Chapter IV Comparative mapping, sequencing and analysis

4.1 Introduction

4.1.1 Benefits of comparative sequence analysis

The identification of the full complement of human genes as a result of the sequencing and analysis of the human genome in isolation seems unlikely, as discussed in chapter III. Currently, the most efficient approach to gene identification utilises expressed sequence evidence (chapter III). However, some genes with a restricted spatial or temporal expression pattern may not be represented in the available EST and cDNA resources. A second limitation of the EST databases is the paucity of 5' UTR sequences in the entries. Currently the sequence available is mainly limited to the 3'UTR of the mRNA as 5' end information is often scarce due to the method of construction of the resources used (section 3.1.3). In addition, most DNA sequences involving regulation of gene expression are in non-transcribed regions, which cannot be accessed through EST sequence.

Alternative transcript mapping methods discussed in chapter III were also noted to have limitations. For example, *ab initio* gene prediction programs require validation by a second line of evidence, as unsupported gene predictions may have only a limited level of accuracy. Additional expression-independent methods, such as exon trapping, may yield only a few exons of a gene, so an additional strategy is required to confirm the full intron/exon structure.

Comparative mapping and sequencing could aid the identification of conserved genomic regions between model organisms and human which are likely to correspond to exonic or regulatory sequences. The premise for such analyses is that functionally important sequences are conserved, whereas other regions will differ as a result of accumulated mutations since their divergence.

As significant amounts of the mouse genome are now being sequenced, the opportunity to use the mouse sequence as an analytical tool to study the human genome has become increasingly attractive. This chapter therefore focuses on utility of mouse sequence for comparative study. The human and mouse species are estimated to have diverged from a common ancestor 100 million years ago (Burt *et al.*, 1999). The level of evolutionary divergence of the two genomes is, in general, great enough to allow identification of functionally conserved regions from the rest of the genomic background, yet small enough that comparison of syntenic linkage is meaningful (Lundin, 1993).

4.1.2 The Mouse Genome Projects

The mouse genome is roughly 3000Mb in size and a number of genetic maps have been constructed. Dietrich *et al*. (1996) (1996) published an intermediate resolution mouse genetic map based on single sequence polymorphisms. A refined map, based on microsatellite markers, was published in 1998 (Rhodes *et al.*). These genetic maps served as the framework for the construction of a YAC map (Nusbaum *et al.*, 1999). An RH map of the mouse genome, incorporating many markers from the genetic map, was produced in 1999 (Van Etten *et al.*, 1999). RH maps have the benefit of allowing incorporation of all sequence-based markers into an ordered framework. These framework maps provide the resources for the construction of bacterial clone contigs, including the determination of the bacterial clone maps of regions of the mouse genome orthologous to human chromosome 22 (section 4.2).

In 1999, the National Human Genome Research Institute (NHGRI) implemented a program to analyse the mouse genome and sequence areas of biological interest. A parallel approach of restriction enzyme fingerprinting (Coulson, 1996; Gregory *et al.*, 1997; Marra *et al.*, 1997;

Olson *et al.*, 1986) and landmark-content mapping (Green & Olson, 1990) is being taken. The *C.elegans* and human mapping projects (Coulson, 1996; Lander *et al.*, 2001) have demonstrated the utility of restriction enzyme fingerprinting. Fingerprinting has the advantage that the overlap between two clones is assessed over the entire length in shared fingerprint bands, thus providing information on the extent of overlap. Landmark content mapping is based on the detection of the presence or absence of a particular small genomic segment in a clone or clones. This can be done by hybridisation experiments in the laboratory or by electronic PCR (ePCR), a sequence comparison to determine if the STS can be detected in the available genomic sequence (Schuler, 1997). The major advantage of landmark content mapping is that it allows the ordering of clones based on their landmark content by integration with existing framework maps. Together, these methods provide an accurate means to assess the extent of overlap between clones and allow the ordering and anchoring of contigs based on their landmark content (figure 4.1).

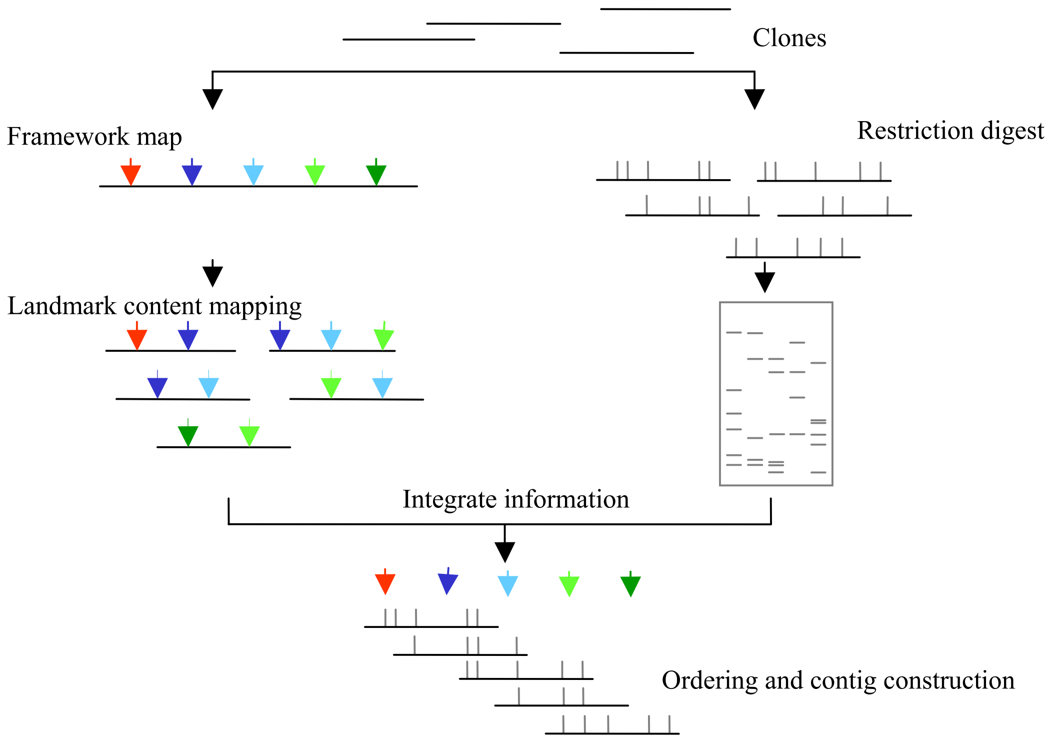


Figure 4.1: Contig construction strategy combining both landmark-content mapping and restriction enzyme fingerprinting (details are explained in the text).

Several different approaches can be used, known collectively as 'walking', to close gaps between contigs. New markers can easily be integrated into the existing framework map, or new markers that localise to the end of existing contigs can be used to isolate new clones. Alternatively, single sequence reads can be generated from clone ends using bacterial vector primers. Those sequences generated from contig ends can be used for STS design.

Resources that are now available for physical mapping projects include a database of over 300,000 fingerprinted clones from two BAC libraries constructed by P. de Jong from C57BL/6J mouse DNA (Marra *et al*., http://www.bcgsc.bc.ca/projects/mouse_mapping). One library, RPCI-23 (Osoegawa *et al.*, 2000) has been constructed from females and the other, RPCI-24, from males of the same strain. A database of sequences from the ends of the cloned genomic fragments has also been produced (Zhao *et al*., http://www.tigr.org/tdb/bac_ends/). These resources have been used to construct both small, regional BAC maps and more recently to assemble a larger physical BAC map of the whole mouse genome, now contained in fewer than 560 contigs. (The Mouse Genome Sequencing Consortium (MGSC), unpublished). The assembly incorporates 1251 framework markers previously placed on genetic and radiation hybrid maps by hybridisation assays or ePCR. A tiling path is currently being selected across the assembled BAC clone contigs, which will be subjected to standard shotgun sequencing, producing a working draft by 2003. The mouse BAC assembly has been imported into the mouse Ensembl database (http://mouse.ensembl.org), which includes predicted transcripts within finished and unfinished mouse sequence clone data.

A parallel effort to sequence the mouse genome was begun in 2000 by a public/private Mouse Sequencing Consortium (MSC). A whole genome shotgun (WGS) strategy has currently generated over 3-fold coverage of the mouse genome sequence. Initial assembly of these sequences has started. Assembled contigs will be anchored to the mouse BAC end sequences and the available RH and genetic marker data by ePCR. The WGS sequence will then be incorporated with the sequence generated from the MGSC mapping project (Collins, http://www.nih.gov/science/models/mouse/genomics/open_letter.html).

The biotechnology company Celera is also currently engaged in work to sequence the mouse genome, using a strategy similar to that used to sequence the human genome (see chapter I), although, in this case, publicly available sequence has not been included in the assembly process. The Celera assembled and annotated mouse genome is sequenced to over 5-fold coverage representing greater than 98% of the genome, but is only available through subscription (http://www.celera.com).

4.1.3 Comparative Analysis

4.1.3.1 Alignment packages

Human and mouse genomic sequence comparison are being increasingly used to search for evolutionarily conserved regions. A variety of programs are available that allow easy identification of conserved sequences that may correspond to functionally important segments and allow the identification of novel genes and possible regulatory elements.

Percentage Identity Plots (PIPs) (Schwartz *et al.*, 2000) have become a popular method of comparing mouse and human sequence, since they allow the display of conserved regions at a range of identity levels. PIPs use the SIM program (Huang *et al.*, 1990) to identify ungapped blocks longer than 50 bp with an identity > 50%. These blocks are then plotted against the length of one of the sequences. PIPs have been used in a number of studies in regional comparisons of human and mouse sequence (for example, Footz *et al.*, 2001; Martindale *et al.*, 2000).

The available mouse whole genome shotgun (WGS) sequence has been aligned with the assembled human draft sequence at the translated nucleotide level, using the BLAT alignment package (Kent, unpublished). The alignment can be viewed at http://genome.cse.ucsc.edu and http://www.ensembl.org. A further large-scale nucleotide alignment of the WGS sequence against the human draft sequence has been undertaken using the algorithm Exonerate (Slater, unpublished) (http://www.ensembl.org/Docs/wiki/html/EnsemblDocs/Exonerate.html).

4.1.3.2 Sequence conservation

A number of comparative sequence studies have been published, which demonstrate the conservation of exonic sequence between human and mouse genomes. Comparative sequencing of a number of regions in mouse and human, including the human and mouse β-globin gene cluster (Collins & Weissman, 1984; Shehee *et al.*, 1989); the human and rat γ-crystallin genes (den Dunnen *et al.*, 1989); the human and murine XRRC1 DNA repair gene regions (Lamerdin *et al.*, 1995); the human, mouse and hamster ERCC2 regions (Lamerdin *et al.*, 1996); a gene rich cluster at human chromosome 12p13 and its syntenic region on murine chromosome 6 (Lamerdin *et al.*, 1996); the mouse and human AIRE regions (Blechschmidt *et al.*, 1999); human and mouse T-cell receptor C-δ and C-α regions (Koop & Hood, 1994); human and hamster α - and β-myosin heavy chain genes (Epp *et al.*, 1995); human and murine Bruton's tyrosine kinase loci (Oeltjen *et al.*, 1997); the human and murine ABCA1 regions (Qiu *et al.*, 2001), has underlined the value of comparative sequence for gene annotation.

Conservation of non-coding sequences may, in some cases, arise due to functional constraint, or may be the result of a lack of divergence time. The latter premise suggests that different portions of the human and rodent genomes may evolve at different rates (Hardison *et al.*, 1997; Koop, 1995; Wolfe *et al.*, 1989). This was supported by Makalowski *et al.*(1998), who demonstrated that protein sequence conservation varied from 36% to 100% in a set of 1196 orthologous mouse and human protein sequences.

Many of the regions conserved between the human and mouse genome may correspond to yet unidentified human genes. A recent study, which described the annotation of 21, 076 full-length mouse cDNAs (Kawai *et al.*, 2001), identified 817 mouse transcripts for which no corresponding human gene had been described. The data indicates that comparative sequence analysis could be an important tool in identification of previously unknown genes.

Additionally, conserved non-coding regions may highlight regulatory sequences. Gumucio *et al*. (1988) described such a comparison of potential human and mouse promoter sequences, in order to identify the determinant of tissue specificity of amylase gene expression. The first large-scale study of non-coding sequences compared 100 kb of human and mouse DNA containing the T-cell receptor family (Hood *et al.*, 1995). The non-coding regions of this gene cluster proved to have an unusually high level of sequence conservation. In a more typical 100 kb segment from chromosome 2p13, 1% of the sequence was accounted for by conserved elements of length >80 bp with sequence identity >75% (Jang *et al.*, 1999). Loots *et al*. (2000) demonstrated the function of a conserved non-coding segment from a multi-species sequence comparison of a 1 Mb region containing an interleukin gene cluster. Deletion of a conserved non-coding element was shown to alter interleukin expression in T cells of transgenic mice.

4.1.3.3 Chromosome evolution

Comparative analysis of human genetic and physical maps with those of other organisms, has allowed mapping of the synteny relationships. Chromosome 22, for example, is a recently formed chromosome that is only found in higher primates. In lemurs and most other primates, information from HSA22 is found on at least two different chromosomes, both of which also contain different subsets of HSA12 (Muller *et al.*, 1999). These human chromosomes are posited to have formed from a single reciprocal translocation involving two ancestral chromosomes (Haig, 1999). In contrast, information from HSA22 is found at 21 different sites on eight different mouse chromosomes.

Several studies have suggested that repeated sequences might be associated with genetic instability, possibly leading to evolutionary rearrangement events. For example, the breakpoint of translocations (HSAXp11; HSA1q21) associated with papillary renal cell carcinoma (RCC) were mapped to a small region of HSA1q21 between SPTA1 and a clustered gene family, including CD1C, CD1B, CD1D, CACY and at least four other members (Weterman *et al.*, 1996). Interestingly, the boundary between two segments of HSA1q21 that are related to MMU1 and MMU3 respectively, is located between SPTA1 and CD1C, a region of <200 kb (Oakey *et al.*, 1992). Amadou *et al*. (1995) also reported a syntenic breakpoint in the HSA6p MHC class I gene region, within a tandemly organised family of genes. Related sequences are found on both MMU13 and MMU17.

Sequence analysis permits finer scale mapping of the human-mouse synteny relationships. Pletcher *et al*. (2000), has described the first sequence level analysis of a synteny breakpoint at one of these sites, an 18 kb region of mouse chromosome 10 (MMU 10) containing the junction of material represented on HSA21 and HSA22. The minimal junction region on MMU10 contains a variety of repeats, including an L32-like ribosomal element and low-copy sequences found on several mouse chromosomes and represented in the mouse EST database. Similar comparative sequence studies could yield further information about the mechanisms of chromosomal evolution.

4.1.4 This chapter

This chapter aims to examine the importance of comparative mapping and sequencing in identifying genes and their control regions. The construction of three mouse clone contigs across the orthologous regions of human chromosome 22 is described. Generated mouse genomic sequences, in both finished and unfinished form, were used in extensive comparative analyses against orthologous human sequences. Dot and percentage identity plots showed extensive conservation of coding regions. The extent of the correlation between the conserved mouse sequence evidence and the annotated transcript map of 22q13.31 was analysed and compared with sequence evidence from other model organisms.

Conserved non-coding sequences were examined for the presence of potential exonic or regulatory features. More detailed analysis of gene structures and sequence content was undertaken on a 0.5 Mb region of finished mouse sequence. This region included sequence from a mouse clone found to span an ‘unclonable’ region in the human chromosome 22 sequence (Dunham *et al.*, 1999).

The utility of mouse genome sequence in the analysis of synteny breakpoints was also examined. A synteny breakpoint junction region between mouse chromosomes 15 and 8 on human chromosome 22q13.1 was refined through comparative analysis of human and unfinished mouse sequence and the sequence of the junction region was analysed.

4.2 Production of regional mouse BAC maps

4.2.1 Bacterial clone contig construction

The initial framework map used for anchoring bacterial clone contigs was the chromosome 22 transcript map (Dunham *et al.*, 1999). BLAST searches were used to identify mouse cDNA sequences orthologous to cDNAs situated within the 3.4 Mb region of human chromosome 22q13.31 and a 1.9 Mb region of 22q13.1. STSs were designed to the 19 mouse mRNA sequences that were identified by this method. To increase marker density, 39 further STSs were designed from mouse ESTs that demonstrated a level of 100% nucleotide identity to the set of human cDNAs.

In order to isolate mouse clones spanning the three orthologous regions of interest, 11.2X genome equivalents of the female mouse BAC library RPCI-23 (strain C57BL/6J) (Osoegawa *et al.*, 2000) were screened by hybridisation (see figure 4.3).

In initial library screens, four pools of STS PCR products were used. The pools identified 111, 135, 199 and 132 clones respectively (table 4.1). In total, 307 clones were identified (taking redundancy into account). The identified BAC clones were transferred into microtitre plates to form a region-specific library subset. To verify the identified clones, arrayed clone filters (polygrids) were screened with all the markers from the pools individually (figure 4.2). Both the verification and the initial screening data were collated and integrated into 22ace.

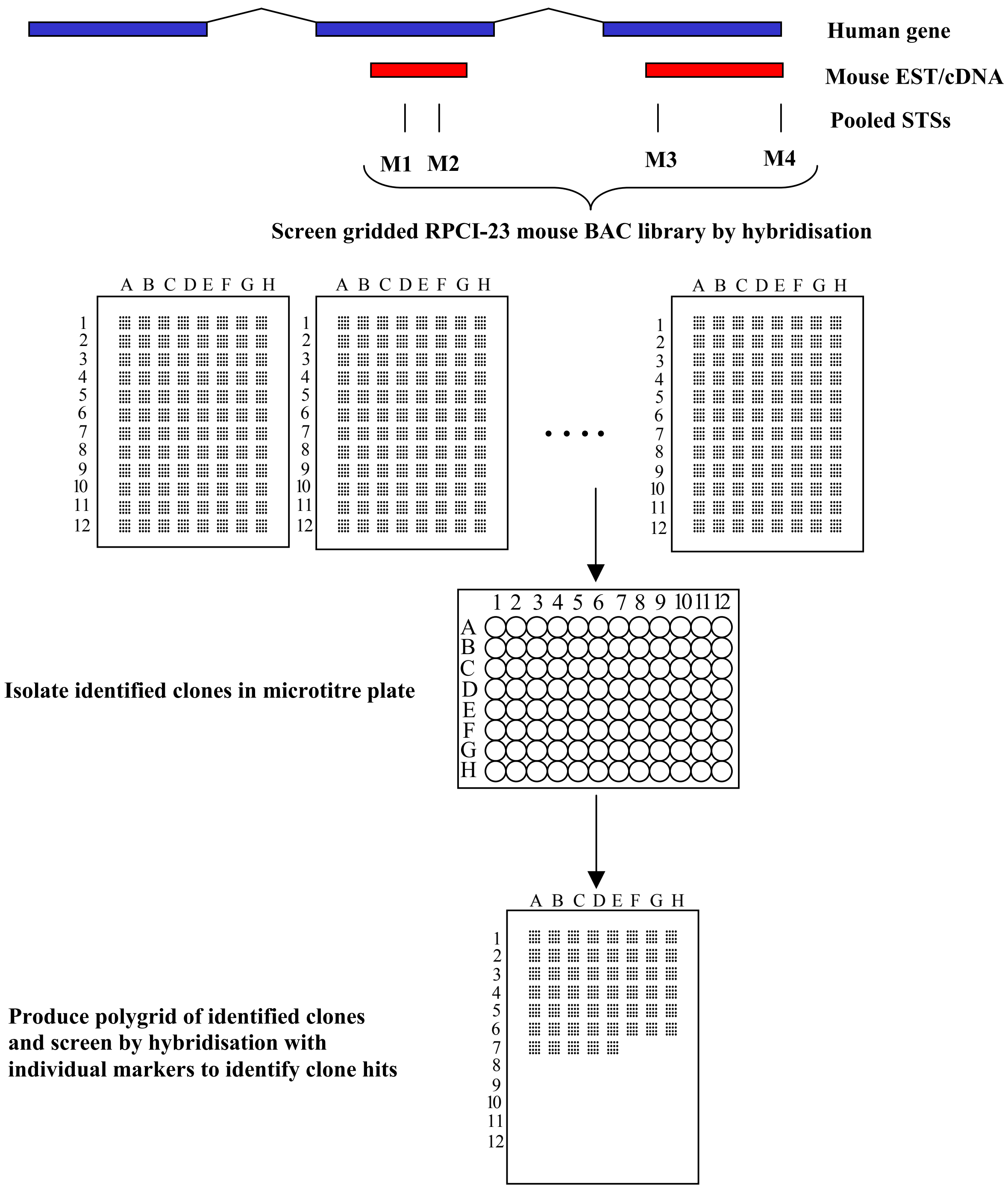

Human gene
Mouse EST/cDNA
Pooled STSs
M1
M2
M3
M4
Screen gridded RPCI-23 mouse BAC library by hybridisation
A B C D E F G H
1 2 3 4 5 6 7 8 9 10 11 12
Isolate identified clones in microtitre plate
Produce polygrid of identified clones and screen by hybridisation with individual markers to identify clone hits

Figure 4.2: Screening strategy. Mouse cDNAs/ESTs homologous to human genes were used to design PCR primers (M1-M4). These were pooled and used to screen arrayed filters of the mouse RPCI-23 bacterial clone library. All identified positive clones were transferred to microtitre plates and gridded onto a specific mouse polygrid. This was then screened with the individual markers to identify specific clone-marker relationships.

Table 4.1: Numbers of pools, markers and isolated clones in the initial library screens

Pool	Contains marker type		BACs
	mRNA	EST	
Pool1	11	0	111
Pool2	1	18	135
Pool3	1	11	199
Pool4	10	10	132
Total	23	39	577
			307*

* Taking into account redundancy between the pools

4.2.2 Fingerprinting

BAC clones from duplicate copies of the microtitre plates were fingerprinted using *Hin*dIII (chapter II). The contigs were built using the program FPC (fingerprinting contig) (Soderlund *et al.*, 1997). FPC automatically clusters fingerprinted clones into contigs using a probability of coincidence score. FPC also allows integration of landmark content data with the fingerprint data, thus providing a workbench for contig assembly, verification and selection of sequence tile path clones.

