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.

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

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.

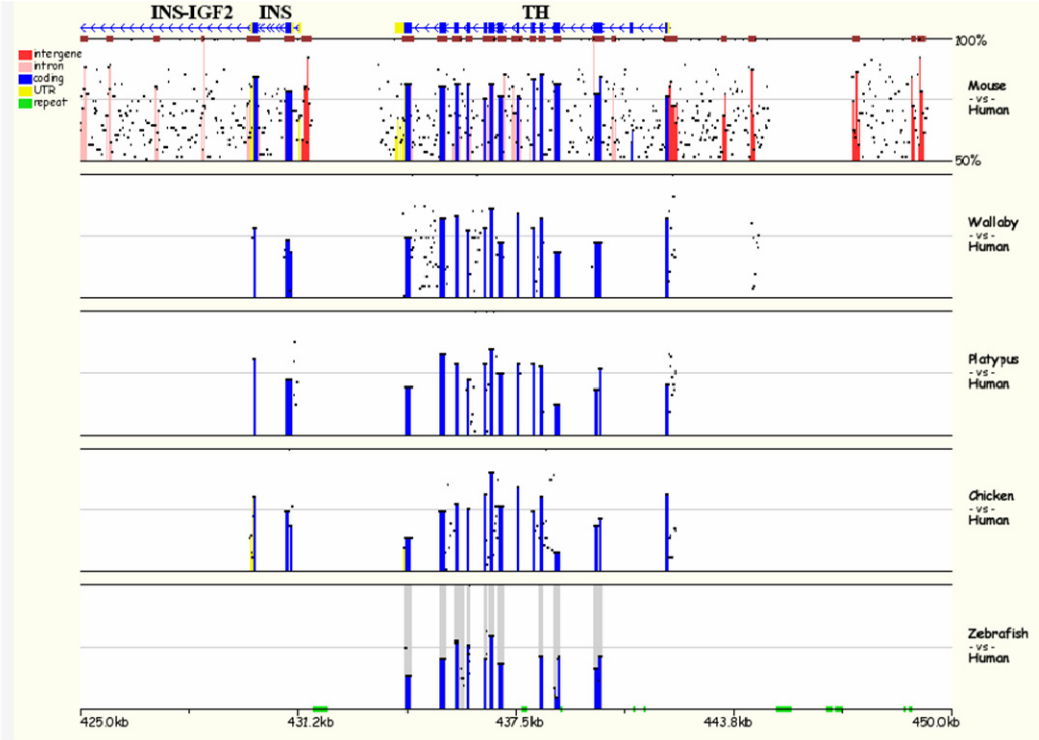

Figure V.2. Example of a zPicture dynamic visualisation plot.

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.

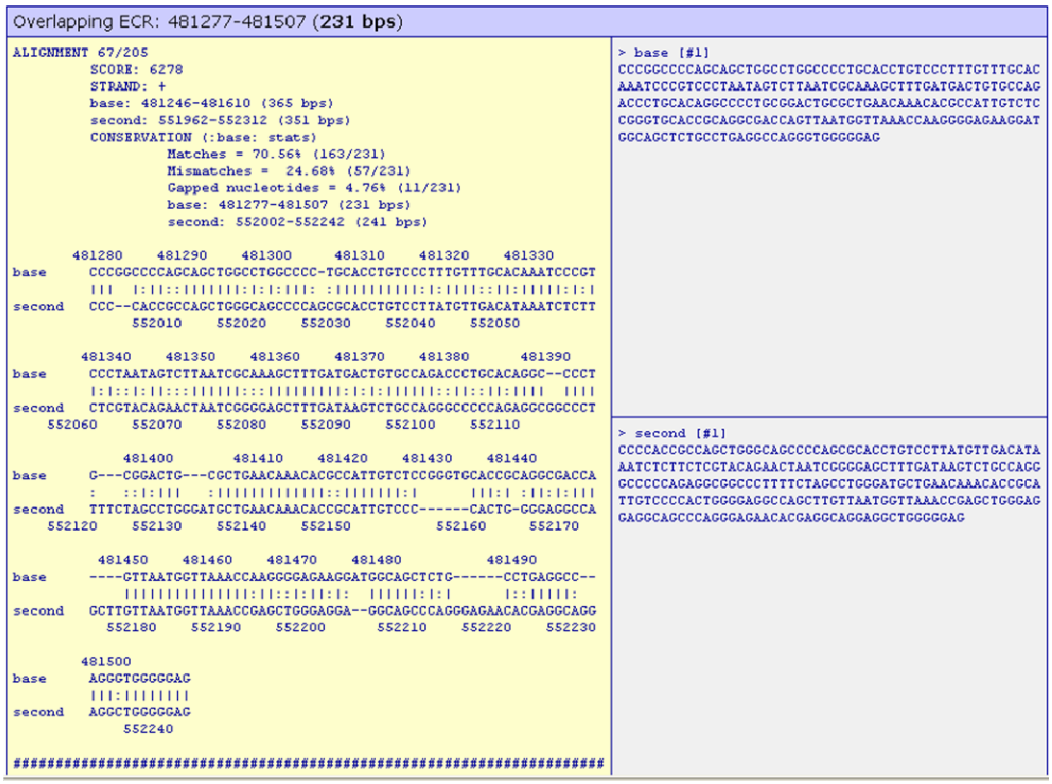

Figure V.3. BLASTZ sequence alignment viewed in zPicture.

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.

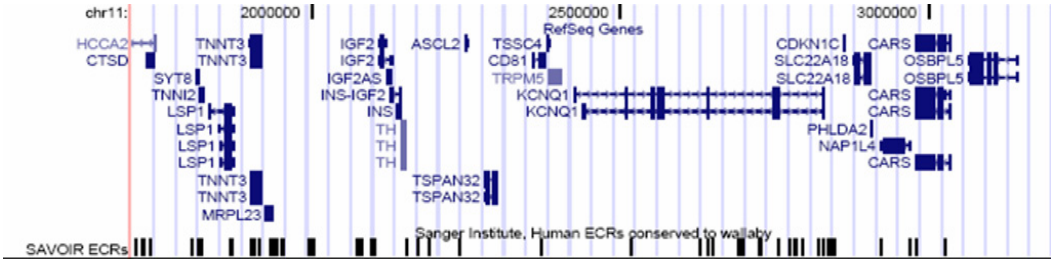

Figure V.4. Overview of the location of non-coding ECRs identified in the human 11p15.5 region.

