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.

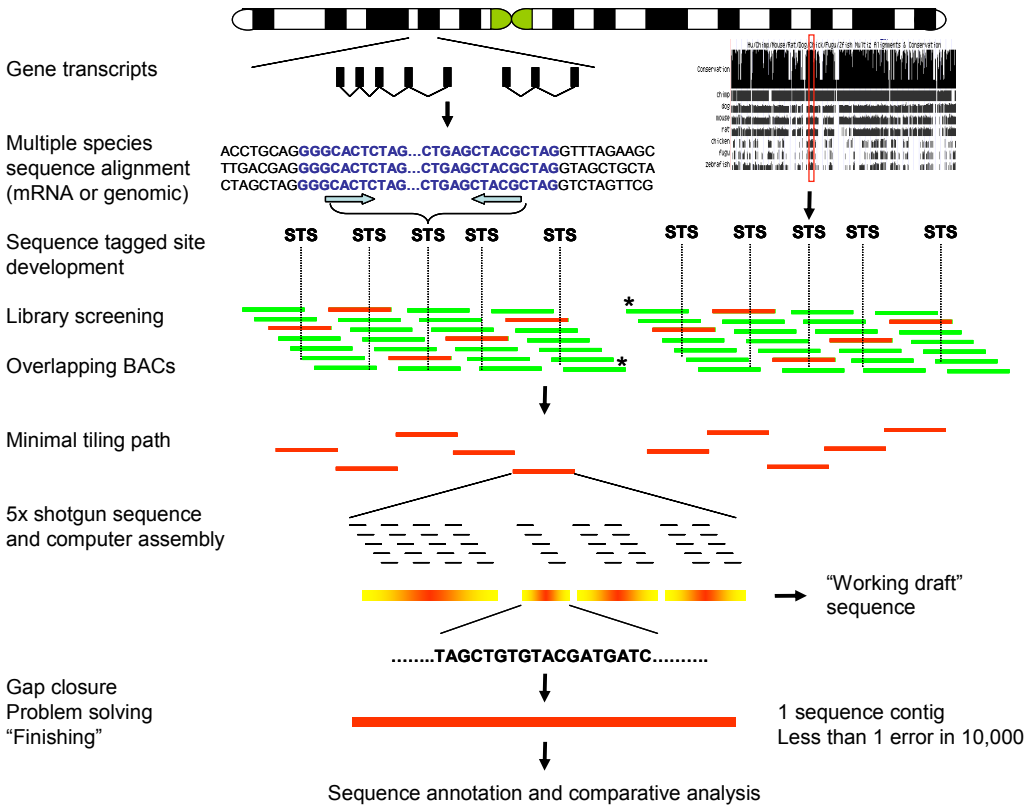

Figure III.1. Mapping and sequencing strategy.

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).