4.2.3 Landmark content mapping

In addition to fingerprinting, maps were also constructed by landmark-content mapping. Polygrids were screened with each of the markers generated from cDNA information. From the hybridisation results, contigs could be constructed based on shared landmark content using the strategy described in figure 4.1. The initial rounds of screening led to the construction of 33 contigs spanning an estimated 3.8 Mb. (Comparison of sequence and fingerprint data determined that for the mouse library clones, a single fingerprint band corresponded to an average of 5 kb. This figure was used to estimate the size of a region based on the number of fingerprinting bands.)

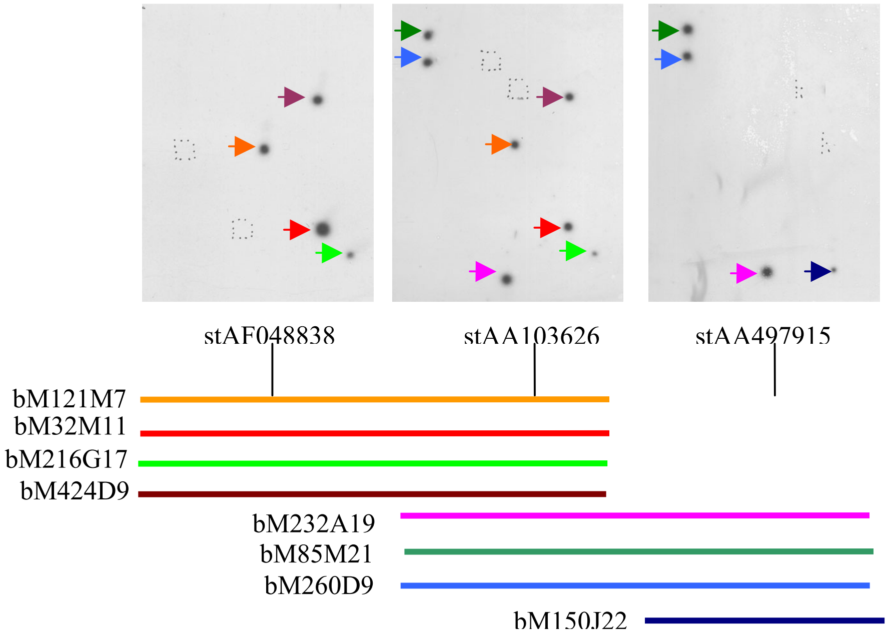


Figure 4.3: Example of landmark-content mapping using three landmarks (stAF048838, stAA103626 and stAA497915). The positives are indicated by coloured arrows, the clones drawn below in corresponding colours.

4.2.1.4 Gap closure

Two strategies were utilised to link the contigs. Initially, the publicly available BAC clone end sequences (Zhao *et al.,* unpublished) were used to design PCR primers to those BACs on the ends of the contigs for further library screens. Five pools of clones were screened in two successive rounds of walking which resulted in the identification of 508 clones. Subsequent fingerprinting and mapping of these clones allowed 25 gaps to be filled.

Table 4.2: Numbers of pools, end STSs and isolated clones in gap closure screens

Pool	End STS	BACs
Pool5	17	137
Pool6	23	203
Pool7	23	122
Pool8	23	186
Pool9	17	132
Total	103	880
		508*

* Taking into account redundancy between the pools

As an increasing number of fingerprints (Marra *et al*., unpublished) and end sequences (Zhao *et al*., unpublished) from the mouse BAC library became available, they were anchored by ePCR and hybridisation using publicly available genetic and radiation hybrid markers (Gregory *et al*., unpublished)(section 4.2.5). Incorporation of this data enabled closure of two gaps. Additionally, the information allowed two spurious contigs, containing 261 clones and 31 markers designed to murine genes or EST sequences, that did not map to the correct mouse chromosome and 68 singletons to be discarded.

NB. The three contigs generated during this project have since been incorporated into the large-scale physical mouse mapping effort. Further work has resulted in joining of the two mouse chromosome 15 contigs, creating a contig spanning approximately 6.7 Mb of mouse sequence.

4.2.4 Tile Path Clones

During contig construction, clones with sufficient mapping information (i.e. both landmark and fingerprinting data) were selected for sequencing (Richard Evans, Sanger Institute and M. Goward). Tiling path clones across the three contigs were selected to ensure that minimal overlap of clones reduced redundant sequencing.

4.2.5 Features of the sequence-ready bacterial clone map

The three contigs incorporated 486 BAC clones in total and the final sequence tile paths, containing 34 clones, cover an estimated 3.96 Mb (excluding overlapping sequences). The clone contigs are depicted in figure 4.4. The division of this set of clones is summarised in table 4.3.

Table 4.3: Clone contig data showing the number of clones within the contigs, the number of clones selected for sequencing and the approximate length of the contig

Contig	Mouse chromosome	Orthologous region	Total # clones	# clones in tile path	Approx. length (Mb)
A	15	22q13.31	229	13	2.00
B	15	22q13.1	164	15	1.59
C	8	22q13.1	93	6*	0.37

*including two clones sequenced by the Albert Einstein College of Medicine Human Genome Research Center (AECOM) and the University of Oklahoma Advanced Center for Genome Technology (UOKNOR) respectively.

The maps also incorporate 54 markers from a range of mouse maps listed in the UniSTS (http://www.ncbi.nlm.nih.gov/genome/sts/index.html) database, that have been positioned by ePCR against available mouse sequence (Gregory *et al*, unpublished). Shared markers between different map types allow integration of the sequence-ready map with previously published mouse maps and confirmed the chromosomal location of the mouse contigs. The incorporation of marker types into the contigs is shown in table 4.4.

Table 4.4: Incorporation of marker information into mouse contigs A, B and C

Contig	Mouse chromosome	Orthologous human region	Marker Type mRNA	EST	End STS	UniSTS	Total no. markers
A	15	22q13.31	4	7	27	15	53
B	8	22q13.1	5	2	5	6	19
C	15	22q13.1	6	6	6	33	55

4.2.6 Sequencing

The tile path clones were shotgun sequenced (chapter I) (Sanger Institute sequencing teams). During the project, sequence was released by other groups for several other clones in the contigs. Where possible, these clones were incorporated into the tile path to minimise redundant sequencing.

At the time of writing, finished sequence was available for nine (26%) clones and unfinished shotgun sequence was available for a further 18 (53%) of the 34 tiling path clones. These clones are highlighted in the FPC display shown in figure 4.5. A table of the sequenced mouse clones showing their genomic location, accession number, author, orthologous human region and current sequencing status is shown in appendix 5.

Approximately 85% of 22q13.31 is spanned by mouse clones that have at least unfinished sequence. Approximately 92% of the region of human chromosome 22q13.1 under investigation is spanned by unfinished/finished mouse sequence (see figures 4.2 and 4.6).

Figure 4.4 (foldout): Bacterial clone contigs containing mouse genomic sequence spanning regions of conserved synteny with a) human chromosome 22q13.31 and b) human chromosome 22q13.1. The human transcript map of each region is depicted at the top of each diagram: full genes are shown in dark blue, partial in light blue and pseudogenes in green. Gene structures orientated 5' to 3' on the DNA strand from centromere (left) to telomere (right) are designated '+' and those on the opposite strand '-'. Markers designed from murine sequences orthologous or similar to the named human genes are shown in black. These markers are positioned relative to both the human transcript map and the mouse clone contigs. Mouse chromosome specific markers from the UniSTS database are shown in pink and are positioned relative to the mouse clone contigs only. The .15 or .8 of these marker names refers to the specific murine chromosome. Conserved mouse genes (identified from dot and PIP analyses (section 4.3) are indicated by red arrows. The mouse clone contigs are shown in red below. Figure a shows part of contig A, a region of MMU15 with conserved synteny to 22q13.31. Figure b. shows parts of contigs B and C, from MMU8 and MMU15 respectively. The hashed red blocks denote clones that extend beyond the region of synteny with HSA22q13.1. Only relevant regions of the contigs are shown: clones that extend these contigs further have been mapped (see figure 4.5) but do not yet have sequence.

TAKE THIS PAGE OUT – foldout figure 4.4a

Take this page out too!!! fig 4.4b

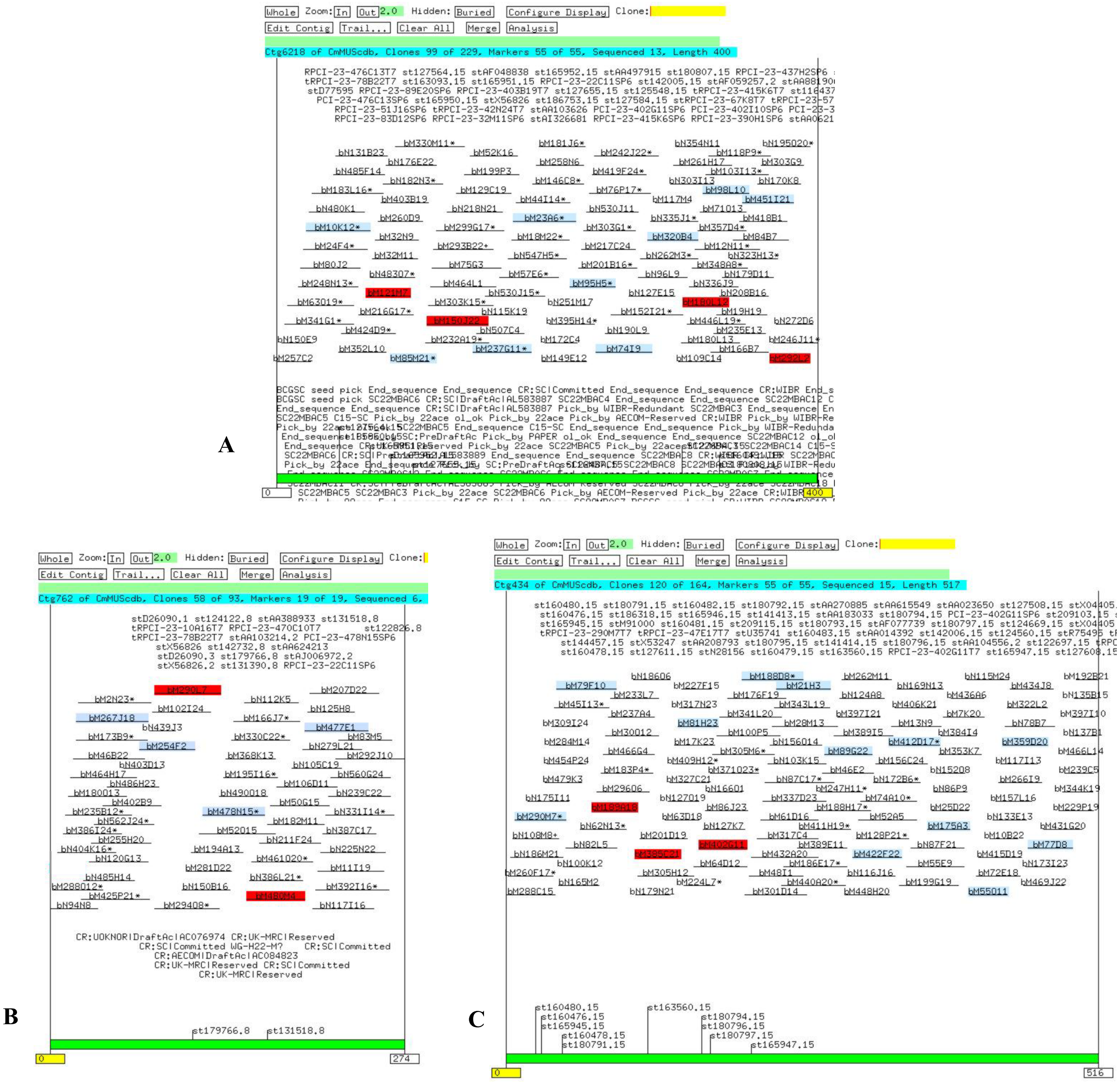

A
Whole Zoom: In Out 2.0 Hidden: Buried Configure Display Clone:
Edit Contig Trail... Clear All Merge Analysis
Ctg6218 of CmMUScdb, Clones 99 of 229, Markers 55 of 55, Sequenced 13, Length 400
0
400
B
Ctg762 of CmMUScdb, Clones 58 of 93, Markers 19 of 19, Sequenced 6,
0
274
C
Ctg434 of CmMUScdb, Clones 120 of 164, Markers 55 of 55, Sequenced 15, Length 517
0
516

Figure 4.5: FPC display of mouse BAC clone contigs spanning orthologous regions of HSA22.

A) Contig spanning region of mouse chromosome 15, orthologous to human chromosome 22q13.31.

B and C) Contigs spanning regions of mouse chromosomes 8 and 15 respectively, encompassing a synteny breakpoint with human chromosome 22. Contig diagrams extracted from FPC (Soderlund *et al.*, 1997). Tiling paths are indicated in blue and finished sequence clones are highlighted in red.

4.3 Comparative sequence analysis

4.3.1 Dot plot analysis

Available sequence from the three mouse clone contigs (appendix 5) was compared against the orthologous human sequence using the dot plot program from the advanced PipMaker analysis tools available at http://bio.cse.psu.edu/pipmaker (Schwartz *et al.*, 2000). This program is similar to Dotter (Sonnhammer & Durbin, 1995), used in chapter III, but reports only matches contained within a statistically significant alignment. Another feature of this program is that unfinished sequence contigs can be ordered according to their alignment to a second, base sequence. Figures 4.6a and 4.6b show annotated dot plots of the two regions of chromosome 22, aligned against the mouse ordered sequence contigs. Of course, the ordering of the mouse unfinished sequence contigs derived from PipMaker is dependent upon the human reference sequence. The order shown is therefore currently unconfirmed.

The dot plots above show that areas of high similarity correspond to single or multiple genes. In regions of finished sequence, gene order and orientation appear to be conserved between human and mouse. This is supported by the distribution of markers within the contigs, shown in figure 4.4. An apparent inversion of APOL2 exists in AL592187.4, but this is likely due to the unfinished nature of this sequence. Figure 4.6a indicates that two mouse clone sequences, AL513354.14 (finished) and AL603714.4 (unfinished), span sequence gaps in the human sequence of 22q13.31. A more detailed analysis of the finished sequence AL513354.14 is shown in section 4.7. Figure 4.6b confirms the existence of a synteny junction region on human chromosome 22, between genes dJ569D19.C22.1 and MB. This is discussed in more detail in section 4.7.

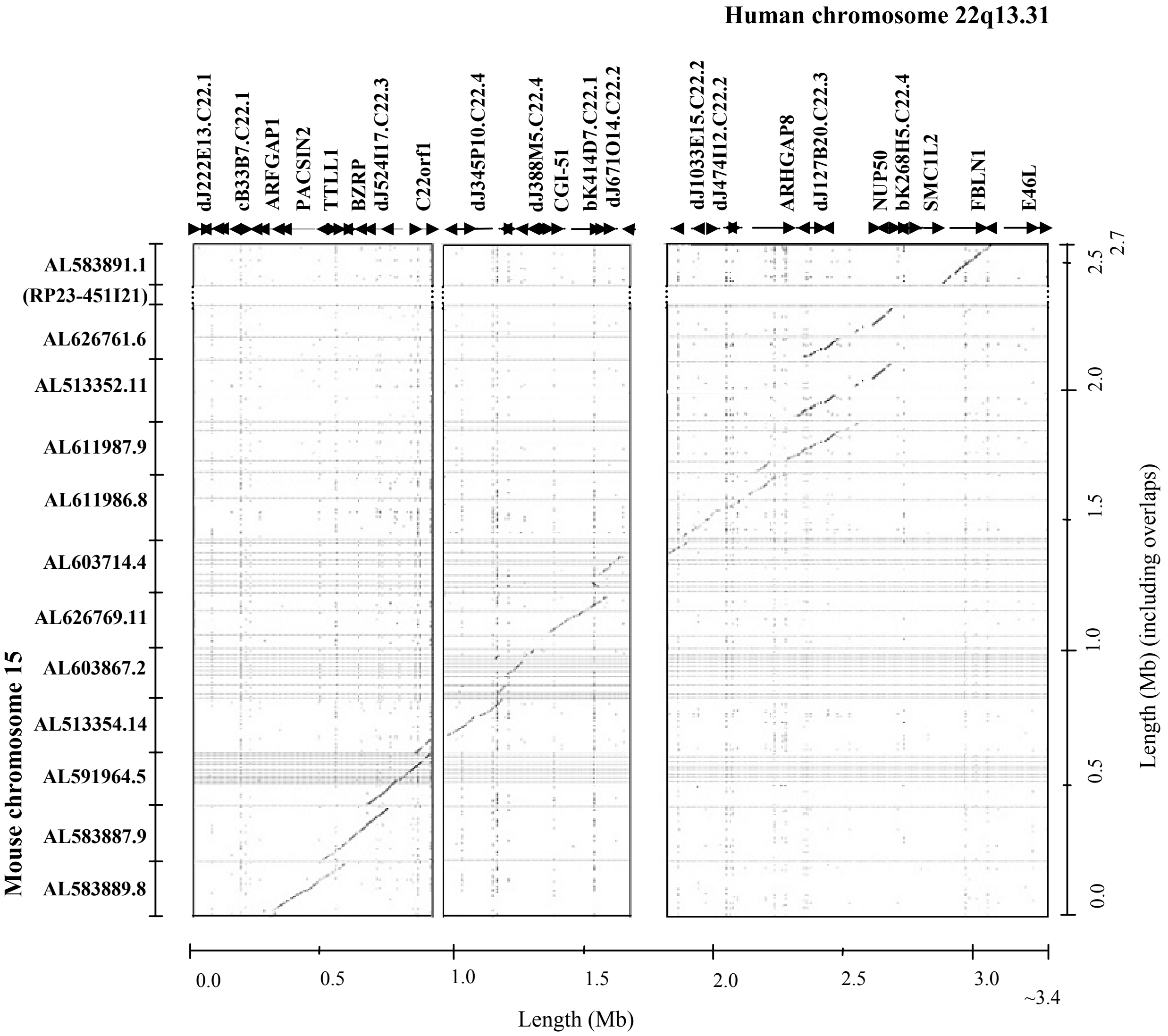
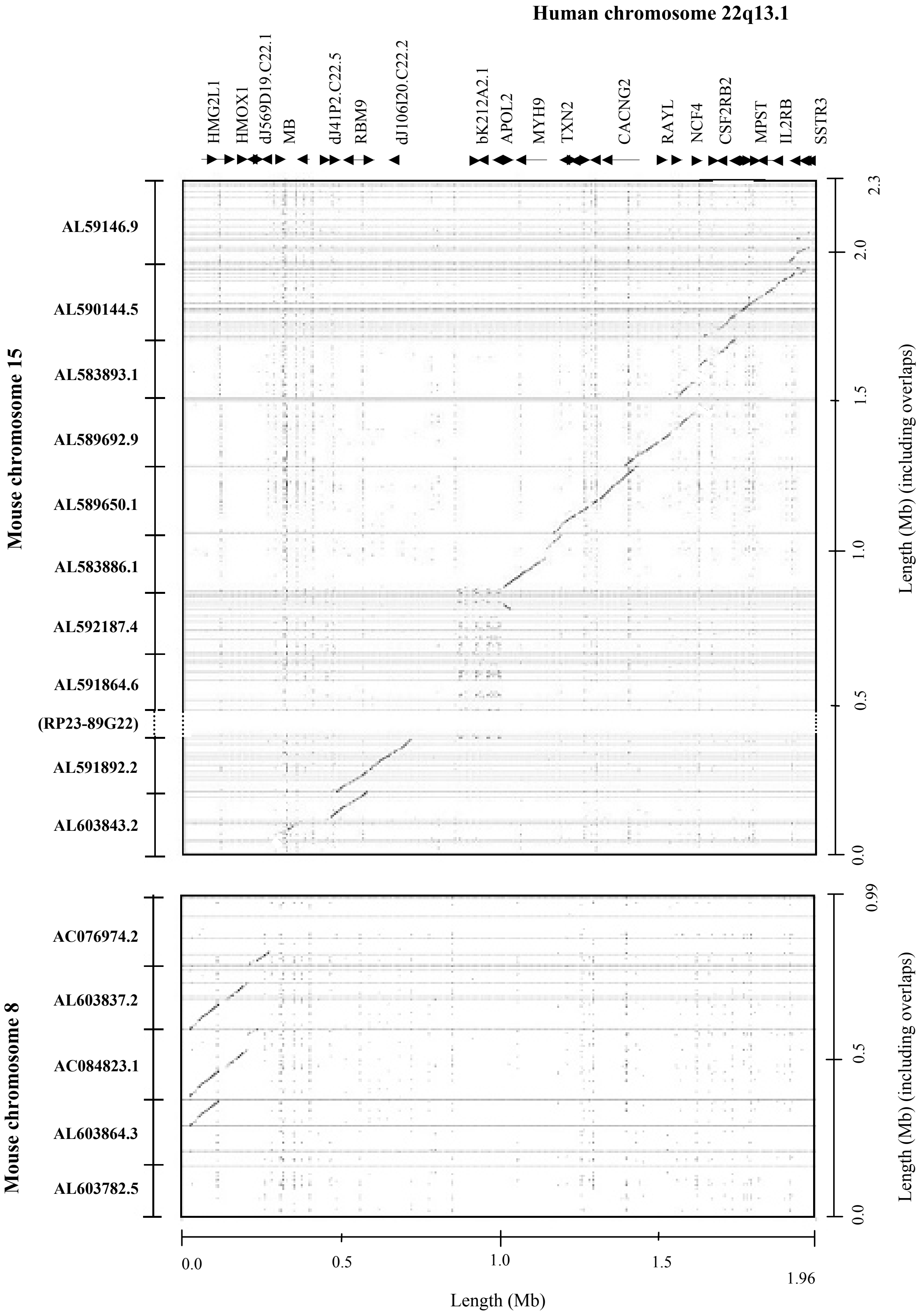


Figure 4.6a: Annotated dot plot of the human sequence of 22q13.31 (X-axis) and orthologous mouse (Y-axis) sequences from MMU 15. Genes present in the human sequence are indicated along the X-axis. Two sequence gaps of approximately ~50kb and ~75kb respectively are shown in the human sequence. The dot plot indicates that these gaps are spanned by the finished mouse sequence AL513354.14 and the unfinished sequence AL603714.4 respectively. Tiling path clone RP23-451I21, for which sequence is not yet available, spans a gap in the mouse sequence.

