Chapter Six - Characterisation of a regional duplication represented on human Xq22-q23 and Xp

6.1 Introduction

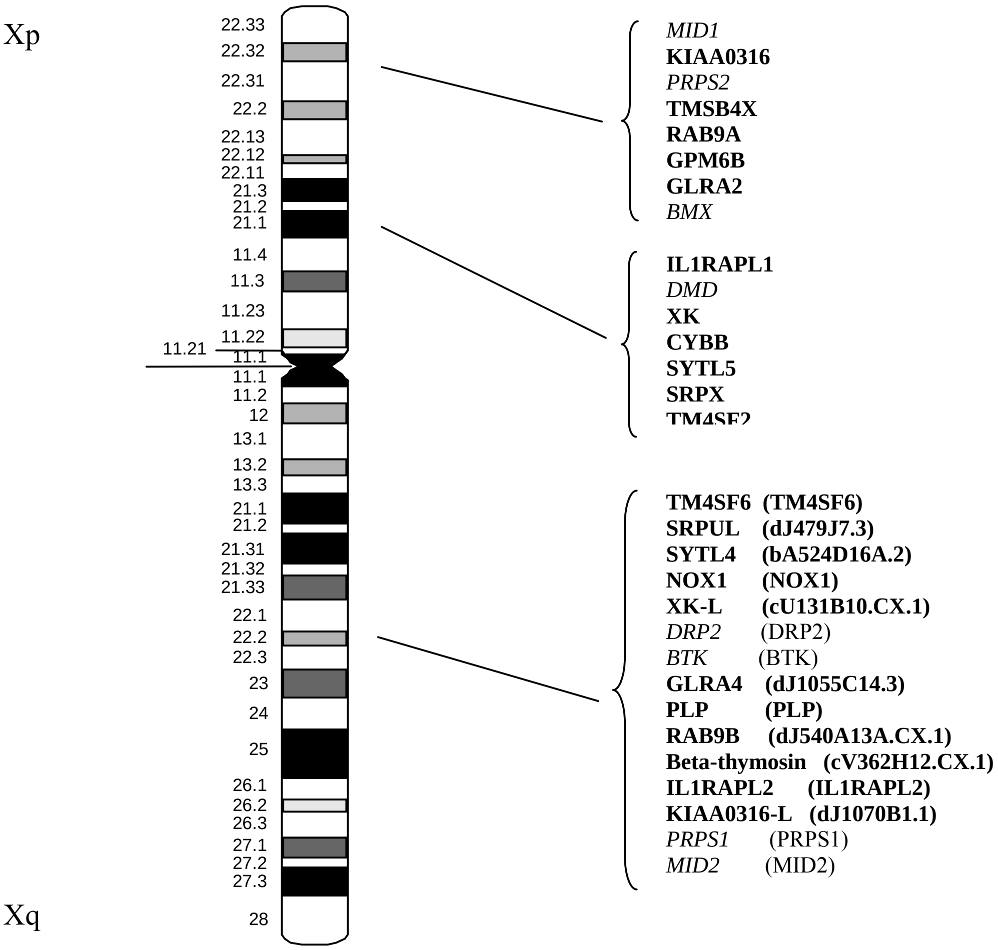
The availability of genomic sequence data has enabled several recent studies of sequence duplications within the human genome (McLysaght *et al.*, 2002), (Gu *et al.*, 2002). These genome-wide studies shed light on the extent of tandem and regional duplications within the human genome, and provide data on the temporal pattern of events and the respective contributions of tandem versus segmental duplications in increasing genome size and content.

During the process of identification of genes within Xq22-q23 described in Chapter 3, it was noted that several genes within Xq22 had paralogues on the X short arm (Xp). Initially, genes with similar names and descriptions were noted, for example MID1 and MID2. The presence of pairs of paralogues shared between the long and short arms of the human X chromosome has already been noted by Perry *et al.* (Perry *et al.*, 1999) in publications describing the MID2 gene (see Chapter 3). The number of gene-pairs noted and their order and direction of transcription strongly suggested a regional duplication leading to the paralogy noted. However, as no systematic characterisation of the extent of paralogy between the two regions has been described, one of the aims of the present study was to identify additional examples of Xp/Xq paralogue pairs.

The presence of paralogues on the short arm of the human X chromosome raises the question of their location in the marsupial genome, as some of the genes (DMD and CYBB) had been localised in the marsupial genome (Spencer *et al.*, 1991). As described in Chapter 1, much of the region represented by the short arm of the human X chromosome is found on an autosome in marsupials.

The work described in this chapter examines the extent of paralogy between Xq22-q23 and Xp, and the genes involved. In addition, the orthologues of the genes, and their chromosomal localisation in the marsupial mouse *Sminthopsis macroura* were investigated. Sequences from selected *Sminthopsis macroura* BAC clones containing orthologues were analysed and compared to the human sequence. Finally, evidence supporting an estimate of the age of the duplication event is presented, in order to place it in context with other studies of regional duplications.

Figure 6-1 Observations of Xp/Xq paralogues. Previously noted paralogues (Perry *et al.*, 1999) are in italic type, new observations are in bold type. Locus names assigned during annotation of Xq22 (Chapter 3) are given in parentheses.



6.2 Characterisation of the Xq22-q23/Xp regional duplication

6.2.1 Extent of the duplication and genes involved

As described in Chapter 3, 15 pairs of paralogues that were shared between Xp and Xq were found. The numbers of exons and exon sizes of the gene pairs were compared, because conservation of gene structure is compelling evidence for a true gene duplication rather than convergent evolution of sequences (Table 6-1). Ensembl and transcript map identifiers, mRNA and gene sizes, and measures of cDNA and protein homology are given in Table 6-1.

As can be seen in Table 6-1, exon size and order is very well conserved for most of the 15 paralogue pairs (a striking outlier is the discordant exon numbers of DMD and DRP2). This provides strong support for the hypothesis that they are true gene duplications. Nucleotide homology between paralogues within coding regions ranges from 54% (XK/XK-L) to 79% (PRPS2/PRPS1), and protein identity/similarity ranges from 43/63% (SYTL5/SYTL4) to 95/98% (PRPS2/PRPS1) (Table 6-2).

One notable feature also apparent from these data is that the gene size is smaller for each of the Xq22 genes in comparison to its Xp paralogue (apart from RAB9A and TMSB4X). Although caution is necessary in interpreting these data as some of the gene structures may be incomplete, it is suggestive of a systematic bias and worth further study when gene structure annotation is complete.

In order to be consistent with the hypothesis that the paralogue pairs arose as a result of a segmental duplication, gene pairs should display the same transcriptional direction and positioning with respect to their neighbours. Examination of the literature and the genomic sequences of the Xp and Xq22 regions shows that the majority of paralogue pairs share the same transcriptional orientation and position with respect to other genes (Figures 6-2 and 6-3).

It can be seen that most of the paralogue pairs are positioned similarly with respect to their neighbouring genes, and share transcriptional direction. There appears to have been a small inversion event involving the PRPS and KIAA0316 genes. The only other exceptions are the IL1RAPL genes, which also appear to have been involved in an inversion (or inversions) (Figures 6-2 and 6-3).

Gene	Xp/Xq	No. exons	Exon sizes (bp)																							
			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
MID1	Xp	10	130	**716**	96	108	149	128	144	162	208	**1609**														
MID2	Xq	10	201	**716**	96	108	149	128	240	162	208	**521**														
KIAA0316	Xp	16	212	117	161	103	46	105	108	132	120	137	127	90	198	139	1065	1289								
KIAA0316-L	Xq																									
PRPS2	Xp	7	**209**	184	99	125	174	160	**1514**																	
PRPS1	Xq	7	**244**	184	99	125	174	160	**1089**																	
TMSB4X	Xp	3	61	116	381																					
cV362H12.CX.1	Xq	3	51	**117**	**436**																					
RAB9A	Xp	1	**940**																							
RAB9B	Xq	3		169	74	**806**																				
GPM6B	Xp	7	**191**	187	157	172	74	66	**671**																	
PLP	Xq	7	**125**	187	262	169	74	66	**2054**																	
GLRA2	Xp	9	**598**	134	68	224	83	138	215	150	**1606**															
GLRA4	Xq	9	**71**	131	68	224	83	141	215	150	**282**															
BMX	Xp	18	**138**	105	82	120	65	242	78	54	55	80	128	75	172	217	65	119	162	**68**						
BTK	Xq	18	**141**	99	69	82	129	68	188	63	55	80	128	75	172	217	65	119	158	**500**						
IL1RAPL1	Xp	10		**82**	280	187	154	75	133	146	144	171	**719**													
IL1RAPL2	Xq	11	737	**101**	274	187	154	75	130	146	144	171	**866**													
DMD	Xp	78			190	173	157	121	269	147	79	61	62	75	202	86	158	167	112	137	39	66	66	159	244	124
DRP2	Xq	24	151	103	**180**	164	157	121	269	147	79	61	62	75	202	86	158	167	112	137	66	66	144	238	121	**432**
XK	Xp	3	**327**	263	**4495**																					
XK-L	Xq	3	239	269	**1639**																					
CYBB	Xp	13	**81**	96	111	85	146	191	130	93	254	163	147	125	**2671**											
NOX1	Xq	13	**251**	96	111	85	152	182	133	93	236	163	147	125	**187**											
SYTL5	Xp	16	**119**	210	116	109	135	142	130	101	93	179	100	162	109	136	209	**143**								
SYTL4	Xq	16	110	216	110	103	102	76	91	104	93	179	103	162	109	100	209	**1683**								
SRPX	Xp	10		**97**	60	192	177	127	122	180	134	122	**556**													
SRPUL	Xq	11	288	**212**	81	192	177	127	122	180	134	122	**493**													
TM4SF2	Xp	7	**150**	189	75	96	156	84	69																	
TM4SF6	Xq	8	**190**	189	75	99	135	84	**108**	1189																

Table 6-1 Gene structure information obtained from Ensembl v15.33.1 (based on the NCBI 33 assembly), and the Xq22-q23 transcript map described in Chapter 3. TMSB4X information was obtained from the UCSC genome browser. Dark row borders separate different Xp/Xq gene pairs. Exon sizes in red font are of equal size in each paralogue within the pair. Exon sizes in blue font differ by a multiple of 3 (preserving coding frame) between each paralogue within the pair. Exons in bold type contain the translation start and stop codons. N.B. To match the gene structure of SRPX with SRPUL, the SRPX gene structure was shifted 3' by one exon (i.e. SRPX exon 1 in Ensembl is allocated to the exon 2 column in the table above – it is possible that the mRNA for SRPX is incomplete). The DMD and DRP2 structures were also shifted accordingly, and only a portion of the DMD structure is shown. As some annotations are incomplete these figures may not represent complete gene structures, but are shown to illustrate exon size similarities.