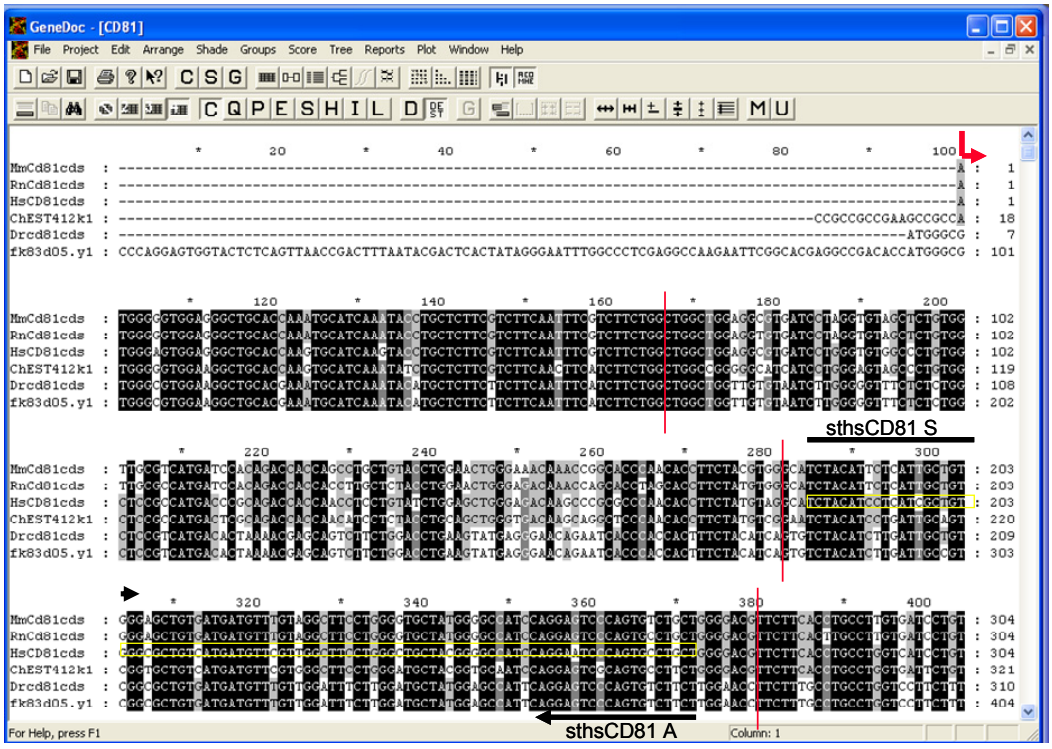

Figure III.2. Multi-species sequence alignment of *CD81* gene sequences.

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).

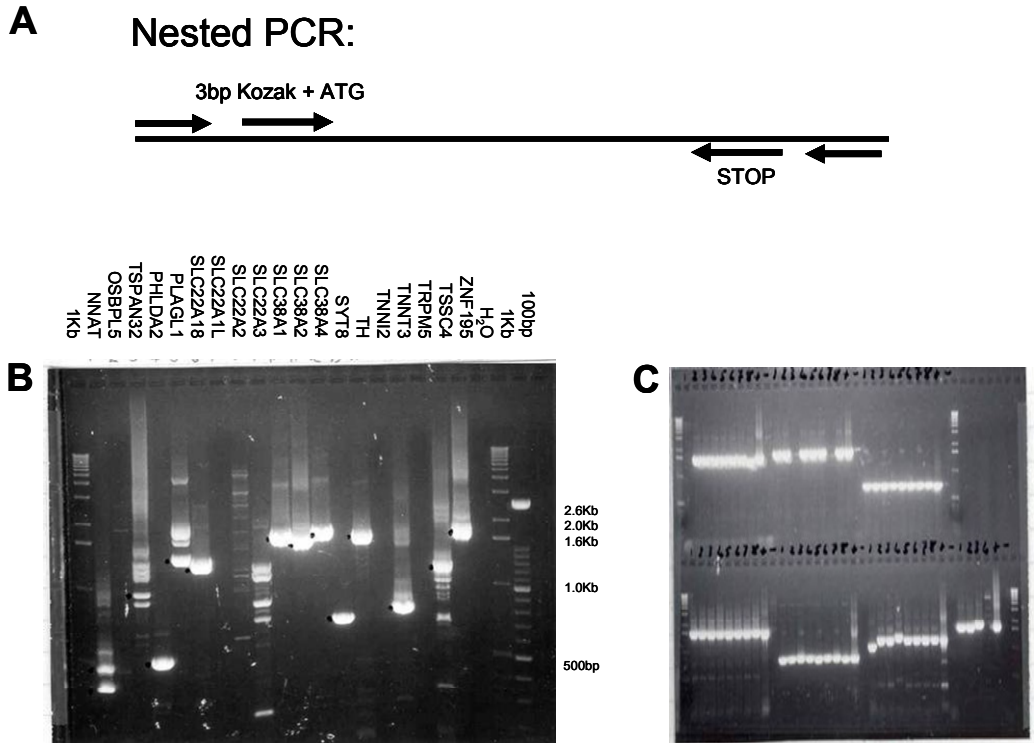

Figure III.3. Strategy for cloning human open reading frames.

