Chapter Four - Genomic landscape of the mouse genomic region equivalent to human Xq22-q23

4.1 Introduction

Much of our understanding of many areas of biology comes from the study of other organisms. Various organisms have been chosen to study different aspects of biology such as genetics and physiology, based on features including their experimental tractability and relationships to other organisms of study (including humans). For instance, much of our understanding of multi-cellular organism development has arisen from studies in the fly.

Particularly well-studied organisms include the mouse, rat, fly, worm and fish, in addition to more distantly related organisms such as plants, yeast, urchins and the sea-squirt. As genome sequencing technologies have advanced, the number of organisms for which genome sequence data are available, or are being generated, has expanded considerably (Ureta-Vidal *et al.*, 2003).

As mentioned in Chapter 3, the human Xq22 region has undergone considerable rearrangements in its evolutionary history, involving multiple gene duplications. For several of the genes such as the thymosin-beta paralogues, levels of sequence similarity were high even in intronic regions. This suggested that the duplications may be relatively recent.

This prompted the mapping and annotation of the orthologous genomic region in mouse, in order to explore the extent of paralogy within the mouse region, and to attempt to determine whether some of the Xq22 paralogy was a representation of duplications occurring relatively recently in the evolution of the human X chromosome.

It was estimated that the mouse genome would provide an appropriate comparison in order to ascertain if a similarly high level of paralogy was present, or, if some of the gene duplications were indeed more recent evolutionary events, as genomic comparative analysis to date in the mouse has demonstrated relatively low levels of homology between intronic sequences between the species. Furthermore, the mouse X chromosome has been shown to be well conserved in terms of gene content with respect to the human X chromosome, although many rearrangements have occurred within the chromosome. The conserved blocks are depicted in Figure 4-1.

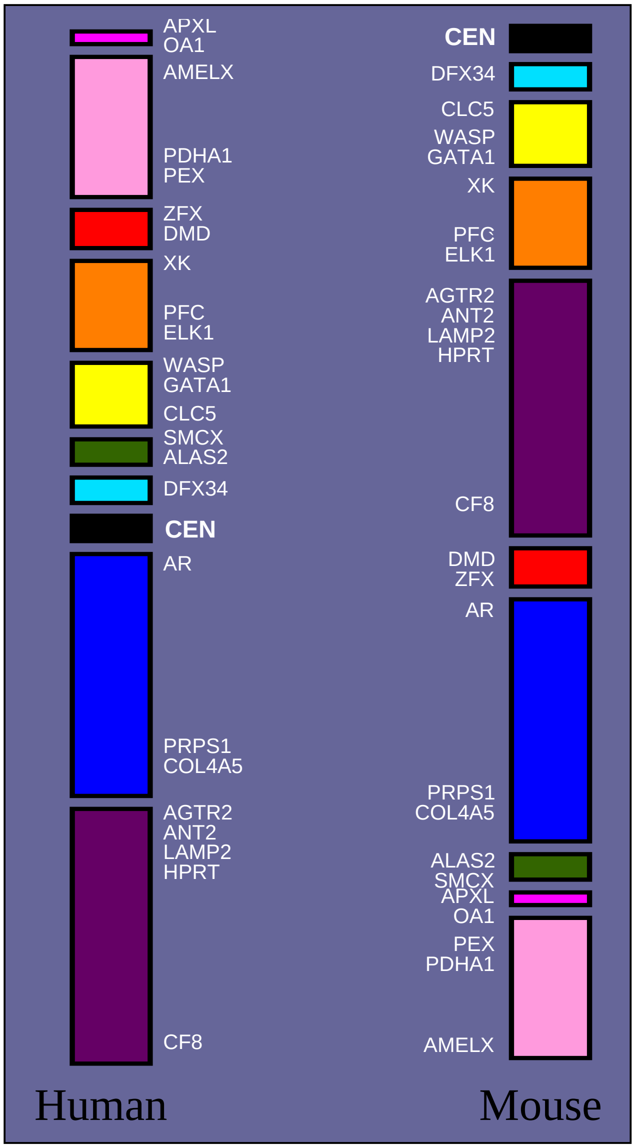


Figure 4-1 Figure illustrating the mouse X chromosome showing blocks with conserved synteny to the human X chromosome (reproduced from MRC Harwell website).

Thus, in order to better understand the evolution of the region, this chapter describes efforts to produce a sequence-ready BAC contig of the region of the mouse genome with shared synteny with human Xq22-q23, and analysis of the genome sequence produced from a tiling-path of BACs from the contig.

During the course of the work undertaken in this chapter, the mouse genome project advanced rapidly due to the framework provided by the draft human genome sequence (Gregory *et al.*, 2002), and a whole-genome shotgun approach generated a draft genome sequence for the mouse (Waterston *et al.*, 2002). This chapter also discusses how these resources were used to expedite production of BAC contigs.

Although the mouse genome shares large regions of shared synteny with the human genome, both organisms' genomes have undergone rearrangements. The X chromosome is particularly conserved between the two species with respect to gene content. Ohno's Law postulates that the X chromosome is protected from rearrangements involving other chromosomes owing to the dosage imbalances that might be created in gene products. This chapter examines species-specific features of the genome regions studied, discovered through analyses of the sequence generated, and conservation of gene content and order between the two species.

4.2 Assembly of a sequence-ready BAC contig for mouse X E3-F2

The aim of this section of work was to produce a sequence-ready BAC tiling path of the *Mus musculus* X chromosome E3-F2 region, which contains genes orthologous to those in human Xq22-q23. Genomic sequence produced from these BACs would then be used to examine the extent of conservation between human and mouse at high resolution. When work began, the following large-scale projects were underway within the mouse genome mapping community (selected references given):

- BAC end sequencing - TIGR, Rockville, MD. RPCI-23 and RPCI-24 libraries (see http://www.tigr.org/tdb/bac_ends/mouse/bac_end_intro.shtml)
- BAC restriction fingerprinting – Genome Sequencing Centre, British Columbia Cancer Research Centre, Vancouver.
- Contig generation (FPC) – Mouse Genome Sequencing Consortium
- Mouse BAC-end vs. human genome BLAST searching – Carol Scott, Wellcome Trust Sanger Institute
- Genetic mapping and EST/gene-based RH mapping - (Dietrich *et al.*, 1996), (Hudson *et al.*, 2001), (Avner *et al.*, 2001)
- WGS generation and assembly – Phusion (Wellcome Trust Sanger Institute) and Arachne (MIT, Cambridge) algorithms, assemblies available via NCBI.

In order to make efficient use of data generated from these various sources, several strategies were adopted for mapping the region, and were adapted as the mouse genome mapping project matured and more information became available.

Initially, two approaches were used to anchor mouse BAC contigs already constructed to the region:

- A gene-based STS hybridisation approach to identify BACs containing mouse orthologues of human Xq22-q23 genes
- Analysis of results from alignments of mouse BAC-end sequences against the human X chromosome sequence

Sequences from genes across the human Xq22-q23 region were used to identify mouse mRNA or EST sequences present in GenBank using BLAST. Where several matches were obtained, the highest-scoring match was retained. In addition, curated information identifying the likely mouse orthologues from LocusLink, Mouse Genome Database (MGD – part of Mouse Genome Informatics at The Jackson Laboratory, Maine) and the scientific literature was utilised where available. This resulted in the identification of potential mouse orthologues of ten human genes, spanned across the Xq22 region (Table 4-1).

These mouse sequences were used for the design of STS primer pairs, using RepeatMasker to avoid repetitive sequences. All STS primer pairs used in this chapter were pre-screened to establish optimal reaction conditions. STS pre-screens were performed on mouse genomic DNA and $T_{0.1}E$. Pre-screens were performed using three different primer annealing temperatures ($55^{\circ}C$, $60^{\circ}C$ and $65^{\circ}C$) to determine the cycling parameters that give a visible and specific DNA product.

These primers were used to generate ten radio-labelled mouse DNA probes (see Chapter 2), which were then pooled and used to screen the RPCI-23 mouse BAC library (see Chapter 2). Positive clones were confirmed and assigned to individual STSs by colony PCR using the same primers (see Chapter 2). An example of the results obtained is shown in Figure 4-2. In this way, 141 clones were identified from screening with 10 different probes. These clones were located in contigs assembled in the mouse genome mapping project FPC database, enabling those contigs to be anchored to the mouse X E3-F2 region.

In parallel, alignments of mouse RPCI-23 and RPCI-24 BAC end sequences (TIGR) against human genomic sequence were analysed to find BACs with matches to human Xq22-q23 (BLAST data were kindly provided by Carol Scott, Wellcome Trust Sanger Institute). The relevant BAC clones were located within the mouse genome mapping project FPC database contigs, as described above.

The combination of these two approaches resulted in a first-generation BAC map of the mouse X E3-F2 region as shown in Figure 4-3.

Human gene name	Mouse gene name	Mouse sequence used for primer design	STS designed and primer sequences	Positive RPCI-23 BAC clones
dJ79P11.1	Bex2	AF097439	stSG136026 TTCTGGTGTCACTTGTTTCCC; TATACTGAGCATCTTCCCATGC	1A3, 29G15, 124C21, 149K3, 172E22, 216I6, 255O6, 260L22, 260L23, 262B17, 268J10, 306G4, 308M8, 313L11, 351C11, 395D23, 396F19, 403A16, 410F19, 431J15, 431N2, 465A15
NXF2 (cU19D8.CX.1)	(Blast hit)	AK005772	stSG136028 AACCAGCATGTGTTTAGCCC; GACCTCTCTTTGGATTCCTGG	8P4, 17M23, 95D9, 96H8, 151O17, 156D7, 183F23, 197H20, 202L24, 250F8, 258J15, 278E22, 340P1, 346I8, 376N8, 394N17, 410F10, 426B2, 441N13, 451E7, 452M10, 456D4
ALEX2 (cV602D8.CX.1)	(Blast hit)	AK014329	stSG136029 GTCACCAGCTTTAAGCTGAACC; AGCTGAGTAGGCCATTCACG	76B8, 121E2, 185C13, 195N13, 223K14, 272J1, 316A19
KIAA0443 (dJ769N13.1)	(Blast hit)	AK014109	stSG136031 ATGCTGGTGGCAATTCTACC; CGAGAACAACATTTAGAAGGGC	27F12, 86I9, 94I24, 249N10, 297L20, 313H21, 376N8
dJ341D10.2	(Blast hit)	AK016872	stSG136032 ATGGACTTTCCACCTGAACG; CCCTGTTGGTCTAAGGCTCA	17C4, 30G6, 116B3, 162B19, 178F23, 182M17, 182N4, 202M20, 219H14, 321B8, 323D8, 400M23, 402E10, 410M19
Pp21-homologue (pp21h)	(Blast hit)	AK002214	stSG136033 AACAAAATGAGCTTCTGATGGG;TGGCAAATACAAATAAGCAGAA	22B16, 22C15, 79P22, 90I22, 105K2, 105O4, 129A14, 132M9, 144O5, 145O13, 147N21, 158N16, 164J10, 168D7, 246O13, 253P12, 284L23, 308H24, 318O9, 325J24, 374B1, 378G5, 421O13, 422N5, 431N2, 443H15, 451H15
IRS4	Irs4	AF087797	stSG136034 GTTGATGCGTTAGTTGGTATGC; GCTAATGTTTTCGCAAAGGC	218M1, 241A1, 262F10, 262N4, 304F16, 413I10, 415G1, 446H18
IL1RAPL2 (Exon 2)	Tigirr-1	AF284437	stSG136291 TGAACAATGAAGCTGCCACT; TTTCTTTTTGACACCATCTTCAA	38M15, 70H21, 85B20, 108O1, 219J3, 246C20, 252F20, 266D19, 290H23, 394A22, 431J23, 435B4, 458L22
COL4A5	Col4A5	Z35168	stSG136970 GCCAAGCCCTAGCCTCTC; ACAGTGGCCAGCCAAAAG	3H7, 218M1, 241A1, 262F10, 262N4, 304F16, 328B7, 328E8, 413I10, 415G1, 446H18
IL1RAPL2 (Final exon)	Tigirr-1	AF284437	stSG136971 CGAACTGGAAAGCAGACTCC; ATTTGCTGCTTTTGGGTCC	5P1, 58K24, 63D17, 101H10, 102O21, 203N11, 204A1, 385K18, 389N3, 434L24

Table 4-1 Table listing human genes, corresponding mouse gene name (or indicated where the mouse sequence represents a BLAST match), the potential mouse orthologous sequences used for design of probes for screening the mouse RPCI-23 BAC library, STS name and primers (Sense; Antisense) and positive clones from screening of the RPCI-23 BAC library (PCR verified).

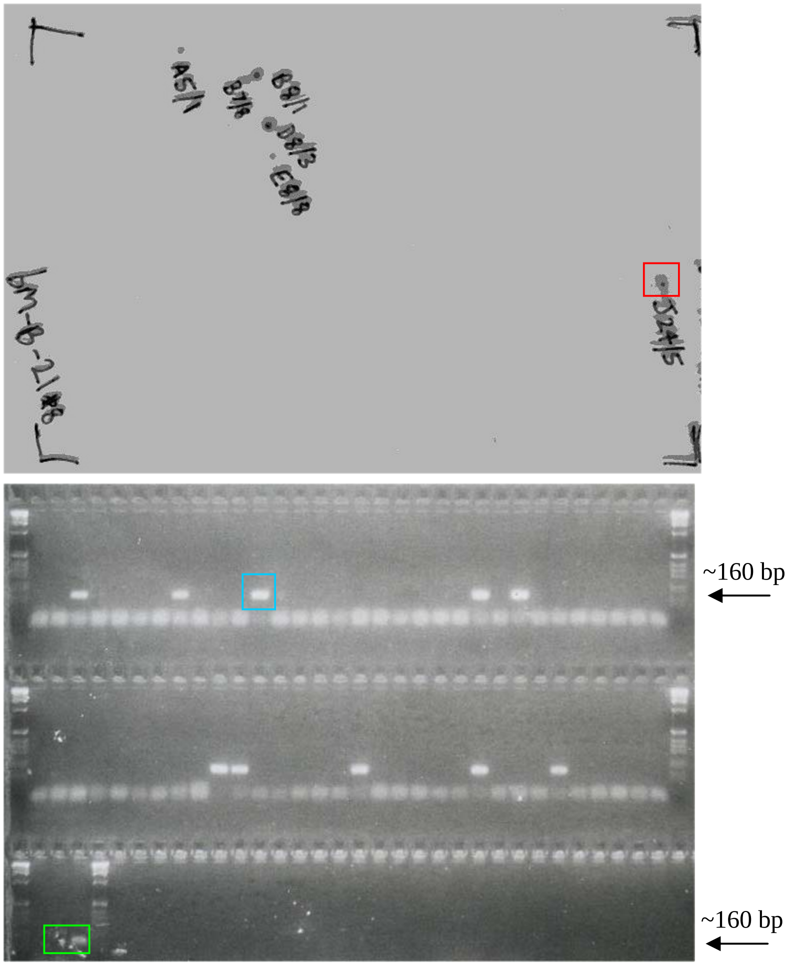


Figure 4-2 Diagram illustrating the STS-based hybridisation strategy used to isolate mouse RPCI-23 BAC clones. The upper image shows an autoradiograph of a mouse BAC filter following hybridisation of pooled radiolabelled STS products and washing. The red box highlights a positive signal for BAC bM325J24. The lower section shows colony PCR results using primers for stSG136033. The blue box highlights a positive result for BAC bM325J24. The green box highlights the $T_{0.1}E$ and genomic DNA controls.

