Chapter 5

Comparative Sequence Analysis

Between Human and Mouse

5.1 Introduction

Humans and mice diverged from a common ancestor approximately 70 million years ago and have a comparable genome size (O'Brien, S. J.*, et al.*, 1999). Comparison of orthologous genes in human and mouse and their function has shown that sequence similarity across much of the coding regions of genes and some of the regulatory elements that control them has been maintained since the split from a common ancestor. Some of the early evidence of conservation between human and mouse came from comparative analysis of 100 kb of human and mouse T-cell receptor DNA (Koop, B. F.*, et al.*, 1994). More recently, regions of conservation have been identified upstream of the SCL gene in human, mouse and chicken, and were later shown to be associated with active regulatory regions (Gottgens, B.*, et al.*, 2001).

The striking sequence similarity between human and mouse in specific genomic regions arises because functionally conserved features between genomes tend to be conserved at the sequence levels. This allows for inferences to be made about one organism using information determined in the other. Comparative sequence analysis is therefore a powerful tool for aiding both human gene identification and understanding the function and control of genes. As discussed in the previous chapter, the function of only a small proportion of human genes identified to date has been experimentally determined. The identification of the orthologous genes between human and mouse will enable function of the human counterpart to be inferred based on the investigation into the function of the orthologue in mouse.

The evolution of mammalian sex chromosomes is characterised by the loss of genes from the Y chromosome by mutation. This, in turn, has led to the development of X inactivation in females in order to achieve dosage compensation for X-linked gene products. Ohno's law states that due to dosage compensation in females, it is thought there is selective pressure to maintain dosage-dependent genes on the X chromosome (Ohno, S., 1967). In agreement with Ohno's suggestion, X-linkage of genes is generally maintained in the eutherian mammals, which is in contrast to what has been observed for the autosomes. The human X chromosome is currently represented by nine syntenic blocks all positioned on the mouse X chromosome (see Figure 5.1). In contrast, human chromosome 6, a similarly sized chromosome to the human X chromosome is represented by at least eight syntenic blocks in the mouse, but present on seven different mouse chromosomes (data taken from http://www.ncbi.nlm.nih.gov/Omim/Homology/human6.html). The availability of orientated X chromosome sequence in both human and mouse allows for a refinement of the synteny map, enabling the precise order and transcriptional orientation of genes to be studied.

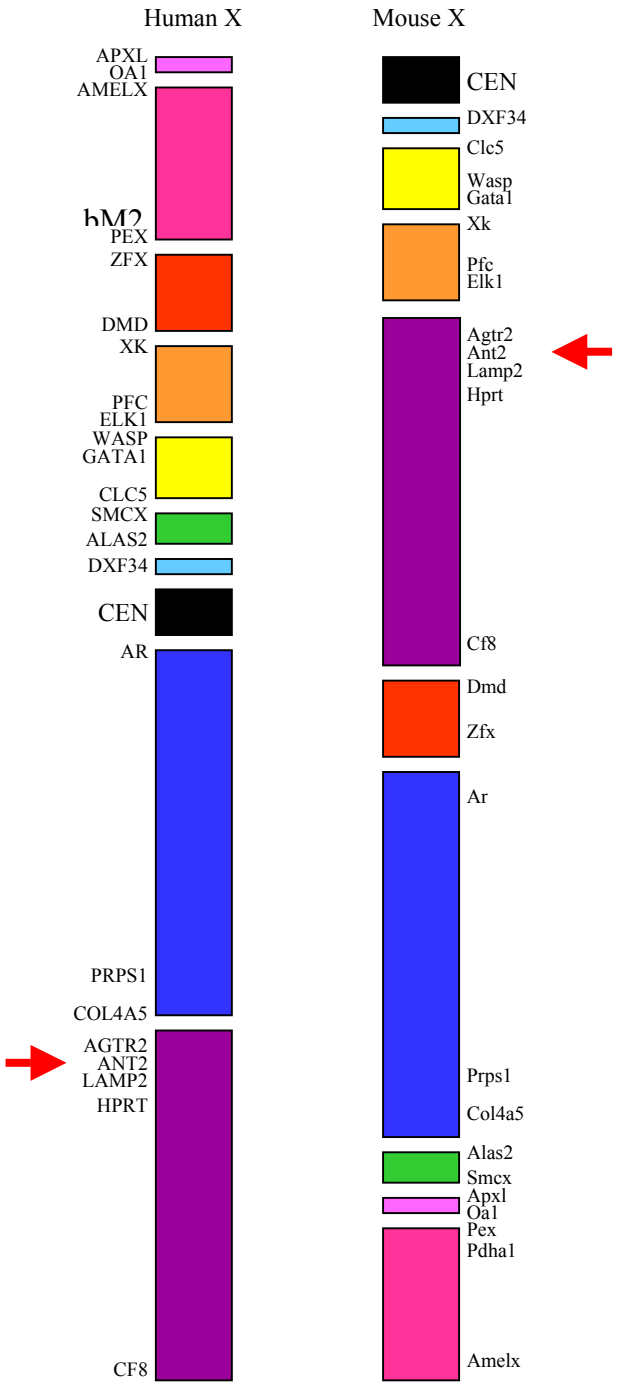
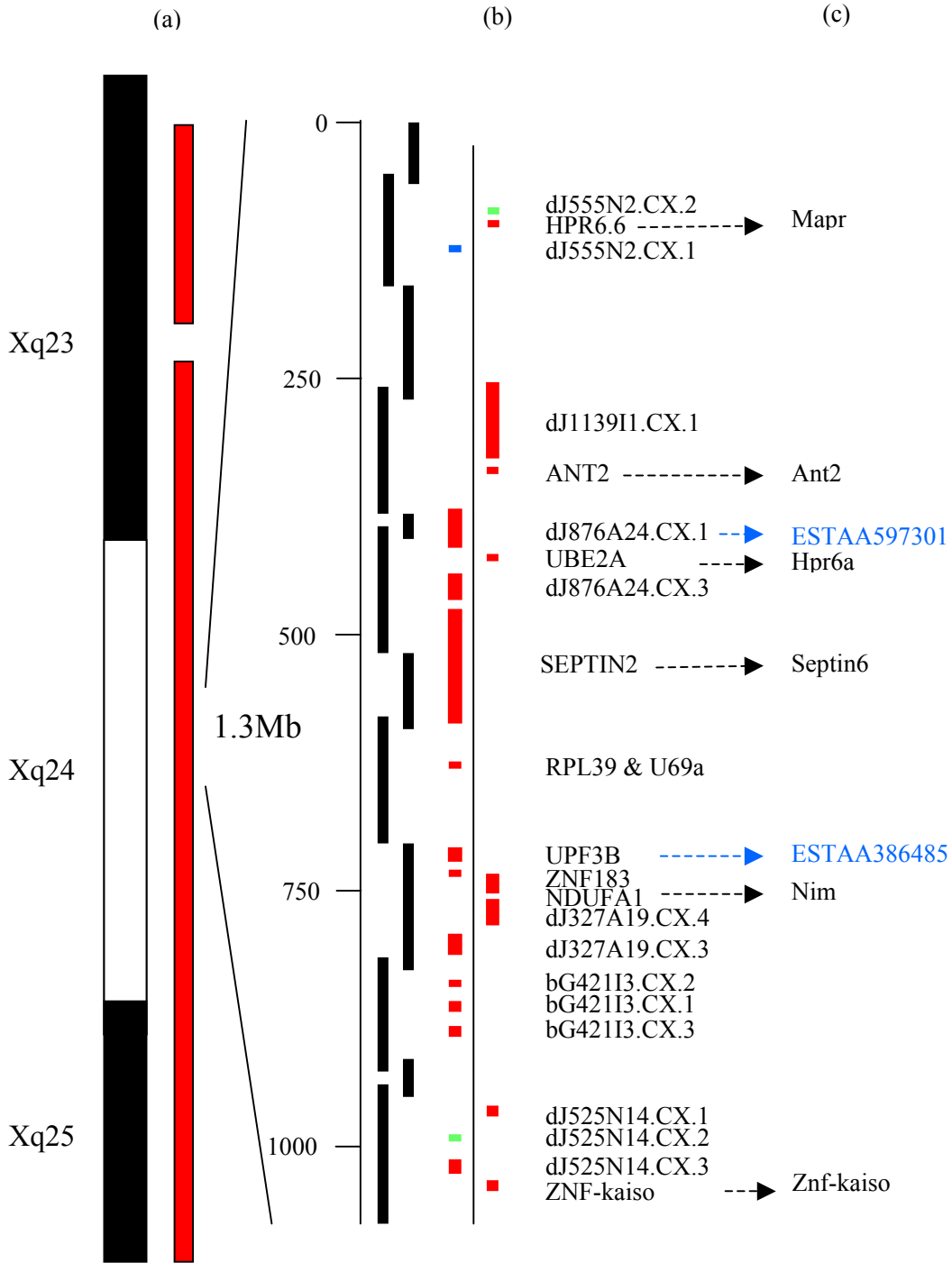


Figure 5.1: *A schematic representation of the syntenic relationship between human and mouse (reproduced from Boyd, Y. et al., 1998). The human X chromosome (left) is divided into 9 syntenic blocks when compared to the mouse X chromosome (right). Syntenic blocks are shown in the same colour. A subset of the markers known to map to each block is also shown. The position of the region studied in this chapter, estimated from the position of mouse Ant2 gene, is shown as a red arrow.*

In this study, a contiguous segment of finished sequence in human Xq24, between HPR6.6 and ZNF-kaiso was chosen for comparative analysis with the mouse. The region was chosen because of the advanced state of the human sequencing and annotation at the time (as discussed in the previous chapter). The 1.3 Mb region contains twenty genes, of which nineteen have been confirmed by cDNA sequence and one remains predicted, and one pseudogene (see Figure 5.2). The syntenic region in mouse is thought to be located in the proximal section of the fifth syntenic block, based on the position of one known orthologous gene, Ant2 (arrowed in Figure 5.1). The aim of the work contained within this chapter was to investigate the usefulness of mouse sequence for annotating human genes, and generate a detailed comparative map of orthologous genes in the region.

