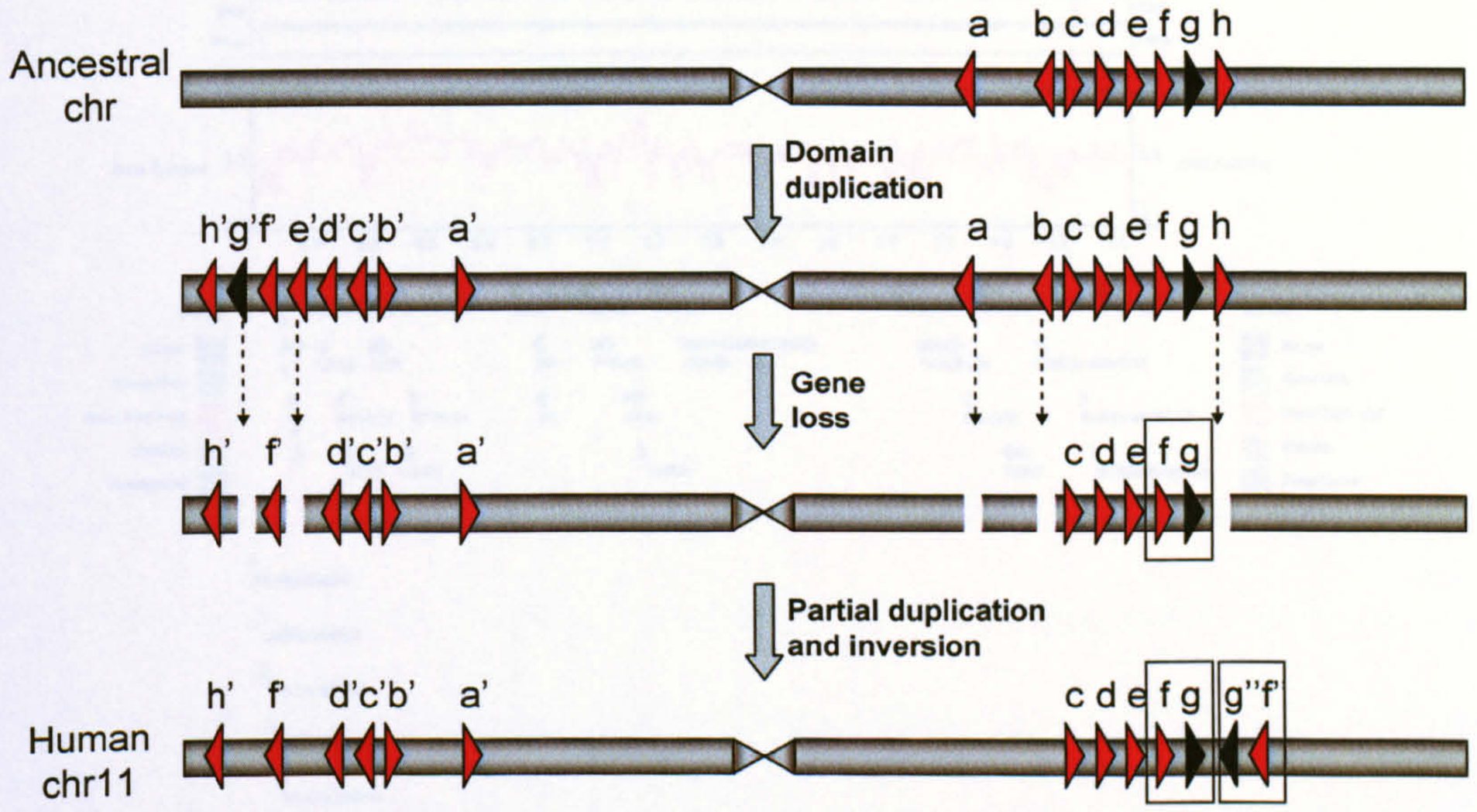


Figure IV.3. Postulated model for the evolution of *KRTAP5* family members.

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

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

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

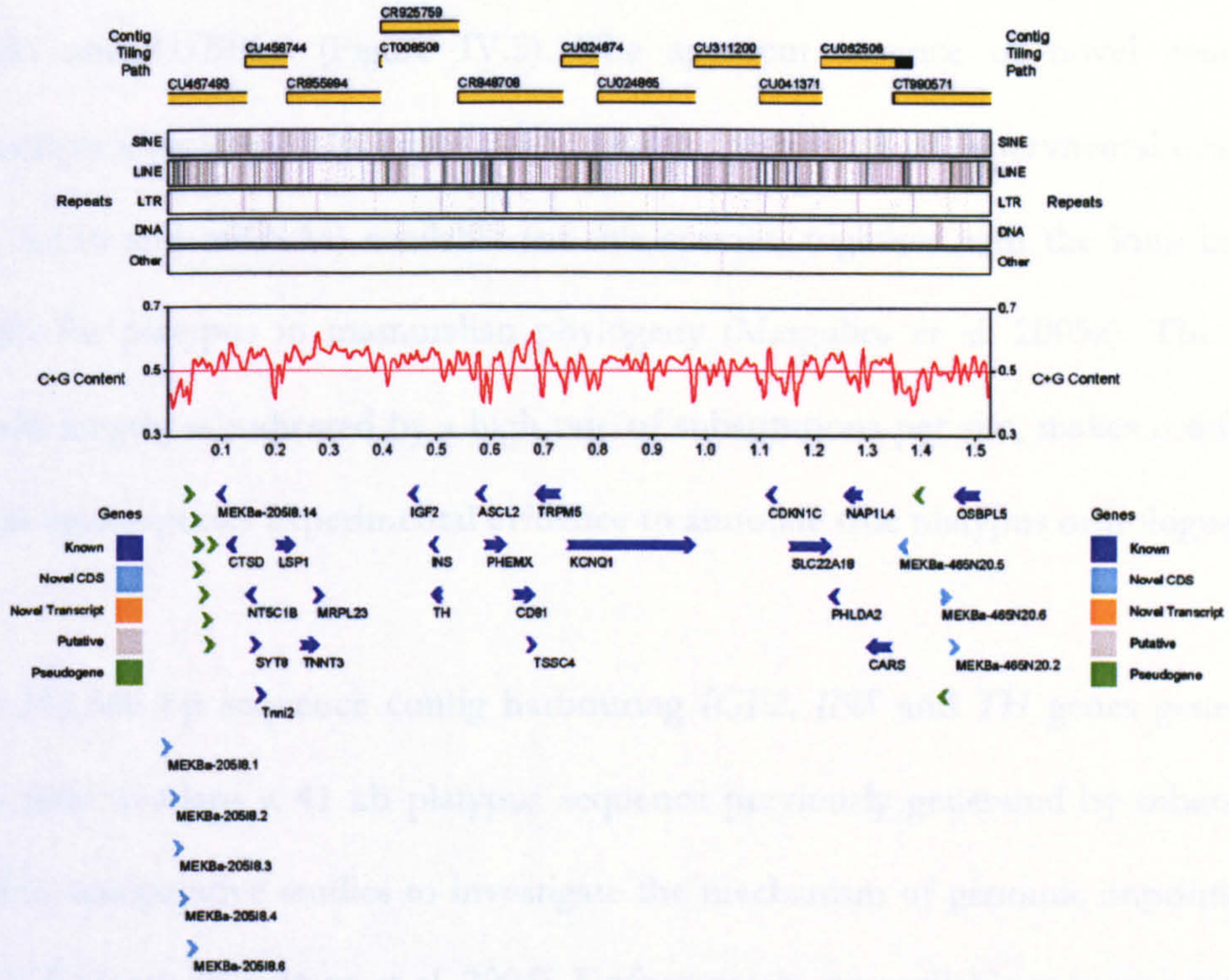


Figure IV.4. Sequence analysis and annotation of wallaby chromosome 2p.

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

4.3.3 Platypus (*Ornithorhynchus anatinus*)

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

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

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

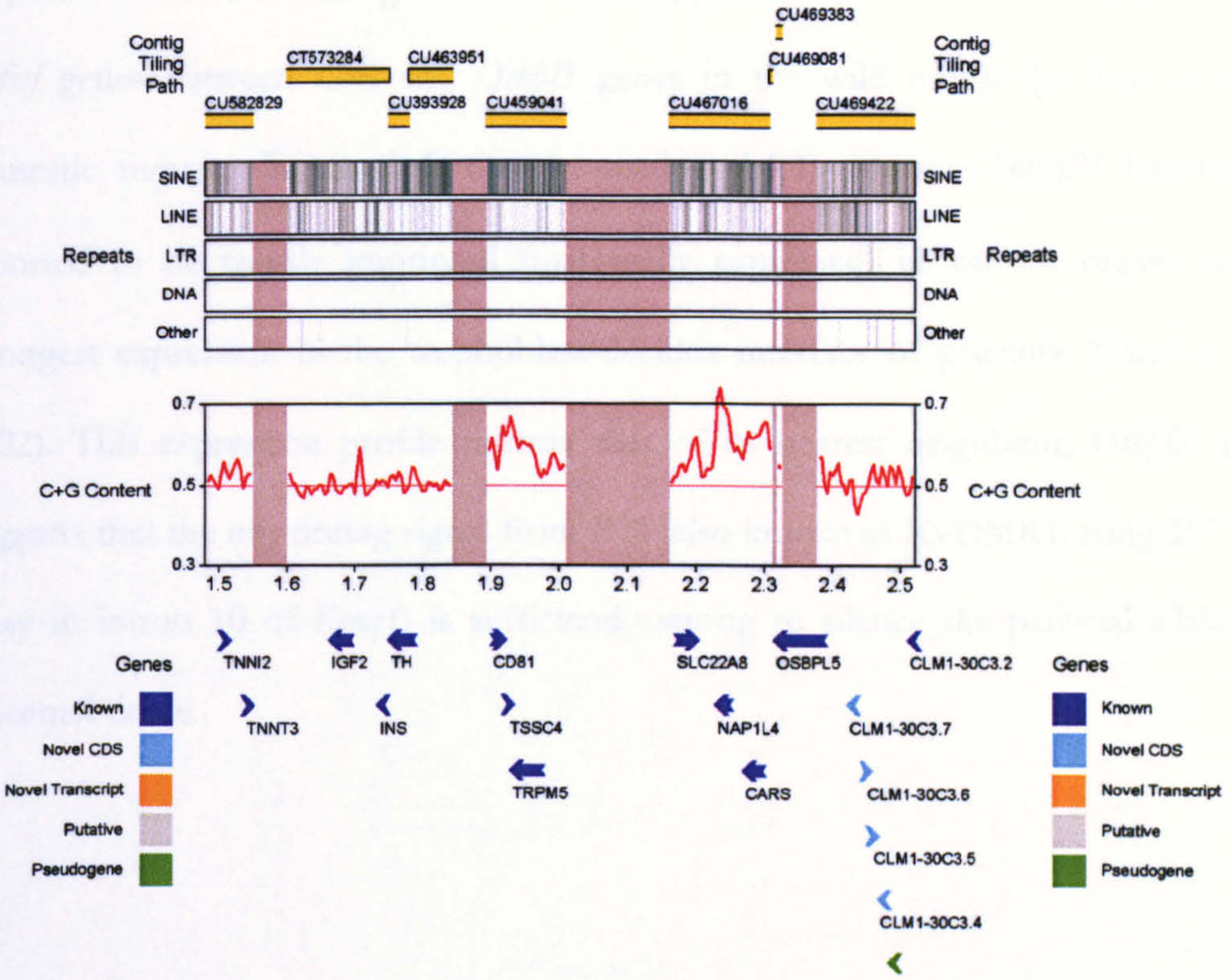


Figure IV.5. Sequence analysis and annotation of platypus chromosome 8p orthologous to human 11p15.5.

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

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

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

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

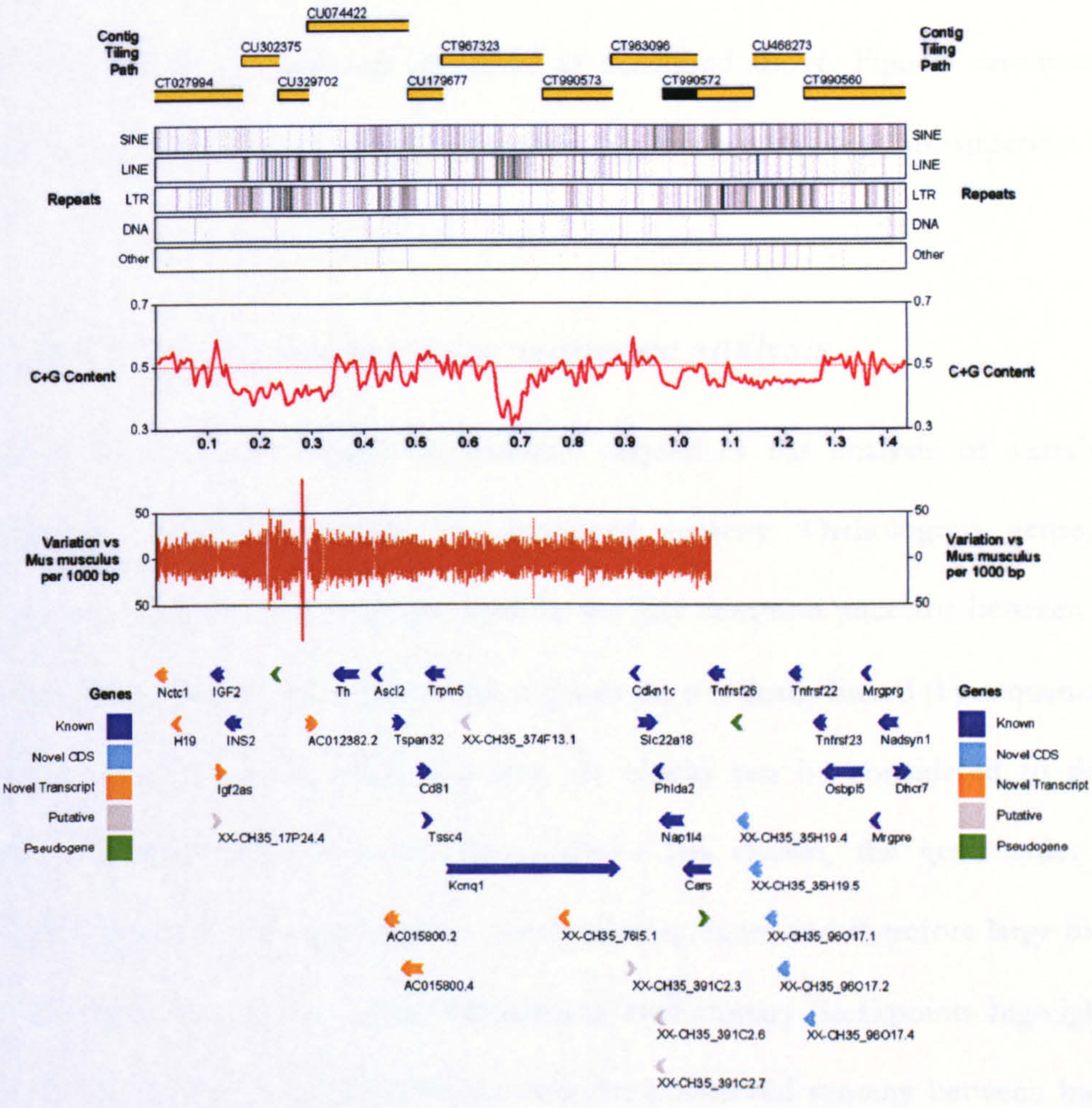


Figure IV.6. Sequence analysis and annotation of *Mus spretus* distal chromosome 7.

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

4.3.5 Analysis and annotation of other SAVOIR regions

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

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

4.4 Multi-species comparative sequence analysis

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

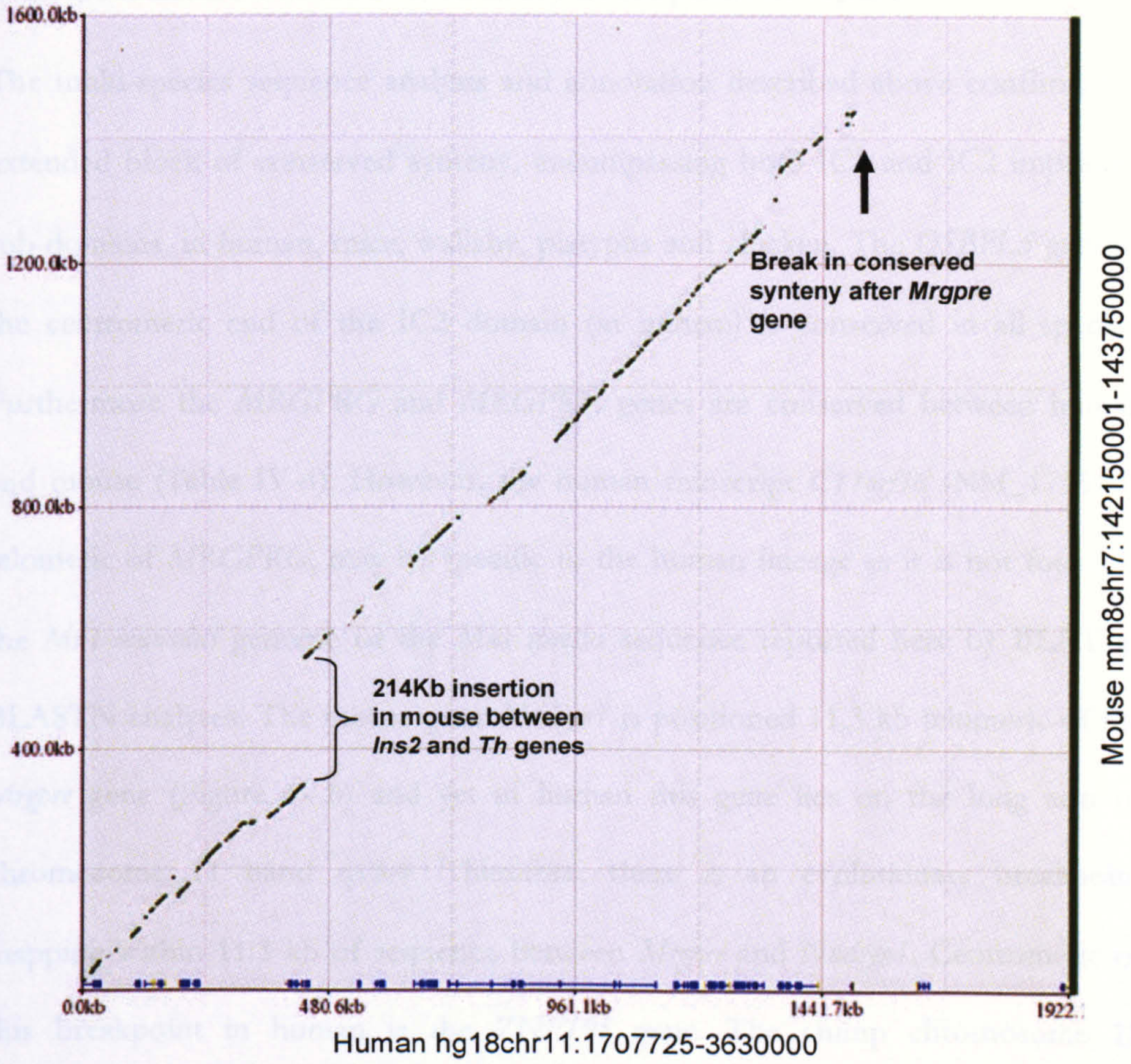


Figure IV.7. An extended block of conserved synteny between human chromosome 11p15.5 and mouse chromosome 7qF5.

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

4.4.1 Localisation of an evolutionary breakpoint at 11p15.5

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

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

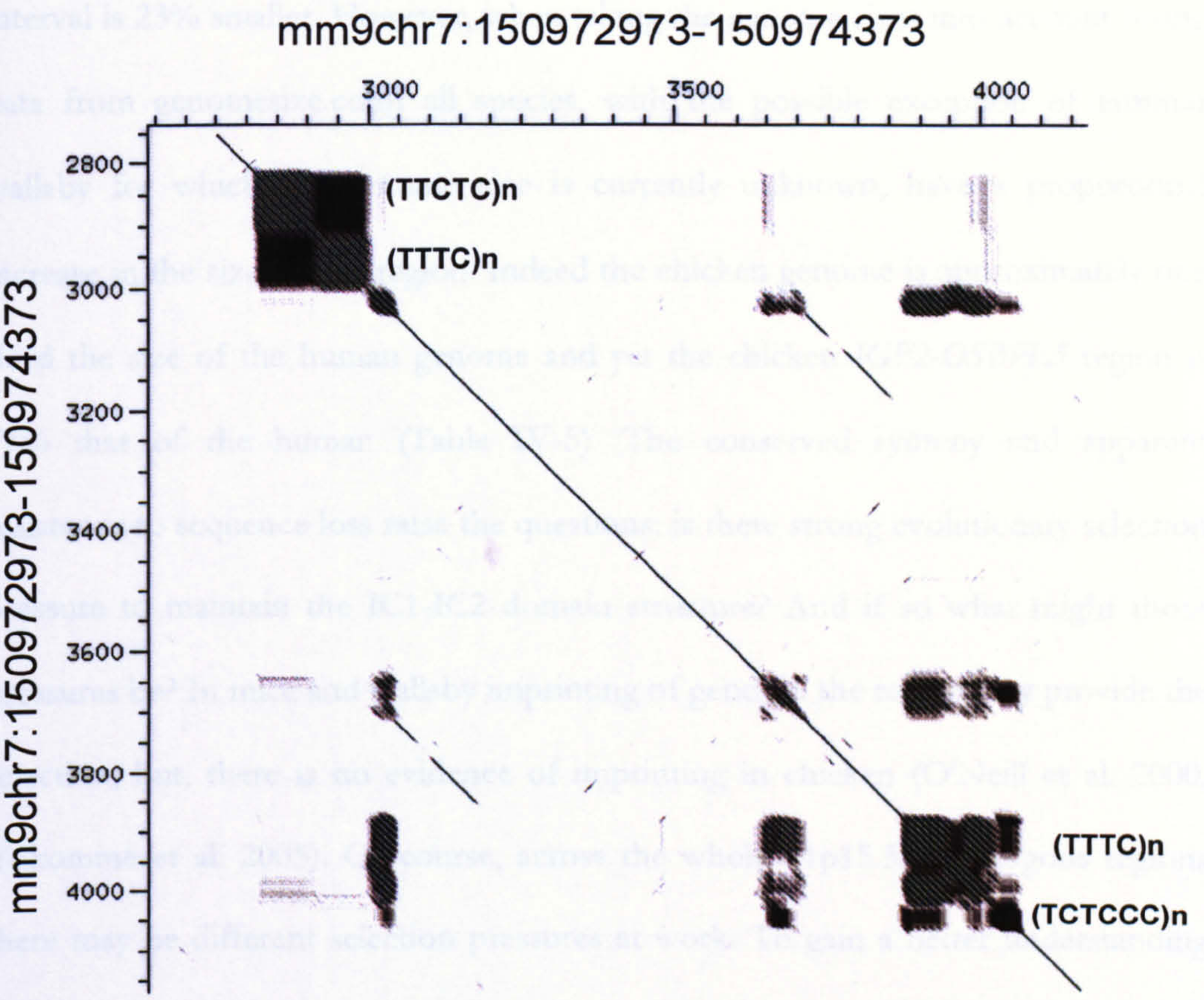


Figure IV.8. *Mus musculus* self–self dot-plot in the distal chromosome 7 evolutionary breakpoint region.

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

4.4.2 Broad scale finished sequence comparisons

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

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

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

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

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

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

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

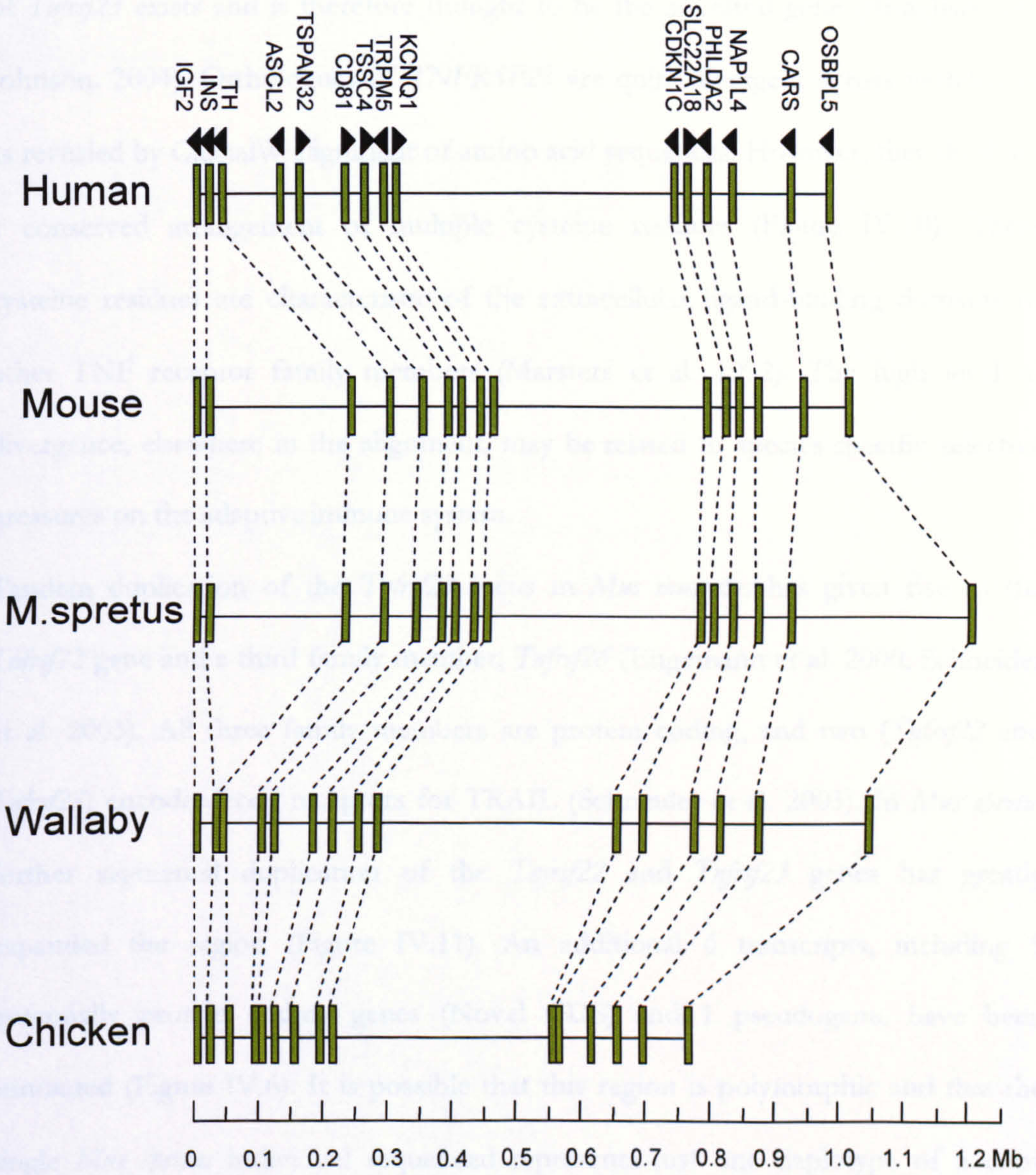


Figure IV.9. Comparison of the genomic structures between *IGF2* and *OSBPL5* genes.

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

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

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

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

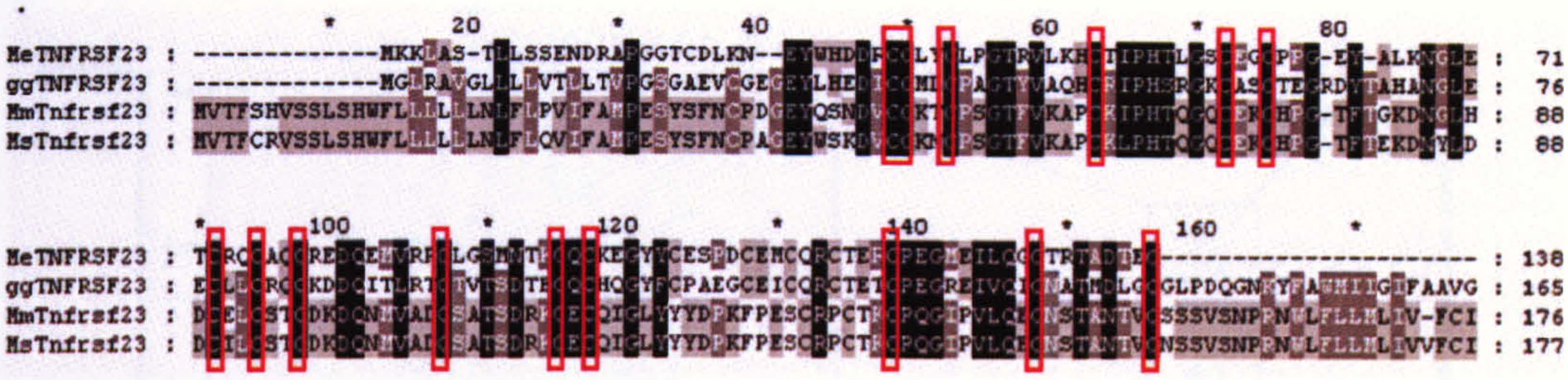


Figure IV.10. Amino acid sequence alignment of TNFRSF23 orthologues.

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

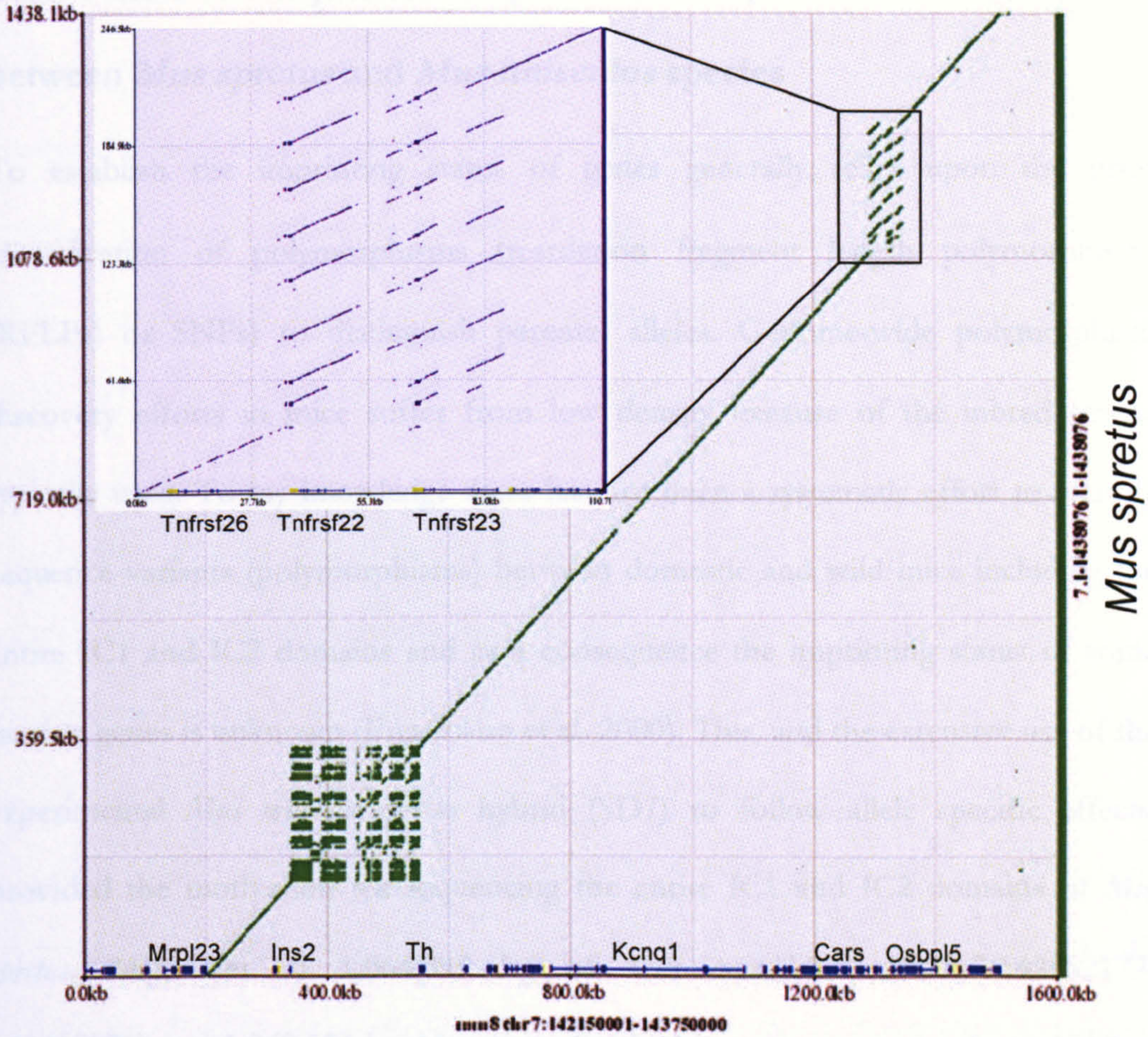


Figure IV.11. Dot-plot of *Mus musculus* and *Mus spretus* sequences in the IC1 and IC2 domains. The green diagonal line represents the extent of pairwise sequence alignment (i.e. one-to-one orthology between the two species). Both sequences were masked for interspersed repeats but not tandem repeats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

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

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

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

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

4.5 Repeat contents of sequences

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

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

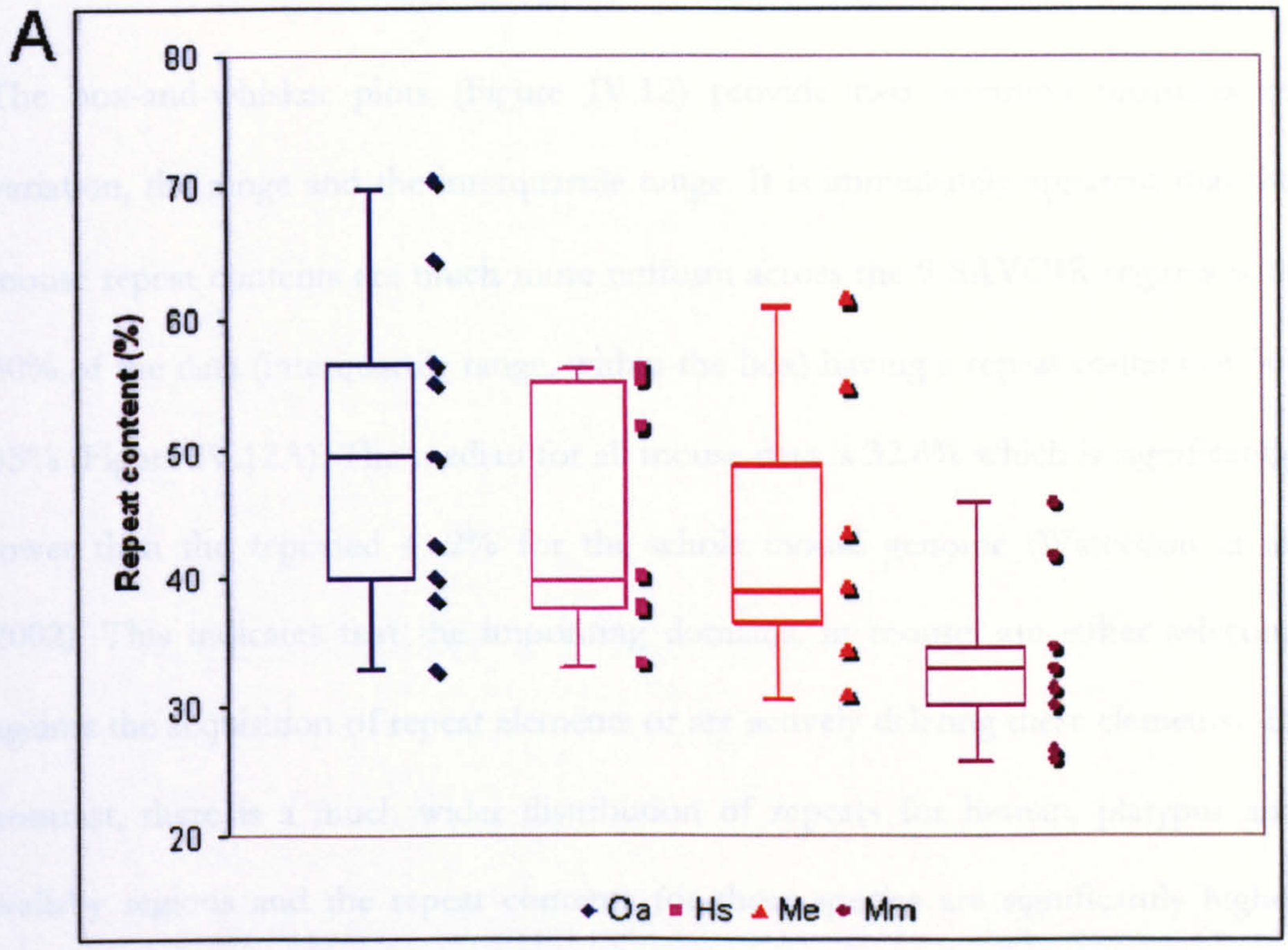
Table IV-6. Repeat and C+G contents of multi-species sequences generated here.

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

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

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

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



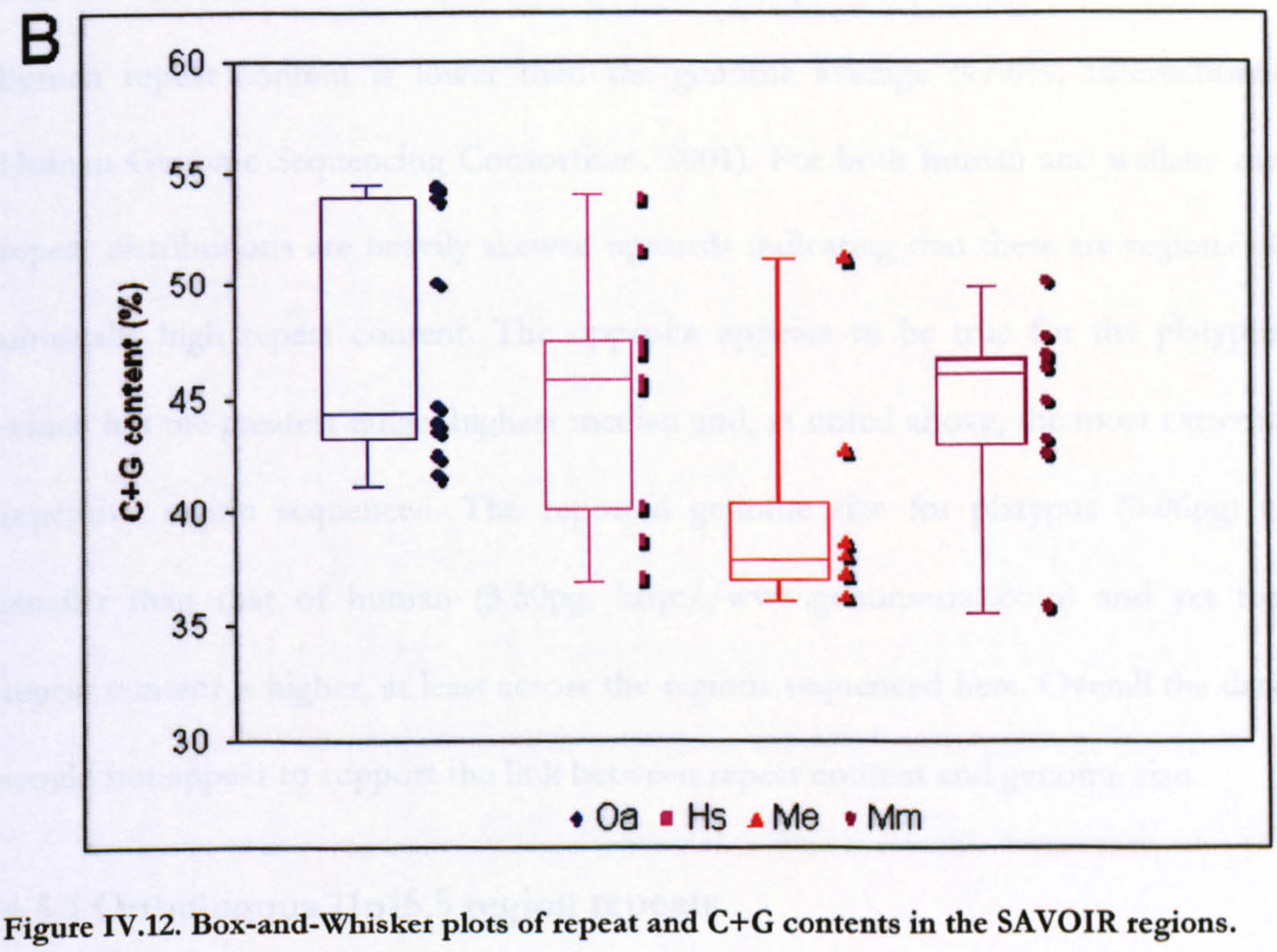


Figure IV.12. Box-and-Whisker plots of repeat and C+G contents in the SAVOIR regions.

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

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

4.5.1 Orthologous 11p15.5 region repeats

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

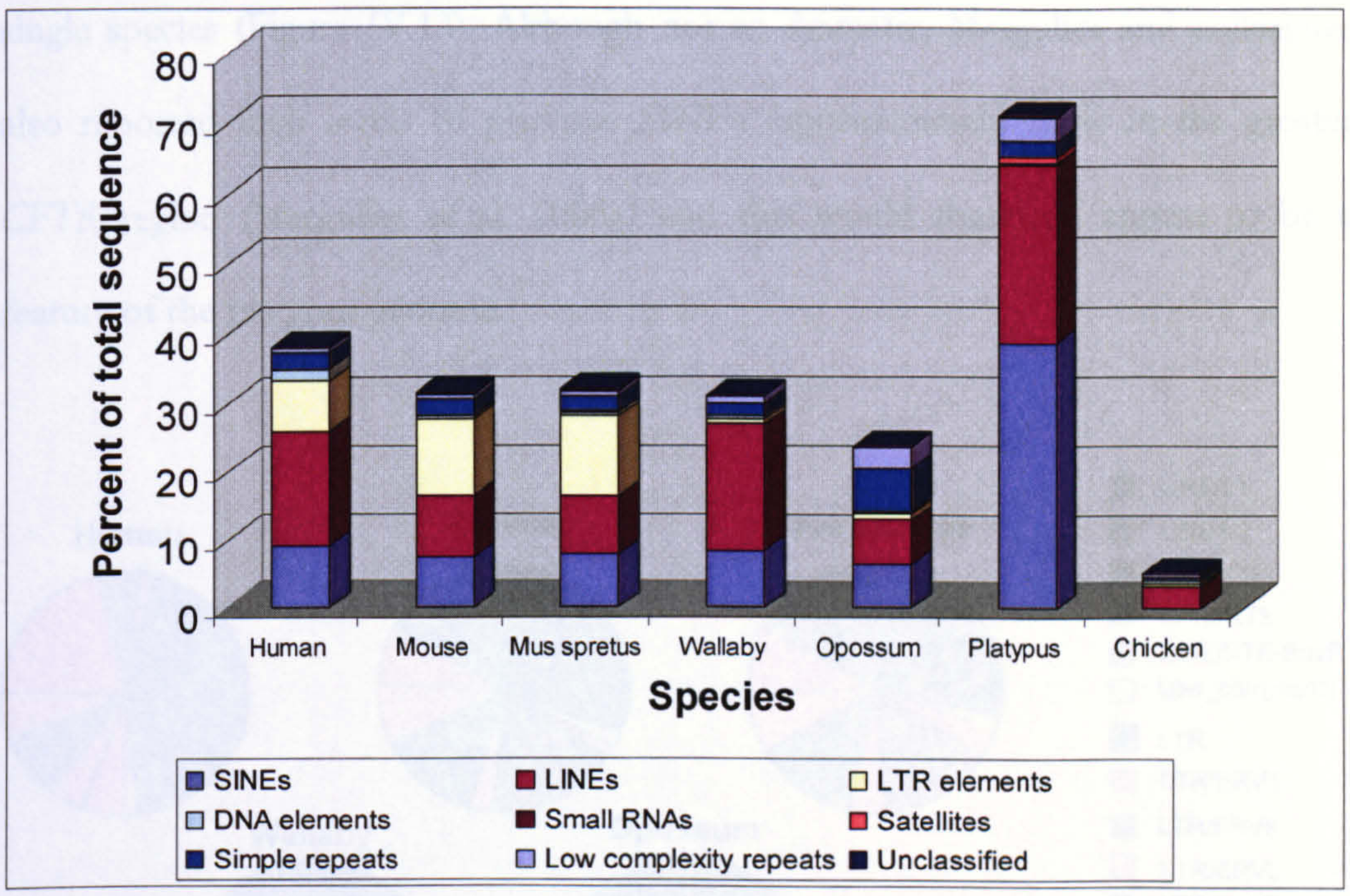


Figure IV.13. Relative content of repeat types within the 11p15.5 orthologous regions.