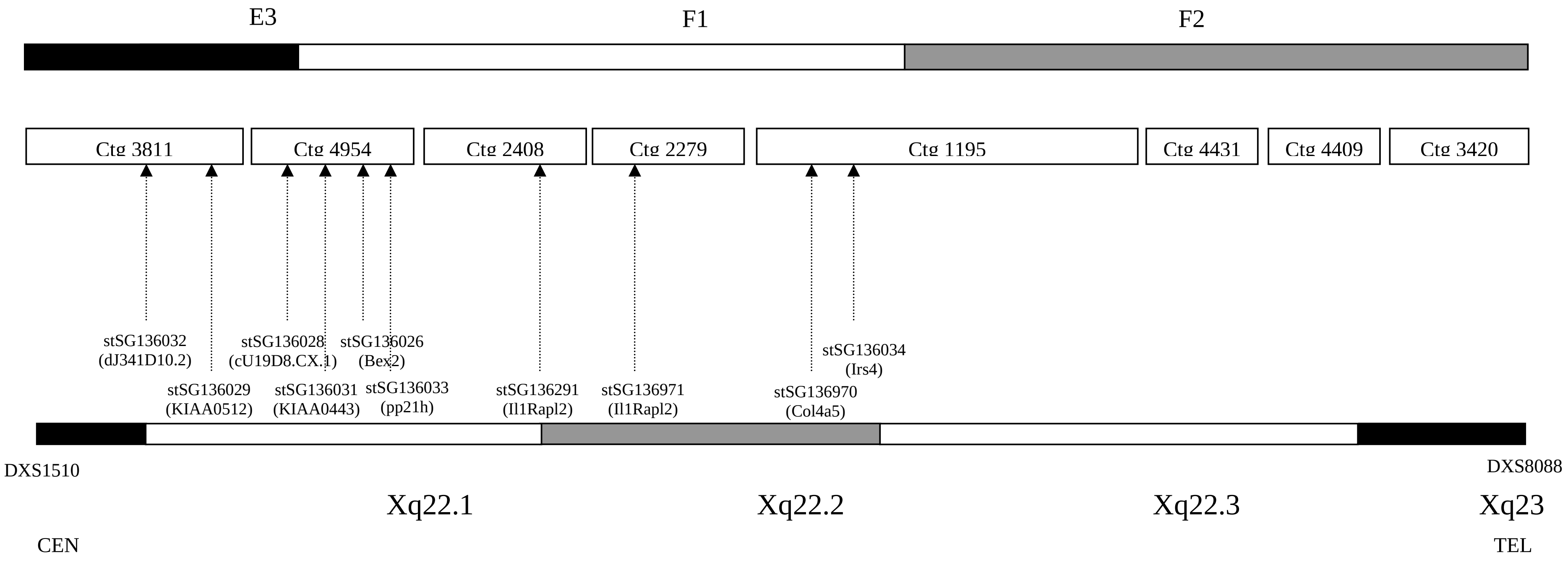


Figure 4-3 A first-generation BAC contig map of the mouse X E3-F2 region. The mouse G-banding pattern of the region is shown at the top of the figure. Initial BAC contigs are shown as open boxes. The approximate positions of BACs positive with mouse gene probes are arrowed. The STS and gene names are drawn beside the G-banded ideogram of the human Xq22 region to indicate the locations of the human orthologous genes.

At this stage the map comprised 8 contigs. Further efforts concentrated on closing the gaps between contigs and on estimating the size of any unclosed gaps using fibre-FISH.

Attempts were made to close gaps using fingerprint information and tools within FPC. This approach used shared bands between BAC fingerprints to determine statistical likelihood of clone overlaps. In this way, fingerprints from BACs at the ends of the contigs were compared to other contigs within the database to identify potential joins. Whilst initial contig assembly did not detect any further contig overlaps, relaxing the stringency criteria used to assess fingerprint overlaps allowed more sensitive searches. This approach can be adopted when initial BAC contig mapping has provided information on contig position, thus contigs which are neighbours would be more likely to represent a true overlap. This approach closed a gap between Ctg4431 and Ctg4409.

Following attempts to ascertain contig overlaps using fingerprint data, efforts were made to close remaining contig gaps, utilising recently generated mouse whole-genome shotgun (WGS) assemblies. End sequences of BACs at the ends of the contigs were used to search the mouse WGS scaffolds by BLAST. WGS scaffolds were used to search the mouse BAC-end sequences (TIGR web site), and the resulting matches and the orientation of the BAC-end sequence alignments were used to ascertain if BACs were likely to overlap. When overlaps were identified, they were then confirmed by colony-PCR. This strategy is outlined in Figure 4-4. In this way, five contig gaps were closed (Figure 4-5).

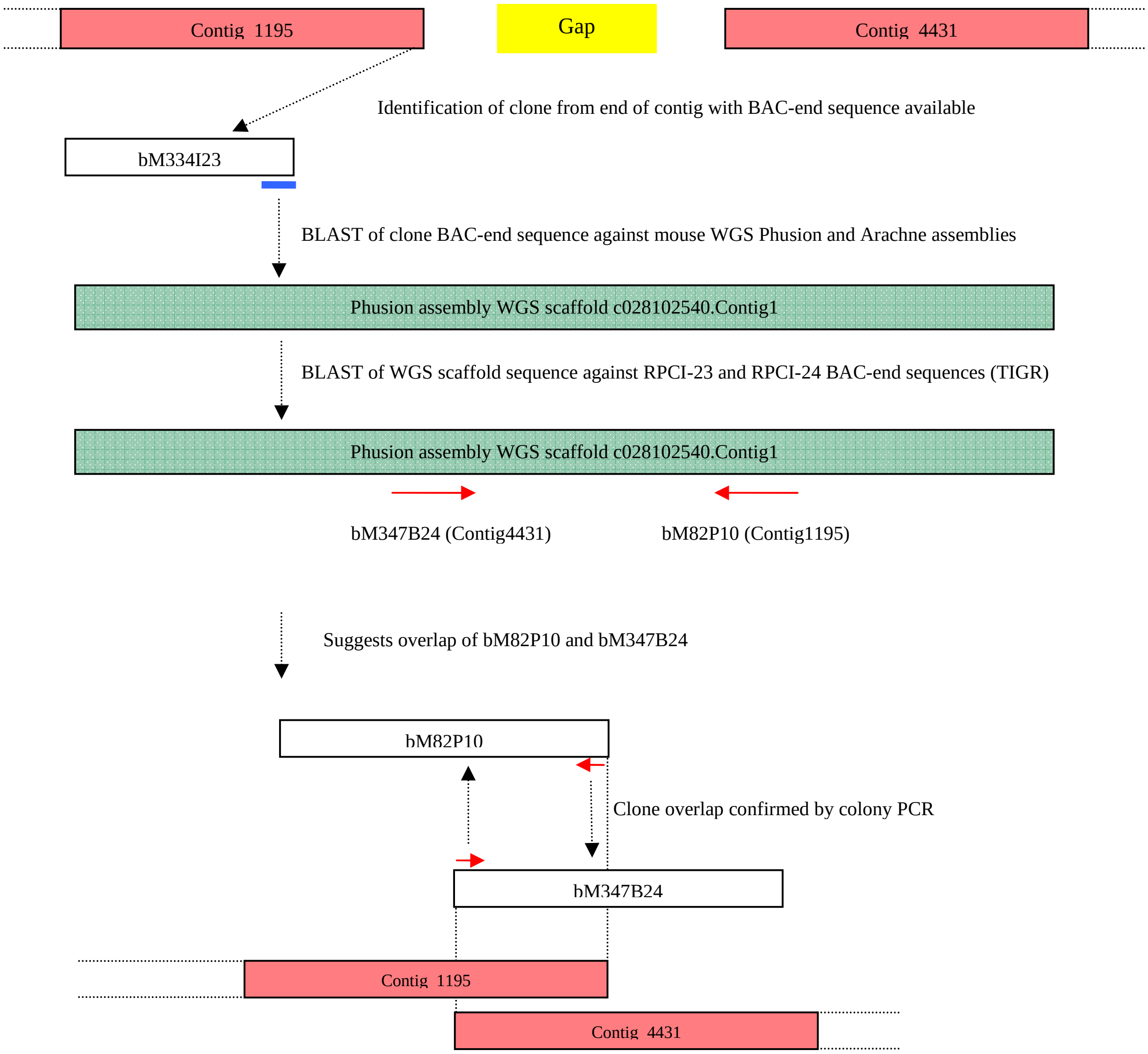

Contig 1195
Gap
Contig 4431
Identification of clone from end of contig with BAC-end sequence available
bM334I23
BLAST of clone BAC-end sequence against mouse WGS Phusion and Arachne assemblies
Phusion assembly WGS scaffold c028102540.Contig1
BLAST of WGS scaffold sequence against RPCI-23 and RPCI-24 BAC-end sequences (TIGR)
Phusion assembly WGS scaffold c028102540.Contig1
bM347B24 (Contig4431)
bM82P10 (Contig1195)
Suggests overlap of bM82P10 and bM347B24
bM82P10
Clone overlap confirmed by colony PCR
bM347B24
Contig 1195
Contig 4431

Figure 4-4 Example of detection of contig overlaps utilising WGS assemblies. This example illustrates a contig overlap undetected by BAC fingerprint analysis. Mouse BAC contigs are shown as pale red boxes, the WGS assembly contig is shown as a green box and mouse BAC clones are open boxes. Red arrows show the orientation of BAC-end sequence matches.

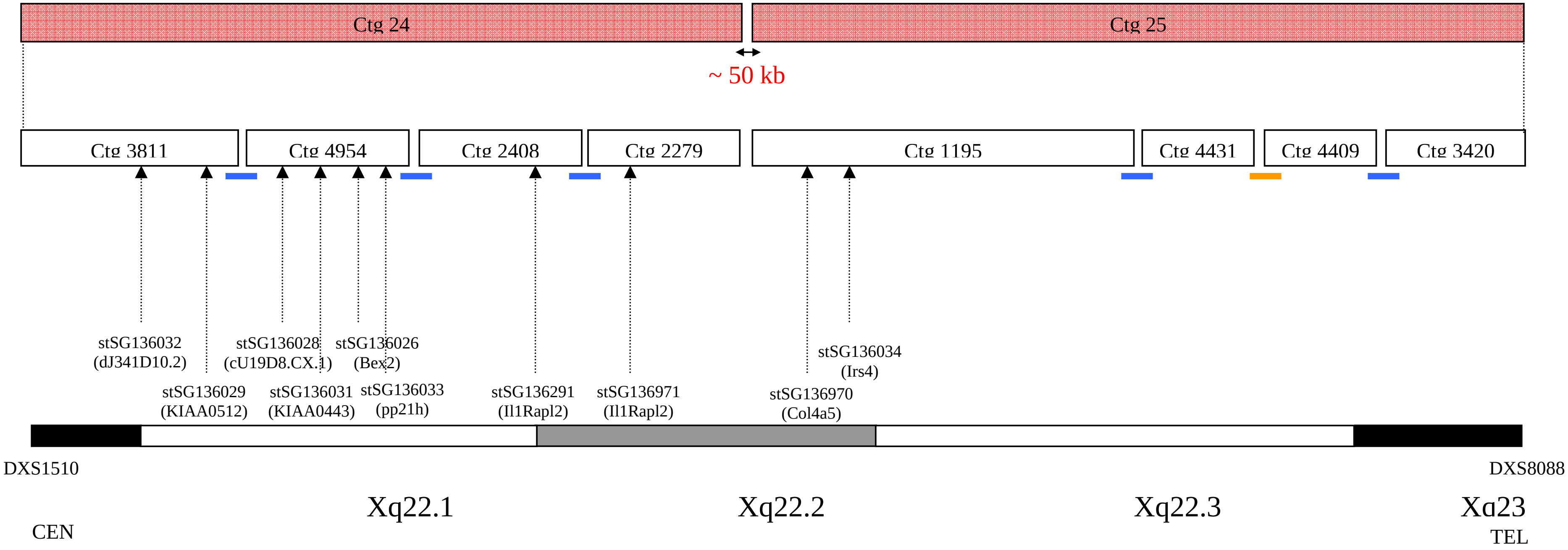


Figure 4-5 Finalised BAC contig map of the mouse X E3-F2 region. Initial BAC contigs are shown as open boxes. The approximate positions of BACs positive with mouse gene probes are arrowed. The STS and gene names are drawn beside the G-banded ideogram of the human Xq22 region to indicate the locations of the human orthologous genes. Gaps closed using WGS data are shown by blue bars and gaps closed by fingerprint data by orange bars. The size of the remaining gap is in red.

Combining these strategies, 6 contig overlaps in total were detected and verified. The remaining gap was sized by fibre-FISH, using clones from either side of the gap between Ctg2279 and Ctg1195. Clones were grown and their BAC DNA isolated. FISH probes were derived by nick-translation and hybridised to mouse DNA fibres prepared from a spleen cell primary culture (see Chapter 2). Results are shown in Figure 4-6.