Gene	Location	Ensembl gene identifier	Ensembl transcript identifier	mRNA cds % identity	protein % identity/ similarity	mRNA length (bp)	gene length (kb)
MID1 MID2	Xp Xq	ENSG00000101871 ENSG00000080561	ENST00000317552 ENST00000262843	70	76/89	3450 2529	172 101
KIAA0316 KIAA0316-L	Xp Xq	ENSG00000169933	ENST00000304087			4149	580
PRPS2 PRPS1	Xp Xq	ENSG00000101911 ENSG00000147224	ENST00000218027 ENST00000276174	79	95/98	2465 2075	33 23
TMSB4X cV362H12.CX.1	Xp Xq	UCSC browser *Xace*	UCSC browser *Xace*	66	68/88	558 *604*	2 *3.3*
RAB9A RAB9B	Xp Xq	ENSG00000123595 ENSG00000123570	ENST00000243325 ENST00000243298	71	76/88	940 1049	0.94 7
GPM6B PLP	Xp Xq	ENSG00000046653 ENSG00000123560	ENST00000050379 ENST00000303958	64	57/73	1518 2937	43 16
GLRA2 GLRA4	Xp Xq	ENSG00000101958 *Xace*	ENST00000218075 *Xace*	72	78/86	3216 *1365*	202 *21*
BMX BTK	Xp Xq	ENSG00000102010 ENSG00000010671	ENST00000311287 ENST00000308731	58	52/71	2025 2408	48 26
IL1RAPL1 IL1RAPL2	Xp Xq	ENSG00000169306 ENSG00000182513	ENST00000302196 ENST00000331930	66	61/80	2091 2061	1170 1110
DMD DRP2	Xp Xq	ENSG00000132438 ENSG00000102385	ENST00000275952 ENST00000263029	60	53/72	11016 2865	1890 29
XK XK-L	Xp Xq	ENSG00000047597 *Xace*	ENST00000051619 *Xace*	54	44/68	5085 *2147*	46 *14.8*
CYBB NOX1	Xp Xq	ENSG00000165168 ENSG00000007952	ENST00000297870 ENST00000217885	62	59/73	4293 1961	33 30
SYTL5 SYTL4	Xp Xq	ENSG00000147041 ENSG00000102362	ENST00000297875 ENST00000276141	58	43/63	2193 3550	93 28
SRPX SRPUL	Xp Xq	ENSG00000101955 ENSG00000102359	ENST00000218072 ENST00000263031	55	44/65	1767 2128	71 27
TM4SF2 TM4SF6	Xp Xq	ENSG00000156298 ENSG00000000003	ENST00000286824 ENST00000003603	63	61/78	819 2069	126 8

Table 6-2 Sequence and structural comparisons of paralogous gene pairs. Gene and transcript identifiers are taken from Ensembl v15.33.1 (based on the NCBI 33 assembly). Percentage identity between mRNAs in the coding region and identity/similarity of protein sequences were calculated as described in Chaper 2. mRNA and gene lengths were derived from Ensembl v15.33.1, or Xace (italics). TMSB4X information was obtained from the UCSC genome browser. As annotation for KIAA0316-L was incomplete, no comparison was made.

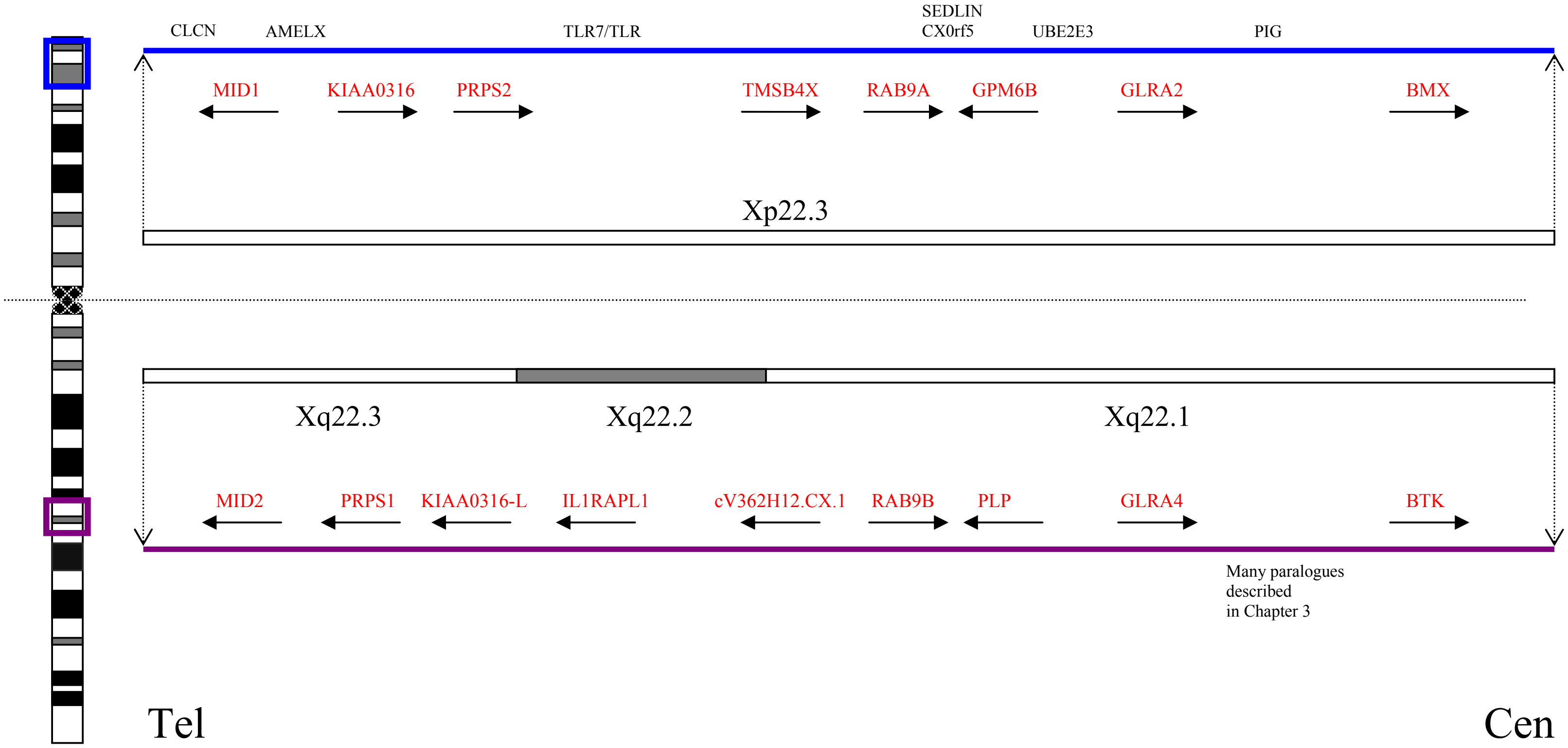


Figure 6- 2 Schematic representation of paralogy between Xp22.3 and Xq22.1-q23 (Block 1). Paralogous genes are represented in red type, with their direction of transcription depicted by a black arrow. Genes are shown in their order along the chromosome (Tel to Cen) relative to one another. Xp genes are represented above the dotted line, Xq genes below. Gene names in black represent selected non-paralogous genes whose positions are shown to provide context.

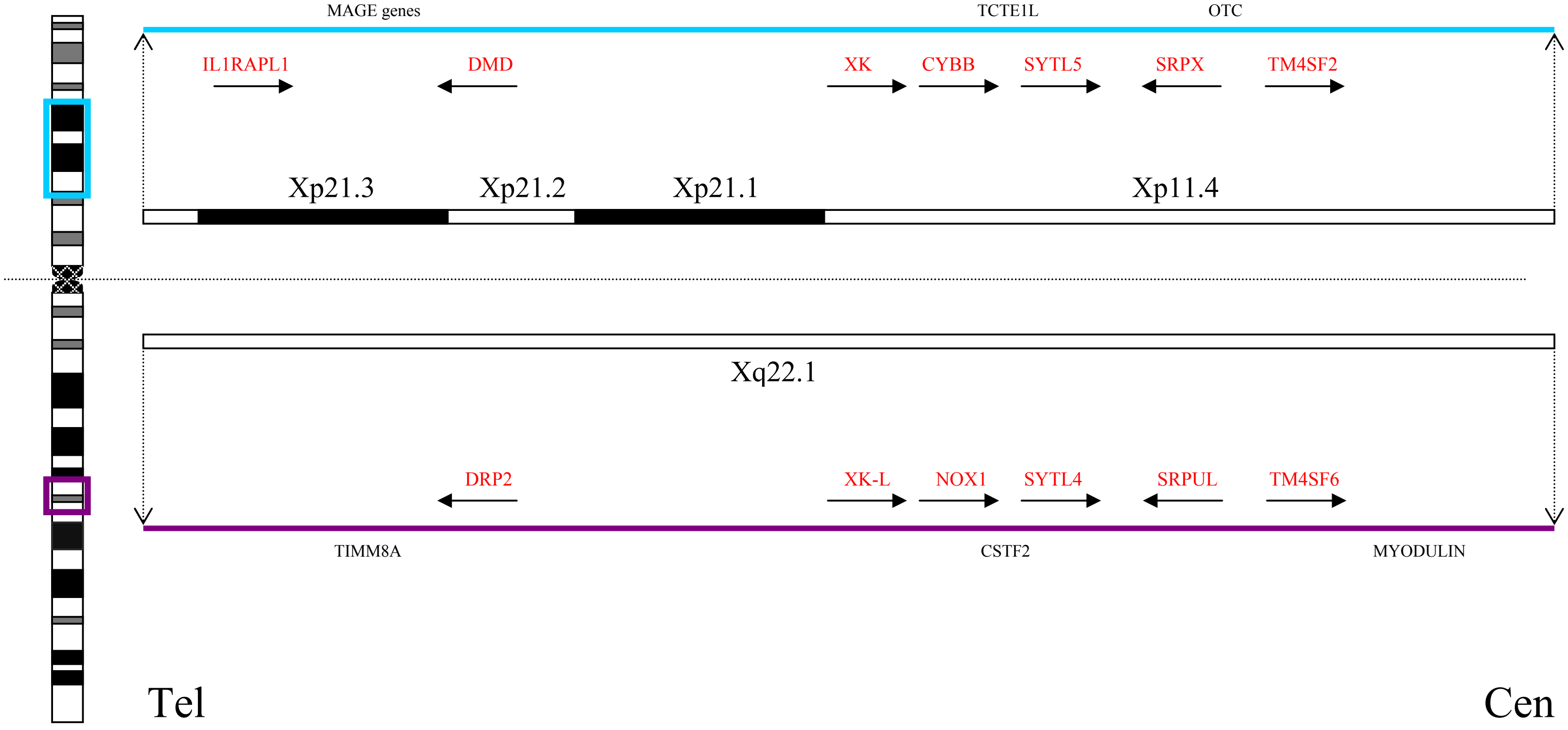


Figure 6-3 Schematic representation of paralogy between Xp21.3-p11.4 and Xq22.1 (Block 2). Paralogous genes are represented in red type, with their direction of transcription depicted by a black arrow. Genes are shown in their order along the chromosome (Tel to Cen) relative to one another. Xp genes are represented above the dotted line, Xq genes below. Gene names in black represent selected genes whose positions are shown to provide context

Examination of genomic sequence information in Ensembl and of members of gene families showed there existed several examples of autosomal paralogues of Xp/Xq genes. Observations are depicted schematically in Figure 6-4.

Several paralogues of Xp genes (e.g. TMSB4Y and XKRY), are seen on the Y chromosome. This would be consistent with the hypothesis that an autosomal block was added to an ancestral pair of sex chromosomes early in the eutherian mammal lineage, which subsequently evolved into the X and Y chromosomes, and with a model in which the genes were part of the original autosome pair that became the X and Y chromosomes.

Some autosomal paralogues retain linkage to one another reflecting their X chromosome counterparts. One example is the UTROPHIN, NOX3, TCTE1 and SYTL3 genes on chromosome 6. They are linked similarly to DMD, CYBB, TCTE1L and SYTL5 on Xp, 3 of which are part of the proposed Xp/q segmental duplication. This suggests that these paralogues were also generated as part of a segmental duplication.