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.

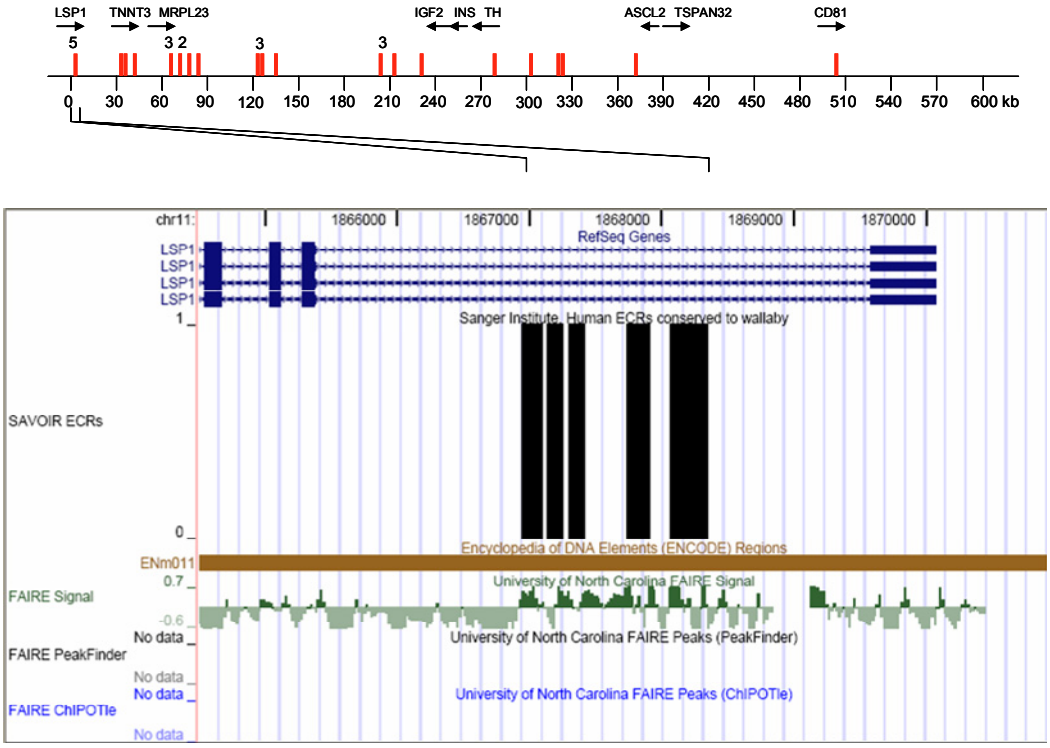

Figure V.5. Clustered ECRs within intron 10 of the *LSP1* gene.

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).

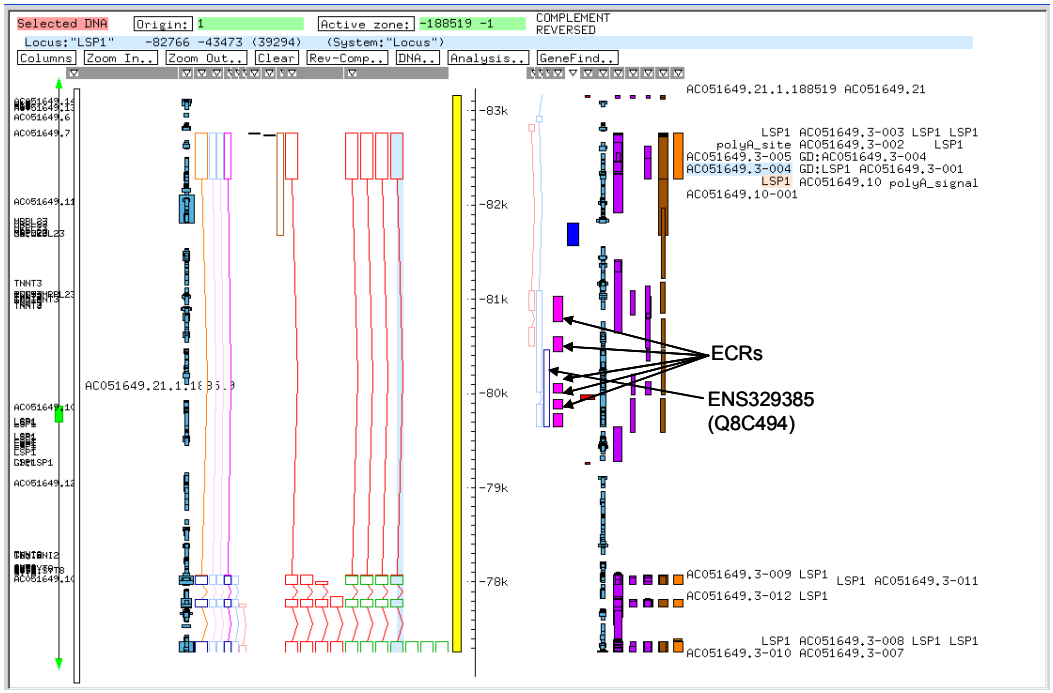

Figure V.6. Location of ECRs relative to annotated features in ACeDB.

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.