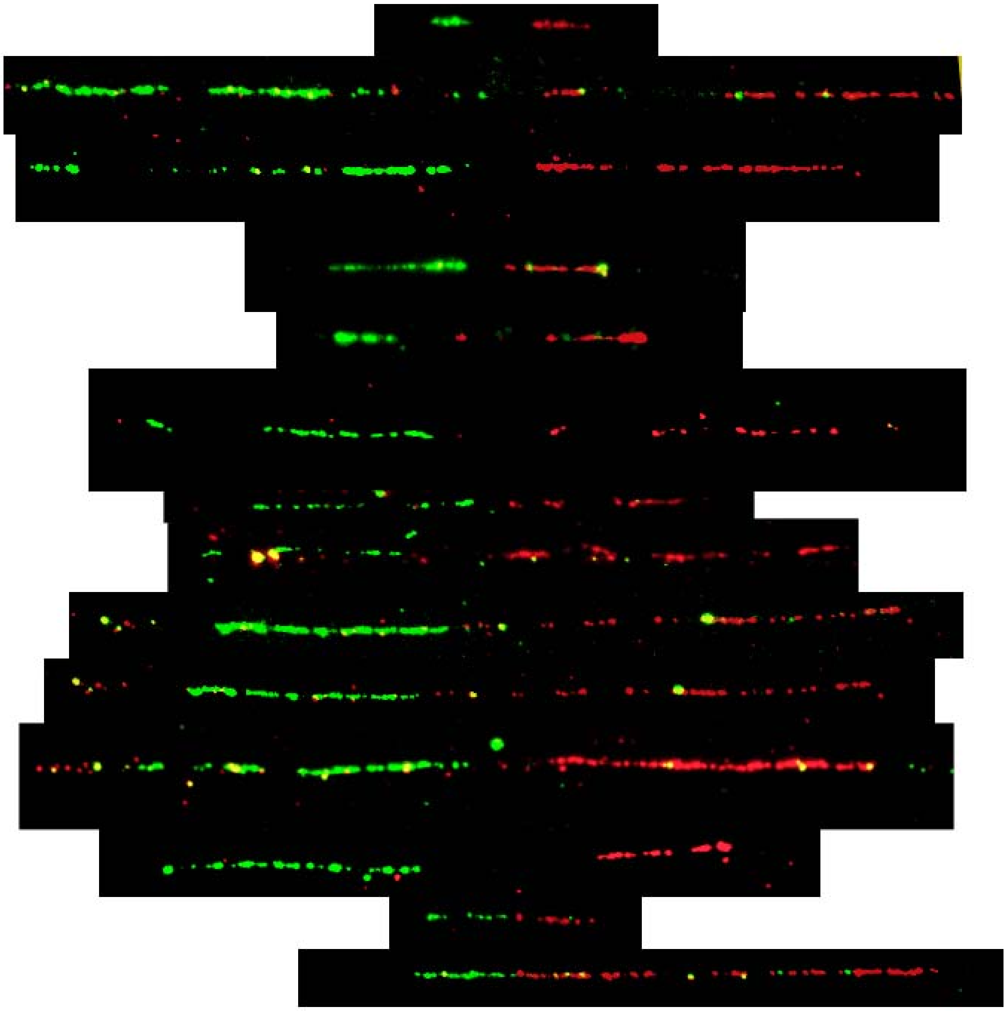
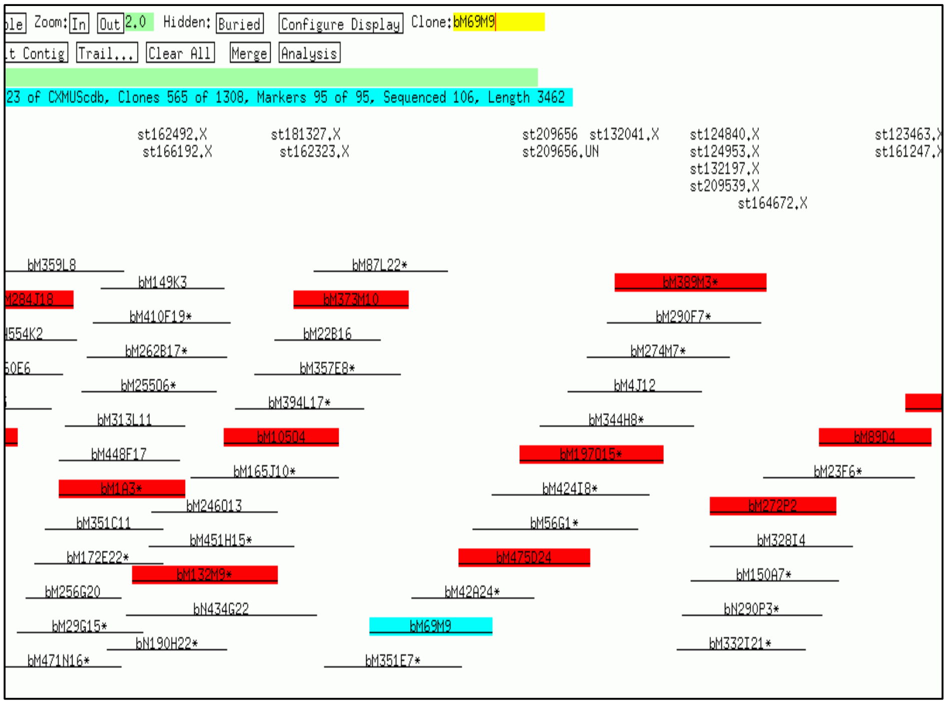


Figure 4-6 Results of fibre-FISH of bM149O3 (Ctg3811/2279 - red) and bM62F12 (Ctg1195 – green). A composite of images captured from 14 separate fibres is shown. A gap of ~ 50kb can be estimated (assuming ~ 150 kb per clone).

Efforts to close the remaining gap were then carried out by the mouse X chromosome mapping group (Glen Threadgold - Wellcome Trust Sanger Institute) as part of their effort to map the entire chromosome.

Gap closures resulted in two sequence-ready BAC contigs covering the mouse X chromosome E3-F2 region, renamed Contig 24 and Contig 25. A tiling path of BAC clones was chosen based on shared fingerprint bands – 68 clones were chosen for Contig 24 and 31 clones for Contig 25. These 99 BACs were picked from the RPCI-23 library (or were ordered if from the RPCI-24 library), grown in 2xTY and submitted to the Sanger Centre sequencing pipeline. Based on sequence available at the time of writing, the size of the region spanned by both contigs was approximately 14.3 Mb. The size of contig 24 was approximately 9.5 Mb, and contig 25 approximately 4.8 Mb.

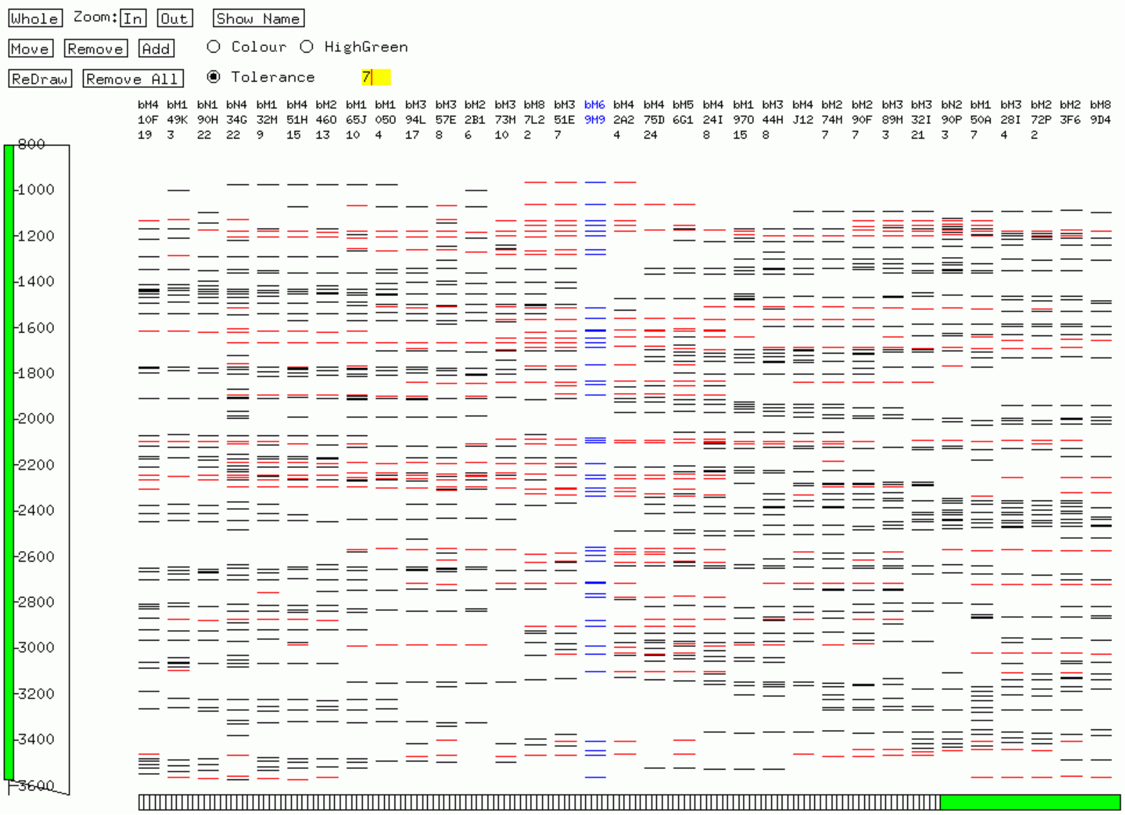


Zoom: 2.0

Clone:bM69M9

23 of CXMUScdb, Clones 565 of 1308, Markers 95 of 95, Sequenced 106, Length 3462

Figure 4-7 Diagram illustrating a section of contig 24 in FPC, illustrating a region of the tiling path of clones chosen. Clone bM69M9 is highlighted at the centre, with adjacent clones selected for sequencing also highlighted.



○ Colour ○ HighGreen

● Tolerance 7

Figure 4-8 Diagram illustrating overlapping clones from contig 24 based on fingerprint data. Bands produced by restriction digest for each clone are displayed vertically in FPC. Clone bM69M9 is highlighted in blue, with neighbouring clones in the contig to left and right. Red bands denote are those shared by the neighbouring clones for the cutoff parameters chosen (Chapter 2).

During sequencing of the region, gaps in the clone tiling path became apparent. Many clones were noted to be rich in repeats causing difficulties in the finishing process (Darren Grafham– personal communication). Some clones were also found to contain repeats that were present in other clones. These repeats could cause false joins within the region by generating fingerprint bands of similar sizes from non-overlapping clones. Additional clones were picked to close sequence gaps (Glen Threadgold, Wellcome Trust Sanger Institute). Table 4-2 lists the sequence clones and status of the region at the time of the study.

contig	seqctg	clone (bM)	accession	status	contig	seqctg	clone (bM)	accession	status
24	1	253D13	AL713898	analysed			bN408L19		sel. Seq
	1	96D6	BX088546	analysed			bN492P10		auto pre-fin
		124P2	BX088537	pre-fin		n/a	5P1	AL672067	analysed
	2	40P1	AL713871	analysed			389N3		shotgun
	2	293N20	AL672096	analysed		n/a	228A20	AL732419	analysed
		13E2		shotgun		n/a	96E23	AL714027	analysed
	3	193L14	AL691418	analysed		n/a	219K12	AL691424	analysed
	3	274A14	AL713982	analysed			130F16		shotgun
	3	305L4	AL713972	analysed		n/a	35I10	AL731648	analysed
	3	373J8	AL713897	analysed		n/a	343M4	AL672243	analysed
		20E14		cleared lib		n/a	351A10	AL672270	analysed
	4	434O7	AL713979	analysed			16O8		cleared lib
	4	60A20	AL672052	analysed		n/a	305F20	AL714021	analysed
		4K22		shotgun		n/a	137E3	AL672297	analysed
	5	161C9	AL672214	analysed		n/a	440B21	AL683809	analysed
	5	330I16	AL713863	analysed			290J11		shotgun
		395D17		shotgun		n/a	149O3	AL672306	analysed
	6	21A16	AL691421	analysed			typeIII		
		78G10		cleared lib	25	n/a	264D18	AL691493	analysed
	7	182N4	AL671915	analysed		n/a	244C21	AL713983	analysed
	7	162B19	AL672215	analysed		n/a	328E8	AL671856	analysed
	7	91G19	AL672064	analysed			typeII		
	7	26D22	BX004852	analysed		n/a	232B3	AL671983	analysed
		195N13		assembly			typeII		
	8	316A19	AL772348	analysed		n/a	294O1	AL671916	analysed
		bN374B8		cleared lib			457L22		shotgun
		65A22	AL672063	ass fin		n/a	161L11	AL731672	analysed
		bN142A19	AL954643	top-up			412I2	BX005213	ass fin
	9	460B8	AL731676	analysed		n/a	39H12	AL731678	analysed
		typeII				n/a	340M18	AL731674	analysed
	10	65C22	AL954640	analysed		n/a	71M18	AL731548	analysed
	10	250F8	AL671911	analysed		n/a	330B20	AL713920	analysed
	10	376N8	AL954646	auto pre-fin			346N16		top-up
	10	94I24	AL683822	analysed			252N4		shotgun
		160E6		streaked		n/a	45O6	AL691499	analysed
	11	1A3	AL671914	analysed			462G16		sel seq
	11	132M9	AL772180	analysed		n/a	48J18	AL713894	analysed
	11	105O4	AL671493	analysed		n/a	159H8	AL713978	analysed
	11	373M10	AL954818	analysed			typeII		
	11	69M9	AL672068	analysed		n/a	367H15	AL713861	analysed
	11	475D24	AL954381	analysed		n/a	140L6	AL731701	analysed
	11	197O15	AL671887	analysed		n/a	142G13	AL713986	analysed
	11	389M3	AL672008	analysed		n/a	18H24	AL808028	analysed
	11	272P2	AL954296	analysed		n/a	319K12	AL807791	analysed
	11	89D4	AL672275	analysed			bN422L8		auto pre-fin
		447K12		shotgun			136N12		shotgun
	n/a	287A19	AL672299	analysed		n/a	359L15	AL807753	analysed
	n/a	462C12	AL683888	analysed		n/a	377K9	AL672267	analysed
	n/a	85B20	AL691422	analysed		n/a	117F22	AL928629	QC
		typeII				n/a	185L10	AL672091	analysed
	n/a	48B17	AL672286	analysed		n/a	405D18	AL732456	analysed
	n/a	150J13	AL831759	analysed		n/a	bN69K11	BX088729	ass fin
	n/a	149B17	AL672205	analysed					

Table 4-2 Clones selected for sequencing and status of the region at the time of the completion of the study. Type II gaps (where there is no clone sequence but the gap is covered by a clone) are noted in yellow, unfinished sequences in grey, and the type III gap (a contig gap) in red. “Ass Fin” – assigned to finisher, “streaked” – clone is streaked, “shotgun” – clone is in shotgun sequencing, “cleared lib” – clone is cleared for library preparation, “pre-fin” – clone is in pre-finishing, “assembly” – shotgun reads are in assembly, “sel seq” – selected for sequencing, “top-up” – further shotgun sequencing is being performed, “QC” – clone is finished and being checked. Clones are mainly from the RPCI-23 library (prefix “bM”), unless noted otherwise (prefix “bN” – RPCI-24 library).