The presence of X chromosome paralogues on the autosomes suggests that further duplications involving genes generated by the Xp/q segmental duplication have occurred, although without further analysis the order of these is unclear. Initial observations also suggest that some of these were also generated by further segmental duplications rather than single gene duplications, as shared synteny is seen for some of the paralogues (e.g. DMD/CYBB/TCTE1L/SYTL5 on chromosome X and UTRN/NOX3/TCTE1/SYTL3 on chromosome 6). Another possibility is that loss of genetic material from the Y chromosome to an autosome occurred during degradation of the Y, which would not require a duplication event.

It is clear that different hypotheses are possible here, and further studies on the genes involved and the extent of the autosomal paralogy with both X and Y would shed further light on the events that generated these regions of the genome, but were not considered further as part of this study due to time constraints.

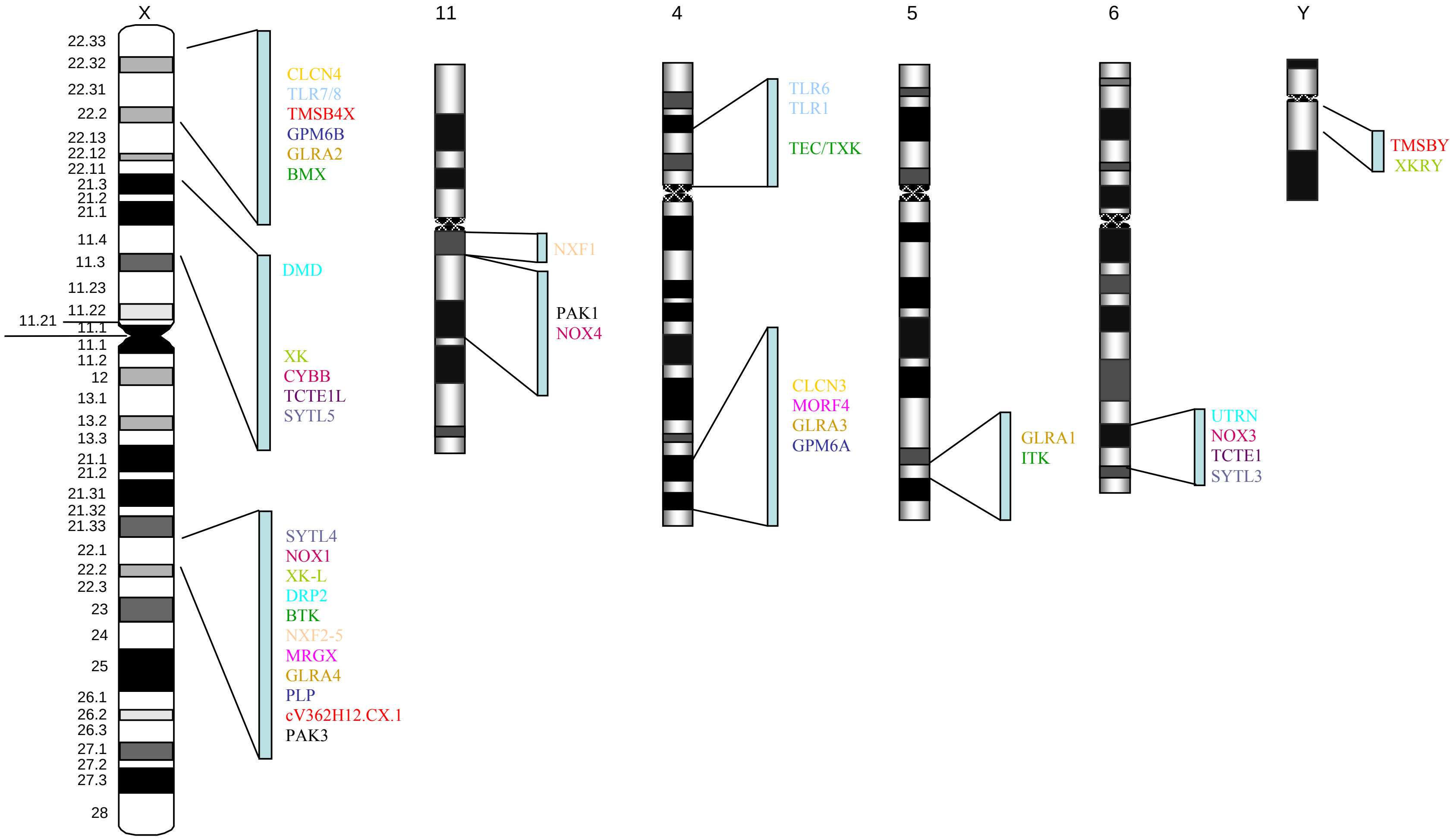


Figure 6-4 Schematic representation of chromosomal locations of autosomal genes with paralogues on the X chromosome, some of which are Xp/Xq paralogues. Names are coloured according to similarity.

6.3 Identification of orthologues of the duplicated genes in the marsupial mouse, *Sminthopsis macroura*

Numerous marsupial orthologues of human genes have previously been isolated using a variety of methods. Sequence information is available for some, and the chromosomal location of many has been determined. These studies have demonstrated that whilst the X chromosome is well conserved with respect to content in eutherian mammals, much of the region represented by human Xp is autosomal in metatherian mammals. This section describes attempts to isolate *Sminthopsis macroura* BAC clones containing orthologues (or parts thereof) of Xp/Xq paralogous genes. These BAC clones could then be localised in the marsupial genome by FISH to determine if Xp paralogues are autosomal as predicted, and sequenced for comparative analysis with human genomic sequence.

A reduced-stringency hybridisation approach was adopted to isolate orthologues of human X chromosome genes involved in the Xp/Xq regional duplication using a genomic BAC library from a male marsupial mouse *Sminthopsis macroura* (Chapman *et al.*, 2003). The library was prepared from the liver of a 20-week old male, and comprised 110,592 clones with an average insert size of 60 kb. Genomic coverage was predicted to be two to three-fold. The hybridisation procedure used for the BAC library screen is described in Chapter 2, and was based on personal communications from Jim Thomas describing his procedures for screening rat genomic DNA libraries (Thomas *et al.*, 2002).

Human DNA probes were designed with the following aims in mind, trying to balance designing probes that would detect marsupial clones whilst attempting to avoid numerous false positives due to the reduced-stringency conditions employed:

- Maximise sequence conservation between species to increase true positives, by aligning nucleotide sequences, annotating exon/intron boundaries and designing STSs to well conserved regions.
- Use coding exon sequences to achieve maximum cross-species conservation
- Minimise location of probes within regions encoding promiscuous protein domains to avoid false-positives from homologous sequences

- Where possible, for paralogous loci design the probe in a common region of the gene structure, to avoid isolation of non-overlapping clones from the same locus with both paralogue probes.
- Avoid repetitive regions.

Human probes were used rather than mouse sequences, as there is some evidence that mouse genomic DNA sequences evolve at a faster rate, thereby potentially reducing sequence conservation with a marsupial orthologue. For example, for the MID2 gene, initially the human and mouse genes' coding regions were aligned (Figure 6-5). Exon/intron boundaries were then annotated, using information from the transcript maps presented in Chapter 3 or the Ensembl web-server (shown by blue arrows in Figure 6-5). The encoded peptide was analysed using InterPro and domain boundaries were annotated (shown by dashed lines underneath the alignment in Figure 6-5).

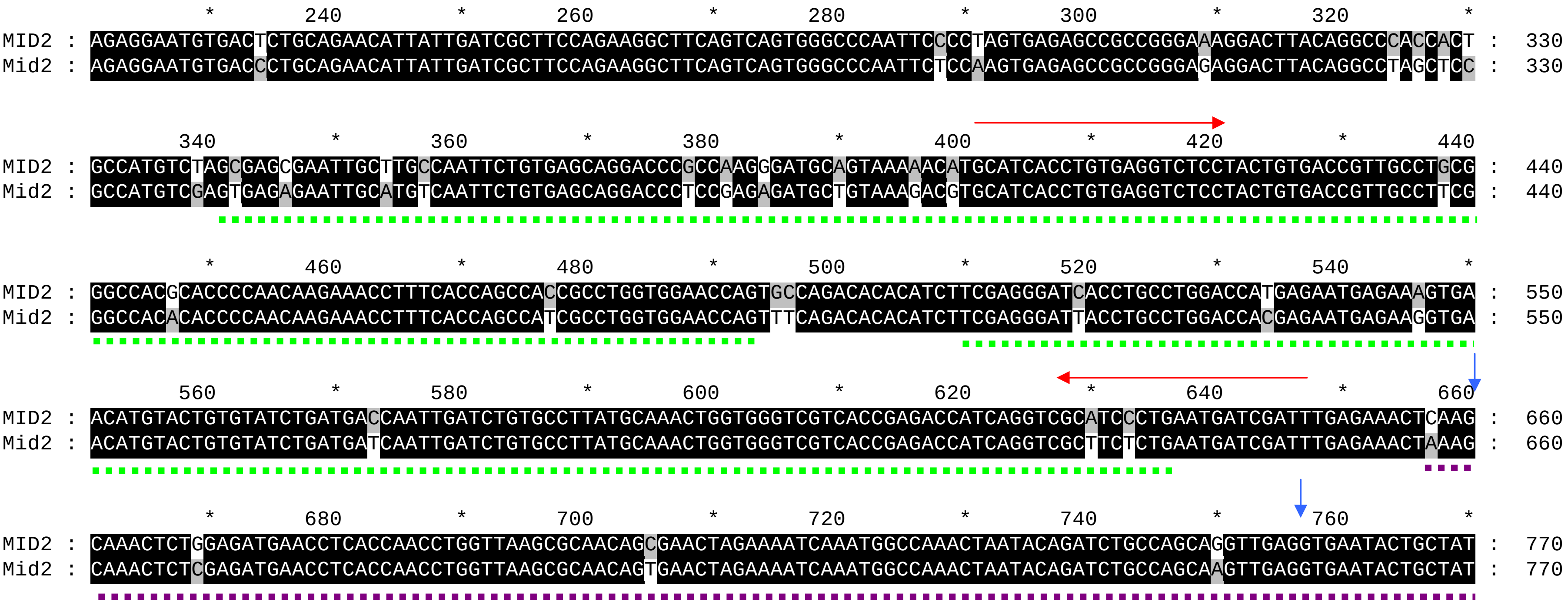
In Figure 6-5, the green line represents domain IPR000315 (Zn-finger B-box, matches 385 proteins) and the purple line domain IPR003649 (Bbox_C, matches 66 proteins). Although encoding protein domains, this region was chosen as further 3', domains with a higher number of protein matches were found. Primers were then designed using Primer3 (shown by red arrows above the alignment for stSG407305). Primers were selected which were contained within a coding exon in a region conserved between human and mouse, but avoiding regions encoding commonly found protein domains were selected. Wherever possible, predicted product sizes were kept between 80-500 bp to try to achieve similarities in probe labelling efficiency. This strategy for probe design attempted to balance sensitivity and specificity. Thus, positive clones were expected due to design of probes to conserved sequences, but it may also result in cross-hybridisation being observed between paralogue pairs.

The primer sequences designed and associated information are given in Appendix D. The genes selected for screening and their positions on the human X chromosome are shown in Figure 6-6. The genes include Xp/Xq paralogue pairs and also genes from intervening non-paralogous segments in Xp and Xq (to assess whether they are also present in similar organisation in the marsupial genome, or may represent subsequent insertions).