Figure 5.2: *(see over) Summary of the region for comparative analysis. (a) The position of the region of interest in relation to the transcript map described in the previous chapter. (b) The minimum set of clones (denoted as vertical black bars) and the genes identified (genes in red, predicted genes in blue, pseudogenes in green). Genes on the left of the thin vertical black line are transcribed on the minus strand, genes on the right are transcribed on the plus strand. A scale is shown in kilobases.*

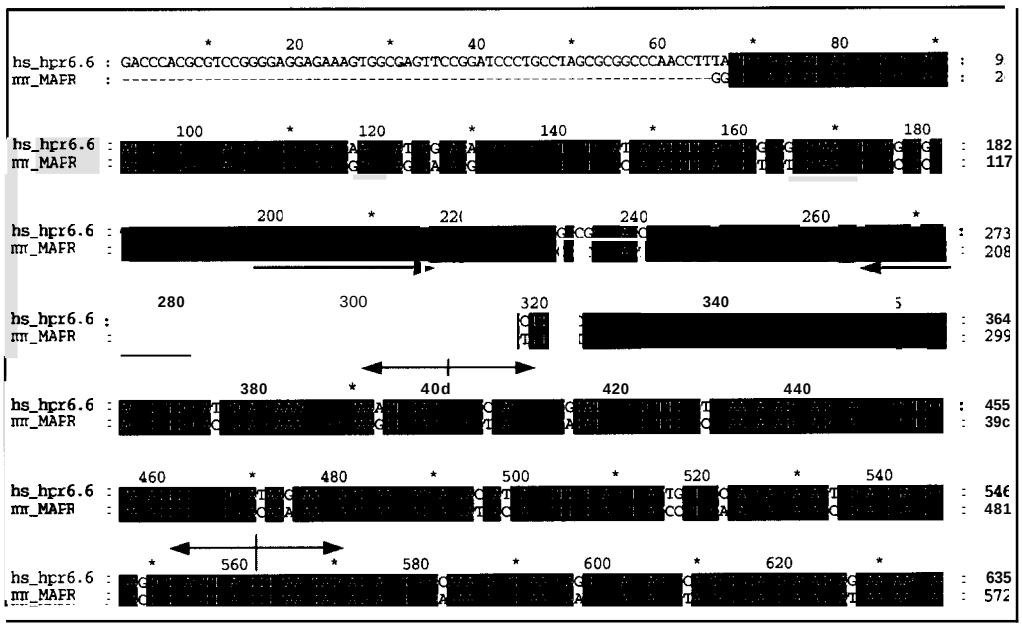
(c) The known orthologous genes (shown in black) or ESTs (shown in blue) used for bacterial clone isolation. Arrows link orthologous sequences.



5.2 Construction of bacterial clone contig

The identification of eight mouse-specific sequences expected to map to the syntenic region in mouse provided the basis for the construction of the bacterial clone contig (see Figure 5.2c). These consisted of six mouse mRNAs known to be orthologous to human genes in the region and two mouse ESTs that were greater than 90% identical at the nucleotide level to human genes. For each orthologous pair, the two sequences were aligned in order to identify the most likely positions of the introns in the mouse gene, based on the positions of the human introns. Examples of the alignments can be seen in Figure 5.3. For each orthologous pair, a PCR assay was then designed within a single exon of the mouse sequence, and used for mouse genomic bacterial clone isolation (see Figure 5.4).

(a)



(b)

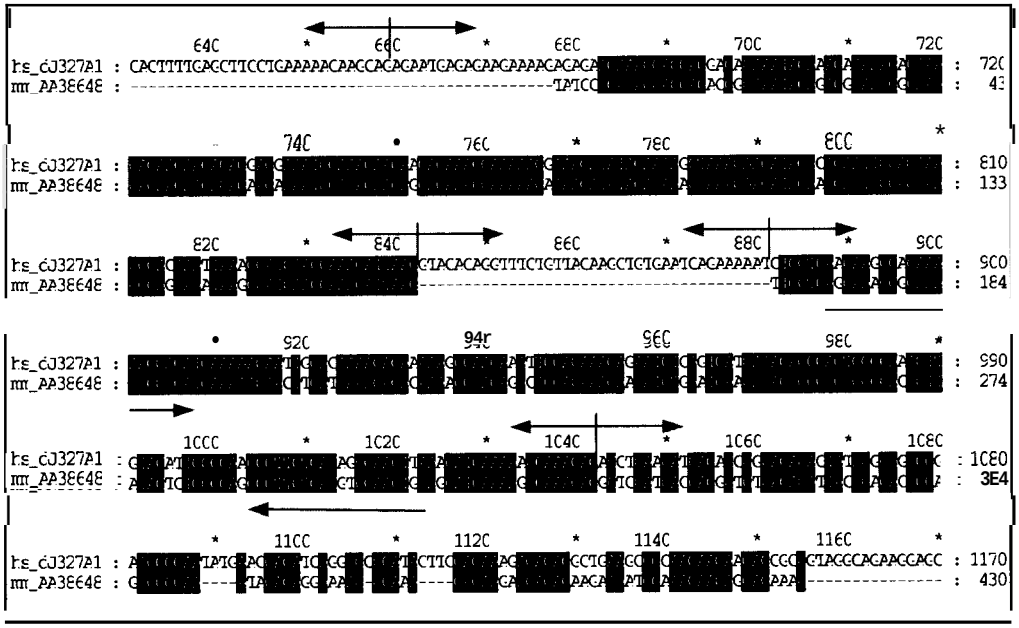
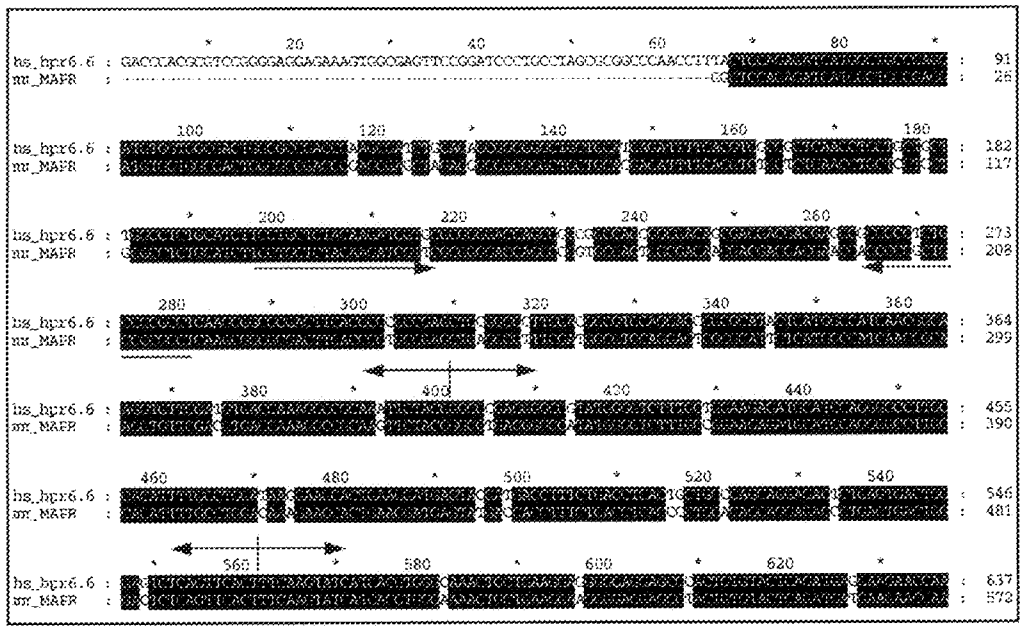


Figure *5.3: Examples of alignment between human and mouse orthologues. Sequences are aligned using CLUSTALW and visualised in GENEDOC. The boundaries between exons are shown as red arrows and the positions of primers for the STS are shown in blue. (a) An alignment between HPR6.6 (human) and MAPR (mouse). (b) An alignment between part of UPF3B and the EST AA386485.*

(a)



(b)