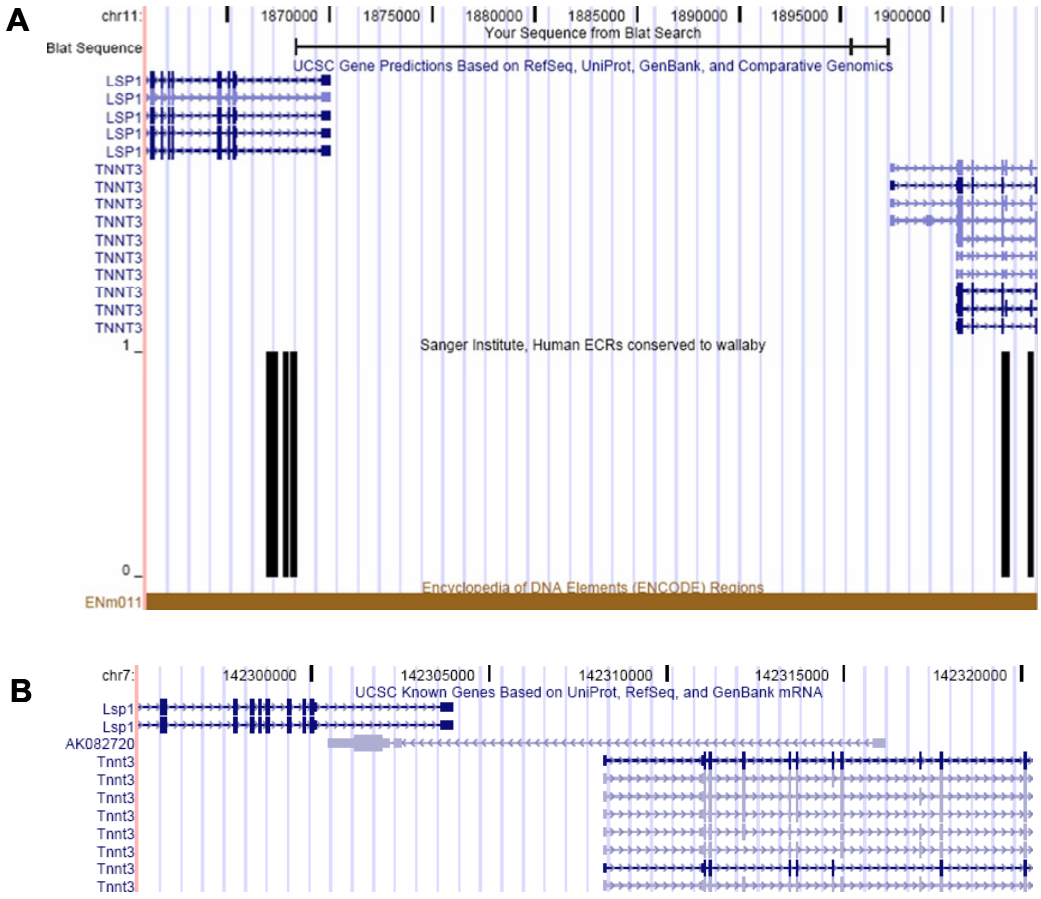

Figure V.8. Extent of human and mouse novel transcripts visualised in the UCSC genome browser.

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.

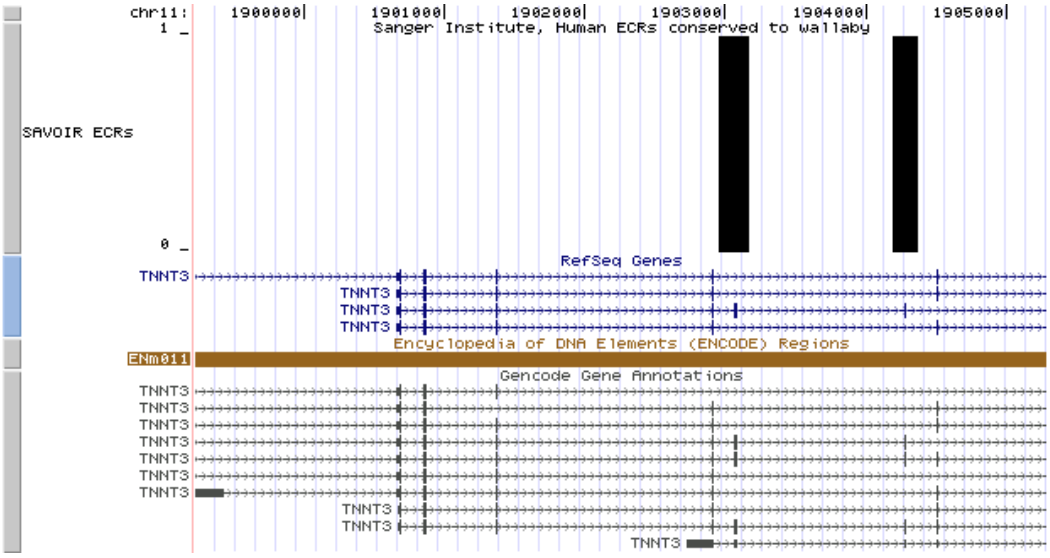

Figure V.9. Example of ECRs highlighting alternative exons.

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).

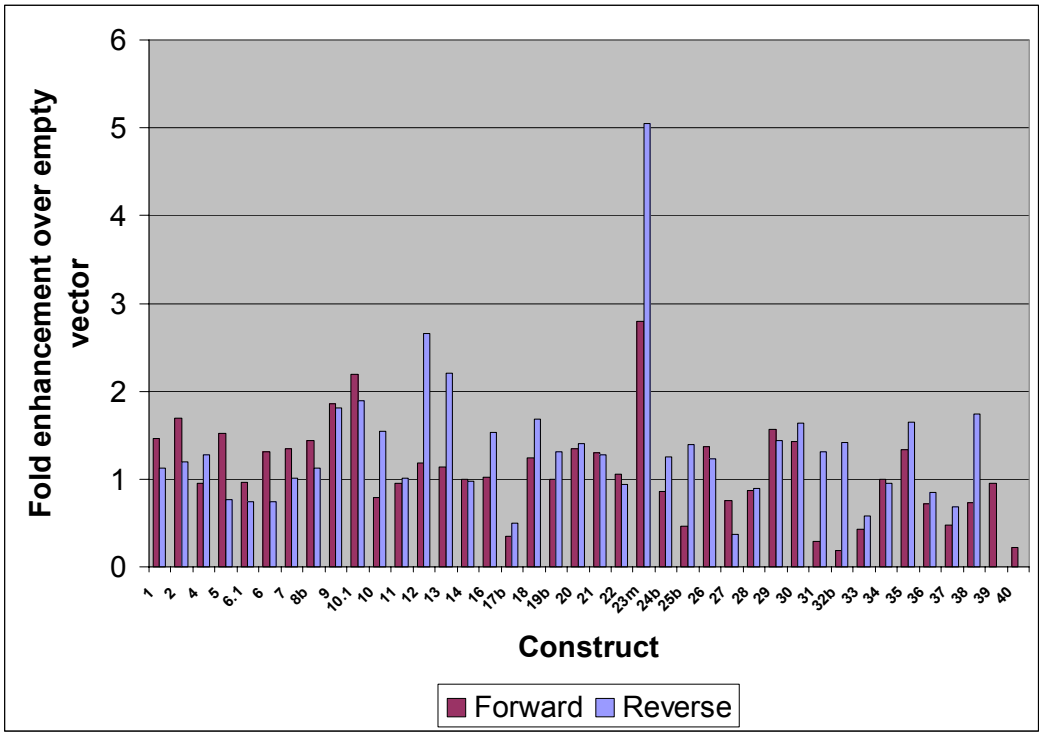

Figure V.10. Testing human 'randomers' for enhancer activity in HepG2 cells.