Primer pairs designed were pre-screened to establish optimal reaction conditions and to confirm localisation of the STS to the human X chromosome. STS pre-screens were performed on the following templates: human genomic DNA, clone 2D (a human-hamster cell hybrid containing the human X chromosome), hamster genomic DNA and $T_{0.1}E$. Pre-screens were performed using three different primer annealing temperatures (55°C, 60°C and 65°C) to determine the cycling parameters that give a visible and specific DNA product.

A total of 40 probes, each representing a single gene, were used to screen the *Sminthopsis macroura* BAC library. Probes were pooled in groups of five (separating paralogue pairs as much as possible to aid interpretation of results in cases of cross-hybridisation) and hybridised to the genomic clone filters at 58^{0}C for greater than 16 hours before washing at a final stringency of 1 x SSC, 1% sarkosyl for 30 minutes at 58^{0}C. An example of the screening is shown in Figure 6-7. A total of 157 positive clones were identified. These positive BAC clones were picked from the library, and re-gridded onto nylon filters (gridding performed by Paul Hunt, Sanger Institute Clone Resources Group).

These filters were then screened using individual probes in order to establish the probe-clone relationships. At this secondary screen stage, the probes were hybridised to the filters as above, then washed to three different levels of stringency in an attempt to reduce the false positive rate. This was achieved by washing first to a final stringency of 1x SSC, 1% sarkosyl for 30 minutes at 58^{0}C and visualising positive clones by autoradiography, then re-washing as above but with 0.5x SSC and then 0.2x SSC. An example of this is shown in Figure 6-8. Results from this secondary screening procedure are given in Table 6-3. A summary of the screening results is given in Table 6-4. Full protocol details are given in Chapter 2.



```
                *       240         *       260         *       280         *       300         *       320         *
MID2 : AGAGGAATGTGACTCTGCAGAACATTATTGATCGCTTCCAGAAGGCTTCAGTCAGTGGGCCCAATTCCCCTAGTGAGAGCCGCCGGGAAAGGACTTACAGGCCCACCACT :  330
Mid2 : AGAGGAATGTGACCCTGCAGAACATTATTGATCGCTTCCAGAAGGCTTCAGTCAGTGGGCCCAATTCTCCAAGTGAGAGCCGCCGGGAGAGGACTTACAGGCCTAGCTCC :  330

              340         *       360         *       380         *       400         *       420         *       440
MID2 : GCCATGTCTAGCGAGCGAATTGCTTGCCAATTCTGTGAGCAGGACCCGCCAAGGGATGCAGTAAAAACATGCATCACCTGTGAGGTCTCCTACTGTGACCGTTGCCTGCG :  440
Mid2 : GCCATGTCGAGTGAGAGAATTGCATGTCAATTCTGTGAGCAGGACCCTCCGAGAGATGCTGTAAAGACGTGCATCACCTGTGAGGTCTCCTACTGTGACCGTTGCCTTCG :  440

                *       460         *       480         *       500         *       520         *       540         *
MID2 : GGCCACGCACCCCAACAAGAAACCTTTCACCAGCCACCGCCTGGTGGAACCAGTGCCAGACACACATCTTCGAGGGATCACCTGCCTGGACCATGAGAATGAGAAAGTGA :  550
Mid2 : GGCCACACACCCCAACAAGAAACCTTTCACCAGCCATCGCCTGGTGGAACCAGTTTCAGACACACATCTTCGAGGGATTACCTGCCTGGACCACGAGAATGAGAAGGTGA :  550

              560         *       580         *       600         *       620         *       640         *       660
MID2 : ACATGTACTGTGTATCTGATGACCAATTGATCTGTGCCTTATGCAAACTGGTGGGTCGTCACCGAGACCATCAGGTCGCATCCCTGAATGATCGATTTGAGAAACTCAAG :  660
Mid2 : ACATGTACTGTGTATCTGATGATCAATTGATCTGTGCCTTATGCAAACTGGTGGGTCGTCACCGAGACCATCAGGTCGCTTCTCTGAATGATCGATTTGAGAAACTAAAG :  660

                *       680         *       700         *       720         *       740         *       760         *
MID2 : CAAACTCTGGAGATGAACCTCACCAACCTGGTTAAGCGCAACAGCGAACTAGAAAATCAAATGGCCAAACTAATACAGATCTGCCAGCAGGTTGAGGTGAATACTGCTAT :  770
Mid2 : CAAACTCTCGAGATGAACCTCACCAACCTGGTTAAGCGCAACAGTGAACTAGAAAATCAAATGGCCAAACTAATACAGATCTGCCAGCAAGTTGAGGTGAATACTGCTAT :  770
```

Figure 6-5 Strategy for design of primers to amplify probes for use in a reduced-stringency hybridisation approach to identify *Sminthopsis macroura* BAC clones, using MID2 as an example. Key – blue arrows represent exon/intron boundaries, red arrows primers designed (stSG407305), green dashed lines the region encoding a Zn-finger B-Box domain and the purple dashed lines the region encoding a Bbox_C domain.

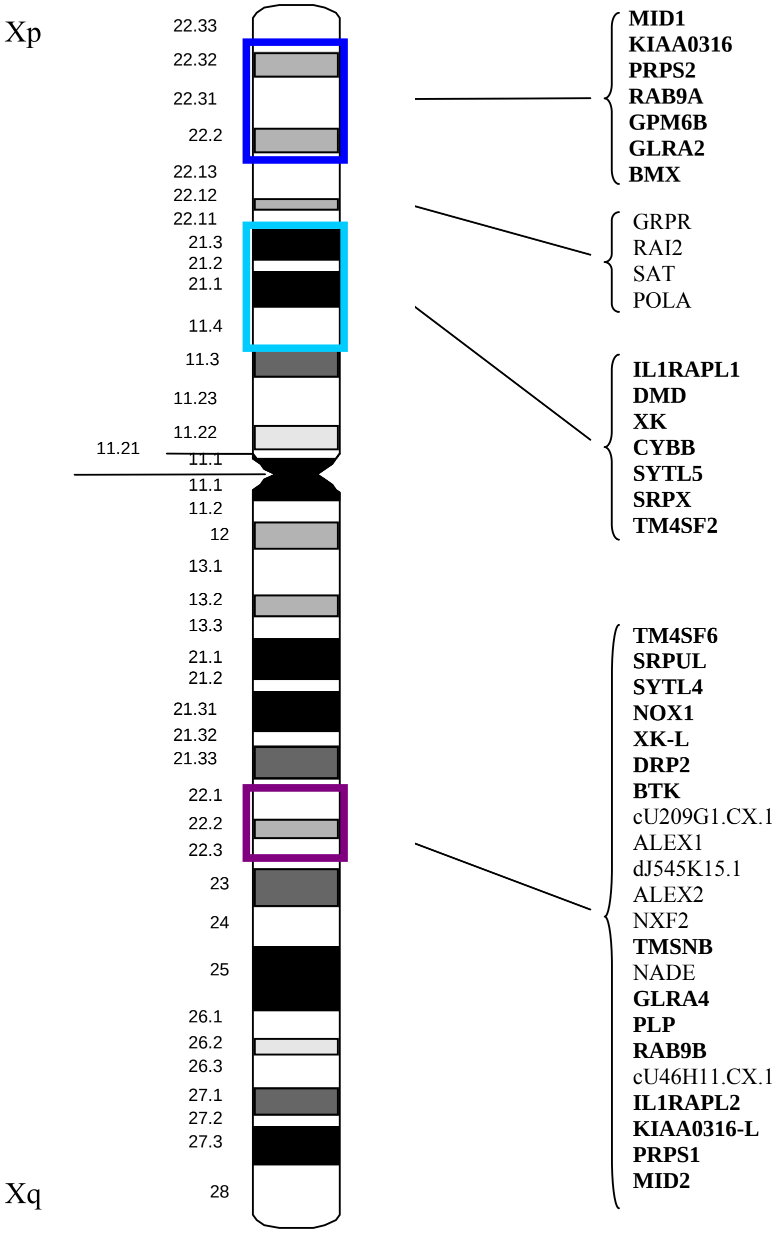


Figure 6-6 Diagram showing the genes for which probes were designed to identify orthologues in *Sminthopsis macroura*, and their positions on the human X chromosome. The genes are listed in order from Xpter to Xqter. The main blocks of Xp/Xq paralogy are denoted by the red, purple and green boxes on the chromosome ideogram. Xp/Xq paralogue gene names are shown in bold.

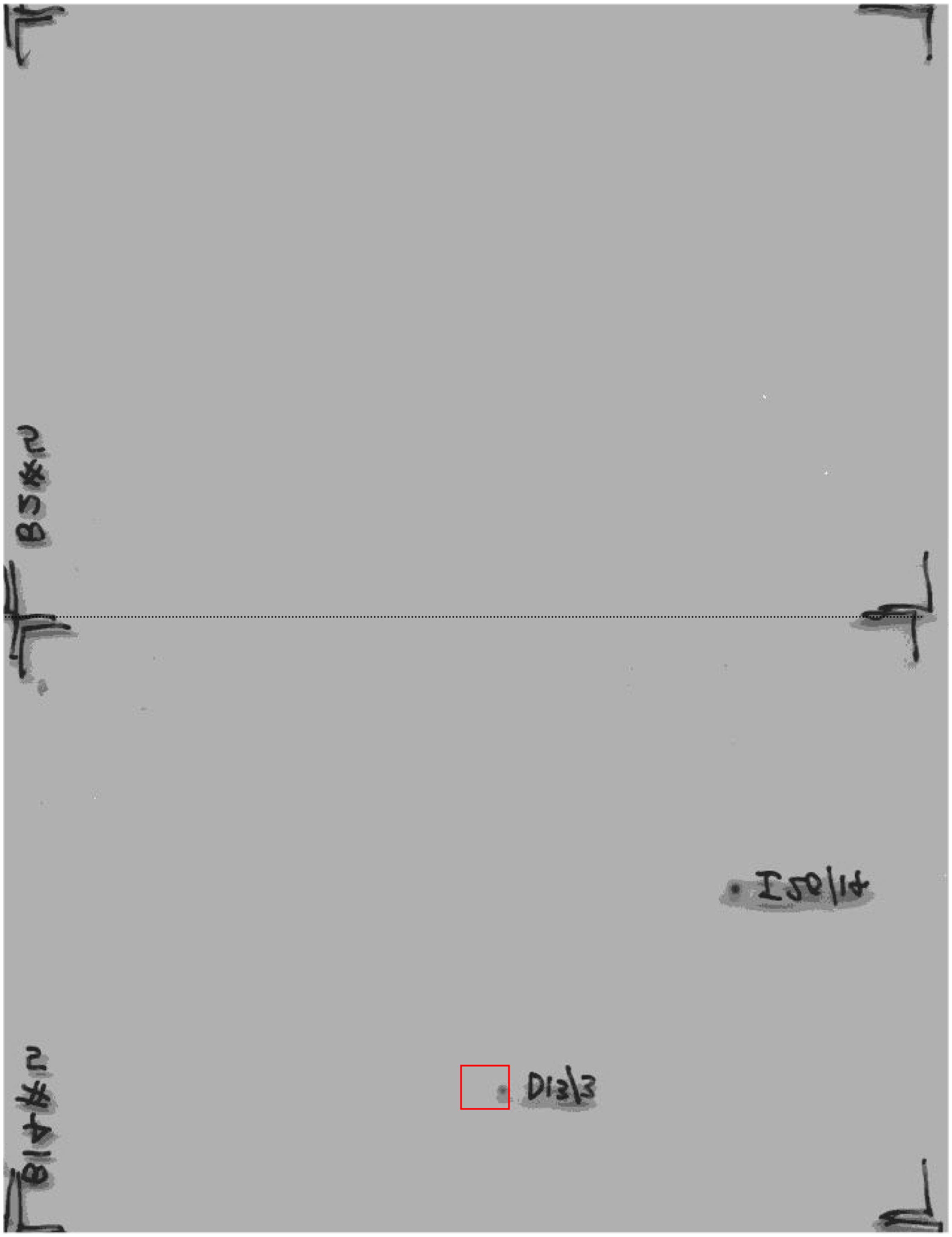


Figure 6-7 An example of a hybridisation of a pool of five probes to filters of the *Sminthopsis macroura* library. The diagram shows two filters of the gridded library (separated by a dotted line) following hybridisation of a pool of five STSs and washing as described in the text. The four corner edge positions of the filters were noted as seen to facilitate scoring. The positive signal on the lower filter in a red box marked “D13/3” represents clone bF211D13.