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

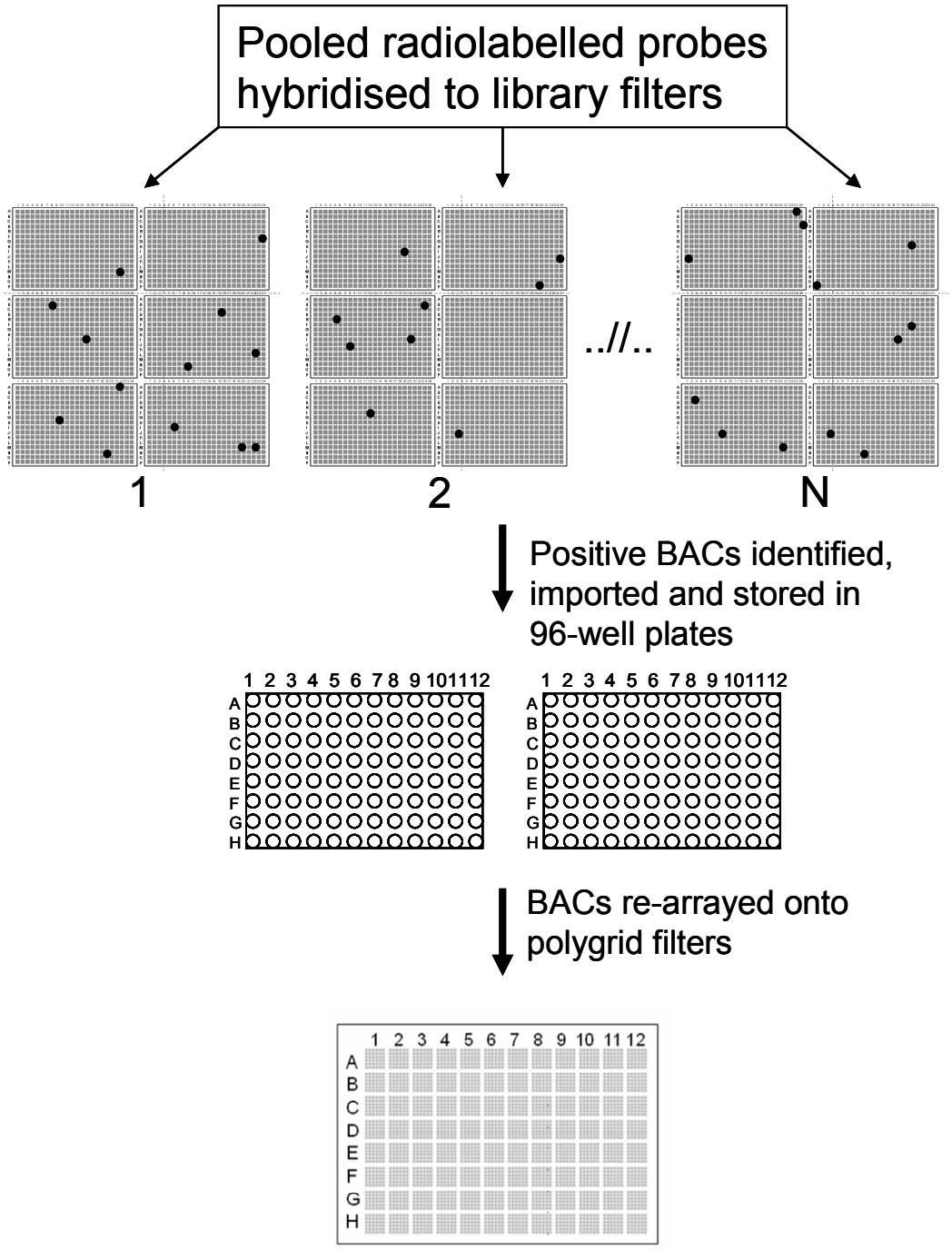

Figure III.4. Library screening strategy.

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.