4.3 Identification of genes and their structures using sequence analysis

Finished sequences were analysed, clone-by-clone, for repeats and BLAST matches to mRNA and protein sequences as described in Chapter 3 (analysis by Stephen Keenan, Wellcome Trust Sanger Institute). Separate sequences were then linked to form sequence contigs and entered into an ACeDb database (kindly performed by Carol Scott, Wellcome Trust Sanger Institute). A total of 71 finished clone sequences were analysed in this manner.

Gene annotation efforts were focussed on a region of approximately 6 Mb bounded by clones bM253D13 (Cen) and bM89D4 (Tel). From preliminary assessment of the sequence analysis and of analysis of unfinished sequences during sequencing of clones from the region (using NIX (RFCGR)), this region was expected to be orthologous to the region of human Xq22 containing the majority of paralogous loci identified in Chapter 3 and described in Chapter 5.

The sequence analysis results for thirty clones were then studied as described in Chapter 3 to identify and annotate genes. Ten sequence gaps were present within this region. Genes were annotated on the basis of identical mouse mRNA matches (loci denoted as "GD_mRNA") or on the basis of similarity to mouse or human mRNA or protein sequences (loci denoted as "GD_supported"). Selected pseudogenes (not all were annotated due to time constraints) were also annotated (loci denoted as "Pseudogene"). Each type of locus was given a locus name, termed with the following syntax: clone name.MX.number (e.g. bM197O15.MX.3). Gene structures were named in a similar fashion.

In this manner, 94 gene structures were annotated, representing 89 loci (the additional gene structures represent splice variants of genes). Of these loci, 46 were classified as "gene" (loci denoted as GD_mRNA, reflecting full or nearly complete mouse mRNA matches supporting the annotated structure), 31 as "predicted_gene" (loci denoted as GD_supported, reflecting incomplete mouse mRNA matches, or other homologies, supporting the annotated structure) and 12 as "pseudogene" (reflecting stop codons or frameshifts suggesting a pseudogene).

This categorisation was adopted to distinguish those genes whose structures were determined via a single transcript (allowing extension of UTRs by EST matches) and those genes whose structures may represent a "composite" transcript or whose splicing pattern was determined from sequence from a different organism.

Examples of each type of gene structure are given in Figure 4-9, Figure 4-10 and Figure 4-11.

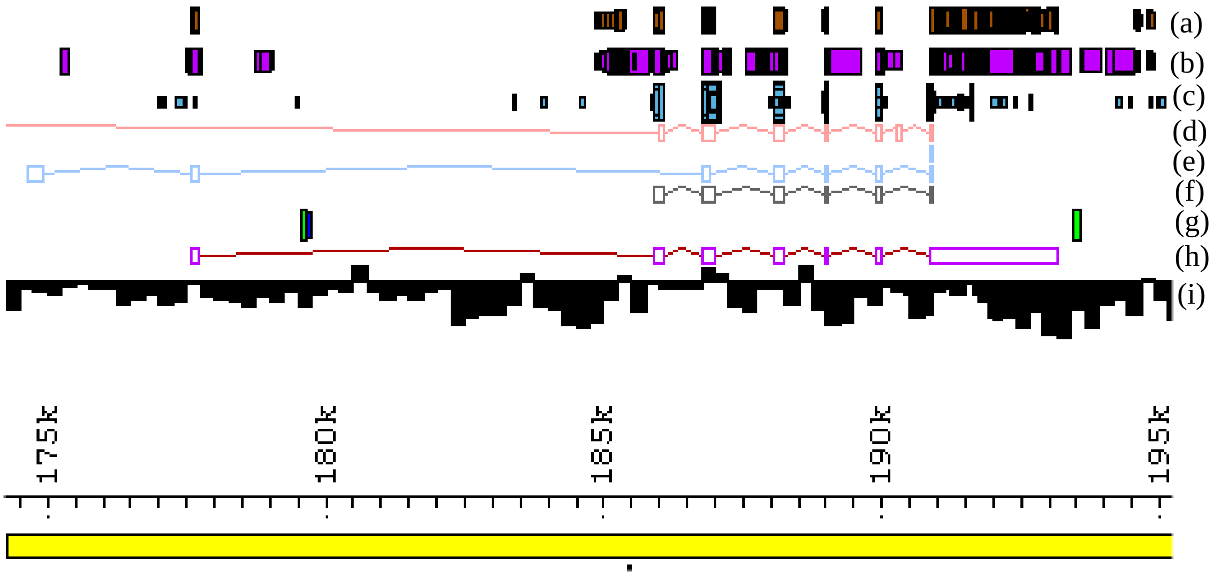


Figure 4-9 Diagram illustrating a "gene" (GD_mRNA structure) structure, for the Plp gene (locus bM197O15.MX.3). The diagram shows an ACeDb representation of the gene structure. Key – (a) mRNA BLASTN matches, (b) EST BLASTN matches, (c) protein BLASTX matches, (d) FGENESH gene prediction, (e) GENSCAN gene predictions, (f) HALFWISE gene prediction, (g) Interspersed repeats (SINEs illustrated), (h) annotated gene structure, (i) GC content (increasing thickness of bars represents increased %GC relative to adjacent sequence). The yellow bar represents the clone sequence with scale (in bp) noted. Exons are depicted as coloured open boxes, with introns represented as coloured lines connecting the exons.

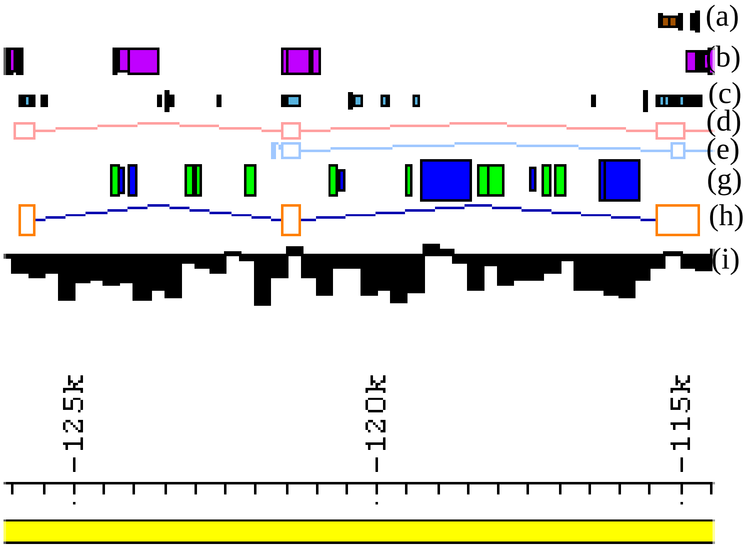


Figure 4-10 Diagram illustrating a “predicted gene” (GD_supported structure) structure, for locus bM182N4.MX.3. The diagram shows an ACeDb representation of the gene structure. Key – as for Figure 4-9. In this case, the gene structure was annotated from BLASTX matches to human XK protein (accession P51811).

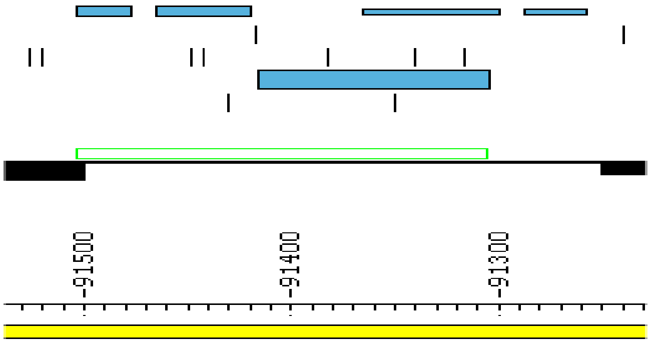


Figure 4-11 Diagram illustrating a “pseudogene” (pseudogene structure) structure, for the pseudogene locus bM389M3.MX.4. The diagram shows an ACeDb representation of the gene structure. Key: green box – annotated pseudogene, blue boxes – BLASTX protein matches, vertical lines – boundaries of open reading frames (one row for each forward strand reading frame). In this case, a BLASTX match to a mouse histone H2B protein skips frames, indicating a frameshift mutation.

The annotated gene structures are shown in context in the region in Figure 4-12 and are listed in Table 4-3.

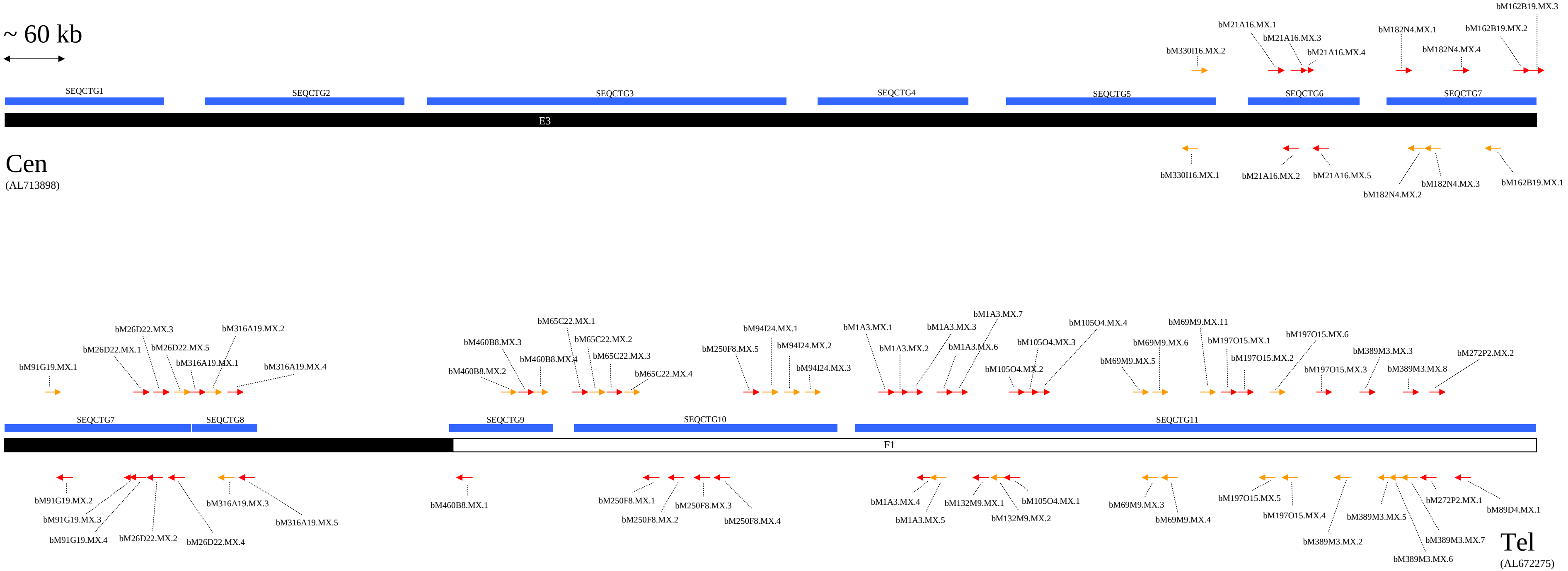


Figure 4-12 Genes annotated on finished sequence of the mouse X E3-F2 region from clone bM253D13 (AL713898) to bM89D4 (AL672275) (in contig 24), annotated as described in this Chapter. The region beginning is at top left, continuing onto the lower section of the diagram. "Cen" denotes the centromeric end, "Tel" the telomeric end. Arrows represent annotated genes, direction indicating transcription direction. Red arrows represent "gene" loci, orange arrows "predicted gene" loci. Sequence contigs are represented by blue bars. The order of clones in the sequence contigs (and their accession numbers) is given in Table 4-2. Approximate boundaries of cytogenetic bands are indicated below the blue bars (from Ensembl mouse v19.30.1).

Locus	Type	Description	Locus	Type	Description
bM330I16.MX.1	predicted	Pcdh19	bM1A3.MX.2	gene	Similar to microsomal signal peptidase
bM330I16.MX.2	predicted		bM1A3.MX.3	gene	Similar to microsomal signal peptidase
bM21A16.MX.1	gene	Sytl4	bM1A3.MX.4	gene	Bex2
bM21A16.MX.2	gene	Srpul	bM1A3.MX.5	predicted	similar to NXF (NXF3?)
bM21A16.MX.3	gene		bM1A3.MX.6	gene	Mouse specific?
bM21A16.MX.4	gene	Tm4sf6	bM1A3.MX.7	gene	Rex3
bM21A16.MX.5	gene	Myodulin	bM132M9.MX.1	gene	pp21-like
bM182N4.MX.1	gene	Cstf2	bM132M9.MX.2	predicted	pp21-like
bM182N4.MX.2	predicted	Nox1	bM105O4.MX.1	gene	Bex1
bM182N4.MX.3	predicted	Xk-L	bM105O4.MX.2	gene	pp21-like
bM182N4.MX.4	gene	ADP-ribosylation factor	bM105O4.MX.3	gene	pp21-like
bM162B19.MX.1	predicted	similar to FLJ12687	bM105O4.MX.4	gene	Bex3
bM162B19.MX.2	gene	similar to FLJ14084	bM105O4.MX.5	pseudogene	Similar to PARL
bM162B19.MX.3	gene	Lrpr1	bM105O4.MX.6	pseudogene	Similar to PARL
bM91G19.MX.1	predicted	Drp2	bM105O4.MX.7	pseudogene	Similar to PARL
bM91G19.MX.2	gene	TafIIq	bM105O4.MX.8	pseudogene	Similar to PARL
bM91G19.MX.3	gene	Timm8a	bM105O4.MX.9	pseudogene	Similar to PARL
bM91G19.MX.4	gene	Btk	bM69M9.MX.1	pseudogene	Similar to PARL
bM26D22.MX.1	gene	Rpl44	bM69M9.MX.2	pseudogene	Similar to PARL
bM26D22.MX.2	gene	Gla	bM69M9.MX.5	predicted	Similar to Kir3DL
bM26D22.MX.3	gene	Hnrnp	bM69M9.MX.6	predicted	Similar to Kir3DL - probably part of bM69M9.MX.5
bM26D22.MX.4	gene		bM69M9.MX.3	predicted	Similar to Kir3DL (this overlaps bM69M9.MX.5/6)
bM26D22.MX.5	predicted	Alex-like	bM69M9.MX.4	predicted	Similar to Kir3DL (this overlaps bM69M9.MX.5/6)
bM316A19.MX.1	gene	Alex-like	bM69M9.MX.7	pseudogene	Similar to PARL
bM316A19.MX.2	predicted	Alex-like	bM69M9.MX.8	pseudogene	Similar to PARL
bM316A19.MX.3	predicted	Alex-like	bM69M9.MX.9	pseudogene	Similar to PARL
bM316A19.MX.4	gene	Alex-like	bM69M9.MX.10	pseudogene	Similar to PARL
bM316A19.MX.5	gene	Alex-like	bM69M9.MX.11	predicted	Probably belongs to AK044164.1 gene
bM460B8.MX.1	gene	pp21-like	bM197O15.MX.1	gene	pp21-like
bM460B8.MX.2	predicted	Pramel3L	bM197O15.MX.2	gene	pp21-like
bM460B8.MX.3	gene	Pramel3L	bM197O15.MX.5	predicted	Mrgx
bM460B8.MX.4	predicted	Pramel3L	bM197O15.MX.6	predicted	
bM65C22.MX.1	gene	Pramel3L	bM197O15.MX.4	predicted	Glra4
bM65C22.MX.2	predicted	Pramel3L	bM197O15.MX.3	gene	Plp
bM65C22.MX.3	gene	Pramel3L	bM389M3.MX.2	predicted	Rab9b
bM65C22.MX.4	predicted	Pramel3L	bM389M3.MX.4	pseudogene	Histone H2B pseudogene
bM250F8.MX.1	gene	similar to NXF (NXF2b?)	bM389M3.MX.3	gene	Histone H2B
bM250F8.MX.2	gene	Pramel3L	bM389M3.MX.5	predicted	Thymosin-beta like
bM250F8.MX.3	gene	similar to TCP11/PBS13	bM389M3.MX.6	predicted	Thymosin-beta like
bM250F8.MX.4	gene	Thymosin-beta	bM389M3.MX.7	predicted	
bM250F8.MX.5	predicted	Similar to KIAA0443	bM389M3.MX.8	gene	Similar to mitochondrial carrier protein
bM94I24.MX.1	predicted	Similar to KIAA0443	bM272P2.MX.1	gene	partly in LINE
bM94I24.MX.2	predicted	Similar to KIAA0443	bM272P2.MX.2	gene	Similar to FLJ33902
bM94I24.MX.3	predicted	Similar to KIAA0443	bM89D4.MX.1	gene	Esx1
bM1A3.MX.1	gene	Intronless, in LINE			