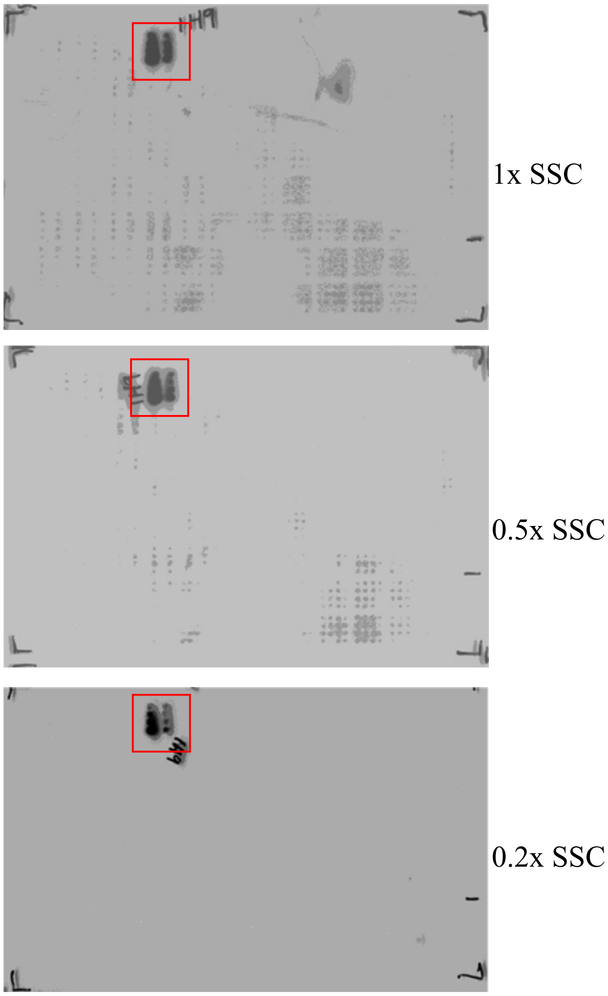


Figure 6-8 An example of the second round of the reduced-stringency hybridisation procedure. Three images of autoradiographs are shown, following hybridisation with a probe generated to the MID1 gene (stSG187894), and filters washed at increasing stringency (1x SSC, 0.5x SSC and 0.2x SSC). The red box highlights positive signal seen for clone bF134C3.

Gene	positive BAC clones
MID1	bF134C3(+++)
PRPS2	bF48C16(+++), bF14N15(+++), bF225I7(+++)
RAB9A	bF147M18(+++), bF89O16(+++), bF244J18(+++), bF65C12(+++), bF20I20(+++), bF144L7(+++), bF265N1(+++), bF45F5(+)
GPM6B	bF153M3(+++)
IL1RAPL1	bF272K20(+++), bF58F4(+++)
CYBB	bF242G1(+++)
SRPX	bF253J14(+++), bF243F20(++), bF252K3(+), bF281H15(+)
TM4SF2	bF99F22(+++), bF39A10(+)
GLRA2	bF149E6(+++), bF139K18(+++), bF50E16(+++), bF36H3(++), bF68P17(++), bF20L6(++), bF111F19(v. weak), bF150F1(+), bF158I6(+), bF65C12(+)
XK	bF255P10(+++), bF78P20(+++), bF231M3(+++), bF123F16(+++), bF255O10(+++), bF135B17(+++)
SYTL5	bF253J14(+++)
SAT	bF211D13(+++)
POLA	bF124A24(+++), bF284I24(+++)
RAI2	bF222I20(+++), bF185E13(+++), bF157M9(+++), bF113I16(+++), bF124C13(v. weak), bF134C3(v. weak)
GRPR	bF146K19(++), bF103A22(++)
ALEX2	NONE
NXF2	bF283J5(+)
NADE	NONE
BMX	NONE
KIAA0316	bF232B10(+++), bF238P19(+++), bF182C13(++), bF182E24(++), bF136O10(+), bF105A9(+)
DMD	bF125G2(+++)
MID2	bF134C3(+++)
PRPS1	bF48C16(+++), bF14N15(+++), bF225I7(+++), bF76L17(+), bF115J9(+)
RAB9B	bF244J18(++), bF265N1(++), bF89O16(++), bF144L7(+)
PLP	NONE (one spot very weak - bF159E15)
BTK	bF168K3(+++)
IL1RAPL2	bF272K20(+), bF48C16(v. weak)
DRP2	bF28C20(+++), bF154M12(+++)
NOX1	bF41P23(+), bF242G1(+), bF106P8(+), bF269L5(+), bF37C21(+), bF177E15(+)
SRPUL	bF281H15(+++), bF99K21(v. weak), bF228K24(++), bF106P8(+)
TM4SF6	bF93H4(+), bF127E19(+)
SYTL4	bF281H15(+++), bF186D19(+), bF231E16(+), bF77O5(v. weak), bF24K10(++), bF124C13(+), bF165M23(v. weak), bF97A19(+), bF13K23(v. weak), bF34O3(v. weak)
XKL	bF106P8(+++)
GLRA4	bF149E6(+)
KIAA0316-L	bF34O3(+++), bF13K23(+++), bF104N15(++), bF57H4(+), bF49K3(+), bF53G1(+)
TMSNB	NONE
cU46H11.CX.1	bF6N3(+++), bF191I22(++), bF82H3(+), bF107F9(+), bF167J13(v. weak)
dJ545K15.1	bF21K1(++)
ALEX1	NONE
cU209G1.CX.1	NONE

Table 6-3 Table showing results from the second round of *Sminthopsis macroura* BAC library screening after increasing stringency washes. The clone names are followed by an indication of the strength of the signal seen on the autoradiograph after the most stringent wash: +++ strong; ++ medium; + weak. Clones in blue are those remaining after the 0.5x SSC wash. Clones in red are those remaining after the 0.5x SSC and 0.2x SSC washes.

Gene	% Mm ID	% incorp.	Probe size (bp)	Number of BAC clones scored		
				1x SSC	0.5x SSC	0.2x SSC
MID1	92	45	307	1	1	1
KIAA0316	na	61	149	6	6	6
PRPS2	87	67	127	3	3	3
RAB9A	na	27	308	8	8	8
GPM6B	94	31	171	1	1	1
GLRA2	90	70	184	9	8	7
BMX	88	47	104	0	0	0
GRPR	91	54	174	2	2	2
RAI2	95	35	209	4	4	4
SAT	93	59	149	1	1	1
POLA	93	56	155	2	2	2
IL1RAPL1	na	33	253	2	2	2
DMD	100	45	102	1	1	1
XK	87	57	304	6	6	6
CYBB	89	61	235	1	1	1
SYTL5	na	38	194	1	1	1
SRPX	89	65	121	4	3	3
TM4SF2	90	63	187	2	2	2
TM4SF6	87	43	180	2	2	1
SRPUL	86	68	107	4	1	2
SYTL4	82	36	169	10	8	3
NOX1	88	55	152	6	6	6
XK-L	na	27	176	1	1	1
DRP2	92	71	180	2	2	2
BTK	94	56	125	1	1	1
cU209G1.CX.1	90	14	212	0	0	0
ALEX1	91	28	307	0	0	0
dJ545K15.1	82	38	285	1	1	1
ALEX2	90	30	280	0	0	0
NXF2	82	66	100	1	1	0
TMSNB	na	62	94	0	0	0
NADE	90	59	104	0	1	0
GLRA4	90	61	259	1	0	0
PLP	98	53	246	1	0	0
RAB9B	na	32	300	4	3	3
cU46H11.CX.1	90	34	282	4	3	4
IL1RAPL2	97	51	188	2	0	0
KIAA0316-L	na	64	128	6	6	4
PRPS1	94	49	144	5	3	3
MID2	96	66	247	1	1	1

Table 6-4 Results from the *Sminthopsis macroura* BAC library screening. Genes are listed in order Xpter-Xqter. The % nucleotide identity between the human probe sequence and the corresponding mouse cDNA sequence where available, % incorporation of radioactivity in the probes used for the first round of screening, probe size in bp and number of positive BACs obtained for each clone after each stringency wash (performed at 58^0C) are given.

The reduced-stringency hybridisation strategy gave positive clones for 30 of the 40 genes selected (as counted after the 0.2xSSC wash). The number of clones obtained per gene, after the 0.2xSSC wash, ranged from 0 to 8, with the average number of clones for probes that gave positive results being 2.7 (calculated for numbers obtained after the 0.2x SSC wash as these are more likely to represent true positives). The number of positives corresponds approximately to that expected, as the library was estimated to provide two to three-fold genome coverage (Chapman *et al.*, 2003). Following the primary screens, there were 157 positive clones, which indicates that the subsequent stringency washes did succeed in removing more weakly-hybridising sequences.

As shown in Table 6-3, for many genes, the increase in wash stringency did not result in a reduction of clones scored, thus increasing confidence that those clones represent true positives and that the sequence conservation appears to be strong between human and marsupial. For some genes (SRPUL, NADE and cU46H11.CX.1), clones were scored at increased stringency conditions where fewer or no positives were scored under less stringent conditions. These instances reflect the detection of weak signals and presumably represent instances where minor differences in exposure times for the autoradiography have resulted in weaker signals being detected after one set of wash conditions, but not another.

For other genes, a reduction in the number of clones scored positive with increased wash stringency was seen. This was most apparent for SYTL4, where 10 clones were scored positive after a 1x SSC wash, but only 3 after a 0.2x SSC wash. This improved confidence that the number of clones remaining after the 0.2x SSC wash represented true positives (either the orthologue or the paralogue).

For some pairs of genes, probes from the two paralogues detected common positive clones. These genes and the clones detected are given in Table 6-5.

These data illustrate two points about the procedure adopted; firstly that the hybridisation conditions employed allowed probes from different paralogues to detect the same marsupial sequence, showing that the procedure was proving to be sufficiently sensitive, at least for some levels of sequence conservation. Secondly, the observation that some of these clones were not scored positive, or decreased in signal intensity, after increased wash stringencies demonstrates that the procedure adopted was also successful in decreasing false positives detected in at least some instances, for example SRPX/SRPUL. In other instances, such as for MID1/MID2, PRPS1/2 and RAB9A/RAB9B, increased wash stringency still failed to discriminate between the paralogues. These three pairs of paralogues are particularly well conserved at the mRNA level (Table 6-2). In these instances, it is likely that the marsupial sequence being detected is equally similar to either paralogue, or that the hybridisation kinetics are particularly favourable for interaction of the probe and target sequence, even at increased wash stringencies. Here, altering other stringency parameters such as increasing the wash temperature may have been effective.