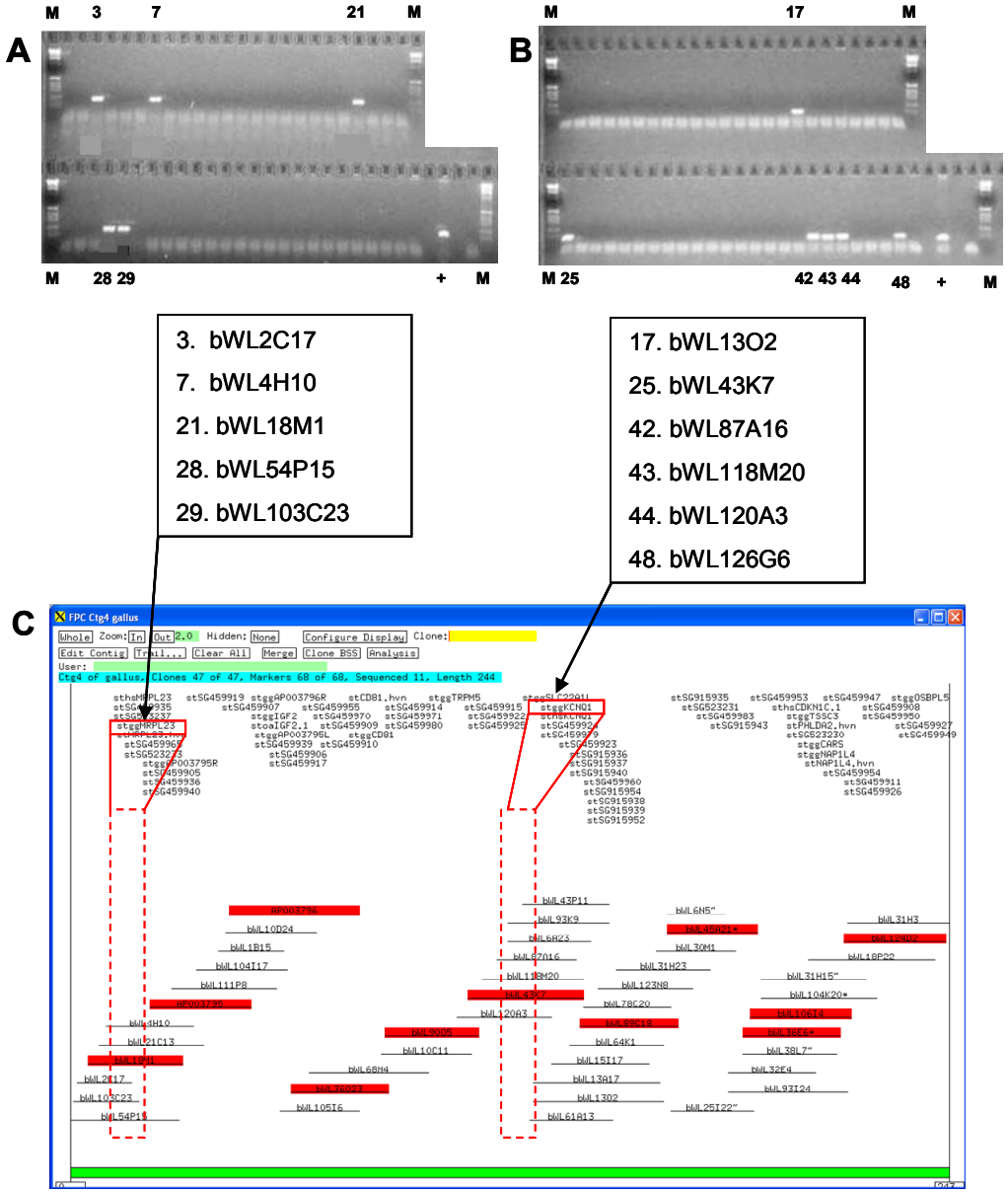

Figure III.5. Landmark content analysis of chicken BACs by colony PCR.

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.