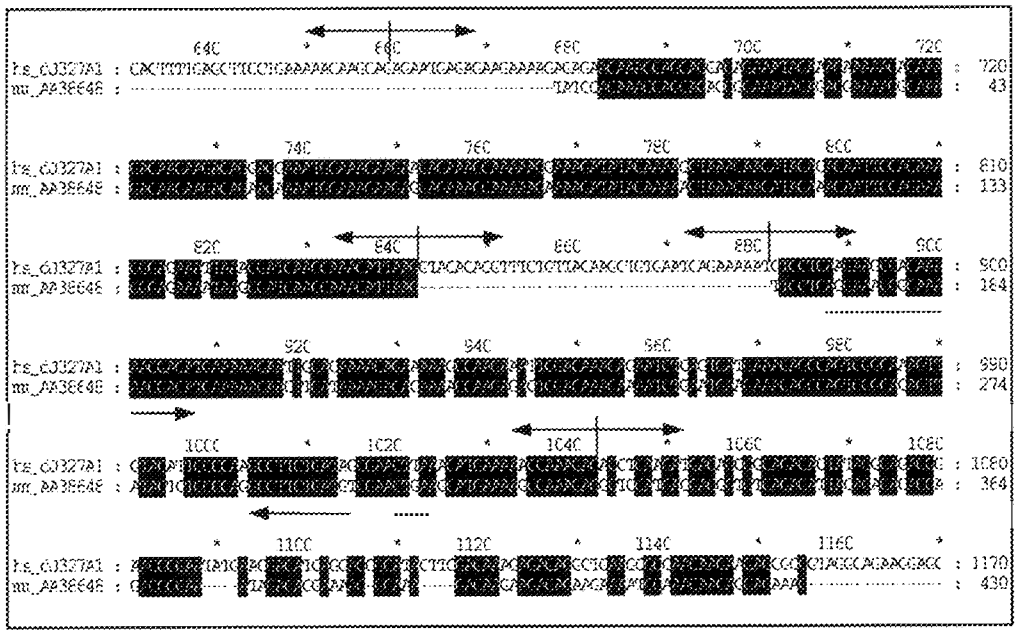


Figure 5.3: *Examples of alignment between human and mouse orthologues. Sequences are aligned using CLUSTALW and visualised in GENEDOC. The boundaries between exons are shown as red arrows and the positions of primers for the STS are shown in blue. (a) An alignment between HPR6.6 (human) and MAPR (mouse). (b) An alignment between part of UPF3B and the EST AA386485.*

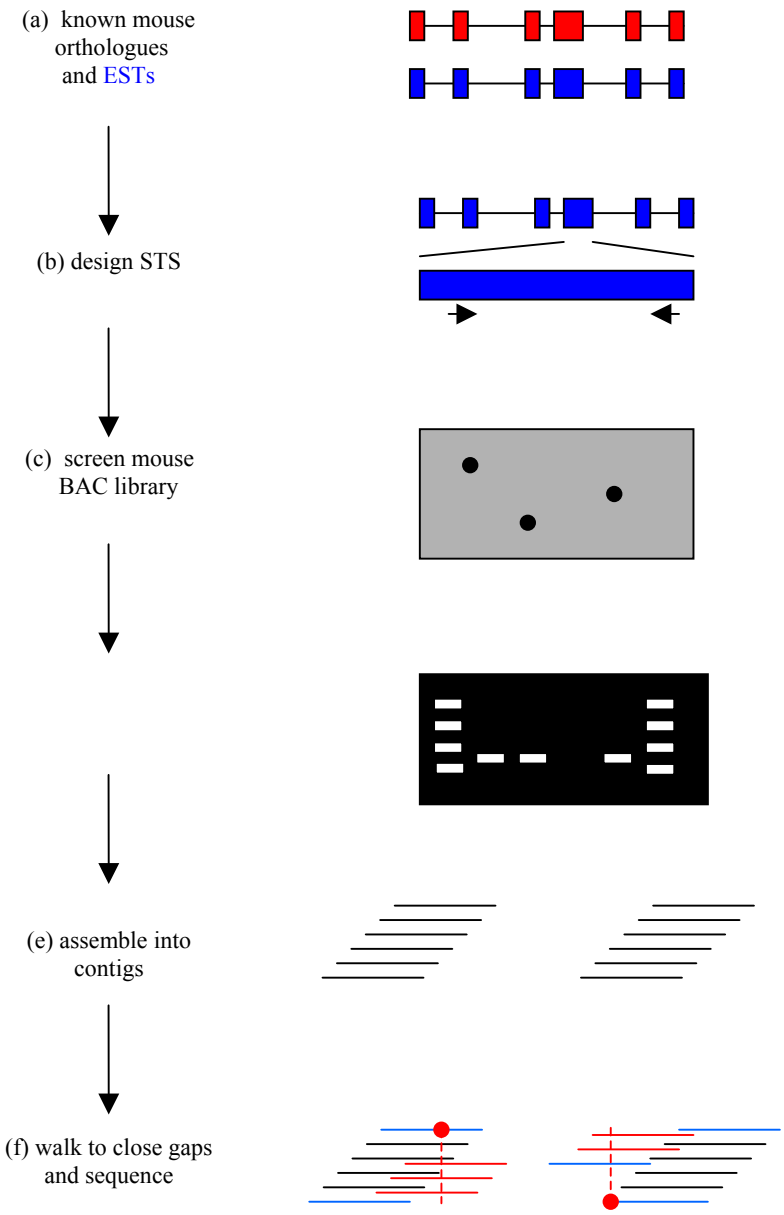
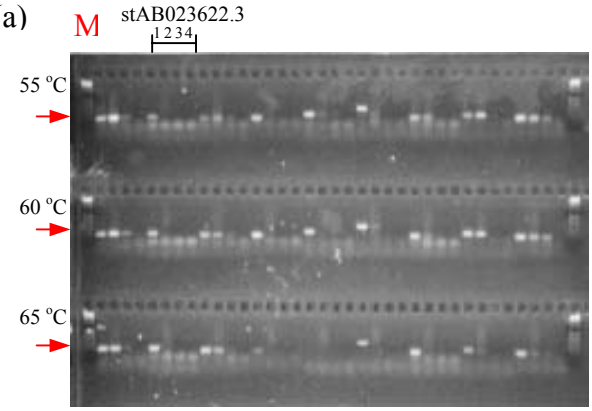


Figure 5.4: *Strategy for contig construction (a) Orthologous genes are aligned (human shown in red and mouse shown in blue) and (b) STSs designed within a single exon (primers shown as black arrows). (c) A pool of STSs is hybridised to a filter (shown as a grey rectangle) and positive BACs (black spots) are identified. (d) BACs positive for each STS are (shown as white rectangles) and (e) clones are assembled into contigs by fingerprinting (horizontal black lines). (f) New STSs (red dot) are identified at the ends of contigs for contig extension and a minimum set of clones identified for sequencing (shown as blue horizontal lines).*

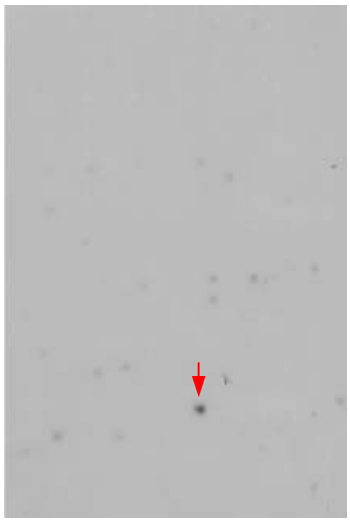
The eight PCR assays, designed to the orthologous mouse sequences, were pooled and hybridised to a gridded array of BAC clones (RPCI-23, see Section 2.8.2) (see Figure 5.5). A total of 45 BACs were identified and positive clones for each PCR assay were confirmed using colony PCR. All BACs identified within the region were assembled into contigs using *Hin*d III fingerprinting (see Section 2.12.3). At this stage, there were 2 contigs covering 1.1 Mb, estimated using fingerprint band sizes (see Section 2.23.2). A section of the contig showing the integration of clones positive for stAB023622.3, designed within an exon of the mouse septin6 gene, is shown in Figure 5.6.

Figure 5.5: *(see over) BAC clone isolation with mouse-specific STSs. (a) Nine STSs designed from sequence generated to nine mouse-specific sequences were tested for their ability to amplify unique sequence in mouse genomic DNA at three different temperatures of the PCR. Reactions included mouse genomic DNA (1), human genomic DNA (2), human X-chromosome hybrid (3), and T0.1E (4). M = marker (b) The product of amplification of mouse genomic DNA using stAB023622.3 designed to the mouse Septin2 gene was labelled, along with eight other products, pooled and used as a hybridisation probe to screen gridded filters containing BAC clones from RPCI-23. The filter shown, bM-18, has one positive clone as marked. (c) Positive clones detected on all filters were streaked and individual colonies tested against stAB023622.3 Positive clones are listed to the side.*

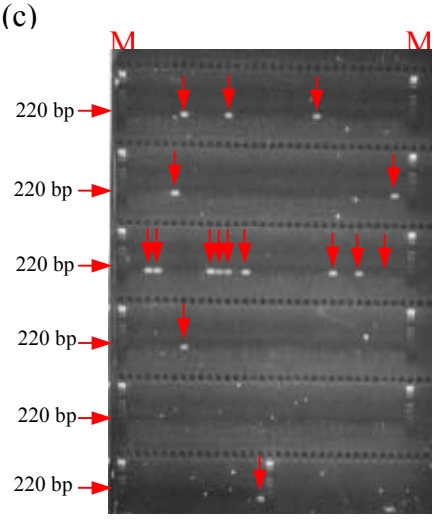
(a)



(b)