Figure 4.6b (overleaf): Annotated dot plot of the human sequence of a 1.96 Mb region of 22q13.1 (X-axis) and orthologous mouse (Y-axis) sequences from MMU15 and MMU8. Genes present in the human sequence are indicated along the X-axis. Tiling path clone RP23-89G22, for which sequence is not yet available, spans a gap in the mouse sequence. Further mapped clones have been selected for sequencing, which extend the tiling path along MMU15. However, sequence is not yet available for these clones and these have not been included in the diagram. The dot plot indicates that a MMU8:15 synteny junction exists between genes dJ569D19.C22.1 and MB on 22q13.1 (section 4.8).


Human chromosome 22q13.1
HMG2L1
HMOX1
dJ569D19.C22.1
MB
dJ41P2.C22.5
RBM9
dJ106I20.C22.2
bK212A2.1
APOL2
MYH9
TXN2
CACNG2
RAYL
NCF4
CSF2RB2
MPST
IL2RB
SSTR3
Mouse chromosome 15
AL59146.9
AL590144.5
AL583893.1
AL589692.9
AL589650.1
AL583886.1
AL592187.4
AL591864.6
(RP23-89G22)
AL591892.2
AL603843.2
2.3
2.0
1.5
1.0
0.5
0.0
Length (Mb) (including overlaps)
Mouse chromosome 8
AC076974.2
AL603837.2
AC084823.1
AL603864.3
AL603782.5
0.99
0.5
0.0
Length (Mb) (including overlaps)
0.0
0.5
1.0
1.5
1.96
Length (Mb)

4.3.2 PIP analysis - investigation of exonic conserved sequences

Repeat elements in the human and mouse sequences were identified and masked using RepeatMasker (Smit and Green, unpublished) and the resulting sequences and exon locations were submitted to the PipMaker website (http://bio.cse.pse.edu/pipmaker) (Schwartz *et al.*, 2000) (section 4.1.3.1). An overview of conserved gene structures, derived from the PIP comparisons, is shown in figure 4.4. An example of a PIP, showing in finer detail the alignments made between a region of the human and mouse sequences, is shown in section 4.7.

The coding exons of conserved genes are easily identified by visual inspection of the PIPs. Untranslated regions of exons often show a decrease in percent identity compared to the protein-coding portion of the gene (see the BZRP gene region from ~112K to 124K in figure 4.12). The number of human gene features from each region demonstrating >50% nucleotide identities to gap-free segments of mouse sequence are listed in table 4.5. Overall, over 75% of the annotated human exons, which lay within regions spanned by finished/unfinished mouse sequence, could be aligned with conserved sequences in the mouse.

Interestingly, no pseudogenes showed homology to the mouse sequence outside of repeat regions. The existence of a human pseudogene on human chromosome 22 (CYKB2-ps) that does not have a murine orthologue, has previously been demonstrated by Lund *et al*. (2000) through comparative sequence analysis. A further study has described non-conservation of the human pseudogene EEF1B3 in the mouse genome, although this research was not performed at sequence level (Chambers *et al.*, 2001). These human pseudogenes may have arisen since the divergence of the human and mouse lineages. Alternatively, these non-functional sequences may have diverged more rapidly in the mouse genome, perhaps because of the shorter murine generation time.

Additionally, no homology was found in the mouse to four human genes: HMG17L1 and dJ1033E15.C22.1 from 22q13.31, and dJ1119A7.C22.4 and dJ1119A7.C22.5 from 22q13.1. This list is not definitive, as analysis of the finished sequence may show further differences. These four gene structures are categorised as partial (see chapter III). It may be that sequence conservation of these genes will be noted when the complete mouse sequence is available. Alternatively, some or all of these features may be pseudogenes (see above) or may be true genes that are not conserved in the mouse sequence.

Table 4.5: Overview of PIP results from comparisons of available mouse genomic sequence to two regions of human chromosome 22.

		No. human gene features spanned by sequenced mouse clones (finished and unfinished sequence)				No. human gene features demonstrating >50% nt. identity to gap-free segments of mouse sequence			
Human Region	Mouse coverage (%)	No. genes	No. exons	No. pseudo -genes	No. pseudo -gene exons	No. genes	No. exons	No. pseudo -genes	No. pseudo -gene exons
22q13.31	85	29	378	12	12	26	243	0	0
22q13.1 (MMU 8)	92	4	55	1	3	4	53	0	0
22q13.1 (MMU15)		29	199	5	5	27	183	0	0
Total	88.5	62	632	18	20	57	479	0	0

Sequence from HSA 22 (6 Mb) was compared against syntenic mouse sequence using the PipMaker website (http://bio.cse.pse.edu/pipmaker) (Schwartz *et al.*, 2000). The resulting PIP was analysed by eye. Coverage shows the estimated amount of the human sequence (%) for which the equivalent orthologous mouse sequence (finished or unfinished) is available. The number of genes and pseudogenes annotated within the human 'covered' region is shown, together with the total number of exons in each category. The numbers of genes, pseudogenes and exons that demonstrate >50% nucleotide identity to gap-free segments of mouse sequence are listed.

4.3.3 Integration of mouse genomic data into 22ace

In order to allow detailed comparison between the mouse genomic data generated during this project, the annotated gene structures described in chapter III and additional data such as

Genscan predictions, it was necessary to generate an alignment of the available mouse genomic sequence with the sequence of 22q13.31 in a format that could be incorporated into the 22ace database.

The program MatchReport (Smink *et al*., unpublished) generates an ace format file from BLAST alignments above a set percentage identity. In order to determine an appropriate value for percentage identity for a local alignment of orthologous mouse unfinished sequence data against human 22q13.31, a preliminary comparison was performed, using three mouse clone sequences against the orthologous human regions using MatchReport at a range of percentage identity values. Repeats in the sequences were masked using RepeatMasker prior to alignment (Smit and Green, unpublished). The compared regions are shown in table 4.6

Table 4.6: Mouse clones and orthologous regions of HSA22q13.31 selected for percentage identity calibration experiment

Mouse clones	Orthologous region of HSA 22q13.31	Size of region (human) (kb)	No. annotated human genes	No. annotated human exons
AL603867, AL513354	dJ345P10.C22.4 – dJ388M5.C22.4	300	3	52
AL583887	TTLL1 – dJ526I14.C22.3	150	6	60
	Total	**450**	**9**	**112**

The generated files were read into 22ace. Values of specificity and sensitivity for each percentage identity value (see chapters II and III) were calculated at a nucleotide level and plotted in figure 4.7.

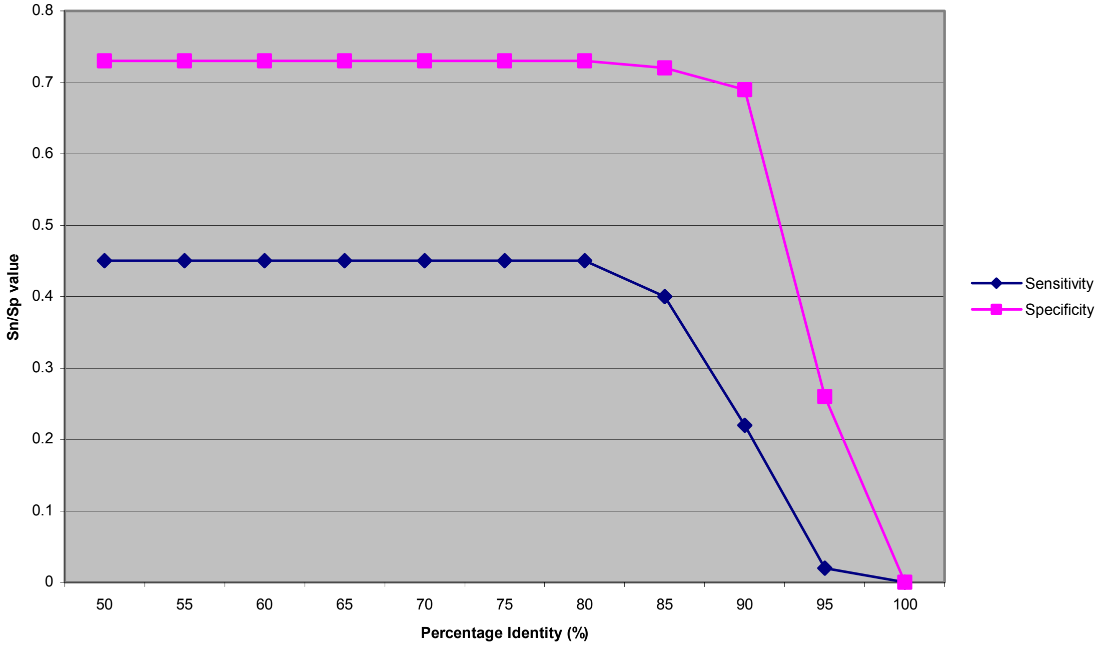


Figure 4.7: Sensitivity and specificity of MatchReport BLAST results from three mouse clone sequences against the equivalent human genomic sequence. The perl script MethComp (D. Beare) was used to calculate specificity and sensitivity of mouse hits to nucleotides contained within exons

These results show that both specificity and sensitivity are compromised if the percentage identity level is raised beyond 80% in this region. Surprisingly, sensitivity did not increase, or specificity decrease, as percentage identity dropped below this level to 50%. A cut-off identity level of 80% was therefore deemed appropriate for a comparative study of this region in order to maximise specificity, without loss of sensitivity. Available mouse sequence from contig A was thus aligned to the human sequence from 22q13.31 using MatchReport at a percentage identity of 80%.

4.4 Correlation of comparative genomic data with 22q13.31 transcript map

The mouse WGS sequence (MSC, unpublished) has been aligned to the draft human genomic sequence using BLAT and Exonerate (section 4.1.3.1). Results specific to HSA22 have been incorporated into 22ace. Additional sequence resources, derived from mouse and other organisms and incorporated into the 22ace database, include sequence from a library of full-length mouse cDNAs (Kawai *et al.*, 2001), output from the ExoFish program (Roest Crollius *et al.*, 2000), which assesses TBLASTX sequence homology to available *T. nigroviridis* genomic sequence, and the translated predicted protein sequences from the *D. melanogaster* (Adams *et al.*, 2000) and *C. elegans* (Coulson *et al.*, 1996) sequencing projects. An example of a 22ace display showing alignment of these features to the gene dJ526I15.C22.2 is shown in figure 4.8. The diagram shows that both mouse genomic sequence resulting from this project and mouse cDNA sequence (Kawai *et al.*, 2001) both align to the human sequence along the full length of the gene dJ526I14.C22.2. Output from the Exofish program (Roest Crollius *et al.*, 2000) aligns to only two exons of this gene.

The perl script MethComp (Dave Beare, unpublished) was used to compare the different methods used for gene identification/annotation against:

A. The set of 39 annotated ‘true’ genes within 22q13.31,

B. The set of 17 annotated pseudogenes within 22q13.31.

Specificity and sensitivity calculations were perfomed at the nucleotide level for all method types. The fraction of exon hits (the number of reference exons hit/total number of reference exons) and gene hits (the number of reference genes hit/total number of reference genes) were also calculated, as before (chapter III). In all cases, multiple hits were counted as one hit.

These results are shown in table 4.7. A plot of the specificity and sensitivity of each type of evidence at the nucleotide level is shown in figure 4.9. Further details of this analysis can be found in chapter II.

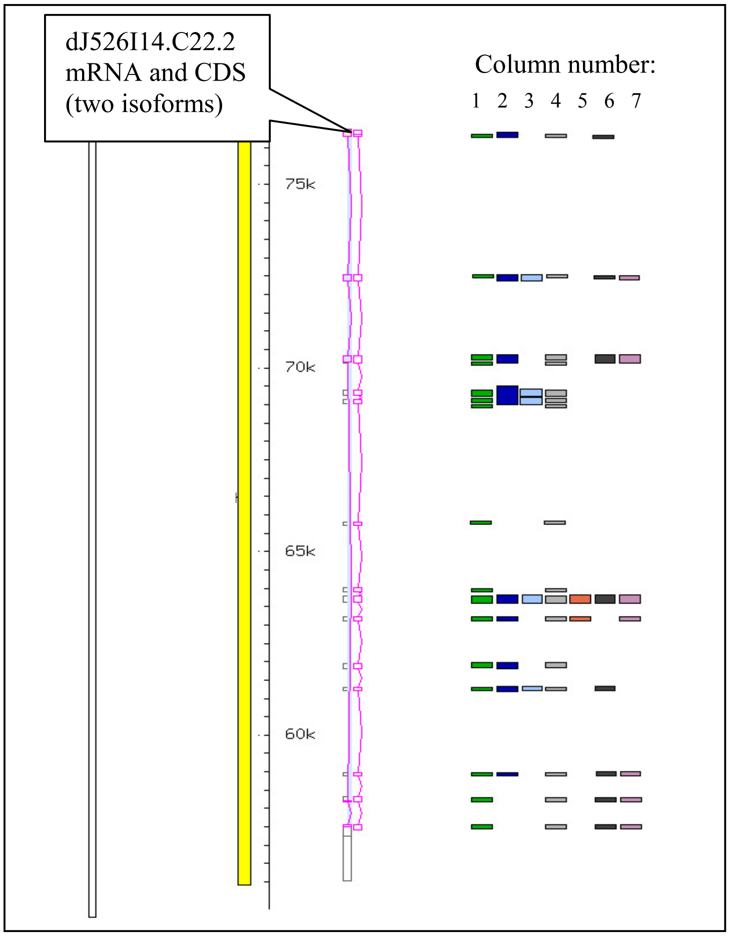


Figure 4.8: 22ace display showing the region surrounding the gene dJ526I14.C22.2. Sequence alignments are shown in columns to the right of the gene structure. Two isoforms of dJ526I14.C22.2 are depicted.
1= Blastn_mus: genomic mouse sequence generated as a result of this project.
2 = Blatmouse: WGS mouse sequence (MSC, unpublished) aligned against the draft human genome sequence with BLAT (Kent, unpublished).
3 = ExoMouse: WGS mouse sequence (MSC, unpublished) aligned against the draft human genome sequence with Exonerate (Slater, unpublished).
4 = fantom: Collection of full-length mouse cDNA sequences (Kawai *et al.*, 2001).
5 = Exofish: Exon prediction program utilising T. nigroviridis genomic sequence (Roest Crollius *et al.*, 2000).
6 = flypep: translated predicted D. melanogaster genes (Adams *et al.*, 2000).
7 = wormpep: translated predicted C. elegans genes (Coulson, 1996).

Additional features have been removed from the display to aid clarity.

Table 4.7 Analysis of the correlation of the evidence types available from different organism genome or gene identification projects used to annotate genes against:

A: 39 annotated true genes in 22q13.31.

Evidence type	Method	Organism	Alignment	Nucleotide			Exon	Gene
				Total Coverage	Sp	Sn		
Genomic	Blastn_mus*	M. musculus	BLASTN	0.016	0.62	0.36	0.60	0.88
Genomic	Blatmouse*	M. musculus	BLAT	0.015	0.51	0.27	0.53	0.78
Genomic	Exomouse*	M. musculus	Exonerate	0.017	0.45	0.26	0.50	0.82
cDNA	fantom*	M. musculus	BLASTN	0.002	0.49	0.03	0.10	0.34
Exon prediction	Exofish*	T. nigroviridis	ExoFish	0.005	0.76	0.12	0.30	0.58
Protein	flypep*	D.melanogaster	BLASTX	0.006	0.69	0.15	0.33	0.56
Protein	wormpep*	C. elegans	BLASTX	0.002	0.58	0.04	0.10	0.17

* Descriptions and references of each method are given in the legend of figure 4.8.

The test region (22q13.31) contained 3,365,293 bp of genomic sequence. The total number of nucleotides contained within the 39 annotated genes structures is 91,249 bp. The total number of reference exons is 400. For more details, see chapter II.

B: 17 annotated pseudogenes in 22q13.31.

Evidence type	Method	Organism	Alignment	Nucleotide			Exon	Pseudogene
				Total Coverage	Sp	Sn		
Genomic	Blastn_mus*	M. musculus	BLASTN	0.016	0.00	0.00	0.00	0.00
Genomic	Blatmouse*	M. musculus	BLAT	0.015	0.12	0.44	0.58	0.76
Genomic	Exomouse*	M. musculus	Exonerate	0.017	0.12	0.45	0.65	0.76
cDNA	fantom*	M. musculus	BLASTN	0.002	0.45	0.18	0.41	0.64
Exon prediction	Exofish*	T. nigroviridis	ExoFish	0.005	0.11	0.11	0.27	0.47
Protein	flypep*	D.melanogaster	BLASTX	0.006	0.13	0.18	0.27	0.47
Protein	wormpep*	C. elegans	BLASTX	0.002	0.24	0.11	0.13	0.23

* Descriptions and references of each method are given in the legend of figure 4.8.

The test region (22q13.31) contained 3,365,293 bp of genomic sequence. The total number of nucleotides contained within the 17 annotated pseudogenes is 6090 bp. The total number of reference exons is 29. For more details, see chapter II.

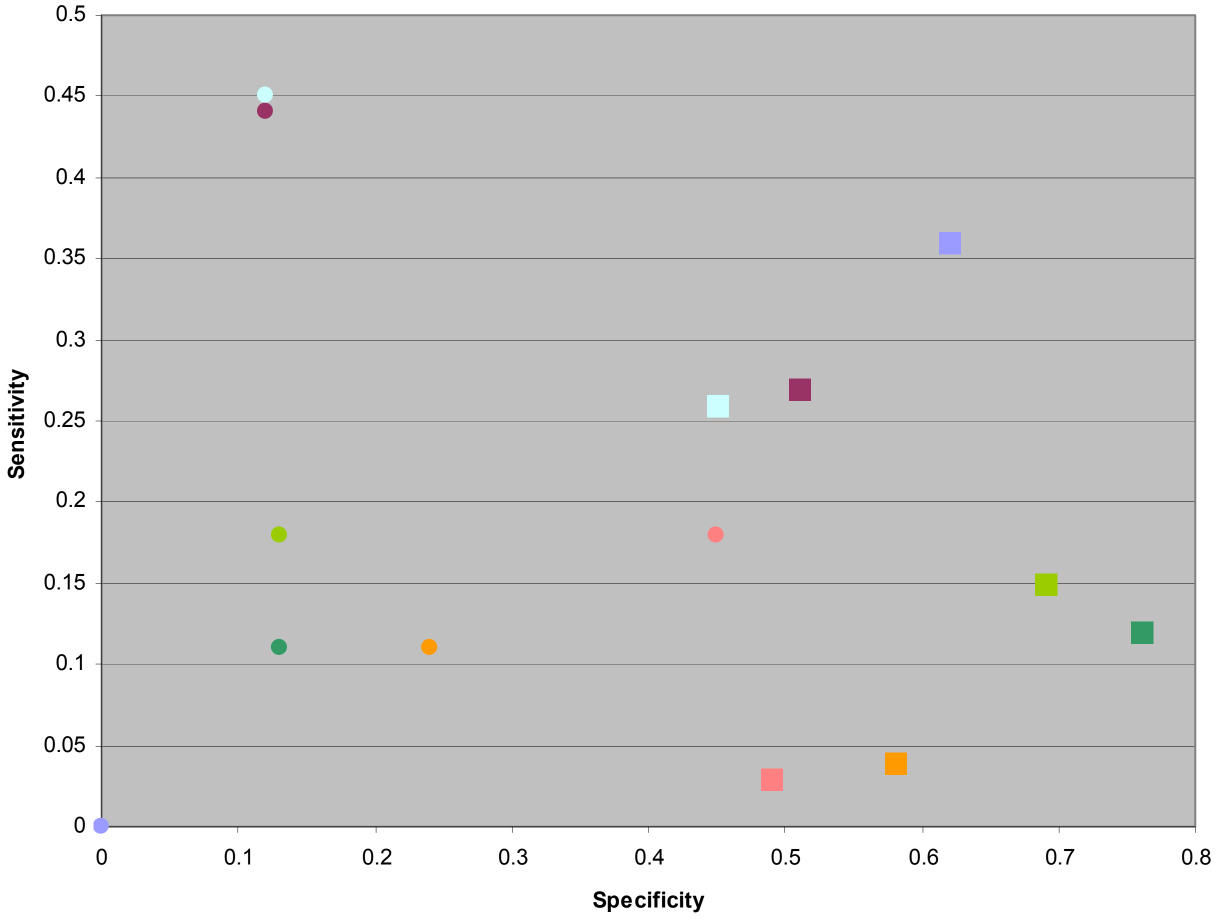


Figure 4.9: Specificity and sensitivity of different comparative sequence data with the 22q13.31 transcript map. Sensitivity and specificity shown are computed at the nucleotide level.

🟦 Blastn_mus
🟪 Blatmouse
🟩 Exomouse
🟥 fantom
🟩 Exofish
🟩 flypep
🟧 wormpep

□ = correlation with 39 annotated genes within 22q13.31

○ = correlation with 17 annotated pseudogenes within 22q13.31

Descriptions and references of the sequence evidence are given in the legend to figure 4.8.

Once again, the sensitivity and specificity of matches to annotated pseudogenes are, in general, lower than the correlation to annotated genes. In the case of Blastn_mus (mapped mouse genomic sequence derived from this project), no alignment to pseudogenes was noted. In comparison, BLAT and Exonerate alignments of the WGS mouse sequence demonstrated relatively high sensitivity of correlation to pseudogene structures: this is because the WGS sequence resource is not limited to the sequence from one particular region. These matches to human pseudogenes may be from sequence of the true mouse gene, orthologous to the true human gene from which the pseudogene is derived.