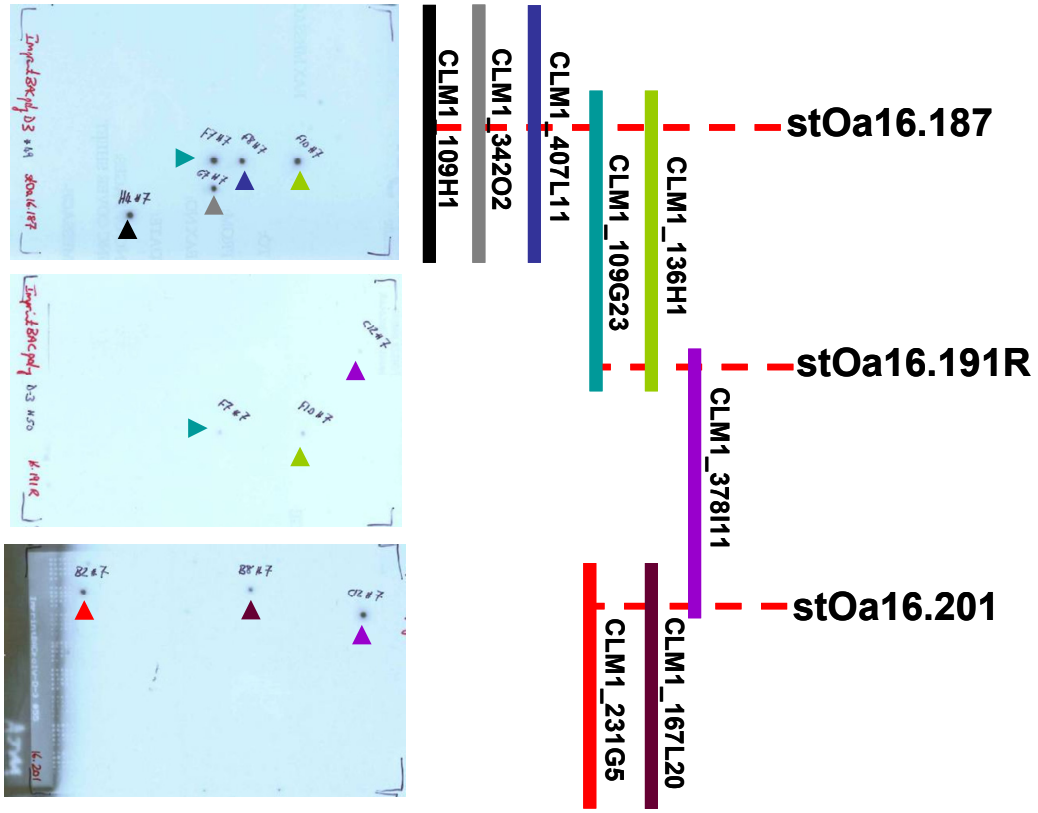

Figure III.6. Landmark content mapping through polygrid screening.

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.

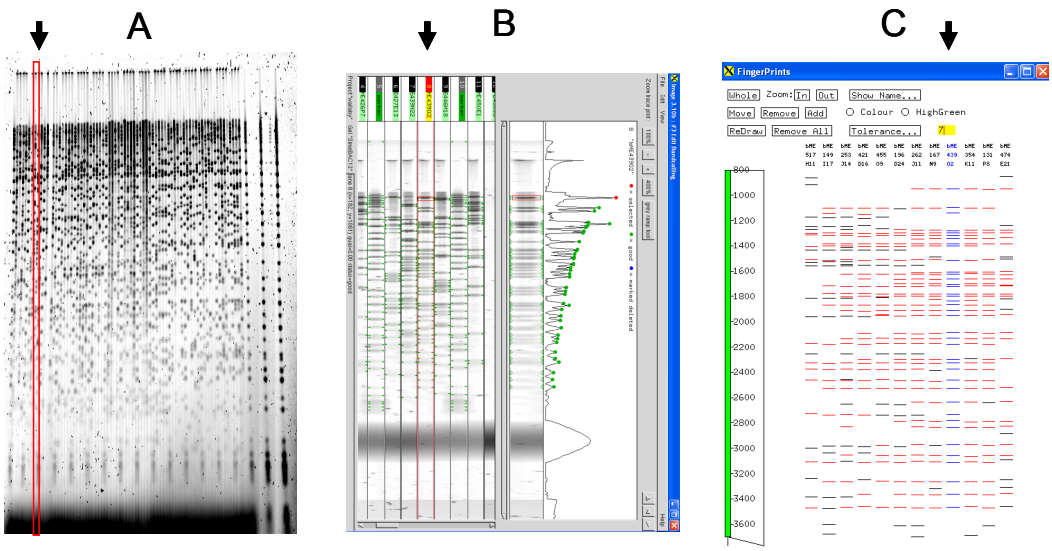

Figure III.7. The process of fingerprint mapping.

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.