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

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

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

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

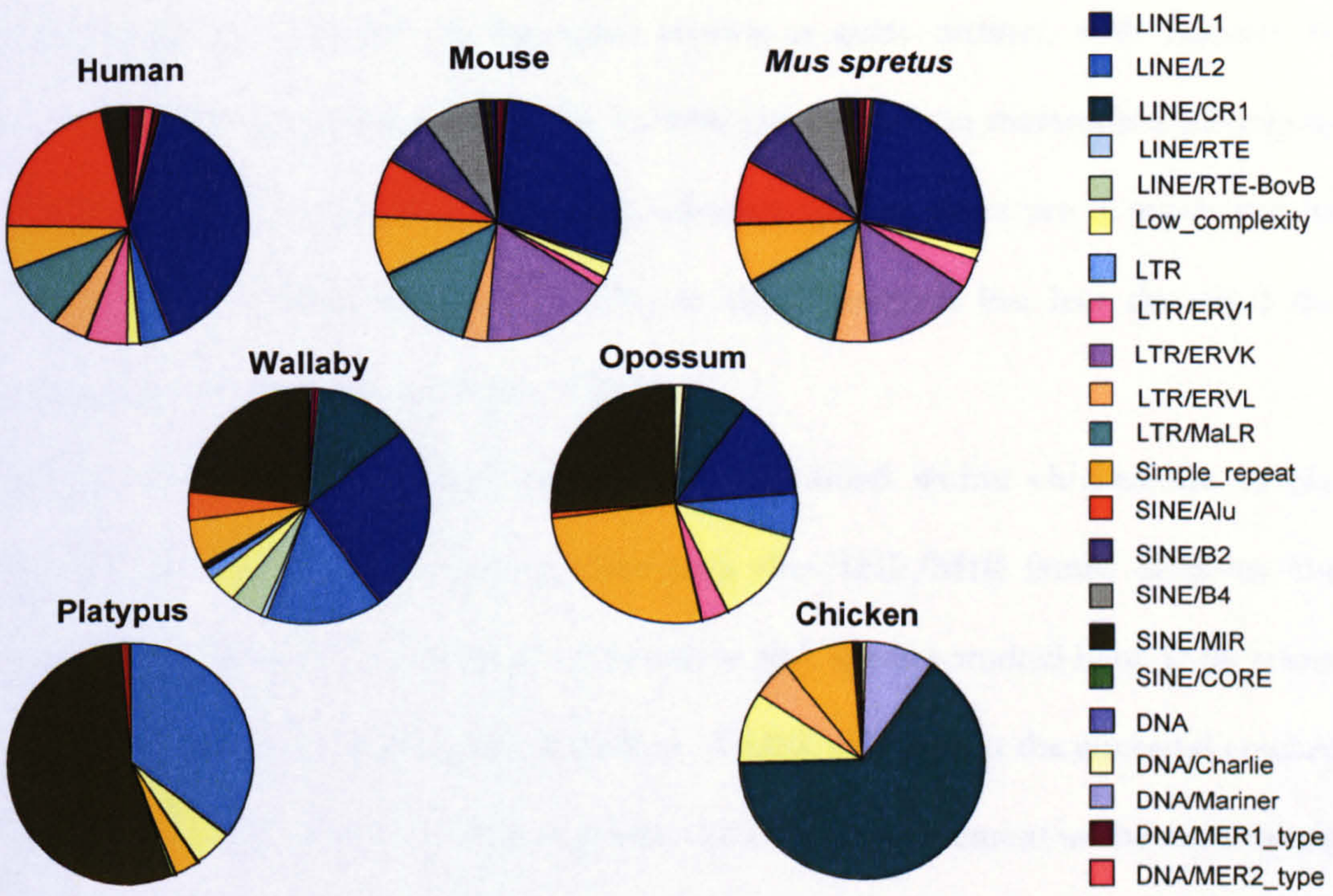


Figure IV.14. Repeat composition of multi-species sequences in the orthologous 11p15.5 region.

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

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

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

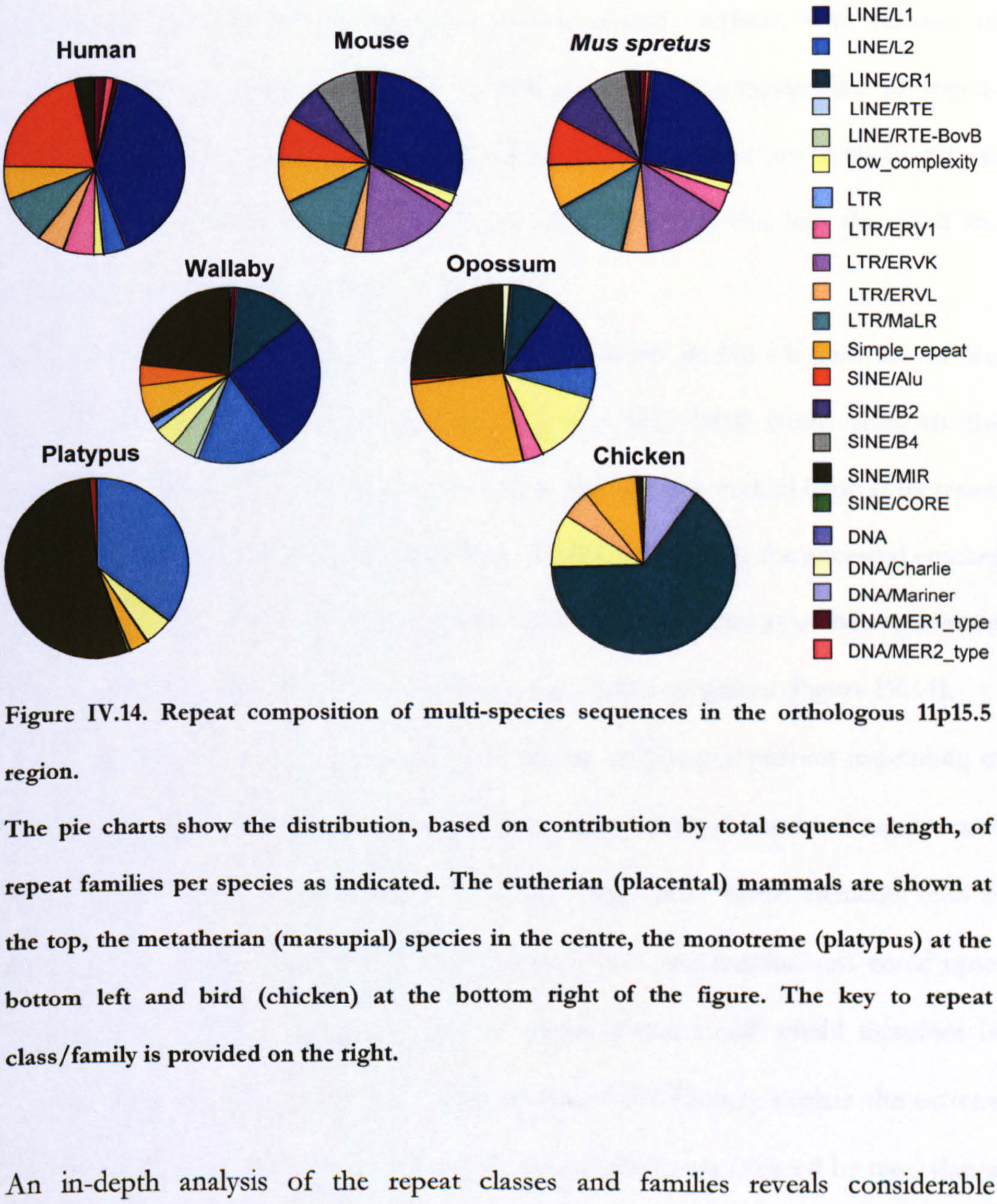


Figure IV.14. Repeat composition of multi-species sequences in the orthologous 11p15.5 region.

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

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

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

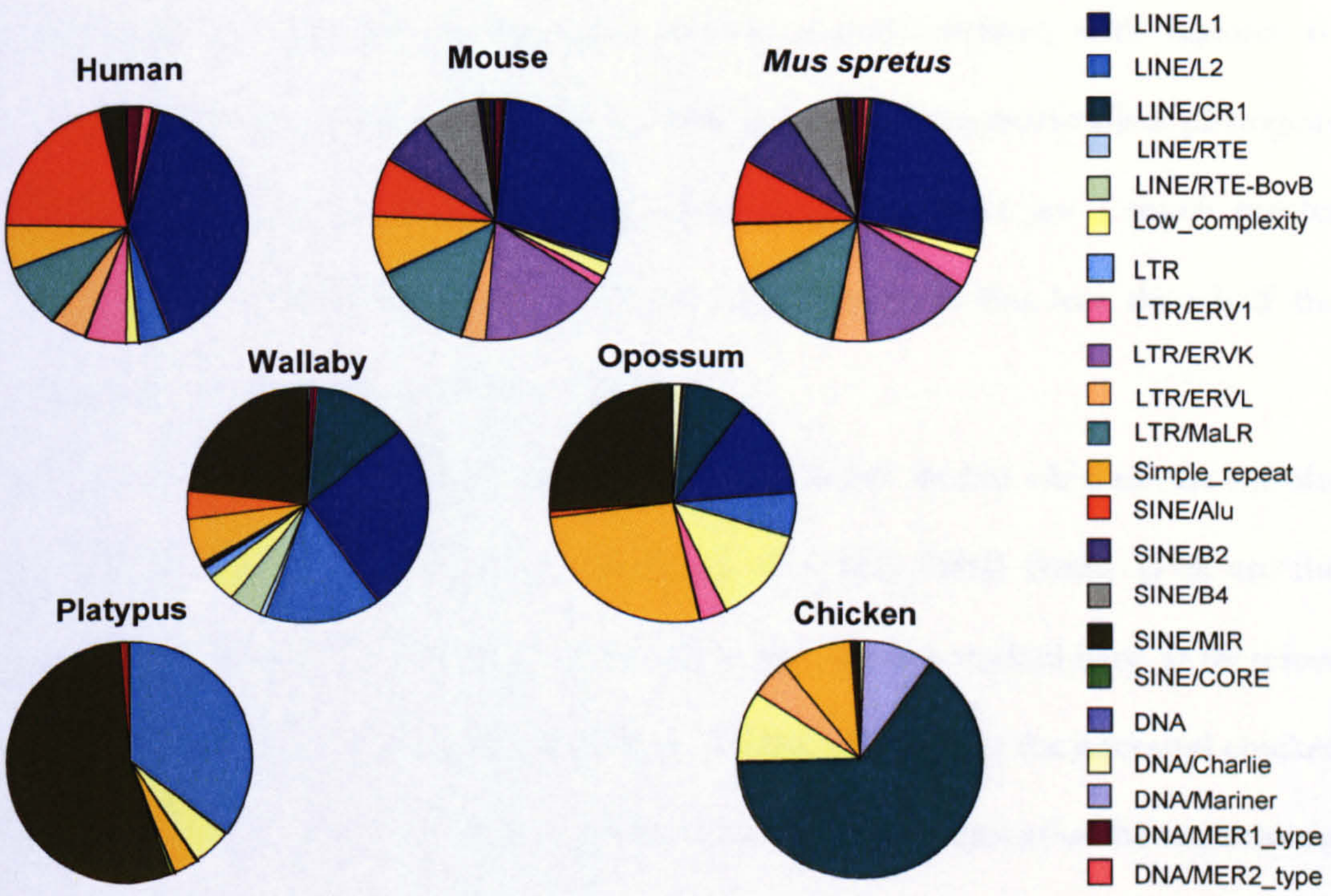


Figure IV.14. Repeat composition of multi-species sequences in the orthologous 11p15.5 region.

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

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

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

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

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

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

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

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

4.6 C+G content and CpG islands

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

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

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

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

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

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

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

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

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

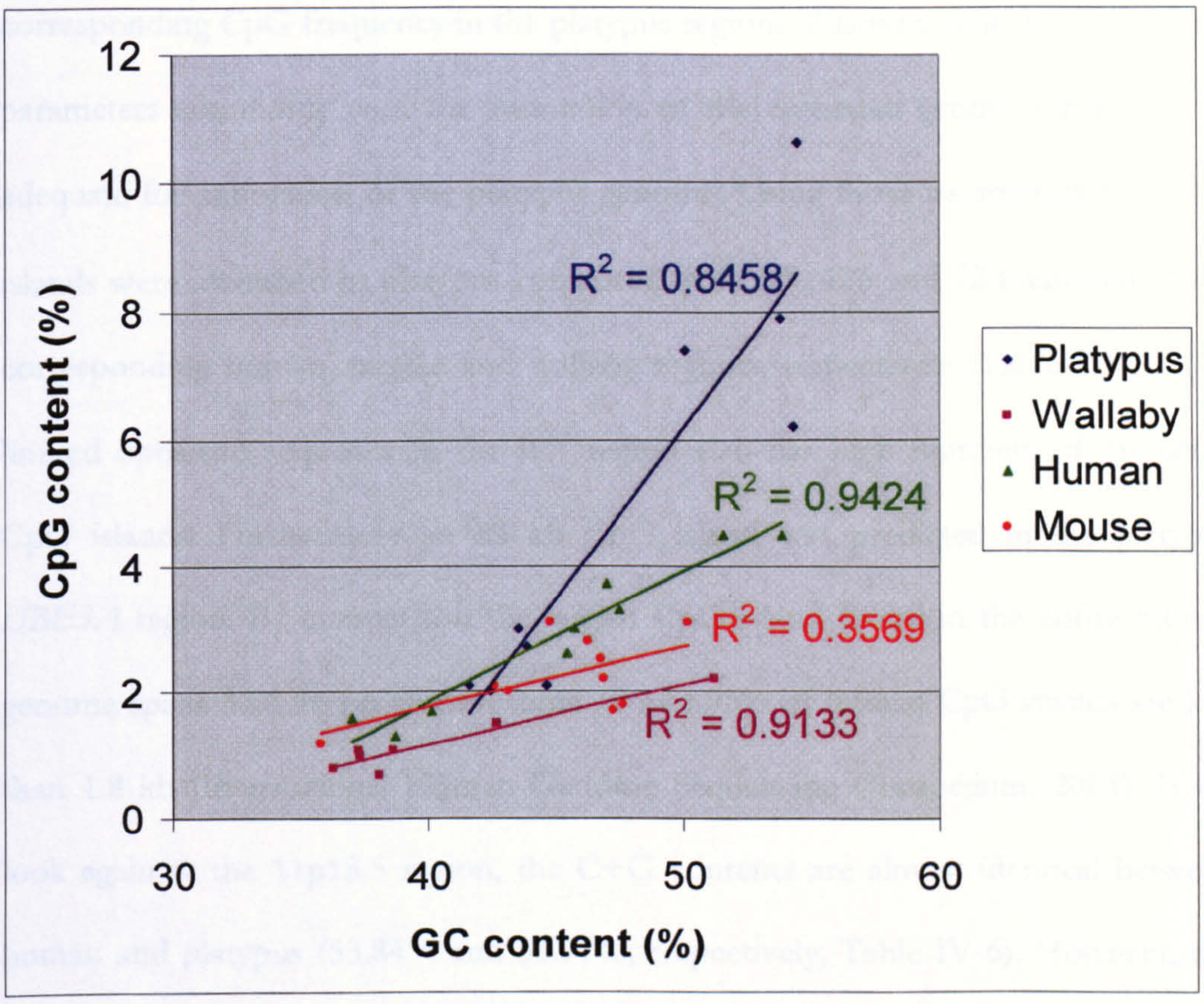


Figure IV.15. Plot of CpG and C+G contents for multi-species regional sequences.

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

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

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

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

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

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

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

4.7 SAVOIR consortium website

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

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

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

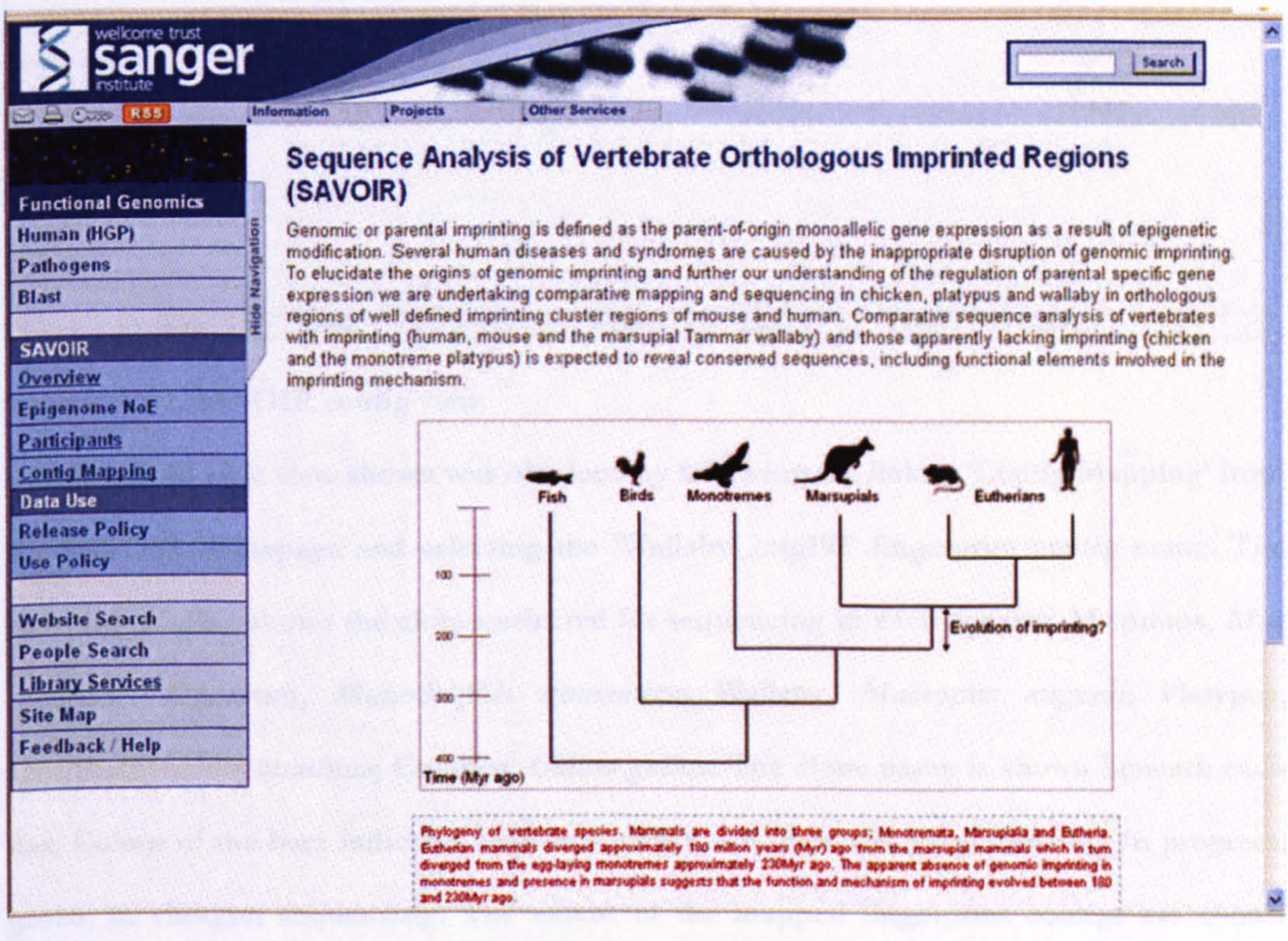


Figure IV.16. The SAVOIR website (http://www.sanger.ac.uk/PostGenomics/epicomp).

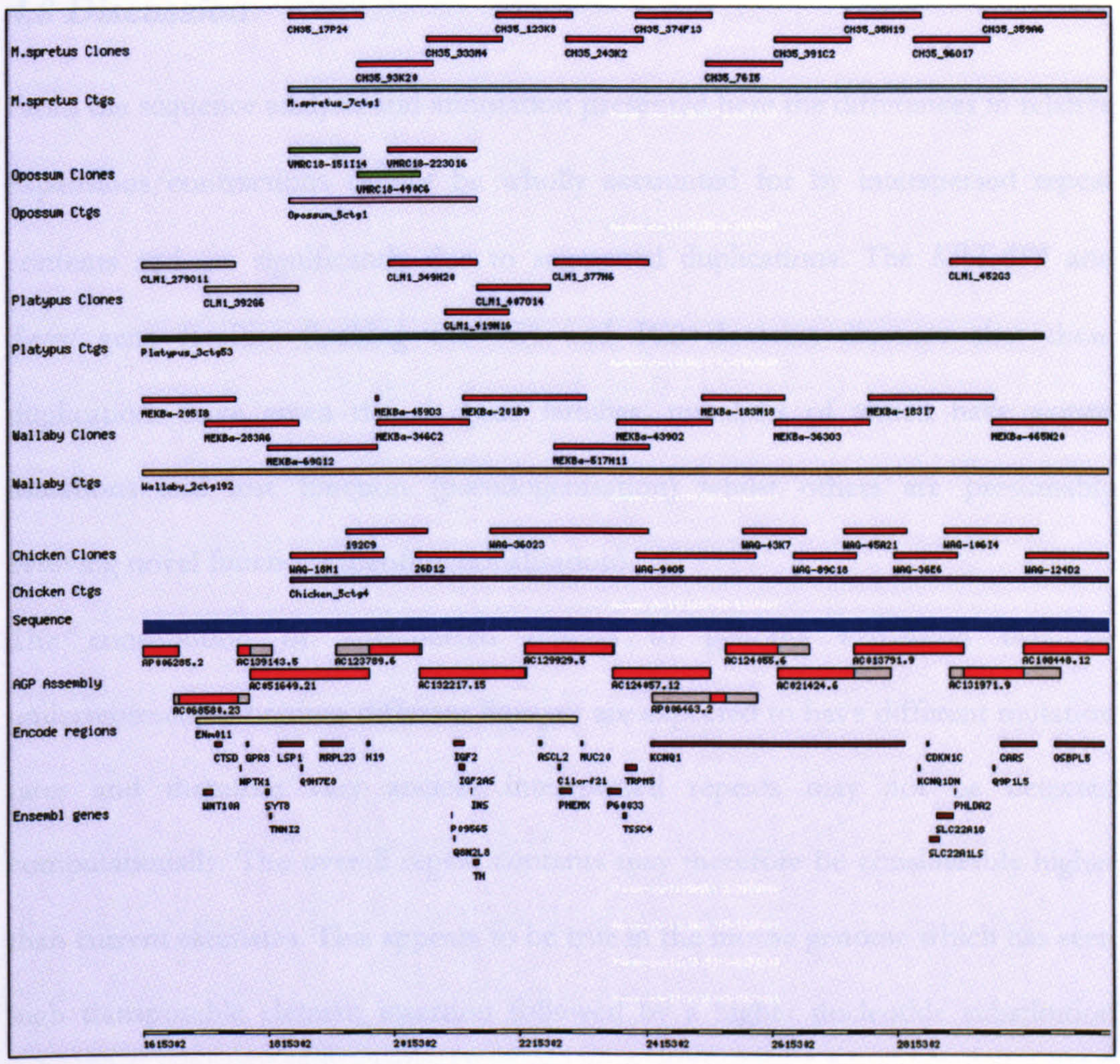


Figure IV.17. SAVOIR contig view.

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

4.8 Discussion

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

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

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

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

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

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

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

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

Chapter V - Establishing function of the non-coding evolutionary conserved regions

5.1 Introduction

5.1.1 Aims of this chapter

The aim of this chapter is to identify function for evolutionary conserved regions (ECRs) identified from the 11p15.5 region. The chapter describes the methods used to identify ECRs from alignments of the vertebrate sequences with conserved synteny to human 11p15.5 generated in chapter III. The process of cloning the ECRs and suitable controls follows. This begins with PCR amplification of ECR and control fragments from human, mouse and wallaby genomic DNA and is followed by the generation, and quality control, of a library of clones created by recombination cloning.

The utility of ECRs in finding previously un-annotated transcripts or alternative exons in the human genome by comparing their sequences to current genome annotation within the genome browsers is also described. The demonstration that non-coding ECRs exhibit enhancer activity is shown using dual luciferase reporter assays following transient transfection of the cloned ECRs into human HepG2 (liver) cells. To address the question "can cross-species enhancer activity be detected in human HepG2 cells?" a sub-set of the ECRs with demonstrated human enhancer activities were cloned from wallaby genomic DNA. The detailed analysis of a novel and highly conserved endodermal enhancer is then described. Finally, the generation of a PCR tiling array from across the 11p15.5 region is described which was included in ENCODE ChIP-chip experiments to study histone modification

profiles and insulator protein DNA binding sites. Correlation of these experimental data and other publicly available datasets with the ECRs is performed in an effort to assign function to the ECRs and better understand gene regulation in the region.

There are those who believe that observed ECRs are not functionally constrained sequences but simply relics of ancestral sequences that have not yet mutated and diverged between the species compared. In cases in which two genomes are evolutionarily relatively close (e.g. divergence between human and mouse, approximately 90 Myr ago) this will be true. However, if comparing say, human and fugu genomes which have diverged more than 400 Myr ago then false positive identification of functional elements is surely negligible. What then of intermediate comparisons between human and wallaby? How many ECRs (100 bp in length with 70% identity) would you expect to see by chance after 148 Myr? Sean Eddy has modelled the statistical power of comparative sequence analysis and provides equations for generating the probability that we erroneously infer that a neutral feature is conserved (false positives, Figure V.1) or false negatives in which conserved features are inferred as neutral (Eddy. 2005). If we substitute values of N=2 (number of genomes compared), L=100 (length of ECR (bp)), D=0.537 (distance between human and wallaby from Margulies et al. 2005a) and C=30 (number of mismatches, 30%) the probability of false positive detection is 1.074497e-14. Therefore, if Eddy's model is accurate the probability of identifying an unconstrained ECR between human and wallaby is exceedingly low and these sequences warrant further investigation.

$$FP = P(\leq C \text{ changes}|\text{neutral}) = \sum_{c=0}^{C}\left[\binom{NL}{c}\left(\frac{3}{4}-\frac{3}{4}e^{-\frac{4D}{3}}\right)^{c}\left(\frac{1}{4}+\frac{3}{4}e^{-\frac{4D}{3}}\right)^{NL-c}\right] \quad (1)$$

Figure V.1. Probability of erroneously inferring that a neutral feature is conserved.

Equation taken from (Eddy. 2005). **FP, false positive; P, probability; D, substitution events; N, genome number; L, length of feature, C, number of mismatches.**

5.1.2 Computational tools for identifying candidate regulatory elements

As discussed in chapter I, comparative sequence analysis is a powerful tool with which to annotate genome sequences. Many of the tools for evolutionary comparisons, sequence alignments (global and local) and detection of functional sequence patterns are accessible from internet servers (reviewed in Frazer et al. 2003 and chapter I). Here I elected to use the local alignment program BLASTZ (Schwartz et al. 2003b) because this program can compute alignments for sequences of any length assuming that the sequences compared share blocks of high conservation, separated by regions lacking homology. The multi-species sequences generated in chapter III satisfy these criteria since they are of known orthology but are highly divergent (e.g. human-chicken comparisons).

A second factor to be considered when choosing algorithms for comparative sequence analysis is the ease of which potentially functional regions are visualised. Some of these algorithms integrate motif finding and dynamic visualisation programs, thus providing practical tools for the analysis of regulatory elements within multi-species conserved sequences. The VISTA portal (http://genome.lbl.gov/vista/index.shtml) was recently described and provides tools to assist in the identification and characterisation of regulatory elements

(Brudno et al. 2007). However, I elected to use the zPicture server (Ovcharenko et al. 2004a) because it incorporates both BLASTZ alignments with highly dynamic visualisation tools and links to the regulatory VISTA (rVISTA) server for conserved TFBS motif discovery (Loots and Ovcharenko. 2004). These tools and many others can be found at http://dcode.org and have been described by Loots and Ovcharenko (Loots and Ovcharenko. 2005).

Many different names and acronyms have been associated with ECRs (reviewed in Aloni and Lancet. 2005) but here I refer to them as ECRs as defined in the zPicture server where they were originally identified. It is of great interest to elucidate the function (if any) of ECRs, which likely includes exonic sequences not previously annotated and elements controlling gene regulation (discussed in chapter I). With recent advances in technology, many of these functions can be experimentally tested in medium to high-throughput.

5.1.3 Assessing function of ECRs

Testing sequences for function can be performed *in vivo* or *in vitro* and each has their merits. Mouse transgenic assays are being used to test for enhancer elements capable of recapitulating spatial and temporal patterns of gene expression. PCR amplified putative functional elements are cloned into a reporter vector containing a heat shock protein 68 promoter and ß-galactosidase reporter gene. Reporter expression in the absence of an enhancer is negligible but in the presence of an enhancer both spatial and temporal gene expression patterns can be characterised. The test construct is injected into fertilised mouse oocytes where random integration into the genomic DNA occurs. Oocytes are then implanted into pseudo-pregnant females and the embryos harvested and stained for ß-galactosidase activity (Pennacchio et al. 2006). Being mammals, mice are an appropriate model in which

to study human candidate functional elements but these experiments are time consuming and expensive. In a dedicated facility approximately 500 elements per year can currently be characterised (Visel et al. 2007). Still higher-throughput *in vivo* transgenic reporter assays have been devised for testing putative enhancer elements in zebrafish (Woolfe et al. 2005) and *Xenopus* (Gottgens et al. 2000). These assays have been used to demonstrate enhancer function of highly conserved vertebrate sequences regulating key developmental genes with evolutionary conserved function and expression patterns. However, these techniques may not detect mammalian-specific function.

In vitro methods for assaying function of DNA elements include DNaseI hypersensitive site (HS) mapping, electrophoretic gel shift assays, Chromatin immunoprecipitation (ChIP) and gene reporter assays. Unlike *in vivo* assays these technologies ideally require prior knowledge of the cell type in which the putative enhancer is active, or a suitably wide survey of cell types. However, they are readily scaleable, cost-effective and can be performed in most molecular biology laboratories. As such, strategies can be devised whereby potentially high-throughput screening of candidate functional elements can be tested *in vitro* to identify elements for subsequent *in vivo* characterisation. In this thesis I have chosen to use gene reporter assays to screen identified ECRs for enhancer activities.

5.1.3.1 Recombinational cloning

When testing multiple ECRs for function in, for example, gene reporter assays the ECRs (and controls) first need to be cloned. Classical restriction endonuclease enzyme cloning methods can be laborious and are not easily scaleable due to the variability of restriction sites within the sequences to be cloned. A method providing a fast and efficient way to move DNA sequences between multiple vector systems is

required. Gateway® technology (Invitrogen) is a universal cloning system that uses site specific recombination (Landy. 1989) to facilitate integration of bacteriophage Lambda into the *E. coli* chromosome and switch between lytic and lysogenic pathways (Ptashne. 1992). The integration of Lambda DNA into the *E. coli* chromosome results from the interaction between Lambda and *E. coli* recombination proteins that mediate recombination between specific sequence attachment (*att)* sites. Details of the recombination cloning can be found in chapter II and references Landy. 1989 and Ptashne. 1992.

The success of recombination cloning was recently demonstrated by members of the human ORFeome consortium who have used Gateway® cloning to build the largest publicly available resource of 12,212 ORFs representing 10,214 human genes (Lamesch et al. 2007).

5.1.3.2 Gene reporter assays

With the exception of promoter elements, enhancers are perhaps the best characterised regulatory element because it is more straightforward to assay for increased gene expression. Reporter genes are typically used to determine whether a gene or element of interest has been taken up by or expressed in a population of cells. In practice the reporter gene and test DNA are introduced into cells in culture (*in vitro*) or cells of a living organism (*in vivo*) in a single DNA construct, typically a plasmid. Reporter genes should be exogenous i.e. not normally expressed in the cells being assayed and readily detectable. Commonly used examples of reporter systems include green fluorescent protein (GFP), ß-galactosidase and luciferase. Selectable markers conveying resistance to antibiotics in cells taking up the plasmids are also in common use.

Dual reporters are commonly used to improve experimental accuracy because the dual reporters provide simultaneous and independent measures of expression within a single system. In the case of the Dual-Luciferase® reporter (DLR™) assay (Promega, E1960) the experimental construct is used to measure the effect on firefly (*Photinus pyralis*) luciferase expression under certain experimental conditions and the co-transfected vector expresses *Renilla* (*Renilla reniformis*) luciferase in a constant manner and therefore provides internal normalisation for transfection efficiency and cell viability.

5.1.3.3 Choice of human cells to test

The choice of human cells or cell-lines to use in transient transfection reporter assays is not always a straightforward one and is best guided by the biological question(s) being addressed. However, technical issues such as the transfectability and ready growth of cells must also be considered. As this study is principally focused on the 11p15.5 region harbouring the co-ordinately expressed *IGF2* and *H19* genes then the identification of enhancers up-regulating these genes requires a cell-line in which they are expressed. *IGF2* and *H19* are strongly expressed (from different parental alleles) in foetal liver and down-regulated in adult liver although transcripts are still detectable (Wu et al. 1997). Previous studies of this region have utilised a human hepatocellular liver carcinoma (HepG2) adherent cell-line in which endodermal enhancer activity was demonstrated (see below and Dannenberg and Edenberg. 2006, Long and Spear. 2004). HepG2 cells exhibit many cellular features of normal adult hepatocytes (Bouma et al. 1989) but also express alpha-foetoprotein, a characteristic marker of foetal liver cells. HepG2 cells therefore offer a suitable environment, with necessary transcription factors, for *IGF2* and *H19* gene regulation. As discussed in section 5.3.1 below, known endodermal enhancers have

been characterised in HepG2 cells. It is therefore anticipated that the action of other, as yet unidentified, endodermal enhancers should be detectable in HepG2 cells. Although the adherent HepG2 cells have a relatively slow growth profile they are readily transfected using, for example, GeneJuice® transfection reagent (Novagen).

5.1.4 Epigenetics

A full understanding of the regulation of transcription will require a thorough appreciation of the interactions between regulatory DNA elements and their chromatin environment. Recent advances in the fields of epigenetics and chromatin biology have led to the histone code hypothesis (Strahl and Allis. 2000). Chemical modifications of histone H3 or H4 N-terminal tails have been associated with specific biological processes such as those mediated from promoter or enhancer elements. For example, acetylation of H3 or H4 residues and tri-methylation of lysine 4 (K4) residues are associated with the promoters of actively transcribed genes (Lachner et al. 2003). In contrast, mono-methylation of K4 on histone H3 was recently linked to enhancer elements (ENCODE Project Consortium et al. 2007, Heintzman et al. 2007). Such epigenetic signatures should therefore be informative for identifying novel *cis*-regulatory elements in the human genome. ChIP products hybridised to DNA microarrays is a technology (known as ChIP-chip) which informs about protein-DNA interactions *in vivo* and maps those interactions to precise regions of the genome. The first use of ChIP-chip in mammals mapped GATA-1 binding sites in the beta-globin locus (Horak et al. 2002). Since then a plethora of TFBSs and sites of histone modifications have been mapped to regions of the human genome (ENCODE Project Consortium et al. 2007 and references within).

The capacity to map known transcription factors (TF) or chromatin-associated proteins to the genome has been a huge advance but what of, as yet, unknown regulatory factors? DNaseI HS sites in the genome are relatively depleted of nucleosomes, indicative of open chromatin, and can be mapped using DNase-chip to give an accurate genomic location of functional regulatory elements (Crawford et al. 2006). Xi and colleagues recently mapped 3,904 DNaseI HS sites from 6 cell lines across the ENCODE regions, 22% of these sites were present in all 6 cell lines studied. Of these ubiquitous sites 86% correspond to promoter regions, lying near annotated transcription start sites (TSS). A further 10% were found to bind CTCF and therefore likely to represent insulator elements (Xi et al. 2007). The large proportion of DNaseI HS sites specific to one or a sub-set of the cell lines tested were found to be enriched for enhancer elements.

5.2 Identifying ECRs

5.2.1 Multi-species sequence alignment

The web-based zPicture server (Ovcharenko et al. 2004a) was used to align multiple sequences, interactively visualise genomic features and identify ECRs. Within zPicture sequence alignments were performed using BLASTZ between a reference sequence (typically human) and one or more orthologous sequences in a given region (Figure V.2). Sequences were either uploaded into the server from links to the UCSC genome browser or could be uploaded from the PC running the application. Uploading sequence from UCSC has the advantage of bringing with it genome annotation including the location of repeats. Typically the human (March 2006, hg18) and mouse (February 2006, mm8) finished sequence assemblies were imported from UCSC and sequences generated in chapter III were uploaded locally.

In all cases sequences were masked for repeats (see chapter II) to avoid incorrect sequence alignments.

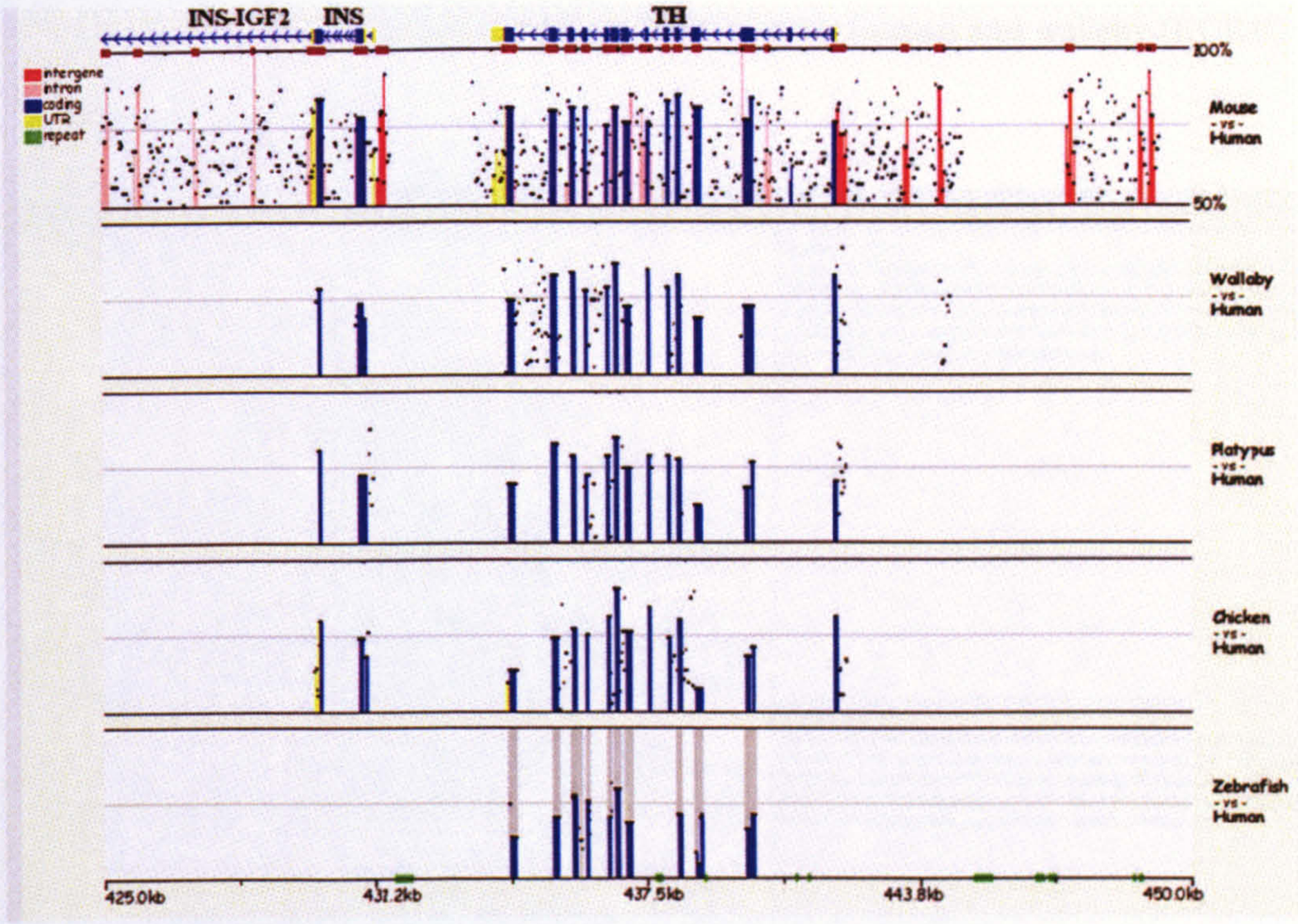


Figure V.2. Example of a zPicture dynamic visualisation plot.

A 25 kb region encompassing the insulin (INS) and tyrosine hydroxylase (TH) genes is shown. The direction of transcription is indicated by blue arrows within introns separating exons (blue rectangles). Percentage identity plots for pairwise comparisons between human and mouse (or wallaby or platypus or chicken or zebrafish) are shown from top to bottom. ECRs, defined as sequence alignments with at least 70% identity over 100 bp, are colour coded to indicate whether they are intergenic (red), intronic (pink), coding (blue), within UTRs (yellow). Repeats are indicated on the bottom axis in green. The grey shading present in the zebrafish *TH* exons reveal the zebrafish sequence to be in the reverse complement. The zebrafish sequence is both unfinished and incomplete in this region.

The optimal thresholds for sequence length and identity used to detect ECRs can be adjusted using the dynamic flexibility of the zPicture browser. ECRs identified from the 11p15.5 region for functional characterisation (below) used the default settings i.e. spanning at least 100 bp with 70% identity between human and wallaby sequences. Conservation between human and wallaby was selected because this

provided a testable number of ECRs of probable function. Importantly these ECRs included all those which were also conserved in available sequences for platypus and chicken. An example of BLASTZ alignment between human and wallaby (ECR#26) is shown in Figure V.3.

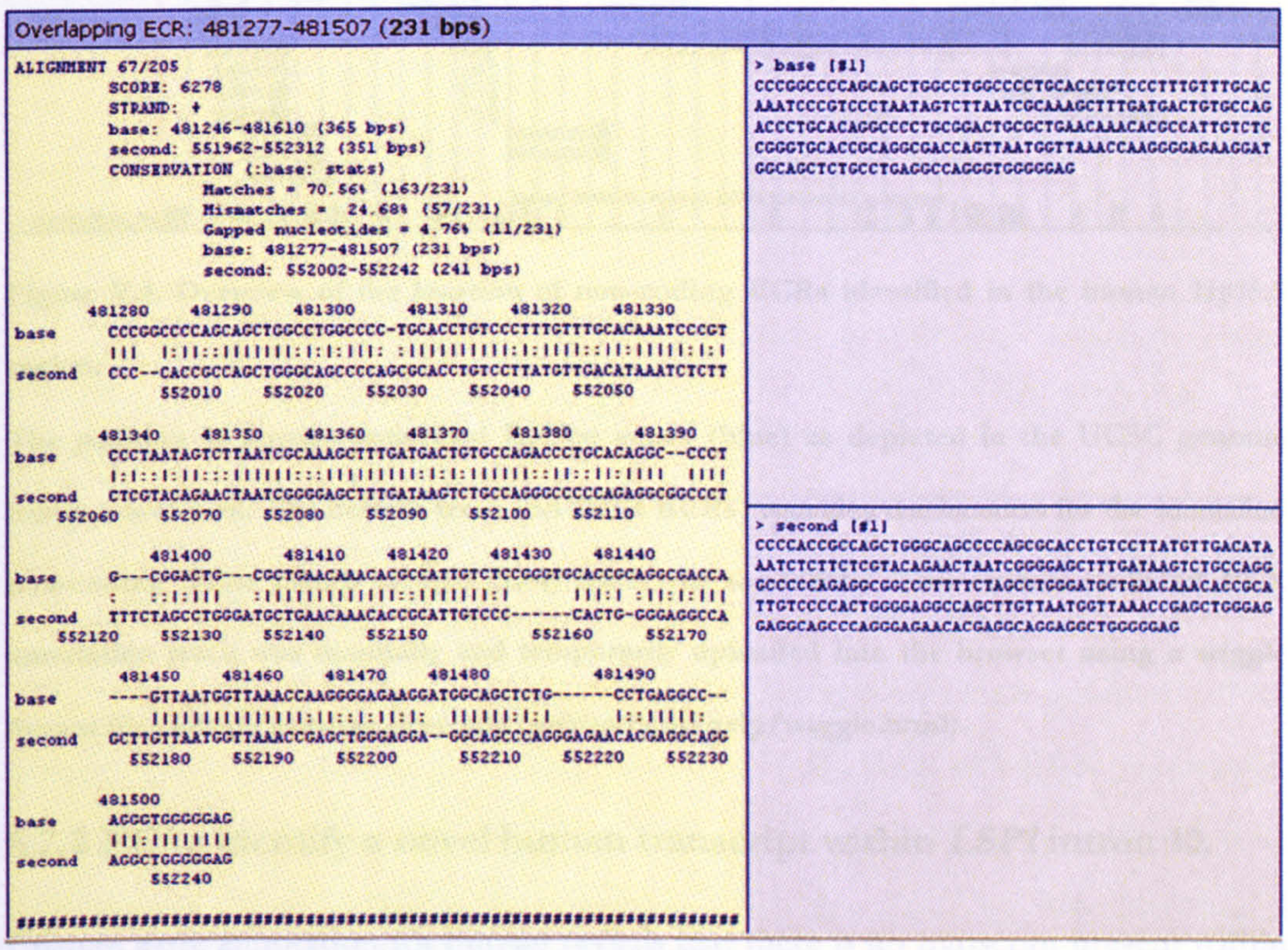


Figure V.3. BLASTZ sequence alignment viewed in zPicture.