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.

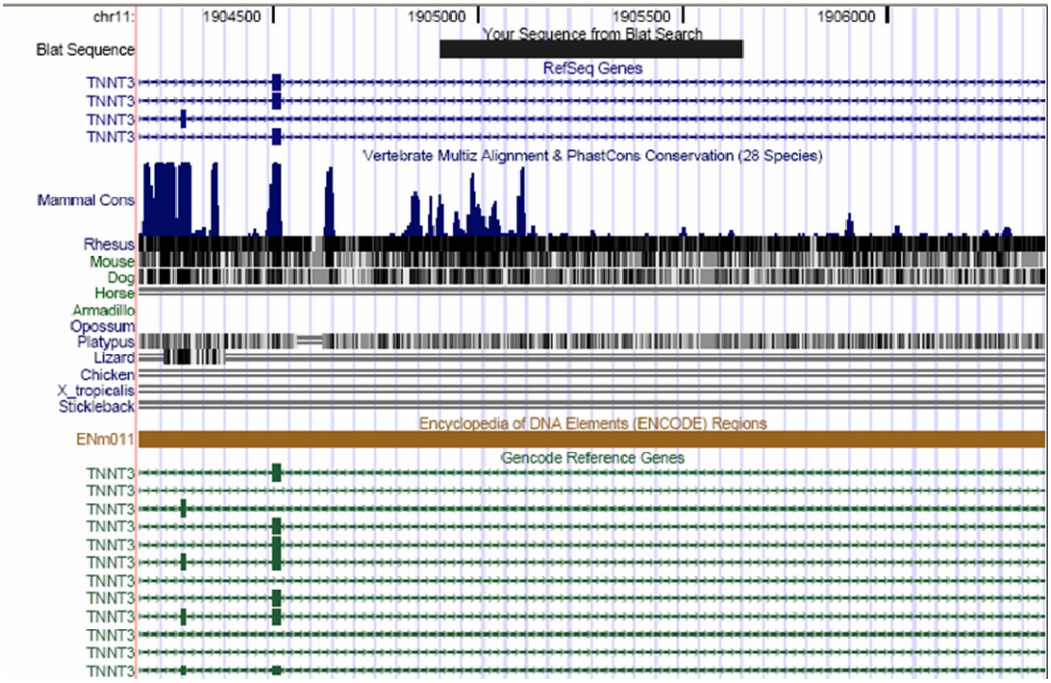

Figure V.11. Sequence conservation overlapping randomer 23m.

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.

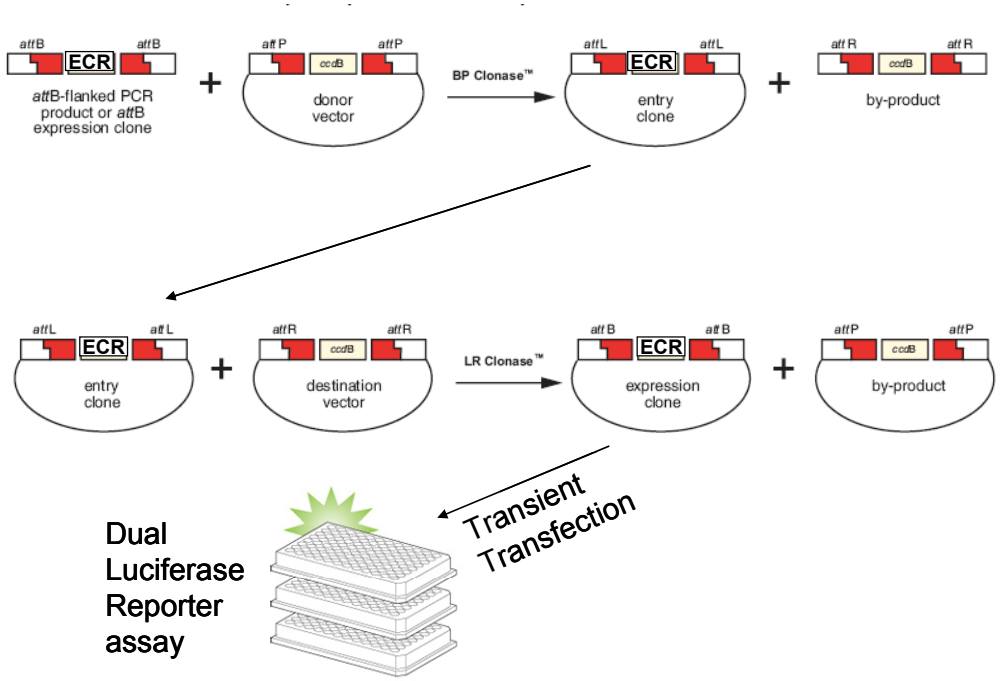

Figure V.12. Gateway® (Invitrogen) recombination cloning strategy.

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.