Table 4-3 List of annotated loci within the region bounded by clones bM253D13 (Cen) and bM89D4 (Tel). The locus names are given in the first and fourth columns, with the annotation type listed in the second and fifth columns. Descriptions, where applicable, are given in the third and sixth columns. Gene annotations are listed from centromere to telomere in the table.

4.4 Comparative analysis of the human and mouse Xq22-q23/E3-F2 region

4.4.1 *Orthologues of human Xq22 genes*

The annotation of the mouse sequence allowed a comparison to be made between the gene complement and organisation of the human Xq22-q23 and mouse X E3-F2 region. Human Xq22 genes and their likely orthologues are listed in Table 4-4.

Human Gene (locus name)	HUGO	other name(s)	Mouse locus	Description
bA99E24.CX.1	PCDH19	KIAA1313	bM330I16.MX.1	Pcdh19
			bM330I16.MX.2	Hits Xq22 by BLAST
			(bM395D17)	
dJ479J7.1		myodulin/TNMD	bM21A16.MX.1	Sytl4
TM4SF6,	TM4SF6	T245	bM21A16.MX.2	Srpul
			bM21A16.MX.3	Hits Xq22 by BLAST
dJ479J7.3		SRPUL	bM21A16.MX.4	Tm4sf6
bA524D16A.2	SYTL4	Granuphilin A	bM21A16.MX.5	Myodulin
			(bM78G10)	
CSTF2	CSTF2		bM182N4.MX.1	Cstf2
NOX1	NOX1	MOX1	bM182N4.MX.2	Nox1
cU131B10.CX.1			bM182N4.MX.3	Xk-L
dJ341D10.1				
dJ341D10.2			bM182N4.MX.4	ADP-ribosylation factor
dJ341D10.3		FLJ12687	bM162B19.MX1	similar to FLJ12687
dJ664K17.CX.1		FLJ14084	bM162B19.MX.2	similar to FLJ14084
FSHPRH1	FSHPRH1	LRPR1	bM162B19.MX.3	Lrpr1
DRP2	DRP2		bM91G19.MX.1	Drp2
dJ738A13.1	TAF7L	TAF2Q/FLJ23157	bM91G19.MX.2	TafIIq
TIMM8A	TIMM8A	DFN1/DDP	bM91G19.MX.3	Timm8a
BTK	BTK	ATK	bM91G19.MX.4	Btk
RPL44	RPL36A	RPL44	bM26D22.MX.1	Rpl44
GLA	GLA		bM26D22.MX.2	Gla
HNRPH2		HNRPH2	bM26D22.MX.3	Hnrnp
dJ164F3.CX.2				
			bM26D22.MX.4	Hits Xq22 by BLAST
cU209G1.CX.1			bM26D22.MX.5	Alex-like
			(bM195N13)	
cU209G1.CX.2				
dJ514P16.CX.1				
cU61B11.CX.1		ALEX1	bM316A19.MX.1	Alex-like
dJ545K15.CX.1		FLJ20811	bM316A19.MX.2	Alex-like
dJ545K15.1		FLJ20811	bM316A19.MX.3	Alex-like
dJ545K15.2		ALEX3	bM316A19.MX.4	Alex-like
cV602D8.CX.1		ALEX2/KIAA0512	bM316A19.MX.5	Alex-like
			(bN374B8, bM65A22, bN142A19)	
NXF5	NXF5			

dJ3E10.CX.1				
dJ122O23.CX.1				
cV351F8.CX.1			bM460B8.MX.1	pp21-like
			bM460B8.MX.2	Pramel3L
			bM460B8.MX.3	Pramel3L
			bM460B8.MX.4	Pramel3L
			bM65C22.MX.1	Pramel3L
			bM65C22.MX.2	Pramel3L
			bM65C22.MX.3	Pramel3L
			bM65C22.MX.4	Pramel3L
cV351F8.CX.2				
cU19D8.CX.1				
NXF2	NXF2			
bA353J17.1			bM250F8.MX.1	similar to NXF
			bM250F8.MX.2	Pramel3L
bA353J17.2			bM250F8.MX.3	similar to TCP11/PBS13
dJ77O19.CX.1		NB thymosin beta/TMSNB	bM250F8.MX.4	Thymosin-beta
dJ1100E15.2	NXF4			
dJ1100E15.CX.3		FLJ12969/FLJ13382	bM250F8.MX.5	Similar to GASP
dJ769N13.1		GASP/KIAA0443	bM94I24.MX.1	Similar to GASP
dJ769N13.CX.1			bM94I24.MX.2	Similar to GASP
dJ769N13.CX.2		KIAA1701	bM94I24.MX.3	Similar to GASP
			(bM160E6)	
			bM1A3.MX.1	Intronless, in LINE
			bM1A3.MX.2	Similar to microsomal signal peptidase
			bM1A3.MX.3	Similar to microsomal signal peptidase
dJ769N13.CX.3				
cU157D4.CX.1				
cU237H1.1				
dJ198P4.CX.1			bM1A3.MX.4	Bex2
NXF3	NXF3		bM1A3.MX.5	similar to NXF (NXF3?)
			bM1A3.MX.6	Hits Xq22 by BLAST
dJ635G19.2		FLJ10097	bM1A3.MX.7	Rex3
cU177E8.CX.1		FLJ22696	bM132M9.MX.1	pp21-like
cU177E8.CX.3			bM132M9.MX.2	pp21-like
dJ79P11.1			bM105O4.MX.1	Bex1
cU105G4.1			bM105O4.MX.2	pp21-like
cU105G4.2			bM105O4.MX.3	pp21-like
NGFRAP1	NGFRAP1	NADE/BEX3/HGR74/DXS6984E	bM105O4.MX.4	Bex3
			bM105O4.MX.5	Similar to PARL
			bM105O4.MX.6	Similar to PARL
			bM105O4.MX.7	Similar to PARL
			bM105O4.MX.8	Similar to PARL
			bM105O4.MX.9	Similar to PARL
			bM69M9.MX.1	Similar to PARL
			bM69M9.MX.2	Similar to PARL
			bM69M9.MX.5	Similar to Kir3DL

			bM69M9.MX.6	Similar to Kir3DL - probably part of above
			bM69M9.MX.3	Similar to Kir3DL (this overlaps pos strand genes)
			bM69M9.MX.4	Similar to Kir3DL (this overlaps pos strand genes)
			bM69M9.MX.7	Similar to PARL
			bM69M9.MX.8	Similar to PARL
			bM69M9.MX.9	Similar to PARL
			bM69M9.MX.10	Similar to PARL
			bM69M9.MX.11	Hits Xq22 by BLAST
cU250H12.CX.1				
cV857G6.CX.1		FLJ21174		
cV857G6.CX.2			bM197O15.MX.1	pp21-like
TCEAL1	TCEAL1	pp21	bM197O15.MX.2	pp21-like
dJ1055C14.2	MORF4L2	MRGX/KIAA0026	bM197O15.MX.5	Mrgx
dJ1055C14.CX.1			bM197O15.MX.6	
dJ1055C14.3			bM197O15.MX.4	Glra4
PLP	PLP1	PLP/PMD	bM197O15.MX.3	
dJ540A13A.CX.1	RAB9B	RAB9L	bM389M3.MX.2	Rab9b
bA370B6.1			bM389M3.MX.4	H2B pseudo
			bM389M3.MX.3	H2B
cU116E7.CX.2		FLJ22859		
cU116E7.CX.3				
cV362H12.CX.1			bM389M3.MX.5	Thymosin-beta
			bM389M3.MX.6	Thymosin-beta
dJ839M11.1				
dJ839M11.2				
cU240C2.1				
cU240C2.2				
cU46H11.CX.1			bM389M3.MX.7	Similar to mitochondrial carrier protein
cU46H11.CX.2			bM389M3.MX.8	
dJ233G16.CX.1			bM272P2.MX.1	partly in LINE
dJ233G16.1			bM272P2.MX.2	Similar to FLJ33902
dJ513M9.1			bM89D4.MX.1	Esx1

Table 4-4 Human Xq22 genes (listed Cen to Tel) and their likely mouse orthologues. Grey rows represent sequence gaps (with clone being sequenced indicated) and the yellow row represents a contig gap. Light blue rows highlight instances where an orthologue is not annotated in one of the species.

From Table 4-4, it is apparent that whilst the two regions are largely orthologous, breaks in conserved synteny are noted. Of the 89 mouse loci annotated, 56 have likely orthologues in the corresponding region in human. Five loci annotated in mouse were not annotated in human, but matched human Xq22 sequence when BLASTN was used to search the human genome for similarities ("Hits Xq22 by BLAST" in description column). Two loci annotated in mouse, encoding histone and thymosin beta genes, were not found in human Xq22.

Of the human loci annotated, 22 were not found in the corresponding mouse region at the time of writing, although as sequence gaps remain it remains to be seen whether these genes are represented in the mouse region. Further comparative sequence analysis may also reveal matches indicating the presence of mouse orthologues.

The main breaks in orthology are the lack of mouse orthologues of human Xq22 genes cU116E7.CX.2 and cU116E7.CX.3, each of which has an additional paralogue within Xq22 (see Chapters 3 and 5), the lack of mouse orthologues of a cluster of human Xq22 histone genes (dJ839M11.1, dJ839M11.2, cU240C2.1 and cU240C2.2) and the lack of human orthologues of a histone and thymosin-beta gene (mentioned above).

The most striking difference between the two regions though is the presence of many PARL and Pramel3L loci within mouse E3-F2, which is not seen within human Xq22. Furthermore, the mouse Kir3DLl gene resides within a cluster of the PARL loci, but the rat and human Kir3DL1 loci are autosomal (NCBI LocusLink), indicating a break in synteny for this gene.

The mouse region studied also contained orthologues of many of the human Xq22 paralogues introduced in Chapter 3. This indicated that many of the duplications leading to the paralogy seen occurred prior to the human-mouse divergence. Detailed comparisons of these human and mouse genes are presented in Chapter 5.

4.4.2 PARL repeats

Analysis of the mouse sequence revealed that in addition to orthologues of paralogous genes described in Chapter 3, multiple members of other gene families were present within the region. One of these families was discovered when initial analysis of the region (using NIX (RFCGR)) revealed several loci with homology to an intra-membrane serine protease, namely, presenilin-associated rhomboid-like protein, or PARL. Human PARL has been mapped to 3q27.3 (NCBI – Locuslink) and has a gene structure containing 10 exons (Ensembl v17.33.1). This is in contrast to the mouse loci described here, which appear to represent retroposed pseudogenes.

A total of 11 loci with similarity to PARL were annotated in the mouse X E3-F2 region (see earlier). The level of homology varies between repeats, as well as the lengths of the sequences annotated. Two separate alignments were performed in order to minimise gaps, and the pairwise identities of the PARL-like sequences with respect to bM105O4.MX.8 and bM105O4.MX.5 were calculated from ungapped regions of these alignments These data are given in Table 4-5. An overview of the locations of these loci and their sequence identity is given in Figure 4-13.

Gene	bM105O4.MX.8	bM105O4.MX.9	bM105O4.MX.7	bM105O4.MX.6	bM69M9.MX.10	bM69M9.MX.1	bM69M9.MX.8
% ID	100	99	96	81	88	74	74
Gene	bM105O4.MX.5	bM69M9.MX.9	bM69M9.MX.2	bM69M9.MX.7			
% ID	100	84	37	37			

Table 4-5 Sequence identities (% ID) of PARL-like nucleotide sequences to PARL-like gene bM105O4.MX.8 (upper row) and bM105O4.MX.5 (lower row), each row calculated from separate sub-alignments.