Human (base) and wallaby (second) sequences align to give a 231bp intergenic ECR (ECR#26). The location of this ECR (481277-481507) relates to its position in the 1.3 Mb human sequence submitted to the zPicture server.

In total 66 ECRs not overlapping the NCBI Reference Sequence (RefSeq) collection (Pruitt et al. 2007) of exons or UTRs (at the time of analysis) were identified from the alignment of 1.3 Mb of human with wallaby sequence, mapping between *HCCA2* and *OSBPL5* genes (Figure V.4 and Table V-1). Despite the fact that the human sequence was masked for repeats, subsequent screening for repeats of individual ECRs using RepeatMasker (RM database version 20050112) revealed that 99 of 118 bp from ECR#27 corresponds to a simple microsatellite ([CA]n) repeat.

This would appear to indicate a reduced sensitivity of repeat masking (at least for tandem repeats) within the zPicture application. ECR#27 was therefore not functionally tested.

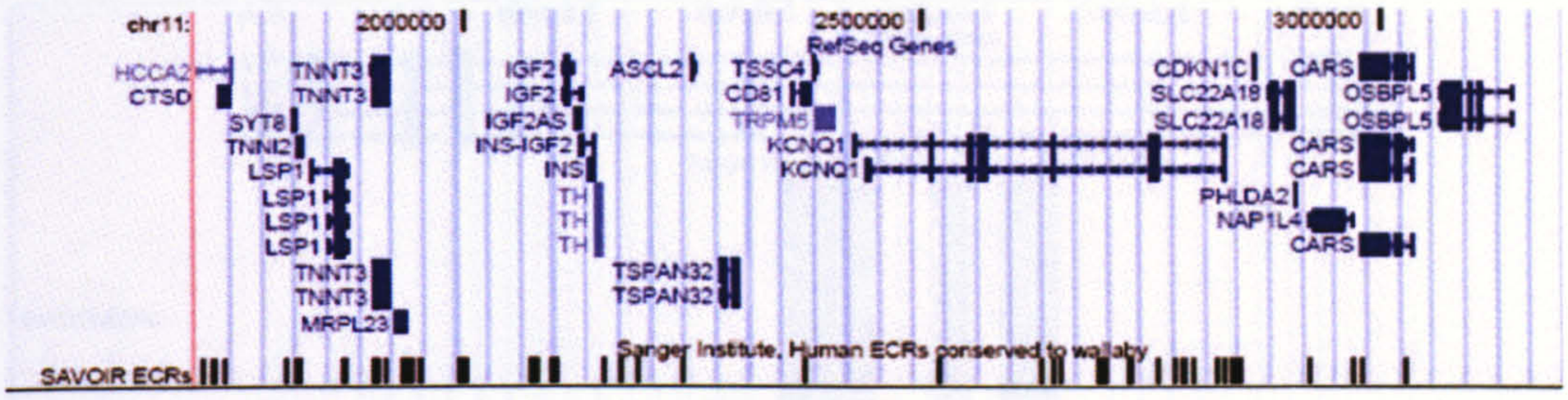


Figure V.4. Overview of the location of non-coding ECRs identified in the human 11p15.5 region.

The position of human annotated RefSeq genes (blue) as depicted in the UCSC genome browser is shown. The bottom track (SAVOIR ECRs) provides the location for the identified non-coding ECRs (black vertical lines) conserved to wallaby. This custom SAVOIR ECR annotation track was manually and temporarily uploaded into the browser using a wiggle format file (http://genome.ucsc.edu/goldenPath/help/wiggle.html).

5.2.2 ECRs identify a novel human transcript within *LSP1* intron 10.

Current gene annotation for human reveals that there is no transcript present within or overlapping the lymphocyte-specific protein 1 (*LSP1*) gene (Figure V.5). However, 5 ECRs (ECR#1-5, Table V-1) are clustered within 1.5 kb of intron 10 of the *LSP1* gene and are conserved in all mammals sequenced but not in chicken.

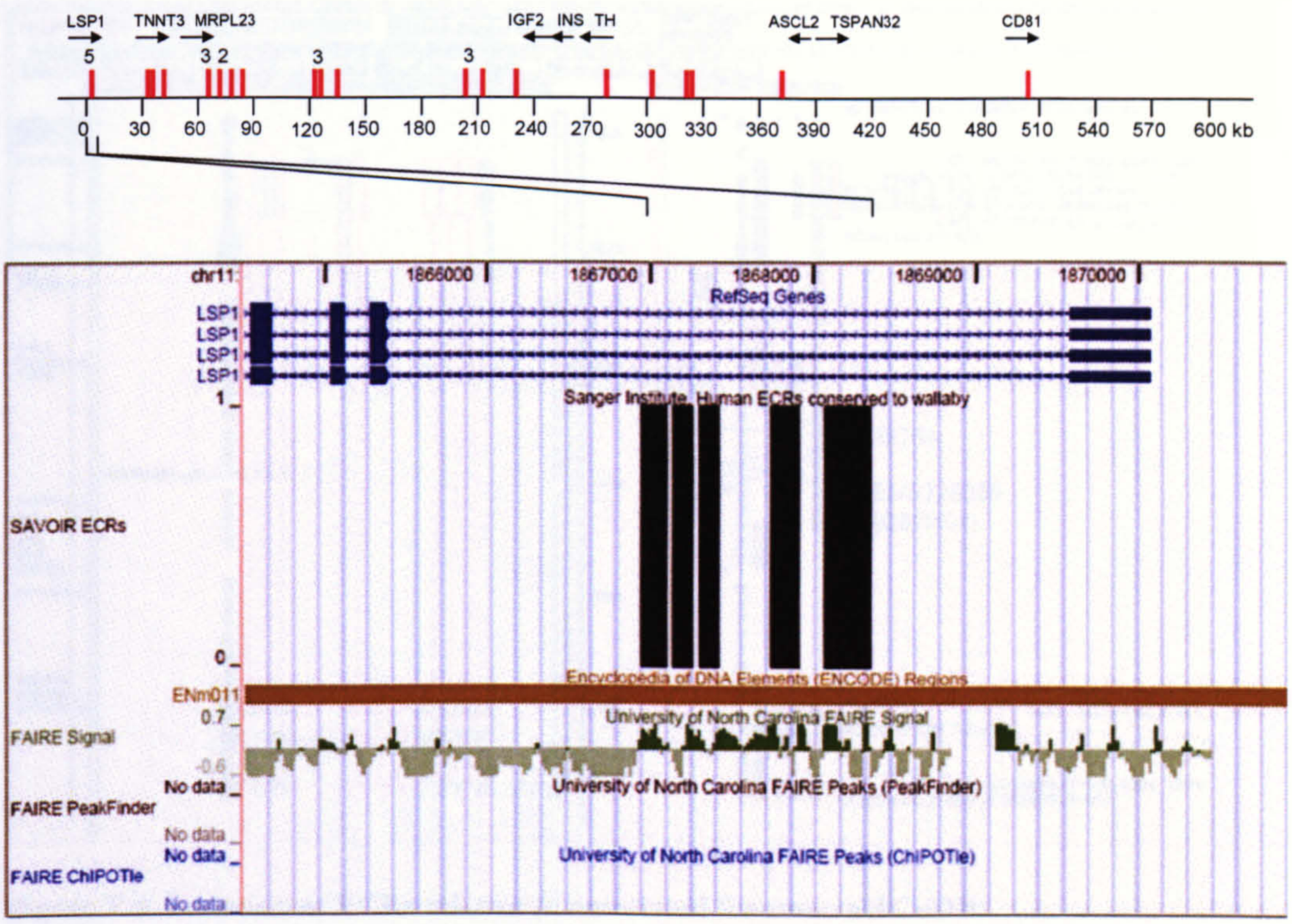


Figure V.5. Clustered ECRs within intron 10 of the *LSP1* gene.

The top schematic shows the position of ECRs (red vertical lines) relative to protein coding genes. Black arrows indicate direction of transcription. A cluster of 5 ECRs lying within intron 10 of the *LSP1* gene is illustrated in the UCSC genome browser. *LSP1* RefSeq annotated transcripts are shown in blue. This region lies within the ENCODE region ENm011 (brown bar). Uploading an ECR track into the UCSC genome browser allows easy correlation with other genome features such as formaldehyde assisted identification of regulatory elements (FAIRE) signals shown in green.

For all species only limited experimental evidence (e.g. ESTs and mRNAs) existed at this locus. However, an EnsEMBL predicted mouse protein (Q8C494) based on a single RIKEN cDNA (AK082720) partially overlaps the ECR cluster (Figure V.6).

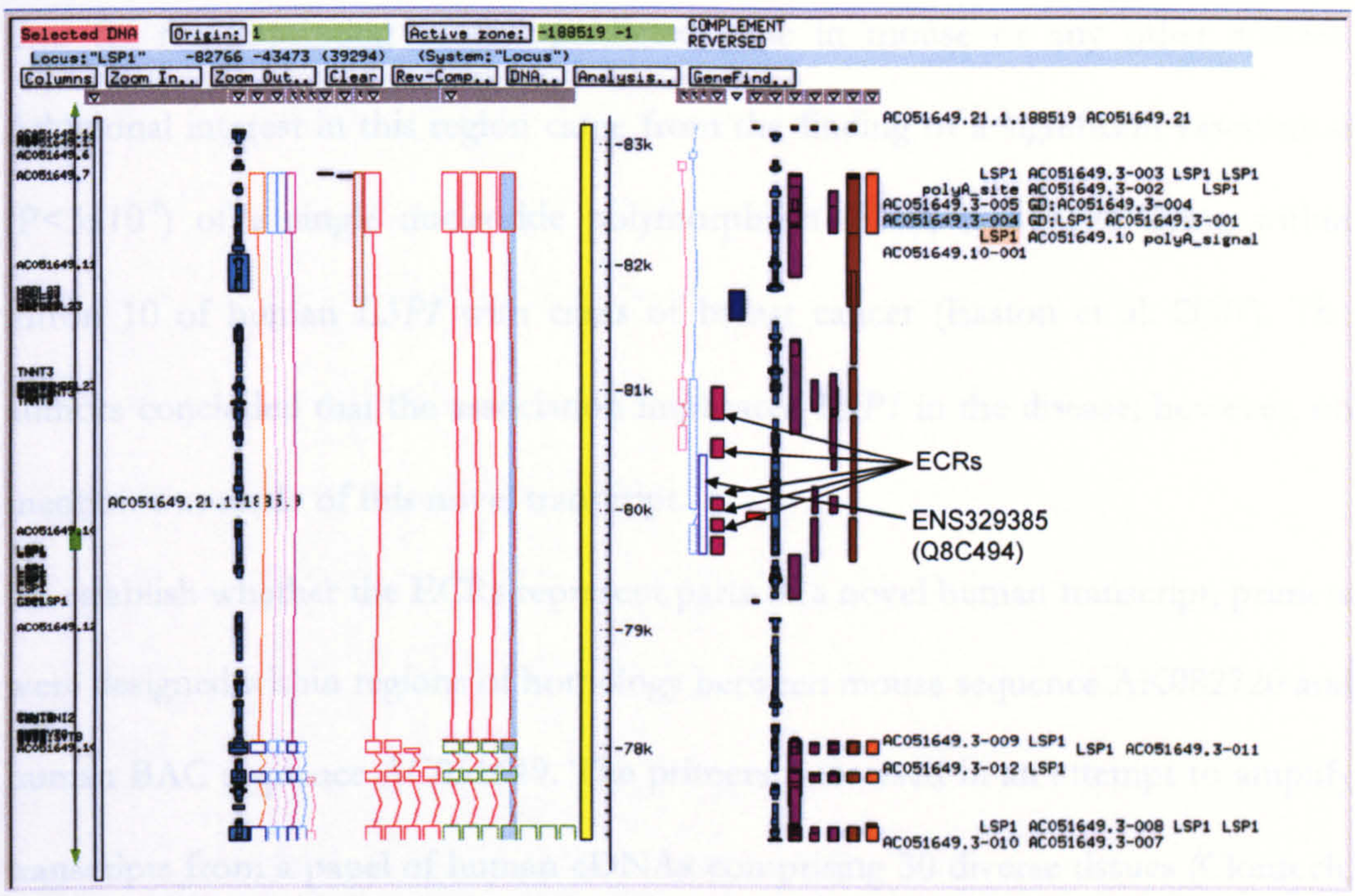


Figure V.6. Location of ECRs relative to annotated features in ACeDB.

The graphical display shows a 6 kb region from the human BAC sequence AC051649 focused on intron 10 of the *LSP1* gene. The HAVANA annotated gene structures are drawn to the left of the yellow vertical bar, illustrating the *LSP1* gene viewed on the reverse strand. The locations of ECRs 1-5 (pink rectangles) are indicated. Three ECRs overlap the Ensembl mouse hypothetical protein Q8C494 outlined in blue. Protein (blue) EST (purple), mRNA (brown) and RefSeq (orange) matches to the BAC sequence are shown and used to annotate gene structures.

The mouse neonate cerebellum cDNA (AK082720) is 2245 bp in length and when translated in frame 3, reveals a hypothetical protein of 260 amino acids with no matches to the Pfam database of protein domains (Finn et al. 2006). The 5' end of AK082720 lies within intron 6 of the *Tnnt3* gene (reverse strand) and splices to the 3' end within intron 10 of *Lsp1* where the potential coding sequence (CDS) is located (Figure V.8B). This mouse predicted gene, which lies outside the previously defined IC1 domain, was further predicted to be maternally expressed in a bioinformatic study (Luedi et al. 2005). There is, however, no experimental evidence

thus far for imprinting of this predicted gene in mouse or any other species. Additional interest in this region came from the finding of a significant association ($P<3x10^{-9}$) of a single nucleotide polymorphism (SNP, rs3817198) lying within intron 10 of human *LSP1* with cases of breast cancer (Easton et al. 2007). The authors concluded that the association implicated *LSP1* in the disease; however, no mention was made of this novel transcript.

To establish whether the ECRs represent parts of a novel human transcript, primers were designed within regions of homology between mouse sequence AK082720 and human BAC sequence AC051649. The primers were used in an attempt to amplify transcripts from a panel of human cDNAs comprising 30 diverse tissues (Clontech, amplification performed by Jackie Bye, formerly in the Sanger Institute experimental gene annotation group). Rapid Amplification of cDNA ends (RACE) PCR products were obtained from prostate, small intestine, testis and retina tissue cDNAs, and sequenced. Alignment of these sequences with human chromosome 11 was performed using BLAT in the UCSC genome browser and revealed two alternate splice forms (Figure V.8A). Transcript variant 1 was observed in prostate, small intestine and retina, whereas variant 2 was observed in testis. Both transcripts reside on the reverse strand and splice (position hg18chr11:1868399) into a 3' exon, lying within intron 10 of the *LSP1* gene and partially overlapping ECR5 (Figure V.8A). Manual alignment of ECRs 1-5 with the longest ORF identified in this region of the human genome indicates that all 5 ECRs are part of the last coding exon for this novel gene (Figure V.7).

```
LongestORF      DTVMLISAASMAPEVCGPSLQGTPGPPPPLLPKPGKDNLRLQKLLRKAARKKMMGGTHLA
HumanECR5                              QGTPGPPPPLLPKPGKDNLRLQKLLRKAARKKMMGGTHLA

LongestORF      PPRAFRTSLSPVSEASHDQEVTAPHAAEGPHPAEAPRLPEAPRPAEAPRMVAALPRSPHT
HumanECR5       PPRAFRTSLSPVSEASHDQEVTAPHAAEGPHPAEAPRLPEAPRPAEAPRMV

LongestORF      PIIHHVASPLQKSTFSIGLTQRRILAAQFRAMGPQVVASAPEPTRPPSGFVPVSGGGGTH
HumanECR4                                                      PSGFVPVSGGGGTH

LongestORF      VTQVHIQLAPSPHNGTPEPPRTAPEVGSNSQDGDATPSPPRAQPLVPVAHIRPLPTTVQA
HumanECR4       VTQVHIQLAPSPHNGTPEPPRTAPEVGSNSQDGDATPS

LongestORF      ASPLPEEPPVPRPPPGFQASVPREASARVVVPIAPTCRSLESSPHSLVPMGPGREHLEEP


LongestORF      PMAGPAAEAERVSSPAWASSPTPPSGPHPCPVPKVAPKPRLSGWTWLKKQLLEEAPEPPC
HumanECR3                                  CPVPKVAPKPRLSGWTWLKKQLLEEAPEPPC

LongestORF      PEPRQSLEPEVPTPTEQEVPAPTEQEVPALTAPRAPASRTSRMWDAVLYRMSVAEAQGRL
HumanECR3-2     PE                                VPALTAPRAPASRTSRMWDAVLYRMSVAEAQGR

LongestORF      AGPSGGEHTPASLTRLPFLYRPRFNARKLQEATRPPPTVRSILELSPQPKNFNRTATGWR
HumanECR1                          LPFLYRPRFNARKLQEATRPPPTVRSILELSPQPKNFNRTATGWR

LongestORF      LQ*
HumanECR1       LQ*
```

Figure V.7. All 5 ECRs comprise a terminal coding exon.

A 3 kb nucleotide sequence (hg18 chr11:1,866,151-1,869,150) centred on ECRs1-5 was reverse complemented and translated in all 3 frames to identify the longest ORF (displayed in black from 5'[top] to 3'[bottom]). ECR nucleotide sequences were also reverse complemented then translated and manually aligned with the human ORF.

Alternative 5' exons for these transcripts which did not correspond to ECRs were identified beginning 106 bp (transcript variant 2) and 1905 bp (transcript variant 1) from the 5' end of the *TNNT3* gene (Figure V.8A). The proximity of the 5' ends of these novel human transcripts to the *TNNT3* gene suggests that they may share a bi-directional promoter with *TNNT3*. Further work will be required to fully characterise the novel gene structure and assess its expression status and possible role in breast cancer. The results presented here show that ECRs can identify novel transcripts.

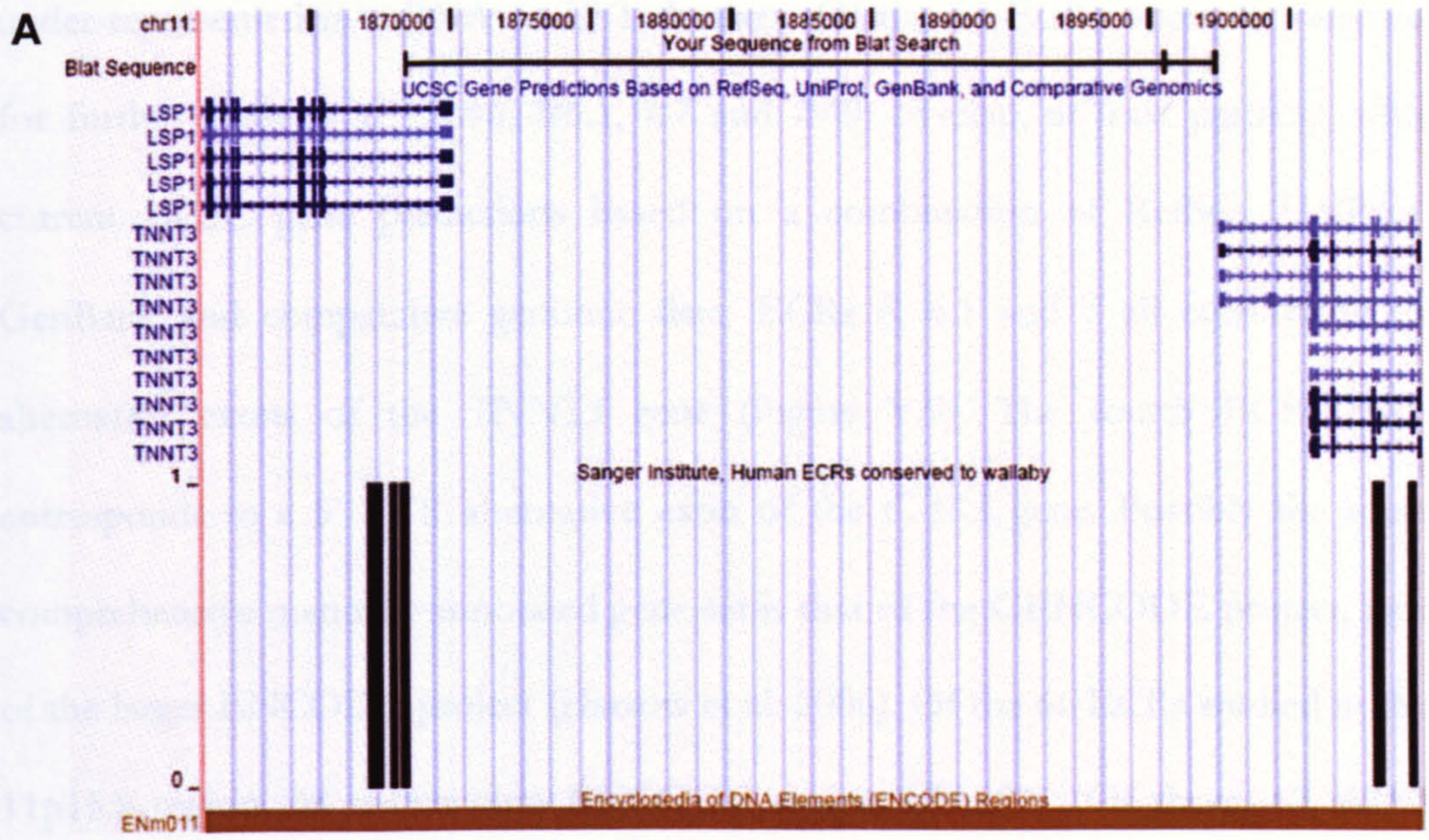


Figure V.8. Extent of human and mouse novel transcripts visualised in the UCSC genome browser.

A) Sequences of PCR amplified human cDNA products were matched by BLAT analysis against the sequence of human chromosome 11. Two splice forms were identified, both of which splice to a 3' exon lying within intron 10 of the *LSP1* gene. This exon corresponds to a cluster of ECRs (black bars). Alternate 5' exons were identified, lying 106 bp and 1,905 bp from the TSS of the *TNNT3* gene. B) In mouse a single mRNA sequence (AK082720) splices between a 5' exon within an intron of the *Tnnt3* gene (opposite strand) and 3' exons, corresponding to the ECRs, within intron 10 of the *Lsp1* gene.

5.2.3 ECRs identify alternative exons

In addition to revealing the presence of entirely novel transcripts in the human (and other) genome(s) ECRs also highlight alternative exons used by transcripts which may have a more restricted temporal or spatial expression pattern and therefore be

under-represented in mRNA or EST datasets. Of the 66 ECRs originally selected for further study, 4 (ECR#6, #6.1, #7 and #40) overlap, at least partially, with current UCSC gene predictions based on a combination of RefSeq, UniProt, GenBank and comparative genomic data. ECRs 6, 6.1 and 7 all correspond to alternative exons of the *TNNT3* gene (Figure V.9). The fourth ECR (#40) corresponds to a 5' UTR alternative exon of the *CARS* gene. Possibly the most comprehensive manually annotated gene set is that of the GENCODE project, part of the larger ENCODE project (Harrow et al. 2006). Of the 66 ECRs studied in the 11p15.5 region, 38 map within ENCODE region ENm011. Of these, 16 (42%) overlap with GENCODE annotated exons and illustrate the need for GENCODE style annotation of the remaining 99% of the human genome, which is now underway. The examples of novel transcripts and exons described here illustrate the potential of comparative sequence analysis to improve genome annotation. However, these exons were identified because of new annotation in at least one species. It would be challenging to identify coding exons for which no experimental evidence yet exists, not least because ECRs provide no splicing information. However, one could devise a strategy in which all ECR sequences (especially those clustered) are translated in all 6 frames (both strands) to identify those containing a minimal length ORF. Reverse-transcriptase (RT)-PCR primers could then be designed between neighbouring ECRs for cDNA library screening.

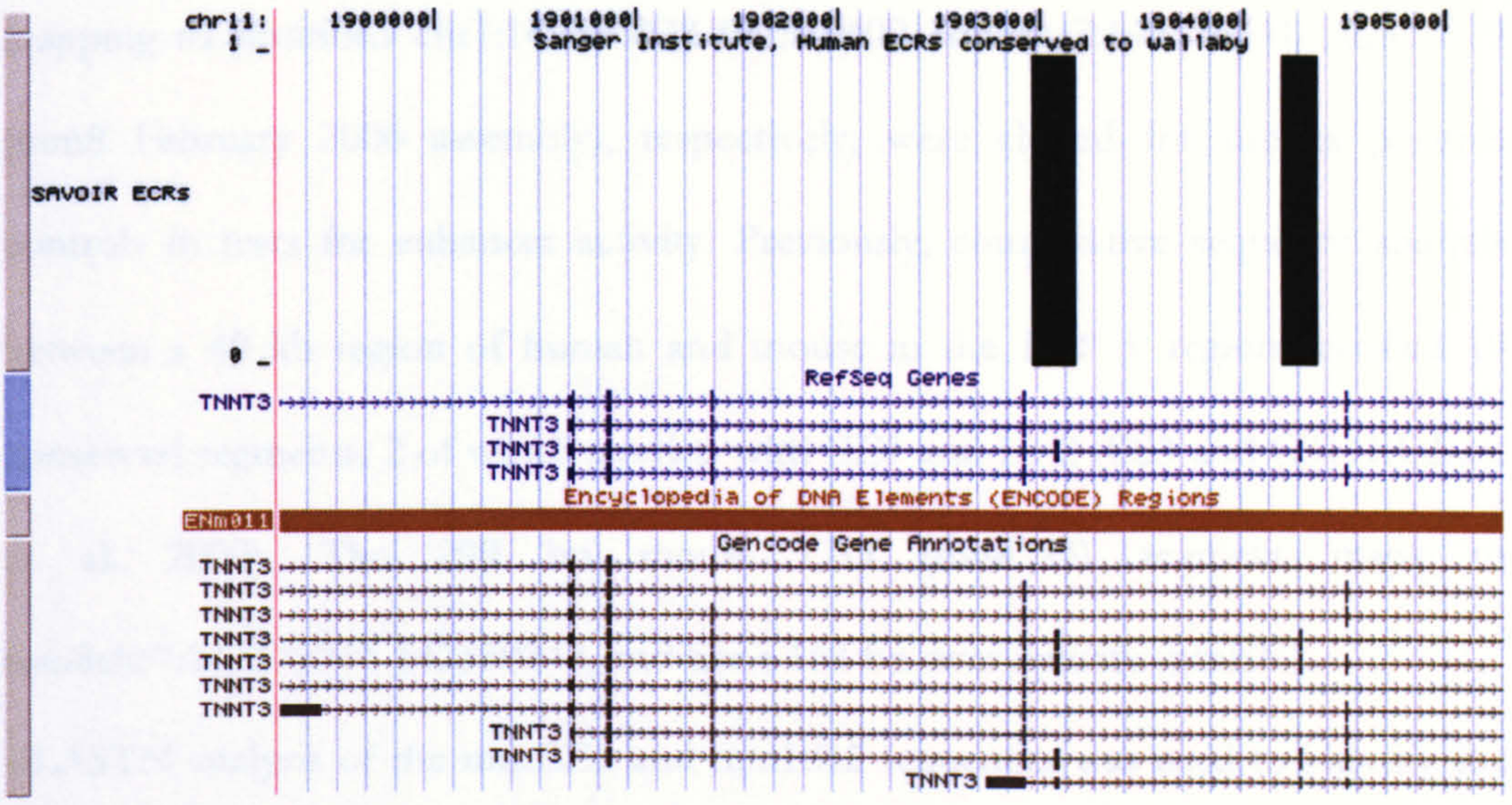


Figure V.9. Example of ECRs highlighting alternative exons.

A 6 kb screenshot of the UCSC genome browser centered on the *TNNT3* gene is shown. The two ECRs (black bars) encompass alternative exons annotated in 1 out of 4 ReSeq gene annotations (blue lines) and 4 out of 10 Gencode gene annotations (green lines). The single RefSeq annotation with alternative exons was not present in the database at the time the ECRs were identified.

5.3 Testing ECRs for enhancer activities

The above analyses have shown that 16 of the initial 66 ECRs correspond to novel exons. However, because much of the new annotation is very recent no ECRs were excluded from the enhancer testing that follows.

5.3.1 Generating enhancer positive controls

Table V-2 details the positive control enhancer sequences which have been cloned for use in this study. Two mouse endodermal enhancer elements (EE1 and EE2) positioned 6 and 8 kb centromeric of the *H19* gene, respectively, have been characterised and shown to interact with the promoter regions of *H19* and *Igf2* genes in an allele-specific manner (Leighton et al. 1995, Yoo-Warren et al. 1988 and chapter VI). Fragments containing mouse EE1 (mmEE1) and EE2 (mmEE2)

mapping to positions chr7:142380078-142380402 and chr7:142378341-142378829 (mm8 February 2006 assembly), respectively, were cloned for use as positive controls in tests for enhancer activity. Previously, comparative sequence analysis between a 40 kb region of human and mouse in the *H19* 3' region revealed 10 conserved segments, 2 of which overlap with EE1 and EE2 (CS3 and CS4, Ishihara et al. 2000). The 299 bp mouse CS3 (mmCS3) segment maps to mm8chr7:142379944-142380242, and has a 164 bp overlap with mmEE1.

BLASTN analysis of the mmEE1 and mmEE2 sequences was used to identify the orthologous human sequences, not previously tested for enhancer activity. PCR amplicons containing human EE1 (hsEE1, hg18chr11:1967732-1968058) and human EE2 (hsEE2, hg18chr11:1965984-1966376) were cloned.

Mouse alpha fetoprotein (*Afp*) gene transcription is activated early in hepatogenesis but is dramatically repressed within several weeks of birth. *Afp* is therefore co-expressed with *H19* in liver and its regulation has been well studied (Godbout et al. 1988). Indeed *H19* was identified due to its coordinate regulation with *Afp*. Three enhancers (EI, EII and EIII) are known to regulate *Afp* expression and lie 2.5, 5.0 and 6.6 kb upstream of the *Afp* TSS respectively. The activity of each enhancer has been localised to minimal enhancer regions (MERs) of 200-300bp (Godbout et al. 1988). These mouse MERs were cloned for additional positive enhancer controls in transient transfection of human HepG2 cells.

The known endodermal enhancers above are relevant positive controls when testing ECRs for enhancer activity in HepG2 cells. For comparison mesodermal enhancers, which are not expected to function in HepG2 cells, were also included. Mouse *H19* upstream conserved region 1 (mmHUC1, mm8chr7:142397711-142398203) and region 2 (mmHUC2, mm8chr7:142396168-142396566) were identified from

human-mouse sequence comparisons (Drewell et al. 2002). Human HUC1 (hsHUC1, hg18chr11:1990054-1990551) and HUC2 (hg18chr11:1988196-1988584) orthologous sequences were also cloned (see below).

5.3.2 Generating negative ('Randomer') controls

Of equal importance to the positive controls described above are negative controls which provide a basal level of firefly luciferase expression for each experiment. In the literature the pGL3-Basic vector (Promega) is frequently used for this purpose, but, unlike the modified pGL3-Promoter vector this vector has no SV40 promoter element upstream of the firefly luciferase reporter gene and is not really a suitable control. The pGL3-Promoter vectors used in enhancer tests were modified to contain a Gateway® cassette (RfC.1) and different antibiotic resistance genes (see below). I therefore elected to use as negative controls the same destination vector into which the test fragments were cloned. For simplicity I refer to these negative control vectors as 'empty' vectors. However, it should be noted that since the 'empty' vectors have not been recombined in an LR reaction with entry clones, containing test fragments (Figure V.12), the 1714 bp Gateway® cassette remains present within the 'empty' vector (see section 5.3.3). Since the cassette contains only prokaryotic sequences it is very unlikely that these sequences could influence firefly luciferase gene expression.

If the empty vectors truly have a negligible effect on luciferase expression then we might expect luciferase levels from a random sampling of sequences cloned into the vectors not to deviate from those of the empty vectors. To address this, 40 randomly selected and repeat-masked sequences from the 11p15.5 region, matched for length and G+C content with the cloned ECRs, were PCR amplified. Like the

ECRs, these 'randomers' were cloned in forward and reverse orientations within the modified pGL3-Promoter vectors and used to transiently transfect HepG2 cells. Except for 4 randomers (10.1, 12, 13 and 23m) all showed less than 2-fold enhancement compared to the empty vector (Figure V.10).

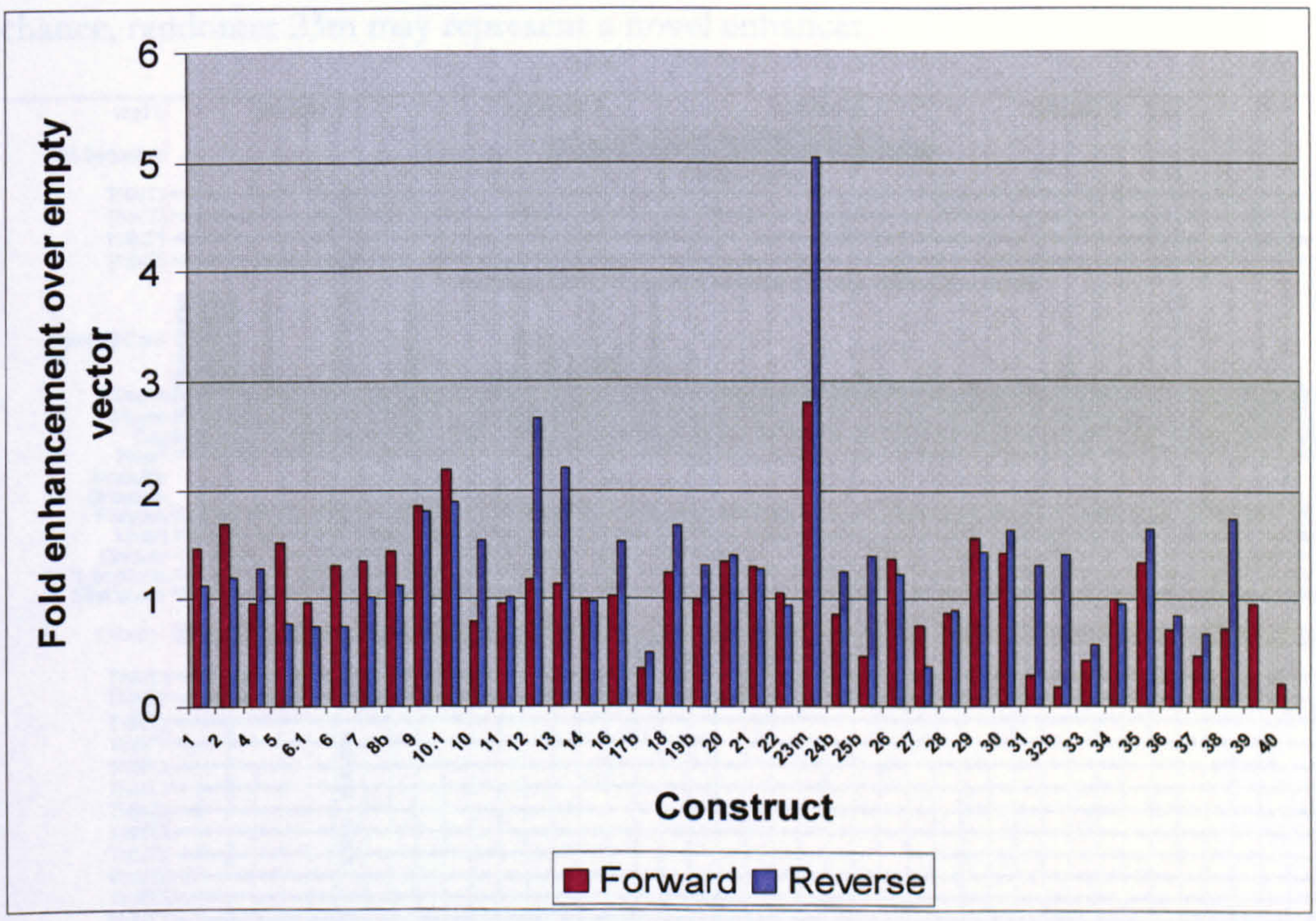


Figure V.10. Testing human 'randomers' for enhancer activity in HepG2 cells.

For each of 40 different constructs average firefly/renilla luciferase ratios, from four replicates, were normalised against average luciferase levels of empty pGL3-Promoter vectors. Randomers cloned in forward (claret) and reverse (blue) orientations were assayed.

Three of the four randomers have apparent but marginal enhancer activity by the selected criteria, meeting the 2-fold threshold in one orientation. The fourth, randomer 23m, does appear to behave as a strong enhancer element in HepG2 cells (Figure V.10). This sequence (736 bp, 67% G+C) maps within an intron of the *TNNT3* gene (hg18chr11:1904912-1905647). It does not overlap with any of the ECRs identified here, or sites of nucleosome depletion, as determined by FAIRE (Formaldehayde Assisted Identification of Regulatory Elements, Giresi et al. 2007),

or DNaseI hypersensitive sites (Crawford et al. 2006, Sabo et al. 2006). However, there is significant evolutionary conservation at the telomeric end of this randomer sequence as observed in the vertebrate MULTIZ alignment and PhastCons conservation track of the UCSC browser (Figure V.11). Therefore, it seems that, by chance, randomer 23m may represent a novel enhancer.

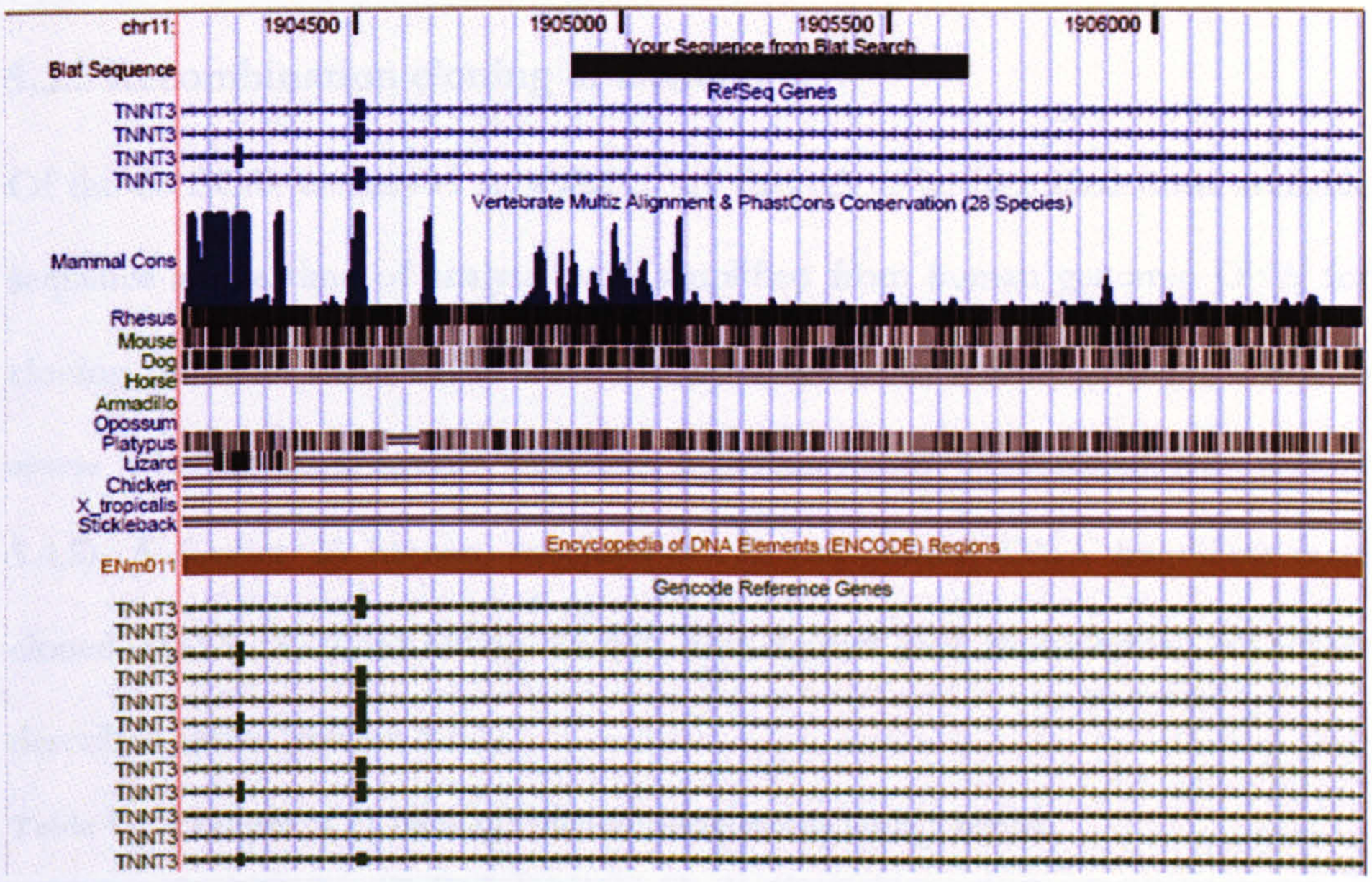


Figure V.11. Sequence conservation overlapping randomer 23m.

The chromosome 11 (chr11) location of the 736bp randomer 23m sequence is shown as a black rectangle. This sequence lies within an intron of the *TNNT3* gene according to both RefSeq (blue) and Gencode (green) gene annotations. The MULTIZ alignment of 28 vertebrate sequences reveals sequence conservation overlapping the left-hand (telomeric) end of randomer 23m.

From Figure V.10 it would appear that the majority of randomers, in either orientation, behave in a similar way as the empty vectors i.e. a fold enhancement over empty vector of around 1. To establish whether there is a statistically significant difference of firefly/*renilla* luciferase ratio values between the empty vectors and randomers, unpaired t-tests were performed. In the forward and reverse

orientations the two-tailed P-values equal 0.3812 and 0.1584, respectively. By conventional criteria these P-values are not considered statistically significant and therefore, the null hypothesis that the compared datasets are the same cannot be rejected. The use of empty vectors as negative controls in subsequent enhancer tests is therefore legitimate.

5.3.3 Recombination cloning of ECRs

Of the 66 ECRs conserved to wallaby, the first 43 to be identified from available sequence at the time of analysis were amplified from human genomic DNA for cloning. Seven of these ECRs were also amplified from wallaby genomic DNA to assess whether cross-species enhancer activities could be detected (see section 5.4.5). A further 23 human, mouse and chicken control DNA fragments were cloned (Table V-2), including known endodermal and mesodermal enhancers described above (section 5.3.1).

Table V-1. Features of ECRs in the human chromosome 11p15.5 region.