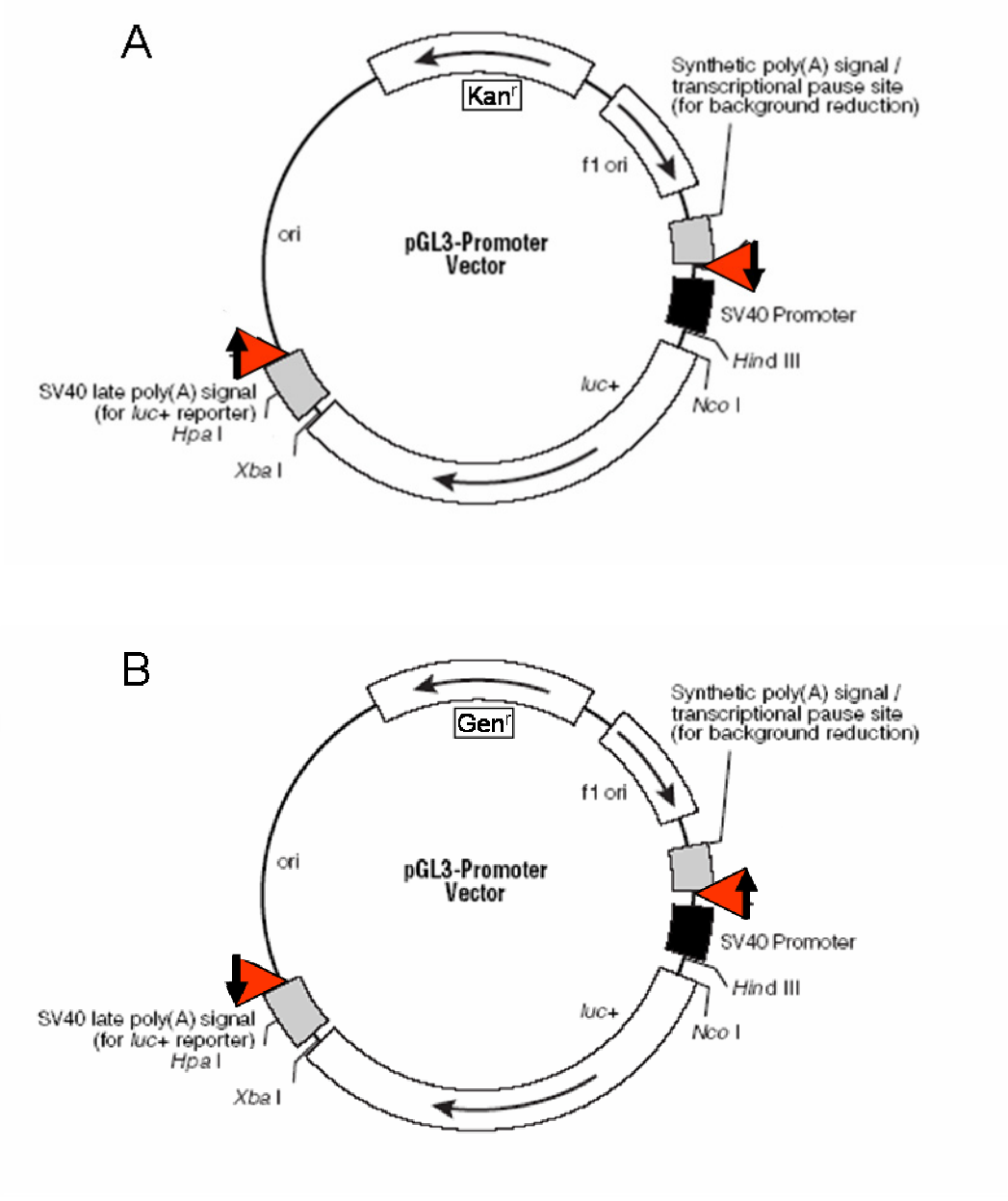

Figure V.13. Gateway® modified pGL3-Promoter vectors for enhancer testing.

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.

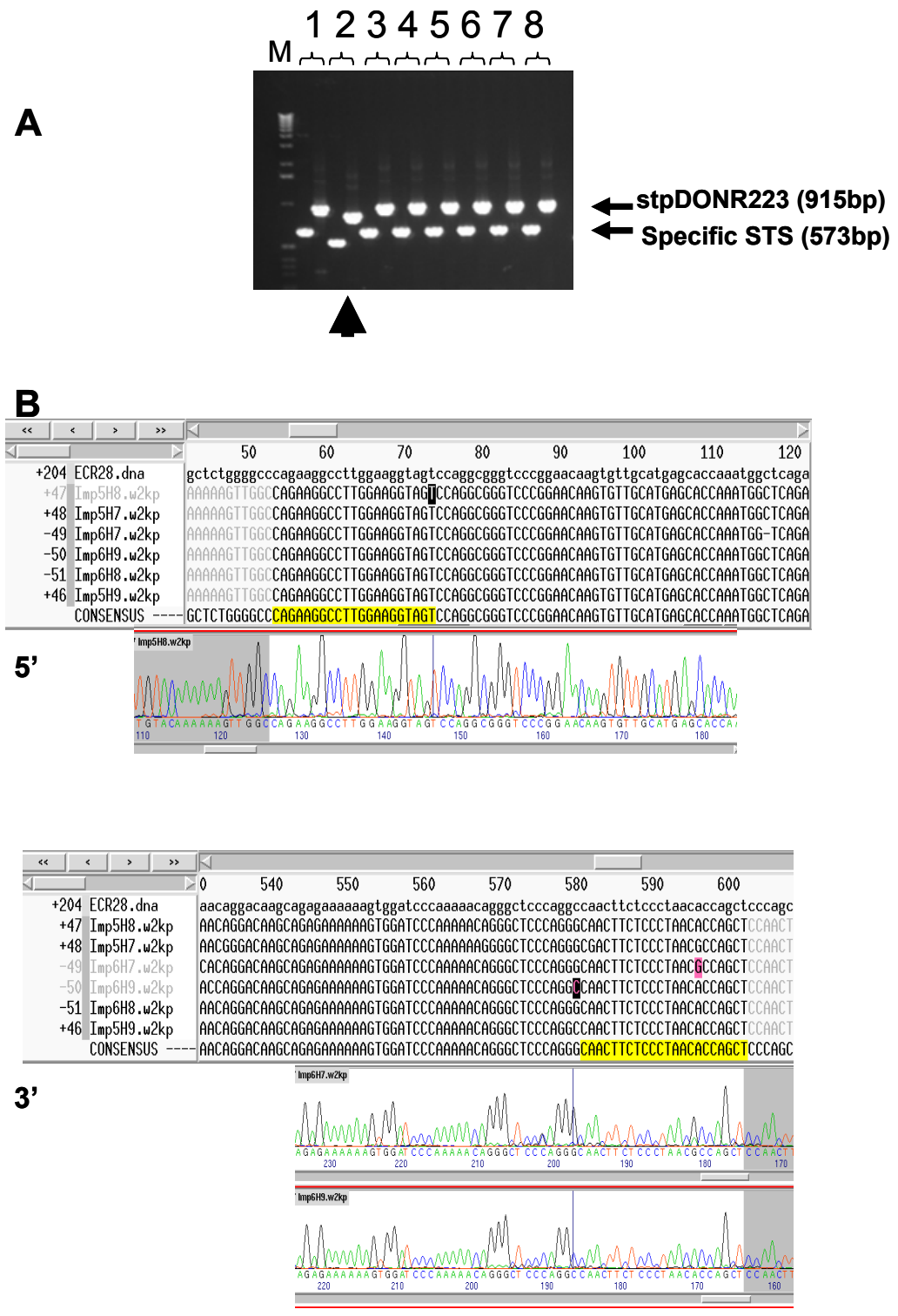

Figure V.14. Cloning verification of the ECR28 pENTR clone.