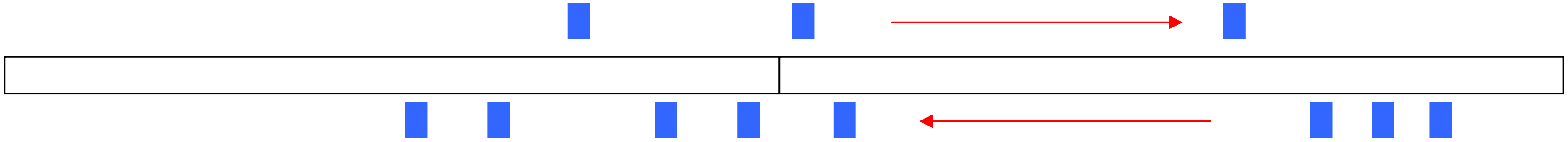
The sequence of the region containing the Parl-like loci was also analysed using Dotter (Sonnhammer and Durbin, 1995) to identify repeats in context with the gene loci (Figure 4-14). The program aligns two sequences against each other, and nucleotide identities are plotted as points to scale along the sequence axes. Thus, in this case because two identical sequences were aligned, the diagonal through the origin reflects complete identity to itself at each nucleotide position. Direct repeats appear as lines parallel to the diagonal and inverted repeats as lines perpendicular to the diagonal.

This plot suggests that the PARL repeats do not reside within large regions of highly conserved sequence and that the intervening sequence has diverged somewhat, although various inverted repeats are seen, identified as lines of longer length than other "noise", perpendicular to the horizontal.

Parl-like loci lie either side of a region that appears to contain an inverted repeat encoding a Kir3Dl1 gene and two PARL-like loci. This repeat is at least partly palindromic, as the Kir3Dl1 gene copies overlap substantially. There is a break here in conserved synteny of the region compared to human (see earlier). A similarity search of human genomic sequence using BLASTN of human PARL sequence accession BC014058 against the human genome assembly 34 (Ensembl release 18.34.1) failed to detect any similar sequences on the X chromosome, but did detect a processed pseudogene (VEGA annotation dJ95L4.4-001) and the PARL gene at 3q27.1.

It is possible that these repeats have arisen during rearrangements of the mouse region during evolution. The lack of multiple PARL-like loci in the human Xq22 region suggests that the *Mus musculus* X E3-F2 region has undergone independent rearrangements.

(a)



(b)

```
                  *       920         *       940         *       960         *       980         *      1000         *      1020         *      1040         *      1060         *      1080
bm105o4mx8 : --------AGGGTTAGGAAGGTCAGGCCAAGGTGGGGCAGAGTTTGGGAAGATGGTATTGCAAGGTTGG-TACAGTGAGG-------CTGGGGCTGCCTTTAGGCCTGTGTGCCTCCCTTCAGAGCTGCCTACCAG-GAGCTCTCTGAGTCCCAGCACCCA-CGACAGCTGGAATGAAGGA :  163
bm105o4mx9 : --------AGGCTTAGGAAGGTCATACCAAGGTGGGGCAGAGTTTGGGAAGATGGTATTGCAAGGTTGG-TACAGTGAGG-------CTGGGGCTGCCTTTAGGCCTGTGTGCCTCCCTTCAGAGCTGCCTACCAG-GAGCTCTCTGAGTCCCAGCACCCA-CGACAGCTGGAATGAAGGA :  163
bm105o4mx7 : --------AGGGTTAGGAAGGTCATACCAAGGTGGGGCAGAGGTTGGGAAGATGGTATTGCAAGGTTGG-TACAGTGAGG-------CTGGGGCTGCCTTTAGCCCTGTGTGCCTCCCTTCAGAGCTGCCTACCAG-GAGCTCTCTGAGTCCCAGCACCCA-CGACAGTTGGAGTGAAGGA :  163
bm105o4mx6 : --------AGGCTTAGGAAGGTCATACCAATGTGGGGCAGAGGTTGGGAAGATGGCATTGCAAGGTTGG-TACAGTGAGG-------CTGGGGCTGCCTTTAGCCCTGTGTGCCTCCCTTCAGAGCTGCCTACCAG-GAGCTTTCTGAGACCCAGATACCA-AGGCTGTTGGAATGCAGGA :  163
bm69m9mx10 : ---------------------------------------------------------------------------------------------------------------GCCTCCCTTGGGGGAAGGCTACCAG-GCGCT--CTGAGACCAAGACCTCA-CGGCTGTTCGAATGCAGGA :   66
bm69m9mx1  : -------------------------------------------------------CATGGTGAGGCAGG-GGTTGGGAAG-------ACGCGGCTGCCTTCAGGCTTG--TGCCTCCCTTTGGCAGTAGCTACCGC-GAGCT----GAGACCGAGACCCCAACGGCTGCTCAGATACAGTT :  111
bm69m9mx8  : --------------------------------------------------------------------------------------------------------------------------------------------------------------------GCTGCTCAGATACAGTT :   17
bm105o4mx5 : --------AGGGTTAGGAAGATCAGGCCAAGATGGGCCAGAGGTTGGGAAGAGGGCATTGCAAGGTTGGATACAGTGAGG-------CTGGGACTGTCTTTAGCTTTCTGTGCCTCCCTTGGGGGAAGGCTACCAG-GAGCTCTCTGAGACCCAAACTGTA-CAGCAGCATGAATATAGGA :  164
bm69m9mx9  : --------AGGCTTAGGAAGGTCATACCAAGGTGGGGCAGAGTTTGGGAAGATGGTATTGCAAGGTTGG-TACAGTGAGG-------CTGGGGCTGCCTTTAGGCCTGTGTGCCTCCCTTCAGAGCTGCCTACCAG-GAGCTCTCTGAG-------------------------------- :  132
bm69m9mx2  : CTGGGGGGAGGCTCACTATCATCTTCCTCCCCGTTTTCCCATTCATGGTAGGCAATGCCTTAAAAGCCATCATTGCCATGGATACATCTGGGGTGATCCTGGGATGGAAACTTTTGACTATGAAGCTGCATCATGGCGAGGTCTCTTAGGATATAGTATAT-CATGTACTGTCATGAATTC :  293
bm69m9mx7  : CTGGGGGGAGGCTCGCTATCATCTTCCTCCCCGTTCTCCCATTCATGGTAGGCAATGCCTTAAAAGCCATCATTGCCATGGATACATCTGGGGTGATCCTGGGATGGAAACTTTTGACTATGGAGCTGCATCATGGCGAGGTCTCTTAGGATATAGTATAT-CATGTACTGTCATGAATTC : 1085

                *      1100         *      1120         *      1140         *      1160         *      1180         *      1200         *      1220         *      1240         *      1260
bm105o4mx8 : TAAAACTATTT-CTTCAGCAAAAAT----GTACATTTATAAAGGTACCGACGAAGAATAACCCTTGAAATC-AGTCAGTGTCATGGGGACAAGCATACAG---AAGTGCCTTGATCCAGTCCCTGAAGGAACCA---GTCTTTTGTCCTCCCCCTAGCTCATAAGAACTCCTGTGAAGCCT :  332
bm105o4mx9 : TAAAACTATTT-CTTCAGCAAAAAT----GTACATTTATAAAGGTACCGACGAAGAATAACCCTTGAAATC-AGTCAGTGTCATGGGGACAAGCATACAG---AAGTGCCTTGATCCAGTCCCTGAAGGAACCA---GTCTTTTGTCCTCCCCCTAGCTCATAAGAACTCCTGTGAAGCCT :  332
bm105o4mx7 : TAAAATTATTT-CTTCAGTAAAAAT----GTACATTTATAAAGGTACCGACGAAGAATAACCCTTGAAATC-AGTCAGTGTCATGGGGACAAGCATACAG---AAGTGCCTTGATCCGGTCCCTGAAGGAACCA---GTCTTTTGTCCTCCCCCTAGCTCATAAGAACTCCTGTGAAGCCT :  332
bm105o4mx6 : TAAAATTTTTTTCTTCAGCAAATAT----TTACATTTATAAAAGTACCAAGGAAAATAAACCTTGAATATC-AGATAGTGTTAAGGGGACAAGCATACAG---AGGCGCCTTGATCCCACGCCTGAAGGAAACA---GTCTTTTGTCCTCCCC-TAGCTTATAAGAAGTATAGTGAAGCGG :  332
bm69m9mx10 : TAAAATTATTT-CTTTAACAAAAAAAAGAGTACATTTATAAAGGTACCGAGGAAGAATAACCCTTGAAATC-AGTCAGTGTCATGGGGACAAGCATACAG---AAGTGCCTTGATCTGGCCCCTGAAGGAAACA---GTCTTTTAACCTCCACCTAGCTCATAAGAACTCCTGTGAA--CT :  237
bm69m9mx1  : TTAACTTATTT-CTTCAGCAAAACT----ATGGATTTAGAAAAGTACTGAGGAAGGCTGAACCTTGAAATTTAAACAGTGTCAAGCTGAGAAGCATACAAGACAAGTGCCTTGATCTTGTCCCTGGAGGAAACGCTTGTGTGCTTTCGTCCCCCTAGCTTATAGGATCTCTCGAGAAGCCT :  287
bm69m9mx8  : TTAACTTATTT-CTTCAGCAAAACT----ATGGATTTAGAAAAGTACTGAGGAAGGCTGAACCTTGAAATTTAAACAGTGTCAAGCTGAGAAGCATACAAGACAAGTGCCTTGATCTTGTCCCTGGAGGAAACGCTTGTGTGCTTTCGTCCCCCTAGCTTATAGGATCTCTCGAGAAGCCT :  193
bm105o4mx5 : TGAAATTATTT----CTTCAGAAAAAA-AGTACATTTATAAAGGTACCGAGGAAGGATAAACTGTGAAGATCAGACAGTGTTATGGCGACAAGCATACAG---ATGTGCCTTGATCCCACCCCTAGAGGAAACA--GTCTTTTATCCTTGCCCTTAGCTTATAAGAACTGCAGTGAAGCCC :  335
bm69m9mx9  : ------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------- :    -
bm69m9mx2  : ATTTGGAAGAA----CAGGGAGCCTCT-AGTGAAAGTCTGGTATAAAATAAGGACTAACAACCCCAAGAATGTAGATGGTTTAAGTGAAGCAAAGTAAGT---CACTATGGTGGTGTGTCTGAACCTTGCTGCC--TG------------------------------------------- :  421
bm69m9mx7  : ATTTGGAAGAA----CAGGGAGCCTCT-AGTGAAAGTCTGGTATAAAATAAGGACTAACAAACCCAAGAATGTAGATGGTTTAAGTGAAGCAAAGTAAGT---CACTATGGTGGTGCATCTGAACCTTGCTGCC--TG------------------------------------------- : 1213
```

Figure 4-13 (a) Schematic diagram of PARL-like loci within the *Mus musculus* X E3-F2 region. Blue boxes represent single-exon PARL-like repeats in approximate locations along the clone sequence (open boxes); genes above the clones are encoded on the forward strand, and those below are on the reverse strand. The red arrows represent the span of the Kir3DL1 multi-exon loci and their transcription orientation (b) part of an alignment of nucleotide sequences of annotated PARL-like loci within the region, illustrating the level of sequence homology seen. Only part of the alignment is shown for clarity, and is representative of the homology seen in the aligned regions of the sequences. Dark grey – 80-100% conservation, light grey – 60-80% conservation.

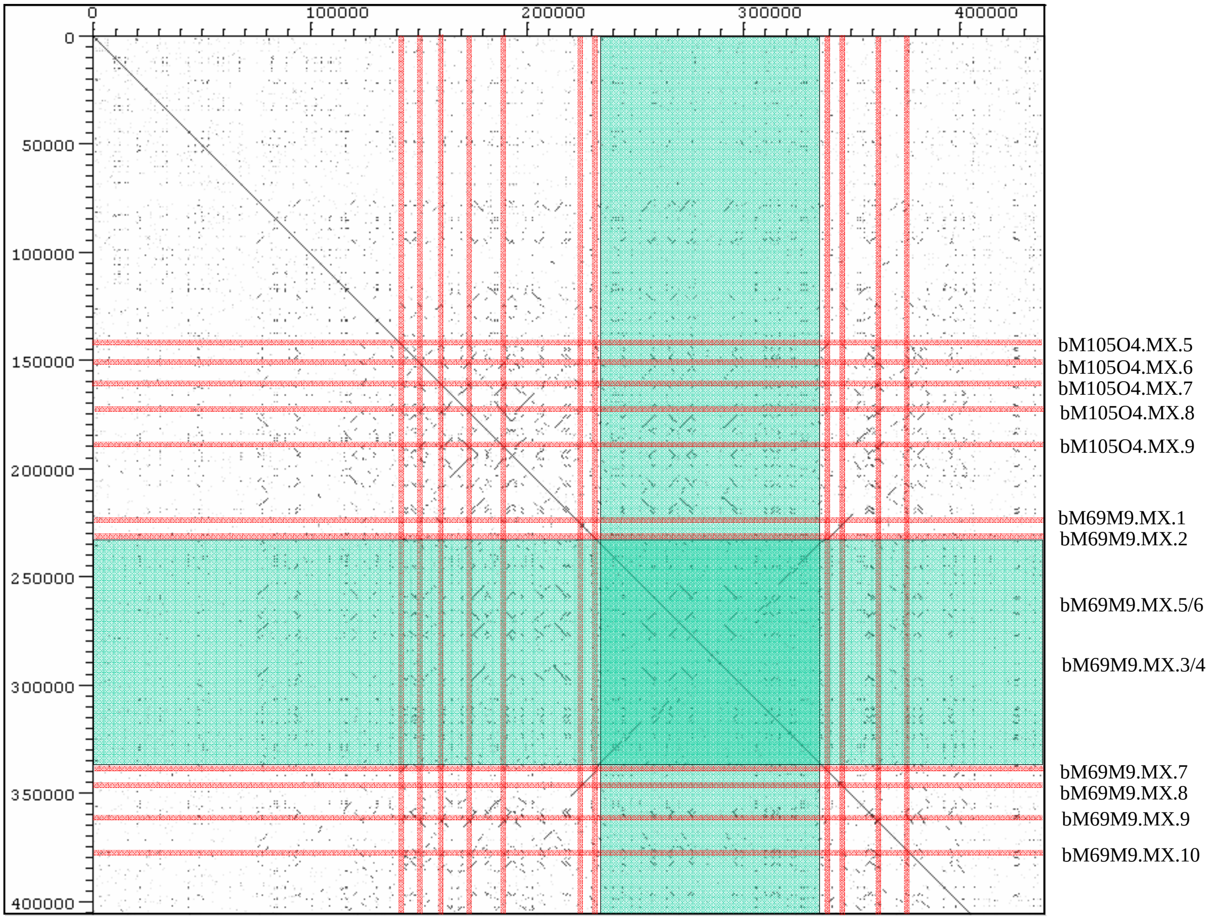


Figure 4-14 Diagram illustrating Dotter analysis of the *Mus musculus* X E3-F2 region containing PARL-like repeats. Approximate positions of PARL-like loci (illustrated in Figure 4-13) are shown by red bars located on both axes. The approximate position of the region containing the Kir3dl1 gene is shown by a green rectangle on each axis. The sequence analysed comprised of linked sequences of clones bM105O4 and bM69M9 (Accession numbers AL671493 and AL672068 respectively). No masking of known repeats was performed. Names are shown for gene positions on the y-axis, and are mirrored on the x-axis.