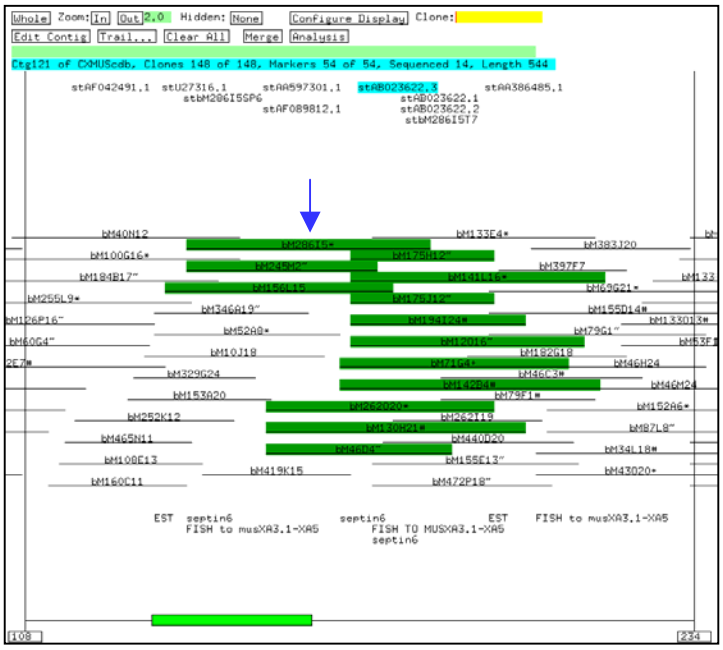


(c)



stAB023622.3 – septin6
1. bM12O16
2. bM26G17
3. bM46D4
4. bM71G4
5. bM130H21
6. bM141L16
7. bM142B4
8. bM156L15
9. bM175H12
10. bM175J12
11. bM194I24
12. bM245M2
13. bM262O20
14. bM272k24 (faint)
15. bM286I5

(a)



(b) bM286I5CX.4

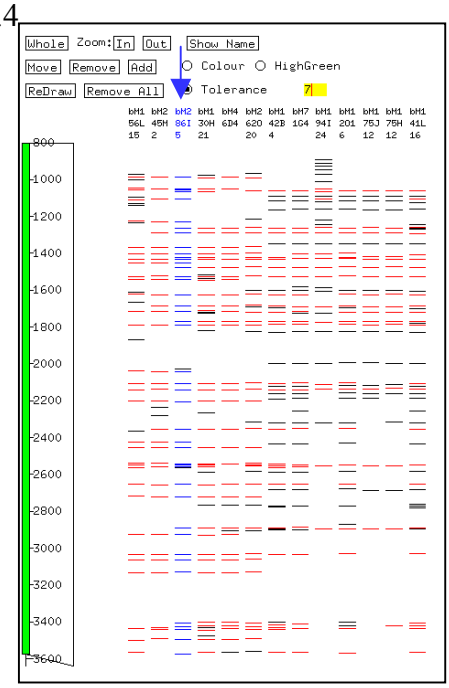
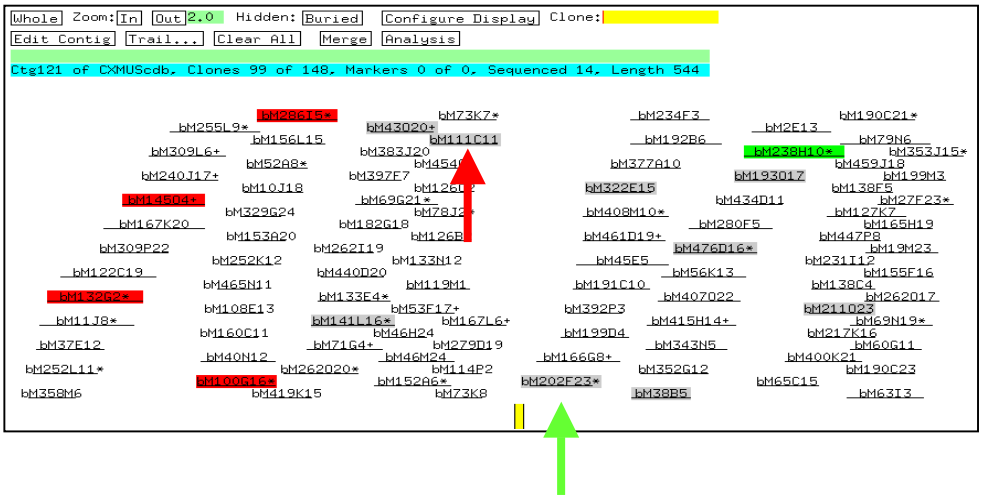


Figure 5.6: *Contig construction by fingerprinting. A section of the contig from FPC, showing the positive clones (highlighted in green) for stAB023622.3 (highlighted in blue) and (b) their fingerprints. The fingerprints of the third positive clone (counting from the left) are shown in blue, and bands with comparable migration distances for the clones in the other lanes are shown in red.*

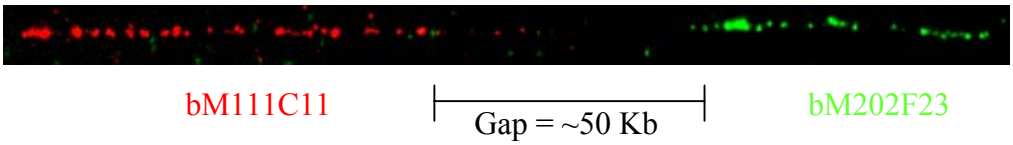
At this time, sequence was being generated at the ends of all BACs in the RPCI-23 library (http://www.tigr.org/tdb/bac_ends/mouse/bac_end_intro.html). In an attempt to close the gap between the two contigs, 13 STSs were designed to the available end sequence from 9 BACs that were positioned close to the ends of each contig. A pooled probe containing the thirteen STSs (see Table 2.7) was hybridised to the gridded array of RPCI-23 BACs, and the positive clones confirmed by colony PCR. The 53 newly identified BACs were fingerprinted and incorporated into the existing contigs. At this stage, there were still two contigs but these had been extended to cover 1.85 Mb. In a further attempt to close the remaining gap, a second pooled probe containing two novel STSs, stbM206F21SP6 and stbM202F23SP6, one from each end of the contigs, failed to identify any new clones when hybridised to the two mouse genomic libraries available at the time (RPCI-23). It was concluded that there were no clones in the available mouse clone libraries that bridged the gap between the two contigs.

In summary, a region covering 1.9 Mb has been covered in two bacterial clone contigs of 1.1 Mb and 0.75 Mb respectively (see Figure 5.7a) and the gap has been sized at approximately 50 kb using fibre-fish (carried out by Pawandeep Dhami) (see Figure 5.7b). A minimum tiling set of seven clones were chosen for sequencing and there are two contiguous segments of finished sequence available covering 714 kb and 193 kb (December 2001) (see Figure 5.7c). A further three clones are available as draft sequence. Four of the clones identified for genomic sequencing (bM100G16, bM38B5, bM43O20 and bM322E15) were localised to XA3.1-XA5 by FISH onto metaphase spreads of mouse chromosomes (carried out by Sheila Clegg), which includes the region syntenic to human Xq24.

(a)



(b)



(c)

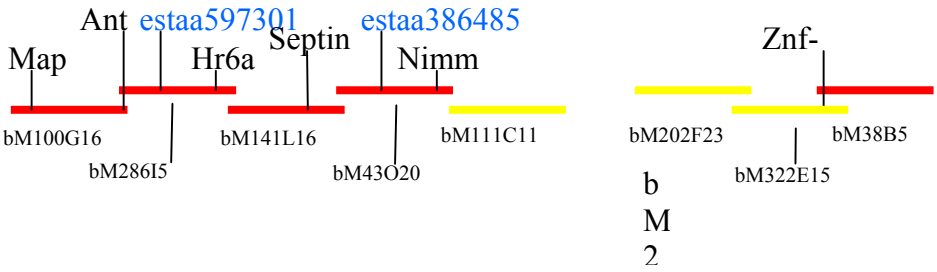


Figure 5.7: *Summary of the mapping. (a) The final contig viewed in FPC. The minimum set of clones identified for sequencing are highlighted (red = finished, grey = draft sequence, green = picked for sequence). Clones used to size the gap are indicated by a red arrow (bM111C11) and a green arrow (bM202F23). (b) The fibre-fish image showing the size of the gap with respect to the length of signal for each clone (approximately 1/3 the length of bM111C11, suggesting the size of the gap is approximately 50 kb). (c) The minimum set of clones sequenced (red = finished, yellow = draft shotgun) and the positions of the mouse-specific genes and ESTs used during the construction of the contig.*

All available genomic sequence data has been analysed as described in the previous chapter (Section 4.2). The analysis used a combination of computational gene prediction and similarity searches, matching genomic sequence to all known DNA and protein sequences. The region was found to contain a total of twenty-four genes, twenty-one genes confirmed by previously available cDNA sequence, one predicted gene and two pseudogenes (see Figure 5.8). No attempts have been made to identify confirmatory cDNA sequences for the predicted gene due the lack of availability of mouse cDNA resources at the time.

Figure 5.8: *(see over) Summary of the gene map constructed in mouse. The red bars indicate the status of the contigs and the black bars indicate the extent of finished sequence. Each link represents a series of individual clones (see appendix to this chapter). Yellow bars indicate clones for which draft sequence was available as of December 2001. A scale is given in kilobase pairs (kb). Approved names are given for known genes. Genes are indicated by arrows (black – complete, blue – predicted, green – pseudogene), the direction of each arrow reflects the direction of*

transcription. Genes on the plus strand are positioned above the dotted line, genes on the minus strand are positioned below the dotted line.

Transcript summary diag.

5.3 Identification of orthologous genes in the region

Orthologous genes are defined as being homologous genes in different organisms derived from the same gene during speciation (Postlethwait, J. H.*, et al.*, 1998). It can be very difficult to determine the true relationship of two genes from different species as both species have been undergoing independent evolution since they diverged from the common ancestor. If a gene has duplicated within a species since the divergence to create a pair of paralogues, sequence similarity alone is not sufficient to be certain that two similar genes in different species are derived from a single gene in a common ancestor. However, a number of features from each gene can be compared in order to ascertain whether two genes are likely to be derived from the same common ancestral gene. These include:

1. Similarity at the nucleotide and amino acid level.
2. Similarity of exon and intron structure.
3. Position with respect to neighbouring genes (and correspondence of identity of neighbouring genes, i.e. synteny).
4. Function if known.
5. Lack of other similarly matching sequence in rest of either genome.