ECR name	Length (bp)	Identity (%)	ECR location on Human chr11 (hg18)	Amplicon location on Human chr11 (hg18)	Amplicon Length (bp)
ECR#0.1	187	80	1713086-1713272	ND	ND
ECR#0.2	486	75	1725434-1725919	ND	ND
ECR#0.3	156	72	1737037-1737192	ND	ND
ECR#0.4	101	70	1805120-1805220	ND	ND
ECR#0.5	127	74	1816667-1816793	ND	ND
ECR#0.6	100	70	1816863-1816962	ND	ND
ECR#0.7	99	71	1817582-1817680	ND	ND
ECR#1	145	72	1866951-1867095	1866770-1867306	561
ECR#2	100	71	1867142-1867241	1866895-1867446	576
ECR#3	101	70	1867313-1867413	1867065-1867623	583
ECR#4	159	69	1867745-1867903	1867503-1868112	634
ECR#5	275	73	1868066-1868340	1867878-1868528	675
ECR#6	214	71	1902954-1903167	1902743-1903405	687
ECR#6.1	173	75	1904192-1904364	1904006-1904583	602
ECR#7	142	74	1913976-1914117	1913756-1914288	557
ECR#8	122	74	1932685-1932806	1932436-1932991	581
ECR#9	240	70	1932869-1933108	1932681-1933287	631
ECR#10	135	69	1933440-1933574	1933208-1933823	640
ECR#10.1	136	69	1940580-1940715	1940407-1940956	574
ECR#11	278	77	1942318-1942595	1942106-1942823	640
ECR#12h	222	71	1947891-1948112	1947642-1948357	740
ECR#13	183	73	1952808-1952990	1952634-1953204	595
ECR#14	267	73	1953127-1953393	1952891-1953612	746
ECR#15h	128	70	1992452-1992579	1992239-1992806	592

ECR#16h	121	70	1992747-1992867	1992525-1993053	553
ECR#17h	257	72	1993014-1993270	1992782-1993508	751
ECR#18	124	74	1996775-1996898	1996541-1997133	617
ECR#19h	183	69	2002222-2002404	2002026-2002653	652
ECR#20	214	80	2073828-2074041	2073636-2074277	666
ECR#21h	138	72	2075179-2075316	2074997-2075528	556
ECR#22h	102	71	2075391-2075492	2075196-2075710	539
ECR#23	110	71	2080518-2080627	2080097-2080415	343
ECR#23m	260	74	2098498-2098757	2098262-2099000	763
ECR#24	100	70	2151943-2152042	2151762-2152236	499
ECR#25	307	75	2171870-2172189	2171636-2172372	762
ECR#26	231	71	2189001-2189231	2188754-2189400	671
ECR#27*	118	73	2194957-2195074	2194482-2195173	716
ECR#28	133	68	2239259-2239391	2239061-2239609	573
ECR#29	100	70	2372955-2373054	2372776-2373254	503
ECR#30	161	71	2517629-2517789	2517401-2517971	595
ECR#30.1	187	71	2629967-2630153	ND	ND
ECR#30.2	116	71	2641961-2642076	ND	ND
ECR#30.21	148	68	2642379-2642526	ND	ND
ECR#30.3	186	70	2642565-2642750	ND	ND
ECR#30.4	128	73	2649934-2650061	ND	ND
ECR#30.5	119	73	2691560-2691678	ND	ND
ECR#31	206	72	2698934-2699139	2698710-2699366	681
ECR#32	259	74	2724965-2725157	2724761-2725347	611
ECR#33	127	69	2728502-2728722	2728313-2728969	681
ECR#34	159	74	2756492-2756650	2756311-2756898	613
ECR#34.1	124	72	2774623-2774746	ND	ND
ECR#34.2	135	76	2776282-2776416	ND	ND
ECR#34.21	102	71	2776457-2776558	ND	ND
ECR#34.3	207	71	2784869-2785075	ND	ND
ECR#34.4	128	71	2785255-2785382	ND	ND
ECR#35	174	72	2794974-2795147	2794794-2795367	598
ECR#36	244	71	2821449-2821692	2821262-2821906	669
ECR#37	536	72	2828273-2828675	2828091-2828923	857
ECR#37.1	103	70	2837386-2837488	ND	ND
ECR#38	256	77	2840268-2840488	2840069-2840681	637
ECR#39	347	76	2846919-2847265	2846740-2847447	732
ECR#39.1	172	81	2922247-2922418	ND	ND
ECR#39.2	278	68	2923186-2923463	ND	ND
ECR#39.3	97	80	2970111-2970207	ND	ND
ECR#39.4	145	68	2978527-2978671	ND	ND
ECR#40	334	72	3025606-3025823	3025183-3025802	642

ECRs conserved at least to wallaby are shown. The ECR name, sequence length in basepairs (bp), percentage (%) sequence identity between human and wallaby, location on human chromosome (chr) 11 (genome build hg18), cloned PCR amplicon location and length are provided. ND, not done.

Table V-2. Cloning known functional elements.

Fragment name	Human, mouse or chicken Genome location	Cloned amplicon Size (bp)	Description (Reference)
hsEE1	hg18chr11:1967732-1968058	327	Human endodermal enhancer 1 (unpublished)
hsEE2	hg18chr11:1965984-1966376	393	Human endodermal enhancer 2 (unpublished)
hsHUC1	hg18chr11:1990054-1990551	498	Human mesodermal enhancer 1 (Drewell et al. 2002)
hsHUC2	hg18chr11:1988196-1988584	389	Human mesodermal enhancer 2 (Drewell et al. 2002)
hsH19minpro	hg18chr11:1975637-1975879	243	Human H19 minimal promoter (Brannan et al. 1990)
hsH19ex1	hg18chr11:1974233-1975690	1458	Human H19 exon 1 (Brannan et al. 1990)
hsDMD_A	hg18chr11:1976751-1978696	1946	Human differentially methylated domain (part A)
hsDMD_B	hg18chr11:1979993-1980973	981	Human differentially methylated domain (part B)
mmCS3	mm8chr7:142379944-142380242	299	Mouse conserved segment 3 (Ishihara et al. 2000)
mmEE1	mm8chr7:142380078-142380402	325	Mouse endodermal enhancer 1 (Yoo-Warren et al. 1988)
mmEE2	mm8chr7:142378341-142378829	489	Mouse endodermal enhancer 2 (Yoo-Warren et al. 1988)
mmHUC1	mm8chr7:142397711-142398203	493	Mouse mesodermal enhancer 1 (Drewell et al. 2002)
mmHUC2	mm8chr7:142396168-142396566	399	Mouse mesodermal enhancer 2 (Drewell et al. 2002)
mmMER1	mm8chr5:91563270-91563831	562	Mouse Afp minimal enhancer region 1 (Godbout et al. 1988)
mmMER2	mm8chr5:91560913-91561313	401	Mouse Afp minimal enhancer region 2 (Godbout et al. 1988)
mmMER3	mm8chr5:91559171-91559708	538	Mouse Afp minimal enhancer region 3 (Godbout et al. 1988)
mmH19ex1	mm8chr7:142386122-142387582	1461	Mouse H19 exon 1 (Zubair et al. 1997)
mmINSex3	mm8chr7:142488001-142488276	276	Mouse INS exon 3 (Wentworth et al. 1986)
mmDMD	mm8chr7:142389487-142391656	2170	Mouse differentially methylated domain (Tremblay et al. 1997)
mmDMD_5'	mm8chr7:142391160-142391457	298	Mouse differentially methylated domain (5' region) (Tremblay et al. 1997)
mmDMD_3'	mm8chr7:142389671-142390194	524	Mouse differentially methylated domain (3' region) (Tremblay et al. 1997)
mmDMD_Sil3'	mm8chr7:142389197-142389773	577	Mouse differentially methylated domain (3' silencer region)(Lyko et al. 1997)
ggCoreIns	gg3chr1:199422883-199423157	621	Chicken core insulator (Chung et al. 1997)

hs, Homo sapiens; mm, Mus musculus; gg, Gallus gallus. hg18chr11, chromosome 11 of the human genome (NCBI build 36, March 2006); mm8chr7, chromosome 7 of the mouse genome (NCBI build 36, February 2006); gg3chr1, chromosome 1 of the chicken genome (WUSTL v2.1, May 2006).

The cloning strategy adopted is depicted in Figure V.12; details of the cloning procedure are described in chapter II. First round recombination reactions result in a library of 'Entry' clones that are subsequently cloned into a selection of 'Destination' vectors designed to measure specific transcriptional activities (Figure V.12). The destination vectors used were derived from the pGL3 series of vectors (Promega) which were modified by John Collins to contain a Gateway® cassette (Invitrogen). The position and orientation of the cassette allows specific functions to be assayed. For example, since enhancer activities are known to be largely position and orientation independent the DNA elements to be tested for enhancer activity can be cloned upstream or downstream of the firefly luciferase reporter gene and in either orientation in the pGL3-Promoter vector i.e. 4 possible constructs (Figure V.13). The pGL3-Promoter vector contains an SV40 promoter element and is used to assay whether luciferase reporter gene expression is increased by the addition of a putative enhancer element. To increase the throughput of this cloning system John Collins also substituted the pGL3 ampicillin resistance gene with gentamycin or kanamycin resistance genes (Figure V.13). Antibiotic selection was therefore used to discriminate between test fragments cloned in forward or reverse orientations following a single LR recombination reaction between one entry clone and multiple destination vectors.

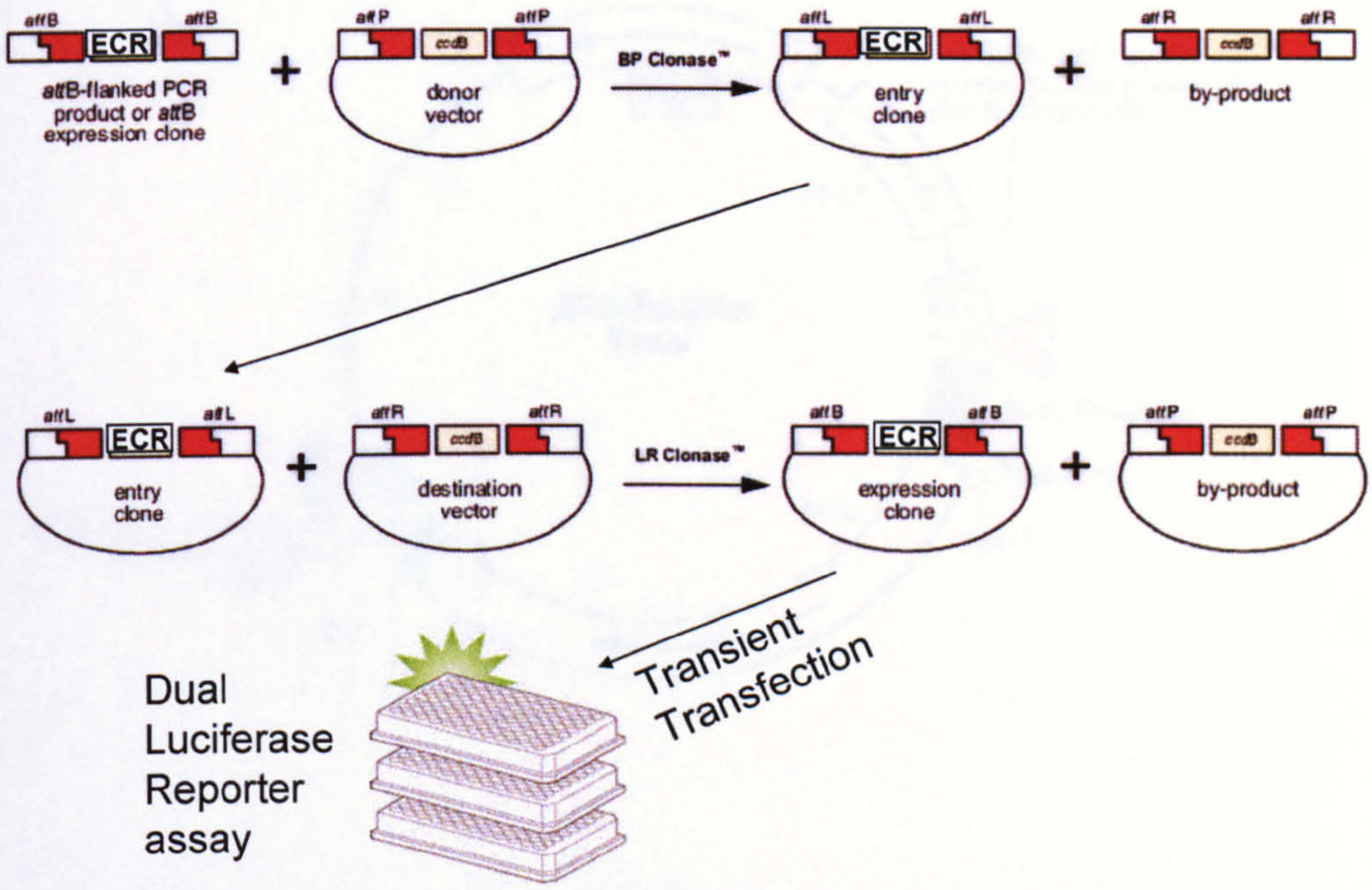


Figure V.12. Gateway® (Invitrogen) recombination cloning strategy.

ECRs to be cloned are PCR amplified with flanking attB sites (top left). Directed recombination mediated by the BP Clonase™ enzyme is performed between the ECR product and donor vector (pDONR223) resulting in an entry clone. A second round recombination reaction between entry clone and destination vector is mediated by the LR Clonase™ enzyme. Resulting expression clones are transiently transfected into human HepG2 cells cultured in 96-well plates. Dual luciferase reporter assays are performed to assess ECR function.

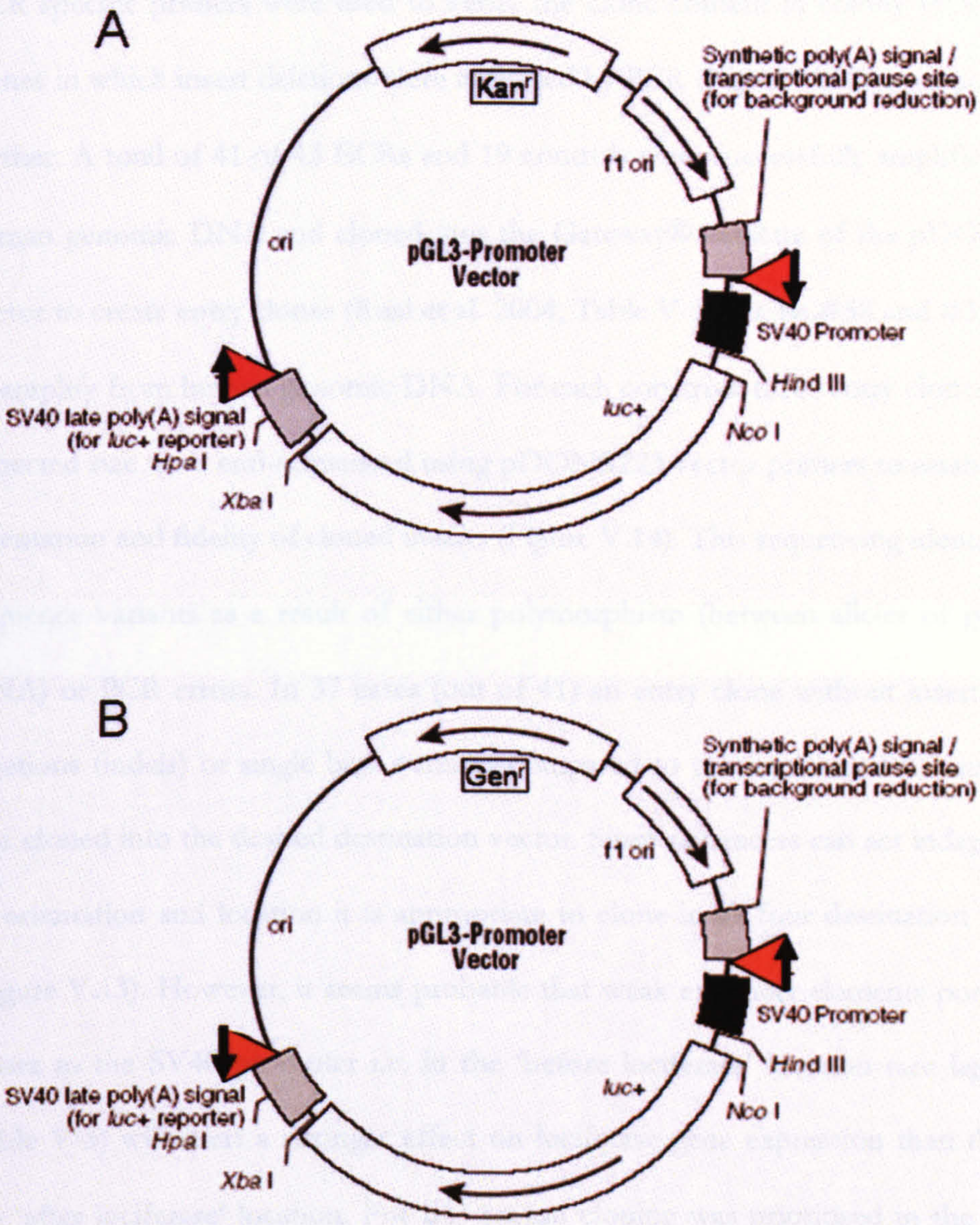


Figure V.13. Gateway® modified pGL3-Promoter vectors for enhancer testing.

pGL3-Promoter vectors purchased from Promega were modified by the addition of Gateway® cassettes (red triangles). In each vector a single Gateway® cassette was cloned either upstream of the SV40 promoter element or downstream of the SV40 late poly(A) signal. Additional modification of the pGL3-Promoter vector includes the replacement of the Ampicillin resistance gene with a Kanamycin resistance (Kan^r) gene (A) or Gentamycin resistance (Gen^r) gene (B).

ECR specific primers were used to verify the clone content in colony PCR. Entry clones in which insert deletions were indicated by PCR (Figure V.14) were not taken further. A total of 41 of 43 ECRs and 19 controls were successfully amplified from human genomic DNA and cloned into the Gateway® cassette of the pDONR223 vector to create entry clones (Rual et al. 2004, Table V-1). ECRs #38 and #39 failed to amplify from human genomic DNA. For each construct three entry clones of the expected size were end-sequenced using pDONR223 vector primers to establish the orientation and fidelity of cloned inserts (Figure V.14). This sequencing identified 17 sequence variants as a result of either polymorphism (between alleles of genomic DNA) or PCR errors. In 37 cases (out of 41) an entry clone without insertions or deletions (indels) or single base variants compared to the reference ECR sequence was cloned into the desired destination vector. Since enhancers can act independent of orientation and location it is appropriate to clone in all four destination vectors (Figure V.13). However, it seems probable that weak enhancer elements positioned closer to the SV40 promoter i.e. in the 'before luciferase' location (see legend to Table V-3) will exert a stronger effect on luciferase gene expression than those in the 'after luciferase' location. For this reason cloning was prioritised in the pGL3-Promoter.KGW.B.F and pGL3-Promoter.G.GW.B.R destination vectors. A total of 279 vector constructs, representing 124 different inserts, were generated in preparation for functional testing in dual luciferase reporter assays (Table V-3).

Table V-3. Details of destination vector cloning.

Destination vectors	Number of Human ECRs	Number of Human ECR26 fragments	Number of Wallaby ECRs	Number of control fragments	Number of Randomers	Total number of cloned fragments per vector
pGL3-Promoter.K.GW.B.F	38	8	7	16	48	117
pGL3-Promoter.G.GW.B.R	39	8	7	15	45	114
pGL3-Promoter.K.GW.A.R	21	0	0	3	0	24
pGL3-Promoter.G.GW.A.F	21	0	0	3	0	24
Total number of cloned elements by category	119	16	14	37	93	279

Modified pGL3-Promoter vector terminology is as follows: K, kanamycin resistance; G, gentamycin resistance; GW, Gateway cassette; B, Gateway cassette placed before SV40 promoter; A, after SV40 late poly(A) signal; F, insert in forward orientation; R, reverse orientation (see Figure V.13). Note; cloning orientations are with respect to the entry clone and do not necessarily reflect orientation in the genome.

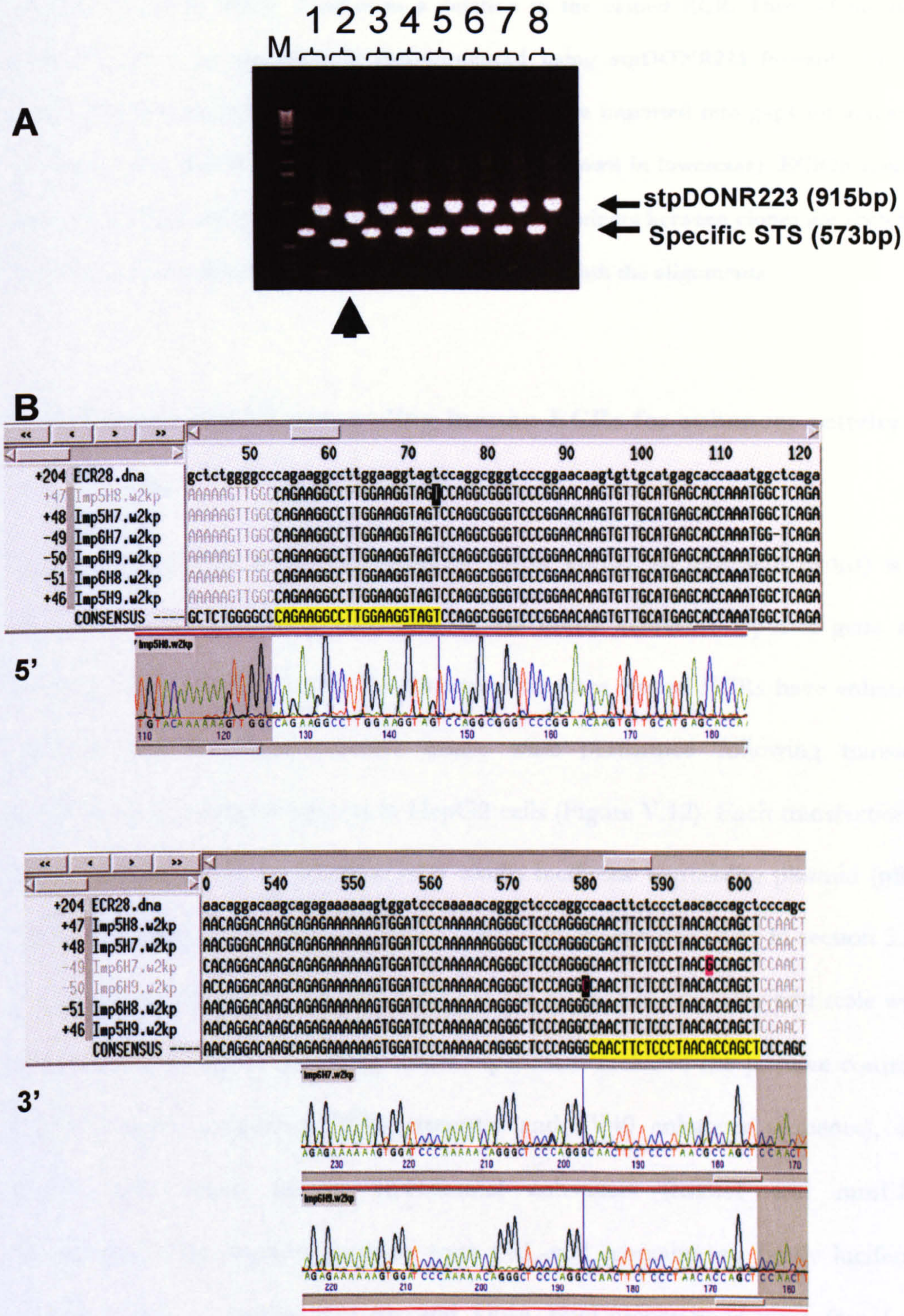


Figure V.14. Cloning verification of the ECR28 pENTR clone.

A, eight colonies resulting from BP recombination of ECR28 into pDONR223 were tested with pDONR223 vector (stpDONR223) and ECR28 specific STSs. PCR products were loaded alongside a size marker (M) in the wells of a 1% agarose gel for electrophoresis. For each colony (1-8) the ECR28 PCR product precedes the stpDONR223 PCR product. The

arrowhead beneath colony 2 indicates a deletion in the cloned ECR. Three of the non-deleted ECRs were subsequently end-sequenced using stpDONR223 forward (5') and reverse (3') primers (B). Individual sequence reads were imported into gap4 for assembly with each other and the reference ECR28 sequence (shown in lowercase). ECR28 specific primer sequences are highlighted in yellow. Sequence variants between clones are shown in pink. Electropherograms for single reads are shown beneath the alignments.

5.3.4 Testing 11p15.5 non-coding human ECRs for enhancer activity in HepG2 cells.

As described above 39 ECRs (including 9 now known to represent exons) were cloned in both orientations upstream of the firefly luciferase reporter gene and SV40 promoter (Figure V.13). To determine whether cloned ECRs have enhancer activities dual luciferase reporter assays were performed following transient transfection of vector constructs in HepG2 cells (Figure V.12). Each transfection is normalised by the co-transfection of a renilla luciferase expressing plasmid (pRL-CMV, Promega). Constructs containing control fragments described in section 5.3.1 were used to validate the reporter assay but for reasons of economy and scale were not included in every experiment. Each experiment included the positive controls: pGL3-Control (containing SV40 promoter and SV40 enhancer sequence), and human and mouse known endodermal enhancers (hsEE1 and mmEE1, respectively). The negative controls were pRL-null (contains no firefly luciferase reporter gene), to ensure that plasmid DNA mini-preparations were free from contamination and to check that no bleed-through of light occurs between wells of the assay plate. Additionally each experiment included the pGL3-Promoter empty vectors containing the Gateway® cassette in both orientations. For each construct transfected average firefly/*renilla* luciferase ratios, from four technical replicates,

were normalised against the orientation-matched empty vector and plotted on a Log2 scale (Figure V.15). pGL3-Control, hsEE1 and mmEE1 constructs reproducibly gave greater than 16-fold (Log2 >4) enhancement of firefly luciferase expression compared with the empty vectors. Indeed, the endodermal enhancers perform almost as well as the SV40 enhancer with matched promoter. Furthermore, the mouse enhancer sequence works well in the human cell-line, demonstrating conservation of enhancer function in spite of approximately 90 Myr of parallel evolution (Figure V.15). Nine of the 39 ECRs (4, 7, 11, 12h, 17h, 21h, 22h, 26 and 36) demonstrate a 2-fold or greater enhancer activity in at least one orientation. ECRs#21h and 22h are physically separated by only 75bp in human and therefore may represent a single functional entity. ECR#26 was the most potent novel enhancer element identified in this study and is further characterised below.

With the ECR and randomer luciferase ratio datasets from enhancer assays we can ask the question; does the ECR method enrich strongly for enhancers over randomly selected sequence? In the forward orientation the two-tailed P-value (from an unpaired t-test) equals 0.0235 which is considered statistically significant. In the reverse orientation the P-value equals 0.0724 which is not quite statistically significant, by conventional criteria. However, if we combine the forward and reverse data then there is a very significant statistical difference between the enhancer activities of ECRs and randomers (P=0.0047). Therefore, despite the fact that I am only assaying for enhancer activity in one cell-line and by chance alone, at least one of the randomers (23m) has enhancer activity in HepG2 cells, there is a clear enrichment for enhancer activity using the ECR method. By testing ECRs for enhancer activities in different cell lines or including assays for other functions we might predict even greater discrimination between evolutionary conserved and random sequences to detect function.

Two ECRs (ECR#14 and #19h) demonstrated apparent silencer activity in the forward orientation only (Figure V.15). Since the experimental system described here was designed for testing enhancer activities no suitable controls for silencer activities were included. Therefore, further work will be required to determine conclusively whether functional silencers can be detected in this way.

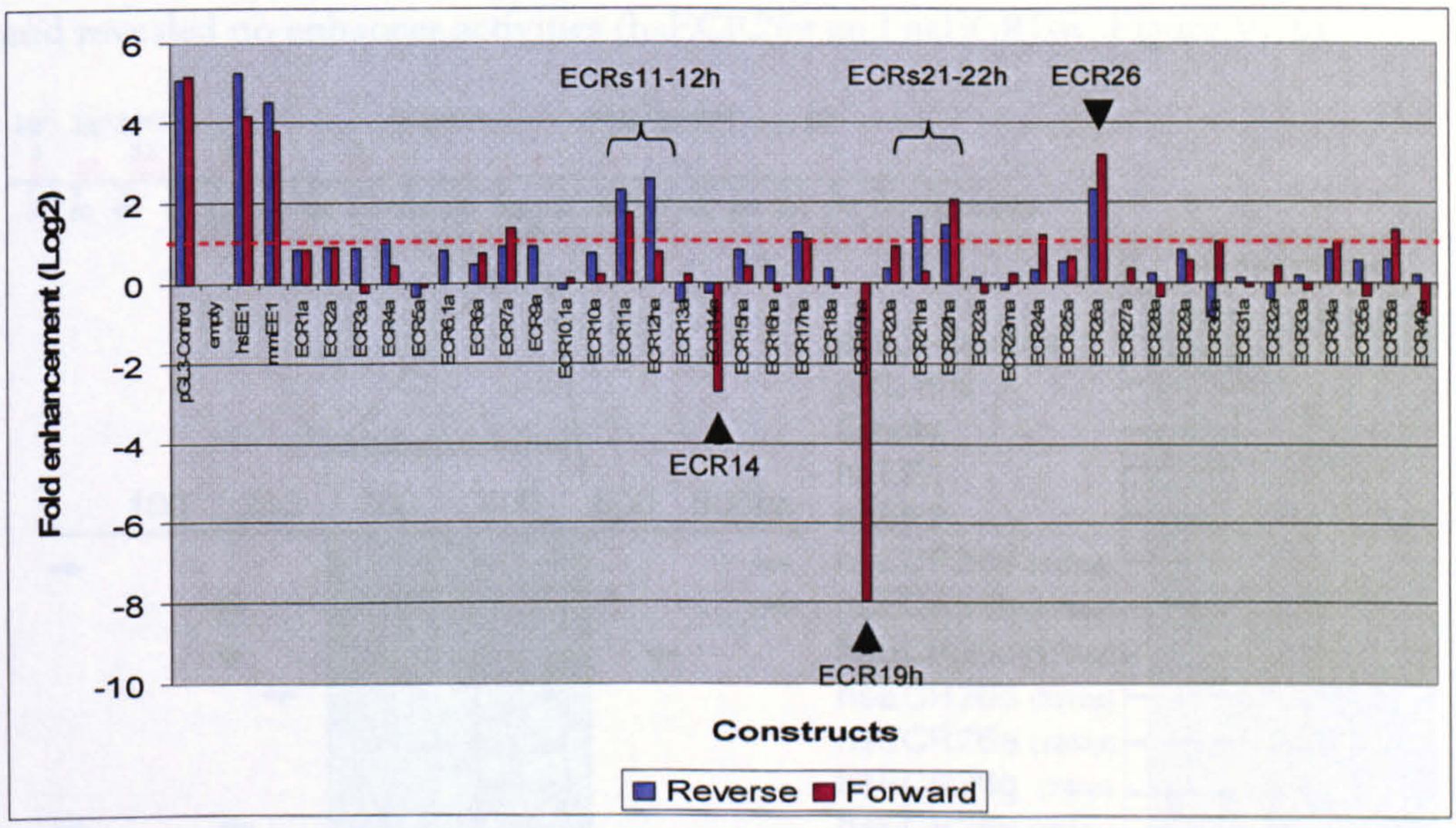


Figure V.15. Testing 11p15.5 ECRs for enhancer activity in human HepG2 cells.

For each construct transfected average dual luciferase ratios were normalised against empty pGL3-Promoter vectors in forward (claret) and reverse (blue) orientations. The fold enhancement over empty vectors is plotted on a Log2 scale. Therefore, bars with values above 0 show relative enhancement over empty vectors and values below indicate a decreased activity (possible suppression). The red dotted line represents a 2-fold enhancement (Log2 of 1.0).

5.3.5 Identifying a core enhancer element (ECR26)

As described above ECR26, lying equidistantly between *TH* and *ASCL2* genes, gave the most potent enhancer activities in HepG2 cells. To further characterise the functional domain of this ECR progressively smaller fragments were cloned and

assayed for enhancer activity (Figure V.16). The originally cloned ECR (hsECR26a, 647bp) contained the 231 bp conserved sequence between human and wallaby together with flanking genomic sequence. Enhancer activity was maintained down to a 73bp core sequence lying within the conserved sequence (hsECR26g, Figure V.16). To demonstrate that enhancer function is determined by this 73bp conserved sequence, cloned fragments flanking the ECR (but within hsECR26a) were assayed and revealed no enhancer activities (hsECR26x and hsECR26y, Figure V.16).

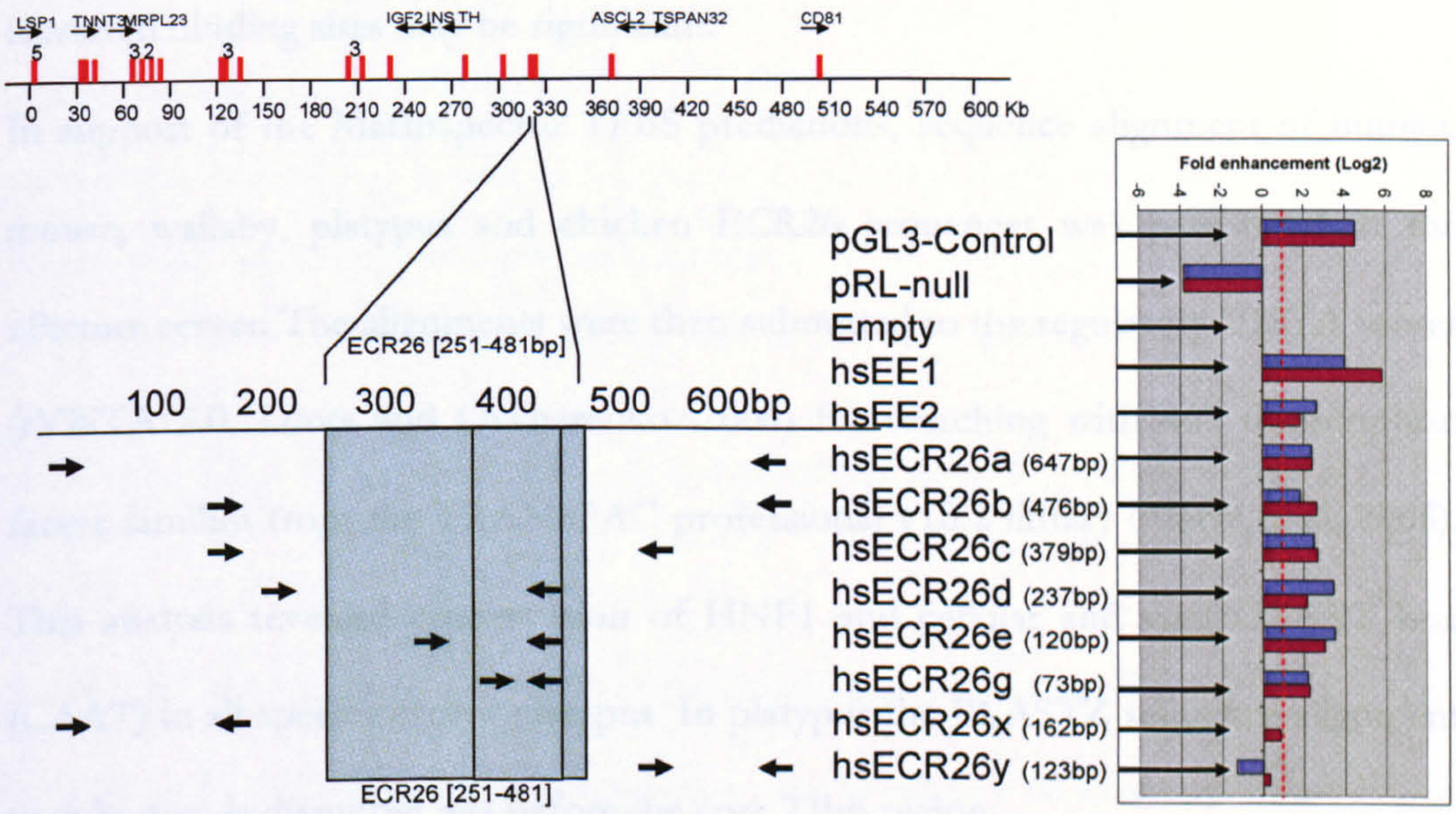


Figure V.16. Identifying a core enhancer element.

The ECR demonstrating highest enhancer activity above (hsECR26a) was serially cloned in progressively smaller fragments (hsECR26b-g) to establish the core functional liver enhancer. The conserved sequence between human and wallaby is indicated by the blue shading. Black arrows within and flanking the blue area represent PCR primers. The bar chart shows enhancer activities of transiently transfected constructs in both forward (blue) and reverse (claret) orientations. The red dashed line marks a two-fold (Log2 of 1) enhancement over the negative control (Empty, see text for details). Cloned fragments not containing conserved sequence (x and y) show no enhancer activity.

Given that the enhancer activity of ECR26 was localised to a 73 bp sequence (hg18chr11:2189129-2189201) I next sought to identify potential TFBSs within this sequence. Using MatInspector software from the Genomatix server (Cartharius et al. 2005) 15 predicted TFBSs were identified and include a cluster of 6 core binding motifs (from longer TFBS sequences) between bases 43-51 of the 73 bp enhancer sequence (Figure V.17). Since the enhancer activity of ECR26 was demonstrated in HepG2 cells the predicted binding of hepatic nuclear factor 1 (HNF1) within the clustered binding sites may be significant.

In support of the MatInspector TFBS predictions, sequence alignment of human, mouse, wallaby, platypus and chicken ECR26 sequences was performed at the zPicture server. The alignments were then submitted to the regulatory VISTA server (rVISTA 2.0, Loots and Ovcharenko. 2004) for searching with 467 transcription factor families from the TRANSFAC professional v10.2 library (Matys et al. 2006). This analysis revealed conservation of HNF1 and cellular and viral CCAAT box (CAAT) in all species except platypus. In platypus the BLASTZ sequence alignment with human is disrupted just before the core 73bp region.

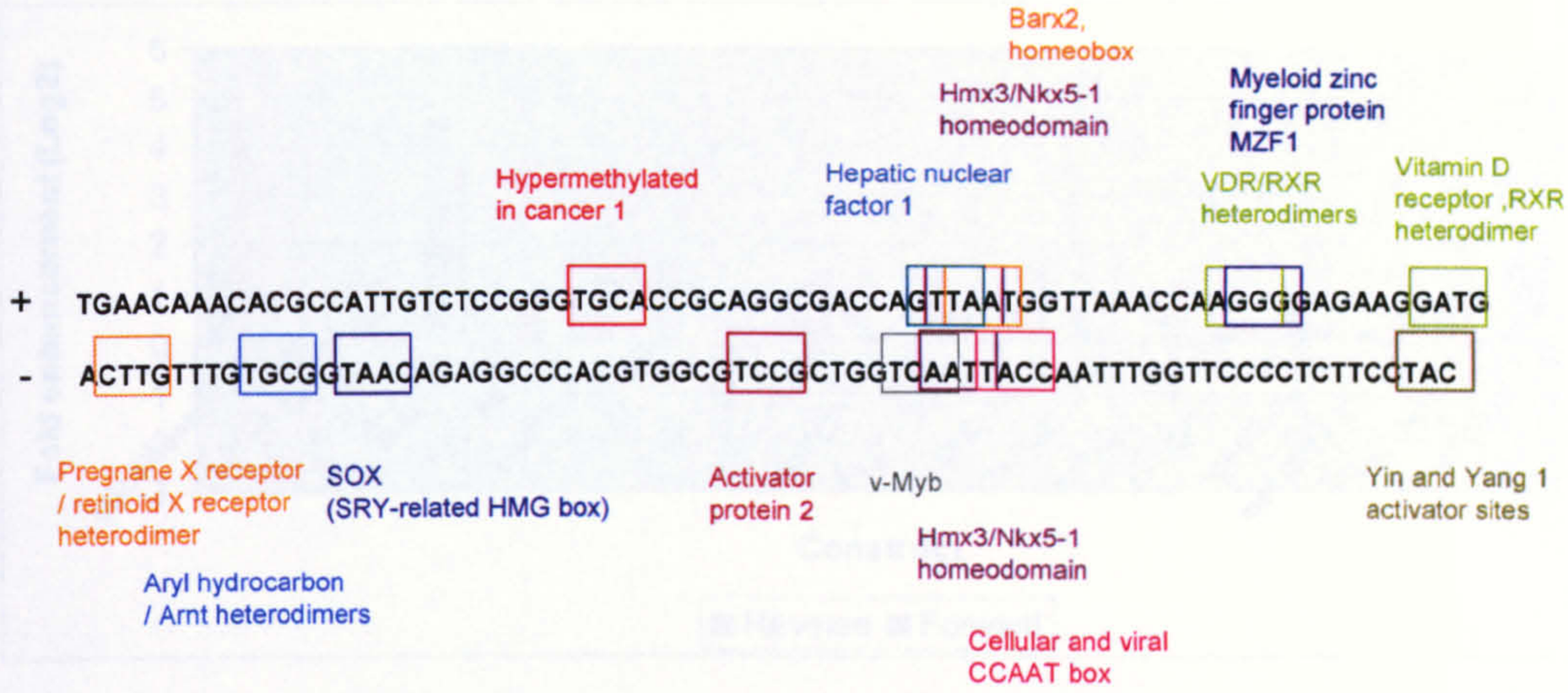


Figure V.17. Predicted TFBSs in the 73bp core enhancer region of ECR26.

Sense (+) and anti-sense (-) strands of the DNA sequence are shown. Coloured boxes indicate the core (4 bp) most highly conserved and consecutive sequence matches from longer TFBS sequences predicted using Genomatix MatInspector software (Cartharius et al. 2005)**. The colour matched transcription factors are also displayed.**

5.3.6 Testing wallaby ECRs for enhancer activity in human HepG2 cells.

With 23% (9/39) of tested 11p15.5 human ECRs demonstrating enhancer function in human HepG2 cells it was of interest to ask whether enhancer function can also be ascertained from the equivalent wallaby sequences for these ECRs tested in human cells. Seven ECRs (numbers 4, 7, 11, 12h, 17h, 21-22h and 26) showing at least 2-fold enhancer activity (in at least one orientation) in the human system were cloned from wallaby genomic DNA and transiently transfected into human HepG2 cells. Dual luciferase reporter assay results for human and wallaby cloned ECRs are shown in Figure V.18. In at least one orientation (both for ECR#26) three of the seven wallaby ECRs (12h, 17h and 26) reach the conservative threshold of 2-fold enhancement over background demonstrating that it is possible to detect cross-species enhancer effects despite the 148 Myr evolutionary separation between human and wallaby.

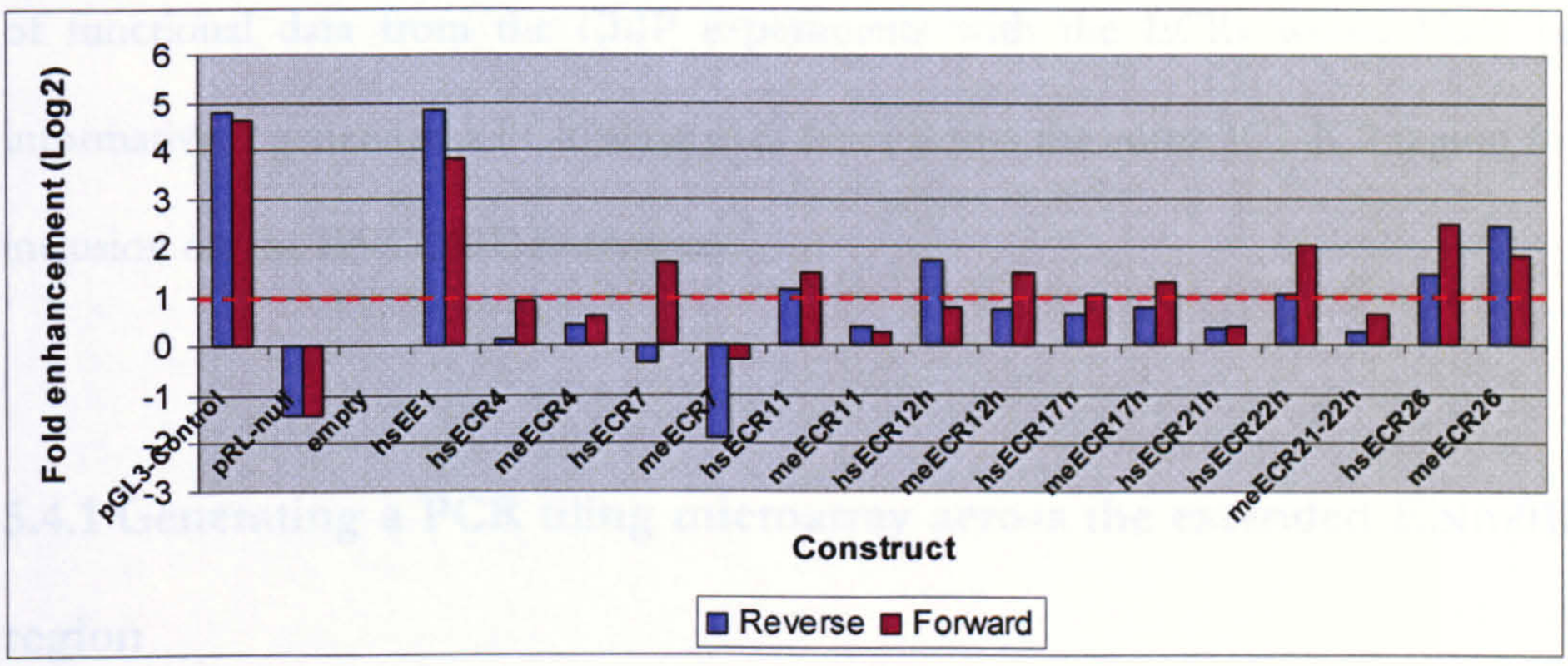


Figure V.18. Enhancer activities of wallaby ECRs in human HepG2 cells.