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.

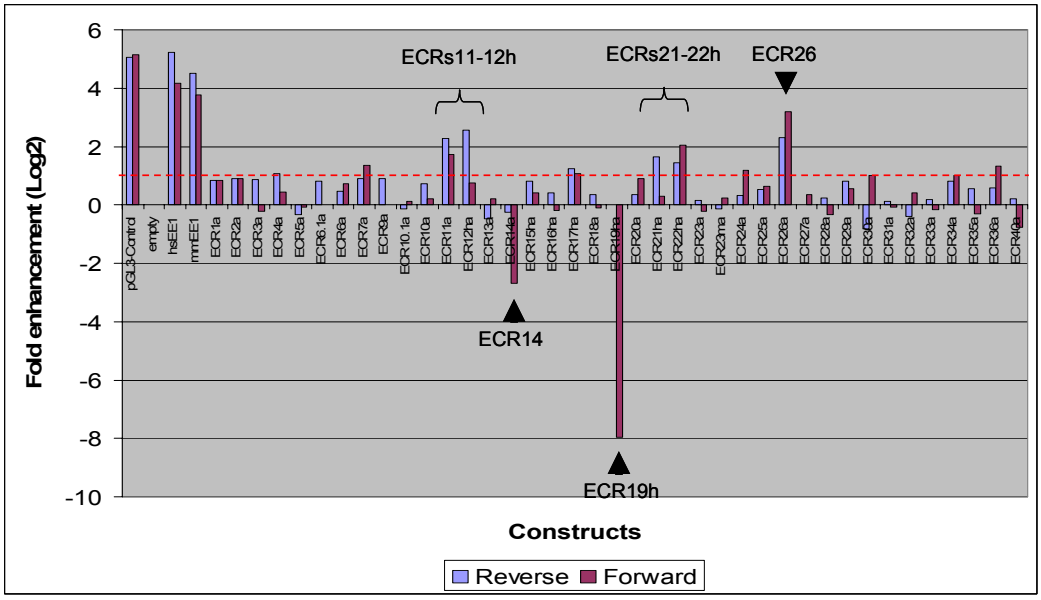

Figure V.15. Testing 11p15.5 ECRs for enhancer activity in human HepG2 cells.

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).

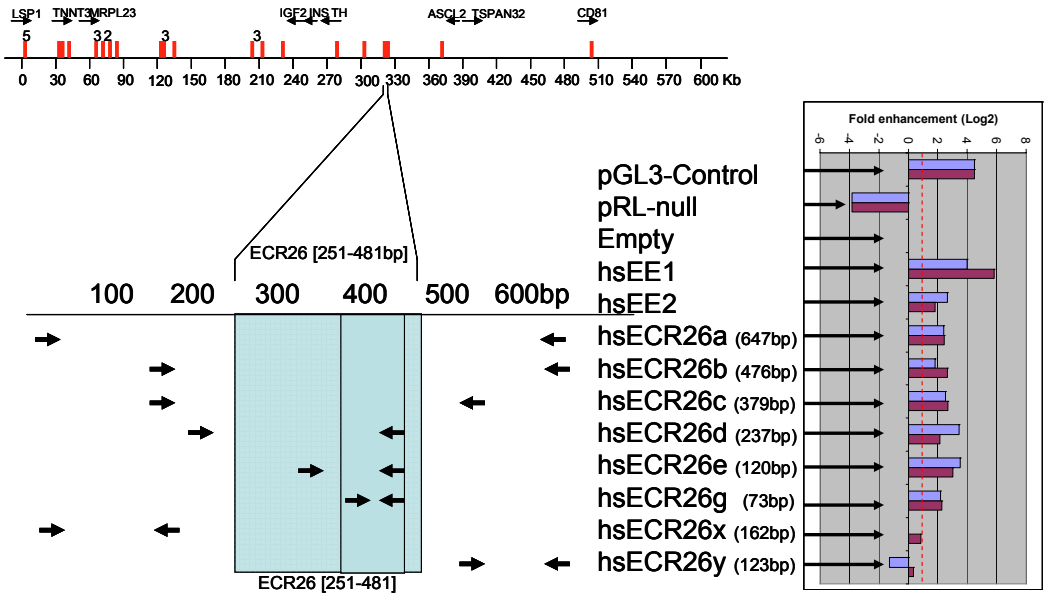

Figure V.16. Identifying a core enhancer element.

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.

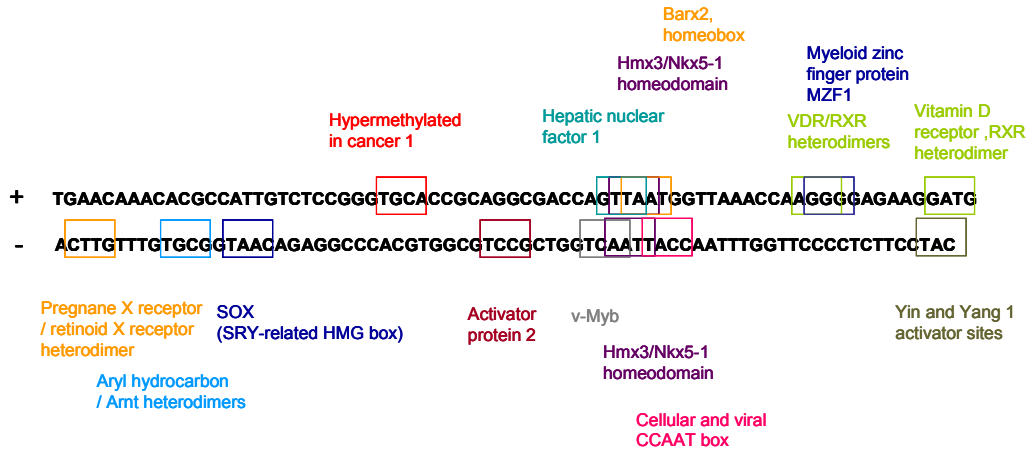

Figure V.17. Predicted TFBSs in the 73bp core enhancer region of ECR26.

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.