Comparison of the genes found between HPR6.6 and ZNF-Kaiso in human with those identified in mouse using the criteria described above, reveals a total of sixteen pairs of genes which appear to be orthologous (see Figure 5.9 and Table 5.1). Six of these

had been previously identified and were used in the construction of the bacterial clone contigs and the remaining nine have been determined in this study. Each orthologous pair shows a high level of similarity at both the nucleotide and amino acid level and good conservation of exon structure. An example of an orthologous pair is shown in Figure 5.10.

Figure 5.9: *(see over) Comparative analysis of the region in human (on the left) and mouse (on the right). The position of genes (red = gene confirmed by cDNA, blue = predicted gene, green = pseudogene) and their direction of transcription (minus strand on the left of vertical line, plus strand on the right) are shown. The names of the genes used during the construction of the contig are shown in blue. Segment 1 (indicated by a vertical purple line) in human and mouse shows a high level of synteny. Segment 2 (indicated by a vertical green line) shows the extent of the inversion of the four genes. Segment 3 (indicated by the veritcal gold line) appears to contain apparent non-orthologous genes. The genes predicted by INTERPRO to contain a homeobox domain are indicated in bold. The genes (ZNF-Kaiso and Znf-kaiso) in segment 4 (blue line) are orthologous*

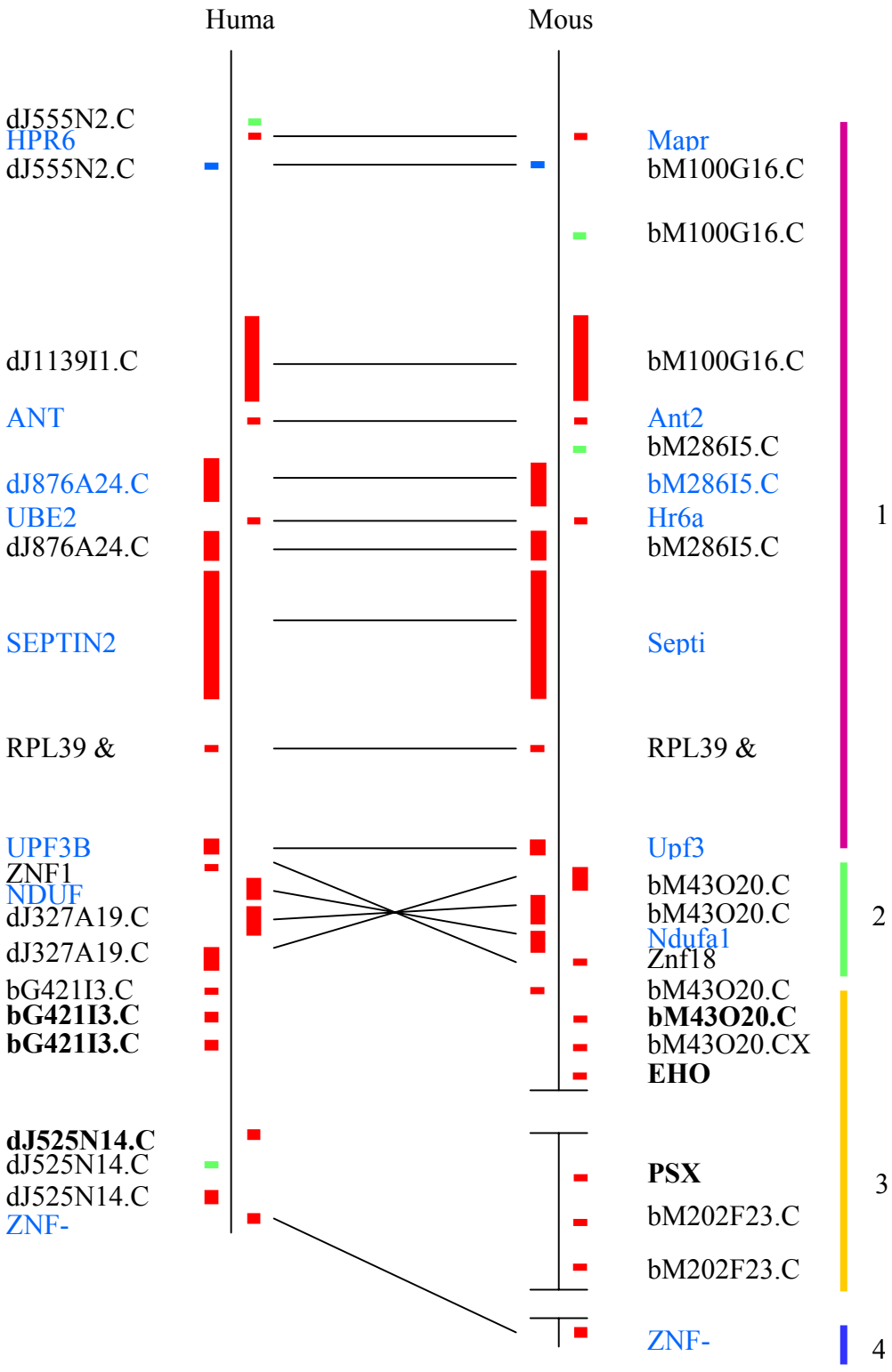
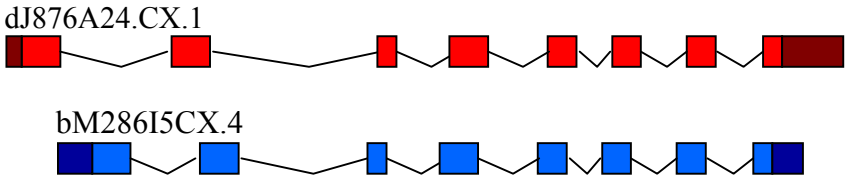

Huma
Mous
dJ555N2.C
HPR6
dJ555N2.C
dJ1139I1.C
ANT
dJ876A24.C
UBE2
dJ876A24.C
SEPTIN2
RPL39 &
UPF3B
ZNF1
NDUF
dJ327A19.C
dJ327A19.C
bG421I3.C
bG421I3.C
bG421I3.C
dJ525N14.C
dJ525N14.C
dJ525N14.C
ZNF-
Mapr
bM100G16.C
bM100G16.C
bM100G16.C
Ant2
bM286I5.C
bM286I5.C
Hr6a
bM286I5.C
Septi
RPL39 &
Upf3
bM43O20.C
bM43O20.C
Ndufa1
Znf18
bM43O20.C
bM43O20.C
bM43O20.CX
EHO
PSX
bM202F23.C
bM202F23.C
ZNF-
1
2
3
4

Table 5.1: *Comparison of orthologous genes*

Human Gene	Mouse Gene	Exon no. human	Exon no. mouse	% identity protein
HPR6.6	Mapr	3	3	98
dJ1139I1.CX.1	bM100G16.CX.4	5	5	77
ANT2	Ant2	4	4	98
dJ555N2.CX.1	bM100G16.CX.2	4	4	64
dJ876A24.CX.1	bM286I5.CX.4	7	7	99
UBE2A	Hr6a	6	6	100
dJ876A24.CX.3	bM286I5.CX.6	2	2	96
SEPTIN2	Septin6	8	8	98
RPL39	Rpl39	3	3	100
UPF3B	Upf3b	11	11	93
ZNF183	Znf183	1	1	90
NDUFA1	Ndufa1	3	3	94
dJ327A19.CX.4	bM43O20.CX.5	5	5	61
dJ327A19.CX.3	bM43O20.CX.4	9	9	94
ZNF-KAISO	Znf-kaiso	2	2	92

A high degree of synteny is apparent between the human and mouse sequence. The proximal portion of the two regions between HPR6.6 and UPF3B in human, and between Mapr and Upf3b in mouse, are exactly conserved in terms of both gene content and gene order (see Figure 5.9, segment 1). The distal portion of the region analysed appears to contain two segments where synteny is disrupted (see Figure 5.9, segments 2 and 3). Segment 2 contains four genes that appear to have undergone an inversion in one of the two species since the divergence from a common ancestor. Analysis of the order of genes in other mammals or vertebrates will enable further investigation into when this inversion took place. There is synteny at the distal end of the region (segment 4) based on the presence of the ZNF-kaiso orthologues.

(a)



(b)

Gene Name	dJ876A24.CX.1	mm_bM286I5.CX.4
5'UTR	46	106
exon 1	124	124
2	118	118
3	42	42
4	139	139
5	90	90
6	93	93
7	63	63
3'UTR	325	187
Total Coding	669	669
Total cDNA size	1040	962
intron 1	4846	267
2	14763	6730
3	971	2430
4	1758	863
5	1084	598
6	1539	1476
Total Intron size	24961	12364

(c)

Figure 5.10: *Comparison of a novel orthologous pair of genes. (a) A schematic represention (not to scale) of the gene structure of the human gene (shown in red) and the mouse gene (shown in blue). Exons are shown as boxes, introns shown as 'v' shaped lines. UTR's are shown as darker coloured boxes (b) A comparison of the sizes of exons and introns, total coding sequences and total cDNA size for the two genes. (c) An alignment of the predicted protein sequences of the two genes showing the amino acid difference (indicated with an arrow).*

Segment 3 contains a number of genes that do not appear to be orthologous between the two species. In both human and mouse, segment three contains three genes predicted by INTERPRO (http://www.ebi.ac.uk/INTERPROSCAN) to contain homeobox domains, these are labelled in bold in Figure 5.9 (bG421I3.CX.1, bG421I3.CX.3, dJ525N14.CX.1, in human, and bM43O20.CX.9, EHOX and PSX1 in mouse). Apart from bG421I3.CX.3, which has only three exons, they all have a similar gene structure with four exons and three introns (see Figure 5.11a). An alignment of the predicted protein sequences of the six genes (using CLUSTALW) shows that although there is similarity between all proteins in the region of the homeobox domains, there is no significant similarity for the rest of the alignment (see Figure 5.11b). This is also the case when individual human proteins are aligned with individual mouse proteins (data not shown).