Some clones were found in common between genes that were not paralogue pairs (Table 6-6). Of the relationships listed in Table 6-6, signals seen for some of the probes were very weak, and may represent commonality of a minor undetected repeat within the probes, rather than a true physical linkage for the genes. This is the case for genes whose probes detected bF48C16 and bF159E15. Other signals were more substantial, suggesting physical linkage of the genes whose probes detected the clone, such as for bF106P8, bF253J14, bF281H15 and to a lesser extent bF13K23 and bF34O3. This indicated that the genes involved were physically closely linked. This information also increased confidence that those clones represented true positives for the respective genes.

This was consistent for example with the close proximity of SRPX and SYTL5 in human, and also SRPUL and SYTL4 (whose 3' UTRs are separated by only ~ 4 kb). Thus a BAC clone, even from a library with an average insert size of 60 kb, could span such loci. However in the human SYTL4 and KIAA0316-L for example are much further separated, and would not be expected to fall within a single BAC.

In order to assess further the relationships between different maraupial genes, all of the BACs isolated in the first round of BAC library screening were subjected to *Hin*d III/*Sau* 3AI fluorescent fingerprinting to detect clone overlaps (Gregory *et al.*, 1997). This approach could also provide further information regarding the hybridisation positives, in order to determine if positive clones for a particular probe came from one locus.

*Hin*d III agarose fingerprinting (Marra *et al.*, 1997) has become the method of choice for large-scale projects such as the mouse and zebrafish genome mapping projects. However, as the average insert size of the *S. macroura* BAC clones was estimated to be only 60 kb (Chapman *et al.*, 2003), *Hin*d III/*Sau* 3AI fluorescent fingerprinting was chosen. This technique was expected to yield more fragments per clone than *Hin*d III fingerprinting and thus to be more informative.

Fingerprinting and fingerprint analysis were performed as described in Chapter 2. Selected contigs containing clones that were positive after the most stringent wash in the hybridisations are given in Table 6-7.

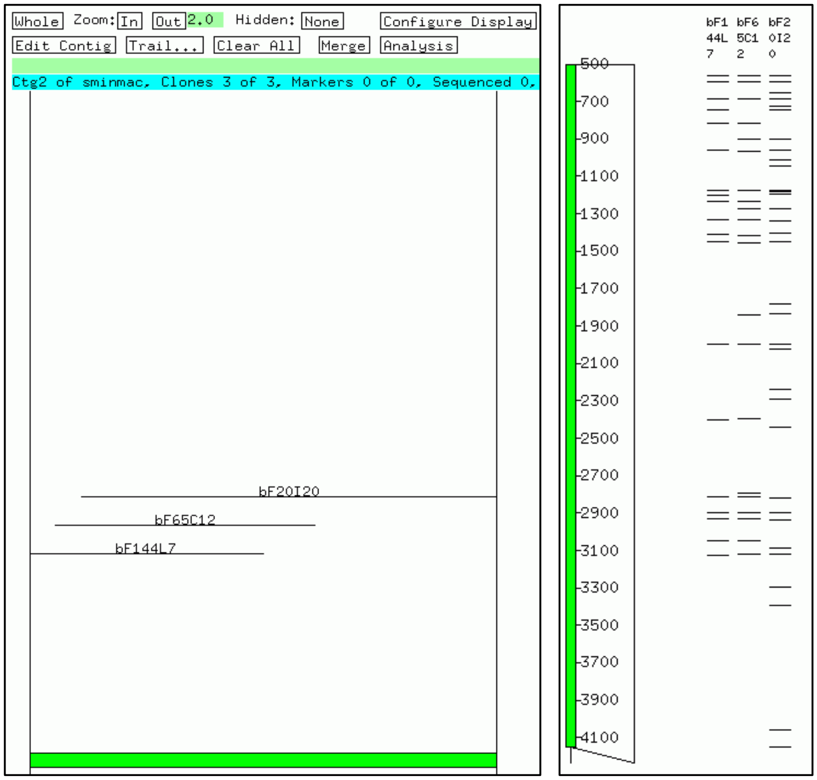
The 157 fingerprints were assembled into contigs in FPC (Chapter 2). Fingerprinting resulted in the incorporation of 37 clones into 11 contigs. It is possible that more contigs may have been generated by lowering the stringency parameters for contig formation, however already one of the contigs, contig 7, suggested that repeats may be present causing clones to appear to overlap, because probes for KIAA0316-L, SYTL4, TM4SF6, GLRA2 and cU46H11.CX.1 were positive for clones in both contigs. These genes are relatively widely separated within human Xq22-q23 (see Chapter 3), suggesting contig 7 may be an artefact. An example of an FPC contig and the associated clone fingerprints is shown in Figure 6-9.

Gene	1xSSC positive clones	0.5x SSC positive clones	0.2x SSC positive clones
MID1	bF134C3(+++)	bF134C3(+++)	bF134C3(+++)
MID2	bF134C3(+++)	bF134C3(+++)	bF134C3(+++)
PRPS2	bF48C16(+++), bF14N15(+++), bF225I7(+++)	bF48C16(+++), bF14N15 (+++), bF225I7 (+++)	bF48C16(+++), bF14N15 (+++), bF225I7 (+++)
PRPS1	bF48C16(+++), bF14N15(+++), bF225I7(+++)	bF48C16(+++), bF14N15(+++), bF225I7(+++)	bF48C16(+++), bF14N15(+++), bF225I7(+++)
RAB9A	bF89O16(+++), bF244J18(+++), bF144L7(+++), bF265N1(+++)	bF244J18(+++), bF89O16(+++), bF265N1(+++), bF144L7(+++)	bF89O16(+++), bF244J18(+++), bF265N1(+++), bF144L7(+++)
RAB9B	bF244J18(++), bF265N1(++), bF89O16(++), bF144L7(+)	bF265N1(++), bF244J18(++), bF89O16(++)	bF244J18(++), bF265N1(++), bF89O16(++)
IL1RAPL1	bF272K20(+++)	bF272K20(+++)	bF272K20(+++)
IL1RAPL2	bF272K20(+)		
CYBB	bF242G1(+++)	bF242G1(+++)	bF242G1(+++)
NOX1	bF242G1(+++)	bF242G1(++)	bF242G1(+)
SRPX	bF281H15(+)		
SRPUL	bF281H15(+++)	bF281H15(+++)	bF281H15(+++)
GLRA2	bF149E6(+++)	bF149E6(+++)	bF149E6(+++)
GLRA4	bF149E6(+ - weak)		

Table 6-5 Paralogous gene pairs for which their respective probes detected clones in common. Clones names in red represent the clones detected by either paralogue probe, clone names in black represent a clone that is still detected by one of the probes, after it fails to be detected by the second probe following an increase in the wash stringency. The clone names are followed by an indication of the strength of the signal seen on the autoradiograph: +++ strong; ++ medium; + weak.

Clone name	Genes whose probes detected the same clone
bF48C16	PRPS2 (+++) or PRPS1(+++), IL1RAPL2 (+ - very weak)
bF159E15	PLP (very weak), NADE (very weak)
bF106P8	NOX1 (++), SRPUL (+), XK-L (+++)
bF253J14	SRPX (+++), SYTL5 (+++)
bF281H15	SRPUL (+++), SYTL4 (+++)
bF13K23	SYTL4 (+), KIAA0316-L (+++)
bF34O3	SYTL4 (+), KIAA0316-L (+++)

Table 6-6 Non-paralogous genes for which their respective probes detected clones in common. The gene names are followed by an indication of the strength of the signal seen on the autoradiograph: +++ strong; ++ medium; + weak.



Ctg2 of sminmac, Clones 3 of 3, Markers 0 of 0, Sequenced 0,

Figure 6-9 The left section shows an FPC representation of contig 2. The right section shows fingerprint bands generated from the 3 clones within the contig.

On the basis of the combined hybridisation and fingerprinting results, BAC clones were selected for FISH experiments and sequencing. In each case, the clone from the contig with strongest signal seen after the most stringent wash condition still giving a signal was chosen, in addition to clones believed to contain multiple genes.

Contig	Clones	Positive Gene STS
1	bF78P20	XK+++
	bF135B17	XK+++
	bF231M3	XK+++
	bF255P10a/b	XK+++
	bF255O10a/b	XK+++
	bF123F16	XK+++
2	bF20I20	RAB9A+++
	bF65C12	RAB9A+++
	bF144L7	RAB9A+++
3	bF264I23	
	bF281H15	SRPUL+++ / SYTL4+++ / SRPX +
4	bF284I24	POLA+++
	bF124A24	POLA+++
5	bF89O16	RAB9B++
	bF244J18	RAB9B++ / RAB9A+++
	bF265N1a/b	RAB9A+++ / RAB9B++
	bF259B14a	
6	bF34O3b	KIAA0316-L+++ / SYTL4+
	bF13K23	KIAA0316-L+++ / SYTL4+
7	bF143A9a	
	bF218J23a	
	bF282H15a	
	bF104F10a	
	bF134H1a	
	bF34O3a	SYTL4+ / KIAA0316-L+++
	bF126H10a	
	bF93H4a	TM4SF6++
	bF158I6a/b	GLRA2+
	bF107F9a	cU46H11.CX.1+
8	bF68P17	GLRA2+++
	bF36H3a/b	GLRA2+++
11	bF157M9a/b	RAI2+++
	bF113I16	RAI2+++
12	bF243F20a/b	SRPX+++
	bF134H1b	
14	bF159K2	
	bF164C3b	
	bF159E15a/b	PLP+ (very weak)

Table 6-7 *Sminthopsis macroura Hin*d III/*Sau* 3A fingerprinting results. For clarity, this table presents only selected contigs formed that contained clones that were found to be positive after the most stringent wash in the reduced stringency hybridisations described earlier. The contig numbers allocated and the clones that the contigs were formed from are listed. The suffix "a" or "b" after a clone name denotes instances where a clone was fingerprinted twice, and is used to discriminate between the two fingerprints generated. Adjacent to the clone names are the names of genes for which the probe used in reduced stringency hybridisation experiments detected that clone. The gene names are followed by an indication of the strength of the signal seen on the autoradiograph: +++ strong; ++ medium; + weak.

6.4 Genomic localisation of the *Sminthopsis macroura* orthologues by FISH

One possibility for the generation of Xp/Xq paralogy is that the regions represent a recent intra-chromosomal duplication within the eutherian lineage; the other possibility is that it represents an older duplication, and hence the Xp paralogues would be autosomal in marsupials.

A FISH approach was undertaken to localise BACs isolated in the previous section within the *Sminthopsis macroura* genome. The hypothesis was that those clones containing orthologues of human genes located on Xp would have an autosomal location in *Sminthopsis macroura*, and those containing orthologues of human genes located on Xq would be located on the X chromosome in *Sminthopsis macroura*. This approach would also demonstrate whether the clones containing orthologues of human genes located on Xp localised to the same autosome, or if they were divided between different autosomes.

If located on the same autosome, it would provide support for the hypothesis that the region corresponding to the portion of human Xp from MID1 (Tel) to TM4SF2 (Cen) was translocated to an ancestral X chromosome as one block in a single event during the time between the divergence of metatherian mammals and eutherian mammals (~130 Mya) and the radiation of eutherian mammals (~90 Mya). Genes from the intervening section between the two Xp paralogy blocks were also chosen, to assess whether these were part of a single duplication event. If co-localised with the Xp paralogues, this would also further support the orthology of these loci.