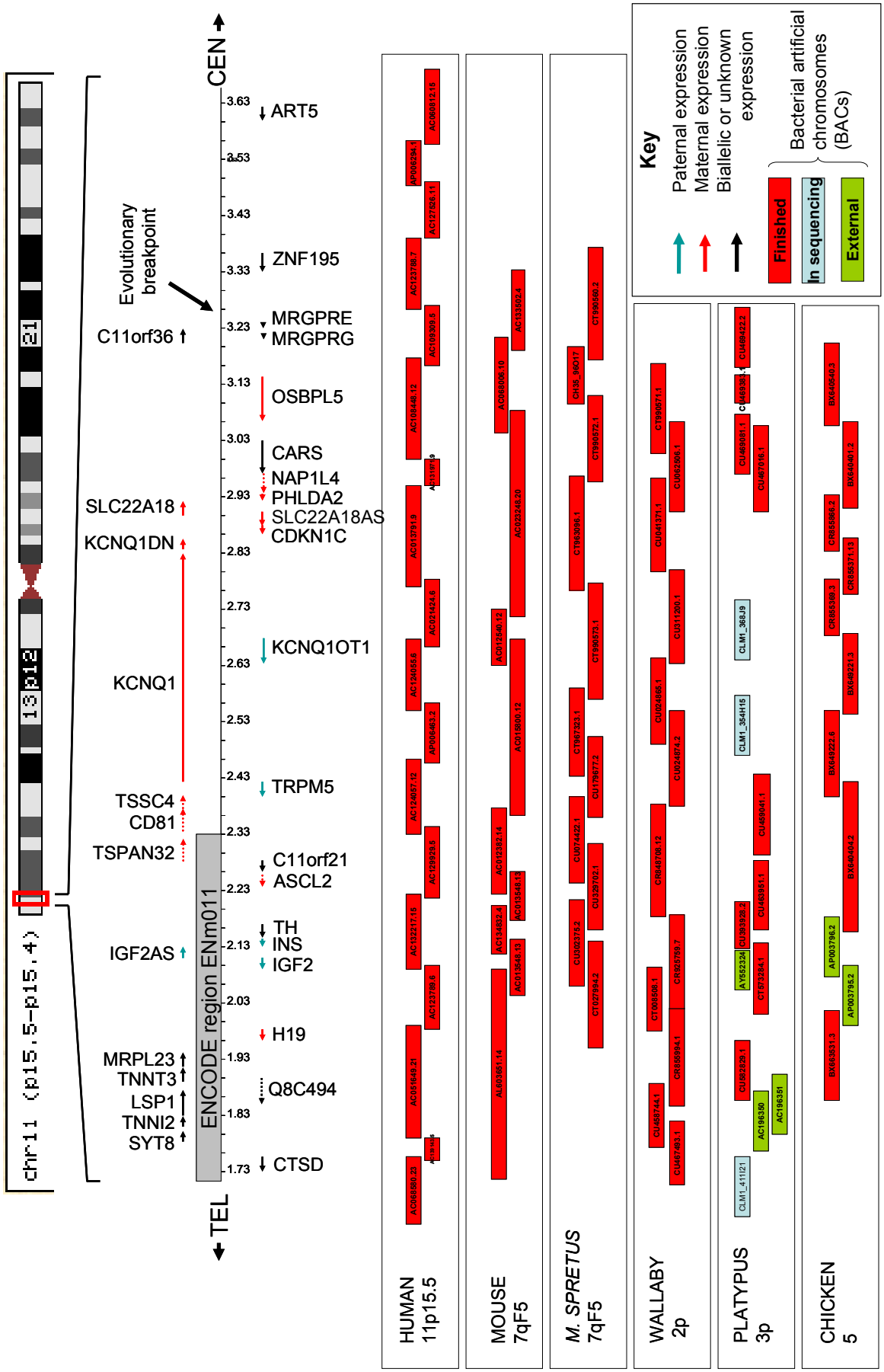
chr11 (p15.5-p15.4)
11p15.5
11p15.4
21
CEN
TEL
Evolutionary breakpoint
ENCODE region ENm011
ART5
ZNF195
MRGPRE
MRGPRG
OSBPL5
CARS
NAP1L4
PHLDA2
SLC22A18AS
CDKN1C
KCNQ1OT1
TRPM5
C11orf21
ASCL2
TH
INS
IGF2
H19
Q8C494
CTSD
C11orf36
SLC22A18
KCNQ1DN
KCNQ1
TSSC4
CD81
TSPAN32
IGF2AS
MRPL23
TNNT3
LSP1
TNNI2
SYT8
3.63
3.53
3.43
3.33
3.23
3.13
3.03
2.93
2.83
2.73
2.63
2.53
2.43
2.33
2.23
2.13
2.03
1.93
1.83
1.73
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).

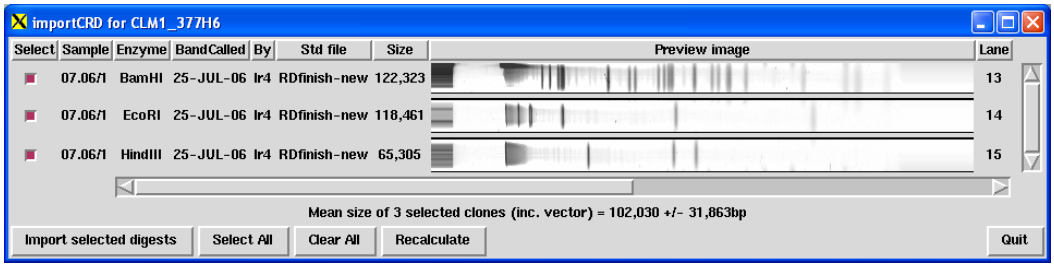

Figure III.9. Restriction endonuclease digests for platypus BAC CLM1_377H6.