(a

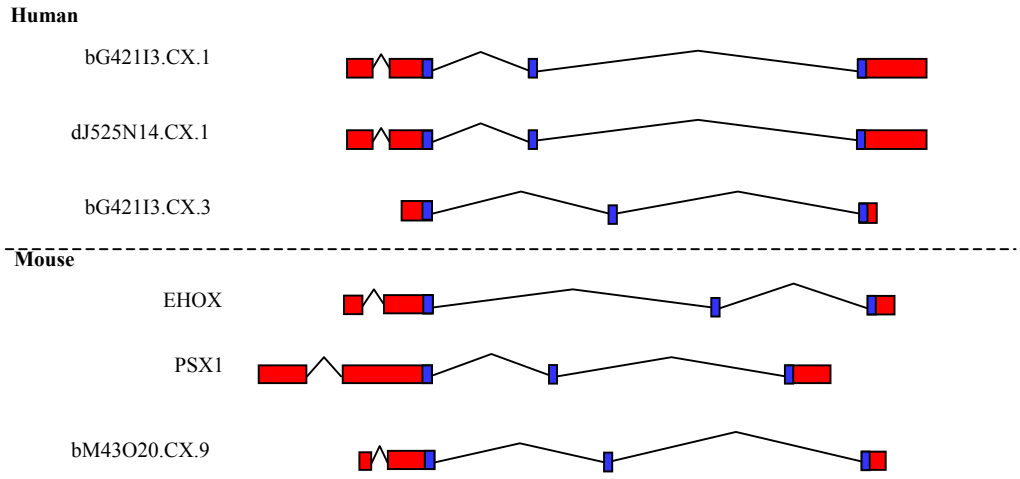
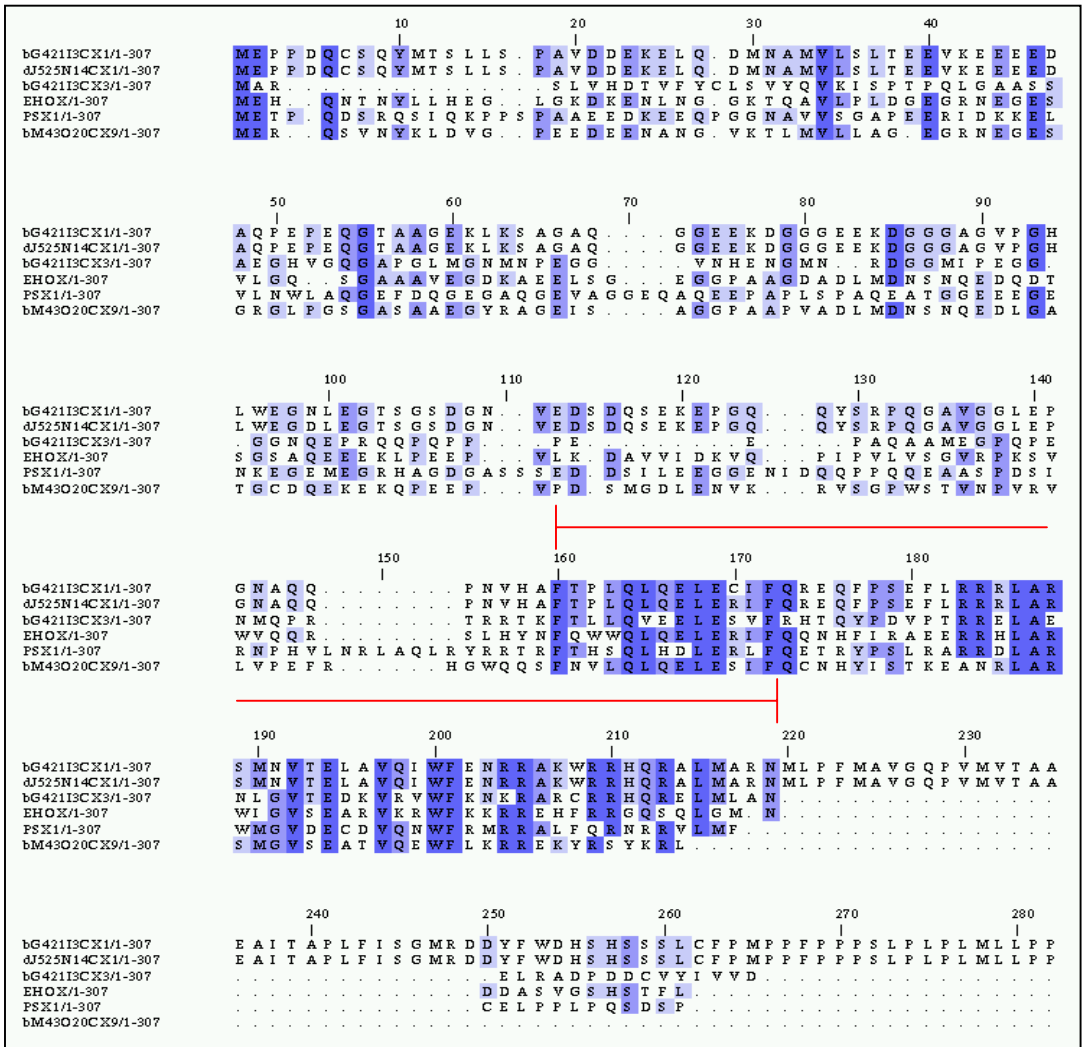


Figure 5.11: *Analysis of the homeobox genes (a) A schematic representation of the six homeobox genes located in segment 3 in human and mouse. Exons are indicated as boxes and introns are indicated as 'v' shaped lines. The region predicted to code for the homeobox domain for each gene is shown in blue.*

Figure 5.11 cont: *(b) An alignment of the predicted protein sequences of the six genes. Amino acids are given in the one letter code. Amino acids shared by at least five of the six proteins in any one position are shaded in dark blue, and those shared by three of the five proteins are shaded in light blue. A high degree of conservation can be seen in the region predicted to contain the homeobox domain (indicated by the red lines).*



Two of the human genes, dJ525N14.CX.1 and bG421I3.CX.1 are almost identical and are present within a 50 kb inverted repeat (discussed in Section 4.3.4). These appear to have arisen by duplication since human and mouse diverged from a common ancestor, due to the conservation in the positioning of an *Alu* element in each human gene, specifically an *Alu*Sx within the first intron of both genes. As *AluSx* elements arose in human between 32 million and 53 million years ago (Jackson, M. S.*, et al.*, 1996), it is likely that the duplication event occurred since humans and mice diverged from a common ancestor, estimated to be approximately 70 million years ago. Therefore, it may have been expected to observe only one orthologue in mouse to these two human genes. However, the comparative analysis in segment 3 fails to identify any true orthologue for dJ525N14.CX.1 and bG421I3.CX.1. In fact, the analysis did not identify any orthologous pairs for any of the human and mouse genes in segment 3.

There are three possible explanations for the lack of synteny observed in segment 3. The first possibility is that the orthologous counterparts of the human genes lie within the gaps present in the mouse sequence. The second possibility is that segment 3 in human and mouse are syntenic to other regions of the mouse and human respectively. However, comparison of the mouse genes with available human genome sequence and human cDNA sequence, and human genes with available mouse genome sequence and mouse cDNA sequence, reveals no other likely candidates for the orthologous genes. A third possibility is that segment 3 in human and mouse are derived from the same region in a common ancestor, but have diverged at a greater rate since the split from the common ancestor than the rest of the region between HPR6.6 and ZNF-Kaiso. This may have occurred if genes in one organism had