This analysis shows that the highest sensitivity of correlation with the annotated genes is currently demonstrated by the mapped mouse genomic sequence resulting from this project. However, as the large-scale murine genome project is completed, and gene identification in this and in other genomes advances, values of sensitivity and specificity will alter. The highest values of specificity here originate from the Exofish gene prediction program, followed by matches to DNA and protein sequence databases. These values are comparable to those derived from human cDNA collections (chapter III) and indicates that comparison to known, or predicted, genes in other species is a powerful tool for accurate gene annotation. However, this high level of specificity is, in general, linked with lower sensitivities than those shown in chapter III and may therefore enable identification of only a subset of genes present in the region of interest.

4.5 Investigation of intronic and intergenic conserved sequences

The results shown in table 4.7 indicate that there are areas where high similarity is observed outside of the annotated human genes. These regions may just be non-functional sequences that have not diverged or could indicate the presence of regulatory element. Some of these conserved features may also be unidentified human exons. This latter possibility was initially investigated through a comparison of the conserved human-mouse sequences and Genscan predicted exons.

4.5.1 Correlation of Genscan predictions with human-mouse conserved sequences

A correlation analysis of Genscan predictions with the gene annotation of 22q13.31 is described in chapter III. From this study, 384 (58%) of the 657 Genscan predicted exons are identified as 'wrong' i.e. do not overlap an annotated true coding exon. Eighteen of the 'wrong' predictions overlap annotated pseudogenes and are therefore discounted from this analysis.

The correlation of the remaining 366 Genscan predicted exons with the Blastn_mus, Blatmouse and Exomouse alignments were manually assessed by eye from the visual display of the 22ace database. Genscan predicted only six exons outside of the annotation, which contained sequence that aligned to mouse genomic DNA. The results of this analysis are shown in detail in table 4.8

Table 4.8: The position of exons predicted by Genscan, which do not overlap annotated true exons, but overlap aligned mouse genomic sequences

Genscan exon no.	Position on human transcript map	Correlates with Human-Mouse genomic alignment:		
		ExoMouse	Blatmouse	Blastn_mus
1	intergenic			•
2	within dJ345P10.C22.4	•	•	•
3	intergenic			•
4	within dJ474I12.C22.2	•	•	•
5	within ARHGAP8	•	•	•

4.5.2 Test of expression

The three intergenic Genscan predictions had previously tested negative for expression in seven cDNA libraries by PCR (see chapter III). In a similar experiment, primers were designed to the remaining three Genscan predictions, as well as to an additional twenty-five exon candidates identified from the Blastn_mus alignment, which were over 30 bp long and contained an ORF. Altogether, six exon candidate regions were not associated with any annotated gene structures, whilst 22, including those supported by Genscan predictions, lay within introns of annotated genes.

The twenty-eight primer pairs were used in PCR screens of seven cDNA vectorette libraries (see chapter II). Only one positive result was obtained from a candidate exon (not supported by a Genscan prediction) within the gene E46L. cDNA sequence from the resulting vectorette PCR product partially matched the existing exon structure, but appeared to result from spurious poly(dT) priming within a repeat. No new human exons or genes were therefore experimentally confirmed in this test.

4.6 Finished mouse sequence analysis

Two finished mouse clone sequences, AL583887.9 (bM121M7) (220050bp) and AL513354.14 (bM150J22) (22703bp) were selected for more detailed analysis. These clones map in close proximity to each other (see figure 4.5) but do not overlap, as a gap of ~60kb (estimated from fingerprint data) exists between them. This gap is spanned by clone bM85M21, which is currently being sequenced.

4.6.1 Mouse gene annotation

Initial annotation of the finished mouse clones was performed by Dr. Laurens Wilming (Sanger Institute) by similarity comparison to:

1. EMBL vertebrate cDNA sequences (see appendix 2)
2. Publicly available EST sequences (see appendix 2)

3. Human annotated gene sequences from 22q13.31.

This initial annotation was extended by similarity comparison to non-publicly available ESTs (appendix 2) and partial, but not submitted, cDNA sequences from 22q13.31 (chapter III) (M. Goward). The approach is similar to the human sequence analysis discussed in chapter III. In total, eight genes were annotated in the mouse clones. The longest isoforms of these genes are summarised in table 4.9. Figure 4.10 shows the genomic distribution of the mouse genes in comparison with the syntenic human region.

Table 4.9: The annotated mouse genes and their exon number, genomic span, transcript size and ORF size.

Mouse gene	Human orthologue	No. of exons	Genomic size (bp)	Transcript size (bp)	ORF size (bp)
bM121M7.1	TTLL1	12(12)	26956(49751)	2003(1684)	1272(1272)
Biklk	BIK	5(5)	17795(19110)	1370(1099)	453(483)
bM121M7.3	bK1191B2.C22.3	4(4)	15727(11180)	1679(2048)	1146(1173)
Bzrp	BZRP	4(4)	10623(11697)	849(850)	510(510)
bM121M7.5	dJ526I14.C22.2	14(14)	19502(20479)	3209(3353)	1920(1935)
Scube1*	dJ526I14.C22.3	>19(22)	>72041(139476)	>4914(5741[#])	2886[#](2967)
bM150J22.1	C22ORF1*	6(>4)	66530(>63349)	3180(2323[#])	981(909[#])
bM150J22.2*	dJ345P10.C22.4	>26(33)	>121975(283449)	>4032(4878)	>3965(4575)

*Gene structure extends beyond available genomic sequence
# Size calculated from EMBL cDNA entry
The equivalent values for the orthologous human genes are shown in brackets.

Figure 4.10 (foldout): Alignment of the human and mouse annotated genes. The figure depicts the human clones (blue boxes) with sequence accession numbers, the human and mouse CpG islands (yellow), the human gene features (genes with orthologues shown in the mouse sequence are shown in dark blue, genes for which equivalent mouse sequence is not yet available in light blue and pseudogenes in green), mouse genes (red) and mouse sequence clones (red boxes) with accession numbers. Similar exons are indicated by the grey lines.

take out this page for figure 4.10

Additionally, five alternative splice forms were annotated based on mouse EST evidence (L. Wilming). Three isoforms of bM121M7.3 have been annotated. Two of these are orthologous to alternative splices verified in human: bK1191B2.C22.3a (Em:AL359401) and bK1191B2.C22.3b (Em:AL359403). The remaining isoform of bM121M7.3 shows a possible alternative 5' end. Additionally, alternative 3' ends are indicated from EST evidence for bM121M7.5 and Scube1. However, there is currently no evidence to support the existence of these isoforms in the orthologous human genes. EST evidence can be unreliable (chapter III) so further experimental evidence is required to confirm these structures.

4.6.2 Human-mouse finished sequence alignment

4.6.2.1 Dot plot

The annotated mouse and human sequences were compared using the PipMaker dot plot program (http://bio.cse.psu.edu/pipmaker) (Schwartz *et al.*, 2000). Figure 4.11 shows the mouse sequence displayed on the x-axis and the human sequence on the y-axis. Drawn along both of the axes are boxes corresponding to each of the annotated genes. Regions of high similarity correspond with gene structures. Gene order and orientation are conserved. The human gene dJ754E20A.C22.4 lies within the mouse sequence gap. The genomic span of the human sequence is approximately 1.6X greater than the equivalent genomic mouse sequence (see sections 4.6.4 and 4.6.5). The mouse clone bM150J22 spans a gap in the human sequence. This is discussed in more detail in section 4.7.

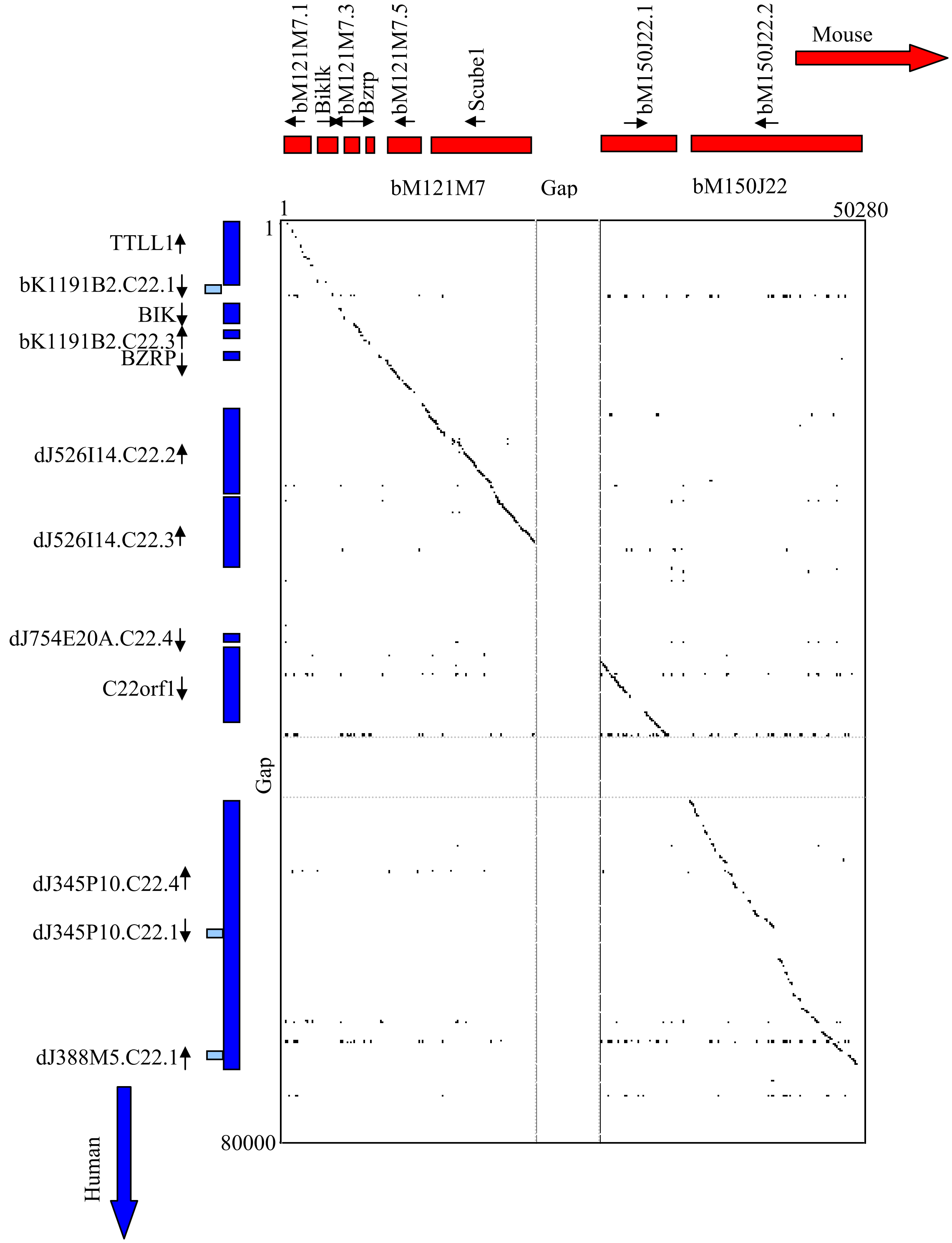


Figure 4.11: Annotated dot plot of the mouse (x-axis) and human (y-axis) sequences. The plot was generated using the PipMaker suite of analysis tools (Schwartz *et al.*, 2000). The boxes along the axes indicate the positions of human (blue) and mouse (red) genes. Light blue boxes depict possible human pseudogenes, which are not conserved in the mouse sequence.

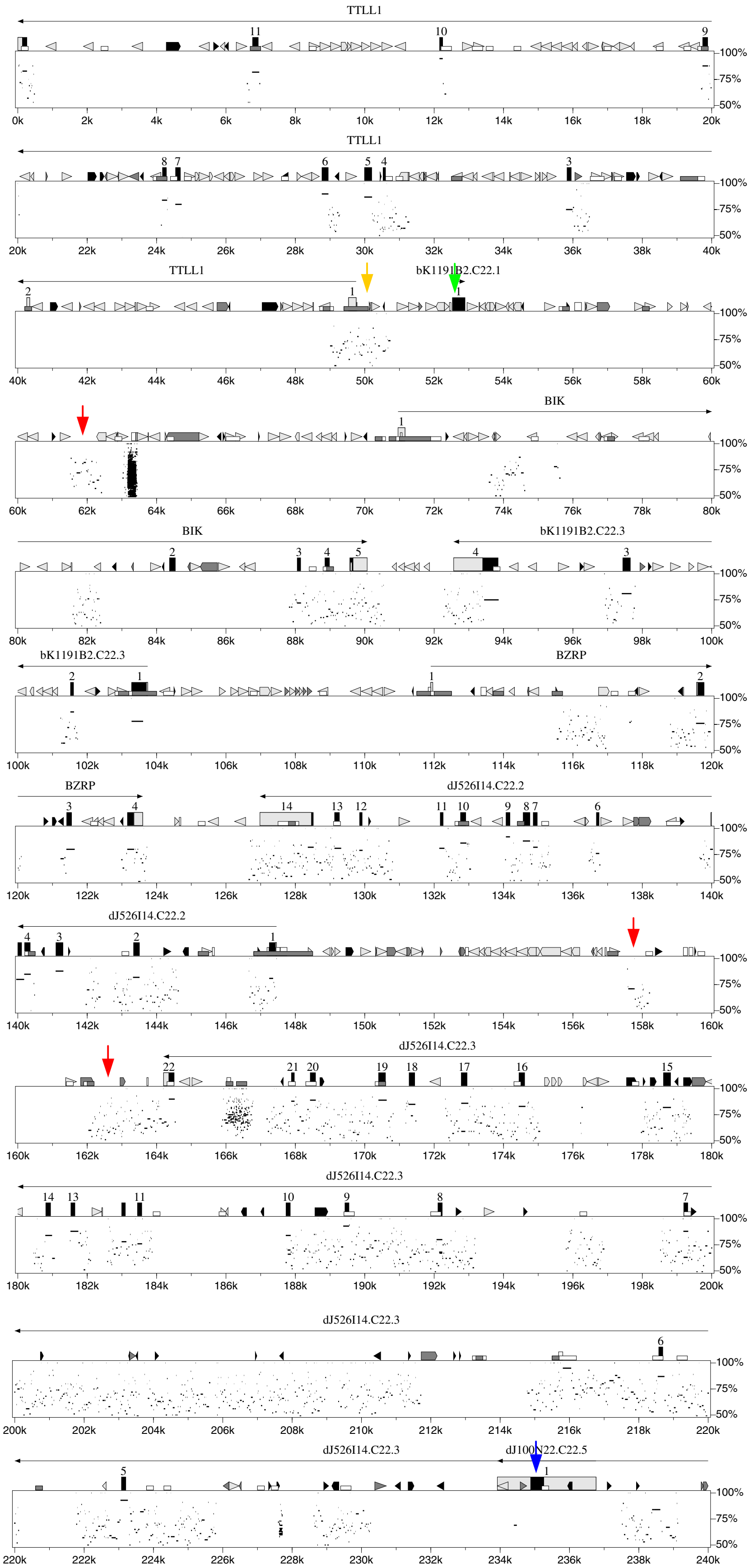
4.6.2.2 PIP analysis

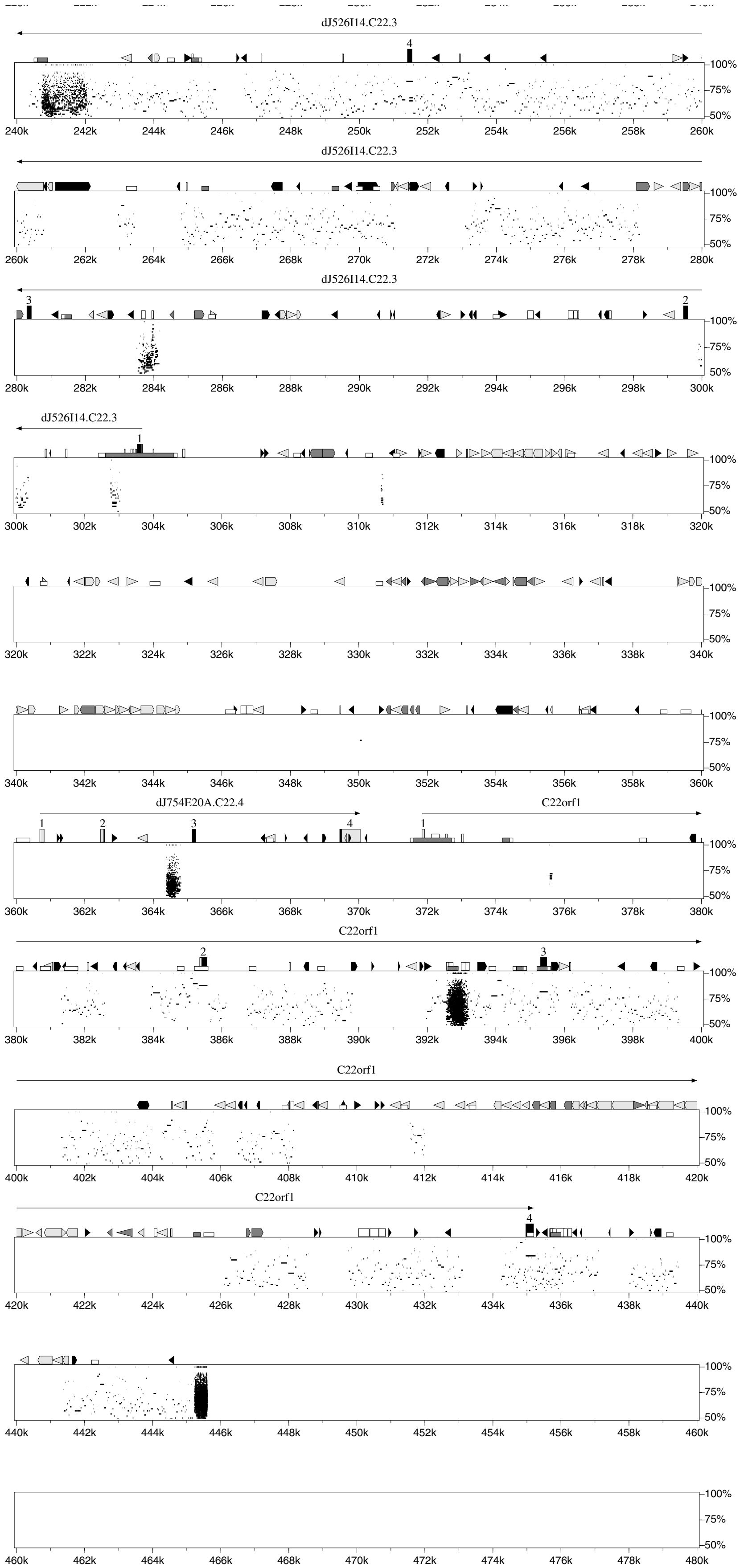
A PIP (Schwartz *et al.*, 2000) was generated to show the conservation of this region between finished human and mouse sequences in more detail. The plot displays the human sequence along the x-axis, incorporating features such as genes, repeats (generated from RepeatMasker output) etc. The y-axis displays the percent identity of the mouse sequence. Figure 4.12 shows that overall the areas of high similarity correspond well with the annotated human genes. There are a few exceptions:

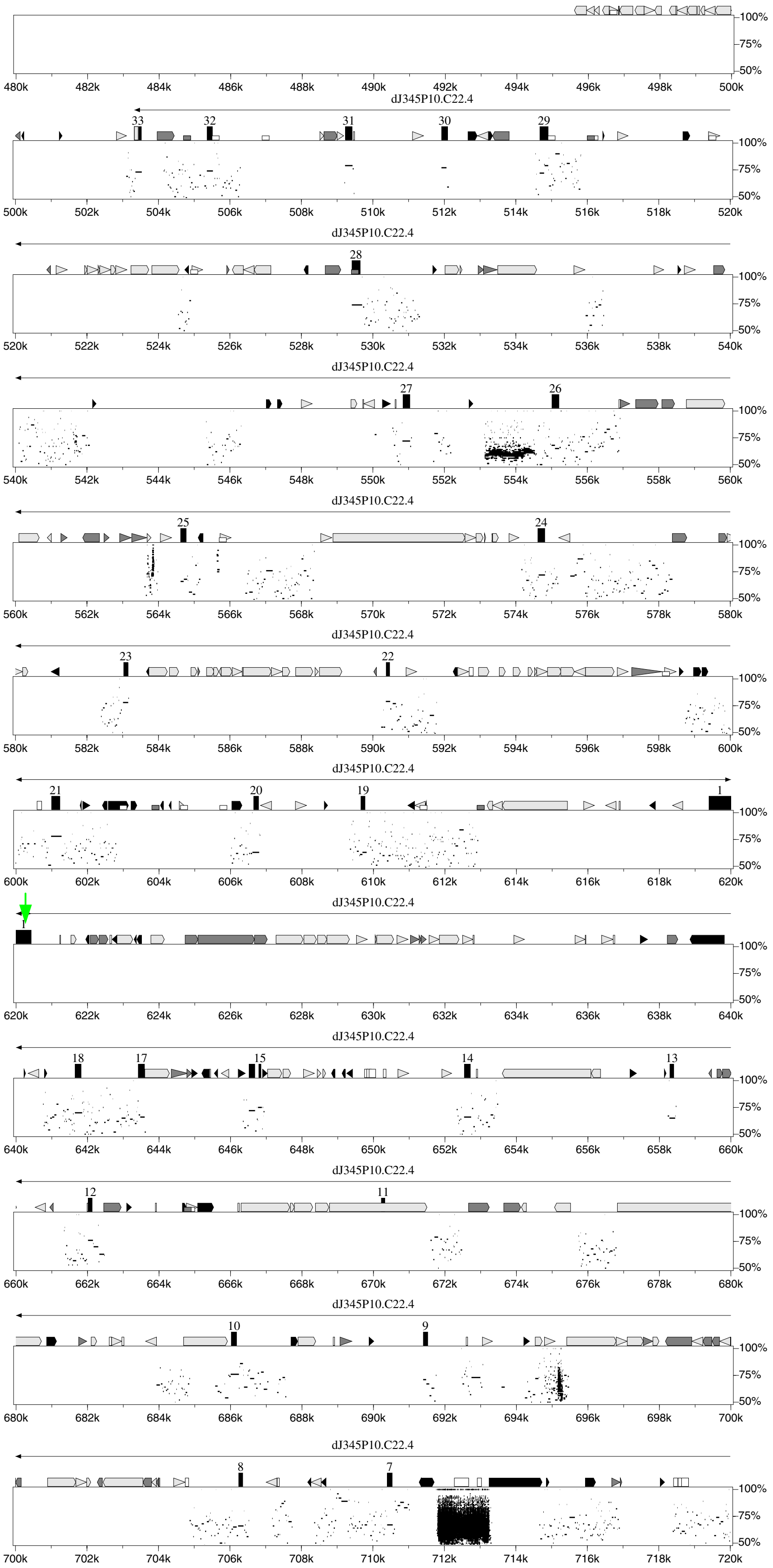
- Conserved sequences are located in an intergenic region around 62K (between TTLL1 and bK1191B2.C22.3) and between 157.5K and 164K (between dJ526I14.C22.2 and dJ526I14.C22.3) (indicated by red arrows).
- Conserved sequences are also found in the 5'UTR of TTLL1 (yellow arrow) and in the introns of most genes.