For each construct transfected average dual luciferase ratios were normalised against empty pGL3-Promoter vectors in forward (claret) and reverse (blue) orientations. The fold enhancement over empty vectors is plotted on a Log2 scale. Therefore, bars with values above 0 show relative enhancement over empty vectors and values below indicate a decreased activity (possible suppression). The red dotted line represents a 2-fold enhancement (Log2 of 1.0). hs, *Homo sapiens* (human); me, *Macropus eugenii* (wallaby).

5.4 Correlating epigenetic features with ECRs across the 11p15.5 region

As discussed in chapter III the IC1 domain lies within the ENCODE region ENm011 (hg18chr11:1699992-2306039). This 606 kb region was targeted because of the wide interest in reciprocally imprinted genes *H19* and *IGF2*. The ENCODE group at the Sanger Institute (Christoph Koch, Gayle Clelland and Sarah Wilcox and the microarray facility) have generated PCR tiling path microarrays across each of the 44 regions comprising approximately 30 Mb (1%) of the human genome. ChIP using antibodies specific for a variety of histone modifications and the insulator-associated protein CTCF was performed and the enriched DNA hybridised to the ENCODE microarray (Koch et al. 2007). As there was space on the ENCODE microarrays for additional features (PCR products), and correlation

of functional data from the ChIP experiments with the ECRs would likely be informative, I generated a PCR tiling path from across the entire IC1-IC2 region for inclusion on the ENCODE microarray.

5.4.1 Generating a PCR tiling microarray across the extended ENm011 region

To assist the ENCODE group, whilst making use of their ChIP-chip data, I generated a PCR tiling microarray spanning not only the ENm011 region but an additional 1.3 Mb region of 11p15.5 encompassing the whole of the IC2 domain. The total length of this ENm011_EXTENDED region was 1,922,276 bp and spanned the genes *CTSD* to *ART5* (hg18chr11:1699992-3622267). Approximately 1.4 kb minimally overlapping amplicons were designed, using PRIMER 3.0 and including repetitive sequence where possible, by Rob Andrews (Microarray bioinformatics department, Sanger Institute). After an initial round of PCR primer design, using pre-determined length and melting temperature parameters, a subsequent round of design was performed allowing for smaller 'gap filling' amplicons. In total the ENm011_EXTENDED region was spanned by 1148 amplicons. Following PCR from human genomic DNA (see chapter II) 1005 (87.5%) of the tiles were successfully amplified representing 71% of the entire region or 86% coverage of the non-repetitive region (Table V-4). These PCR products were supplied, together with all other ENCODE region products, to the Sanger microarray facility for immobilisation on CodeLink (GE Healthcare) glass slides via 5' aminolinks incorporated in the forward primer in each PCR product (Dhami et al. 2005). Slides were processed to generate single-stranded array features, as described at http://www.sanger.ac.uk/Projects/Microarrays.

Table V-4. Features of the ENm011_EXTENDED microarray.

Description	Feature
Number of bases in region (bp)	1922240
Number of non-repetitive bases (bp)	1209264
Non-repetitive DNA in region (%)	62.9
Number of bases covered in tiles (bp)	1563040
Tile coverage of whole region (%)	81.3
Number of non-repetitive bases covered in tiles (bp)	1183581
Tile coverage of non-repetitive region (%)	97.9
Number of PCR amplicons designed	1148
Number of successfully PCR amplified products	1005 (87.5%)

5.4.2 ChIP-chip experiments

ChIP-chip experiments were performed by the Sanger Institute ENCODE group using antibodies for 8 histone modifications (Table V-5) and CTCF. For the ENm011_EXTENDED region histone modification (Figure V.19) and CTCF binding (Figure V.20) profiles were obtained for GM06990 (lymphoblastoid) cells. Data from only one biological replicate for the histone modifications was available for analysis but within the ENm011 region the profiles were identical to those obtained previously for multiple replicates (Koch et al. 2007). Three biological replicates were available for CTCF (Figure V.20).

Table V-5. Histone modifications tested across the ENm011_EXTENDED array.

Histone modification	Antibody used	Function (epigenetic mark for)	Reference(s) to function defined in the third column
H3 acetylation (K9/14)	06-599, Millipore	Transcriptional activation (active promoters)	(Roth et al. 2001)
H4 acetylation (K5/8/12/16)	06-866, Millipore	Transcriptional activation	(Schiltz et al. 1999)
H3K4 mono-methylation	ab8895, Abcam	Enhancer signature	(ENCODE Project Consortium et al. 2007, Heintzman et al. 2007)
H3K4 tri-methylation	ab8580, Abcam	Transcriptional activation (active promoters)	Reviewed in Vakoc et al. 2006
H3K9 tri-methylation	ab8898, Abcam	Transcriptional repression (throughout transcript)	Reviewed in Vakoc et al. 2006
H3K27 tri-methylation	07-449, Millipore	Polycomb repression	Reviewed in Vakoc et al. 2006
H3K36 tri-methylation	ab9050, Abcam	Transcriptional activation (transcription elongation – 3' ends)	Reviewed in Vakoc et al. 2006
H3K79 tri-methylation	ab2621, Abcam	Transcriptional activation (telomeric silencing)	Reviewed in Vakoc et al. 2006

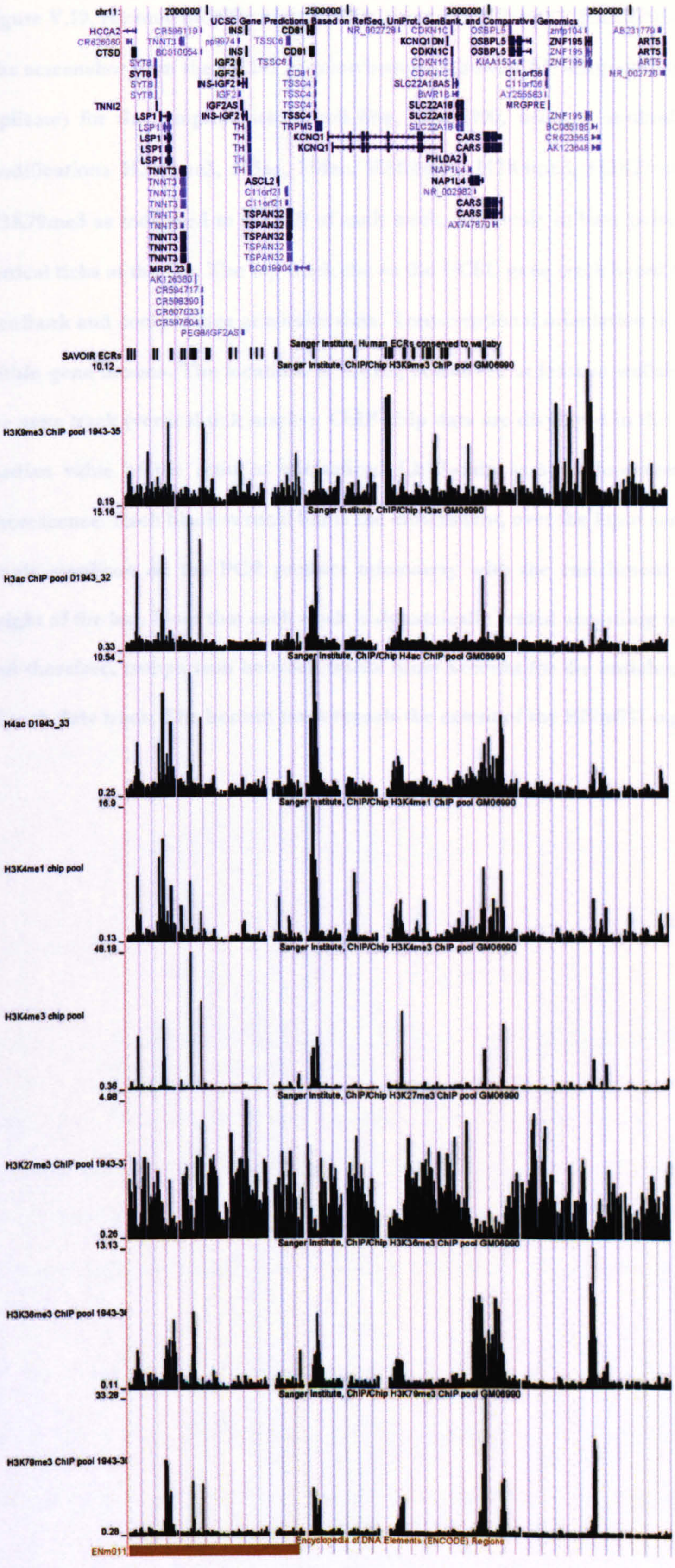

chr11:
2000000
2500000
3000000
3500000
UCSC Gene Predictions Based on RefSeq, UniProt, GenBank, and Comparative Genomics
Sanger Institute, Human ECRs conserved to wallaby
SAVOIR ECRs
Sanger Institute, ChIP/Chip H3K9me3 ChIP pool GM06990
H3K9me3 ChIP pool 1943-35
Sanger Institute, ChIP/Chip H3ac GM06990
H3ac ChIP pool D1943_32
Sanger Institute, ChIP/Chip H4ac ChIP pool GM06990
H4ac D1943_33
Sanger Institute, ChIP/Chip H3K4me1 ChIP pool GM06990
H3K4me1 chip pool
Sanger Institute, ChIP/Chip H3K4me3 ChIP pool GM06990
H3K4me3 chip pool
Sanger Institute, ChIP/Chip H3K27me3 ChIP pool GM06990
H3K27me3 ChIP pool 1943-3
Sanger Institute, ChIP/Chip H3K36me3 ChIP pool GM06990
H3K36me3 ChIP pool 1943-3
Sanger Institute, ChIP/Chip H3K79me3 ChIP pool GM06990
H3K79me3 ChIP pool 1943-3
Encyclopedia of DNA Elements (ENCODE) Regions
ENm011

Figure V.19. Histone modification profiles across the ENm011_EXTENDED region.

The screenshot from the UCSC genome browser shows ChIP-chip data (from one biological replicate) for the lymphoblastoid cell line, GM06990, using 8 antibodies for the histone modifications H3K9me3, H3ac, H4ac, H3K4me1, H3K4me3, H3K27me3, H3K36me3 and H3K79me3 as indicated to the left of each track. The scale in base pairs is indicated by the vertical ticks at the top. The top track shows the UCSC gene track based on RefSeq, Uniprot, GenBank and comparative genomics data. Transcriptional orientation is indicated by arrows within gene introns. The locations of ECRs, conserved at least to wallaby, are shown below the gene track (vertical tick marks). ChIP-chip data are displayed in the next 8 tracks as the median value of the ratio of normalized ChIP-chip sample fluorescence to input DNA fluorescence. Each black vertical bar is the enrichment, over the input sample, measured at a single amplicon on the PCR product microarray with the enrichment represented by the height of the bar. Note that each track is dynamically scaled according to the data displayed and therefore, comparison between tracks must account for the enrichment scale at the left of each data track. The bottom track reveals the extent of the ENm011 region (brown bar).

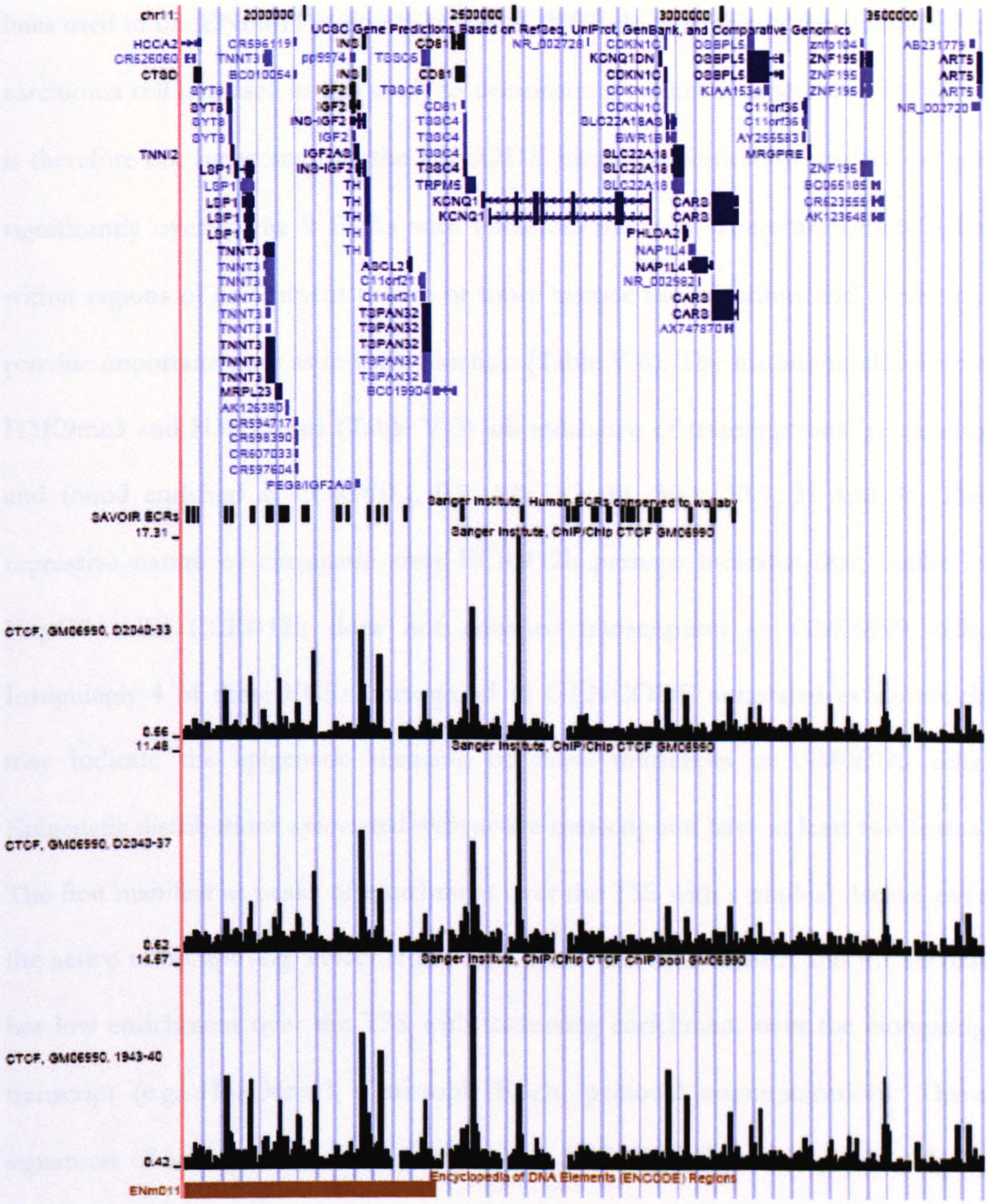


Figure V.20. CTCF profiles across the ENm011_EXTENDED region.

The screenshot from the UCSC genome browser shows ChIP-chip data for the lymphoblastoid cell line, GM06990, using an antibody for CTCF. The data from three biological replicate experiments are shown. The legend is otherwise the same as in Figure V.19.

It has been shown that enhancer elements are frequently cell type-specific (Heintzman et al. 2007, Xi et al. 2007) and therefore predominantly do not overlap with ubiquitous or common sites of open chromatin. The common ENCODE cell

lines used in the ENm011 region (Koch et al. 2007) do not include the HepG2 liver carcinoma cell line used in this study to demonstrate enhancer function of ECRs. It is therefore not surprising that the ENCODE histone modification profiles do not significantly overlap the 9 ECRs with enhancer function. There are other ECRs within regions of enrichment to one or more histone modifications and these may provide important clues as to their function (Table V-6). The histone modifications H3K9me3 and H3K27me3 (Table V-5) are indicative of transcriptional repression and found enriched at ECRs#0.1, 0.2, 12h, 13, 14, 30.1, 30.5, 31 and 35. The repressive nature of chromatin over ECR#12h perhaps indicates that, unlike in HepG2 cells, ECR#12h does not enhance transcription in GM06990 cells. Intriguingly 4 of these ECRs correspond to GENCODE annotated exons which may indicate the epigenetic silencing of these transcripts in GM06990 cells. Epigenetic distributions associated with active transcription have at least two forms. The first manifest as peaks of enrichment over the TSS with a gradual decline over the active transcript (e.g. H4ac, H3ac, H3K4me1 and H3K79me3) and the second has low enrichment over the TSS with increasing enrichment over the elongating transcript (e.g. H3K36me3, Christoph Koch, personal communication). These signatures of active chromatin are observed at ECRs#0.5, 0.6, 0.7, 8, 9, 10, 29, 34, 39.1, 39.2, 39.3 and 40. Specifically the H3K4me1 enhancer signature identifies ECRs#0.5, 0.6, 0.7 and 34 as potential enhancer elements in GM06990 cells.

Table V-6. Assigning probable function to the ECRs.

Probable human function	Evidence	ECRs from Table V-1
Novel gene	Mouse mRNA and human RACE PCR	1-5
Alternative coding exon	Overlap with current annotation (e.g. Gencode)	0.1, 0.2, 0.7, 6, 6.1, 7, 10, 13, 14, 29, 40
Promoters	H3K4me3	None
Endodermal enhancers	Reporter assays	4, 7, 11, 12h, 17h, 21h, 22h, 26, 36
Other enhancers	Enhancer signature (e.g. H3K4me1)	0.5, 0.6, 0.7, 34
Insulators	Overlap with CTCF binding	0.4, 25
Unknown		0.3, 8, 9, 10.1, 15h-16h#, 18, 19, 20, 23m#, 24, 28, 30, 30.1, 30.2, 30.21, 30.3, 30.4, 30.5, 31, 32, 33, 35, 37, 37.1, 38, 39, 39.1, 39.2, 39.3, 39.4

#These ECRs have DNaseI hypersensitivity sites.

Although the histone modification ChIP-chip data from the GM06990 cell line does not identify much overlap with those ECRs demonstrating enhancer activities in HepG2 cells, there is some ChIP-chip data for HepG2 accessible from the UCSC genome browser. A group at the University of Uppsala have performed ChIP-chip across the ENCODE regions to study the binding of transcriptional activator factors; forkhead box A2 (FOXA2 or HNF3b), hepatocyte nuclear factor 4, alpha (HNF4a) and upstream transcription factor 1 (USF1) in HepG2 cells. The same group also mapped sites of enrichment of H3 acetylation in HepG2 cells (Rada-Iglesias et al. 2005). Only three peaks of enrichment were reported in the ENm011 region. However, one of these, with HNF3b enrichment, encompasses ECR#26, the strongest enhancer identified here. The Crawford and Collins groups at Duke University and the National Health Genome Research Institute, respectively, have also used HepG2 cells to map DNaseI HS sites (Crawford et al. 2006). Inspection of these HepG2 datasets, within the UCSC genome browser, reveals that 7 out of 38 ECRs mapping to the ENm011 region overlap with DNaseI HS sites. Significantly, 5 of these 7 ECRs (ECR#12h, 17h, 21h, 22h and 26) gave over 2-fold

enhancement in the reporter assays above. The two ECRs with no apparent HepG2 enhancer activities are ECR#16h, which lies only 147 bp from ECR#17h, and ECR#23m which also overlaps with a FAIRE signal from a human foreskin fibroblast cell line (2091). Although these two ECRs do not appear to be enhancer elements in the HepG2 reporter assay, the multiple lines of experimental evidence support a functional role.

CTCF binding is associated with enhancer blocking function of insulator elements (Bell et al. 1999). The CTCF ChIP-chip data presented here (Figure V.20) confirms previous findings of CTCF binding to the *H19* upstream differentially methylated domain critical to the imprinting of *IGF2* (discussed in detail in chapter VI). Only ECRs#0.4 and 25 overlap CTCF binding sites which may indicate a role in insulation.

To summarise, of the 65 non-repetitive ECRs studied here 16 overlap recently annotated exons. 9 of 39 ECRs tested in *in vitro* enhancer assays display strong enhancer activity in HepG2 cells and a further 4 ECRs have the H3K4me1 enhancer signature in GM06990 cells. Two of the ECRs indicate an insulator function in GM06990 cells and the function of 30 ECRs (46%) remains enigmatic.

5.5 Discussion

This chapter has described the identification, using comparative sequence analysis, of candidate regulatory elements in the region of human chromosome 11 (band p15.5) that harbours IC1 and IC2 imprinting domains. Enhancer activity of cloned ECRs was demonstrated in HepG2 cells for 23% of the tested ECRs using dual luciferase reporter assays. Furthermore, enhancer activity was demonstrated to be evolutionarily conserved through the testing of wallaby sequences in human cells. The functional enhancer element in ECR#26 was localised to a 73bp sequence

containing conserved binding sites for the TFs HNF1 and CAAT. Finally, epigenetic profiles of histone modifications and TF binding sites were correlated with ECR locations to gain further insight into the potential function of these sequences. Despite the limitations imposed by studying different cell lines a general epigenetic and regulatory profile of the ENm011 extended region is emerging.

Missing data and/or assembly mistakes in draft quality sequences will inevitably result in alignment gaps. In some cases this will result in the loss of biological information. For this reason gap-free finished sequence has been generated for each species studied. Comparison of identified ECRs with pre-computed ECRs in the ECRbase database (Loots and Ovcharenko. 2007) reveal that only 11 of the 66 (17%) ECRs conserved, at least, between human and wallaby are identified between human and opossum (monDom1) sequence alignments. The low correspondence between human-wallaby and human-opossum non-coding ECRs likely reflects the poor coverage of the draft opossum genome sequence in this region (see chapter III) and illustrates the importance of finished sequence.

No ultra-conserved elements, defined as 100% identical over 200 bp or more in human and rodent alignments (Bejerano et al. 2004, Woolfe et al. 2005) or coreECRs (350bp, 77% identity, Ovcharenko et al. 2004b) were identified in the 11p15.5 region. These elements have been associated with regulatory elements controlling transcription of key vertebrate developmental genes (Woolfe et al. 2005). The 11p15.5 region is notably devoid of such key developmental genes or master regulatory elements.

The approach adopted here has identified novel endodermal enhancer elements and importantly, the set of highly conserved sequences is significantly enriched for

enhancer function (P<0.005) compared with length and C+G content matched random sequences. The identification of ECRs, cloning into gene reporter assays and transfection of cell-lines are all scalable and therefore this strategy could be widely adopted. It is interesting that, by chance, randomer 23m does represent a novel enhancer despite no other functional and limited conservational data for this sequence. This implies that with sufficient resources it may be valid to clone sequences representing a tiling path across entire regions or indeed genomes in the search for regulatory elements. The enhancer experiments reported here were all performed using HepG2 cells because of the known expression patterns of genes in the IC1 domain. However, to identify all enhancer elements it will be necessary to test the ECRs in multiple cell lines containing required TFs for the correct spatial and temporal expression of genes active in those cell-lines. To test tiling paths of cloned sequences in multiple cell-lines would be an enormous undertaking with many sequences (possibly 90-97.5% based on the randomer data here) not reporting function.

Epigenetic techniques are now available to aid in deciphering the regulatory code of transcription including the identification and characterisation of interactions between TFs, their co-factors and DNA in its native chromatin structure. However, each method has its limitations. The study of TF and co-factor binding using ChIP-chip requires not only prior knowledge of the proteins but ChIP-grade (i.e. highly specific recognition of an epitope in free solution) antibodies. DNaseI HS site mapping, whilst indicating regions of open chromatin, does not inform us which regulatory elements are present. Finally, expression arrays can be used to address which genes are expressed in a given cell type. However, the factors bringing about such cell-specific expression are unknown using this technology. Possibly the best approach to build-up a complete picture of transcriptional regulation is one

integrating these epigenetic methods, together with analyses of evolutionary conserved sequences.

For 1% of the genome this has been done through the ENCODE pilot project (ENCODE Project Consortium et al. 2007) and efforts are now underway to apply these technologies to the remainder of the human genome. It should be noted that even genome-wide experiments are only testing a limited number of cell types at specific developmental time points and environmental conditions. More cell-lines, or better still primary cells, from different developmental stages and under varied environmental conditions will be required to complete our understanding of transcriptional regulation.

Perhaps one of the most surprising findings to come from the ENCODE pilot project was that many (approximately 50%) non-coding functional elements do not appear to be evolutionarily constrained (ENCODE Project Consortium et al. 2007). To a small degree this might reflect the underestimation of sequence constraint, currently approximately 5% of the human genome, or the overestimation of experimentally identified functional elements. Theories to account for the large proportion of unconstrained functional elements are discussed by the ENCODE consortium (ENCODE Project Consortium et al. 2007). The opposite is also true that many constrained sequences are not readily explained by functional elements. As noted above this might reflect the spatial and temporal limitations of functional assays used to date. Additionally there will undoubtedly be genome functions we have yet to recognise, let alone test for.

The ENCODE observation is not the first to reveal conservation of function without conservation of sequence. The receptor tyrosine kinase (*RET*) gene is expressed in several developmental tissues and both tissue and temporal expression

is highly conserved across vertebrate species. Non-coding regulatory elements 5' or within the intron 1 of the *RET* gene are conserved in mammals but noticeably absent from human-fish sequence alignments. However, when testing these human sequences in a transgenic zebrafish model upregulation of the *RET* gene was observed and furthermore reflected the endogenous tissue expression even in cell types not present in human (Elgar. 2006, Fisher et al. 2006). How this remarkable conservation of function is achieved is unclear. However, it does inform us that our knowledge of TF binding of DNA is at best incomplete and that we should take a more evolutionary neutral view of genome function, whilst striving to understand the sequence constraint. Despite the fact that not all functional elements will be detected through comparative sequence analysis it is undeniable that these approaches have, and will continue, to refine genome annotation.

Increasingly, experimental datasets are required to 'train' computational prediction algorithms as has been highly successful for protein-coding gene prediction. Although a significant challenge, the same reasoning applies for regulatory element identification. In this regard having the ECR and randomer datasets provides an opportunity to search for sequence motif over-representation. Thomas Down (Gurdon Institute, Cambridge) has used the following procedure to identify known motifs from the JASPER database (http://jaspar.genereg.net/). Using a scoring scheme that takes into account local mononucleotide composition, the best match to motifs in the JASPER database for each ECR and randomer sequence was calculated. Next, an empirical statistical test was performed to determine whether, or not, the distribution of motif scores differs significantly between the ECR (test) and randomer (control) sets. Four motifs appear to be enriched in the set of ECRs (Figure V.21). Prrx2 (paired related homeobox 2) and FOXD1 (forkhead box D1)

are highly significant (P<0.001) whereas the significance of TBP (TATA box binding protein) and SRF (serum response factor) is less certain (P<0.005). Simulations on the data, performed by Thomas, indicate that a P-value of approximately 0.001 equates to a false discovery rate of 5%. The enrichment of Prrx2 and FOXD1 in the ECR sequences therefore appears to be real but the specific functional role of these TFs is not known. There was no over-representation of the CTCF motif (Kim et al. 2007) associated with insulator function. This would suggest that, in general, the ECRs identified here do not correspond with boundary elements, consistent with the observation that only two ECRs overlap with CTCF-bound DNA from ChIP-chip studies. The importance of insulator function in the genomic imprinting mechanism is further discussed in chapter VI.

Figure V.21. Over-represented known motifs in the ECR set.

The overall height of a single stack of bases indicates the sequence conservation at that position. The height of each base within the stack represents the information content. P-values for enrichment of motifs in the ECR set compared with randomer set are given in brackets.

To address whether we can identify potentially novel sequence motifs enriched in the set of ECRs, the NestedMICA method was used (Down and Hubbard. 2005). Although the number of ECRs is relatively small for such an analysis, four novel

motifs appear to be over-represented (Figure V.22). These motifs do not match any known motifs and larger datasets will be required to validate them.

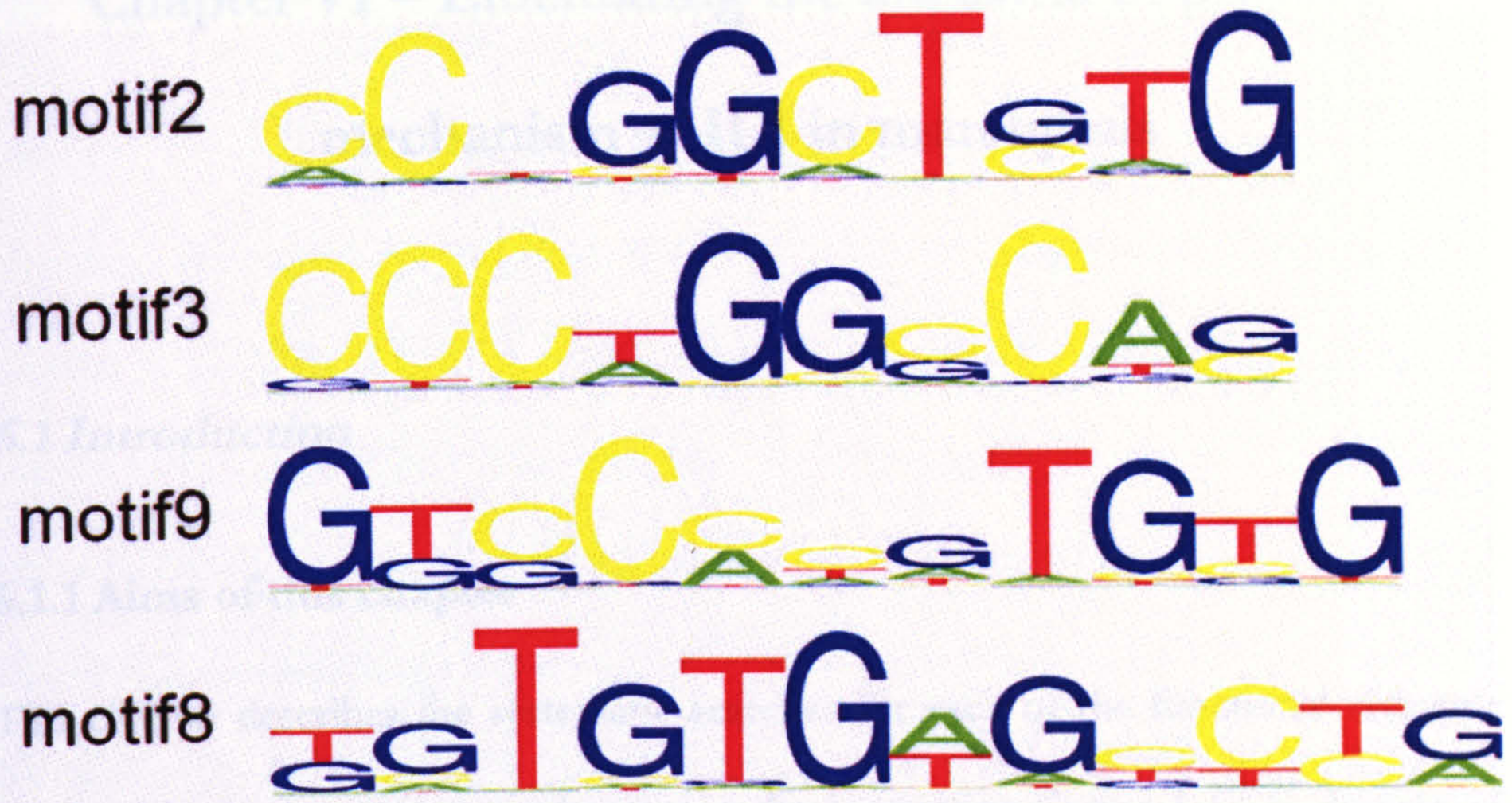


Figure V.22. Identifying novel sequence motifs over-represented in the ECR set.

The overall height of a single stack of bases indicates the sequence conservation at that position. The height of each base within the stack represents the frequency of the base at that position.

Chapter VI – Elucidating the ancestral imprinting mechanism at IC1 in marsupials

6.1 Introduction

6.1.1 Aims of this chapter

This chapter describes the systematic searches for each of the functional elements involved in the imprinting mechanism in the Imprinting Centre 1 (IC1) domain of wallaby. The results presented were obtained from a close collaboration between me and Guillaume Smits (Babraham Institute). This division of labour and the accompanying exchange of ideas greatly contributed to the generation of a more complete picture as described below. Specifically Guillaume performed the reverse transcription polymerase chain reaction (RT-PCR) amplification of wallaby samples and subsequent RACE experiments to determine the wallaby *H19* gene structure. SNPs identified by Guillaume in the wallaby H19 and *IGF2* genes were used by him to determine imprinting status. The characterisation of the wallaby *H19* differentially methylated region in a variety of tissues was also performed by Guillaume. All other experiments and bioinformatic analyses presented were performed by me.

To further our knowledge of the origin of genomic imprinting and the imprinting mechanism in ancestral mammals I have focused on the *H19/IGF2* or IC1 locus, located towards the telomere of the short arm of human chromosome 11 (11p15.5)

and mouse distal chromosome 7 (7qF5). The murine *Igf2* and *H19* genes were amongst the first imprinted genes to be identified (Bartolomei et al. 1991, DeChiara et al. 1991) and as a consequence the IC1 locus has been well characterized in eutherian mammals. In recent years there has also been a growing body of literature regarding the imprinting status of marsupial, monotreme and bird *Igf2* orthologues (Killian et al. 2001, O'Neill et al. 2000, Suzuki et al. 2005). The *Igf2* gene has been shown to be imprinted in the marsupial species *Monodelphis domestica* (grey short-tailed opossum, O'Neill et al. 2000) and *Macropus eugenii* (tammar wallaby, Suzuki et al. 2005) but biallelically expressed in the monotreme species *Ornithorhynchus anatinus* (duck-billed platypus) and *Tachyglossus aculeatus* (short-beaked echidna) as well as the birds such as *Gallus gallus* (chicken, Killian et al. 2001, O'Neill et al. 2000, Weidman et al. 2004). In contrast, the expression status of the long, non-coding RNA (hereafter termed macroRNA) *H19* gene in ancestral vertebrate species has not been elucidated as this non-coding RNA gene has yet to be identified in any non-eutherian species (Paulsen et al. 2005, Yokomine et al. 2005).

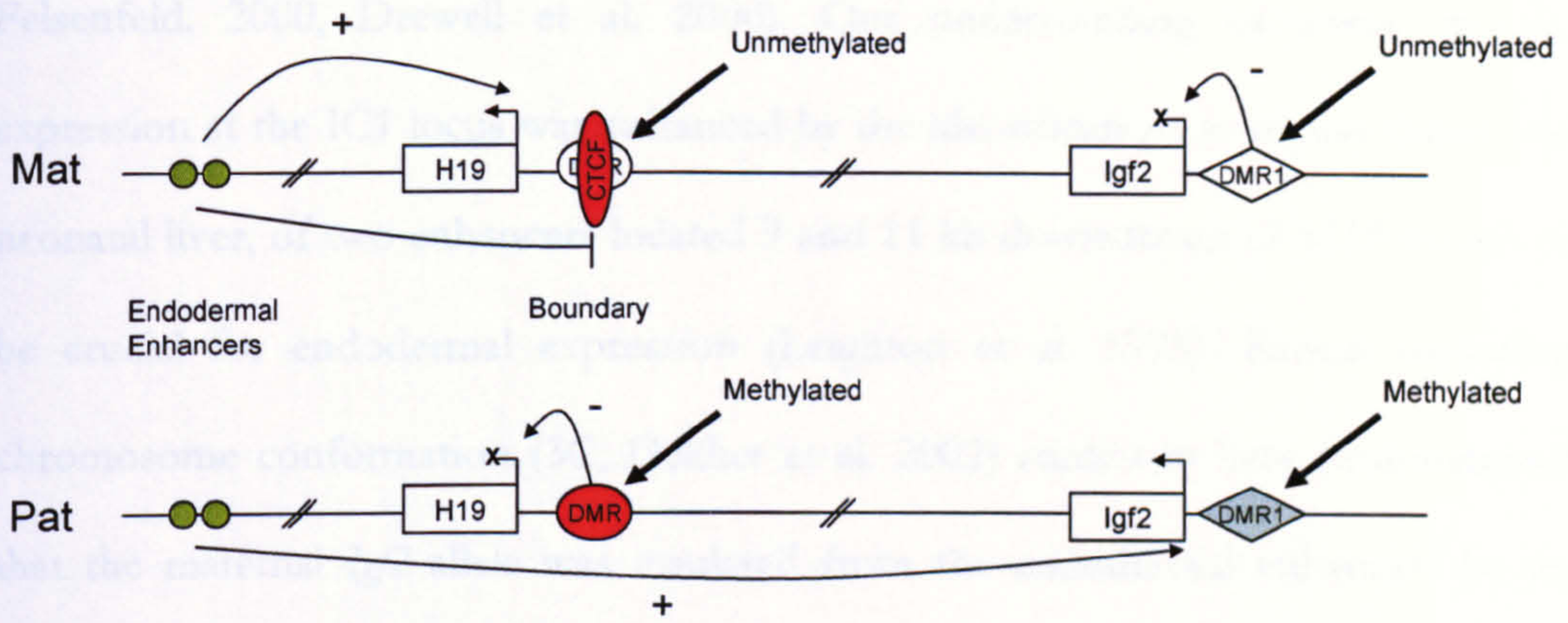


Figure VI.1. Boundary model of *Igf2/H19* gene regulation.

On the maternal allele the lack of methylation at a differentially methylated region (DMR) permits the binding of the insulator protein CTCF which acts as an enhancer blocker, preventing access of downstream endodermal enhancers to the *Igf2* promoter region. These enhancers are therefore able to drive expression of *H19*, a non-coding RNA of unknown function. On the paternal allele the DMR is methylated preventing CTCF binding and therefore permitting access of the enhancers to the *Igf2* promoter. The result is paternal expression of *Igf2*.

Whilst the integrity of the functional elements controlling *H19* transcription is essential for the correct imprinting of *Igf2* (at least in mice) the *H19* RNA itself is not, as illustrated by experiments to generate transgenic mice lacking the *H19* transcript (Jones et al. 1998, Thorvaldsen et al. 2002). The balancing paternal expression of *Igf2* and maternal expression of *H19*, in eutherian species studied to date, has given rise to the accepted 'boundary model' of imprinting at the IC1 locus (Arney. 2003, Hark et al. 2000 and Figure VI.1). In all tissues studied, eutherian *Igf2* imprinting is dependent upon a differentially methylated region (DMR) located 2 kb 5' of the *H19* gene. This 2.3 kb element (in mouse) is an unmethylated insulator on the maternal chromosome (resulting in *Igf2* insulation and *H19* expression) and a methylated silencer on the paternal chromosome (resulting in *H19* promoter repression and *Igf2* expression, Figure VI.1, Bartolomei et al. 1991, Bell and

Felsenfeld. 2000, Drewell et al. 2000). Our understanding of tissue specific expression at the IC1 locus was enhanced by the identification, in mouse foetal and neonatal liver, of two enhancers located 9 and 11 kb downstream of *H19* shown to be crucial for endodermal expression (Leighton et al. 1995). Recent capturing chromosome conformation (3C, Dekker et al. 2002) studies in liver demonstrated that the maternal *Igf2* allele was insulated from the endodermal enhancers by an epigenetic switch. The formation of chromatin loops caused by the binding of CTCF (CCCTC binding factor) protein to the unmethylated *H19* DMR results in the *Igf2* promoter residing in an inactive chromatin domain, insulated from the endodermal enhancers which lie with *H19* in an active chromatin domain (Kurukuti et al. 2006, Murrell et al. 2004). CTCF is a highly evolutionary conserved protein with 11-zinc finger domains and was found to bind to the prototypical chicken insulator in the β-globin locus (Bell et al. 1999). The precise means by which CTCF binding confers chromatin insulation and therefore segregates distinct regulatory domains is unclear. However, recent studies have shown that other proteins, such as nucleophosmin and CHD8 may bind to CTCF to generate a protein complex tethered to the nuclear matrix (Ishihara et al. 2006, Yusufzai and Felsenfeld. 2004). Binding to the nuclear matrix is thought to result in chromatin loops that may partition regulatory elements and thus prevent their interaction. Such a model in the IC1 locus was proposed by Murrell and colleagues (Murrell et al. 2004).

To gain further insights into the very early stages of genomic imprinting evolution and the gene regulatory mechanisms operating to bring about this unusual phenomenon it is important to study extant ancestral mammalian species in which genomic imprinting as been demonstrated. In order to establish the mechanism of imprinting in marsupial species it is therefore of critical importance to look for and

characterize the *H19* non-coding RNA and its regulatory elements in these species. Despite a report in 1996 claiming the fluorescence *in situ* hybridisation mapping of a phage clone containing tammar wallaby *H19* to chromosome 2p (Toder et al. 1996), sequence of this clone was not obtained (Jenny Marshall-Graves, personal communication). The recent whole genome shotgun draft assembly of the grey, short-tailed opossum also failed to reveal the presence of *H19*. To determine whether *H19* does exist in marsupial species BACs mapping to the IC1 orthologous region were sequenced in both tammar wallaby and opossum (chapters III and IV). As discussed in chapter V there is a striking conservation of both the order of the ECRs and the physical size of the inter ECR regions, particularly between human and wallaby, which permits the ECRs to serve as landmarks between the *Lsp1* gene and the 3' region of *Igf2* (Figure VI.2). In human and mouse, the region between ECR14 and ECR15 (39059 and 31162 bp, respectively) contains two endodermal enhancers (EE1 and EE2, Leighton et al. 1995, Yoo-Warren et al. 1988), the *H19* gene and the differentially methylated region (DMR) which are critical for the reciprocal imprinting of *Igf2* and *H19* genes (Figure VI.2). Unexpectedly, these features were not found by standard BLAST analyses in the finished orthologous wallaby ECR14-15 region sequence which is very similar in length (38186 bp) to the human interval. In addition these features were not readily identifiable in the whole wallaby BAC sequence (174698 bp, CR855994). The wallaby ECR14-15 region was therefore studied in fine detail for the presence of *H19* and associated functional elements.

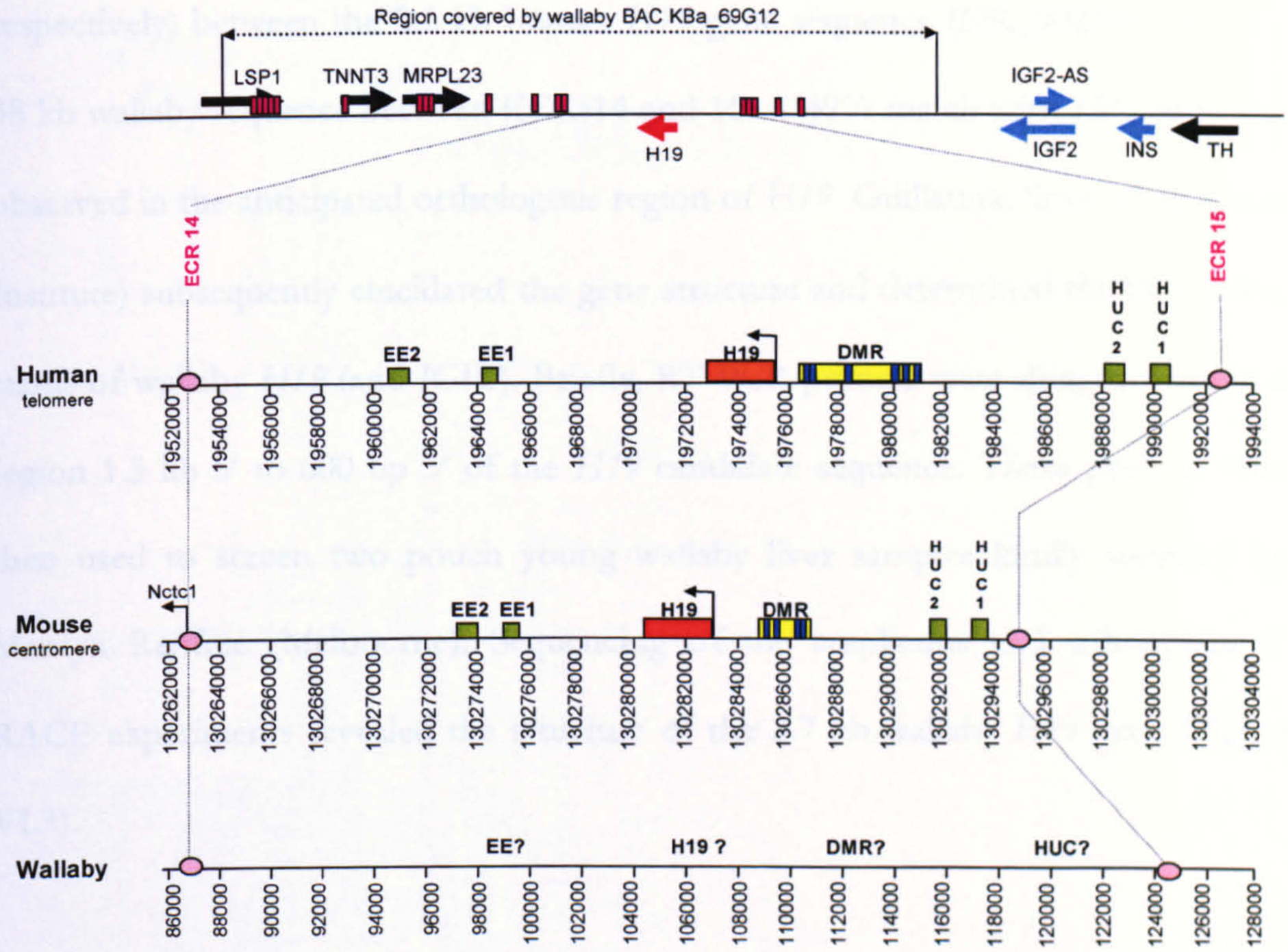


Figure VI.2. Sequence analysis in the *H19* region.