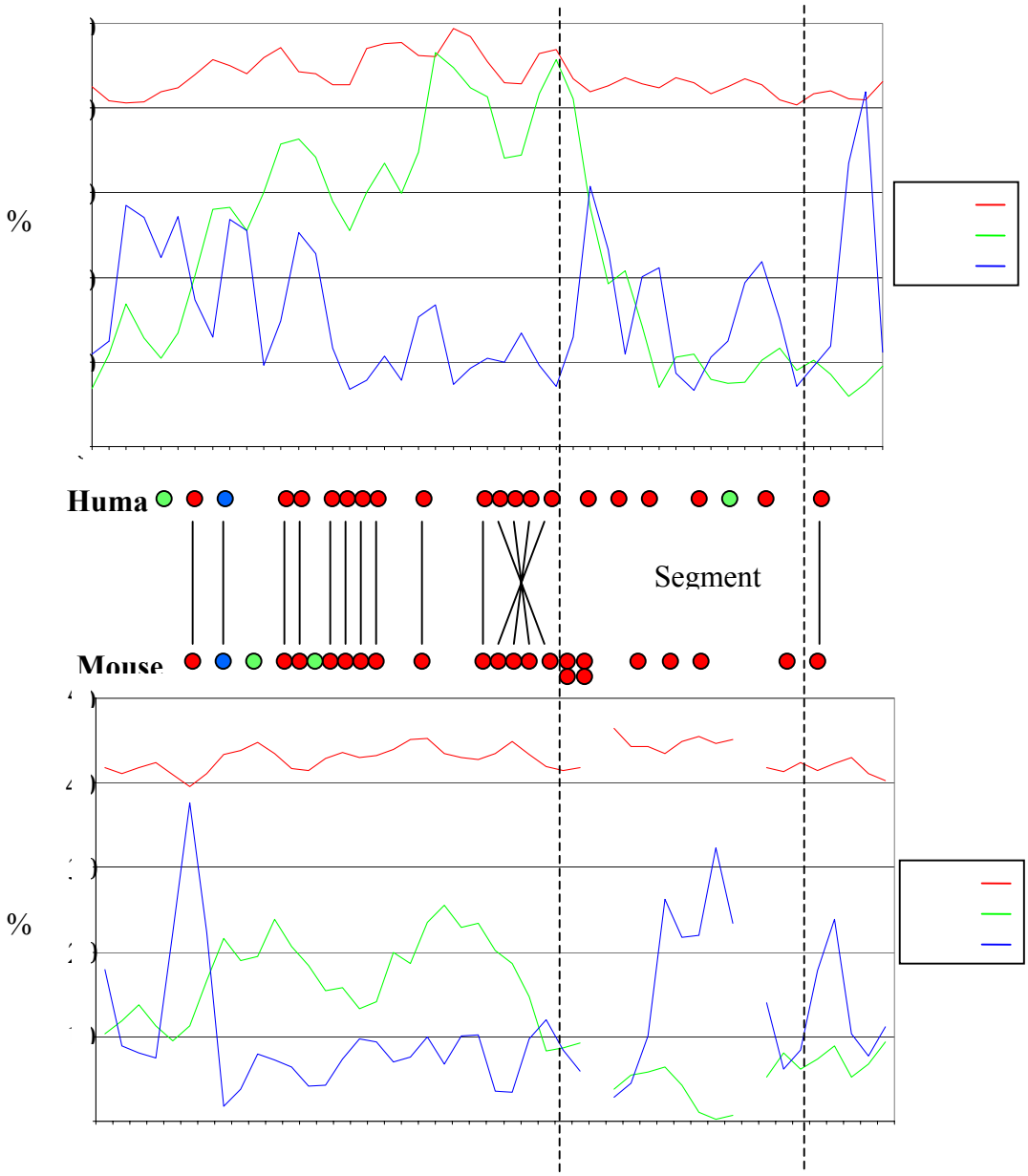
acquired new function. Analysis of segment 3 in other organisms will provide more data to further the understanding of the evolution of the region to support or reject this third explanation.

5.4 Comparison of the genome landscape in human and mouse

Finished sequence in both human and mouse was analysed for GC content and the content of SINEs and LINEs. A series of 50 kb sequence segments overlapping by 25 kb were generated and analysed for repeat content and GC content using RepeatMasker (Smit, AFA & Green, P. RepeatMasker at http://ftp.genome.washington.edu/RM/RepeatMasker.html) and the values plotted. The results for the regions in human and mouse were aligned based on the relative positions of the previously identified orthologous genes (see Section 5.3).

Comparison of the genome landscapes in human and mouse shows that there is a correlation between both GC content, and LINE and SINE (see Figure 5.12). The GC content remains above 40% in both human and mouse. Regions of relatively high SINE density were seen to correspond to gene rich regions. In general, the mouse sequence seems to have a lower repeat content, but this may be due to the reduced amount of information currently available for mouse repeats, so that some may remain unidentified. The SINE content both in human and mouse decreases in segment 3, the region where there is no apparent synteny between human and mouse. This correlates with a local change from a gene rich to a gene poor region in both species.

Figure 5.12: *(see over) Comparison of the genome landscape in human and mouse. A 50 kb window, moving in 25 kb increments was analysed for GC content (red line), SINE (green line) and LINE content (blue line), figures are given as a percentage. A break in the line represents a gap in the sequence. The gene content is also shown. Genes shown as red circles, predicted genes shown as blue circles and pseudogenes shown as green circles, orthologous genes are linked with a thin black line. The repeat content is generally lower for the mouse sequence which may reflect the level of understanding of repeat sequences in the two organisms. There appears to be a high SINE content in the region containing the majority of the orthologous pairs of genes, and a lower SINE content in the region containing no obvious orthologous pairs of genes (segment 3 indicated by a vertical dotted black lines).*


%
Huma
Segment
Mouse
%

5.7 Discussion

Two mouse-specific bacterial clone contigs containing 98 BAC clones covering 1.9 Mb between MAPR and ZNF-kaiso have been generated. The remaining gap has been sized at approximately 50 kb by fibre-fish. In this study, the generation of bacterial clone contigs across the syntenic portions of the mouse genome relied on the identification of a sufficient number of orthologous sequences in mouse cDNA resources. A second approach has been developed using the available mouse BAC end sequence and the human genome sequence (Simon Gregory, personal communications). Using this method the human sequence between HPR6.6 and ZNF-kaiso was compared using BLAST to a database of mouse BAC end sequences from the RPCI-23 and RPCI-24 libraries (generated by TIGR). The analysis identified the same set of BACs as was identified by the hybridisation method described in Section 5.2. By this time, fingerprints were available for all the identified BACs and so the contigs described in this chapter could be constructed without the need for identification of mouse-specific expressed sequences known to be orthologous to the region in human (a similar analysis has been carried out, comparing the whole of human genome to the mouse genome and is described in Section 7.1).

The mouse sequence generated from the minimum set of BAC clones between Mapr and ZNF-kaiso, was shown to contain twenty-three genes and two pseudogenes. Comparison of this region to the syntenic portion between HPR6.6 and ZNF-kaiso in mouse has identified 16 pairs of orthologous genes. The proximal portion of the region, between HPR6.6 and UPF3B in human, appears to be entirely syntenic with the proximal portion in mouse, between Mapr and Up3b. There is evidence of an

inversion of four genes in either human or mouse since the divergence from a common ancestor. The region between bG421I3.CX.2 and dJ525N14.CX.3 in human does not appear to be syntenic with the equivalent region in mouse between bM43O20.CX.8 and bM202F23.CX.3 as no orthologous pairs could be identified.

The region of human sequence studied between HPR6.6 and ZNF-Kaiso contains one gene with a 5' end that could not be confirmed by cDNA sequence (dJ1139I1.CX.1) and one predicted gene for which no cDNA could be detected in the available resources (dJ555N2.CX.1). In both cases, mouse sequence greater than 90% identical across predicted exons was identified (see Figure 5.17). Although the predicted transcribed regions are still to be confirmed with human cDNA sequence, the identification of the mouse orthologue for each gene will provide added confidence to the presence of a real exon or gene. This data will also enable further analyses to be carried out using a wider variety of mouse tissues.

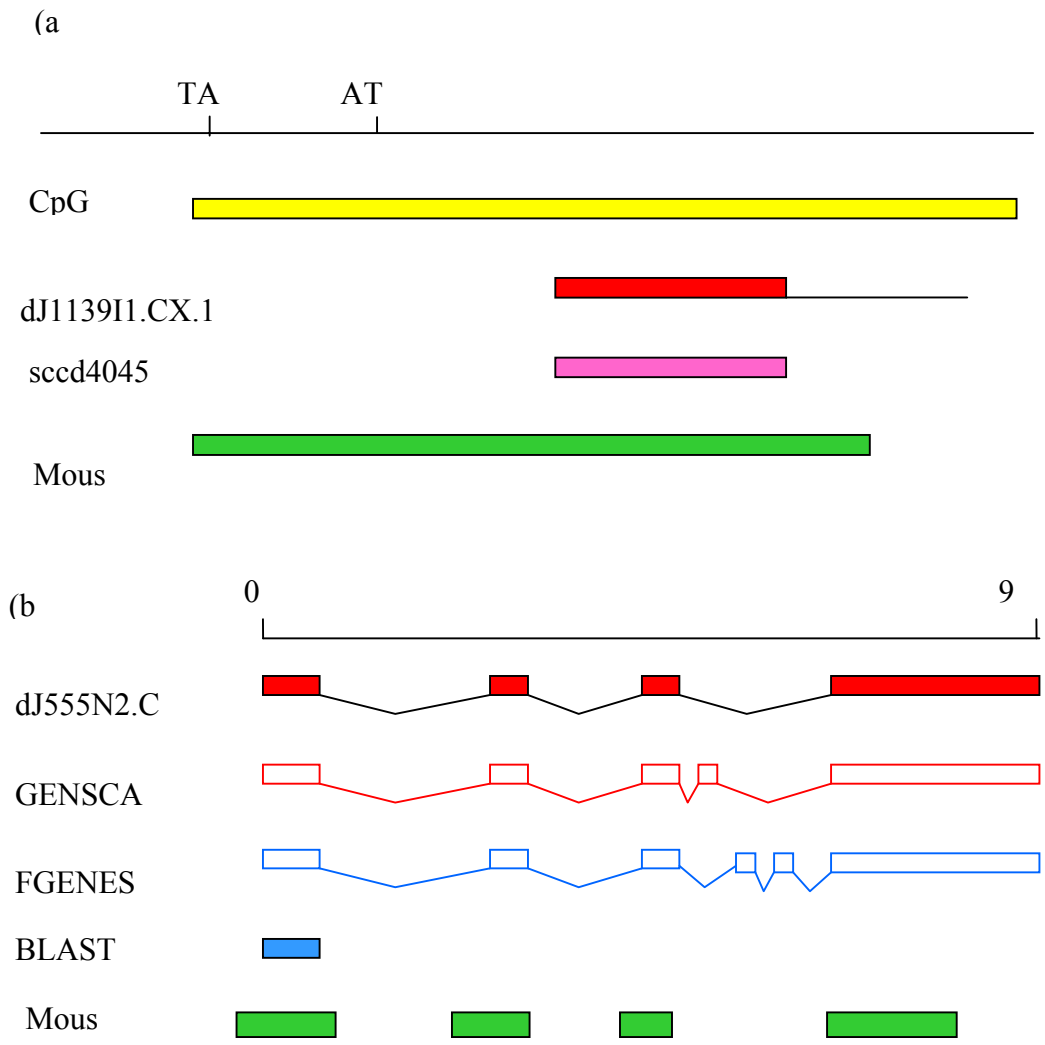


Figure 5.17: *Analysis of predicted and incomplete genes. (a) The 5' end of dJ1139I1.CX.1 (shown as red box). The most likely start site (ATG) is upstream of the confirmed cDNA sequence (pink box). A predicted CpG island is shown as a yellow box. Conserved sequence between human and mouse (green box) extends past the ATG site. (b) The predicted gene dJ555N2.CX.1 (red boxes linked by black lines) and the evidence supporting the prediction (GENSCAN prediction shown as open red boxes linked with red lines, FGENESH prediction shown as open blue boxes linked with blue lines, BLASTX match shown as filled blue box). Conserved sequences between human and mouse (filled green boxes) align with the predicted exons.*