These sequences may highlight additional exons that have not been annotated in the human sequence, or may indicate the presence of regulatory regions.

- The cDNA sequence Em:AL442096 (Bloecker et al., unpublished), was previously noted as possibly resulting from spurious priming of an adjacent genomic poly(A) tract (chapter III). The sequence is not conserved in mouse (blue arrow), which supports the premise that this cDNA does not originate from a true gene.
- Similarly, the human pseudogenes bK1191B2.C22.1 and dJ345P10.C22.1 were not conserved in the mouse sequence (green arrows).


TTLL1
11
10
9
0k
2k
4k
6k
8k
10k
12k
14k
16k
18k
20k
100%
75%
50%
TTLL1
8 7
6
5 4
3
20k
22k
24k
26k
28k
30k
32k
34k
36k
38k
40k
TTLL1
2
1
bK1191B2.C22.1
1
40k
42k
44k
46k
48k
50k
52k
54k
56k
58k
60k
BIK
1
60k
62k
64k
66k
68k
70k
72k
74k
76k
78k
80k
BIK
2
3
4
5
bK1191B2.C22.3
4
3
80k
82k
84k
86k
88k
90k
92k
94k
96k
98k
100k
bK1191B2.C22.3
2
1
BZRP
1
2
100k
102k
104k
106k
108k
110k
112k
114k
116k
118k
120k
BZRP
3
4
dJ526I14.C22.2
14
13
12
11
10
9
8 7
6
120k
122k
124k
126k
128k
130k
132k
134k
136k
138k
140k
dJ526I14.C22.2
4
3
2
1
140k
142k
144k
146k
148k
150k
152k
154k
156k
158k
160k
dJ526I14.C22.3
22
21
20
19
18
17
16
15
160k
162k
164k
166k
168k
170k
172k
174k
176k
178k
180k
dJ526I14.C22.3
14
13
11
10
9
8
7
180k
182k
184k
186k
188k
190k
192k
194k
196k
198k
200k
dJ526I14.C22.3
6
200k
202k
204k
206k
208k
210k
212k
214k
216k
218k
220k
dJ526I14.C22.3
5
dJ100N22.C22.5
1
220k
222k
224k
226k
228k
230k
232k
234k
236k
238k
240k


dJ526I14.C22.3
4
100%
75%
50%
240k
242k
244k
246k
248k
250k
252k
254k
256k
258k
260k
dJ526I14.C22.3
262k
264k
266k
268k
270k
272k
274k
276k
278k
280k
dJ526I14.C22.3
3
2
282k
284k
286k
288k
290k
292k
294k
296k
298k
300k
dJ526I14.C22.3
1
302k
304k
306k
308k
310k
312k
314k
316k
318k
320k
322k
324k
326k
328k
330k
332k
334k
336k
338k
340k
342k
344k
346k
348k
350k
352k
354k
356k
358k
360k
dJ754E20A.C22.4
C22orf1
1
2
3
4
1
362k
364k
366k
368k
370k
372k
374k
376k
378k
380k
C22orf1
2
3
382k
384k
386k
388k
390k
392k
394k
396k
398k
400k
C22orf1
402k
404k
406k
408k
410k
412k
414k
416k
418k
420k
C22orf1
4
422k
424k
426k
428k
430k
432k
434k
436k
438k
440k
442k
444k
446k
448k
450k
452k
454k
456k
458k
460k
462k
464k
466k
468k
470k
472k
474k
476k
478k
480k


100%
75%
50%
480k
482k
484k
486k
488k
490k
492k
494k
496k
498k
500k
dJ345P10.C22.4
33
32
31
30
29
502k
504k
506k
508k
510k
512k
514k
516k
518k
520k
28
522k
524k
526k
528k
530k
532k
534k
536k
538k
540k
27
26
542k
544k
546k
548k
550k
552k
554k
556k
558k
560k
25
24
562k
564k
566k
568k
570k
572k
574k
576k
578k
580k
23
22
582k
584k
586k
588k
590k
592k
594k
596k
598k
600k
21
20
19
1
602k
604k
606k
608k
610k
612k
614k
616k
618k
620k
622k
624k
626k
628k
630k
632k
634k
636k
638k
640k
18
17
15
14
13
642k
644k
646k
648k
650k
652k
654k
656k
658k
660k
12
11
662k
664k
666k
668k
670k
672k
674k
676k
678k
680k
10
9
682k
684k
686k
688k
690k
692k
694k
696k
698k
700k
8
7
702k
704k
706k
708k
710k
712k
714k
716k
718k
720k

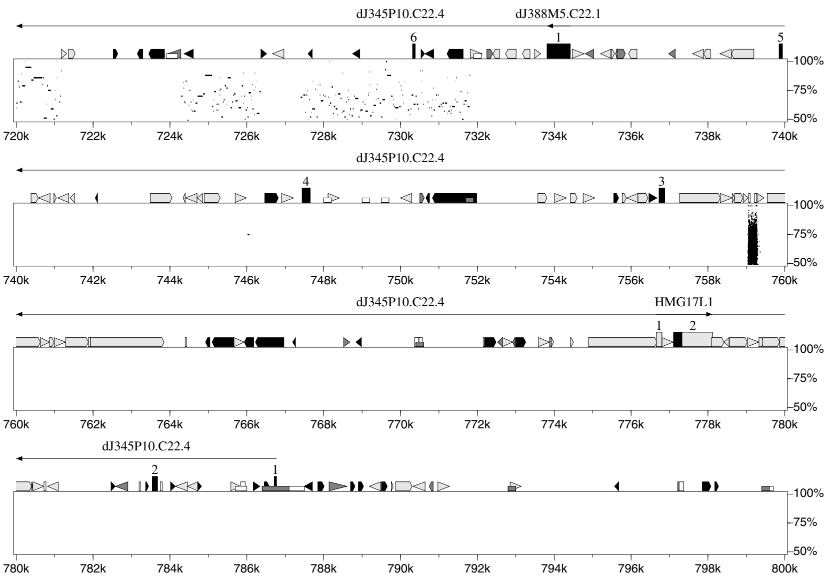


Figure 4.12: Percentage identity plot calculated by PipMaker for the human interval TTLL1 to dJ345P10.C22.4, compared with sequence from the region of conserved synteny on mouse chromosome 15. Black horizontal bars beneath the graphical depictions of interspersed repeats and gene structures indicate gap-free segments demonstrating> 50% nucleotide identities. Exons are numbered from the 5'-most annotated exon. A single gap-free alignment underneath a protein-coding exon indicates the mouse exon is conserved, and thus the mouse locus maintains a homologous ORF.

4.6.3 GC content

4.6.3.1 Comparison of human and mouse GC content

The fraction GC content in 1kb intervals was calculated by GC profile (Gillian Durham) and the GC content profiles plotted (Figure 4.13). The two GC profiles are similar, although direct comparison is complicated by the expansion of the human sequence to 1.6X the length of the equivalent mouse sequence. The 5' ends of genes align well with peaks in GC content. The human sequence has a higher overall GC content of 51% compared with the mouse sequence value of 49%.

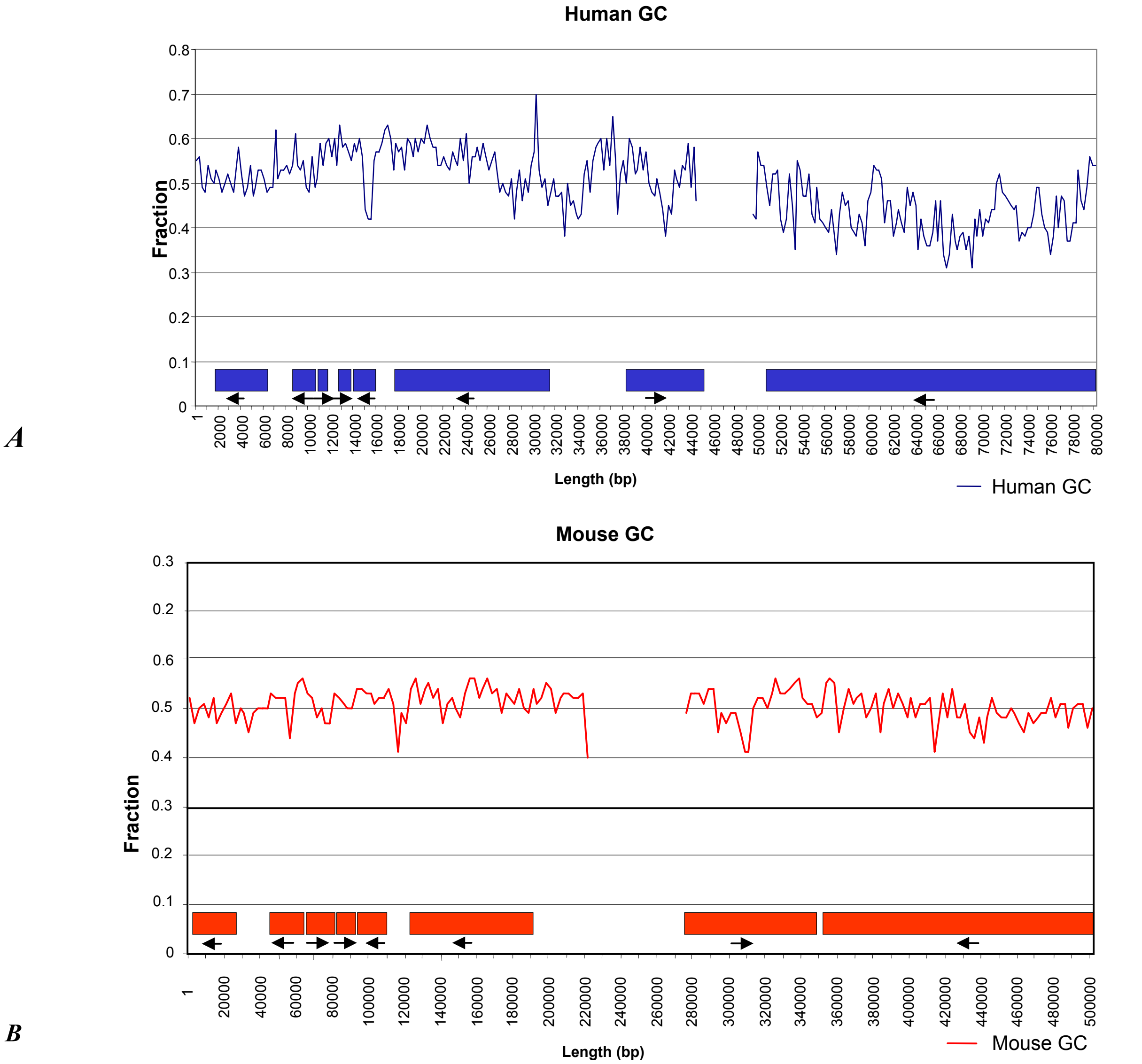


Figure 4.13: Human and mouse GC distribution, calculated using GC profile (G. Durham), with a window size of 1 kb. Human and mouse genes are depicted by blue and red boxes respectively, along the x-axes.

4.6.3.2 CpG islands

The 5' UTRs of six of the eight genes shown above are contained in the available finished mouse sequence. In human, all six genes contain a CpG island, but four of the mouse genes lack a CpG islands, using the criteria of the CpG island prediction package CPGFIND (Micklem, unpublished) (chapter III). An additional predicted CpG island does correspond to exon 2 of bM121M7.3 however. Antequera and Bird (1993) suggested that approximately 20% of mouse genes lack a CpG island. In this region, 66% of genes lack a CpG island at the 5'UTR, although the sample size is very small and figure 4.13 indicates that there are still peaks in the GC content associated with the starts of all genes. Details of the CpG islands are summarised in figure 4.14.

4.6.4 Repeat content

The repeat content of the human and mouse regions was analysed using RepeatMasker (Smit and Green, unpublished), with human- and rodent-specific repeats as appropriate. Figure 4.15 shows that the human and mouse SINE density are similar. The coverage of the SINEs in human, however, is four times that of mouse. This greater genomic coverage contributes to the difference in size noted between the equivalent regions of the human and mouse genomes: the human region is 1.6X larger than the mouse region. One third of this difference is caused by the greater coverage of the human SINE repeats. Simple sequence repeats and MaLRs are far more abundant in the mouse sequence. The MaLRs in mouse are still actively expanding, which is the most likely reason for the higher density of these repeats in mouse (Smit & Riggs, 1995).

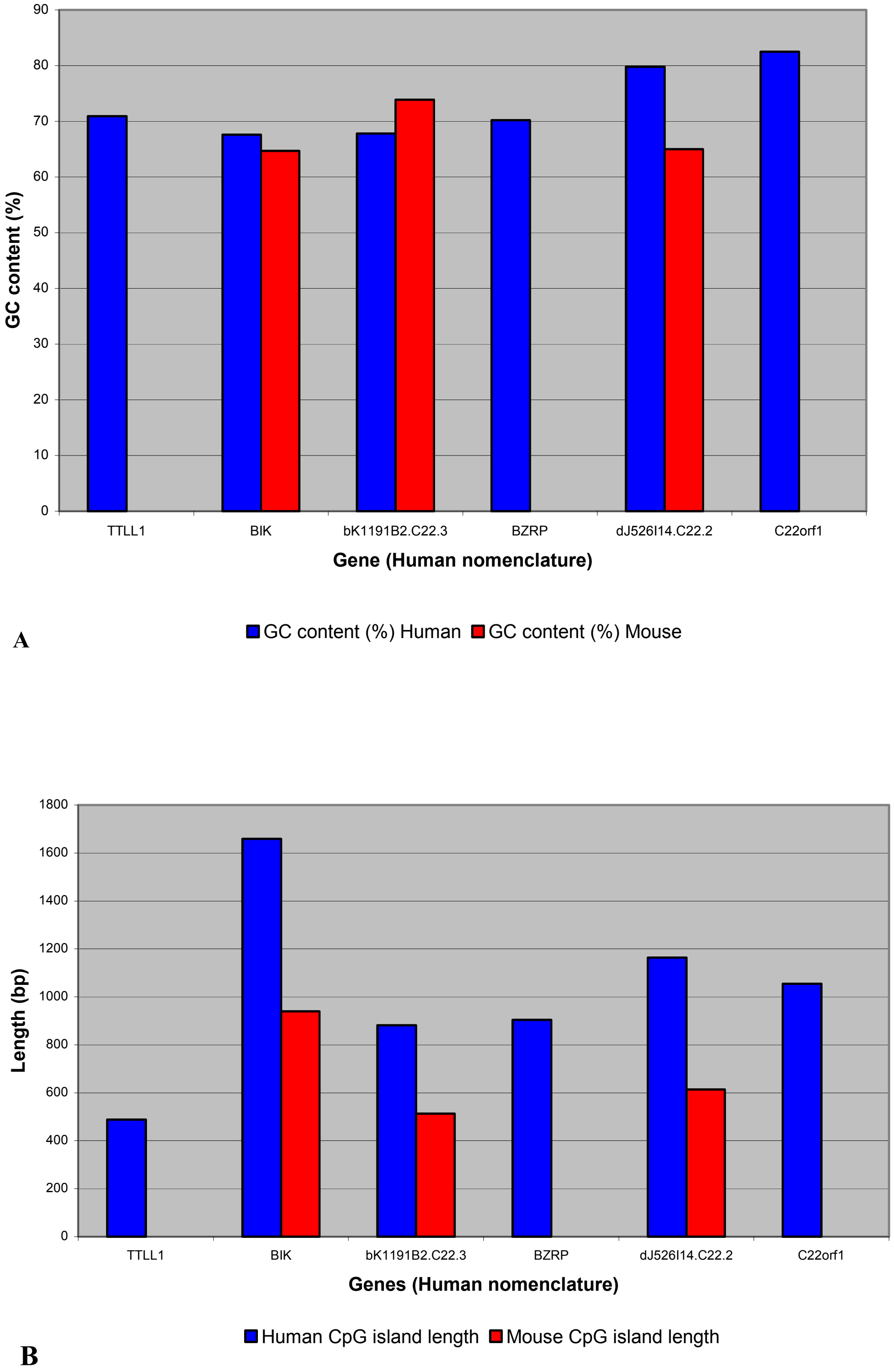


Figure 4.14: Comparison of human and mouse CpG island GC content (A) and length (B). CpG islands were predicted using CPGFIND (Micklem, unpublished).

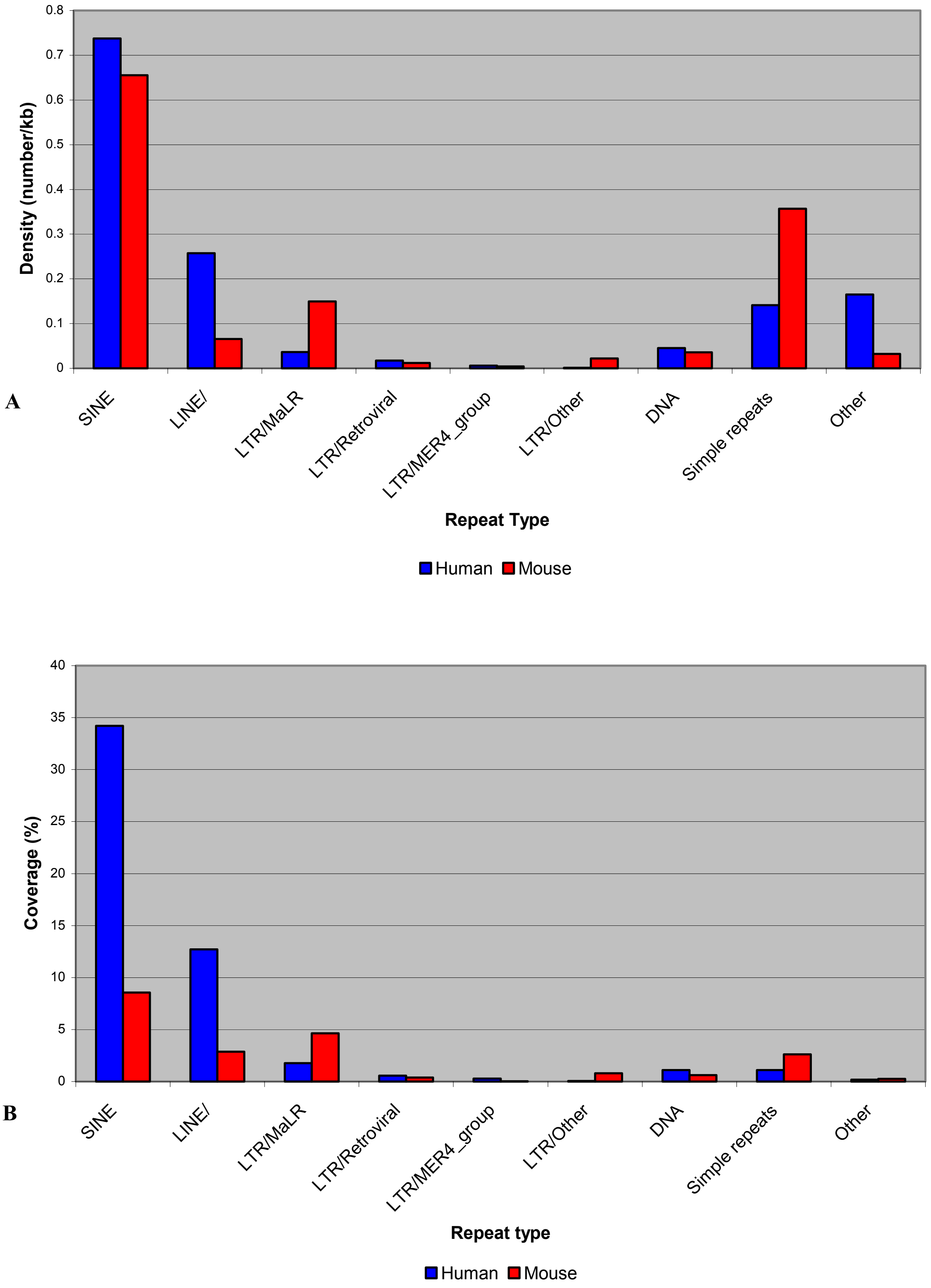
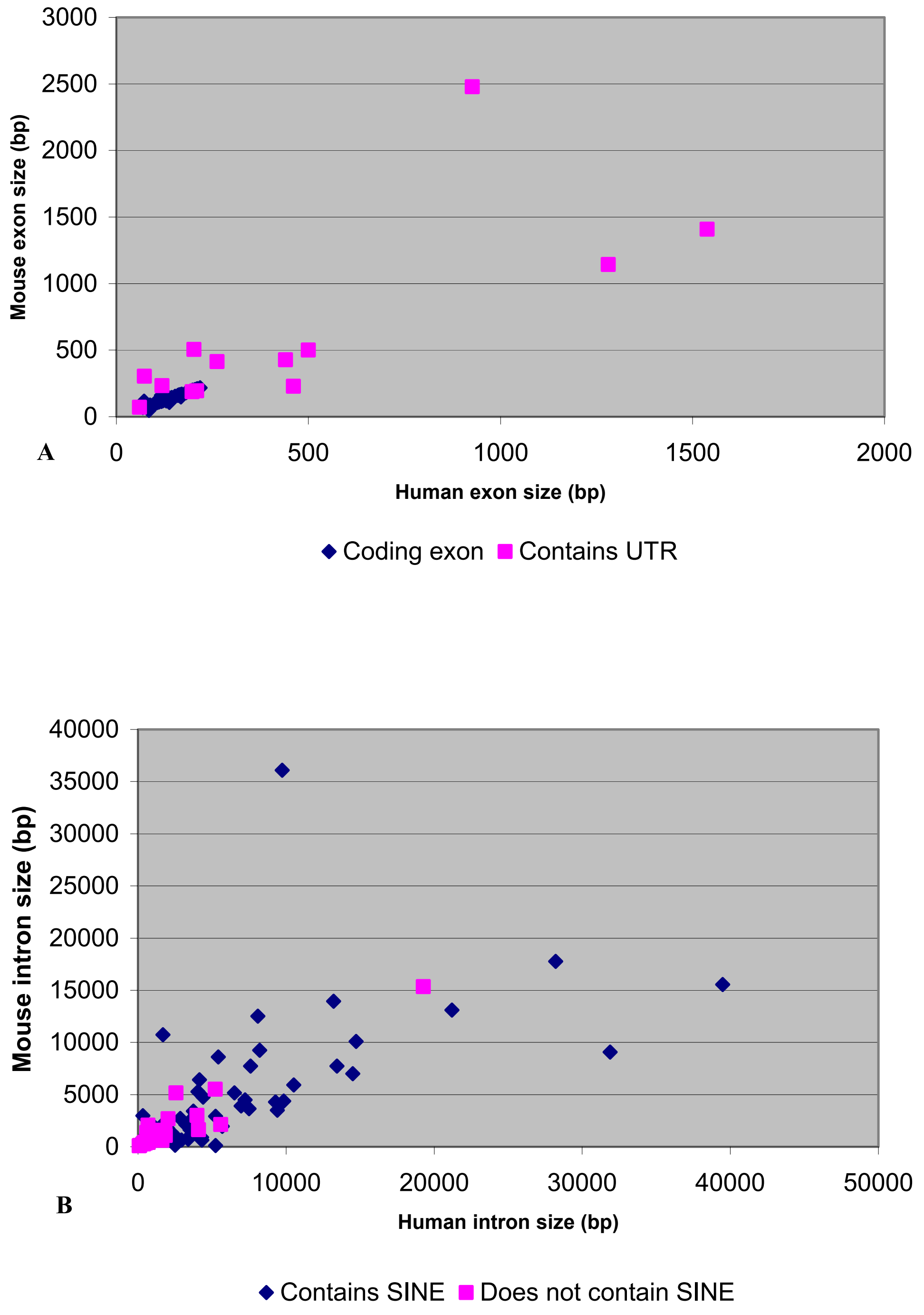


Figure 4.15: Repeat density (A) and genomic coverage by repeats (B) for human and mouse.