The human region of 11p15.5 encompassing the IC1 locus is presented from telomeric (left) to centromeric (right) end with the genes encoded on the sense strand above the line and those encoded on the antisense strand below the line. The genes depicted by black arrows have not been shown to be imprinted in any species (*LSP1*, *TNNT3*, *MRPL23*, *TH*), the genes depicted by blue arrows (*IGF2*, *IGF2-AS*, *INS*) are imprinted and paternally expressed (at least in some tissues). The maternally expressed *H19* gene is shown in red. The protein coding genes *LSP1*, *TNNT3* and *MRPL23* are all present in the wallaby BAC ME_KBa69G12. ECRs conserved between human, mouse and wallaby are illustrated by pink rectangles. The ECRs 14 and 15 define a region of approximately 40 kb in the wallaby BAC sequence where the *H19* gene and regulatory elements reside in human and mouse.

6.2 Identifying wallaby H19 and establishing its imprinting status

BLAST 2 sequences (BL2SEQ) analysis was performed at NCBI (http://www.ncbi.nlm.nih.gov/blast/bl2seq/wblast2.cgi) using high sensitivity (low specificity) parameters (word size of 10, match and mismatch penalties of +3 and -2

respectively) between the 2.3 kb human *H19* gene sequence (NR_002196) and the 38 kb wallaby sequence between ECRs14 and 15. A 49% match with 13% gaps was observed in the anticipated orthologous region of *H19*. Guillaume Smits (Babraham Institute) subsequently elucidated the gene structure and determined the imprinting status of wallaby *H19* (and *IGF2*). Briefly, RT-PCR primers were designed within a region 1.5 kb 5' to 600 bp 3' of the *H19* candidate sequence. These primers were then used to screen two pouch young wallaby liver samples kindly supplied by Marilyn Renfree (Melbourne). Sequencing of the amplicons and subsequent 5' RACE experiments revealed the structure of the 2.7 kb wallaby *H19* gene (Figure VI.3).

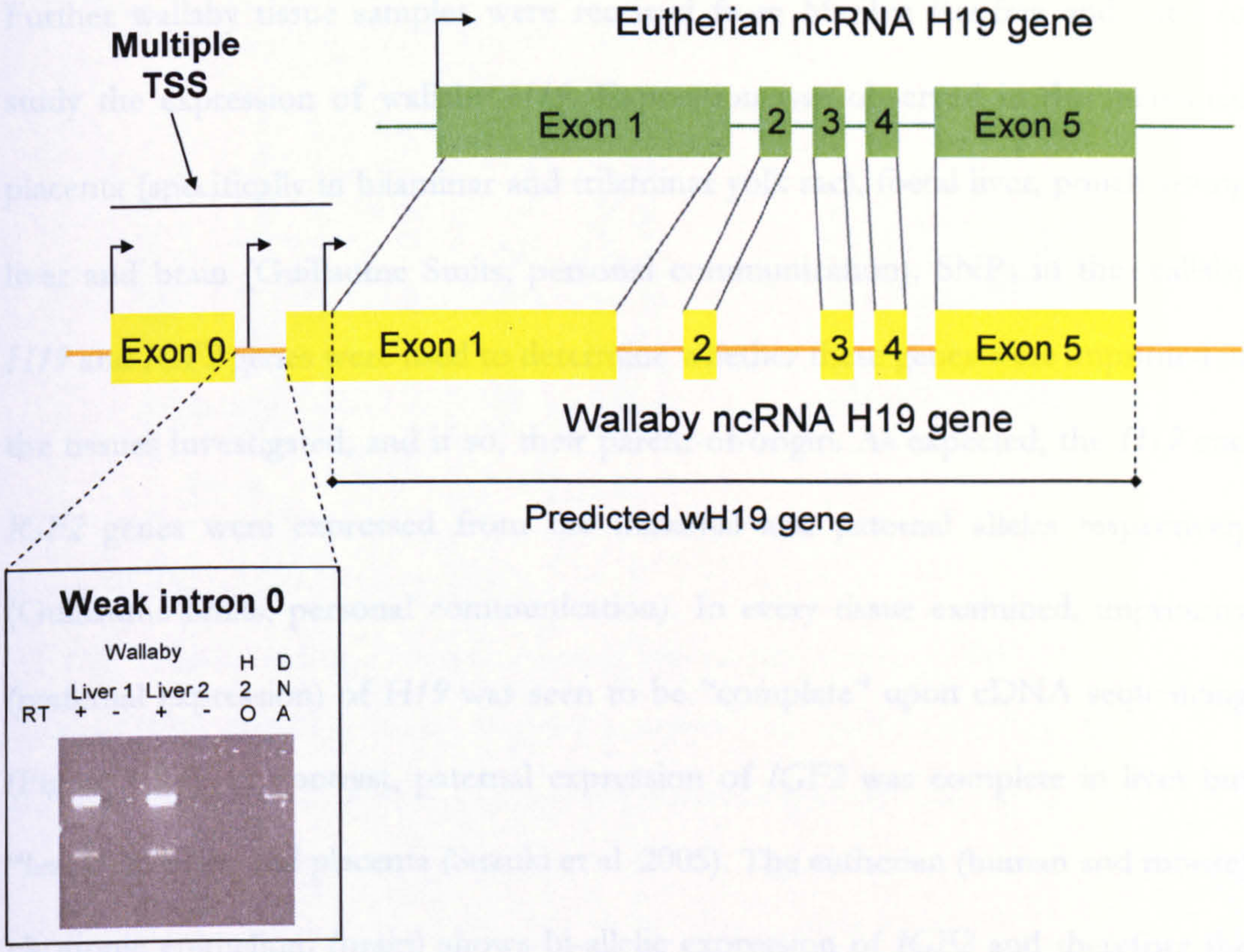


Figure VI.3. Elucidated structure of the wallaby *H19* gene.

Weak amplification products visible upon gel electrophoresis in both wallaby pouch young liver samples following Reverse Transcriptase (RT) PCR indicates the presence of an additional intron (intron 0) in the wallaby *H19* (wH19) gene. TSS, transcription start site; ncRNA, non-coding RNA.

Despite 148 Myr of parallel evolution the exon/intron structure (but not the sequence) of this non-coding RNA gene is remarkably well conserved between wallaby and eutherians with one notable exception; In wallaby a supplementary 5' exon (exon 0) is separated from exon 1 by a weak intron (intron 0, Figure VI.3). The wallaby *H19* gene has a major transcription start site (TSS) at the beginning of exon 0 and a series of apparently minor TSSs throughout exon 0, intron 0 and the 5' region of exon 1, consistent with the recent definition of a broad promoter (Sandelin et al. 2007).

Further wallaby tissue samples were received from Marilyn Renfree and used to study the expression of wallaby *H19*. Expression was observed in the marsupial placenta (specifically in bilaminar and trilaminar yolk sac), foetal liver, pouch young liver and brain (Guillaume Smits, personal communication). SNPs in the wallaby *H19* and *IGF2* genes were used to determine whether these genes were imprinted in the tissues investigated, and if so, their parent-of-origin. As expected, the *H19* and *IGF2* genes were expressed from the maternal and paternal alleles respectively (Guillaume Smits, personal communication). In every tissue examined, imprinting (maternal expression) of *H19* was seen to be "complete" upon cDNA sequencing (Figure VI.4). In contrast, paternal expression of *IGF2* was complete in liver but "leaky" in brain and placenta (Suzuki et al. 2005). The eutherian (human and mouse) chorionic epithelium (brain) shows bi-allelic expression of *IGF2* and therefore the finding of leaky imprinted expression of wallaby *IGF2* in brain is not entirely surprising. However, eutherian imprinted *IGF2* expression in placenta and yolk sac is complete, at odds with the findings reported here of leaky wallaby *IGF2* expression. One explanation for this observation is that two cell populations (one imprinted and another with bi-allelic expression of *IGF2*) exist in the wallaby placenta, reminiscent of the eutherian brain. The difference in *IGF2* imprinting between wallaby liver and placenta parallels that observed in the eutherian *Kcnq1ot1* locus (Reik and Lewis. 2005). Leaky wallaby *IGF2* expression in the placenta is indicative that the therian placenta was likely not the major driving force for the emergence of *IGF2* imprinting. However, the possibility that wallaby has lost the need for complete *IGF2* imprinting cannot be excluded.

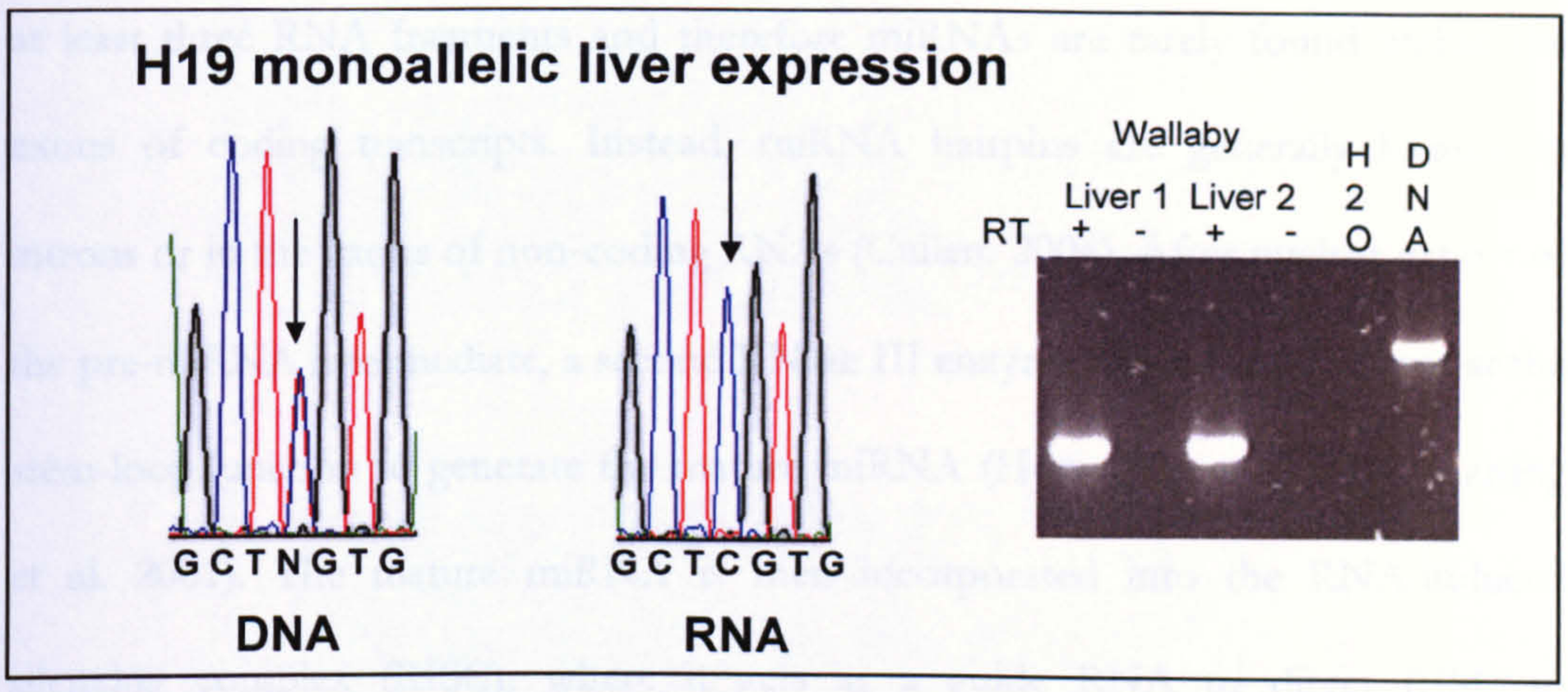


Figure VI.4. Expression of wallaby *H19*.

RT-PCR amplification from two pouch young liver samples is shown (right). A C/T SNP in genomic DNA (DNA) was used to show 'complete' monoallelic expression of *H19*. The cytosine base marked by the arrow in the RNA corresponds to the maternal allele.

6.3 Identifying wallaby micro RNA (miR-675) within exon 1 of the H19 gene

MicroRNAs (miRNAs) are short (18-25 nucleotides (nt)) non-coding RNA sequences that have been shown to regulate a wide range of biological processes including vertebrate development (reviewed in Zhao and Srivastava. 2007). Cellular miRNAs are initially expressed as part of one arm of an approximately 80-nt RNA hairpin derived from a longer primary miRNA transcript transcribed by RNA polymerase II (Cullen. 2004). Primary miRNA hairpins have a characteristic structure, consisting of an approximately 30-base-pair (bp) imperfect stem, a large terminal loop, and flanking unstructured RNA sequences, that is both necessary and sufficient for recognition by the nuclear RNase III enzyme Drosha acting in concert with its cofactor DGCR8 (Han et al. 2004, Zeng et al. 2005). Drosha then cleaves approximately 22 bp down the stem to excise the approximately 60-nt pre-miRNA intermediate. The cleavage of the primary miRNA by Drosha processing results in

at least three RNA fragments and therefore miRNAs are rarely found within the exons of coding transcripts. Instead, miRNA hairpins are generally located in introns or in the exons of non-coding RNAs (Cullen. 2004). After nuclear export of the pre-miRNA intermediate, a second RNase III enzyme called Dicer cleaves at the stem-loop junction to generate the mature miRNA (Hutvagner et al. 2001, Ketting et al. 2001). The mature miRNA is then incorporated into the RNA-induced silencing complex (RISC), where it acts as a guide RNA to direct RISC to complementary mRNA species for degradation or translational inhibition (Hammond et al. 2000, Martinez et al. 2002, Schwarz et al. 2002).

Mineno and colleagues (Mineno et al. 2006) used massively parallel signature sequencing (Brenner et al. 2000) to identify novel miRNA species from whole embryos of mouse. A total of 390 miRNAs were identified using this approach, of which 195 (50%) were previously known. The novel miRNAs were deposited into the public miRBase (Griffiths-Jones et al. 2006). One of these novel murine miRNAs (miR-675) was independently identified and further characterized in mouse and human by Cai and Cullen (Cai and Cullen. 2007). In mouse there are two mature 22 nt miRNAs lying within exon 1 of the *H19* gene (Positions chr7:142386502-142386523 and 142386468-142386489 in NCBI build 36 of mouse). A single mature 23 nt human miRNA lies at position chr11:1974606-1974628 in NCBI build 36 and position 1014-1036 of human *H19* RNA (NR_002196.1) The *H19* gene itself was therefore demonstrated to be the primary miRNA transcript in humans and mice and sheds some light on the elusive function of *H19*.

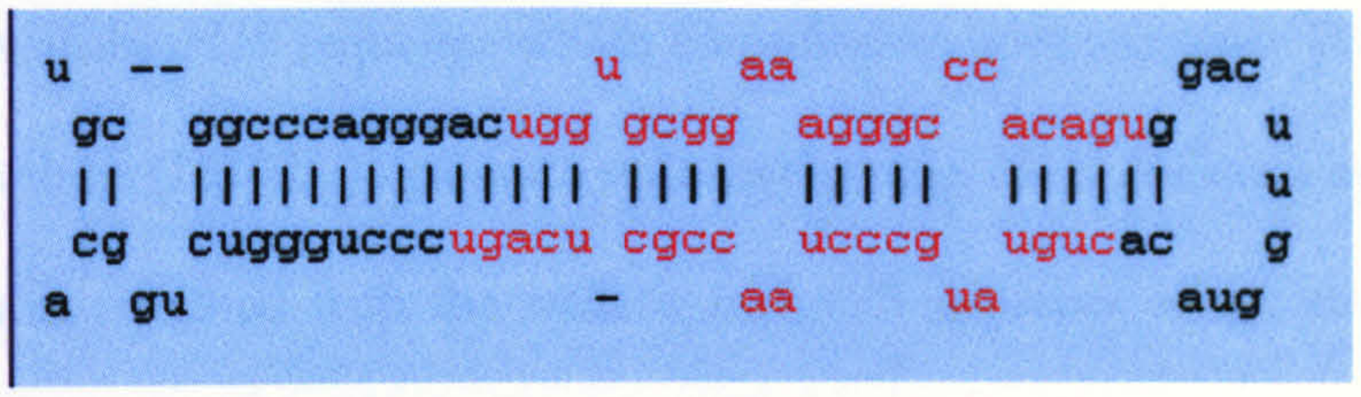


Figure VI.5. *Mus musculus* (top) and *Homo sapiens* (bottom) miR-675 stem-loop sequences.

As illustrated in miRBase. Mature miR-675 sequences are highlighted in red.

The approximately 80 nt miR-675 precursor is highly conserved between human and mouse (Figure VI.5). Alignments of derived eutherian miR-675 sequences were performed by Sam Griffiths-Jones (University of Manchester). The most highly conserved sequences from this alignment were used to search the wallaby BAC sequence (CR855994) using the 'DNA analysis' feature of ACeDB. This analysis revealed the presence of a candidate wallaby miR-675 lying on the reverse strand at position 109093-109178 of the BAC sequence (CR855994). This location is consistent with that of the wallaby *H19* gene (Figure VI.2).

6.4 Opossum H19 and miR-675

In order to assess whether the findings above are specific to wallaby or can be applied more generally to marsupial mammals BACs representing the IC1 domain of opossum were mapped and sequenced (see chapters III and IV). The unfinished consensus sequence for BAC VMRC18-490C6 was exported from a gap4 database and a BLASTN database created (chapter II). This database was searched with 2260

bp of the wallaby *H19* sequence which identified an approximately 70% nucleotide match to a short (1610 bp) opossum sequence contig. Next the same opossum BAC sequence was searched with the wallaby miR-675 sequence which revealed a 95% (82/86 bp) identity. Predicted mature miRNA sequences were identical in the two marsupial species. The eutherian and marsupial alignments of miR-675 are shown in Figure VI.6.

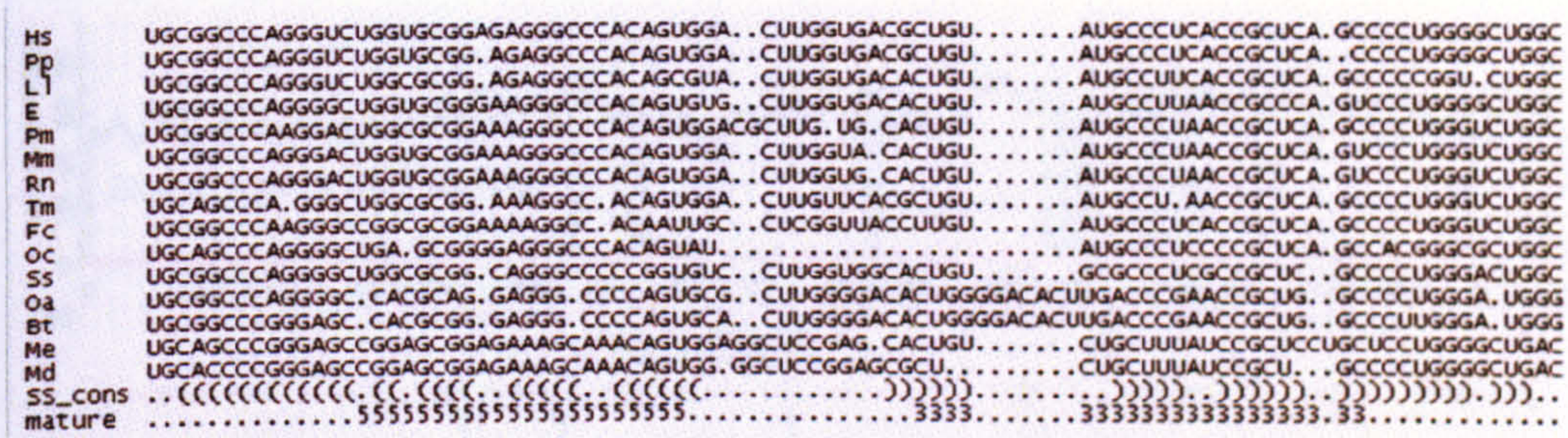


Figure VI.6. Sequence alignment of therian miR-675 sequences.

Sequence alignment produced by Sam Griffiths-Jones using the STOCKHOLM 1.0 software package. The *Monodelphis domestica* derived miR-675 sequence from BAC VMRC18-490C6 sequence was manually aligned. Hs, *Homo sapiens* (human); Pp, *Pongo pygmaeus* (orangutan); Ll, *Lynx lynx* (Eurasian lynx); E, *Elephantidae* (elephant); Pm, *Peromyscus maniculatus* (deer mouse); Mm, *Mus musculus* (house mouse); Rn, *Rattus norvegicus* (Norway rat); Tm, *Thomomys monticola* (mountain pocket gopher); Fc, *Felis catus* (domestic cat); Oc, *Oryctolagus cuniculus* (domestic rabbit); Ss, *Sus scrofa* (wild boar); Oa, *Ovis aries* (domestic sheep); Bt, *Bos taurus* (domestic cow); Me, *Macropus eugenii* (tammar wallaby); Md, *Monodelphis domestica* (opossum).

6.5 Identification and characterization of the H19 differentially methylated region in wallaby

An analysis of the CpG content of the wallaby *H19* gene and flanking sequences demonstrated a high density of CpG dinucleotides extending from 4 kb 5' to exon 0

to the 3' end of exon 1 (Figure VI.7). Differential methylation of this region was confirmed in liver, yolk sac, brain and primary ear fibroblasts by Guillaume Smits and colleagues (Babraham Institute, Cambridge). The region was fully methylated in testis and therefore a paternal germ-line DMR as in eutherians (Olek and Walter. 1997, Tremblay et al. 1997).

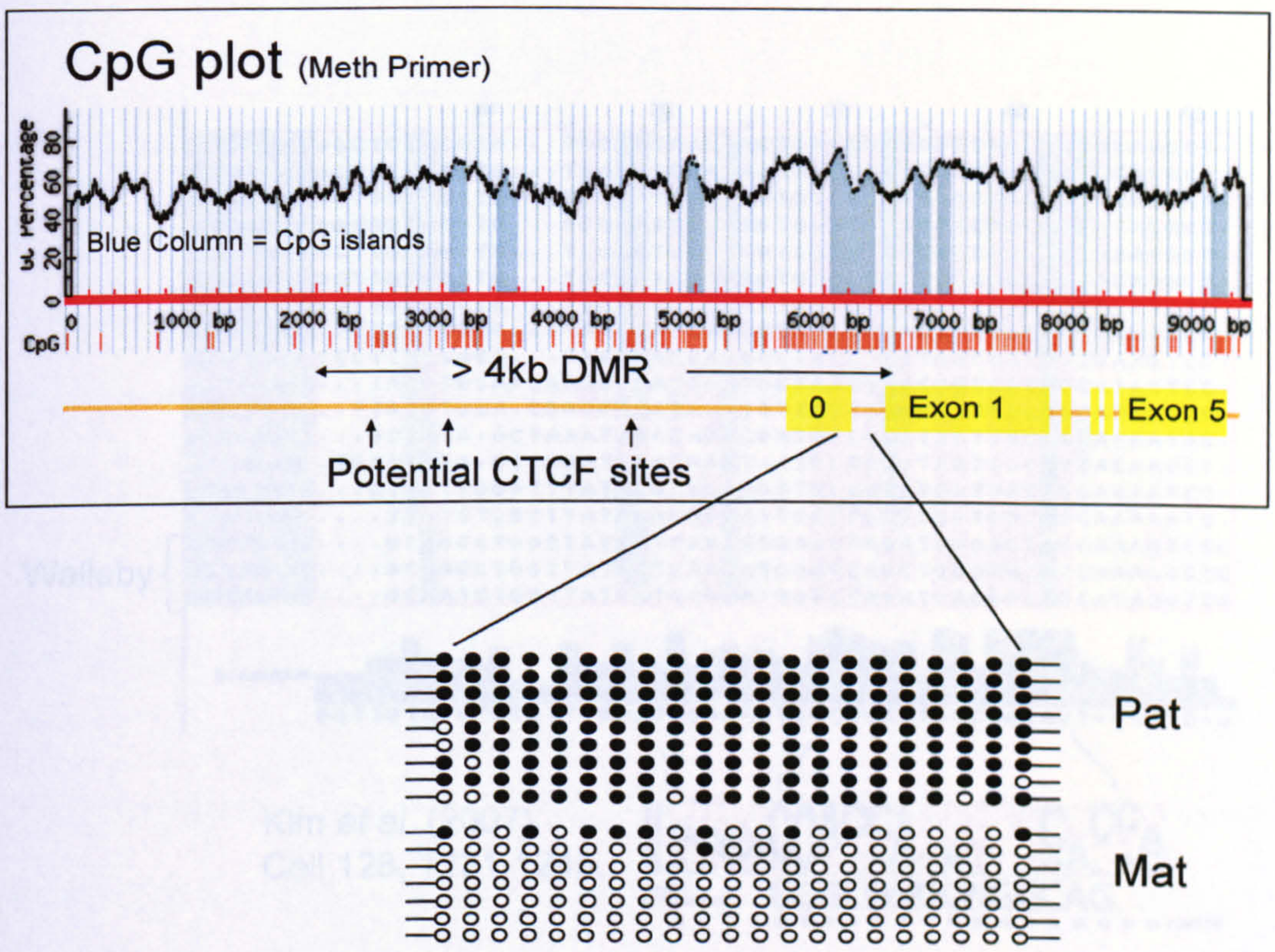


Figure VI.7. Identification of the wallaby *H19* DMR.

The C+G content of the wallaby *H19* gene and upstream region is depicted on the graph. Blue shading beneath the C+G plot indicates CpG islands. Individual CpG dinucleotides are shown by red vertical bars. Bisulphite sequencing of TA clones from the CpG island overlapping with exon 0 reveal differential methylation of this region. Filled circles, methylated; open circles unmethylated. Pat, paternal allele; Mat, maternal allele.

To address whether the wallaby DMR might function as an insulator, as is the case in eutherian mammals, I searched for potential CTCF binding sites. The 14 bp

consensus sequence CCGCGNGGNGGCAG (Bell and Felsenfeld. 2000) was searched against the wallaby BAC sequence CR855994 in ACeDB. Allowing for two mismatches a total of 25 matches was obtained including a cluster of 3 lying within the previously defined DMR, positions -4.2, -3.4 and -2.1 kb from the major H19 TSS. These putative CTCF binding sites lay within a 48 bp sequence motif (Figure VI.8 and Kim et al. 2007, Xie et al. 2007).

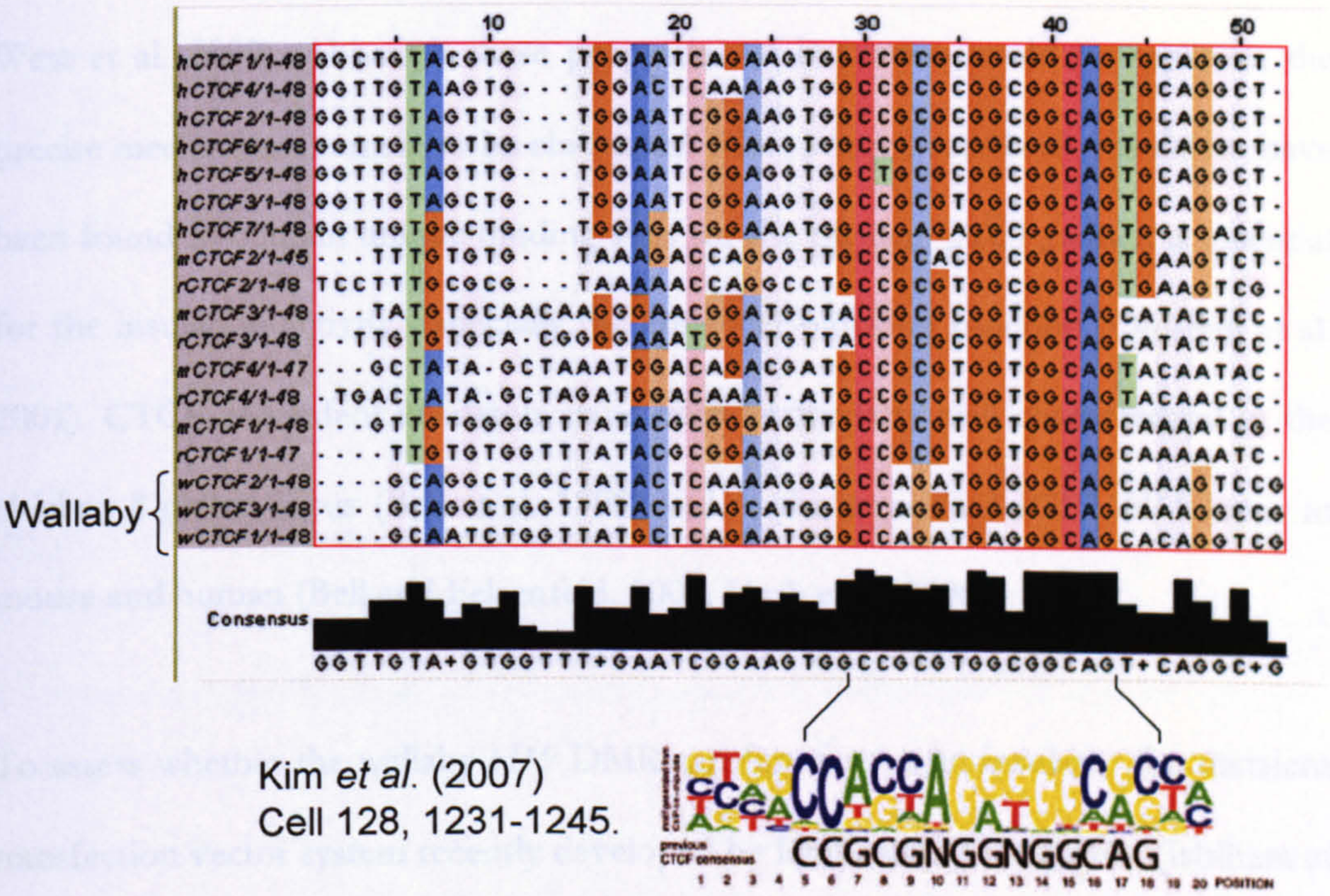


Figure VI.8. Identifying potential CTCF binding sites in wallaby.

ClustalW (version 1.83) sequence alignment of human CTCF (hCTCF), mouse CTCF (mCTCF), rat CTCF (rCTCF) and wallaby CTCF (wCTCF) binding sites in the *H19* DMR region. Alignments are viewed using JalView software (Clamp et al. 2004)**. A conserved 20 bp CTCF binding site logo is reproduced from** (Kim et al. 2007)**. Under the logo is the original 14 bp consensus** (Bell and Felsenfeld. 2000)**.**

6.6 Testing the wallaby DMR for insulator barrier activity

Chromatin insulators, also known as boundary elements, partition the genome into independent chromosomal domains to control individual gene expression. The boundaries produced by these insulators have the ability to block the interaction between enhancer(s) and promoter when positioned between them (enhancer-blocking activity). They also have the ability to block repressive chromatin effects on flanking regions (barrier activity, Bell et al. 2001, Gaszner and Felsenfeld. 2006, West et al. 2002). Although these properties of insulation could be separate, the precise mechanisms remain to be elucidated. Most known vertebrate insulators have been found to contain unique binding sites for the protein CTCF, which is essential for the insulation activity, especially the enhancer-blocking function (Ohlsson et al. 2001). CTCF-dependent insulators have been particularly well characterized in the chicken β-globin locus (Bell et al. 1999) and in the imprinted *IGF2/H19* locus in mouse and human (Bell and Felsenfeld. 2000, Hark et al. 2000).

To assess whether the wallaby *H19* DMR can function as an insulator the transient transfection vector system recently developed by Ishihara and colleagues (Ishihara et al. 2006) was used. This vector ('pIHLE', Figure VI.9) contains the firefly luciferase reporter gene under the control of the mouse *H19* promoter (-818 to +6 relative to the *H19* transcription start site) and stimulated by a downstream SV40 enhancer. Fragments to be tested were cloned into the *Bam*HI site between the luciferase reporter gene and enhancer (Figure VI.9). If a DNA fragment tested has enhancer blocking activity then it will block the interaction between the SV40 enhancer and *H19* promoter and therefore a reduction in luciferase gene expression will result. In contrast cloned fragments with no insulator function should have no effect on luciferase expression.

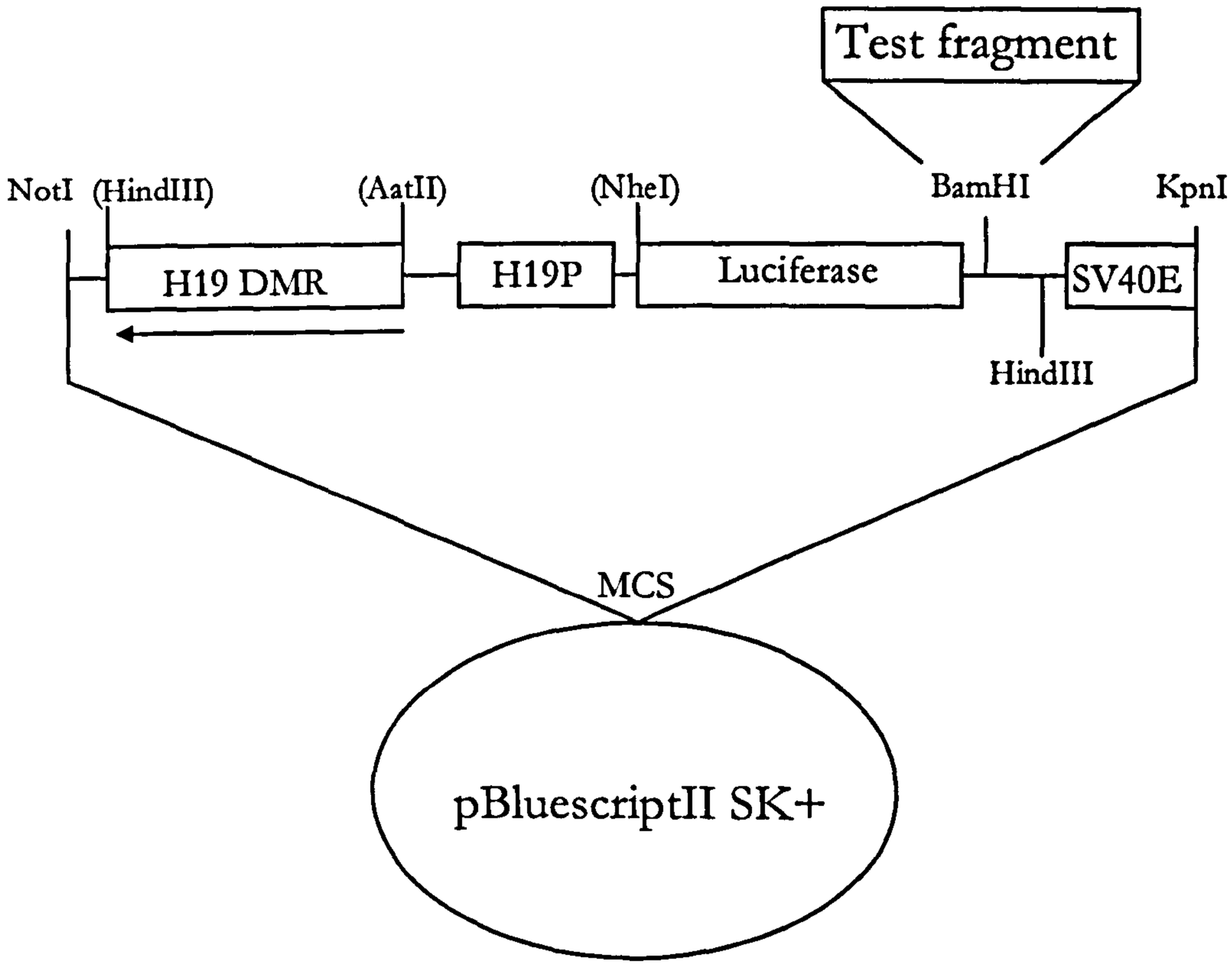


Figure VI.9. Testing for insulator function.

The pIHLE vector was a kind gift of Ko Ishihara and Mitsuyoshi Nakao. *H19* DMR, mouse 1.8 kb insulator; H19P, mouse *H19* promoter; Luciferase, firefly luciferase reporter gene from pGL3-Basic (Promega) vector; SV40E, SV40 enhancer; MCS, multiple cloning site. Test and control insulator sequences were cloned into the *Bam*HI site between luciferase reporter and SV40 enhancer sequences.

Purified DNA from plasmid constructs were co-transfected with a renilla luciferase expressing plasmid (pRL-CMV, for internal normalisation) in human hepatocellular liver carcinoma cell line (HepG2) cells. HepG2 cells were used because both *IGF2* and *H19* genes are transcribed at high levels in liver. Indeed the name *H19* comes from the fact that it was clone number 19 from a foetal hepatic library. Luciferase levels were detected in a luminometer and insulator activity was indicated by a decrease in relative luciferase expression levels when compared with the pIHLE

vector containing random sequence insertions. The pIHLE vector containing the 1.8 kb mouse H19 DMR insulator (Ishihara et al. 2006) cloned into the insulator site showed a greater than 5-fold (>80%) reduction in luciferase expression (Figure VI.10), compared to the 60% reduction observed by Ishihara and colleagues (Ishihara et al. 2006). One possible explanation for the improvement in insulation in this study is that liver (HepG2) cells were transfected in contrast to HeLa cells used by Ishihara and co-workers. *H19* RNA is strongly expressed in HepG2 cells and therefore the *H19* promoter driving luciferase expression is in a more native context. The pIHLE vector containing the 2.3 kb candidate wallaby insulator sequence demonstrated a 40% reduction in luciferase expression (Figure VI.10). To demonstrate whether the apparent insulator effect could be the result of a position effect i.e. increasing the distance between the SV40 enhancer and *H19* promoter, random 2.3-2.4 kb wallaby fragments were also cloned into the *Bam*HI site of the pIHLE vector. Only a marginal decrease in relative luciferase expression was observed indicating that the 2.3 kb cloned DNA fragments have little or no effect on luciferase activity (Figure VI.10).

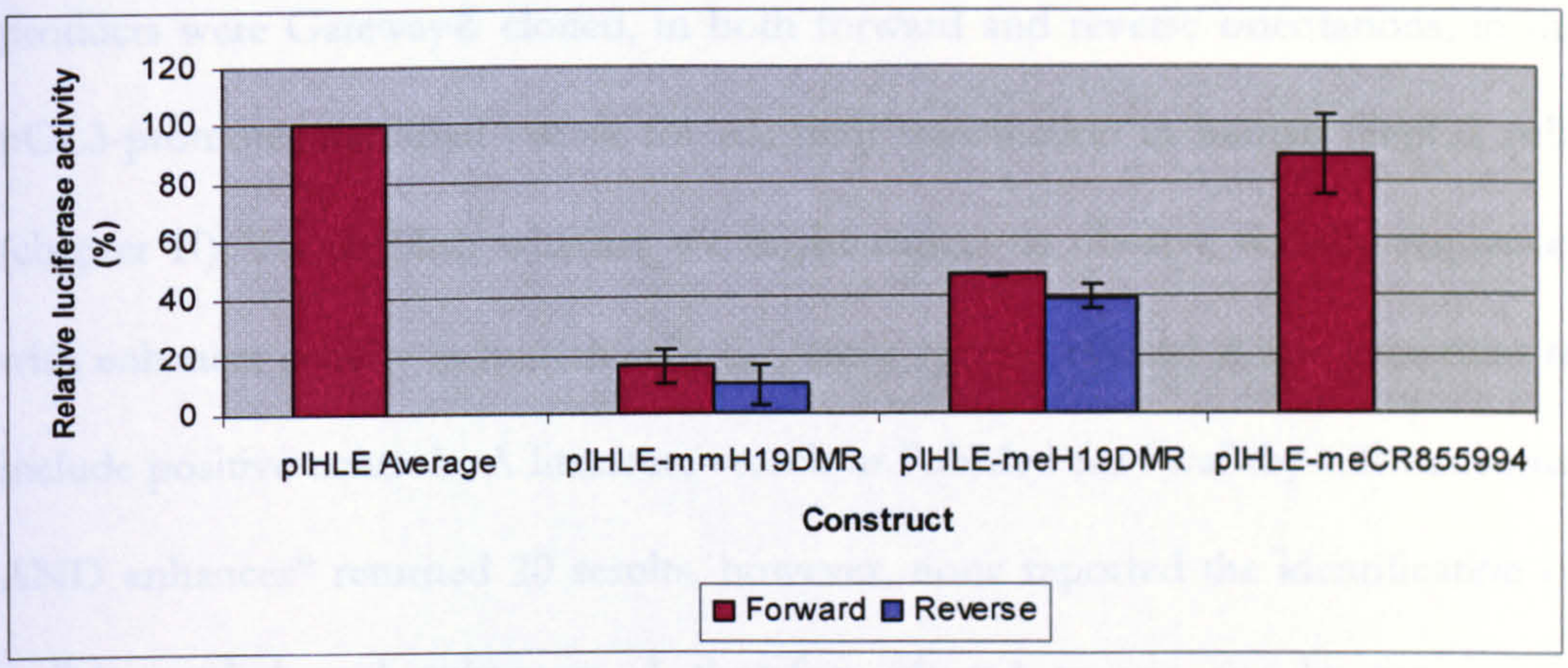


Figure VI.10. Insulator activity of the wallaby *H19* DMR.

Luciferase levels were normalized against the pIHLE vector with no insert. The pIHLE-mmH19DMR vector contains the mouse *H19* DMR with known enhancer-blocker activity. The pIHLE-meH19DMR vector contains a 2.3 kb wallaby sequence encompassing 3 putative CTCF binding sites. The pIHLE-meCR855994 bar represents an average of two independent constructs containing different randomly cloned 2.3 kb PCR products amplified from wallaby BAC sequence CR855994. Mouse and wallaby *H19* DMR constructs were cloned in both forward (claret) and reverse (blue) orientations.

6.7 Searching for wallaby endodermal enhancers

The *IGF2* gene is imprinted (paternally expressed) in wallaby tissues; yolk sac placenta, foetal and pouch young livers (Guillaume Smits personal communication). We have also established the presence of wallaby *H19* and the upstream DMR which has CTCF sites and insulator (enhancer-blocker) activity analogous to the situation in human and mouse. According to the 'boundary model' the enhancer-blocking activity of the CTCF occupied *H19* DMR on the maternal allele prevents access of the endodermal enhancers to the *IGF2* promoter. So could endodermal enhancers be identified in wallaby? To address this question I designed 20, approximately 3 kb wallaby PCR amplicons, overlapping by 1 kb, from across the 38 kb wallaby region between ECR14 and ECR15 (Figure VI.2). The PCR amplified

products were Gateway® cloned, in both forward and reverse orientations, in the pGL3-promoter modified vector for transient transfection in human HepG2 cells (chapter II). To establish whether we might expect to observe wallaby sequences with enhancer activity in human cells (i.e. cross-species effects) it was important to include positive controls. A literature search in PubMed for "wallaby OR marsupial AND enhancer" returned 20 results, however, none reported the identification of wallaby endodermal enhancers. I therefore elected to use the known mouse endodermal enhancer 1 (EE1) element from the *H19* downstream region. In human HepG2 cells this mouse enhancer increased relative luciferase activity by approximately 16-fold (log2 of 4) over the empty Gateway® modified pGL3-Promoter construct (Figure VI.11). To control for size effects a 3 kb human tile containing both endodermal enhancers EE1 and EE2 (positive control) and a neighbouring, but non-overlapping, 3 kb tile (negative control) was cloned and transfected. The positive control shows that cloning relatively large (3 kb) fragments into an already large plasmid (approximately 8 kb) has little, if any, effect on detecting enhancer activities (Figure VI.11). Indeed there is an apparent additive effect of cloning EE1 and EE2 together compared with the enhancer activities of EE1 and EE2 cloned independently (chapter V). None of the 20 wallaby tiles showed enhancer activity in human HepG2 cells (Figure VI.11). Whether this is the result of sequence divergence between wallaby and human rendering the human gene regulatory machinery incapable of recognising wallaby enhancers is unknown. However, despite having last shared a common ancestor with mouse over 90 Myr ago, mouse EE1 behaves as a potent enhancer in human HepG2 cells with enhancer activity equal to that of the human EEs. Perhaps a more likely explanation for the lack of EE activity is that the enhancers are not located in this region of conserved synteny but lie in a more remote location.