4.4.3 Pramel3L repeats

Another family of genes with homology to the Prame-like 3 gene was discovered and annotated within the region. These genes were termed the Prame-like3-like l (Pramel3L) loci. The PRAME– like (Preferentially Expressed Antigen in Melanoma) genes have six mouse loci noted in Locuslink (NCBI), of which two are mapped to mouse chromosome 2, three to chromosome 4 and one, the Prame-like 3 gene, to mouse X E3. Human PRAME is mapped to 22q11.22 (Locuslink-NCBI). The human PRAME gene comprises 6 exons (Ensembl v17.33.1), and is expressed in testis as well as many different tumour types.

A total of 7 loci with similarity to Pramel3, together with Pramel3 itself (gene bM460B8.MX.3), were annotated. Seven of the eight Pramel3L loci are located on the same DNA strand. There is a high level of homology between the gene family members. The pairwise identities of the sequences with respect to Pramel3 (bM460B8.MX.3) were calculated from ungapped regions of an alignment of the sequences, and are given in Table 4-6 below. The numbers of exons in the different genes ranged from 3 to 10. This may reflect alternative transcripts, different gene structures or partial duplications. An overview of the locations of these loci and their sequence homology is given in Figure 4-15.

Gene	bM460B8. MX.3	bM460B8. MX.2	bM460B8. MX.4	bM65C22. MX.1	bM65C22. MX.2	bM65C22. MX.3	bM65C22. MX.4	bM250F8. MX.2
% ID	100	71	69	99	69	99	70	73

Table 4-6 Sequence identities (% ID) of Pramel3L nucleotide sequences to Pramel3 (bM460B8.MX.3).

The sequence of the region containing the loci was also analysed using Dotter to identify genomic repeats in context with the Prame-like3-like loci (Figure 4-16). As described above, using Dotter the sequence of the region was aligned against itself, and several direct repeats are seen as dark lines parallel to the diagonal through the origin, some encompassing Pramel3L loci. This is consistent with the observation that seven of the eight Pramel3L loci are on the same strand.

From the Dotter, genes bM65C22.MX.2 and bM460B8.MX.4 are localised within a direct repeat, as are genes bM65C22.MX.3 and bM65C22.MX.1, appearing as intersecting red lines in a direct repeat in the dotter diagram. Their sequence identities to one another are 98% and 99% respectively, which is consistent with their localisation within a direct repeat.

During the annotation of human Xq22-q23 (see Chapter 3), two PRAME3L loci were found, within the NXF2 duplicon and between the TCP11-like and NXF2 loci. These two PRAME3L loci appear to be pseudogenes based on the presence of a stop codon within the frame with BLASTX homology to PRAME. The same mutation is found in both copies indicating that it is likely to have arisen prior to the NXF2 duplication event (subsequent to the human-mouse divergence, see Chapter 3). This would imply that humans and mice differ in functionality of the Prame3l gene product, as mice have retained functional copies of the Pramel3L genes, whilst in humans they are very likely non-functional.

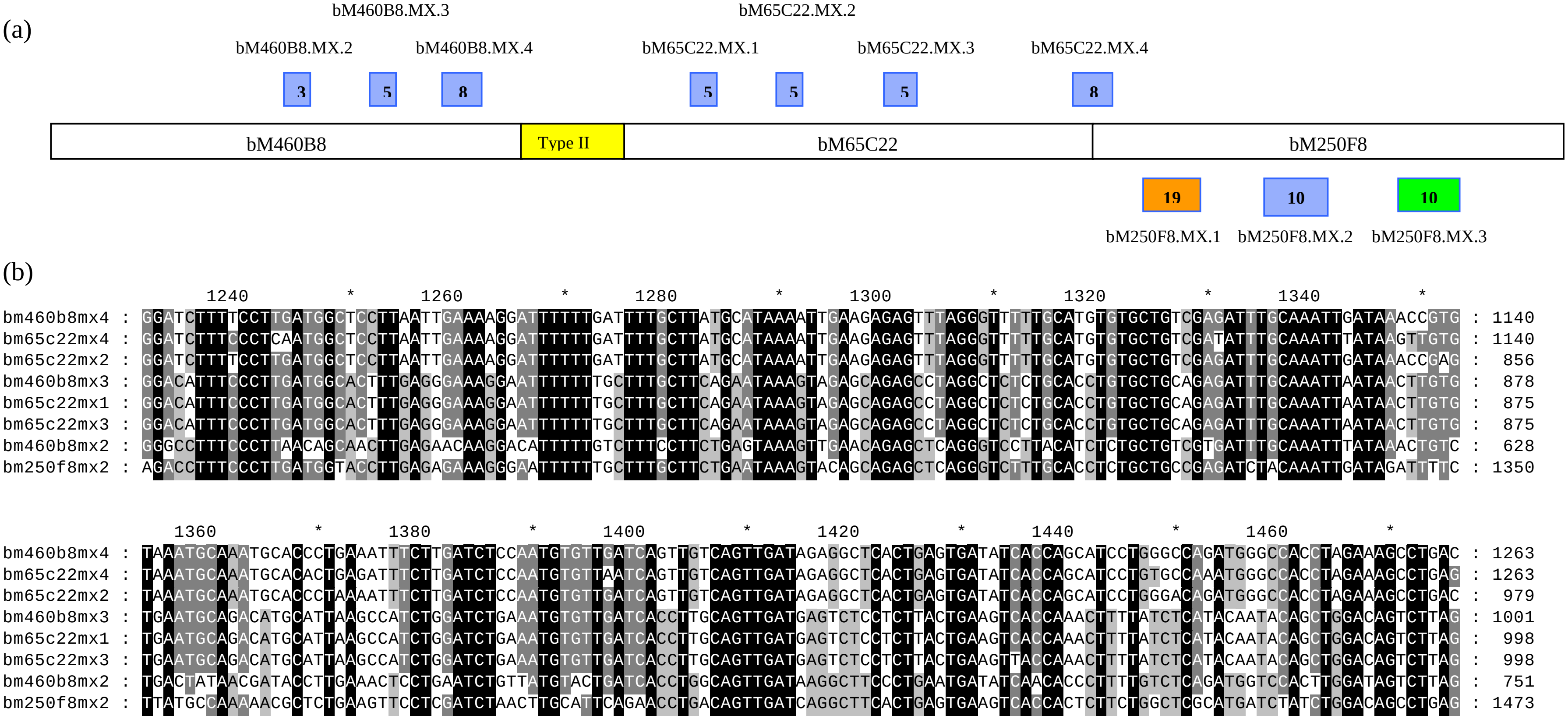

(a)
bM460B8.MX.3
bM65C22.MX.2
bM460B8.MX.2
bM460B8.MX.4
bM65C22.MX.1
bM65C22.MX.3
bM65C22.MX.4
3
5
8
5
5
5
8
bM460B8
Type II
bM65C22
bM250F8
19
10
10
bM250F8.MX.1
bM250F8.MX.2
bM250F8.MX.3
(b)
1240 * 1260 * 1280 * 1300 * 1320 * 1340 *
bm460b8mx4 : GGATCTTTTCCTTGATGGCTCCTTAATTGAAAAGGATTTTTTGATTTTGCTTATGCATAAAATTGAAGAGAGTTTAGGGTTTTTGCATGTGTGCTGTCGAGATTTGCAAATTGATAAACCGTG : 1140
bm65c22mx4 : GGATCTTTCCCTCAATGGCTCCTTAATTGAAAAGGATTTTTTGATTTTGCTTATGCATAAAATTGAAGAGAGTTTAGGGTTTTTGCATGTGTGCTGTCGATATTTGCAAATTTATAAGTTGTG : 1140
bm65c22mx2 : GGATCTTTTCCTTGATGGCTCCTTAATTGAAAAGGATTTTTTGATTTTGCTTATGCATAAAATTGAAGAGAGTTTAGGGTTTTTGCATGTGTGCTGTCGAGATTTGCAAATTGATAAACCGAG : 856
bm460b8mx3 : GGACATTTCCCTTGATGGCACTTTGAGGGAAAGGAATTTTTTTGCTTTGCTTCAGAATAAAGTAGAGCAGAGCCTAGGCTCTCTGCACCTGTGCTGCAGAGATTTGCAAATTAATAACTTGTG : 878
bm65c22mx1 : GGACATTTCCCTTGATGGCACTTTGAGGGAAAGGAATTTTTTTGCTTTGCTTCAGAATAAAGTAGAGCAGAGCCTAGGCTCTCTGCACCTGTGCTGCAGAGATTTGCAAATTAATAACTTGTG : 875
bm65c22mx3 : GGACATTTCCCTTGATGGCACTTTGAGGGAAAGGAATTTTTTTGCTTTGCTTCAGAATAAAGTAGAGCAGAGCCTAGGCTCTCTGCACCTGTGCTGCAGAGATTTGCAAATTAATAACTTGTG : 875
bm460b8mx2 : GGGCCTTTCCCTTAACAGCAACTTGAGAACAAGGACATTTTTGTCTTTCCTTCTGAGTAAAGTTGAACAGAGCTCAGGGTCCTTACATCTCTGCTGTCGTGATTTGCAAATTTATAAACTGTC : 628
bm250f8mx2 : AGACCTTTCCCTTGATGGTACCTTGAGAGAAAGGGAATTTTTTGCTTTGCTTCTGAATAAAGTACAGCAGAGCTCAGGGTCTTTGCACCTCTGCTGCCGAGATCTACAAATTGATAGATTTC : 1350
1360 * 1380 * 1400 * 1420 * 1440 * 1460 *
bm460b8mx4 : TAAATGCAAATGCACCCTGAAATTTCTTGATCTCCAATGTGTTGATCAGTTGTCAGTTGATAGAGGCTCACTGAGTGATATCACCAGCATCCTGGGCCAGATGGGCCACCTAGAAAGCCTGAC : 1263
bm65c22mx4 : TAAATGCAAATGCACACTGAGATTTCTTGATCTCCAATGTGTTAATCAGTTGTCAGTTGATAGAGGCTCACTGAGTGATATCACCAGCATCCTGTGCCAAATGGGCCACCTAGAAAGCCTGAG : 1263
bm65c22mx2 : TAAATGCAAATGCACCCTAAAATTTCTTGATCTCCAATGTGTTGATCAGTTGTCAGTTGATAGAGGCTCACTGAGTGATATCACCAGCATCCTGGGACAGATGGGCCACCTAGAAAGCCTGAC : 979
bm460b8mx3 : TGAATGCAGACATGCATTAAGCCATCTGGATCTGAAATGTGTTGATCACCTTGCAGTTGATGAGTCTCCTCTTACTGAAGTCACCAAACTTTTATCTCATACAATACAGCTGGACAGTCTTAG : 1001
bm65c22mx1 : TGAATGCAGACATGCATTAAGCCATCTGGATCTGAAATGTGTTGATCACCTTGCAGTTGATGAGTCTCCTCTTACTGAAGTCACCAAACTTTTATCTCATACAATACAGCTGGACAGTCTTAG : 998
bm65c22mx3 : TGAATGCAGACATGCATTAAGCCATCTGGATCTGAAATGTGTTGATCACCTTGCAGTTGATGAGTCTCCTCTTACTGAAGTTACCAAACTTTTATCTCATACAATACAGCTGGACAGTCTTAG : 998
bm460b8mx2 : TGACTATAACGATACCTTGAAACTCCTGAATCTGTTATGTACTGATCACCTGGCAGTTGATAAGGCTTCCCTGAATGATATCAACACCCTTTTGTCTCAGATGGTCCACTTGGATAGTCTTAG : 751
bm250f8mx2 : TTATGCCAAAAACGCTCTGAAGTTCCTCGATCTAACTTGCATTCAGAACCTGACAGTTGATCAGGCTTCACTGAGTGAAGTCACCACTCTTCTGGCTCGCATGATCTATCTGGACAGCCTGAG : 1473

Figure 4-15 (a) Schematic diagram of Pramel3L loci within the *Mus musculus* X E3-F2 region. Blue boxes represent Pramel3L repeats in approximate locations along the clone sequence (open boxes); those above the clones represent genes on the forward strand and those below are genes on the reverse strand. Exon number is given in each box. The orange and green boxes represent the likely NXF2 and TCP11-like orthologues respectively. A type II gap is shown in yellow. (b) part of an alignment of nucleotide sequences of annotated Pramel3L loci within the region, illustrating the level of sequence similarity seen. Gene bM460B8.MX.3 is the Pramel3 gene. Only part of the alignment is shown for clarity, and is representative of the homology seen in the aligned regions of the sequences.