4.6.5 Comparison of coding regions

Exon number is conserved for all of the complete genes shown in table 4.9. The conservation of exon and intron sizes between mouse and human was examined by plotting the mouse exon sizes against the human (figure 4.16a); the equivalent comparison was carried out for intron size (figure 4.16n), and included analysis of the SINE content of the intron. A more detailed depiction of the 500 bp window of the human-mouse exon sizes is shown in figure 4.16c.

Generally, most of the internal coding exons are exactly the same length. The lengths of the 5' and 3' UTR exons, however, do show differences, as illustrated in table 4.9. The intron sizes are less well correlated (figure 4.16b). Introns containing SINEs generally tend to be larger in human genes, which contributes to the difference in sizes of the two equivalent regions (section 4.6.4). This is also reflected in figure 4.10 where the intron-exon structures are shown for all the genes. Together, this evidence reflects a high degree of conservation of the coding exons, with a lesser degree of conservation of gene structure.


Mouse exon size (bp)
3000
2500
2000
1500
1000
500
0
A
0
500
1000
1500
2000
Human exon size (bp)
◆ Coding exon
■ Contains UTR
Mouse intron size (bp)
40000
35000
30000
25000
20000
15000
10000
5000
0
0
10000
20000
30000
40000
50000
B
Human intron size (bp)
◆ Contains SINE
■ Does not contain SINE

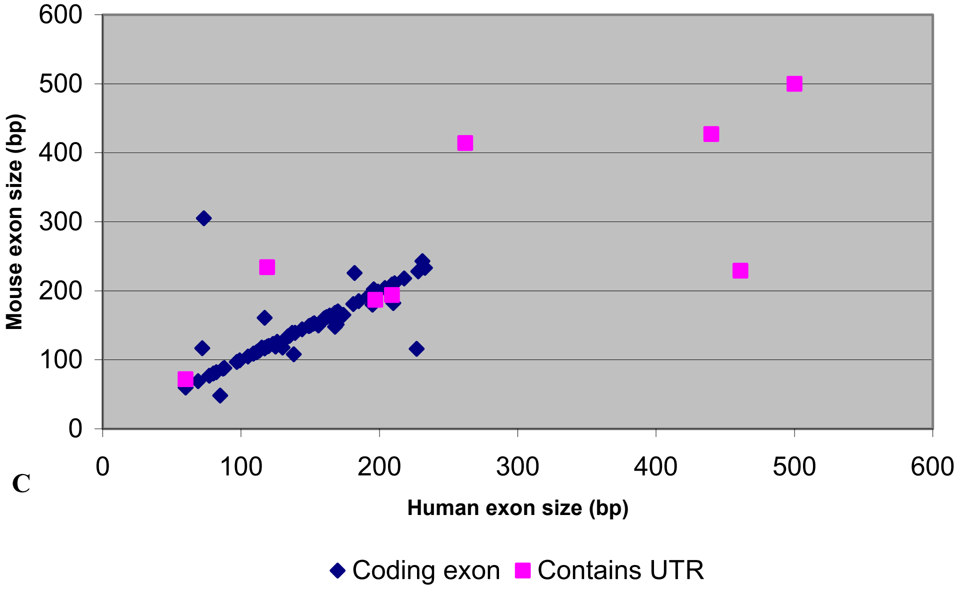


Figure 4.16: Scatter plots depicting (A) exon sizes and (B) intron sizes between human and mouse gene structures. (C) A more detailed view of the 500 bp exon interval is also shown.

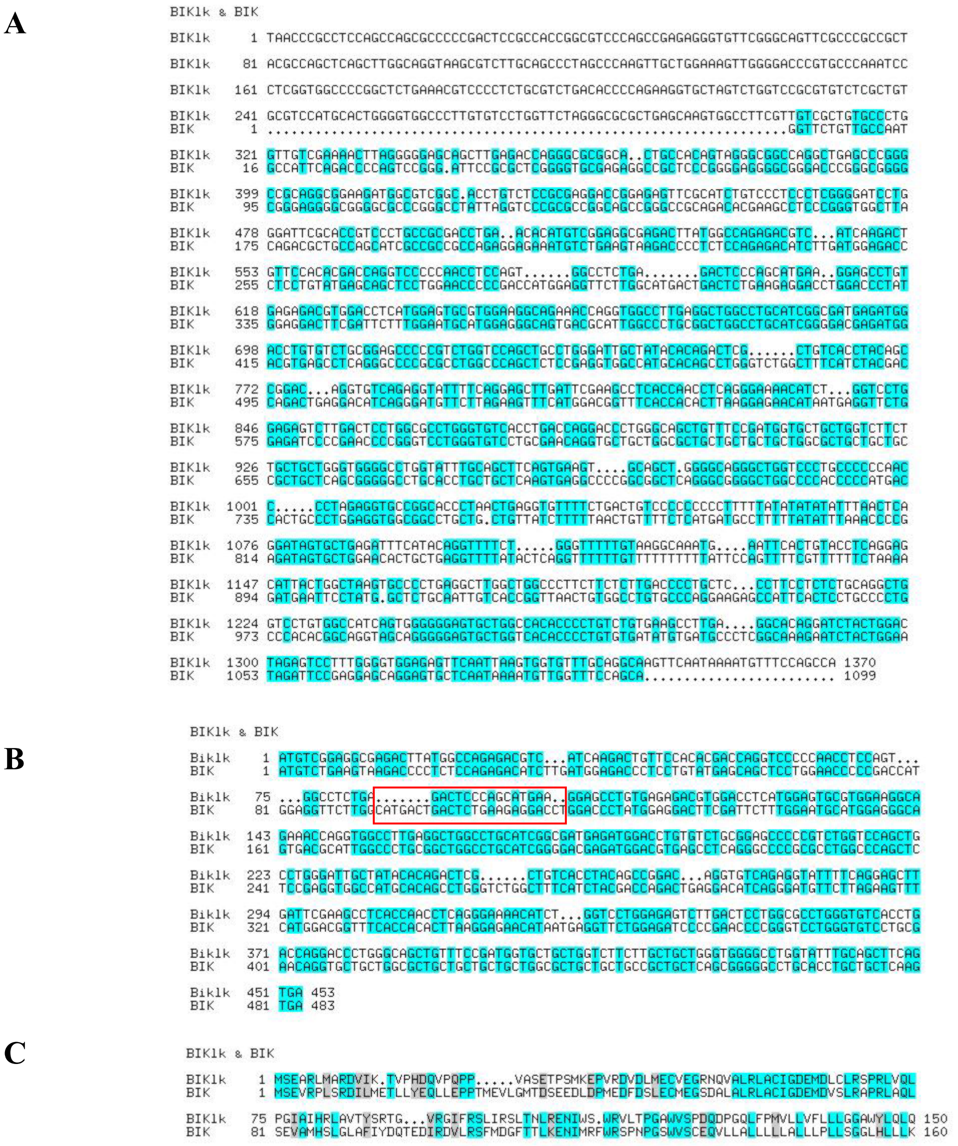
Nucleotide and amino acid sequence conservation was examined using clustalw (Thompson *et al.*, 1994) and sequence identities calculated (belvu; Sonnhammer, unpublished). These results are shown below.

Table 4.10: Percentage identities of mouse and human gene sequences

	mRNA nt. sequence	ORF nt. sequence	Amino acid sequence
Orthologous gene pair	identity (%)	identity (%)	identity (%)
bM121M7.1 & TTLL1	79.4	86.7	96.9
Biklk & BIK	57.6	64.0	41.3
bM121M7.3 & bK1191B2.C22.3	69.7	78.1	75.9
Bzrp & BZRP	75.5	81.8	81.1
bM121M7.5 & dJ526I14.C22.2	76.4	85.7	86.2
Scube1 and dJ526I14.C22.3*	81.7	87.8	87.1
bM150J22.1 & C22ORF1	70.8	90.2	98.2
bM150J22.2* & dJ345P10.C22.4*	72.4	72.6	78.0

*Gene currently incomplete; only partial sequences aligned

The percentage identity of all nucleotide sequences was increased by the exclusion of the 5’ and 3’ UTR sequences, which contain more divergent sequences. In four cases the level of conservation of the predicted amino acid sequence was lower than the equivalent nucleotide value. This was most marked between the human BIK gene and mouse Biklk (figure 4.17). This is due to a reading frame shift, caused by the insertion or deletion of a 7 bp sequence (highlighted in red). The conserved reading frame is restored by a 2 bp insertion/deletion downstream of the 7 bp difference. Five other in-frame insertions/deletions are also present. Altogether, these changes have the effect of lengthening the human protein, or shortening the mouse protein, by 10 amino acids. Additionally, there are 142 nucleotide changes (excluding deletions/insertions), of which only 28 are synonymous changes (do not alter the amino acid sequence). However, the number of amino acid changes that result from non-synonymous nucleotide changes is less than 114, as some changes occur in two different positions within the same codon. The existence of insertions/deletions in the sequence means that other, although perhaps less parsimonious, codon alignments exist in addition to the one shown below.



A

```
BIKlk & BIK

BIKlk     1 TAACCCGCCTCCAGCCAGCGCCCCCGACTCCGCCACCGGCGTCCCAGCCGAGAGGGTGTTCGGGCAGTTCGCCCGCCGCT

BIKlk    81 ACGCCAGCTCAGCTTGGCAGGTAAGCGTCTTGCAGCCCTAGCCCAAGTTGCTGGAAAGTTGGGGACCCGTGCCCAAATCC

BIKlk   161 CTCGGTGGCCCCGGCTCTGAAACGTCCCCTCTGCGTCTGACACCCCAGAAGGTGCTAGTCTGGTCCGCGTGTCTCGCTGT

BIKlk   241 GCGTCCATGCACTGGGGTGGCCCTTGTGTCCTGGTTCTAGGGCGCGCTGAGCAAGTGGCCTTCGTTGTCGCTGTGCCCTG
BIK       1 .................................................................GGTTCTGTTGCCAAT

BIKlk   321 GTTGTCGAAAACTTAGGGGGAGCAGCTTGAGACCAGGGCGCGGCA..CTGCCACAGTAGGGCGGCCAGGCTGAGCCCGGG
BIK      16 GCCATTCAGACCCCAGTCCGGG.ATTCCGCGCTCGGGGTGCGAGAGGCCGCTCCCGGGGAGGGGCGGGACCCGGGCGGGG

BIKlk   399 CCGCAGGCGGAAGATGGCGTCGGC.ACCTGTCTCCGCGAGGACCGGAGAGTTCGCATCTGTCCCTCCCTCGGGGATCCTG
BIK      95 CGGGAGGGGCGGGGCGCCCGGGCCTATTAGGTCCCGCGCCGGCAGCCGGGCCGCAGACACGAAGCCTCCCGGGTGGCTTA

BIKlk   478 GGATTCGCACCGTCCCTGCCGCGACCTGA..ACACATGTCGGAGGCGAGACTTATGGCCAGAGACGTC...ATCAAGACT
BIK     175 CAGACGCTGCCAGCATCGCCGCCGCCAGAGGAGAAATGTCTGAAGTAAGACCCCTCTCCAGAGACATCTTGATGGAGACC

BIKlk   553 GTTCCACACGACCAGGTCCCCCAACCTCCAGT......GGCCTCTGA.......GACTCCAGCATGAA..GGAGCCTGT
BIK     255 CTCCTGTATGAGCAGCTCCTGGAACCCCCGACCATGGAGGTTCTTGGCATGACTGACTCTGAAGAGGACCTGGACCCTAT

BIKlk   618 GAGAGACGTGGACCTCATGGAGTGCGTGGAAGGCAGAAACCAGGTGGCCTTGAGGCTGGCCTGCATCGGCGATGAGATGG
BIK     335 GGAGGACTTCGATTCTTTGGAATGCATGGAGGGCAGTGACGCATTGGCCCTGCGGCTGGCCTGCATCGGGGACGAGATGG

BIKlk   698 ACCTGTGTCTGCGGAGCCCCCGTCTGGTCCAGCTGCCTGGGATTGCTATACACAGACTCG......CTGTCACCTACAGC
BIK     415 ACGTGAGCCTCAGGGCCCCGCGCCTGGCCCAGCTCTCCGAGGTGGCCATGCACAGCCTGGGTCTGGCTTTCATCTACGAC

BIKlk   772 CGGAC...AGGTGTCAGAGGTATTTTCAGGAGCTTGATTCGAAGCCTCACCAACCTCAGGGAAAACATCT...GGTCCTG
BIK     495 CAGACTGAGGACATCAGGGATGTTCTTAGAAGTTTCATGGACGGTTTCACCACACTTAAGGAGAACATAATGAGGTTCTG

BIKlk   846 GAGAGTCTTGACTCCTGGCGCCTGGGTGTCACCTGACCAGGACCCTGGGCAGCTGTTTCCGATGGTGCTGCTGGTCTTCT
BIK     575 GAGATCCCCGAACCCCGGGTCCTGGGTGTCCTGCGAACAGGTGCTGCTGGCGCTGCTGCTGCTGCTGGCGCTGCTGCTGC

BIKlk   926 TGCTGCTGGGTGGGGCCTGGTATTTGCAGCTTCAGTGAAGT....GCAGCT.GGGGCAGGGCTGGTCCCTGCCCCCCAAC
BIK     655 CGCTGCTCAGCGGGGGCCTGCACCTGCTGCTCAAGTGAGGCCCCGGCGGCTCAGGGCGGGGCTGGCCCCACCCCCATGAC

BIKlk  1001 C.....CCTAGAGGTGCCGGCACCCTAACTGAGGTGTTTTCTGACTGTCCCCCCCCTTTTTATATATATATTTAACTCA
BIK     735 CACTGCCCTGGAGGTGGCGGCCTGCTG.CTGTTATCTTTTTAACTGTTTTCTCATGATGCCTTTTTATATTTAAACCCCG

BIKlk  1076 GGATAGTGCTGAGATTTCATACAGGTTTTCT.....GGGTTTTTGTAAGGCAAATG....AATTCACTGTACCTCAGGAG
BIK     814 AGATAGTGCTGGAACACTGCTGAGGTTTTATACTCAGGTTTTTTGTTTTTTTTTTATTCCAGTTTTCGTTTTTTCTAAAA

BIKlk  1147 CATTACTGGCTAAGTGCCCCTGAGGCTTGGCTGGCCCTTCTTCTCTTGACCCCTGCTC...CCTTCCTCTCTGCAGGCTG
BIK     894 GATGAATTCCTATG.GCTCTGCAATTGTCACCGGTTAACTGTGGCCTGTGCCCAGGAAGAGCCATTCACTCCTGCCCCTG

BIKlk  1224 GTCCTGTGGCCATCAGTGGGGGGAGTGCTGGCCACACCCCTGTCTGTGAAGCCTTGA....GGCACAGGATCTACTGGAC
BIK     973 CCCACACGGCAGGTAGCAGGGGGAGTGCTGGTCACACCCCTGTGTGATATGTGATGCCCTCGGCAAAGAATCTACTGGAA

BIKlk  1300 TAGAGTCCTTTGGGTGGAGAGTTCAATTAAGTGGTGTTTGCAGGCAAGTTCAATAAAATGTTTCCAGCCA 1370
BIK    1053 TAGATTCCGAGGAGCAGGAGTGCTCAATAAAATGTTGGTTTCCAGCA........................ 1099
```

B

```
BIKlk & BIK

Biklk     1 ATGTCGGAGGCGAGACTTATGGCCAGAGACGTC...ATCAAGACTGTTCCACACGACCAGGTCCCCCAACCTCCAGT...
BIK       1 ATGTCTGAAGTAAGACCCCTCTCCAGAGACATCTTGATGGAGACCCTCCTGTATGAGCAGCTCCTGGAACCCCCGACCAT

Biklk    75 ...GGCCTCTGA.......GACTCCAGCATGAA..GGAGCCTGTGAGAGACGTGGACCTCATGGAGTGCGTGGAAGGCA
BIK      81 GGAGGTTCTTGGCATGACTGACTCTGAAGAGGACCTGGACCCTATGGAGGACTTCGATTCTTTGGAATGCATGGAGGGCA

Biklk   143 GAAACCAGGTGGCCTTGAGGCTGGCCTGCATCGGCGATGAGATGGACCTGTGTCTGCGGAGCCCCCGTCTGGTCCAGCTG
BIK     161 GTGACGCATTGGCCCTGCGGCTGGCCTGCATCGGGGACGAGATGGACGTGAGCCTCAGGGCCCCGCGCCTGGCCCAGCTC

Biklk   223 CCTGGGATTGCTATACACAGACTCG......CTGTCACCTACAGCCGGAC...AGGTGTCAGAGGTATTTTCAGGAGCTT
BIK     241 TCCGAGGTGGCCATGCACAGCCTGGGTCTGGCTTTCATCTACGACCAGACTGAGGACATCAGGGATGTTCTTAGAAGTTT

Biklk   294 GATTCGAAGCCTCACCAACCTCAGGGAAAACATCT...GGTCCTGGAGAGTCTTGACTCCTGGCGCCTGGGTGTCACCTG
BIK     321 CATGGACGGTTTCACCACACTTAAGGAGAACATAATGAGGTTCTGGAGATCCCCGAACCCCGGGTCCTGGGTGTCCTGCG

Biklk   371 ACCAGGACCCTGGGCAGCTGTTTCCGATGGTGCTGCTGGTCTTCTTGCTGCTGGGTGGGGCCTGGTATTTGCAGCTTCAG
BIK     401 AACAGGTGCTGCTGGCGCTGCTGCTGCTGCTGGCGCTGCTGCTGCCGCTGCTCAGCGGGGGCCTGCACCTGCTGCTCAAG

Biklk   451 TGA 453
BIK     481 TGA 483
```

C

```
BIKlk & BIK

BIKlk     1 MSEARLMARDVIK.TVPHDQVPQPP.....VASETPSMKEPVRDVDLMECVEGRNQVALRLACIGDEMDLCLRSPRLVQL
BIK       1 MSEVRPLSRDILMETLLYEQLLEPPTMEVLGMTDSEEDLDPMEDFDSLECMEGSDALALRLACIGDEMDVSLRAPRLAQL

BIKlk    75 PGIAIHRLAVTYSRTG...VRGIFRSLIRSLTNLRENIWS.WRVLTPGAWVSPDQDPGQLFPMVLLVFLLLGGAWYLQLQ 150
BIK      81 SEVAMHSLGLAFIYDQTEDIRDVLRSFMDGFTTLKENIMRFWRSPNPGSWVSCEQVLLALLLLALLLPLLSGGLHLLLK 160
```

Figure 4.17: A) Alignment of Biklk and BIK including 5' and 3' UTRs. B) Greater conservation is shown in the alignment of the cDNA sequences without the UTRs. An insertion/deletion of 7bp causes a frameshift, which is corrected downstream by a further 2bp insertion/deletion (red box). C) Alignment of Biklk and BIK peptide sequences. Alignments were created with clustalw (Thompson *et al.*, 1994). The alignments were formatted for printing using belvu (Sonnhammer, unpublished).