One of the major challenges of comparative genome analysis is the identification and functional analysis of sequences that are conserved between different organisms. Conserved segments could be genes, regulatory elements or other biologically important features such as origins of replication. It is also possible that not all biologically important regions will necessarily show conservation at the primary sequence level, e.g. non-coding RNA genes and regulatory elements. Fourteen conserved sequences of unknown function were identified in the region by three different methods for comparing DNA, PIPMAKER, VISTA and BLAST. Evaluating whether any of these sequences are expressed in a wider variety of cDNA resources in both human and mouse may show that some of these conserved sequences are parts of novel genes.

The conserved sequences may also represent regulatory elements. It has been shown that regulatory elements are conserved in a number of species such as human, mouse and chicken (Gottgens, B*., et al.*, 2000). Experimental analysis is required to determine whether any of these conserved regions function as regulatory elements. For instance, DNA from each conserved region could be cloned into an expression vector in order to test for promoter or enhancer activity. Observing conservation of the regions in other species, such as other mammals or other vertebrates will increase the confidence that these regions are functional.

One of the major challenges of comparative genome analysis is the identification and functional analysis of sequences that are conserved between different organisms. Conserved segments could be genes, regulatory elements or other biologically important features such as origins of replication. It is also possible that not all biologically important regions will necessarily show conservation at the primary sequence level, e.g. non-coding RNA genes and regulatory elements. Fourteen conserved sequences of unknown function were identified in the region by three different methods for comparing DNA, PIPMAKER, VISTA and BLAST. Evaluating whether any of these sequences are expressed in a wider variety of cDNA resources in both human and mouse may show that some of these conserved sequences are parts of novel genes.

The conserved sequences may also represent regulatory elements. It has been shown that regulatory elements are conserved in a number of species such as human, mouse and chicken (Gottgens, B*., et al.*, 2000). Experimental analysis is required to determine whether any of these conserved regions function as regulatory elements. For instance, DNA from each conserved region could be cloned into an expression vector in order to test for promoter or enhancer activity. Observing conservation of the regions in other species, such as other mammals or other vertebrates will increase the confidence that these regions are functional.

5.8 Appendix

Table 5.4: *Information on clone names and links shown in Figure 5.8*

Link/Status	Accession	Clone Name
Link_bM100G16	AL450397	RP23-100G16
	AL450399	RP23-286I5
	AL589767	RP23-141L16
	AL451076	RP23-451076
Draft	AL589623	RP23-111C11
Draft	AL590629	RP23-202F23
Draft	AL123456	RP23-322E15
Finished	AL450391	RP23-38B5

5.6 Evaluation of whole genome shotgun (WGS)

The human sequence is nearing completion and the emphasis for large scale sequencing has shifted to generating sequence from genomes of other organisms. In order to produce sequence representing as much of the mouse genome as possible, an initial whole genome shotgun (WGS) has been carried out (data being generated by the Mouse Sequencing Consortium) and made available via trace repositories and BLAST databases on the WWW (http://trace.ensembl.org, http://www.ncbi.nlm.nih.gov/Traces/trace.cgi?, http://www.ncbi.nlm.nih.gov/genome/seq/MmBlast.html). There are currently almost 30 million reads available representing five genome equivalents. This is calculated by multiplying the number of reads by the average length of a sequence read, in this case 500 bp, and dividing by the genome size, in this case assumed to be $3x10^9$. The initial aim of the consortium was to produce three genome equivalents in WGS reads, but this figure was recently revised and extended to produce six genome equivalents. In parallel, a clone by clone sequencing effort is also underway, which in combination with the WGS sequencing will provide high quality finished sequence for the mouse genome by 2005.

In an attempt to evaluate to what extent whole genome shotgun from the mouse will aid human sequence annotation, an analysis of unfinished sequence was carried out to assess the contribution of the different levels of shotgun sequence depth (see Figure 5.16). In order to generate the equivalent of different amounts of coverage from WGS, a varying number of sequence reads representing a series of 'coverage equivalents' were used from four mouse BAC clones bM100G16, bM286I5, bM43O20 and

bM38B5. Table 5.3 shows the total number of reads available for each clone and the number of reads required for each 'coverage equivalent'.

Table 5.3: *Comparison of read number for various genome equivalents (RD = restriction digest)*

Clone Name	Size by RD (kb)	1/3x	1x	2x	3x	4x	6x
bM100G16	210	141	424	848	1272	1696	2544
bM286I5	230	152	458	916	1374	1832	2748
bM43O20	175	117	352	704	1056	1408	2112
bM38B5	180	118	356	712	1068	1424	2136

The figures were calculated based on an average read length of 500 bp. Analysis of the finished sequence for the four mouse clones showed 73 exons were present from 16 orthologous genes.

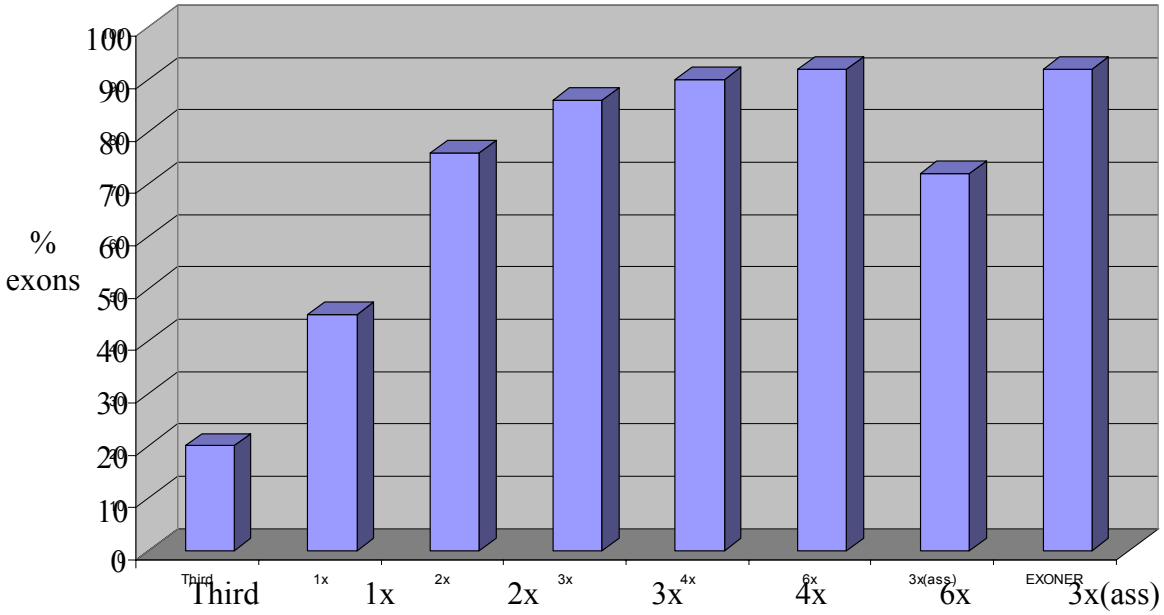


Figure 5.16: *Evaluation of whole genome shotgun. Percentage of matches to human exons in the region (vertical axis) at increasing amounts of coverage of mouse sequence (horizontal axis). 86% of all human exons are present in three genome equivalents of mouse sequence, and 92% are present in six genome equivalents*

(3x(ass) = 3x assembled). The final bar represents the amount of exons hit by mouse sequence traces currently positioned in the region by EXONERATE.

The sequence reads from each ‘coverage equivalent’ were then compared using BLAST to the total number of exons covered by the four clones. The results are shown in Figure 5.16. As the number of reads increases the number of exons present in the mouse sequence increases. This information would suggest that the original target of three genome equivalents of whole genome shotgun data would contain 86% all exons, whereas the revised target of six genome equivalents would contain 92% of all exons. Even though the original reads from the bacterial clones were randomised initially, they were still constrained to lie within a single clone and so cannot be considered precisely comparable. However, analysis of the whole genome shotgun data currently available (approximately five genome equivalents) using EXONERATE, a program that aligns mouse sequence reads to human sequence (courtesy of Michelle Clamp), shows that 92% of the exons are present (see Figure 5.16). A very similar set of exons were identified by both methods, EXONERATE identifying six exons not observed in the sequence generated by the clone based evaluation, which in turn identified six that were not detected by EXONERATE.