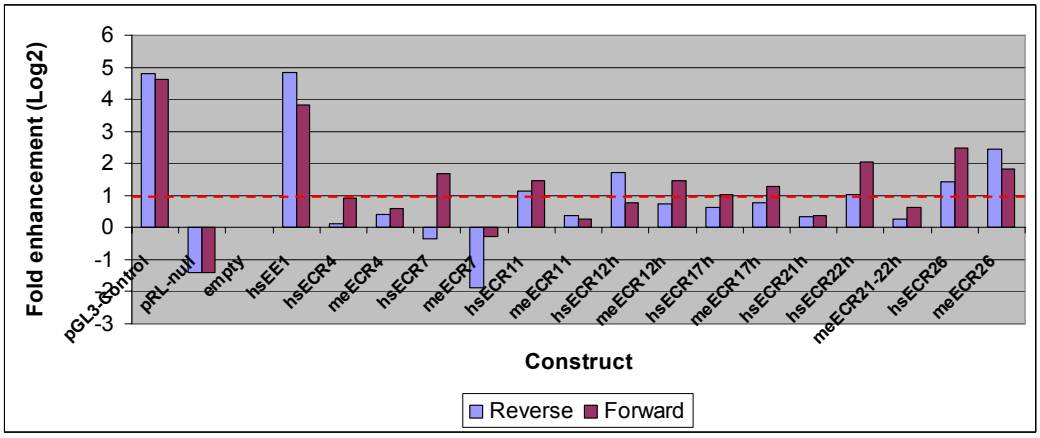

Figure V.18. Enhancer activities of wallaby ECRs in human HepG2 cells.

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

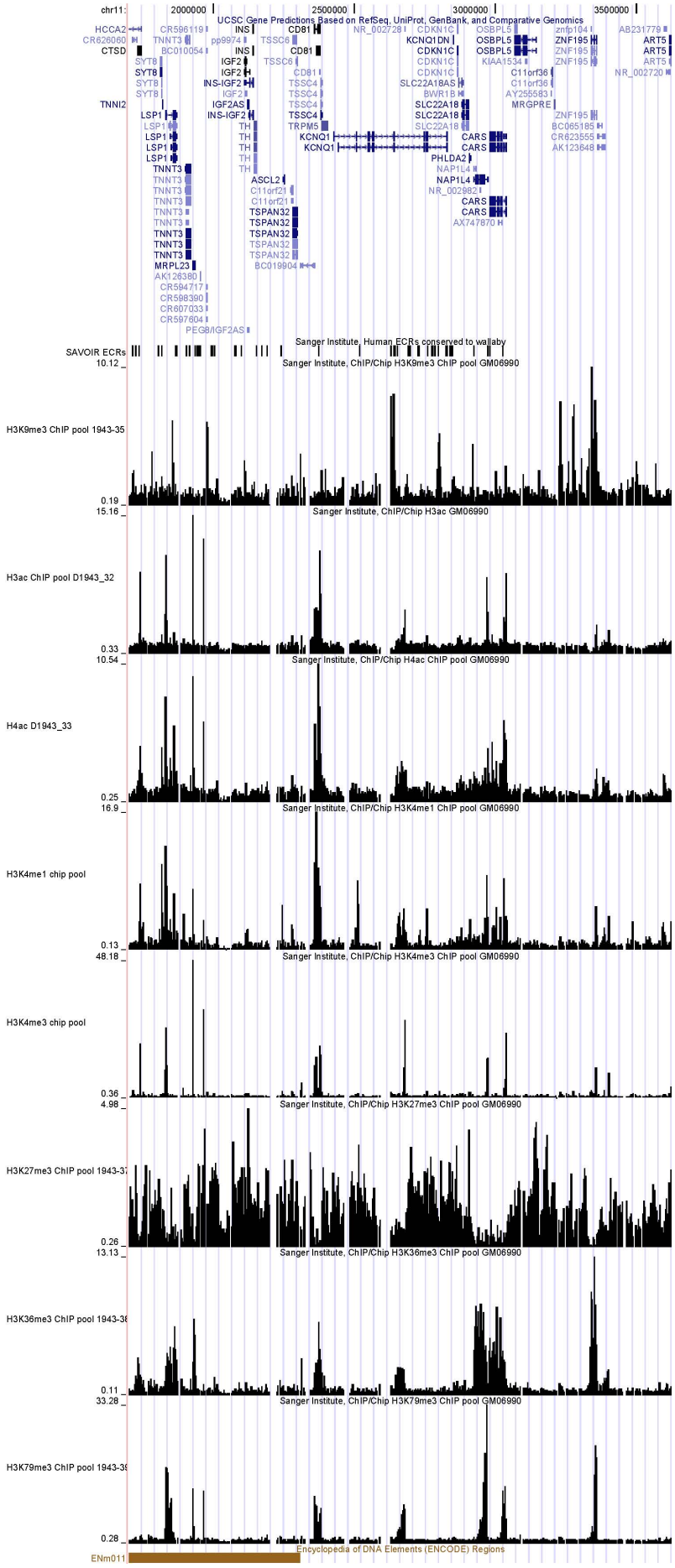
chr11:
2000000
2500000
3000000
3500000
UCSC Gene Predictions Based on RefSeq, UniProt, GenBank, and Comparative Genomics
Sanger Institute, Human ECRs conserved to wallaby
SAVOIR ECRs
Sanger Institute, ChIP/Chip H3K9me3 ChIP pool GM06990
H3K9me3 ChIP pool 1943-35
Sanger Institute, ChIP/Chip H3ac GM06990
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).