4.6.6 Splice site comparison

The splice sites of both the human and mouse genes were compared using the sequence logo technique described in chapter III. Eighty splice acceptor and donor sequences from equivalent introns were extracted from gff files and used to generate sequence logos (D. Beare). The cumulative height of each position reflects the importance of this position in the splice consensus sequence. The height of each nucleotide reflects the frequency of that nucleotide at that particular position. Figure 4.18 shows the human splice donor and acceptor (A),and mouse splice donor and acceptor (B). This shows that, overall, the splice consensus is well conserved between human and mouse. The important GT nucleotides (positions 7 and 8) in the splice donor and AT (24 and 25) in the acceptor are well conserved between human and mouse. Differences are limited to the C/T tail where a C is more commonly found at position 14 in mouse whereas T is commonly found in human. These results support those of a previous study of 84 human and mouse introns (Smink, 2001).

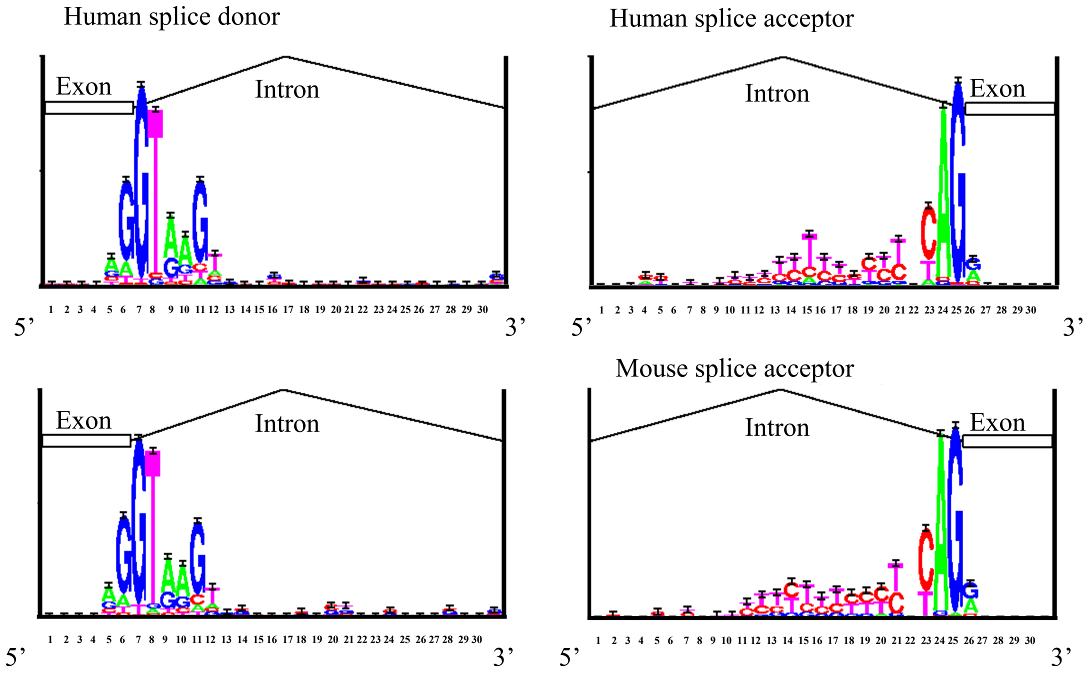


Figure 4.18: The splice acceptor and donor sites for human (A) and mouse (B). The splice site sequences were extracted by D. Beare (Sanger Institute) and visualised using Sequence Logo (Steven Brenner) (http://www.bio.cam.ac.uk/cgi-bin/seqlogo/logo.cgi).

4.6.7 Regulatory regions

Sequence conservation between human and mouse DNA in inter- and intragenic regions may indicate the existence of functional features, such as exons or regulatory regions, or may be non-functional sequence inherited from a common ancestor. CpG islands are associated with the promoter of ~50% of all mammalian genes (Antequera & Bird, 1993; Larsen *et al.*, 1992) and often contain multiple binding sites for transcription factors (Somma *et al.*, 1991). General conservation of the GC profile and peaks seems to suggest conservation of possible CpG islands (see section 4.6.3). The PIP (figure4.12), however, demonstrated conservation upstream of only one gene, TTLL1.

DBA (DNA Block Aligner) (Jareborg *et al.*, 1999) is an alignment algorithm designed to identify conserved collinear blocks in two DNA sequences. The main difference between DBA and PIP alignments is that DBA identifies gapped blocks. Also, blocks identified by DBA can be shorter than 50 bp, although the nucleotide identity must be greater than 60%, whereas PIPs will highlight only ungapped alignments longer than 50 bp with an identity >50% (section 4.1.3.1). Jarebourg *et al*. propose that these features of DBA make the program particularly suitable to identify small conserved functional motifs whose relative positioning may not be conserved and which may be separated by large pieces of non-functional DNA sequence due to random insertions in one species compared with another.

To investigate whether any further sequence conservation could be observed in these putative regulatory regions, three kilobases of sequence was extracted upstream of the transcription start site for both human and mouse, containing the entire length of any CpG islands predicted at this position. The human and mouse sequences were aligned with DBA. DBA identified significant alignments 5’ of the transcription start sites of the genes TTLL1, BIK, BZRP and C22orf1 (see appendix 6). An example of a region aligned by DBA is shown in figure 4.19.

The consensus sequences were used to scan the TRANSFAC 4.0 transcription factor database (Wingender *et al.*, 2000), using MatInspector V2.2 (Quandt *et al.*, 1995). Thresholds were set so that only exact matches to the core sequence of the matrix (capitalised) and overall matrix similarity >0.9 were listed, in order to enhance accuracy of the search results. The sites found are shown in table 4.11

```
bM121M7.1   -582     CCGCCTGCTTCTGCCTCCCAAAGTGCTGGGATTAAAGGCATGCGCCACC
Consensus      D     CC CC GC TCTGCCTCCC AAGTGCTGGGATTA AGGC TG GCCACC
TTLL1      -1559     CCACCCGCCTCTGCCTCCC-AAGTGCTGGGATTACAGGCGTGAGCCACC
```

Figure 4.19: Sequence alignment (DBA, Jareborg *et al.*, 1999) of mouse and human sequence upstream of TTLL1 (human gene) and bM121M7.1 (mouse orthologue). A potential binding site for the zinc finger protein Ik-2 is highlighted in red (Molnar & Georgopoulos, 1994)(see table 4.11).

The expression patterns of the human genes (chapter III) were examined in order to determine if there was a relationship between tissue distribution of the human transcript and what is currently known about the putative functional regions listed in table 4.11. TTLL1, BIK and BZRP are expressed in a wide variety of tissues. Examination of the TRANSFAC sites preceding these genes did not preclude this expression pattern. C22orf1 demonstrated a more limited expression pattern in RT-PCR screens of RNA from human tissues and previous research has shown that C22orf1 is predominantly expressed in adult brain (Schwartz & Ota, 1997). However, examination of the 24 sites found did not suggest specific involvement with adult brain transcription.

Table 4.11: Resulting sites from TRANSFAC screen with consensus sequences from DBA alignment of putative promoter regions.

Gene (human nomenclature)	Matrix	Orientation	Matrix similarity	Sequence
TTLL1	GFI1_01	-	0.905	angcctntAATCccagcacttngg
TTLL1	IK2_01	-	0.911	cttnGGGAggca
TTLL1	IK2_01	+	0.946	tgctGGGAttan
TTLL1	LYF1_01	-	0.911	ttnGGGAgg
TTLL1	RFX1_01	-	0.922	nggngncctnGCAAccn
BIK	IK2_01	+	0.928	cttnGGGAtntt
BZRP	DELTAEF1_01	-	0.954	ncacACCTnta
BZRP	GFI1_01	-	0.911	acacctntAATCccagcacttngn
BZRP	HFH2_01	+	0.911	nttTGTTtnntt
BZRP	HNF3B_01	+	0.908	ttnttTGTTtnnttn
BZRP	IK2_01	+	0.946	tgctGGGAttan
BZRP	SRY_02	-	0.931	nnaaACAAanaa
C22orf1	AP4_Q5	-	0.94	ctCAGCagtt
C22orf1	BRN2_01	+	0.923	aagatttgTAATgagt
C22orf1	BRN2_01	-	0.93	ctcattacAAATcttt
C22orf1	CREL_01	-	0.98	gggnntTTCC
C22orf1	DELTAEF1_01	+	0.953	cnccACCTgcn
C22orf1	E47_01	-	0.933	nnnGCAGgtggngac
C22orf1	FREAC2_01	-	0.912	attttgTAAAcaggnn
C22orf1	GFI1_01	-	0.902	tcattacaAATCtttccanctcag
C22orf1	GKLF_01	-	0.93	aaagagggagAGGG
C22orf1	GKLF_01	-	0.927	aanggagggaGGGG
C22orf1	IK2_01	-	0.917	nntgGGGAacag
C22orf1	LMO2COM_01	-	0.969	nngCAGGtggng
C22orf1	MYOD_01	-	0.926	nngCAGGtggng
C22orf1	MYOD_Q6	+	0.947	ncCACCtgcn
C22orf1	MZF1_01	-	0.975	nntGGGGa
C22orf1	MZF1_01	-	0.982	ggaGGGGa
C22orf1	NFAT_Q6	+	0.944	agntgGAAAgat
C22orf1	NFKAPPAB65_01	-	0.958	gggnntTTCC
C22orf1	NKX25_02	+	0.951	caTAATta
C22orf1	S8_01	+	0.968	ngcacataATTAaaat
C22orf1	S8_01	-	0.968	acattttaATTAtgtg
C22orf1	S8_01	-	0.934	ngacaaaaATTAgaga
C22orf1	S8_01	-	0.948	naaaacaaATTAgatt
C22orf1	SRY_02	-	0.925	naaaACAAatta

Core sequences are capitalised

4.7 Chromosome 22 sequence gap

Figure 4.11 shows that the mouse BAC bM150J22 spans one of the few remaining ‘unclonable’ gaps in the human genomic sequence of chromosome 22. This gap has been estimated to be approximately 50 kb long by fibre-FISH (Dunham *et al.*, 1999) and is known to contain the 3’ end of the C22orf1 gene at the centromeric end. The telomeric end of the gap is adjacent to the gene dJ345P10.C22.4. The mouse sequence spanning the gap is approximately 34 kb long. The sequence was analysed in more detail in order to identify any possible reasons why the region may be unclonable in human. To obtain equal start- and end-points for this comparison, sequence from bM150J22.1 to the 3’ exons of bM150J22.2 was analysed. These features are equivalent to the closest gene features annotated in the human genome sequence flanking the gap. The mouse ‘gap’ region, shown in figure 4.20, contains the 3’ end of the murine C22orf1 gene and provides evidence that the full human gene may be arranged in six exons. No further mouse EST or cDNA evidence was found to map to this region.

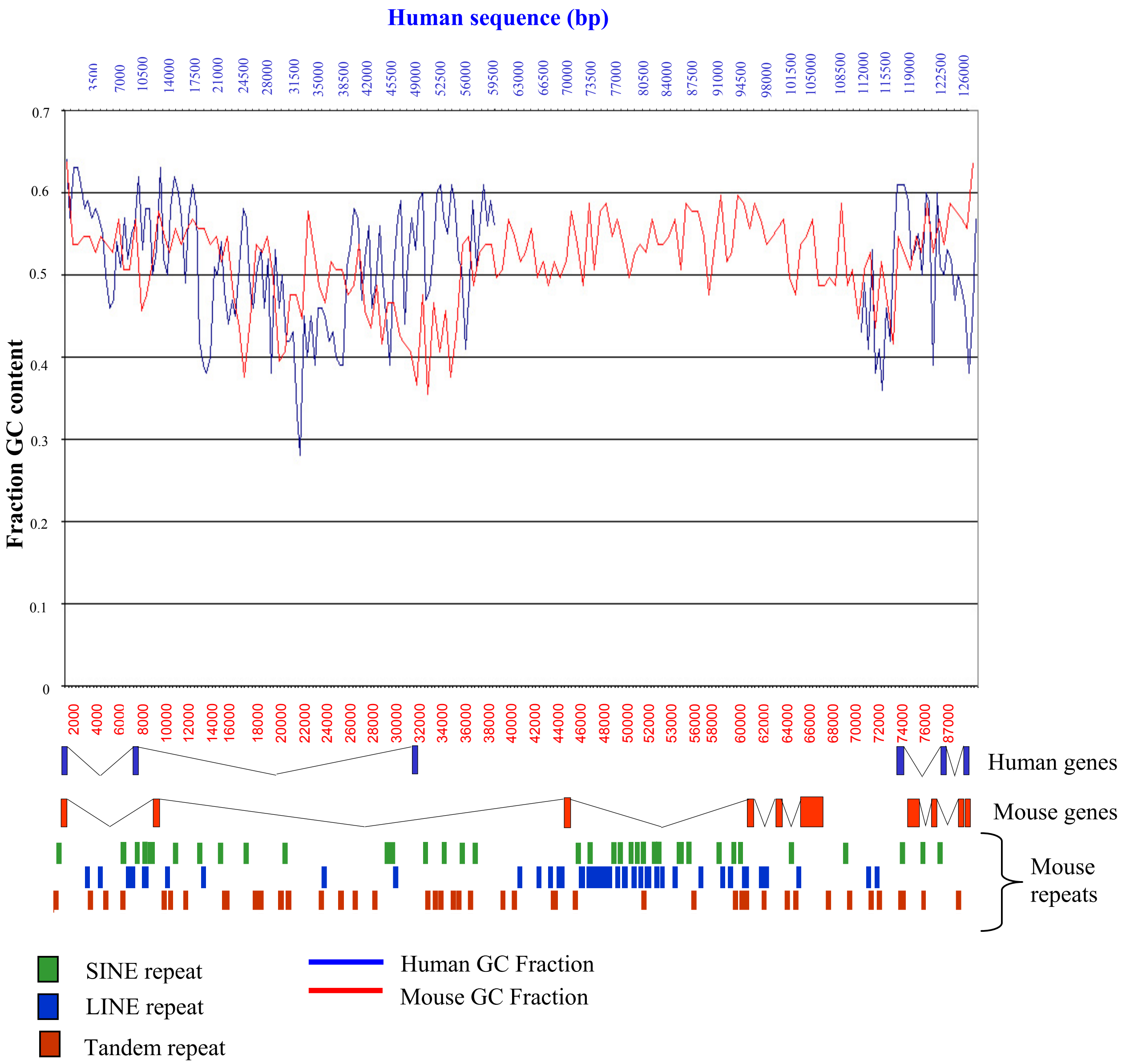


Figure 4.20: Diagram showing GC content, gene content and repeat content (mouse sequence only) of sequence spanning an 'unclonable' sequence gap in human chromosome 22. Human GC content and genes are shown in blue and mouse GC content and genes in red. GC fraction was calculated for 1kb windows using gc profile (Gillian Durham, unpublished). The distribution of mouse SINE, LINE and tandem repeats are also shown.

The graph of mouse GC content shows that a high proportion of GC dinucleotides are found throughout the region spanning the human sequence gap. The overall human GC content of the region of interest is higher than that of mouse (section 4.6.3.1). Extrapolation of the graph indicates that human GC content is maintained above a level of 50% throughout the gap region. This high GC distribution may have an adverse affect on the 'clonability' of this DNA segment (section 4.9).

The repeat content of the 30216bp of mouse sequence that spans the human sequence gap was analysed in more detail using RepeatMasker. Results are shown below.

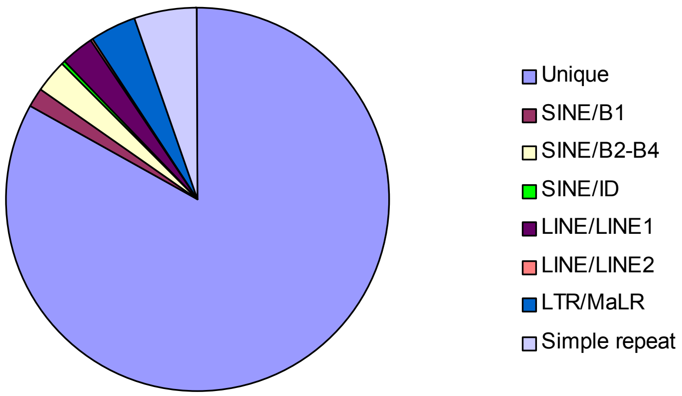


Figure 4.21: Repetitive and non-repetitive DNA distribution of 30216bp of mouse sequence, spanning an equivalent 'unclonable' sequence gap in human chromosome 22.

This region of mouse sequence contains no LTR elements or DNA transposon repeats. Although figure 4.21 shows that this region contains a greater coverage of SINE and LINE repeats than the immediately flanking sequences, the coverage and density of these repeats is comparable to the analysis of 50.2 kb of finished mouse sequence shown in section 4.6.4. No specific repetitive features were identified that could result in instability of this chromosomal region, leading to the difficulties in cloning the equivalent human DNA.

4.8 Localisation of synteny breakpoint

4.8.1 Definition of the junction region

A synteny breakpoint between HSA 22q13.1 and mouse chromosomes 15 and 8 was previously identified by Dunham *et al*. (1999), by combining data from the genomic sequence of HSA22 with information from the Mouse Genome Database (MGD) (http://www.informatics.jax.org/). The genes, HMOX and MB, situated 160 kb apart on HSA22, and their murine orthologues Hmox1 on MMU8 and Mb on MMU15, were identified as flanking the syntenic breakpoint.

In order to further narrow the breakpoint region boundaries, two mouse BAC contigs were constructed across the syntenic regions of mouse chromosomes 8 and 15 (section 4.2). Figure 4.4 shows that marker data from the two contigs localised the synteny breakpoint to a 130 kb region in the human sequence between genes MCM5 and MB. The available sequence from the contig tiling paths was compared with corresponding finished sequence from HSA22 using dot and PIP plots. Mouse BACs were identified that contained both conserved regions and sequence that extended beyond the syntenic breakpoint.

Currently, only unfinished sequence is available from the majority of adjacent mouse clones (see table 4.12) but detailed analysis is still possible.

Table 4.12: Mouse BAC genomic sequence clones adjacent to and spanning the syntenic breakpoint with human chromosome 22q13.1

		Sequencing Centre	Genomic	
Clone name	Author		location	Accession number
bM290L7	Grills *et al.*	AECOM*	MMU8	AC084823.10 (finished)
bM254F2	Sims	Sanger Institute	MMU8	AL603837.2 (unfinished)
bM267J18	Deschamps et al.	UOKNOR#	MMU8	AC076974.23 (unfinished)
bM422F22	Sims	Sanger Institute	MMU15	AL591892.2 (unfinished)
bM412D17	Sims	Sanger Institute	MMU15	AL603843.2 (unfinished)

* AECOM – Albert Einstein College of Medicine. # UOKNOW – University of Oklahoma

A dot plot comparison of these mouse sequences with the finished sequence of the orthologous region of human chromosome 22 is shown below (figure 4.22). The syntenic breakpoint junction is clearly delineated between genes dJ569D19.C22.1 and MB. Gene order and orientation also appear to be conserved. Intergenic sequences are generally divergent, although strong conservation is noted in the genomic sequence 5' to the RBM9 gene, which may denote conserved regulatory regions or a novel gene structure.

The genes APOL5 and APOL6, however, do not appear to be conserved in this dot plot alignment. The nucleotide and protein sequences of these human genes were therefore compared against the available mapped mouse sequence (http://mouse.ensembl.org) using BLAST. The best matches for the protein sequences were found to be within Em:AL603843 (23% and 27% sequence identity respectively), but no matches were found at the nucleotide level, which may explain their absence in the dot plot. Analysis of the finished sequence, when available, may allow annotation of these genes within the mouse sequence. Alternatively, these genes may not exist in mouse, perhaps having arisen from duplication events in the human genome after divergence from the mouse lineage.

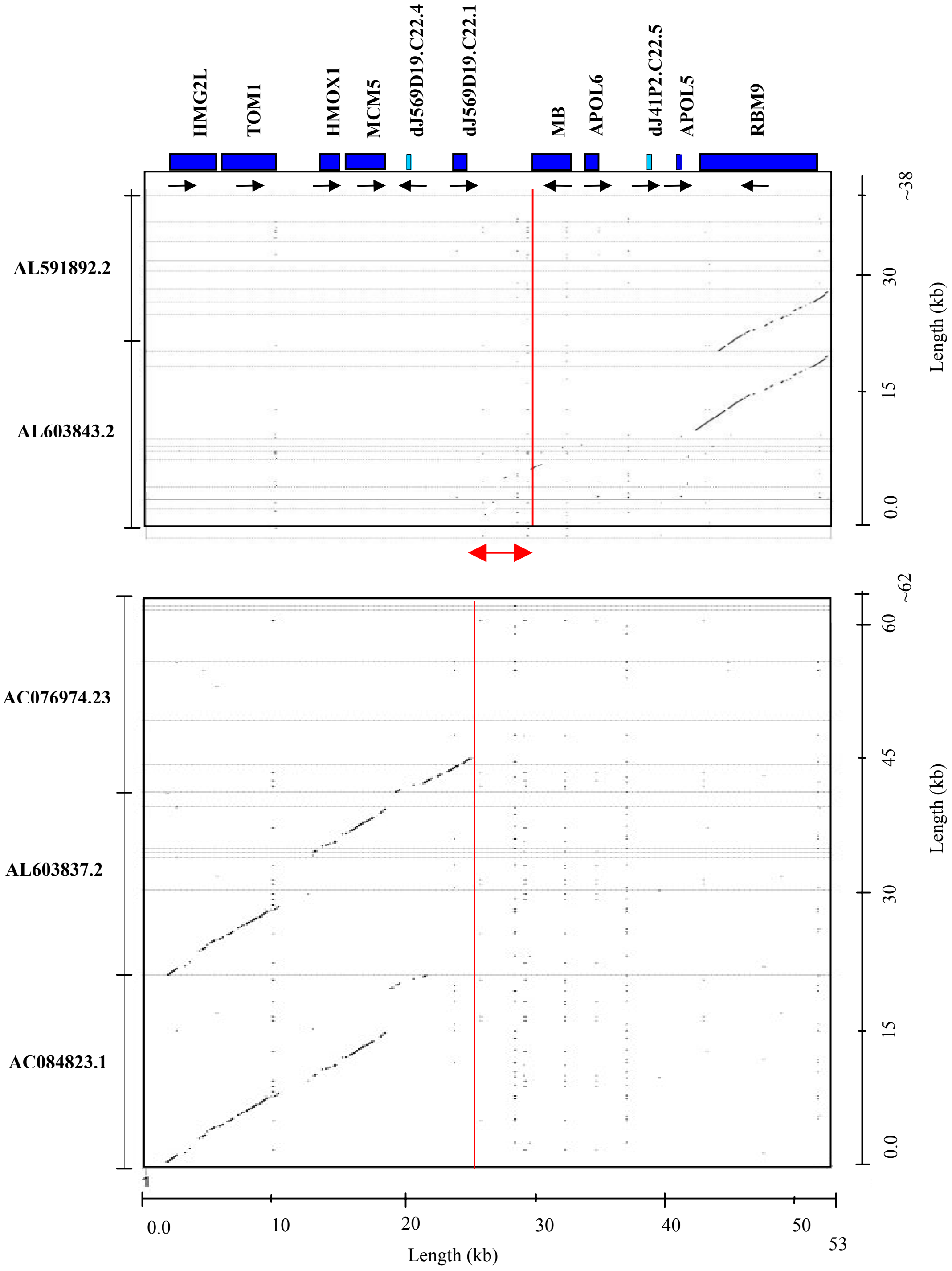


Figure 4.22 : Annotated dot plot of regions of mouse chromosome 8 and 15 available sequences (Y-axis) against the syntenic region of human chromosome 22 sequence (X-axis). The boxes along the X-axis indicate the human genes (dark blue). Human pseudogenes are indicated in light blue. The MMU8:15 syntenic breakpoint on HSA22 lies between dJ569D19.C22.2 and MB (indicated in red). The dot plot was generated using the PipMaker suite of analysis tools (http://bio.cse.psu.edu/pipmaker)

The schematic in Figure 4.23 shows the genes found adjacent to the junction region in the human and mouse chromosomes.