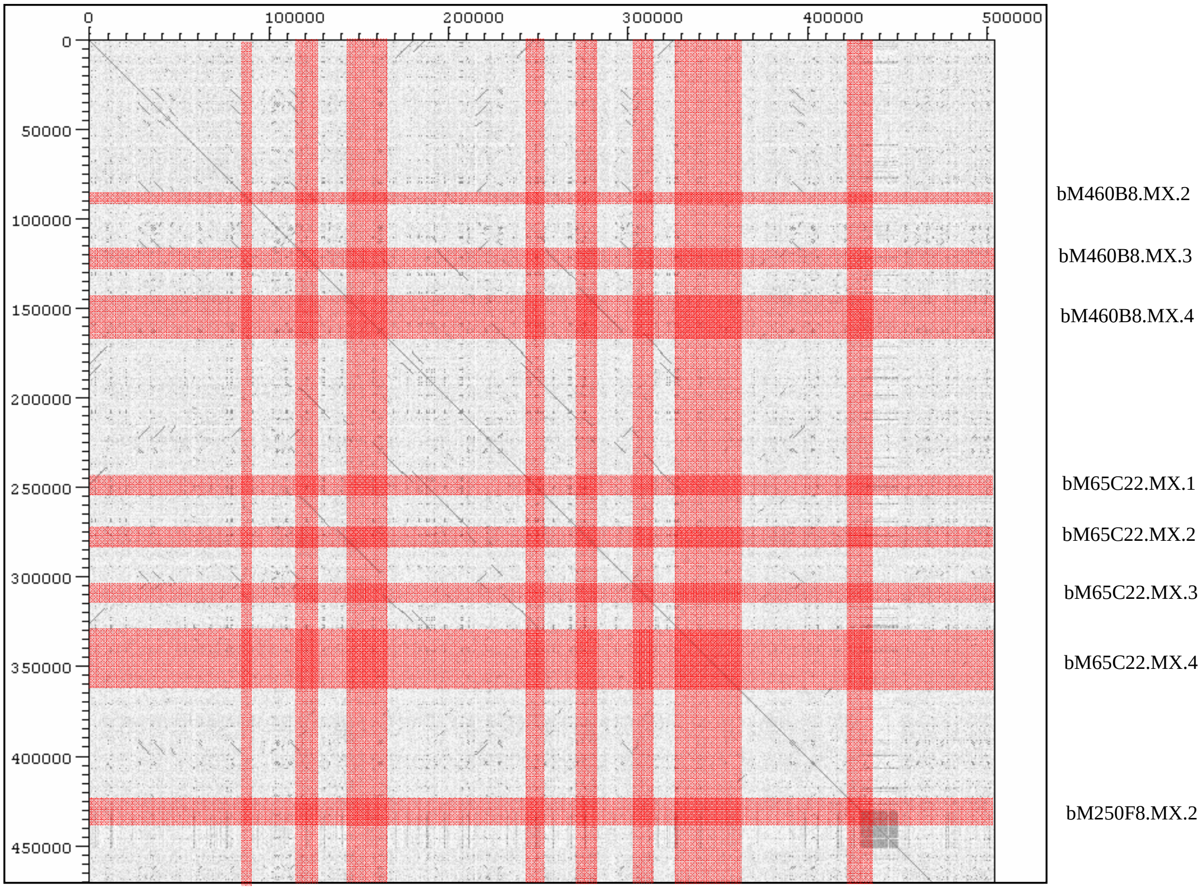


Figure 4-16 Diagram illustrating Dotter analysis of the *Mus musculus* X E3-F2 region containing Pramel3L repeats. Approximate positions of annotated loci (illustrated in Figure 4-15) are shown by red bars located on both axes. The sequence analysed comprised of linked sequences of clones bM460B8, (gap of ~50 kb), bM65C22 and bM250F8 (accession numbers AL731676, AL954640 and AL671911 respectively). No masking of known repeats was performed. As for Figure 4-14, direct repeats are visible as dark lines parallel to the diagonal through the origin. Names are shown for gene positions on the y-axis, and are mirrored on the x-axis.

4.4.4 The mouse Nxf2 locus

As was discussed in Chapter 3, the human NXF2 locus may have undergone duplication since the human and mouse lineages diverged. As expected, therefore, only one locus with homology to NXF2 was annotated within the *Mus musculus* X E3-F2 region here. The caveat remains however that an additional mouse Nxf2 locus could reside in the sequence gap proximal to the annotated locus. In common with human NXF2/NXF2a, a TCP11-like locus was found just upstream of the Nxf2 gene.

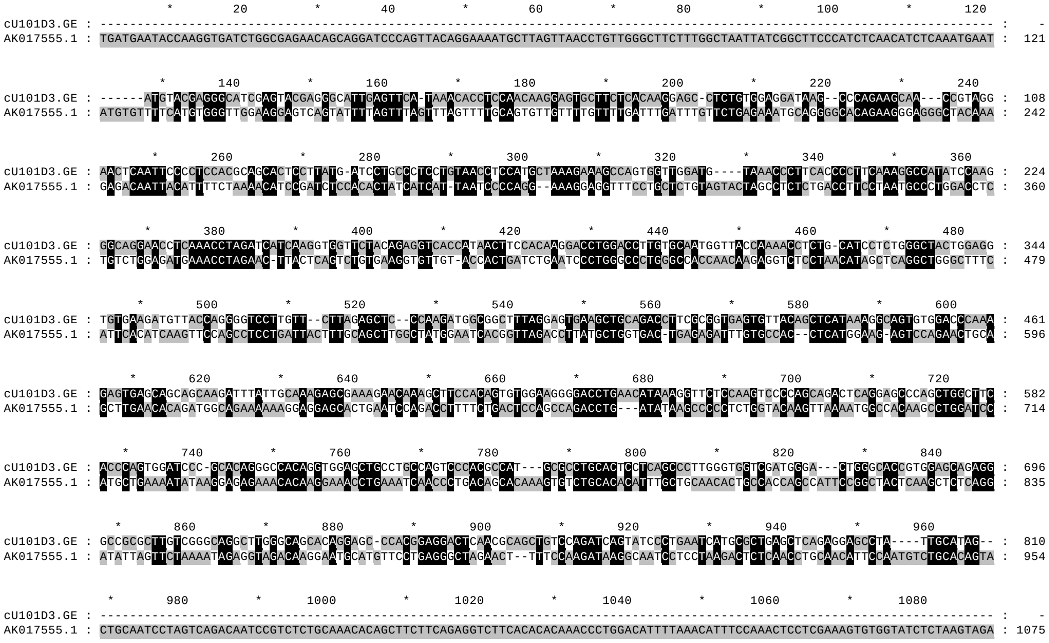
Unlike the human situation however, a Pramel3L gene (named for its similarity to the mouse Prame-like 3 gene) was annotated between the Nxf2 and Tcp11-like loci that appears functional, from the identity to the mRNA sequence used to annotate the gene. As discussed earlier, the human PRAMEL3L loci in the NXF2 region appear to be pseudogenes. This suggests different requirements for the functionality of this locus in human and mouse.

4.4.5 A mouse gene supporting the presence of a novel gene in human Xq22

Locus bM1A3.MX.6 was annotated from mouse mRNA AK017555.1. BLASTN analysis of the human genome with AK017555.1 (NCBI – HTGS and nr subsets, no filter) found no significant similarity. Initially it was thought that this may reflect a further mouse-specific gene. However, a TBLASTX search with AK017555.1 against the NCBI non-redundant dataset found homology to two genomic clones within Xq22 (RP11-522L3, RP13-349O20).

In the corresponding region to bM1A3.MX.6 in human Xq22, overlapping GENSCAN and GRAIL predictions in the sequence of genomic clone Z85998 (cosmid cU101D3) were noted. These were used to design primers for cDNA screening, as described in Chapter 3. These primers, which define STS stcU101D3.1, failed to give positive results, and no gene structure could be confirmed. An alignment of AK017555.1 and the human Xq22 genscan prediction (cU101D3.GENSCAN.3) does show significant homology between the sequences (see Figure 4-17), and suggests that this genscan prediction may in fact represent a gene within human Xq22. A search for expressed sequences representing the human gene (BLASTN against NCBI nr database, filtered for human repeats) failed to find mRNA or EST matches. However, several matches to human Xq22 genomic clones were detected, which may indicate repeats within the region.

This demonstrates the utility of model organism sequence resources in gene identification studies, uncovering potential genes missed by other methodologies. The mouse sequence AK017555.1 was derived from an 8-day embryo whole-body cDNA library, and, as such, may represent a developmentally restricted transcript. It is possible that the lack of human mRNA sequence for this locus reflects the more limited cDNA coverage of developmentally restricted and tissue specific transcripts. In order to confirm expression of this locus in human tissues, a direct RT-PCR approach, without a cDNA cloning step, using cDNA templates derived from a wider variety of tissues (particularly embryonic tissues) could be employed.



```
                      *        20         *        40         *        60         *        80         *       100         *       120
cU101D3.GE : ------------------------------------------------------------------------------------------------------------------------- :    -
AK017555.1 : TGATGAATACCAAGGTGATCTGGCGAGAACAGCAGGATCCCAGTTACAGGAAAATGCTTAGTTAACCTGTTGGGCTTCTTTGGCTAATTATCGGCTTCCCATCTCAACATCTCAAATGAAT :  121

                     *       140         *       160         *       180         *       200         *       220         *       240
cU101D3.GE : ------ATGTACGAGGGCATCGAGTACGAGGGCATTGAGTTCA-TAAACACCTCCAACAAGGAGTGCTTCTCACAAGGAGC-CTCTGTGGAGGATAAG--CCCAGAAGCAA---CCGTAGG :  108
AK017555.1 : ATGTGTTTTCATGTGGGTTGGAAGGAGTCAGTATTTTAGTTTAGTTTAGTTTTGCAGTGTTGTTTTGTTTTGATTTGATTTGTTCTGAGAAATGCAGGGGCACAGAAGGGAGGGCTACAAA :  242

                    *       260         *       280         *       300         *       320         *       340         *       360
cU101D3.GE : AACTCAATTCCCCTCCACGCAGCACTCCTTATG-ATCCTGCCCTCCTGTAACCTCCATGCTAAAGAAAGCCAGTGGTTGGATG----TAAACCCTTCACCCCTTCAAAGGCCATATCCAAG :  224
AK017555.1 : GAGACAATTACATTTTCTAAAACATCCGATCTCCACACTATCATCAT-TAATCCCCAGG--AAAGGAGGTTTCCTGCTCTGTAGTACTAGCCTCTCTGACCTTCCTAATGCCCTGGACCTC :  360

                   *       380         *       400         *       420         *       440         *       460         *       480
cU101D3.GE : GGCAGGAACCTCAAACCTAGATCATCAAGGTGGTTCTACAGAGGTCACCATAACTTCCACAAGGACCTGGACCTTGTGCAATGGTTACCAAAACCTCTG-CATCCTCTGGGCTACTGGAGG :  344
AK017555.1 : TGTCTGGAGATGAAACCTAGAAC-TTACTCAGTCTGTGAAGGTGTTGT-ACCACTGATCTGAATCCTGGGCCCTGGGCCACCAACAAGAGGTCTCCTAACATAGCTCAGGCTGGGCTTTC :  479

                  *       500         *       520         *       540         *       560         *       580         *       600
cU101D3.GE : TGTGAAGATGTTACCAGGGGTCCTTGTT--CTTAGAGCTC--CCAAGATGGCGGCTTTAGGAGTGAAGCTGCAGACCTTCGCGGTGAGTGTTACAGCTCATAAAGGCAGTGTGGACCCAAA :  461
AK017555.1 : ATTCACATCAAGTTCCAGCCTCCTGATTACTTTGCAGCTTGGCTATGGAATCACGGTTAGACCTTATGCTGGTGAC-TGAGAGATTTGTGCCAC--CTCATGGAAG-AGTCCAGAACTGCA :  596

                 *       620         *       640         *       660         *       680         *       700         *       720
cU101D3.GE : GAGTGAGCAGCAGCAAGATTTATTGCAAAGAGCGAAAGAACAAAGCTTCCACAGTGTGGAAGGGGACCTGAACATAAAGGTTCTCCAAGTCCCCAGCAGACTCAGGAGCCCAGCTGGCTTC :  582
AK017555.1 : GCTTGAACACAGATGGCAGAAAAAGGAGGAGCACTGAATCCAGACCTTTTCTGACTCCAGCCAGACCTG---ATATAAGCCCCTCTGGTACAAGTTAAAATGGCCACAAGCCTGGATCC :  714

                *       740         *       760         *       780         *       800         *       820         *       840
cU101D3.GE : ACCCAGTGGATCCC-GCACAGGGCCACAGGTGGAGCTGCCTGCCAGTCCCACGCCAT---GCGCCTGCACTCCTCAGCCCTTGGGTGGTCGATGGGA---CTGGGCACCGTGGAGCAGAGG :  696
AK017555.1 : ATGCTGAAAATATAAGGAGAGAAACACAAGGAAACCTGAAATCAACCCTGACAGCACAAAGTGTCTGCACACATTTGCTGCAACACTGCCACCAGCCATTCCGGCTACTCAAGCTCTCAGG :  835

               *       860         *       880         *       900         *       920         *       940         *       960
cU101D3.GE : GCCGCGCTTGTCGGGCAGGCTTGGGCAGCACAGGAGC-CCACGGAGGACTCAACGCAGCTGTCCAGATCAGTATCCCTGAATCATGCGCTGAGCTCAGAGGAGCCTA----TTGCATAG-- :  810
AK017555.1 : ATATTAGTTCTAAAATAGAGGTAGACAAGGAATGCATGTTCCTGAGGGCTAGAACT--TTTCCAAGATAAGGCAATCCTCCTAAGACTCTCAACCTGCAACATTCCAATGTCTGCACAGTA :  954

              *       980         *      1000         *      1020         *      1040         *      1060         *      1080
cU101D3.GE : ------------------------------------------------------------------------------------------------------------------------- :    -
AK017555.1 : CTGCAATCCTAGTCAGACAATCCGTCTCTGCAAACACAGCTTCTTCAGAGGTCTTCACACACAAACCCTGGACATTTTAAACATTTCCAAACTCCTCGAAAGTGTGGTATCTCTAAGTAGA : 1075
```

Figure 4-17 Alignment of human gene prediction cU101D3.GENSCAN.3 (labelled cU101D3.GE) and part of mouse mRNA AK017555.1.

4.5 Discussion

The studies presented in this Chapter have demonstrated how the comprehensive mapping resources generated for the mouse by the scientific community facilitated rapid production of two sequence-ready BAC contigs covering the entire X E3-F2 region. This was further aided by the availability of the human genomic sequence to act as a framework on which to position mouse contigs via mouse BAC end sequences. The strength of this approach was also demonstrated in the subsequent publication of a BAC map of the *Mus musculus* genome (Gregory *et al.*, 2002).

Whilst the conserved synteny of genes on the human and other eutherian mammalian X chromosomes appears to be the general rule, the annotation of the *Mus musculus* X E3-F2 region highlighted subtle differences between human and mouse. Differences in copy number for some duplicated genes within the region (see also Chapter 5) were seen. For the Kir3DL1 gene within the mouse region, the rat orthologue appears to be autosomal and a human orthologue does not appear in the Xq22 region. Together these loci represent examples of incomplete conservation of the regions between the two species. Knowledge of such breaks in orthology are of importance in studies using data from model organisms.

Finally, large families of repeats were found within the *Mus musculus* X E3-F2 region that appear to be functional genes or processed pseudogenes. In the case of the Pramel3L loci, the only human copies noted in Xq22 appear to be pseudogenes, and suggest different requirements for this gene in the different species. For the PARL-like loci, no human Xq22 counterparts were discovered. Sequence repeats have been described within the human Xq22 region (Gareth Howell, PhD thesis, Open University), and whilst apparently different to the mouse repeats, may reflect common features between the species that predispose these regions to rearrangement. More detailed analyses of these repeats may shed further light on these observations.

The studies presented in this chapter illustrate that even when comparing regions between human and mouse for such highly conserved chromosomes as the X chromosomes, differences are apparent and must be taken into account in interpreting studies based on mouse models. The completion of a BAC map of the mouse genome and progression of sequencing of the mouse genome will allow a detailed annotation and comparison of the human and mouse genomes in order to aid studies using the mouse as a model organism.