The localisation of the marsupial orthologues of the human Xp/Xq paralogue pairs would also provide further information regarding the timing of the segmental duplication event leading to creation of the human Xp/Xq paralogues. If both Xp and Xq representative genes were found within the marsupial, it would support the hypothesis that the duplication occurred prior to separation of the therian lineages.

The BACs selected for the FISH analysis and sequencing, the potential orthologues they contain, and their positions relative to the human X chromosome are given in Table 6-8 and shown in Figure 6-10.

Clone	Gene	Comment relating to clone choice
bF134C3	MID1/MID2	Both MID1 and MID2 probes detect clone equally well. Strong signal after 0.2x SSC wash.
bF232B10	KIAA0316	Strong signal after 0.2x SSC wash.
bF14N15	PRPS2/PRPS1	Both PRPS2 and PRPS1 probes detect clone equally well. Strong signal after 0.2x SSC wash.
bF48C16	PRPS2/PRPS1	Both PRPS2 and PRPS1 probes detect clone equally well. Strong signal after 0.2x SSC wash.
bF20I20	RAB9A/RAB9B	Strong signal after 0.2x SSC wash.
bF153M3	GPM6B	Strong signal after 0.2x SSC wash.
bF149E6	GLRA2	Strong signal after 0.2x SSC wash.
bF103A22	GRPR	Medium signal after 0.2x SSC wash.
bF185E13	RAI2	Strong signal after 0.2x SSC wash.
bF211D13	SAT	Strong signal after 0.2x SSC wash.
bF284I24	POLA	Strong signal after 0.2x SSC wash.
bF272K20	IL1RAPL1	Strong signal after 0.2x SSC wash.
bF125G2	DMD	Strong signal after 0.2x SSC wash.
bF231M3	XK	Strong signal after 0.2x SSC wash.
bF242G1	CYBB	Strong signal after 0.2x SSC wash.
bF253J14	SYTL5 and SRPX	Detected by probes from two genes closely linked in human. Strong signal after 0.2x SSC wash.
bF99F22	TM4SF2	Strong signal after 0.2x SSC wash.
bF93H4	TM4SF6	Weak signal after 0.2x SSC. The only clone detected at this stringency.
bF281H15	SRPUL and SYTL4	Detected by probes from two genes closely linked in human. Strong signal after 0.2x SSC wash.
bF106P8	NOX1, XK-L and SRPUL	Detected by probes from three genes closely linked in human. Strong signal after 0.2x SSC wash for XK-L probe, weak for NOX1 and only weakly after a 1x SSC wash for SRPUL.
bF28C20	DRP2	Strong signal after 0.2x SSC wash.
bF168K3	BTK	Strong signal after 0.2x SSC wash.
bF21K1	dJ545K15.1	Medium signal after 0.2x SSC wash.
bF283J5	NXF2	Weak signal after 0.5x SSC wash.
bF159E15	PLP	Very weak signal after 1x SSC wash.
bF89O16	RAB9A/RAB9B	Medium signal after 0.2x SSC wash. Fingerprint data suggest different locus to that for bF20I20.
bF6N3	cU46H11.CX.1	Strong signal after 0.2x SSC wash.
bF13K23	KIAA0316-L	Strong signal after 0.2x SSC wash.

Table 6-8 Table listing *Sminthopsis macroura* BAC clones chosen for FISH analysis and sequencing. The clone selected and the hybridising gene probe are shown. Clone names in bold represent clones selected for whole-insert genomic sequencing. Clones are listed by genes contained within them and the order of location of these orthologues on the human X chromosome, Xpter (top) to Xqter (bottom). Comments relating to choice of the clone thought most likely to represent the *Sminthopsis macroura* orthologue are noted.

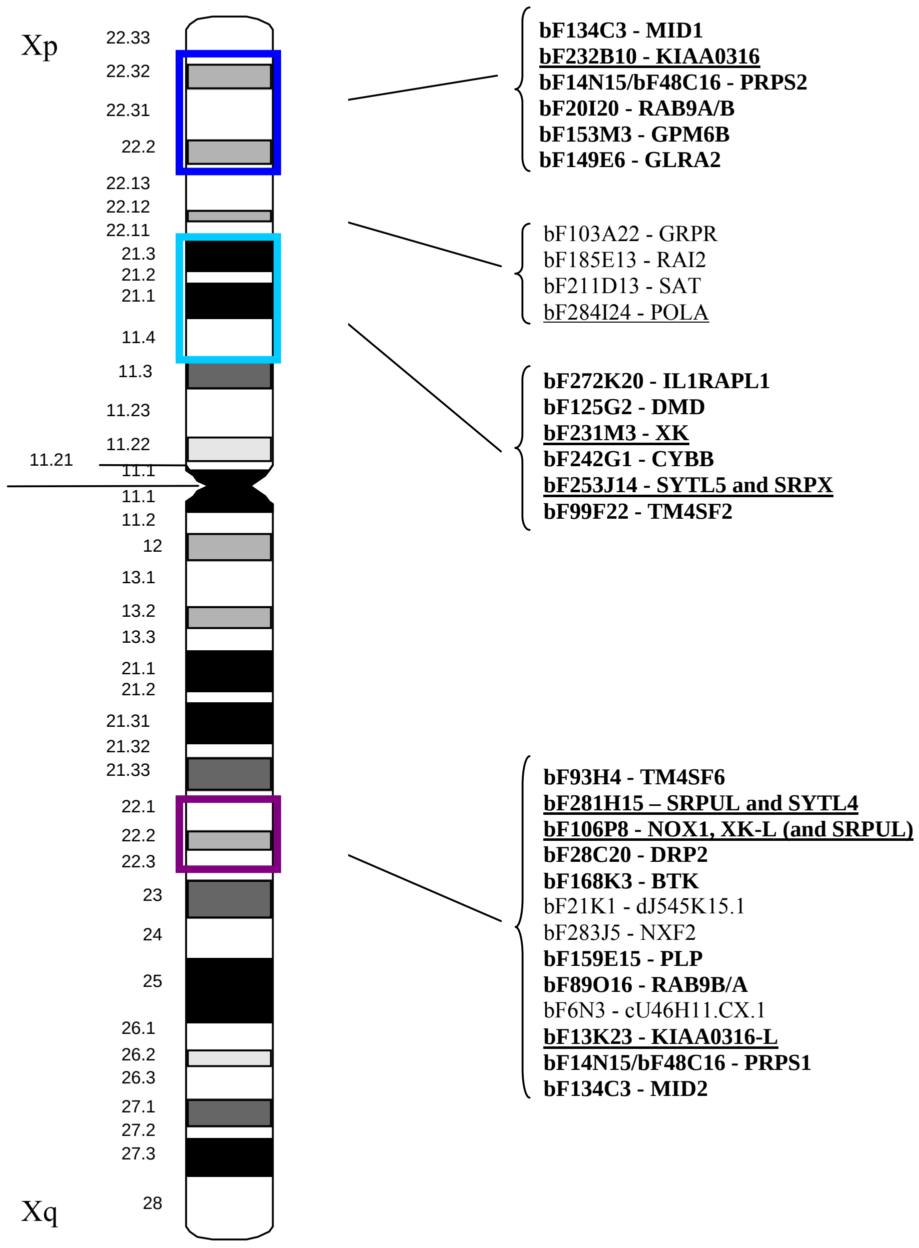


Figure 6-10 Diagram illustrating genes for which *S. macroura* positive BACs were selected for FISH analysis and sequencing. Positions of the human genes relative to the human X chromosome are illustrated, together with their selected BACs. The genes are listed in order from Xpter-Xqter. The main blocks of Xp/Xq paralogy are denoted by the blue, turquoise and purple boxes on the chromosome ideogram. Xp/Xq paralogue gene names are shown in bold. Clones being sequenced are underlined.

For FISH analysis, *Sminthopsis macroura* metaphase chromosome preparations were obtained as a kind gift from Dr. Willem Rens (Cambridge Resource Centre for Comparative Genomics, Centre for Veterinary Science, University of Cambridge). The chromosome preparations were made from a male *Sminthopsis macroura* cell line, whose karyotype has undergone rearrangement and aneuploidy. The chromosome changes have been characterised by chromosome painting using flow-sorted chromosomes from a related marsupial, *Sminthopsis crassicaudata* (Dr. Willem Rens, personal communication). This information was utilised in interpretation of the *Sminthopsis macroura* FISH results, and is illustrated in a DAPI-stained karyogram shown in Figure 6-11. From this information, re-arrangements were not detected that involved the X chromosome, hence localisation of a BAC to either an autosome or the X chromosome should be straightforward and valid.

Initial experiments established that hybridisation of BAC clones to the metaphase chromosome preparations without the use of sheared genomic DNA to suppress repeats gave the best signal-to-background ratio, and these conditions were then employed for all subsequent FISH experiments (data not shown).

BAC clones were initially hybridised to metaphase chromosome spreads in pairs, each clone labelled using a different fluorophore, or singly. This set of experiments aimed to determine whether a BAC localised to an autosome or the X chromosome in the *Sminthopsis macroura* genome.