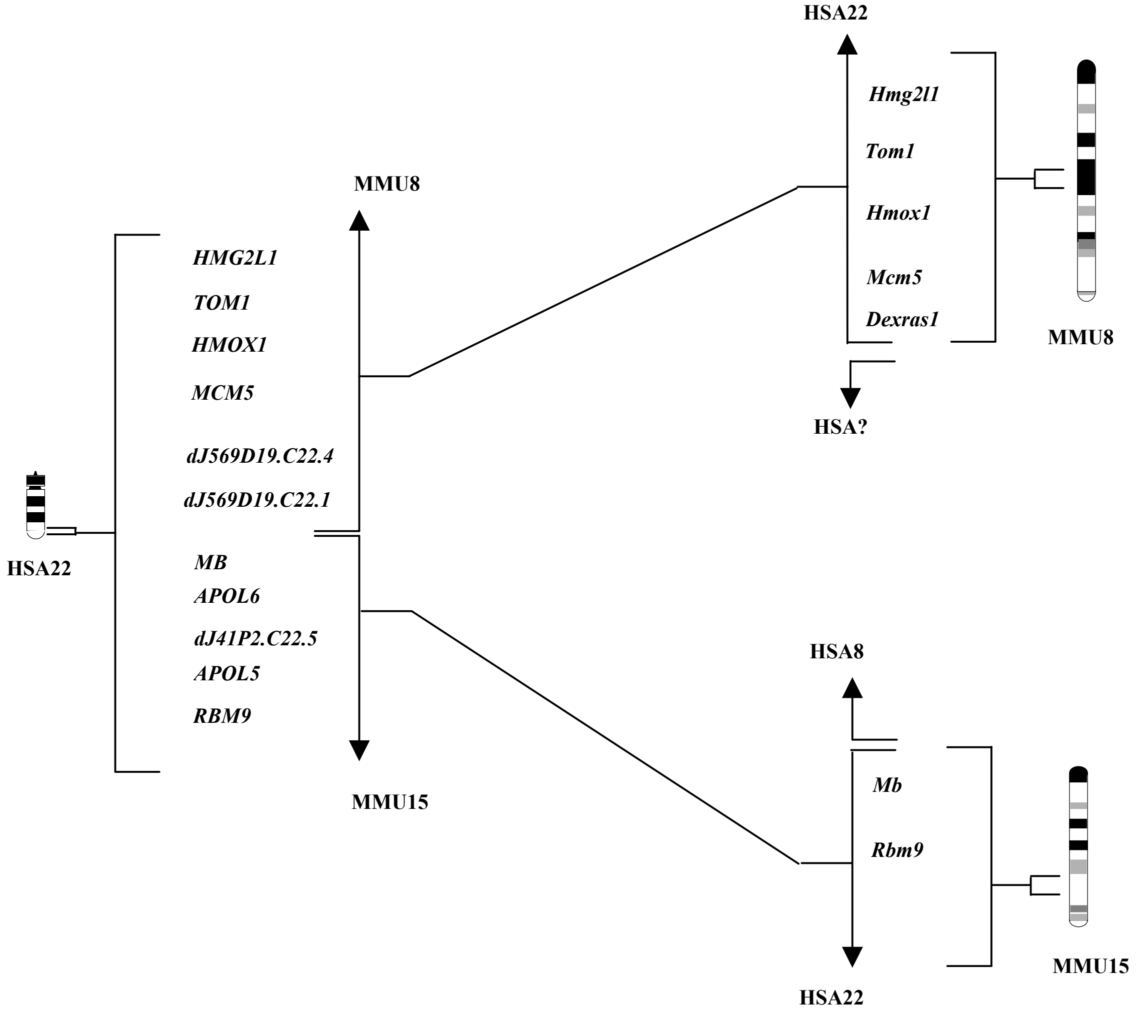


Figure 4.23: Comparative maps define the MMU8:15 chromosome junction region on human chromosome 22. HSA22 gene order is used as the reference. Apart from the apparent absence genes APOL5 and APOL6 and pseudogenes dJ569D19.C22.4 and dJ41P2.C22.5 in the mouse sequence, linkage is conserved within the two mouse chromosomal regions.

Sequence similarity between HSA22 and MMU15 decreases after the gene MB. BLAST experiments using the mouse sequence against the NCBI human genome database show that mouse sequence after this point may be syntenic with HSA8. Additionally, gene predictions in the unfinished sequence provided by the mouse Ensembl website (http://mouse.ensembl.org) also matched HSA8 sequences in similar BLASTP experiments. This finding correlates with data from the NCBI human-mouse homology map (http://www.ncbi.nlm.nih.gov/Homology).

Similarly, sequence similarity between HSA22 and MMU8 decreases after dJ569D19.C22.1. BLASTP experiments of the mouse sequence against the NCBI human genome database showed low-level similarity to HSA13 and HSA20. However, no genes have been predicted to lie within bM267J18 by Ensembl prediction methods (http://mouse.ensembl.org) and no further information is available on the NCBI human mouse homology map for this region.

4.8.2 The junction region

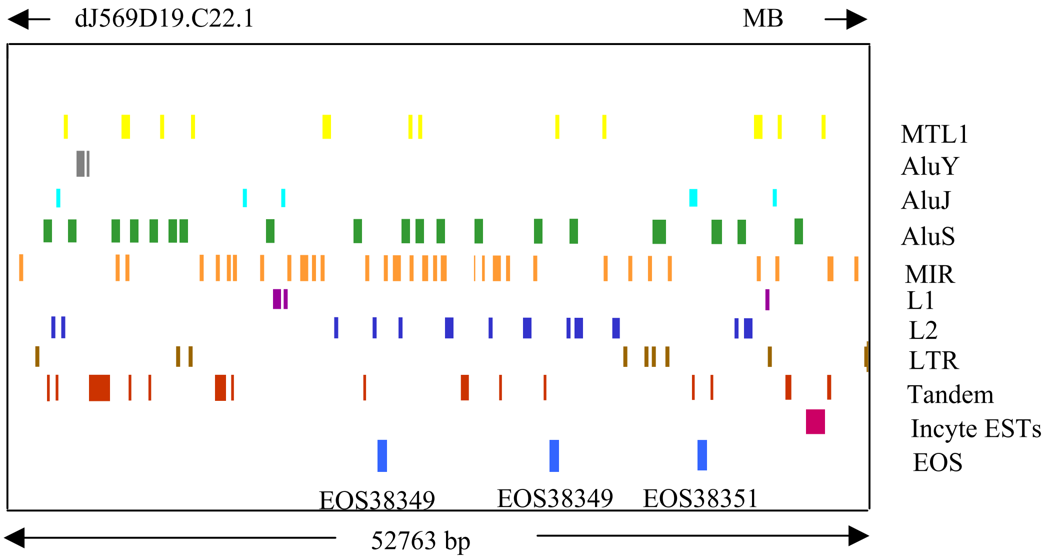


Figure 4.24: Comparative sequence analysis defines the MMU8:15 junction region on human chromosome 22. The junction region is composed of a variety of human repetitive DNA sequences. A cluster of Incyte EST sequences and 3 EOS sequences (see chapter III and appendix 2) are also included within the region.

Repeat sequences make up 40.65% of the 52763 bp MMU8:15 junction region on HSA22 (figure 4.24) and consist of several classes of repetitive DNA elements. Thirty-three mammalian-wide interspersed repeats (MIRs) were found, distributed throughout the region. The current unfinished nature of much of the mouse sequence in this region, however, makes it difficult to ascertain if these MIR repeats are conserved in the mouse genome. MIRs are believed to have amplified before the radiation of mammals, and their transposition has been implicated in gene control and evolution (Hughes, 2000). A single MIR repeat has also been observed in a HSA21:22 junction region on MMU10 (Pletcher *et al.*, 2000), although no similarity is noted in the distribution of repeat sequences between these two examples.

Three 'EOS' sequences, that have been predicted to be coding by Genscan and which have tested positive for expression by microarray hybridisation (R. Glynne, personal communication) (chapter III and appendix 2), were also contained within the region. Two showed a high level of conservation with sequences on mouse chromosomes 5 (EOS38349), 15, 11, 3, 18 and 6 (EOS38350). EOS38351, along with seven overlapping ESTs from the Incyte database (J. Seilhamer, personal communication) (chapter III and appendix 2) identified in this region, but did not show significant similarity to any other human or mouse DNA or protein sequence by using BLASTN and BLASTX. The remaining 27980 kb of unique sequence was not similar to any known human or mouse sequences.

The sequence analysis of this region and of evolutionary chromosomal breakpoints previously described at the sequence level by both Lund *et al*. (2000) and Pletcher *et al*. (2000), has so far revealed no unusual sequences or repeat structure that might suggest chromosomal instability underlying the rearrangements. As increasing amount of mouse genomic sequence become available, perhaps further examination of similar regions will identify common features of evolutionary chromosomal breakpoint regions.

4.9 Discussion

This chapter has described the construction, sequencing and comparative sequence analysis of approximately 3.5 Mb of the mouse genome, spanning regions of conserved synteny with human chromosome 22q13.31 and with a syntenic breakpoint between mouse chromosomes 8 and 15, within a region of human chromosome 22q13.1.

The use of both fingerprinting and landmark content mapping initially contributed to the construction of three contigs across regions of interest on mouse chromosomes 15 and 8. Restriction enzyme fingerprinting allows analysis over the length of the clone and the construction of contigs relies on the number of bands shared between overlapping clones. The disadvantage of fingerprinting is that it does not allow the orientation of the contigs relative to each other, nor does it allow integration with the framework map. Initial landmark STSs were designed from known orthologous mouse mRNA sequences. Increased marker density was achieved by including STSs to mouse ESTs that demonstrated high similarity to the remaining human genes. The increasing availability of marker and fingerprint data from the mouse physical mouse mapping effort (MGSC, unpublished) anchored the initial contigs to existing mouse framework maps. This combined approach offered the best strategy for contig construction, determining accurately the overlap between clones and integration of the constructed contigs with the framework maps. The resulting BAC maps from this effort provide a resource for the genomic sequencing of these regions of mouse chromosomes 15 and 8 and have been incorporated into the mouse physical map produced by the MGSC (http://mouse.ensembl.org).

PIP analysis of regions of available sequence, show that approximately 90% of annotated gene features within 22q13.31 and 22q13.1 are conserved. 76% of the annotated exons within these regions of HSA22 demonstrate >50% sequence identity with mouse genomic sequence. Interestingly, no mouse sequence homology was noted, outside of repeat regions, of the 18 human pseudogenes annotated in these regions. It may be that these non-functional sequences have diverged more quickly in the mouse genome, possibly because of the much shorter generation time of mouse. Alternatively, some, or all, of the pseudogenes may have arisen in the human lineage after divergence from the common mouse-human ancestor. Otherwise, gene order is generally conserved in these regions. Exceptions were seen with the genes APOL5 and APOL6, which were not found in the available mouse sequence and the APOL2 gene, which may be inverted in mouse. However, a large part of this analysis is based on unfinished sequence and is therefore unconfirmed.

A percentage identity level of 80% was selected for alignment of the mouse genomic sequence generated from this project against the sequence of 22q13.31 and incorporation into 22ace for further analysis. The basis for this choice was the result of preliminary alignment experiments on a subset of the region at a range of identity levels, which suggested that beyond a level of 80% identity, specificity and sensitivity were compromised. This observation is supported by Makalowski and Boguski (1998), who reported that protein-coding exons show an average percent identity of ~85% for many comparisons between human and mouse genes.

The alignment of the 39 annotated gene structures within 22q13.31 (chapter III), with both the mouse genomic sequence generated from this project and other examples of sequence evidence from model organisms, was analysed using MethComp (D. Beare) (chapter II). Higher levels of specificity and sensitivity were noted for genomic sequence resulting from BLASTN comparison at a level of 80% nucleotide identity of sequence generated by a clone-by-clone shotgun approach than from the WGS mouse project (MSC, unpublished). This may be because the clone-by-clone approach has generated more complete data over the region than the current stage of the WGS project. Interestingly, BLAT alignments (Kent, unpublished) of the output from the WGS project showed greater sensitivity and specificity than alignments from the Exonerate program (Slater, unpublished). The completion of the mouse genome project and reanalysis of these alignments should provide a definitive measure of the correlation of human and mouse sequence in this region.

Overall, these results and those from the equivalent calculations described in chapter II, indicated that the most efficient approach to annotation is through comparison to known gene or protein sequences, both from human and from model organisms. However, this study showed that mouse genomic sequence has the potential to provide an important tool in annotation of the human genome sequence, although comparative sequence analysis utilising mouse genomic sequence supported, but did not add to, the annotation of this already well-studied region (see below). The utility of mouse genomic sequence in this field may therefore lie in the annotation of human genes in previously unstudied regions.

The two regions of human chromosome 22, unlike other examples (Epp *et al.*, 1995; Koop & Hood, 1994; Oeltjen *et al.*, 1997) do not show extensive conservation of intronic and intergenic sequences with mouse, although several isolated examples were noted. Only six conserved regions were also predicted to contain exons by the gene prediction program Genscan (Burge & Karlin, 1997). Three of these predicted exons had already tested negative for expression by

PCR screening of cDNA libraries (chapter III). The remaining three predictions, together with a further 25 candidate exons identified from the human-mouse alignment were tested for expression in seven cDNA libraries. No new exons were confirmed. It is possible, however, that these conserved regions could be transcribed in different tissues or under different conditions than the seven cDNA populations tested. A benefit of mouse sequence comparison is that, unlike EST and cDNA evidence, identification of putative coding regions is not limited by spatial or temporal restrictions on transcription. However, this also means that expression of these regions is difficult to confirm. Analysis of the finished mouse sequence, using techniques similar to those described in chapter III, including detailed comparison to the related human sequence, additional homology searches and use of gene prediction algorithms, may provide additional evidence that these conserved regions encode genes.

The conserved non-coding sequences may also indicate the presence of regulatory elements. The putative promoter regions of six genes, present in both human and mouse finished sequences, were examined for the presence of potential transcription factor binding sites. Thirty-six putative sites were identified in conserved sequences upstream of the annotated transcription start sites of four genes. This investigation represents only a preliminary *in silico* analysis and identification of these regions represents a starting point for further analysis (see chapter I). Many of the consensus sequences listed for possible transcription factor binding sites are very short – only a few nucleotides long in some cases. These could be expected to occur frequently in both functional and non-functional genomic sequence. Recent studies by Göttgens *et al*. (2000) and Frazer *et al*. (2001) have demonstrated the utility of including a third vertebrate species in comparisons of non-coding sequences. Potentially, inclusion of, for example, genomic sequence from chicken or dog, will increase the specificity of this analysis of potential regulatory regions. Non-coding sequences conserved in all three species will provide strong candidates for future investigation.

Investigation of a 0.5 Mb region of finished mouse sequence showed that the gene structure overall is well conserved in this region between the two species. Comparison of exon and intron size in mouse and human shows that coding regions are more stringently conserved. Increased variation is noted in the sizes of UTR exons. Within coding regions, most insertions/deletions of nucleotides occur in multiples of three, so the reading frame is maintained. Exceptions, such as the shift in reading frame shown between the human and murine versions of BIK, result in a decrease in identity between the predicted protein sequences. It would be interesting to determine if this change has an affect on the functions of the orthologous BIK genes.

The comparison of splice donor and acceptor sites has shown that human and mouse splice sites in this region are highly conserved. The consensus donor and acceptor sites reported in this study are very similar to those reported by Stephens and Schneider(1992) from a study of 1800 human introns, and by Smink (2001) from a study of 84 human and mouse introns. It is therefore clear from the studies that the core splice donor and acceptor sites are strongly conserved in mouse and human.

The repeat density of the 0.5 Mb finished sequence region in mouse (1.33 repeats/kb) is higher than in human (1.26 repeats/kb). This may be explained by the faster murine generation time. Most of the higher repeat density is attributable to the increase in numbers of simple and MaLR repeats. MaLRs retrotransposons are known to be still active within the mouse genome (Smit, 1996). The overall repeat coverage is greater in the human (41.28%) than in mouse (31.90%). This is mainly attributable to the larger size of the human *Alu* repeat, in comparison to the mouse B1 and B2 repeats (Ansari-Lari *et al.*, 1998). The increased coverage of human repeats contributes to the 1.6X expansion of the sequence length in human compared to mouse. The overall coverage of the repeats in this region are slightly higher than those found in other comparative studies. Ansari-Lari *et al*., (1998) have shown an overall repeat coverage of 33.36% (human) and 26.39% (mouse) whilst Oeltjen *et al*., (1997)(1997) have shown values in the BTK region to be 31.22% (human) and 16.49% (mouse). An additional study by Smink (2000), found repeat coverage to be 39.2% (human) and 11% (mouse) over a 150kb region of human 22q13.3/mouse 15.

The GC content of both human and mouse genomes in this region follow a similar pattern, although the difference in length of the equivalent sequences prohibits direct comparison. This is also reflected by the distribution of predicted CpG islands in the region. All of the six human genes fully annotated in the mouse sequence are associated with a CpG island at the 5' end, whereas only two of the mouse genes start in a predicted CpG island. Peaks in GC content can still be observed for the genes lacking a CpG island, indicating that these regions are relatively GC rich, but not sufficiently so to be predicted as a CpG island. Erosion of mouse CpG islands is generally observed due to deamination of methylated cytosine to thymidine (Cooper & Krawczak, 1989; Coulondre *et al.*, 1978). This also occurs in humans, but the shorter generation time of mouse may account for the faster rate of cytosine deamination and CpG island erosion observed in this and other studies (Aissani & Bernardi, 1991; Antequera & Bird, 1993; Matsuo *et al.*, 1993).

This region of finished mouse sequence was also interesting as it was found to span an 'unclonable' gap in the sequence of human chromosome 22. Analysis of the repeat content of the mouse 'gap' subregion showed no obvious deviation from that of the total analysed sequence. GC content, however, was maintained at a high level throughout this section. The human GC levels are estimated to be maintained above 50% throughout the gap region. This observation could be a reason why efforts to identify a clone containing the equivalent region in human have so far been unsuccessful. In *Escherichia coli*, (CpG)n repetitive sequences have been shown to be deletion prone (Bichara *et al.*, 1995, 2000). Two pathways have been suggested by which this could occur;

(1) (CpG)n tracts are potential Z-forming DNA sequences and this DNA structure could be processed by an unknown cellular mechanism to give rise to the observed deletions
(2) (CpG)n monotonous runs can be considered as a succession of direct or palindromic repeats, allowing formation of DNA structures that are known to participate in frameshift mutagenesis.

The sequence of the mouse clone and putative structure of the human C22orf1 gene identified by this study could be used in the design of new hybridisation experiments in attempts to identify a human genomic clone spanning this gap from the available libraries.

Examination of unfinished sequence from mouse chromosomes 8 and 15 enabled a more precise definition of the MMU8:15 synteny junction on human chromosome 22q13.1. Investigation of the finished mouse sequence, when available, may further reduce this region. Analysis of the finished human sequence of this junction region identified a range of different repetitive features, including MIR repeats. MIRs are thought to have arisen before the radiation of mammals, and their transposition has been implicated in gene control and evolution (Hughes, 2000). Comparison of this region with the synteny breakpoints analysed by Pletcher *et al.* (2000) and Lund *et al.* (2000), identified no similarity in the distribution of repetitive sequences. As additional comparative sequence information becomes available, analyses of a range of such synteny breakpoint junction sequences may enable identification of common elements.

In summary, this chapter has shown that comparative sequencing is a powerful tool for the annotation of genomic sequence. Although all the genes annotated during this project were identified without the aid of mouse genomic sequence, the high levels of correlation of the mouse-human sequence alignments with the human transcript map indicate that a completed mouse genome sequence resource will provide a useful gene-finding resource. Comparison of human and mouse genomic sequence will therefore speed the annotation of both genomes. Comparative sequence analysis also enhances *in silico* prediction of conserved regulatory sequences. As the genomic sequence from other vertebrate model organisms becomes available, this process may become more efficient. Comparative analysis also enables detailed, sequence-level analysis of chromosome evolution. This study showed that the availability of genomic sequence permits a level of definition of evolutionary breakpoints that was previously unavailable. An understanding of the mechanism behind these evolutionary changes may develop as more of these detailed comparisons are perfomed.