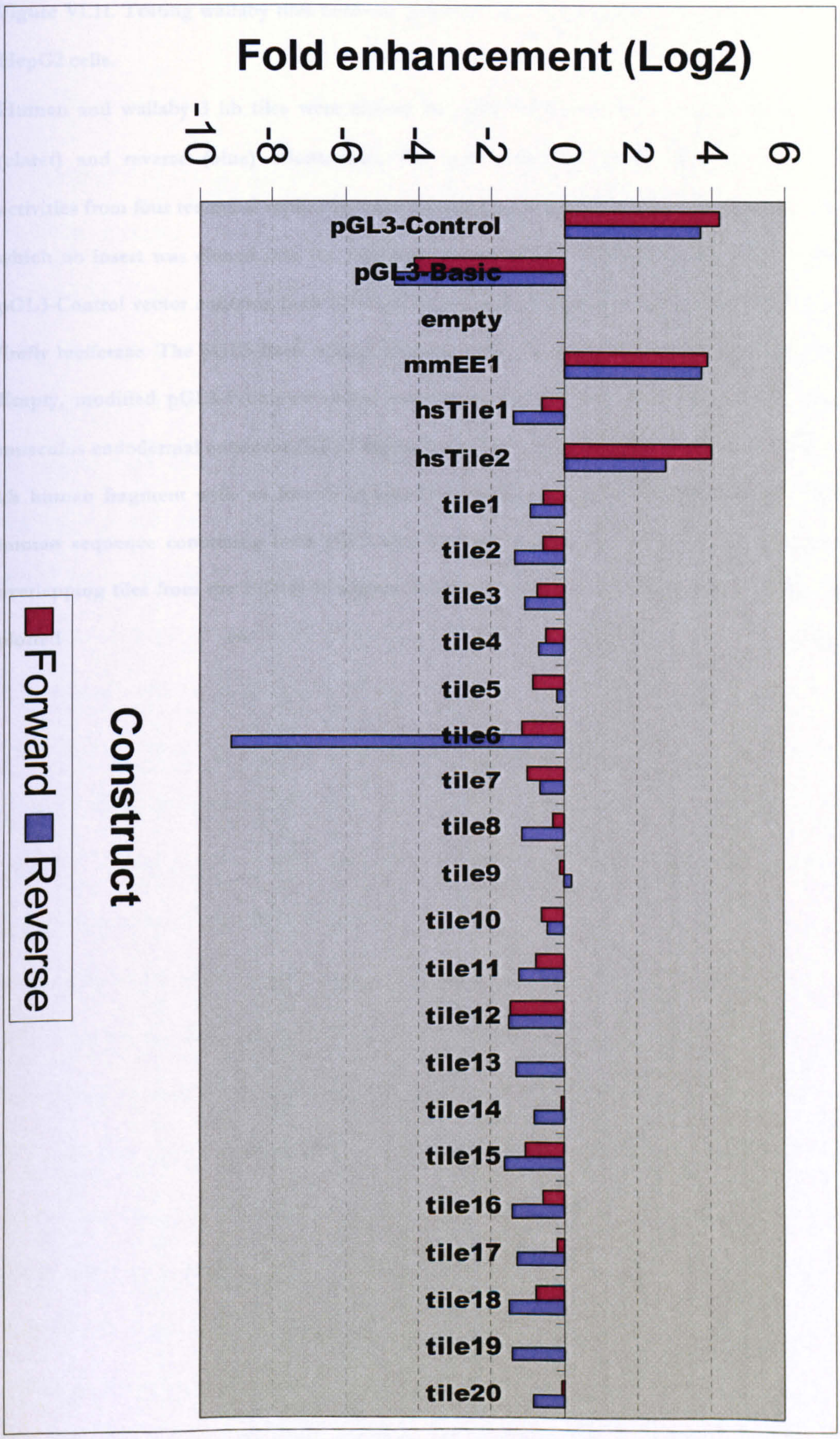

Fold enhancement (Log2)
-10
-8
-6
-4
-2
0
2
4
6
pGL3-Control
pGL3-Basic
empty
mmEE1
hsTile1
hsTile2
tile1
tile2
tile3
tile4
tile5
tile6
tile7
tile8
tile9
tile10
tile11
tile12
tile13
tile14
tile15
tile16
tile17
tile18
tile19
tile20
Construct
Forward
Reverse

Figure VI.11. Testing wallaby tiles between ECRs 14 and 15 for enhancer activity in human HepG2 cells.

Human and wallaby 3 kb tiles were cloned into pGL3-Promoter vectors in both forward (claret) and reverse (blue) orientations. For each construct transfected the luciferase activities from four technical replicates were averaged and normalized against a construct in which no insert was cloned into the Gateway cassette (labeled 'empty' in the chart). The pGL3-Control vector contains both SV40 enhancer and promoter for optimal expression of firefly luciferase. The pGL3-Basic vector contains neither promoter nor enhancer elements. Empty, modified pGL3-Promoter vector containing the Gateway cassette. mmEE1, *Mus musculus* endodermal enhancer (EE) 1 element cloned into the Gateway cassette. hsTile1, 3 kb human fragment with no known enhancer elements (negative control). hsTile2, 3 kb human sequence containing both EE 1 and 2 (positive control). Tile1-20, wallaby 3 kb overlapping tiles from the ECR14-15 region. Fold enhancement over the empty construct is plotted on a log2 scale.

6.8 Discussion

Taken together the results presented in this chapter demonstrate that, with the possible exception of endodermal enhancers, all the major hallmarks of the eutherian *H19-IGF2* imprinting system are present in marsupials and must therefore have been present in the therian ancestor in existence prior to the divergence of marsupials and eutherians. One of the few apparent differences between the marsupial and eutherian IC1 domains is the absence of endodermal enhancers in the region of conserved synteny between ECR14 and ECR15. As discussed above this could be a sensitivity problem with the transient transfection assay in that the sequence divergence between wallaby and human gives a false negative result. Alternatively the enhancer position may be less well conserved between therians than, for example, the *H19* DMR. It should be noted that ECR11 and ECR12h, lying 40 kb telomeric of *H19* do show significant enhancer activity in HepG2 cells (chapter V) and therefore these endodermal enhancers could be enhancing wallaby *IGF2* expression on the paternal allele.

Every imprinted gene cluster characterized to date harbours at least one non-coding RNA and yet the function of many remains elusive (Pauler and Barlow. 2006). The processing of miR-675 from exon 1 of the *H19* pri-miRNA precursor RNA is evidently not highly efficient because full-length *H19* transcripts are abundant in developing embryos (Bartolomei et al. 1991). miRNAs themselves are subject to developmental regulation (Obernosterer et al. 2006, Thomson et al. 2006) and therefore it is possible that processing of miRNAs from *H19* varies both spatially (different tissue types) and temporally (different developmental stages). If the full-length *H19* transcript does have a function, which would seem likely given the

evolutionary conservation of the exon/intron structure, then it is possible that miR-675 serves to affect this function.

Kinship theory (Haig. 2004) predicts that maternally expressed genes such as *H19* will generally act to suppress excessive foetal growth stimulated by paternally expressed growth factors such as *IGF2*. It will therefore be interesting to see whether the post-transcriptional targets of miR-675 include growth factors, their receptors or downstream pathway members. A search for predicted microRNA targets for mouse miR-675-5p in the TargetScan (release 4.0, July 2007) database (Grimson et al. 2007) reveals 8 conserved targets; CALN1, OSR2, RNF165, WIT1, TLX3, MLL2, SLC22A3 and EGR3. Some caution must be exercised when interpreting the results of miRNA target prediction before experimental evidence exists. However, it is of note that Wilms tumour upstream neighbour 1 is in the list. This intronless gene lies at 11p13 in a head to tail fashion with the Wilms tumour 1 (WT1) gene and shares a bidirectional promoter with it. WT1 transcripts have been reported to be imprinted (Jinno et al. 1994, Mitsuya et al. 1997, Dallosso et al. 2004). Furthermore WIT1 hypomethylation has been implicated in chemoresistant acute myeloid leukaemia (Plass et al. 1999).

The absence of *IGF2* imprinting in monotreme mammals makes it of particular interest to search for the features discussed above. As described in chapters III and IV physical BAC mapping and sequencing in the IC1 (and IC2) region of platypus has been hampered by clone coverage and stability. As a consequence a BAC spanning ECR14 to ECR15 has yet to be identified. Likewise, no whole genome shotgun contig(s) have been identified from this orthologous region. New platypus genomic resources such as markers and/or BAC libraries may be required to extend existing sequences in this region. A BAC library is also now available for the short-

beaked echidna (*Tachyglossus aculeatus*). Further work will be required to determine whether the IC1 region can be mapped and sequenced in this monotreme.

To conclude, the *H19-IGF2* imprinting system must have been present in the therian ancestor making it the most ancient mammalian epigenetic system identified to date. The marsupial *H19* gene is also the most ancient macroRNA known and for at least 148 million years has served as a miRNA precursor. These findings should help in ascribing elusive function to the *H19* gene.

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.

Chapter VIII - References

Abbasi, A.A., Paparidis, Z., Malik, S., Goode, D.K., Callaway, H., Elgar, G. and Grzeschik, K.H. (2007) Human GLI3 Intragenic Conserved Non-Coding Sequences Are Tissue-Specific Enhancers. *PLoS ONE,* **2,** e366.

Adams, D.J., Dermitzakis, E.T., Cox, T., Smith, J., Davies, R., Banerjee, R., Bonfield, J., Mullikin, J.C., Chung, Y.J., Rogers, J., et al (2005) Complex haplotypes, copy number polymorphisms and coding variation in two recently divergent mouse strains. *Nat. Genet.,* **37,** 532-536.

Ager, E., Suzuki, S., Pask, A., Shaw, G., Ishino, F. and Renfree, M.B. (2007) Insulin is imprinted in the placenta of the marsupial, Macropus eugenii. *Dev. Biol.,* **309,** 317-328.

Allen, E., Horvath, S., Tong, F., Kraft, P., Spiteri, E., Riggs, A.D. and Marahrens, Y. (2003) High concentrations of long interspersed nuclear element sequence distinguish monoallelically expressed genes. *Proc. Natl. Acad. Sci. U. S. A.,* **100,** 9940-9945.

Aloni, R. and Lancet, D. (2005) Conservation anchors in the vertebrate genome. *Genome Biol.,* **6,** 115.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.,* **215,** 403-410.

Ambatipudi, K., Joss, J., Raftery, M. and Deane, E. (2007) A proteomic approach to analysis of antimicrobial activity in marsupial pouch secretions. *Dev. Comp. Immunol.,*

Anderson, R.J. and Spencer, H.G. (1999) Population models of genomic imprinting. I. Differential viability in the sexes and the analogy with genetic dominance. *Genetics,* **153,** 1949-1958.

Arney, K.L. (2003) H19 and Igf2--enhancing the confusion? *Trends Genet.,* **19,** 17-23.

Ashurst, J.L., Chen, C.K., Gilbert, J.G., Jekosch, K., Keenan, S., Meidl, P., Searle, S.M., Stalker, J., Storey, R., Trevanion, S., et al (2005) The Vertebrate Genome Annotation (Vega) database. *Nucleic Acids Res.,* **33,** D459-65.

Bailey, J.A., Baertsch, R., Kent, W.J., Haussler, D. and Eichler, E.E. (2004) Hotspots of mammalian chromosomal evolution. *Genome Biol.*, **5,** R23.

Banerji, J., Olson, L. and Schaffner, W. (1983) A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. *Cell,* **33,** 729-740.

Banerji, J., Rusconi, S. and Schaffner, W. (1981) Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell,* **27,** 299-308.

Barlow, D.P. (1995) Gametic imprinting in mammals. *Science,* **270,** 1610-1613.

Barlow, D.P. (1993) Methylation and imprinting: from host defense to gene regulation? *Science,* **260,** 309-310.

Barlow, D.P., Stoger, R., Herrmann, B.G., Saito, K. and Schweifer, N. (1991) The mouse insulin-like growth factor type-2 receptor is imprinted and closely linked to the Tme locus. *Nature,* **349,** 84-87.

Bartolomei, M.S., Zemel, S. and Tilghman, S.M. (1991) Parental imprinting of the mouse H19 gene. *Nature,* **351,** 153-155.

Barton, S.C., Surani, M.A. and Norris, M.L. (1984) Role of paternal and maternal genomes in mouse development. *Nature,* **311,** 374-376.

Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L., et al (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32,** D138-41.

Batzoglou, S. (2005) The many faces of sequence alignment. *Brief Bioinform,* **6,** 6-22.

Beaudet, A.L. and Jiang, Y.H. (2002) A rheostat model for a rapid and reversible form of imprinting-dependent evolution. *Am. J. Hum. Genet.*, **70,** 1389-1397.

Beck, S., Olek, A. and Walter, J. (1999) From genomics to epigenomics: a loftier view of life. *Nat. Biotechnol.*, **17,** 1144.

Bejerano, G., Lowe, C.B., Ahituv, N., King, B., Siepel, A., Salama, S.R., Rubin, E.M., Kent, W.J. and Haussler, D. (2006) A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature,* **441,** 87-90.

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W.J., Mattick, J.S. and Haussler, D. (2004) Ultraconserved elements in the human genome. *Science,* **304,** 1321-1325.

Bell, A.C. and Felsenfeld, G. (2000) Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. *Nature*, **405,** 482-485.

Bell, A.C., West, A.G. and Felsenfeld, G. (2001) Insulators and boundaries: versatile regulatory elements in the eukaryotic genome. *Science*, **291,** 447-450.

Bell, A.C., West, A.G. and Felsenfeld, G. (1999) The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. *Cell*, **98,** 387-396.

Benson, G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*, **27,** 573-580.

Bentley, D.R. (2006) Whole-genome re-sequencing. *Curr. Opin. Genet. Dev.*, **16,** 545-552.

Bentley, D.R. (1996) Genomic sequence information should be released immediately and freely in the public domain. *Science*, **274,** 533-534.

Bentley, D.R., Deloukas, P., Dunham, A., French, L., Gregory, S.G., Humphray, S.J., Mungall, A.J., Ross, M.T., Carter, N.P., Dunham, I., et al (2001) The physical maps for sequencing human chromosomes 1, 6, 9, 10, 13, 20 and X. *Nature*, **409,** 942-943.

Berg, P. (2006) Origins of the human genome project: why sequence the human genome when 96% of it is junk? *Am. J. Hum. Genet.*, **79,** 603-605.

Bernardi, G. (2000) Isochores and the evolutionary genomics of vertebrates. *Gene*, **241,** 3-17.

Bernardi, G. (1995) The human genome: organization and evolutionary history. *Annu. Rev. Genet.*, **29,** 445-476.

Bernardi, G., Mouchiroud, D., Gautier, C. and Bernardi, G. (1988) Compositional patterns in vertebrate genomes: conservation and change in evolution. *J. Mol. Evol.*, **28,** 7-18.

Bernstein, B.E., Meissner, A. and Lander, E.S. (2007) The mammalian epigenome. *Cell*, **128,** 669-681.

Bertone, P., Stolc, V., Royce, T.E., Rozowsky, J.S., Urban, A.E., Zhu, X., Rinn, J.L., Tongprasit, W., Samanta, M., Weissman, S., et al (2004) Global identification of human transcribed sequences with genome tiling arrays. *Science*, **306,** 2242-2246.

Bininda-Emonds, O.R., Cardillo, M., Jones, K.E., MacPhee, R.D., Beck, R.M., Grenyer, R., Price, S.A., Vos, R.A., Gittleman, J.L. and Purvis, A. (2007) The delayed rise of present-day mammals. *Nature*, **446,** 507-512.

Bird, A.P. (1986) CpG-rich islands and the function of DNA methylation. *Nature*, **321,** 209-213.

Birney, E., Clamp, M. and Durbin, R. (2004) GeneWise and Genomewise. *Genome Res.*, **14,** 988-995.

Blakesley, R.W., Hansen, N.F., Mullikin, J.C., Thomas, P.J., McDowell, J.C., Maskeri, B., Young, A.C., Benjamin, B., Brooks, S.Y., Coleman, B.I., et al (2004) An intermediate grade of finished genomic sequence suitable for comparative analyses. *Genome Res.*, **14,** 2235-2244.

Blanchette, M., Kent, W.J., Riemer, C., Elnitski, L., Smit, A.F., Roskin, K.M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E.D., et al (2004) Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.*, **14,** 708-715.

Boffelli, D., McAuliffe, J., Ovcharenko, D., Lewis, K.D., Ovcharenko, I., Pachter, L. and Rubin, E.M. (2003) Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science*, **299,** 1391-1394.

Bonfield, J.K., Smith, K. and Staden, R. (1995) A new DNA sequence assembly program. *Nucleic Acids Res.*, **23,** 4992-4999.

Bouma, M.E., Rogier, E., Verthier, N., Labarre, C. and Feldmann, G. (1989) Further cellular investigation of the human hepatoblastoma-derived cell line HepG2: morphology and immunocytochemical studies of hepatic-secreted proteins. *In Vitro Cell. Dev. Biol.*, **25,** 267-275.

Bourque, G., Pevzner, P.A. and Tesler, G. (2004) Reconstructing the genomic architecture of ancestral mammals: lessons from human, mouse, and rat genomes. *Genome Res.*, **14,** 507-516.

Brannan, C.I., Dees, E.C., Ingram, R.S. and Tilghman, S.M. (1990) The product of the H19 gene may function as an RNA. *Mol. Cell. Biol.*, **10,** 28-36.

Bray, N. and Pachter, L. (2004) MAVID: constrained ancestral alignment of multiple sequences. *Genome Res.*, **14,** 693-699.

Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D.H., Johnson, D., Luo, S., McCurdy, S., Foy, M., Ewan, M., et al (2000) Gene expression analysis by

massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.*, **18,** 630-634.

Bridgham, J.T. and Johnson, A.L. (2004) Alternatively spliced variants of Gallus gallus TNFRSF23 are expressed in the ovary and differentially regulated by cell signaling pathways. *Biol. Reprod.*, **70,** 972-979.

Brudno, M., Do, C.B., Cooper, G.M., Kim, M.F., Davydov, E., NISC Comparative Sequencing Program, Green, E.D., Sidow, A. and Batzoglou, S. (2003) LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.*, **13,** 721-731.

Brudno, M., Poliakov, A., Minovitsky, S., Ratnere, I. and Dubchak, I. (2007) Multiple whole genome alignments and novel biomedical applications at the VISTA portal. *Nucleic Acids Res.*, **35,** W669-74.

Bruford, E.A., Lush, M.J., Wright, M.W., Sneddon, T.P., Povey, S. and Birney, E. (2007) The HGNC Database in 2008: a resource for the human genome. *Nucleic Acids Res.*,

Burge, C. and Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, **268,** 78-94.

Burgio, G., Szatanik, M., Guenet, J.L., Arnau, M.R., Panthier, J.J. and Montagutelli, X. (2007) Interspecific recombinant congenic strains between C57BL/6 and mice of the Mus spretus species : a powerful tool to dissect genetic control of complex traits. *Genetics*,

Burt, A. and Trivers, R. (1998) Genetic conflicts in genomic imprinting. *Proc. Biol. Sci.*, **265,** 2393-2397.

Cai, X. and Cullen, B.R. (2007) The imprinted H19 noncoding RNA is a primary microRNA precursor. *RNA*, **13,** 313-316.

Carninci, P. (2006) Tagging mammalian transcription complexity. *Trends Genet.*, **22,** 501-510.

Cartharius, K., Frech, K., Grote, K., Klocke, B., Haltmeier, M., Klingenhoff, A., Frisch, M., Bayerlein, M. and Werner, T. (2005) MatInspector and beyond: promoter analysis based on transcription factor binding sites. *Bioinformatics*, **21,** 2933-2942.

Cattanach, B.M. and Kirk, M. (1985) Differential activity of maternally and paternally derived chromosome regions in mice. *Nature*, **315,** 496-498.

Cavaille, J., Seitz, H., Paulsen, M., Ferguson-Smith, A.C. and Bachellerie, J.P. (2002) Identification of tandemly-repeated C/D snoRNA genes at the imprinted human 14q32 domain reminiscent of those at the Prader-Willi/Angelman syndrome region. *Hum. Mol. Genet.*, **11,** 1527-1538.

Celniker, S.E. and Drewell, R.A. (2007) Chromatin looping mediates boundary element promoter interactions. *Bioessays,* **29,** 7-10.

Chapman, M.A., Charchar, F.J., Kinston, S., Bird, C.P., Grafham, D., Rogers, J., Grutzner, F., Graves, J.A., Green, A.R. and Gottgens, B. (2003) Comparative and functional analyses of LYL1 loci establish marsupial sequences as a model for phylogenetic footprinting. *Genomics,* **81,** 249-259.

Cheng, J., Kapranov, P., Drenkow, J., Dike, S., Brubaker, S., Patel, S., Long, J., Stern, D., Tammana, H., Helt, G., et al (2005) Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science,* **308,** 1149-1154.

Chimpanzee Sequencing and Analysis Consortium (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature,* **437,** 69-87.

Choi, S.H., Kim, I.C., Kim, D.S., Kim, D.W., Chae, S.H., Choi, H.H., Choi, I., Yeo, J.S., Song, M.N. and Park, H.S. (2006) Comparative genomic organization of the human and bovine PRNP locus. *Genomics,* **87,** 598-607.

Chow, J.C., Yen, Z., Ziesche, S.M. and Brown, C.J. (2005) Silencing of the mammalian X chromosome. *Annu. Rev. Genomics Hum. Genet.*, **6,** 69-92.

Chung, J.H., Bell, A.C. and Felsenfeld, G. (1997) Characterization of the chicken beta-globin insulator. *Proc. Natl. Acad. Sci. U. S. A.*, **94,** 575-580.

Clamp, M., Cuff, J., Searle, S.M. and Barton, G.J. (2004) The Jalview Java alignment editor. *Bioinformatics,* **20,** 426-427.

Clark, L., Wei, M., Cattoretti, G., Mendelsohn, C. and Tycko, B. (2002) The Tnfrh1 (Tnfrsf23) gene is weakly imprinted in several organs and expressed at the trophoblast-decidua interface. *BMC Genet.*, **3,** 11.

Collins, J.E., Wright, C.L., Edwards, C.A., Davis, M.P., Grinham, J.A., Cole, C.G., Goward, M.E., Aguado, B., Mallya, M., Mokrab, Y., et al (2004) A genome annotation-driven approach to cloning the human ORFeome. *Genome Biol.*, **5,** R84.

Constancia, M., Kelsey, G. and Reik, W. (2004) Resourceful imprinting. *Nature,* **432,** 53-57.

Cooper, D.W., VandeBerg, J.L., Sharman, G.B. and Poole, W.E. (1971) Phosphoglycerate kinase polymorphism in kangaroos provides further evidence for paternal X inactivation. *Nat. New Biol.*, **230,** 155-157.

Cooper, G.M. and Sidow, A. (2003) Genomic regulatory regions: insights from comparative sequence analysis. *Curr. Opin. Genet. Dev.*, **13,** 604-610.

Cooper, G.M., Stone, E.A., Asimenos, G., NISC Comparative Sequencing Program, Green, E.D., Batzoglou, S. and Sidow, A. (2005) Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.*, **15,** 901-913.

Coulondre, C., Miller, J.H., Farabaugh, P.J. and Gilbert, W. (1978) Molecular basis of base substitution hotspots in Escherichia coli. *Nature,* **274,** 775-780.

Crawford, G.E., Davis, S., Scacheri, P.C., Renaud, G., Halawi, M.J., Erdos, M.R., Green, R., Meltzer, P.S., Wolfsberg, T.G. and Collins, F.S. (2006) DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays. *Nat. Methods,* **3,** 503-509.

Cullen, B.R. (2004) Transcription and processing of human microRNA precursors. *Mol. Cell,* **16,** 861-865.

Dallosso, A.R., Hancock, A.L., Brown, K.W., Williams, A.C., Jackson, S. and Malik, K. (2004) Genomic imprinting at the WT1 gene involves a novel coding transcript (AWT1) that shows deregulation in Wilms' tumours. *Hum. Mol. Genet.*, **13,** 405-415.

Dannenberg, L.O. and Edenberg, H.J. (2006) Epigenetics of gene expression in human hepatoma cells: expression profiling the response to inhibition of DNA methylation and histone deacetylation. *BMC Genomics,* **7,** 181.

Darling, A.C., Mau, B., Blattner, F.R. and Perna, N.T. (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.*, **14,** 1394-1403.

Day, T. and Bonduriansky, R. (2004) Intralocus sexual conflict can drive the evolution of genomic imprinting. *Genetics,* **167,** 1537-1546.

Deakin, J.E., Siddle, H.V., Cross, J.G., Belov, K. and Graves, J.A. (2007) Class I genes have split from the MHC in the tammar wallaby. *Cytogenet. Genome Res.*, **116,** 205-211.

Dean, W., Bowden, L., Aitchison, A., Klose, J., Moore, T., Meneses, J.J., Reik, W. and Feil, R. (1998) Altered imprinted gene methylation and expression in completely

ES cell-derived mouse fetuses: association with aberrant phenotypes. *Development,* **125,** 2273-2282.

DeChiara, T.M., Robertson, E.J. and Efstratiadis, A. (1991) Parental imprinting of the mouse insulin-like growth factor II gene. *Cell,* **64,** 849-859.

Deininger, P.L. and Batzer, M.A. (1999) Alu repeats and human disease. *Mol. Genet. Metab.,* **67,** 183-193.

Dekker, J., Rippe, K., Dekker, M. and Kleckner, N. (2002) Capturing chromosome conformation. *Science,* **295,** 1306-1311.

Delbridge, M.L., Lingenfelter, P.A., Disteche, C.M. and Graves, J.A. (1999) The candidate spermatogenesis gene RBMY has a homologue on the human X chromosome. *Nat. Genet.,* **22,** 223-224.

Dewey, C.N. and Pachter, L. (2006) Evolution at the nucleotide level: the problem of multiple whole-genome alignment. *Hum. Mol. Genet.,* **15 Spec No 1,** R51-6.

Dhami, P., Coffey, A.J., Abbs, S., Vermeesch, J.R., Dumanski, J.P., Woodward, K.J., Andrews, R.M., Langford, C. and Vetrie, D. (2005) Exon array CGH: detection of copy-number changes at the resolution of individual exons in the human genome. *Am. J. Hum. Genet.,* **76,** 750-762.

Down, T.A. and Hubbard, T.J. (2005) NestedMICA: sensitive inference of over-represented motifs in nucleic acid sequence. *Nucleic Acids Res.,* **33,** 1445-1453.

Down, T.A. and Hubbard, T.J. (2002) Computational detection and location of transcription start sites in mammalian genomic DNA. *Genome Res.,* **12,** 458-461.

Drewell, R.A., Arney, K.L., Arima, T., Barton, S.C., Brenton, J.D. and Surani, M.A. (2002) Novel conserved elements upstream of the H19 gene are transcribed and act as mesodermal enhancers. *Development,* **129,** 1205-1213.

Drewell, R.A., Brenton, J.D., Ainscough, J.F., Barton, S.C., Hilton, K.J., Arney, K.L., Dandolo, L. and Surani, M.A. (2000) Deletion of a silencer element disrupts H19 imprinting independently of a DNA methylation epigenetic switch. *Development,* **127,** 3419-3428.

Duncan, B.K. and Miller, J.H. (1980) Mutagenic deamination of cytosine residues in DNA. *Nature,* **287,** 560-561.

Dunham, I., Shimizu, N., Roe, B.A., Chissoe, S., Hunt, A.R., Collins, J.E., Bruskiewich, R., Beare, D.M., Clamp, M., Smink, L.J., et al (1999) The DNA sequence of human chromosome 22. *Nature,* **402,** 489-495.

Dunzinger, U., Nanda, I., Schmid, M., Haaf, T. and Zechner, U. (2005) Chicken orthologues of mammalian imprinted genes are clustered on macrochromosomes and replicate asynchronously. *Trends Genet.,* **21,** 488-492.

Duret, L., Chureau, C., Samain, S., Weissenbach, J. and Avner, P. (2006) The Xist RNA gene evolved in eutherians by pseudogenization of a protein-coding gene. *Science,* **312,** 1653-1655.

Easton, D.F., Pooley, K.A., Dunning, A.M., Pharoah, P.D., Thompson, D., Ballinger, D.G., Struewing, J.P., Morrison, J., Field, H., Luben, R., et al (2007) Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature,* **447,** 1087-1093.

Eddy, S.R. (2005) A model of the statistical power of comparative genome sequence analysis. *PLoS Biol.,* **3,** e10.

Edwards, C.A. and Ferguson-Smith, A.C. (2007) Mechanisms regulating imprinted genes in clusters. *Curr. Opin. Cell Biol.,* **19,** 281-289.

Edwards, C.A., Rens, W., Clark, O., Mungall, A.J., Hore, T., Marshall Graves, J.A., Dunham, I., Ferguson-Smith, A.C. and Ferguson-Smith, M.A. (2007) The evolution of imprinting: chromosomal mapping of orthologues of mammalian imprinted domains in monotreme and marsupial mammals. *BMC Evol. Biol.,* **7,** 157.

Elgar, G. (2006) Different words, same meaning: understanding the languages of the genome. *Trends Genet.,* **22,** 639-641.

ENCODE Project Consortium. (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science,* **306,** 636-640.

ENCODE Project Consortium, Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigo, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T., et al (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature,* **447,** 799-816.

Engemann, S., Strodicke, M., Paulsen, M., Franck, O., Reinhardt, R., Lane, N., Reik, W. and Walter, J. (2000) Sequence and functional comparison in the Beckwith-Wiedemann region: implications for a novel imprinting centre and extended imprinting. *Hum. Mol. Genet.,* **9,** 2691-2706.

Estivill, X., Cheung, J., Pujana, M.A., Nakabayashi, K., Scherer, S.W. and Tsui, L.C. (2002) Chromosomal regions containing high-density and ambiguously mapped putative single nucleotide polymorphisms (SNPs) correlate with segmental duplications in the human genome. *Hum. Mol. Genet.*, **11,** 1987-1995.

Feinberg, A.P. and Vogelstein, B. (1983) A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity. *Anal. Biochem.*, **132,** 6-13.

Fichant, G.A. and Burks, C. (1991) Identifying potential tRNA genes in genomic DNA sequences. *J. Mol. Biol.*, **220,** 659-671.

Finn, R.D., Mistry, J., Schuster-Bockler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R., et al (2006) Pfam: clans, web tools and services. *Nucleic Acids Res.*, **34,** D247-51.

Fisher, S., Grice, E.A., Vinton, R.M., Bessling, S.L. and McCallion, A.S. (2006) Conservation of RET regulatory function from human to zebrafish without sequence similarity. *Science,* **312,** 276-279.

Fitzpatrick, G.V., Pugacheva, E.M., Shin, J.Y., Abdullaev, Z., Yang, Y., Khatod, K., Lobanenkov, V.V. and Higgins, M.J. (2007) Allele-specific binding of CTCF to the multipartite imprinting control region KvDMR1. *Mol. Cell. Biol.*, **27,** 2636-2647.

Fitzpatrick, G.V., Soloway, P.D. and Higgins, M.J. (2002) Regional loss of imprinting and growth deficiency in mice with a targeted deletion of KvDMR1. *Nat. Genet.*, **32,** 426-431.

Frazer, K.A., Elnitski, L., Church, D.M., Dubchak, I. and Hardison, R.C. (2003) Cross-species sequence comparisons: a review of methods and available resources. *Genome Res.*, **13,** 1-12.

Frazer, K.A., Pachter, L., Poliakov, A., Rubin, E.M. and Dubchak, I. (2004) VISTA: computational tools for comparative genomics. *Nucleic Acids Res.*, **32,** W273-9.

Fredlake, C.P., Hert, D.G., Mardis, E.R. and Barron, A.E. (2006) What is the future of electrophoresis in large-scale genomic sequencing? *Electrophoresis*, **27,** 3689-3702.

Gardiner-Garden, M. and Frommer, M. (1987) CpG islands in vertebrate genomes. *J. Mol. Biol.*, **196,** 261-282.

Gaszner, M. and Felsenfeld, G. (2006) Insulators: Exploiting transcriptional and epigenetic mechanisms. *Nat. Rev. Gen.*, **7,** 703-713.

Gerstein, M.B., Bruce, C., Rozowsky, J.S., Zheng, D., Du, J., Korbel, J.O., Emanuelsson, O., Zhang, Z.D., Weissman, S. and Snyder, M. (2007) What is a gene, post-ENCODE? History and updated definition. *Genome Res.*, **17,** 669-681.

Gibbs, R.A., Weinstock, G.M., Metzker, M.L., Muzny, D.M., Sodergren, E.J., Scherer, S., Scott, G., Steffen, D., Worley, K.C., Burch, P.E., et al (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature*, **428,** 493-521.

Gilbert, N. and Labuda, D. (2000) Evolutionary inventions and continuity of CORE-SINEs in mammals. *J. Mol. Biol.*, **298,** 365-377.

Gilbert, N. and Labuda, D. (1999) CORE-SINEs: eukaryotic short interspersed retroposing elements with common sequence motifs. *Proc. Natl. Acad. Sci. U. S. A.*, **96,** 2869-2874.

Giresi, P.G., Kim, J., McDaniell, R.M., Iyer, V.R. and Lieb, J.D. (2007) FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res.*, **17,** 877-885.

Glass, J.L., Thompson, R.F., Khulan, B., Figueroa, M.E., Olivier, E.N., Oakley, E.J., Van Zant, G., Bouhassira, E.E., Melnick, A., Golden, A., et al (2007) CG dinucleotide clustering is a species-specific property of the genome. *Nucleic Acids Res.*,

Godbout, R., Ingram, R.S. and Tilghman, S.M. (1988) Fine-structure mapping of the three mouse alpha-fetoprotein gene enhancers. *Mol. Cell. Biol.*, **8,** 1169-1178.

Gottgens, B., Barton, L.M., Gilbert, J.G., Bench, A.J., Sanchez, M.J., Bahn, S., Mistry, S., Grafham, D., McMurray, A., Vaudin, M., et al (2000) Analysis of vertebrate SCL loci identifies conserved enhancers. *Nat. Biotechnol.*, **18,** 181-186.

Gould, S.J. and Vrba, E.S. (1982) Exaptation; a missing term in the science of form. *Paleobiology*, **8,** 4-15.

Graves, J.A. and Westerman, M. (2002) Marsupial genetics and genomics. *Trends Genet.*, **18,** 517-521.

Greally, J.M. (2002) Short interspersed transposable elements (SINEs) are excluded from imprinted regions in the human genome. *Proc. Natl. Acad. Sci. U. S. A.*, **99,** 327-332.

Green, P. (2007) 2x Genomes--does Depth Matter? *Genome Res.*, **17,** 1547-1549.

Gregory, S.G., Sekhon, M., Schein, J., Zhao, S., Osoegawa, K., Scott, C.E., Evans, R.S., Burridge, P.W., Cox, T.V., Fox, C.A., et al (2002) A physical map of the mouse genome. *Nature,* **418,** 743-750.

Griffiths-Jones, S., Grocock, R.J., van Dongen, S., Bateman, A. and Enright, A.J. (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.,* **34,** D140-4.

Grimson, A., Farh, K.K., Johnston, W.K., Garrett-Engele, P., Lim, L.P. and Bartel, D.P. (2007) MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol. Cell,* **27,** 91-105.

Grutzner, F., Rens, W., Tsend-Ayush, E., El-Mogharbel, N., O'Brien, P.C., Jones, R.C., Ferguson-Smith, M.A. and Marshall Graves, J.A. (2004) In the platypus a meiotic chain of ten sex chromosomes shares genes with the bird Z and mammal X chromosomes. *Nature,* **432,** 913-917.

Guenet, J.L. and Bonhomme, F. (2003) Wild mice: an ever-increasing contribution to a popular mammalian model. *Trends Genet.,* **19,** 24-31.

Guigo, R., Flicek, P., Abril, J.F., Reymond, A., Lagarde, J., Denoeud, F., Antonarakis, S., Ashburner, M., Bajic, V.B., Birney, E., et al (2006) EGASP: the human ENCODE Genome Annotation Assessment Project. *Genome Biol.,* **7 Suppl 1,** S2.1-31.

Haig, D. and Westoby, M. (1989) Parent-Specific Gene Expression and the Triploid Endosperm. *Am. Nat.,* **134,** 147-155.

Haig, D. (2004) Genomic imprinting and kinship: how good is the evidence? *Annu. Rev. Genet.,* **38,** 553-585.

Haig, D. (1994) Refusing the ovarian time bomb. *Trends Genet.,* **10,** 346-7; author reply 348-9.

Haig, D. and Westoby, M. (2006) An earlier formulation of the genetic conflict hypothesis of genomic imprinting. *Nat. Genet.,* **38,** 271.

Hammond, S.M., Bernstein, E., Beach, D. and Hannon, G.J. (2000) An RNA-directed nuclease mediates post-transcriptional gene silencing in Drosophila cells. *Nature,* **404,** 293-296.

Han, J., Lee, Y., Yeom, K.H., Kim, Y.K., Jin, H. and Kim, V.N. (2004) The Drosha-DGCR8 complex in primary microRNA processing. *Genes Dev.,* **18,** 3016-3027.

Hardison, R.C., Oeltjen, J. and Miller, W. (1997) Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. *Genome Res.*, **7,** 959-966.

Hardison, R.C., Roskin, K.M., Yang, S., Diekhans, M., Kent, W.J., Weber, R., Elnitski, L., Li, J., O'Connor, M., Kolbe, D., et al (2003) Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. *Genome Res.*, **13,** 13-26.

Hark, A.T., Schoenherr, C.J., Katz, D.J., Ingram, R.S., Levorse, J.M. and Tilghman, S.M. (2000) CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. *Nature*, **405,** 486-489.

Harrow, J., Denoeud, F., Frankish, A., Reymond, A., Chen, C.K., Chrast, J., Lagarde, J., Gilbert, J.G., Storey, R., Swarbreck, D., et al (2006) GENCODE: producing a reference annotation for ENCODE. *Genome Biol.*, **7 Suppl 1,** S4.1-9.

Hayashizaki, Y., Shibata, H., Hirotsune, S., Sugino, H., Okazaki, Y., Sasaki, N., Hirose, K., Imoto, H., Okuizumi, H. and Muramatsu, M. (1994) Identification of an imprinted U2af binding protein related sequence on mouse chromosome 11 using the RLGS method. *Nat. Genet.*, **6,** 33-40.

Heard, E. (2005) Delving into the diversity of facultative heterochromatin: the epigenetics of the inactive X chromosome. *Curr. Opin. Genet. Dev.*, **15,** 482-489.

Heintzman, N.D., Stuart, R.K., Hon, G., Fu, Y., Ching, C.W., Hawkins, R.D., Barrera, L.O., Van Calcar, S., Qu, C., Ching, K.A., et al (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.*, **39,** 311-318.

Hillier, L.W., Miller, W., Birney, E., Warren, W., Hardison, R.C., Ponting, C.P., Bork, P., Burt, D.W., Groenen, M.A., Delany, M.E., et al (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, **432,** 695-716.

Hodgson, C.P. and Fisk, R.Z. (1987) Hybridization probe size control: optimized 'oligolabelling'. *Nucleic Acids Res.*, **15,** 6295.

Horak, C.E., Mahajan, M.C., Luscombe, N.M., Gerstein, M., Weissman, S.M. and Snyder, M. (2002) GATA-1 binding sites mapped in the beta-globin locus by using mammalian chIp-chip analysis. *Proc. Natl. Acad. Sci. U. S. A.*, **99,** 2924-2929.

Humphray, S.J., Scott, C.E., Clark, R., Marron, B., Bender, C., Camm, N., Davis, J., Jenks, A., Noon, A., Patel, M., et al (2007) A high utility integrated map of the pig genome. *Genome Biol.*, **8,** R139.

Hurst, L.D. (1998) Peromysci, promiscuity and imprinting. *Nat. Genet.*, **20,** 315-316.

Hurst, L.D. and McVean, G.T. (1998) Do we understand the evolution of genomic imprinting? *Curr. Opin. Genet. Dev.*, **8,** 701-708.

Hurst, L.D. and McVean, G.T. (1997) Growth effects of uniparental disomies and the conflict theory of genomic imprinting. *Trends Genet.*, **13,** 436-443.

Hutvagner, G., McLachlan, J., Pasquinelli, A.E., Balint, E., Tuschl, T. and Zamore, P.D. (2001) A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. *Science,* **293,** 834-838.

Huynh, K.D. and Lee, J.T. (2005) X-chromosome inactivation: a hypothesis linking ontogeny and phylogeny. *Nat. Rev. Genet.*, **6,** 410-418.

International Human Genome Sequencing Consortium. (2004) Finishing the euchromatic sequence of the human genome. *Nature,* **431,** 931-945.

International Human Genome Sequencing Consortium. (2001) Initial sequencing and analysis of the human genome. *Nature,* **409,** 860-921.

Ishihara, K., Hatano, N., Furuumi, H., Kato, R., Iwaki, T., Miura, K., Jinno, Y. and Sasaki, H. (2000) Comparative genomic sequencing identifies novel tissue-specific enhancers and sequence elements for methylation-sensitive factors implicated in Igf2/H19 imprinting. *Genome Res.*, **10,** 664-671.

Ishihara, K., Oshimura, M. and Nakao, M. (2006) CTCF-dependent chromatin insulator is linked to epigenetic remodeling. *Mol. Cell,* **23,** 733-742.

Jabbari, K. and Bernardi, G. (2004) Cytosine methylation and CpG, TpG (CpA) and TpA frequencies. *Gene,* **333,** 143-149.

Jabbari, K., Caccio, S., Pais de Barros, J.P., Desgres, J. and Bernardi, G. (1997) Evolutionary changes in CpG and methylation levels in the genome of vertebrates. *Gene,* **205,** 109-118.

Jinno, Y., Yun, K., Nishiwaki, K., Kubota, T., Ogawa, O., Reeve, A.E. and Niikawa, N. (1994) Mosaic and polymorphic imprinting of the WT1 gene in humans. *Nat. Genet.*, **6,** 305-309.

Jones, B.K., Levorse, J.M. and Tilghman, S.M. (1998) Igf2 imprinting does not require its own DNA methylation or H19 RNA. *Genes Dev.*, **12,** 2200-2207.

Ke, X., Thomas, N.S., Robinson, D.O. and Collins, A. (2002) The distinguishing sequence characteristics of mouse imprinted genes. *Mamm. Genome,* **13,** 639-645.

Kellis, M., Patterson, N., Endrizzi, M., Birren, B. and Lander, E.S. (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature,* **423,** 241-254.

Kent, W.J. (2002) BLAT--the BLAST-like alignment tool. *Genome Res.,* **12,** 656-664.

Ketting, R.F., Fischer, S.E., Bernstein, E., Sijen, T., Hannon, G.J. and Plasterk, R.H. (2001) Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in C. elegans. *Genes Dev.,* **15,** 2654-2659.

Killian, J.K., Byrd, J.C., Jirtle, J.V., Munday, B.L., Stoskopf, M.K., MacDonald, R.G. and Jirtle, R.L. (2000) M6P/IGF2R imprinting evolution in mammals. *Mol. Cell,* **5,** 707-716.

Killian, J.K., Nolan, C.M., Stewart, N., Munday, B.L., Andersen, N.A., Nicol, S. and Jirtle, R.L. (2001) Monotreme IGF2 expression and ancestral origin of genomic imprinting. *J. Exp. Zool.,* **291,** 205-212.

Kim, T.H., Abdullaev, Z.K., Smith, A.D., Ching, K.A., Loukinov, D.I., Green, R.D., Zhang, M.Q., Lobanenkov, V.V. and Ren, B. (2007) Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell,* **128,** 1231-1245.