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 *Hin*dIII.

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).

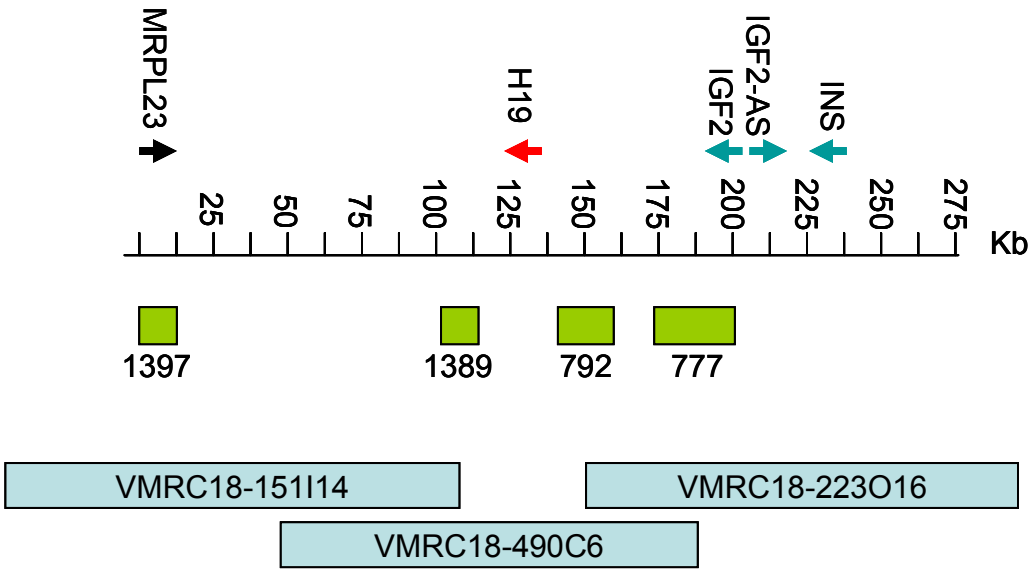

Figure III.10. Schematic of opossum mapping in orthologous IC1 region.

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.

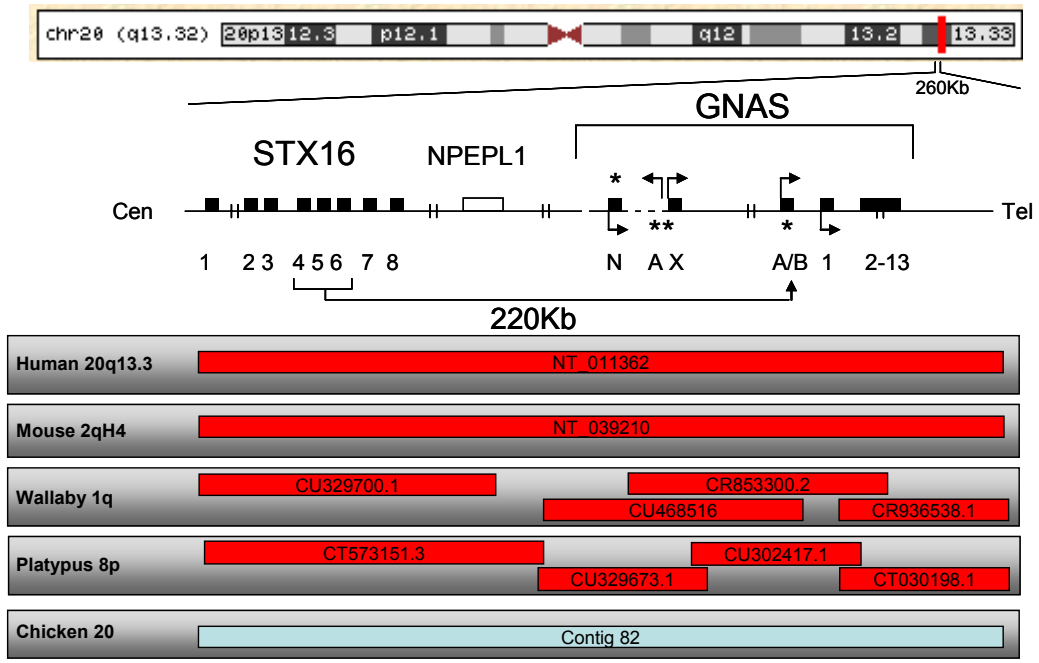

Figure III.11. Schematic of the GNAS complex region.