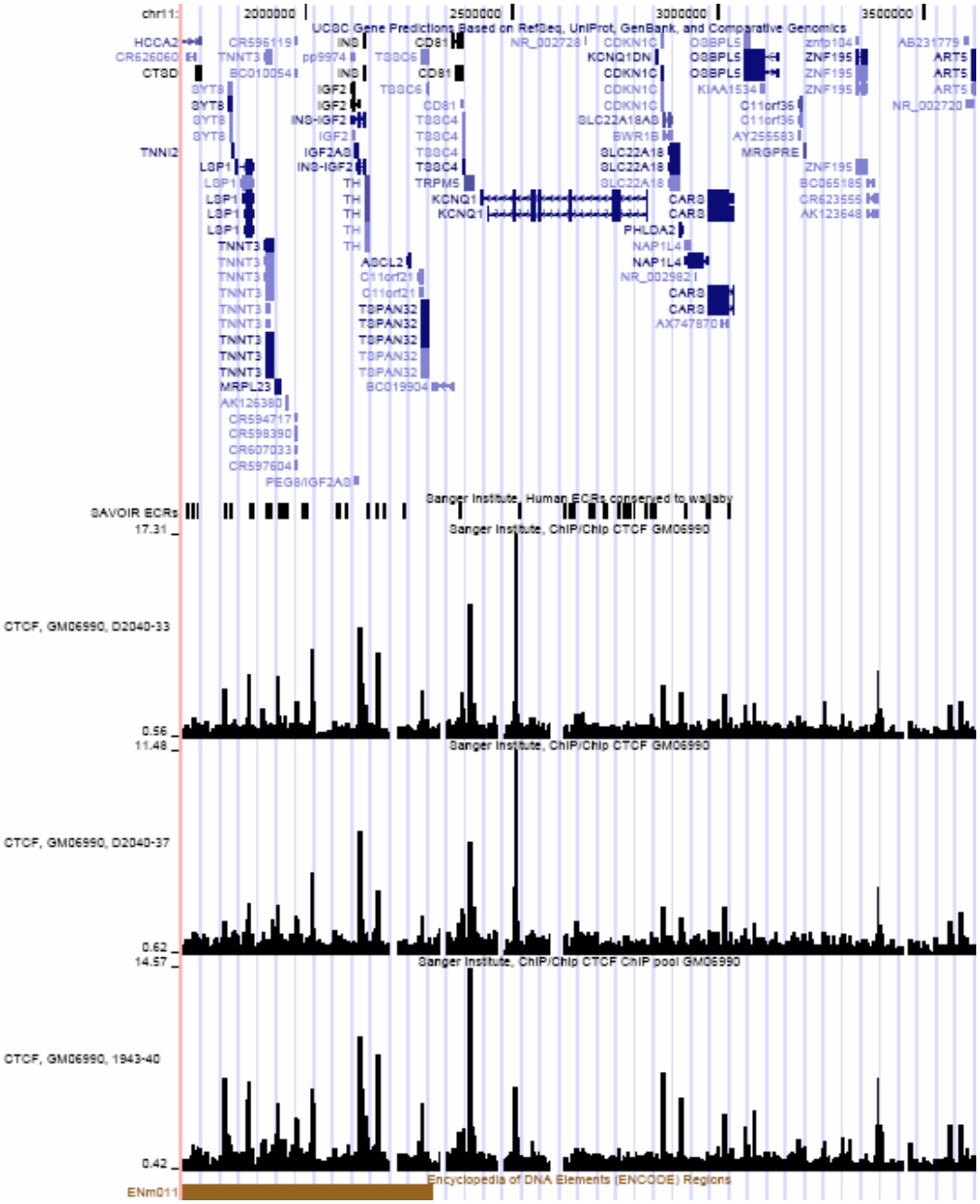

Figure V.20. CTCF profiles across the ENm011_EXTENDED region.

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.

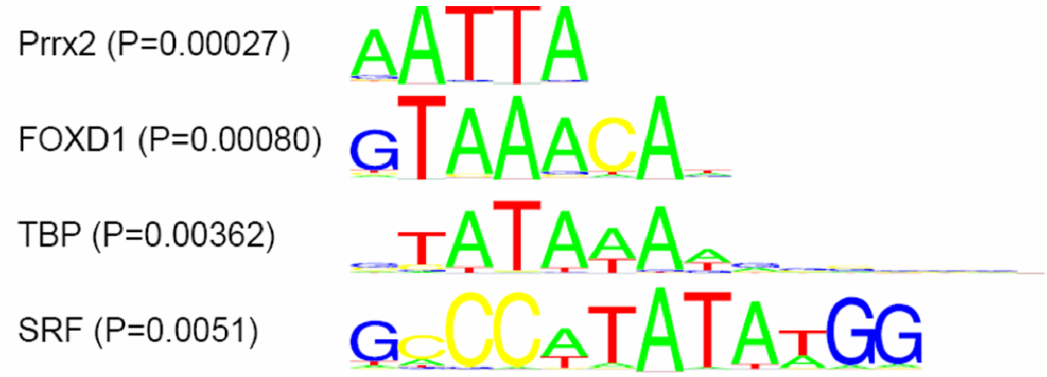

Figure V.21. Over-represented known motifs in the ECR set.

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.

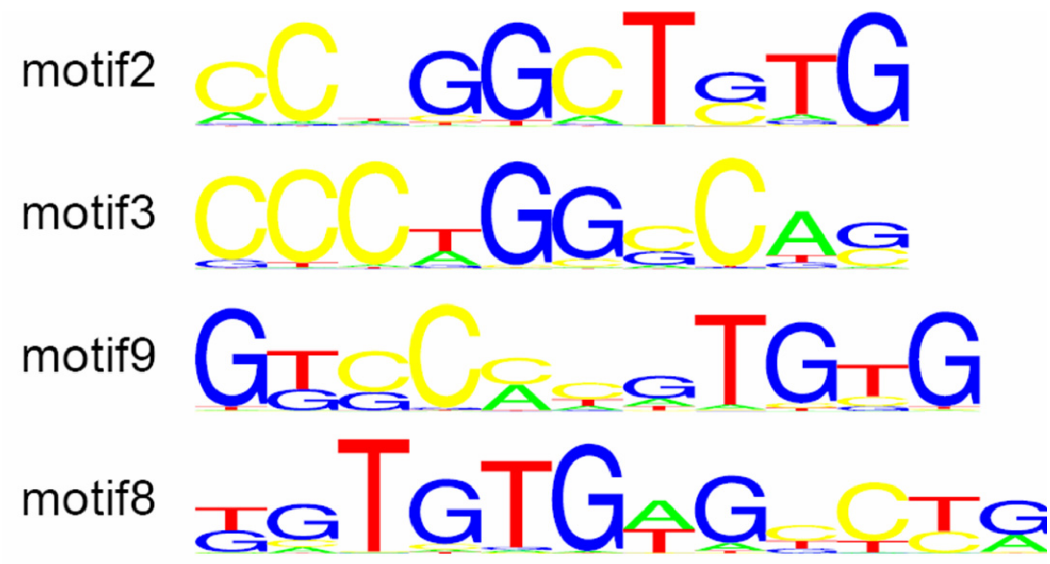

Figure V.22. Identifying novel sequence motifs over-represented in the ECR set.

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.