Kimura, M. (1983) *The Neutral Theory of Molecular Evolution.* Cambridge University Press, Cambridge.

Koch, C.M., Andrews, R.M., Flicek, P., Dillon, S.C., Karaoz, U., Clelland, G.K., Wilcox, S., Beare, D.M., Fowler, J.C., Couttet, P., et al (2007) The landscape of histone modifications across 1% of the human genome in five human cell lines. *Genome Res.,* **17,** 691-707.

Kono, T. (2006) Genomic imprinting is a barrier to parthenogenesis in mammals. *Cytogenet. Genome Res.,* **113,** 31-35.

Kono, T., Obata, Y., Wu, Q., Niwa, K., Ono, Y., Yamamoto, Y., Park, E.S., Seo, J.S. and Ogawa, H. (2004) Birth of parthenogenetic mice that can develop to adulthood. *Nature,* **428,** 860-864.

Kumar, S. and Hedges, S.B. (1998) A molecular timescale for vertebrate evolution. *Nature,* **392,** 917-920.

Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C. and Salzberg, S.L. (2004) Versatile and open software for comparing large genomes. *Genome Biol.*, **5,** R12.

Kurukuti, S., Tiwari, V.K., Tavoosidana, G., Pugacheva, E., Murrell, A., Zhao, Z., Lobanenkov, V., Reik, W. and Ohlsson, R. (2006) CTCF binding at the H19 imprinting control region mediates maternally inherited higher-order chromatin conformation to restrict enhancer access to Igf2. *Proc. Natl. Acad. Sci. U. S. A.*, **103,** 10684-10689.

Lachner, M., O'Sullivan, R.J. and Jenuwein, T. (2003) An epigenetic road map for histone lysine methylation. *J. Cell. Sci.*, **116,** 2117-2124.

Laemmli, U.K., Kas, E., Poljak, L. and Adachi, Y. (1992) Scaffold-associated regions: cis-acting determinants of chromatin structural loops and functional domains. *Curr. Opin. Genet. Dev.*, **2,** 275-285.

Lalande, M. (1996) Parental imprinting and human disease. *Annu. Rev. Genet.*, **30,** 173-195.

Lamesch, P., Li, N., Milstein, S., Fan, C., Hao, T., Szabo, G., Hu, Z., Venkatesan, K., Bethel, G., Martin, P., et al (2007) hORFeome v3.1: a resource of human open reading frames representing over 10,000 human genes. *Genomics,* **89,** 307-315.

Landy, A. (1989) Dynamic, structural, and regulatory aspects of lambda site-specific recombination. *Annu. Rev. Biochem.*, **58,** 913-949.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., et al (2007) Clustal W and Clustal X version 2.0. *Bioinformatics,* **23,** 2947-2948.

Lawton, B.R., Obergfell, C., O'Neill, R.J. and O'Neill, M.J. (2007) Physical mapping of the IGF2 locus in the South American opossum Monodelphis domestica. *Cytogenet. Genome Res.*, **116,** 130-131.

Lee, J.T. (2003) Molecular links between X-inactivation and autosomal imprinting: X-inactivation as a driving force for the evolution of imprinting? *Curr. Biol.*, **13,** R242-54.

Lefevre, C.M., Digby, M.R., Whitley, J.C., Strahm, Y. and Nicholas, K.R. (2007) Lactation transcriptomics in the Australian marsupial, Macropus eugenii: transcript sequencing and quantification. *BMC Genomics*, **8,** 417.

Leighton, P.A., Saam, J.R., Ingram, R.S., Stewart, C.L. and Tilghman, S.M. (1995) An enhancer deletion affects both H19 and Igf2 expression. *Genes Dev.*, **9,** 2079-2089.

Lettice, L.A., Heaney, S.J., Purdie, L.A., Li, L., de Beer, P., Oostra, B.A., Goode, D., Elgar, G., Hill, R.E. and de Graaff, E. (2003) A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum. Mol. Genet.*, **12,** 1725-1735.

Lewis, A. and Reik, W. (2006) How imprinting centres work. *Cytogenet. Genome Res.*, **113,** 81-89.

Li, E., Beard, C., Forster, A.C., Bestor, T.H. and Jaenisch, R. (1993a) DNA methylation, genomic imprinting, and mammalian development. *Cold Spring Harb. Symp. Quant. Biol.*, **58,** 297-305.

Li, E., Beard, C. and Jaenisch, R. (1993b) Role for DNA methylation in genomic imprinting. *Nature*, **366,** 362-365.

Lindblad-Toh, K., Winchester, E., Daly, M.J., Wang, D.G., Hirschhorn, J.N., Laviolette, J.P., Ardlie, K., Reich, D.E., Robinson, E., Sklar, P., et al (2000) Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse. *Nat. Genet.*, **24,** 381-386.

Lingenfelter, P.A., Delbridge, M.L., Thomas, S., Hoekstra, H.E., Mitchell, M.J., Graves, J.A. and Disteche, C.M. (2001) Expression and conservation of processed copies of the RBMX gene. *Mamm. Genome*, **12,** 538-545.

Long, L. and Spear, B.T. (2004) FoxA proteins regulate H19 endoderm enhancer E1 and exhibit developmental changes in enhancer binding in vivo. *Mol. Cell. Biol.*, **24,** 9601-9609.

Loots, G. and Ovcharenko, I. (2007) ECRbase: database of evolutionary conserved regions, promoters, and transcription factor binding sites in vertebrate genomes. *Bioinformatics*, **23,** 122-124.

Loots, G.G., Locksley, R.M., Blankespoor, C.M., Wang, Z.E., Miller, W., Rubin, E.M. and Frazer, K.A. (2000) Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science*, **288,** 136-140.

Loots, G.G. and Ovcharenko, I. (2005) Dcode.org anthology of comparative genomic tools. *Nucleic Acids Res.*, **33,** W56-64.

Loots, G.G. and Ovcharenko, I. (2004) rVISTA 2.0: evolutionary analysis of transcription factor binding sites. *Nucleic Acids Res.*, **32,** W217-21.

Lowe, C.B., Bejerano, G. and Haussler, D. (2007) Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. *Proc. Natl. Acad. Sci. U. S. A.*, **104,** 8005-8010.

Lowe, T.M. and Eddy, S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, **25,** 955-964.

Luedi, P.P., Dietrich, F.S., Weidman, J.R., Bosko, J.M., Jirtle, R.L. and Hartemink, A.J. (2007) Computational and experimental identification of novel human imprinted genes. *Genome Res.*, **17,** 1723-1730.

Luedi, P.P., Hartemink, A.J. and Jirtle, R.L. (2005) Genome-wide prediction of imprinted murine genes. *Genome Res.*, **15,** 875-884.

Lyko, F., Brenton, J.D., Surani, M.A. and Paro, R. (1997) An imprinting element from the mouse H19 locus functions as a silencer in Drosophila. *Nat. Genet.*, **16,** 171-173.

Lyon, M.F. (1998) X-chromosome inactivation: a repeat hypothesis. *Cytogenet. Cell Genet.*, **80,** 133-137.

Ma, B., Tromp, J. and Li, M. (2002) PatternHunter: faster and more sensitive homology search. *Bioinformatics,* **18,** 440-445.

Maglott, D., Ostell, J., Pruitt, K.D. and Tatusova, T. (2007) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **35,** D26-31.

Maizel, J.V.,Jr and Lenk, R.P. (1981) Enhanced graphic matrix analysis of nucleic acid and protein sequences. *Proc. Natl. Acad. Sci. U. S. A.*, **78,** 7665-7669.

Mancini-Dinardo, D., Steele, S.J., Levorse, J.M., Ingram, R.S. and Tilghman, S.M. (2006) Elongation of the Kcnq1ot1 transcript is required for genomic imprinting of neighboring genes. *Genes Dev.*, **20,** 1268-1282.

Margueron, R., Trojer, P. and Reinberg, D. (2005) The key to development: interpreting the histone code? *Curr. Opin. Genet. Dev.*, **15,** 163-176.

Margulies, E.H., Blanchette, M., Haussler, D., Green, E.D. and NISC Comparative Sequencing Program. (2003a) Identification and characterization of multi-species conserved sequences. *Genome Res.*, **13,** 2507-2518.

Margulies, E.H., Green, E.D. and NISC Comparative Sequencing Program. (2003b) Detecting highly conserved regions of the human genome by multispecies sequence comparisons. *Cold Spring Harb. Symp. Quant. Biol.*, **68,** 255-263.

Margulies, E.H., Maduro, V.V., Thomas, P.J., Tomkins, J.P., Amemiya, C.T., Luo, M., Green, E.D. and NISC Comparative Sequencing Program. (2005a) Comparative sequencing provides insights about the structure and conservation of marsupial and monotreme genomes. *Proc. Natl. Acad. Sci. U. S. A.*, **102,** 3354-3359.

Margulies, E.H., Vinson, J.P., Miller, W., Jaffe, D.B., Lindblad-Toh, K., Chang, J.L., Green, E.D., Lander, E.S., Mullikin, J.C., Clamp, M., et al (2005b) An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing. *Proc. Natl. Acad. Sci. U. S. A.*, **102,** 4795-4800.

Marra, M.A., Kucaba, T.A., Dietrich, N.L., Green, E.D., Brownstein, B., Wilson, R.K., McDonald, K.M., Hillier, L.W., McPherson, J.D. and Waterston, R.H. (1997) High throughput fingerprint analysis of large-insert clones. *Genome Res.*, **7,** 1072-1084.

Marsters, S.A., Frutkin, A.D., Simpson, N.J., Fendly, B.M. and Ashkenazi, A. (1992) Identification of cysteine-rich domains of the type 1 tumor necrosis factor receptor involved in ligand binding. *J. Biol. Chem.*, **267,** 5747-5750.

Martinez, J., Patkaniowska, A., Urlaub, H., Luhrmann, R. and Tuschl, T. (2002) Single-stranded antisense siRNAs guide target RNA cleavage in RNAi. *Cell*, **110,** 563-574.

Maston, G.A., Evans, S.K. and Green, M.R. (2006) Transcriptional Regulatory Elements in the Human Genome. *Annu. Rev. Genomics Hum. Genet.*, **7,** 29-59.

Matys, V., Kel-Margoulis, O.V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., et al (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, **34,** D108-10.

Mayor, C., Brudno, M., Schwartz, J.R., Poliakov, A., Rubin, E.M., Frazer, K.A., Pachter, L.S. and Dubchak, I. *(2000)* VISTA : visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics*, **16,** 1046-1047.

McDonald, J.F., Matzke, M.A. and Matzke, A.J. (2005) Host defenses to transposable elements and the evolution of genomic imprinting. *Cytogenetic and Genome Research,* **110,** 242-249.

McGowan, R.A. and Martin, C.C. (1997) DNA methylation and genome imprinting in the zebrafish, Danio rerio: some evolutionary ramifications. *Biochem. Cell Biol.*, **75,** 499-506.

McGrath, J. and Solter, D. (1984) Completion of mouse embryogenesis requires both the maternal and paternal genomes. *Cell,* **37,** 179-183.

McPherson, J.D., Marra, M., Hillier, L., Waterston, R.H., Chinwalla, A., Wallis, J., Sekhon, M., Wylie, K., Mardis, E.R., Wilson, R.K., et al (2001) A physical map of the human genome. *Nature,* **409,** 934-941.

MENDEL, G. (1950) Gregor Mendel's letters to Carl Nageli, 1866-1873. *Genetics,* **35,** 1-29.

Mikkelsen, T.S., Wakefield, M.J., Aken, B., Amemiya, C.T., Chang, J.L., Duke, S., Garber, M., Gentles, A.J., Goodstadt, L., Heger, A., et al (2007) Genome of the marsupial Monodelphis domestica reveals innovation in non-coding sequences. *Nature,* **447,** 167-177.

Mineno, J., Okamoto, S., Ando, T., Sato, M., Chono, H., Izu, H., Takayama, M., Asada, K., Mirochnitchenko, O., Inouye, M., et al (2006) The expression profile of microRNAs in mouse embryos. *Nucleic Acids Res.*, **34,** 1765-1771.

Mitsuya, K., Sui, H., Meguro, M., Kugoh, H., Jinno, Y., Niikawa, N. and Oshimura, M. (1997) Paternal expression of WT1 in human fibroblasts and lymphocytes. *Hum. Mol. Genet.*, **6,** 2243-2246.

Moore, T. (1994) Refusing the ovarian time bomb. *Trends Genet.*, **10,** 347-349.

Moore, T. and Haig, D. (1991) Genomic imprinting in mammalian development: a parental tug-of-war. *Trends Genet.*, **7,** 45-49.

Morgan, T.H. (1915) Localization of the Hereditary Material in the Germ Cells. *Proc. Natl. Acad. Sci. U. S. A.*, **1,** 420-429.

Morgenstern, B. (1999) DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics,* **15,** 211-218.

Morison, I.M., Ramsay, J.P. and Spencer, H.G. (2005) A census of mammalian imprinting. *Trends Genet.*, **21,** 457-465.

Morrow, C.P., Curtin, J.P. and Townsend, D.E. (1993) Tumors of the ovary: classification; the adnexal mass. *Synopsis of gynecologic oncology.4th ed.New York: Churchill Livingstone,* **224,**

Mott, R. (1997) EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. *Comput. Appl. Biosci.*, **13,** 477-478.

Mungall, A.J. and Humphray, S.J. (2003) Assembling physical maps and sequence clone selection. In Dunham, I. (ed) *Genome Mapping and Sequencing.* Horizon Scientific Press, Cambridge, UK, 167-200.

Mungall, A.J., Humphray, S.J., Ranby, S.A., Edwards, C.A., Heathcott, R.W., Clee, C.M., Holloway, E., Peck, A.I., Harrison, P., Green, L.D., et al (1997) From long range mapping to sequence-ready contigs on human chromosome 6. *DNA Seq.*, **8,** 151-154.

Murphy, S.K. and Jirtle, R.L. (2003) Imprinting evolution and the price of silence. *Bioessays,* **25,** 577-588.

Murphy, W.J., Larkin, D.M., Everts-van der Wind, A., Bourque, G., Tesler, G., Auvil, L., Beever, J.E., Chowdhary, B.P., Galibert, F., Gatzke, L., et al (2005) Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science,* **309,** 613-617.

Murrell, A., Heeson, S. and Reik, W. (2004) Interaction between differentially methylated regions partitions the imprinted genes Igf2 and H19 into parent-specific chromatin loops. *Nat. Genet.*, **36,** 889-893.

Nardone, J., Lee, D.U., Ansel, K.M. and Rao, A. (2004) Bioinformatics for the 'bench biologist': how to find regulatory regions in genomic DNA. *Nat. Immunol.*, **5,** 768-774.

Needleman, S.B. and Wunsch, C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48,** 443-453.

Nilsson, M.A., Arnason, U., Spencer, P.B. and Janke, A. (2004) Marsupial relationships and a timeline for marsupial radiation in South Gondwana. *Gene,* **340,** 189-196.

Nobrega, M.A., Ovcharenko, I., Afzal, V. and Rubin, E.M. (2003) Scanning human gene deserts for long-range enhancers. *Science,* **302,** 413.

Nolan, C.M., Killian, J.K., Petitte, J.N. and Jirtle, R.L. (2001) Imprint status of M6P/IGF2R and IGF2 in chickens. *Dev. Genes Evol.*, **211,** 179-183.

Obernosterer, G., Leuschner, P.J., Alenius, M. and Martinez, J. (2006) Post-transcriptional regulation of microRNA expression. *RNA*, **12,** 1161-1167.

Ohlsson, R., Renkawitz, R. and Lobanenkov, V. (2001) CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease. *Trends Genet.*, **17,** 520-527.

Ohshima, K. and Okada, N. (2005) SINEs and LINEs: symbionts of eukaryotic genomes with a common tail. *Cytogenet. Genome Res.*, **110,** 475-490.

Olek, A. and Walter, J. (1997) The pre-implantation ontogeny of the H19 methylation imprint. *Nat. Genet.*, **17,** 275-276.

O'Neill, M.J., Ingram, R.S., Vrana, P.B. and Tilghman, S.M. (2000) Allelic expression of IGF2 in marsupials and birds. *Dev. Genes Evol.*, **210,** 18-20.

Ono, R., Kobayashi, S., Wagatsuma, H., Aisaka, K., Kohda, T., Kaneko-Ishino, T. and Ishino, F. (2001) A retrotransposon-derived gene, PEG10, is a novel imprinted gene located on human chromosome 7q21. *Genomics*, **73,** 232-237.

Osoegawa, K., Tateno, M., Woon, P.Y., Frengen, E., Mammoser, A.G., Catanese, J.J., Hayashizaki, Y. and de Jong, P.J. (2000) Bacterial artificial chromosome libraries for mouse sequencing and functional analysis. *Genome Res.*, **10,** 116-128.

Ovcharenko, I., Loots, G.G., Giardine, B.M., Hou, M., Ma, J., Hardison, R.C., Stubbs, L. and Miller, W. (2005) Mulan: multiple-sequence local alignment and visualization for studying function and evolution. *Genome Res.*, **15,** 184-194.

Ovcharenko, I., Loots, G.G., Hardison, R.C., Miller, W. and Stubbs, L. (2004a) zPicture: dynamic alignment and visualization tool for analyzing conservation profiles. *Genome Res.*, **14,** 472-477.

Ovcharenko, I., Stubbs, L. and Loots, G.G. (2004b) Interpreting mammalian evolution using Fugu genome comparisons. *Genomics*, **84,** 890-895.

Owen, R. (1849) *On parthenogenesis: or the successive production of procreating individuals from a single ovum.* John Van Voorst, London, London.

Pauler, F.M. and Barlow, D.P. (2006) Imprinting mechanisms--it only takes two. *Genes Dev.*, **20,** 1203-1206.

Pauler, F.M., Koerner, M.V. and Barlow, D.P. (2007) Silencing by imprinted noncoding RNAs: is transcription the answer? *Trends Genet.*, **23,** 284-292.

Paulsen, M., Khare, T., Burgard, C., Tierling, S. and Walter, J. (2005) Evolution of the Beckwith-Wiedemann syndrome region in vertebrates. *Genome Res.*, **15,** 146-153.

Pennacchio, L.A., Ahituv, N., Moses, A.M., Prabhakar, S., Nobrega, M.A., Shoukry, M., Minovitsky, S., Dubchak, I., Holt, A., Lewis, K.D., et al (2006) In vivo enhancer analysis of human conserved non-coding sequences. *Nature,* **444,** 499-502.

Pennisi, E. (2003) Human genome. A low number wins the GeneSweep Pool. *Science,* **300,** 1484.

Peters, A.H. and Schubeler, D. (2005) Methylation of histones: playing memory with DNA. *Curr. Opin. Cell Biol.,* **17,** 230-238.

Pevzner, P. and Tesler, G. (2003) Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. *Proc. Natl. Acad. Sci. U. S. A.,* **100,** 7672-7677.

Plass, C., Yu, F., Yu, L., Strout, M.P., El-Rifai, W., Elonen, E., Knuutila, S., Marcucci, G., Young, D.C., Held, W.A., et al (1999) Restriction landmark genome scanning for aberrant methylation in primary refractory and relapsed acute myeloid leukemia; involvement of the WIT-1 gene. *Oncogene,* **18,** 3159-3165.

Prabhakar, S., Poulin, F., Shoukry, M., Afzal, V., Rubin, E.M., Couronne, O. and Pennacchio, L.A. (2006) Close sequence comparisons are sufficient to identify human cis-regulatory elements. *Genome Res.,* **16,** 855-863.

Pruitt, K.D., Tatusova, T. and Maglott, D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.,* **35,** D61-5.

Ptashne, M. (1992) *A Genetic Switch: Phage [lambda] and Higher Organisms.* Blackwell Publishing, Cambridge, MA.

Puttagunta, R., Gordon, L.A., Meyer, G.E., Kapfhamer, D., Lamerdin, J.E., Kantheti, P., Portman, K.M., Chung, W.K., Jenne, D.E., Olsen, A.S., et al (2000) Comparative maps of human 19p13.3 and mouse chromosome 10 allow identification of sequences at evolutionary breakpoints. *Genome Res.,* **10,** 1369-1380.

Rada-Iglesias, A., Wallerman, O., Koch, C., Ameur, A., Enroth, S., Clelland, G., Wester, K., Wilcox, S., Dovey, O.M., Ellis, P.D., et al (2005) Binding sites for metabolic disease related transcription factors inferred at base pair resolution by chromatin immunoprecipitation and genomic microarrays. *Hum. Mol. Genet.,* **14,** 3435-3447.

Rapkins, R.W., Hore, T., Smithwick, M., Ager, E., Pask, A.J., Renfree, M.B., Kohn, M., Hameister, H., Nicholls, R.D., Deakin, J.E., et al (2006) Recent assembly of an imprinted domain from non-imprinted components. *PLoS Genet.*, **2,** e182.

Reik, W., Constancia, M., Fowden, A., Anderson, N., Dean, W., Ferguson-Smith, A., Tycko, B. and Sibley, C. (2003) Regulation of supply and demand for maternal nutrients in mammals by imprinted genes. *J. Physiol.*, **547,** 35-44.

Reik, W. and Lewis, A. (2005) Co-evolution of X-chromosome inactivation and imprinting in mammals. *Nat. Rev. Genet.*, **6,** 403-410.

Reik, W. and Walter, J. (2001) Genomic imprinting: parental influence on the genome. *Nat. Rev. Genet.*, **2,** 21-32.

Reik, W. and Walter, J. (1998) Imprinting mechanisms in mammals. *Curr. Opin. Genet. Dev.*, **8,** 154-164.

Renda, M., Baglivo, I., Burgess-Beusse, B., Esposito, S., Fattorusso, R., Felsenfeld, G. and Pedone, P.V. (2007) Critical DNA binding interactions of the insulator protein CTCF: a small number of zinc fingers mediate strong binding, and a single finger-dna interaction controls binding at imprinted loci. *J. Biol. Chem.*, **282,** 33336-33345.

Renfree, M.B. and Blanden, D.R. (2000) Progesterone and oestrogen receptors in the female genital tract throughout pregnancy in tammar wallabies. *J. Reprod. Fertil.*, **119,** 121-128.

Renfree, M.B. and Shaw, G. (2000) Diapause. *Annu. Rev. Physiol.*, **62,** 353-375.

Rens, W., O'Brien, P.C., Grutzner, F., Clarke, O., Graphodatskaya, D., Tsend-Ayush, E., Trifonov, V.A., Skelton, H., Wallis, M.C., Johnston, S., et al (2007) The multiple sex chromosomes of platypus and echidna are not completely identical and several share homology with the avian Z. *Genome Biol.*, **8,** R243.

Rice, P., Longden, I. and Bleasby, A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.*, **16,** 276-277.

Rice, W.R. and Chippindale, A.K. (2001) Sexual recombination and the power of natural selection. *Science*, **294,** 555-559.

Ross, M.T., LaBrie, S., McPherson, J. and Stanton, V.P. (1999) Screening large-insert libraries by hybridization. In Dracopoli, N. C., Haines, J. L., Korf, B. R., et al (eds) *In Current protocols in human genetics.* John Wiley and Sons, New York, 1-52.

Roth, S.Y., Denu, J.M. and Allis, C.D. (2001) Histone acetyltransferases. *Annu. Rev. Biochem.*, **70,** 81-120.

Rozen, S. and Skaletsky, H. (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.*, **132,** 365-386.

Rual, J.F., Hirozane-Kishikawa, T., Hao, T., Bertin, N., Li, S., Dricot, A., Li, N., Rosenberg, J., Lamesch, P., Vidalain, P.O., et al (2004) Human ORFeome version 1.1: a platform for reverse proteomics. *Genome Res.*, **14,** 2128-2135.

Ruiz-Herrera, A., Castresana, J. and Robinson, T.J. (2006) Is mammalian chromosomal evolution driven by regions of genome fragility? *Genome Biol.*, **7,** R115.

Ryan, D., Rahimi, M., Lund, J., Mehta, R. and Parviz, B.A. (2007) Toward nanoscale genome sequencing. *Trends Biotechnol.*, **25,** 385-389.

Sabo, P.J., Kuehn, M.S., Thurman, R., Johnson, B.E., Johnson, E.M., Cao, H., Yu, M., Rosenzweig, E., Goldy, J., Haydock, A., et al (2006) Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays. *Nat. Methods,* **3,** 511-518.

Salamov, A.A. and Solovyev, V.V. (2000) Ab initio gene finding in Drosophila genomic DNA. *Genome Res.*, **10,** 516-522.

Salcedo, T., Geraldes, A. and Nachman, M.W. (2007) Nucleotide variation in wild and inbred mice. *Genetics,* **177,** 2277-2291.

Sandelin, A., Carninci, P., Lenhard, B., Ponjavic, J., Hayashizaki, Y. and Hume, D.A. (2007) Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nat. Rev. Genet.*, **8,** 424-436.

Sanger, F., Nicklen, S. and Coulson, A.R. (1977) DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.*, **74,** 5463-5467.

Santangelo, A.M., de Souza, F.S., Franchini, L.F., Bumaschny, V.F., Low, M.J. and Rubinstein, M. (2007) Ancient exaptation of a CORE-SINE retroposon into a highly conserved mammalian neuronal enhancer of the proopiomelanocortin gene. *PLoS Genet.*, **3,** 1813-1826.

Sasaki, H., Jones, P.A., Chaillet, J.R., Ferguson-Smith, A.C., Barton, S.C., Reik, W. and Surani, M.A. (1992) Parental imprinting: potentially active chromatin of the repressed maternal allele of the mouse insulin-like growth factor II (Igf2) gene. *Genes Dev.*, **6,** 1843-1856.

Schiltz, R.L., Mizzen, C.A., Vassilev, A., Cook, R.G., Allis, C.D. and Nakatani, Y. (1999) Overlapping but distinct patterns of histone acetylation by the human coactivators p300 and PCAF within nucleosomal substrates. *J. Biol. Chem.*, **274,** 1189-1192.

Schmutz, J., Wheeler, J., Grimwood, J., Dickson, M., Yang, J., Caoile, C., Bajorek, E., Black, S., Chan, Y.M., Denys, M., et al (2004) Quality assessment of the human genome sequence. *Nature*, **429,** 365-368.

Schneider, P., Olson, D., Tardivel, A., Browning, B., Lugovskoy, A., Gong, D., Dobles, M., Hertig, S., Hofmann, K., Van Vlijmen, H., et al (2003) Identification of a new murine tumor necrosis factor receptor locus that contains two novel murine receptors for tumor necrosis factor-related apoptosis-inducing ligand (TRAIL). *J. Biol. Chem.*, **278,** 5444-5454.

Schuler, G.D. (1997) Sequence mapping by electronic PCR. *Genome Res.*, **7,** 541-550.

Schwartz, S., Elnitski, L., Li, M., Weirauch, M., Riemer, C., Smit, A., NISC Comparative Sequencing Program, Green, E.D., Hardison, R.C. and Miller, W. (2003a) MultiPipMaker and supporting tools: Alignments and analysis of multiple genomic DNA sequences. *Nucleic Acids Res.*, **31,** 3518-3524.

Schwartz, S., Kent, W.J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R.C., Haussler, D. and Miller, W. (2003b) Human-mouse alignments with BLASTZ. *Genome Res.*, **13,** 103-107.

Schwartz, S., Zhang, Z., Frazer, K.A., Smit, A., Riemer, C., Bouck, J., Gibbs, R., Hardison, R. and Miller, W. (2000) PipMaker--a web server for aligning two genomic DNA sequences. *Genome Res.*, **10,** 577-586.

Schwarz, D.S., Hutvagner, G., Haley, B. and Zamore, P.D. (2002) Evidence that siRNAs function as guides, not primers, in the Drosophila and human RNAi pathways. *Mol. Cell*, **10,** 537-548.

Searle, S.M., Gilbert, J., Iyer, V. and Clamp, M. (2004) The otter annotation system. *Genome Res.*, **14,** 963-970.

Seitz, H., Royo, H., Bortolin, M.L., Lin, S.P., Ferguson-Smith, A.C. and Cavaille, J. (2004) A large imprinted microRNA gene cluster at the mouse Dlk1-Gtl2 domain. *Genome Res.*, **14,** 1741-1748.

Seitz, H., Youngson, N., Lin, S.P., Dalbert, S., Paulsen, M., Bachellerie, J.P., Ferguson-Smith, A.C. and Cavaille, J. (2003) Imprinted microRNA genes

transcribed antisense to a reciprocally imprinted retrotransposon-like gene. *Nat. Genet.*, **34,** 261-262.

Sharman, G.B. (1971) Late DNA replication in the paternally derived X chromosome of female kangaroos. *Nature,* **230,** 231-232.

She, X., Jiang, Z., Clark, R.A., Liu, G., Cheng, Z., Tuzun, E., Church, D.M., Sutton, G., Halpern, A.L. and Eichler, E.E. (2004) Shotgun sequence assembly and recent segmental duplications within the human genome. *Nature,* **431,** 927-930.

Shen, J.C., Rideout, W.M.,3rd and Jones, P.A. (1994) The rate of hydrolytic deamination of 5-methylcytosine in double-stranded DNA. *Nucleic Acids Res.*, **22,** 972-976.

Shendure, J., Mitra, R.D., Varma, C. and Church, G.M. (2004) Advanced sequencing technologies: methods and goals. *Nat. Rev. Genet.*, **5,** 335-344.

Shirohzu, H., Yokomine, T., Sato, C., Kato, R., Toyoda, A., Purbowasito, W., Suda, C., Mukai, T., Hattori, M., Okumura, K., et al (2004) A 210-kb segment of tandem repeats and retroelements located between imprinted subdomains of mouse distal chromosome 7. *DNA Res.*, **11,** 325-334.

Shizuya, H., Birren, B., Kim, U.J., Mancino, V., Slepak, T., Tachiiri, Y. and Simon, M. (1992) Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in Escherichia coli using an F-factor-based vector. *Proc. Natl. Acad. Sci. U. S. A.*, **89,** 8794-8797.

Siddle, H.V., Deakin, J.E., Baker, M.L., Miller, R.D. and Belov, K. (2006) Isolation of major histocompatibility complex Class I genes from the tammar wallaby (Macropus eugenii). *Immunogenetics,* **58,** 487-493.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15,** 1034-1050.

Sleutels, F., Zwart, R. and Barlow, D.P. (2002) The non-coding Air RNA is required for silencing autosomal imprinted genes. *Nature,* **415,** 810-813.

Smith, T.F. and Waterman, M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147,** 195-197.

Soderlund, C., Longden, I. and Mott, R. (1997) FPC: a system for building contigs from restriction fingerprinted clones. *Comput. Appl. Biosci.*, **13,** 523-535.

Soejima, H. and Wagstaff, J. (2005) Imprinting centers, chromatin structure, and disease. *J. Cell. Biochem.*, **95,** 226-233.

Solovyev, V.V., Salamov, A.A. and Lawrence, C.B. (1994) Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. *Nucleic Acids Res.*, **22,** 5156-5163.

Solter, D. (1994) Refusing the ovarian time bomb. *Trends Genet.*, **10,** 346; author reply 348-9.

Solter, D. (1988) Differential imprinting and expression of maternal and paternal genomes. *Annu. Rev. Genet.*, **22,** 127-146.

Sonnhammer, E.L. and Durbin, R. (1995) A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene,* **167,** GC1-10.

Spillane, C., Schmid, K.J., Laoueille-Duprat, S., Pien, S., Escobar-Restrepo, J.M., Baroux, C., Gagliardini, V., Page, D.R., Wolfe, K.H. and Grossniklaus, U. (2007) Positive darwinian selection at the imprinted MEDEA locus in plants. *Nature,* **448,** 349-352.

Staden, R. (1979) A strategy of DNA sequencing employing computer programs. *Nucleic Acids Res.*, **6,** 2601-2610.

Staden, R., Beal, K.F. and Bonfield, J.K. (2000) The Staden package, 1998. *Methods Mol. Biol.*, **132,** 115-130.

Stankiewicz, P., Park, S.S., Inoue, K. and Lupski, J.R. (2001) The evolutionary chromosome translocation 4;19 in Gorilla gorilla is associated with microduplication of the chromosome fragment syntenic to sequences surrounding the human proximal CMT1A-REP. *Genome Res.*, **11,** 1205-1210.

Strahl, B.D. and Allis, C.D. (2000) The language of covalent histone modifications. *Nature,* **403,** 41-45.

Sueoka, N. (1988) Directional mutation pressure and neutral molecular evolution. *Proc. Natl. Acad. Sci. U. S. A.*, **85,** 2653-2657.

Surani, M.A., Barton, S.C. and Norris, M.L. (1984) Development of reconstituted mouse eggs suggests imprinting of the genome during gametogenesis. *Nature,* **308,** 548-550.

Suzuki, S., Ono, R., Narita, T., Pask, A.J., Shaw, G., Wang, C., Kohda, T., Alsop, A.E., Marshall Graves, J.A., Kohara, Y., et al (2007) Retrotransposon silencing by DNA methylation can drive mammalian genomic imprinting. *PLoS Genet.*, **3,** e55.

Suzuki, S., Renfree, M.B., Pask, A.J., Shaw, G., Kobayashi, S., Kohda, T., Kaneko-Ishino, T. and Ishino, F. (2005) Genomic imprinting of IGF2, p57(KIP2) and PEG1/MEST in a marsupial, the tammar wallaby. *Mech. Dev.*, **122,** 213-222.

Sved, J. and Bird, A. (1990) The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. *Proc. Natl. Acad. Sci. U. S. A.*, **87,** 4692-4696.

Szabo, Z., Levi-Minzi, S.A., Christiano, A.M., Struminger, C., Stoneking, M., Batzer, M.A. and Boyd, C.D. (1999) Sequential loss of two neighboring exons of the tropoelastin gene during primate evolution. *J. Mol. Evol.*, **49,** 664-671.

Tagle, D.A., Koop, B.F., Goodman, M., Slightom, J.L., Hess, D.L. and Jones, R.T. (1988) Embryonic epsilon and gamma globin genes of a prosimian primate (Galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *J. Mol. Biol.*, **203,** 439-455.

The Arabidopsis Genome Initiative. (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature,* **408,** 796-815.

The C. elegans Sequencing Consortium. (1998) Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology. *Science,* **282,** 2012-2018.

Thomas, J.W., Touchman, J.W., Blakesley, R.W., Bouffard, G.G., Beckstrom-Sternberg, S.M., Margulies, E.H., Blanchette, M., Siepel, A.C., Thomas, P.J., McDowell, J.C., et al (2003) Comparative analyses of multi-species sequences from targeted genomic regions. *Nature,* **424,** 788-793.

Thomson, J.M., Newman, M., Parker, J.S., Morin-Kensicki, E.M., Wright, T. and Hammond, S.M. (2006) Extensive post-transcriptional regulation of microRNAs and its implications for cancer. *Genes Dev.*, **20,** 2202-2207.

Thorvaldsen, J.L., Mann, M.R., Nwoko, O., Duran, K.L. and Bartolomei, M.S. (2002) Analysis of sequence upstream of the endogenous H19 gene reveals elements both essential and dispensable for imprinting. *Mol. Cell. Biol.*, **22,** 2450-2462.

Toder, R., Wilcox, S.A., Smithwick, M. and Graves, J.A. (1996) The human/mouse imprinted genes IGF2, H19, SNRPN and ZNF127 map to two conserved autosomal clusters in a marsupial. *Chromosome Res.*, **4,** 295-300.

Tremblay, K.D., Duran, K.L. and Bartolomei, M.S. (1997) A 5' 2-kilobase-pair region of the imprinted mouse H19 gene exhibits exclusive paternal methylation throughout development. *Mol. Cell. Biol.*, **17,** 4322-4329.

Tuiskula-Haavisto, M. and Vilkki, J. (2007) Parent-of-origin specific QTL--a possibility towards understanding reciprocal effects in chicken and the origin of imprinting. *Cytogenet. Genome Res.*, **117,** 305-312.

Tycko, B. and Morison, I.M. (2002) Physiological functions of imprinted genes. *J. Cell. Physiol.*, **192,** 245-258.

Tyndale-Biscoe, H. and Renfree, M. (1987) *Reproductive Physiology of Marsupials.* Cambridge University Press,

Vakoc, C.R., Sachdeva, M.M., Wang, H. and Blobel, G.A. (2006) Profile of histone lysine methylation across transcribed mammalian chromatin. *Mol. Cell. Biol.*, **26,** 9185-9195.

Varmuza, S. and Mann, M. (1994) Genomic imprinting--defusing the ovarian time bomb. *Trends Genet.*, **10,** 118-123.

Veevers, J.J. (1991) Phanerozoic Australia in the changing configuration of Proto-Pangea through Gondwanaland and Pangea to the present dispersed continents. *Aust. Syst. Bot.*, **4,** 1-11.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al (2001) The sequence of the human genome. *Science,* **291,** 1304-1351.

Visel, A., Bristow, J. and Pennacchio, L.A. (2007) Enhancer identification through comparative genomics. *Semin. Cell Dev. Biol.*, **18,** 140-152.

Vollrath, D. (1999) DNA Markers for Physical Mapping. In Birren, B., Green, E. D., Hieter, P., et al (eds) *Genome Analysis: A Laboratory Manual.* Cold Spring Harbor Laboratory Press, 187-216.

Vu, T.H., Jirtle, R.L. and Hoffman, A.R. (2006) Cross-species clues of an epigenetic imprinting regulatory code for the IGF2R gene. *Cytogenet. Genome Res.*, **113,** 202-208.

Waddington, C. (1942) The Epigenotype. *Endeavour,* **1,** 18-20.

Wakefield, M.J. and Graves, J.A. (2005) Marsupials and monotremes sort genome treasures from junk. *Genome Biol.*, **6,** 218.

Walter, J., Hutter, B., Khare, T. and Paulsen, M. (2006) Repetitive elements in imprinted genes. *Cytogenet. Genome Res.*, **113,** 109-115.

Walter, J. and Paulsen, M. (2003) The potential role of gene duplications in the evolution of imprinting mechanisms. *Hum. Mol. Genet.*, **12 Spec No 2,** R215-20.

Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., et al (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420,** 520-562.

Weber, M., Hagege, H., Murrell, A., Brunel, C., Reik, W., Cathala, G. and Forne, T. (2003) Genomic imprinting controls matrix attachment regions in the Igf2 gene. *Mol. Cell. Biol.*, **23,** 8953-8959.

Weidman, J.R., Maloney, K.A. and Jirtle, R.L. (2006) Comparative phylogenetic analysis reveals multiple non-imprinted isoforms of opossum Dlk1. *Mammalian Genome*, **17,** 157-167.

Weidman, J.R., Murphy, S.K., Nolan, C.M., Dietrich, F.S. and Jirtle, R.L. (2004) Phylogenetic footprint analysis of IGF2 in extant mammals. *Genome Res.*, **14,** 1726-1732.

Wentworth, B.M., Schaefer, I.M., Villa-Komaroff, L. and Chirgwin, J.M. (1986) Characterization of the two nonallelic genes encoding mouse preproinsulin. *J. Mol. Evol.*, **23,** 305-312.

West, A.G., Gaszner, M. and Felsenfeld, G. (2002) Insulators: many functions, many mechanisms. *Genes Dev.*, **16,** 271-288.

Wilkins, J.F. (2005) Genomic imprinting and methylation: epigenetic canalization and conflict. *Trends Genet.*, **21,** 356-365.

Wilkins, J.F. and Haig, D. (2003) What good is genomic imprinting: the function of parent-specific gene expression. *Nat. Rev. Genet.*, **4,** 359-368.

Willson, M.F. and Burley, N. (1983) *Mate Choice in Plants: Tactics, Mechanisms, and Consequences.* Princeton University Press,

Wolfe, K.H., Sharp, P.M. and Li, W.H. (1989) Mutation rates differ among regions of the mammalian genome. *Nature*, **337,** 283-285.

Wood, A.J., Roberts, R.G., Monk, D., Moore, G.E., Schulz, R. and Oakey, R.J. (2007) A screen for retrotransposed imprinted genes reveals an association between X chromosome homology and maternal germ-line methylation. *PLoS Genet.*, **3,** e20.

Woolfe, A., Goodson, M., Goode, D.K., Snell, P., McEwen, G.K., Vavouri, T., Smith, S.F., North, P., Callaway, H., Kelly, K., et al (2005) Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol.*, **3,** e7.

Wu, H.K., Squire, J.A., Song, Q. and Weksberg, R. (1997) Promoter-dependent tissue-specific expressive nature of imprinting gene, insulin-like growth factor II, in human tissues. *Biochem. Biophys. Res. Commun.*, **233,** 221-226.

Xi, H., Shulha, H.P., Lin, J.M., Vales, T.R., Fu, Y., Bodine, D.M., McKay, R.D., Chenoweth, J.G., Tesar, P.J., Furey, T.S., et al (2007) Identification and characterization of cell type-specific and ubiquitous chromatin regulatory structures in the human genome. *PLoS Genet.*, **3,** e136.

Xie, X., Mikkelsen, T.S., Gnirke, A., Lindblad-Toh, K., Kellis, M. and Lander, E.S. (2007) Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. *Proc. Natl. Acad. Sci. U. S. A.*, **104,** 7145-7150.

Yahagi, S., Shibuya, K., Obayashi, I., Masaki, H., Kurata, Y., Kudoh, J. and Shimizu, N. (2004) Identification of two novel clusters of ultrahigh-sulfur keratin-associated protein genes on human chromosome 11. *Biochem. Biophys. Res. Commun.*, **318,** 655-664.

Yang, P.K. and Kuroda, M.I. (2007) Noncoding RNAs and intranuclear positioning in monoallelic gene expression. *Cell,* **128,** 777-786.

Ye, L. and Huang, X. (2005) MAP2: multiple alignment of syntenic genomic sequences. *Nucleic Acids Res.*, **33,** 162-170.

Ylstra, B., van den Ijssel, P., Carvalho, B., Brakenhoff, R.H. and Meijer, G.A. (2006) BAC to the future! or oligonucleotides: a perspective for micro array comparative genomic hybridization (array CGH). *Nucleic Acids Res.*, **34,** 445-450.

Yoder, J.A., Walsh, C.P. and Bestor, T.H. (1997) Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet.*, **13,** 335-340.

Yokomine, T., Hata, K., Tsudzuki, M. and Sasaki, H. (2006) Evolution of the vertebrate DNMT3 gene family: a possible link between existence of DNMT3L and genomic imprinting. *Cytogenet. Genome Res.*, **113,** 75-80.

Yokomine, T., Kuroiwa, A., Tanaka, K., Tsudzuki, M., Matsuda, Y. and Sasaki, H. (2001) Sequence polymorphisms, allelic expression status and chromosome locations of the chicken IGF2 and MPR1 genes. *Cytogenet. Cell Genet.*, **93,** 109-113.

Yokomine, T., Shirohzu, H., Purbowasito, W., Toyoda, A., Iwama, H., Ikeo, K., Hori, T., Mizuno, S., Tsudzuki, M., Matsuda, Y., et al (2005) Structural and functional analysis of a 0.5-Mb chicken region orthologous to the imprinted mammalian Ascl2/Mash2-Igf2-H19 region. *Genome Res.*, **15,** 154-165.

Yoo-Warren, H., Pachnis, V., Ingram, R.S. and Tilghman, S.M. (1988) Two regulatory domains flank the mouse H19 gene. *Mol. Cell. Biol.*, **8,** 4707-4715.

Youngson, N.A., Kocialkowski, S., Peel, N. and Ferguson-Smith, A.C. (2005) A small family of sushi-class retrotransposon-derived genes in mammals and their relation to genomic imprinting. *J. Mol. Evol.*, **61,** 481-490.

Yusufzai, T.M. and Felsenfeld, G. (2004) The 5'-HS4 chicken beta-globin insulator is a CTCF-dependent nuclear matrix-associated element. *Proc. Natl. Acad. Sci. U. S. A.*, **101,** 8620-8624.

Zaratiegui, M., Irvine, D.V. and Martienssen, R.A. (2007) Noncoding RNAs and gene silencing. *Cell,* **128,** 763-776.

Zeng, Y., Yi, R. and Cullen, B.R. (2005) Recognition and cleavage of primary microRNA precursors by the nuclear processing enzyme Drosha. *EMBO J.*, **24,** 138-148.

Zhao, Y. and Srivastava, D. (2007) A developmental view of microRNA function. *Trends Biochem. Sci.*, **32,** 189-197.

Zubair, M., Hilton, K., Saam, J.R., Surani, M.A., Tilghman, S.M. and Sasaki, H. (1997) Structure and expression of the mouse L23mrp gene downstream of the imprinted H19 gene: biallelic expression and lack of interaction with the H19 enhancers. *Genomics,* **45,** 290-296.