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

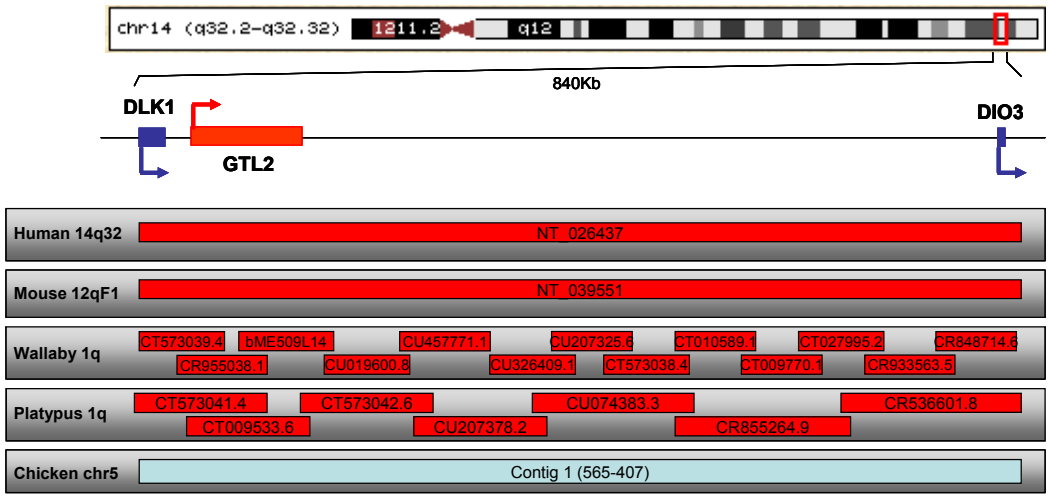

Figure III.12. Schematic of the DLK1-DIO3 region.

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).

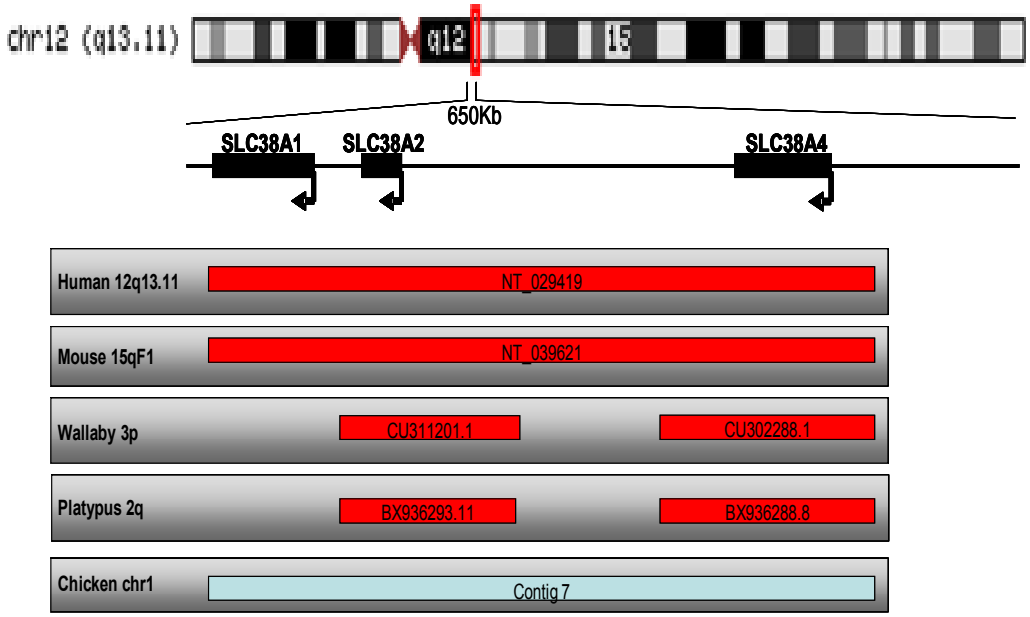

Figure III.13. Schematic of the solute carrier gene family 38 region.

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.