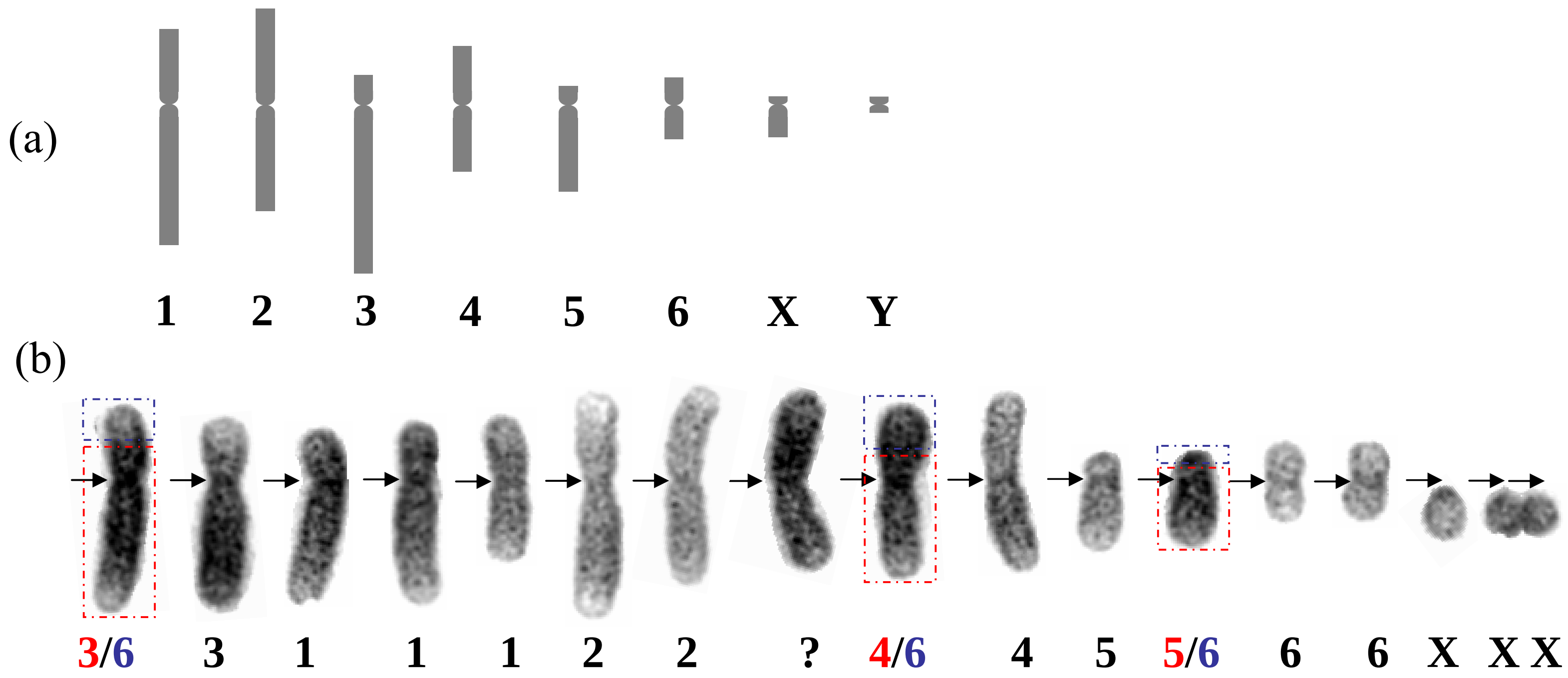


Figure 6-11 Karyogram showing (a) *Sminthopsis macroura* normal karyotype ideogram (from (De Leo *et al.*, 1999)), (b) Representative DAPI-stained chromosomes from metaphase chromosome preparations from a male *Sminthopsis macroura* cell line (2n=18) used for FISH analyses, obtained as a kind gift from Dr. Willem Rens (University of Cambridge). It includes interpretations of chromosome assignment, using information from cross-species chromosome painting using paints derived from flow-sorted chromosomes of a related marsupial, *Sminthopsis crassicaudata* (performed by Dr. Willem Rens, personal communication). Black arrows denote centromere position. Numbers beneath chromosomes denote the allocated chromosome number, however these are only guides and are often ambiguous, due to poor morphology of marsupial metaphase chromosomes. Coloured dashed boxes correspond to coloured chromosome numbers beneath, to illustrate rearrangements. The Y chromosome appears only as a dot. Deviation from the ancestral Sminthopsis macroura 2n=14 karyotype is explained by re-arrangements and aneuploidy occurring during the cultivation of the cell-line.

These experiments succeeded in localising BAC clones to the *Sminthopsis macroura* X chromosome or autosomes, and results are shown in Figures 6-12 to 6-16, and Table 6-9. Thirteen BACs representing fourteen Xp genes, ten BACs representing thirteen Xq genes and five BACs whose orthologue could not be distinguished at present were hybridised and localised. Thirteen of the Xp gene BACs localised to autosomes, eleven of which appeared to localise to chromosome 3 or 1. Five of the Xq gene BACs localised to autosomes (not chromosome 3 or 1) and one, (DRP2) co-localised with its' Xp paralogue. As the probes designed to DMD and DRP2 were located in different regions of the genes that would explain why the probes failed to detect clones in common. Four of the Xq gene BACs localised to the X chromosome.

Of the five BACs whose orthologue could not be distinguished, bF20I20 localised to the X chromosome indicating it contained the orthologue of RAB9B; bF134C3 localised to chromosome 3 or 1, indicating it contained the orthologue of MID1; bF89O16 localised to an autosome that did not appear to be chromosome 3 or 1; and clones bF14N15 and bF48C16 co-localised to chromosome 3 or 1, suggesting they both contain the orthologue of PRPS2.

The localisation information obtained increases confidence that certain BAC clones selected contain true *Sminthopsis macroura* orthologues of the human genes. However in some cases, the localisation information suggests that either a minor rearrangement has occurred, or that the BAC clone does not contain the true orthologue. From the present data, it cannot be ascertained which of these statements is correct. For DRP2 and DMD, both BACs co-localised. The localisation to chromosome 3 or 1 suggests that both of the BACs contain DMD, and that the DRP2 probe cross-hybridised.

Of the Xq22 genes, 6 were localised to autosomes that did not seem to be chromosome 3 or 1. Of these, NXF2 has an autosomal paralogue in human (NXF1 on chromosome 11) and thus the BAC could represent an NXF1 locus instead of NXF2. The BAC could also be a false positive, as it was only weakly positive after the 0.5x SSC wash. Similarly the BAC for PLP was only weakly positive after the 1x SSC wash, and is likely a false positive, as is the BAC for TM4SF6.

The BACs for dJ545K15.1, RAB9B/A and cU46H11.CX.1 hybridised more strongly. For RAB9B/A, as there are many Rab family members, it is most likely the BAC represents a different paralogue. For dJ545K15.1 and cU46H11.CX.1, as these are involved in the Xq22 paralogy described in Chapter 5, further work could be performed using other genes from the region to determine if they confirm these results.

The BACs for TM4SF2 and GLRA2 hybridised strongly, but localised to autosomes other than 3 or 1. Further work would be required to determine whether these represent additional paralogues or the true orthologues.

In general, more of the Xp genes localised as expected. This is partly accounted for by the less convincing hybridisation results seen for some of the Xq22 genes, and cross-hybridisation for DRP2 (and possibly for RAB9B/A). For the remaining two genes, additional experiments could be performed to determine the localisations of the other genes involved in the extensive Xq22 paralogy (Chapter 5) and help assess the likelihood of these being true autosomal orthologues or different paralogues.

These data support the hypothesis that the duplication event leading to generation of the human Xp/Xq paralogues was a relatively ancient segmental duplication, occurring before the divergence of metatherian mammals and eutherian mammals (~130 Mya) as all four of the Xp non-paralogous genes appeared to localise to the same autosome as Xp paralogues. This argues against the duplication occurring as an intra-chromosomal event within the eutherian mammal lineage.

Clone	**Gene**	**Human chromosomal location**	***Sminthopsis macroura* chromosomal location**
bF134C3	**MID1/MID2**	Xp22.2 - p22.3/ Xq22	3 or 1
bF232B10	**KIAA0316**	Xp22.2 - p22.3	3 or 1
bF14N15	**PRPS2/PRPS1**	Xp22.2 - p22.3	3 or 1
bF48C16	**PRPS2/PRPS1**	Xp22.2 - p22.3	3 or 1
bF20I20	**RAB9B**	Xp22.2 - p22.3	X
bF153M3	**GPM6B**	Xp22.2 - p22.3	3 or 1
bF149E6	**GLRA2**	Xp22.2 - p22.3	autosome
bF103A22	GRPR	Xp22.1	3 or 1
bF185E13	RAI2	Xp22.1	3 or 1
bF211D13	SAT	Xp22.1	3 or 1
bF284I24	POLA	Xp22.1	3 or 1
bF272K20	**IL1RAPL1**	Xp11.3 - p21.3	3 or 1
bF125G2	**DMD**	Xp11.3 - p21.3	3 or 1
bF231M3	**XK**	Xp11.3 - p21.3	3 or 1
bF242G1	**CYBB**	Xp11.3 - p21.3	3 or 1
bF253J14	**SYTL5 and SRPX**	Xp11.3 - p21.3	3 or 1
bF99F22	**TM4SF2**	Xp11.3 - p21.3	autosome
bF93H4	**TM4SF6**	Xq22 - q23	autosome
bF281H15	**SRPUL and SYTL4**	Xq22 - q23	X
bF106P8	**NOX1, XK-L and SRPUL**	Xq22 - q23	X
bF28C20	**DRP2**	Xq22 - q23	3 or 1
bF168K3	**BTK**	Xq22 - q23	X
bF21K1	dJ545K15.1	Xq22 - q23	autosome
bF283J5	NXF2	Xq22 - q23	autosome
bF159E15	**PLP**	Xq22 - q23	autosome
bF89O16	**RAB9A/RAB9B**	Xq22 - q23/ Xp22.2	autosome
bF6N3	cU46H11.CX.1	Xq22 - q23	autosome
bF13K23	**KIAA0316-L**	Xq22 - q23	X

Table 6-9 Localisation data for FISH of *Sminthopsis macroura* BAC clones against spreads of metaphase chromosomes. The table lists the BAC clone used for FISH, the gene it contains, the chromosomal location of the human gene, and the *Sminthopsis macroura* chromosomal assignment from FISH. In cases where the autosome did not appear to be chromosome 3 or 1, it was simply termed "autosome". Bold gene names denote human Xp/Xq paralogues. Table borders are coloured as in Figure 6-10.

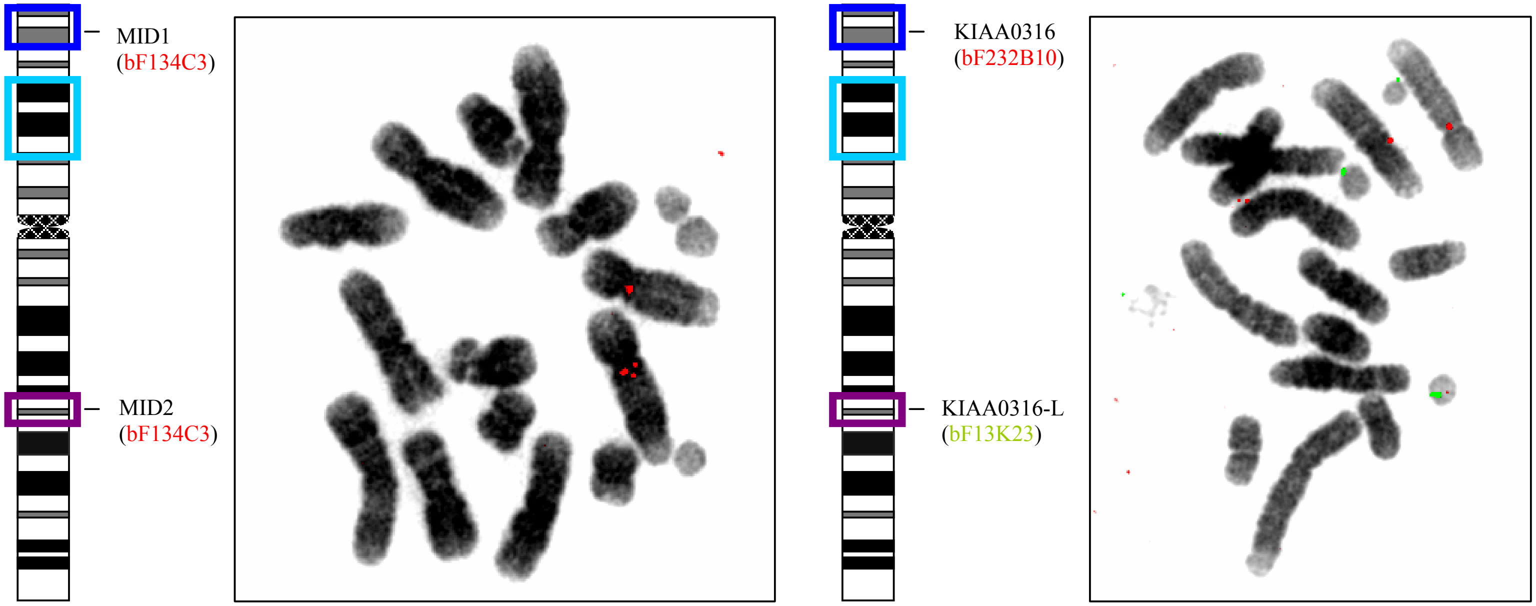


Figure 6-12 FISH of *Sminthopsis macroura* BAC clones against spreads of metaphase chromosomes. The human gene and the hybridisation-positive *Sminthopsis macroura* BAC clone used for FISH are shown against an ideogram of the human X chromosome to illustrate positioning. The colour of the BAC clone name reflects the label colour for that clone seen in the image. To the right of the ideogram is a representative FISH image. At least 10 metaphase images were studied for each FISH experiment

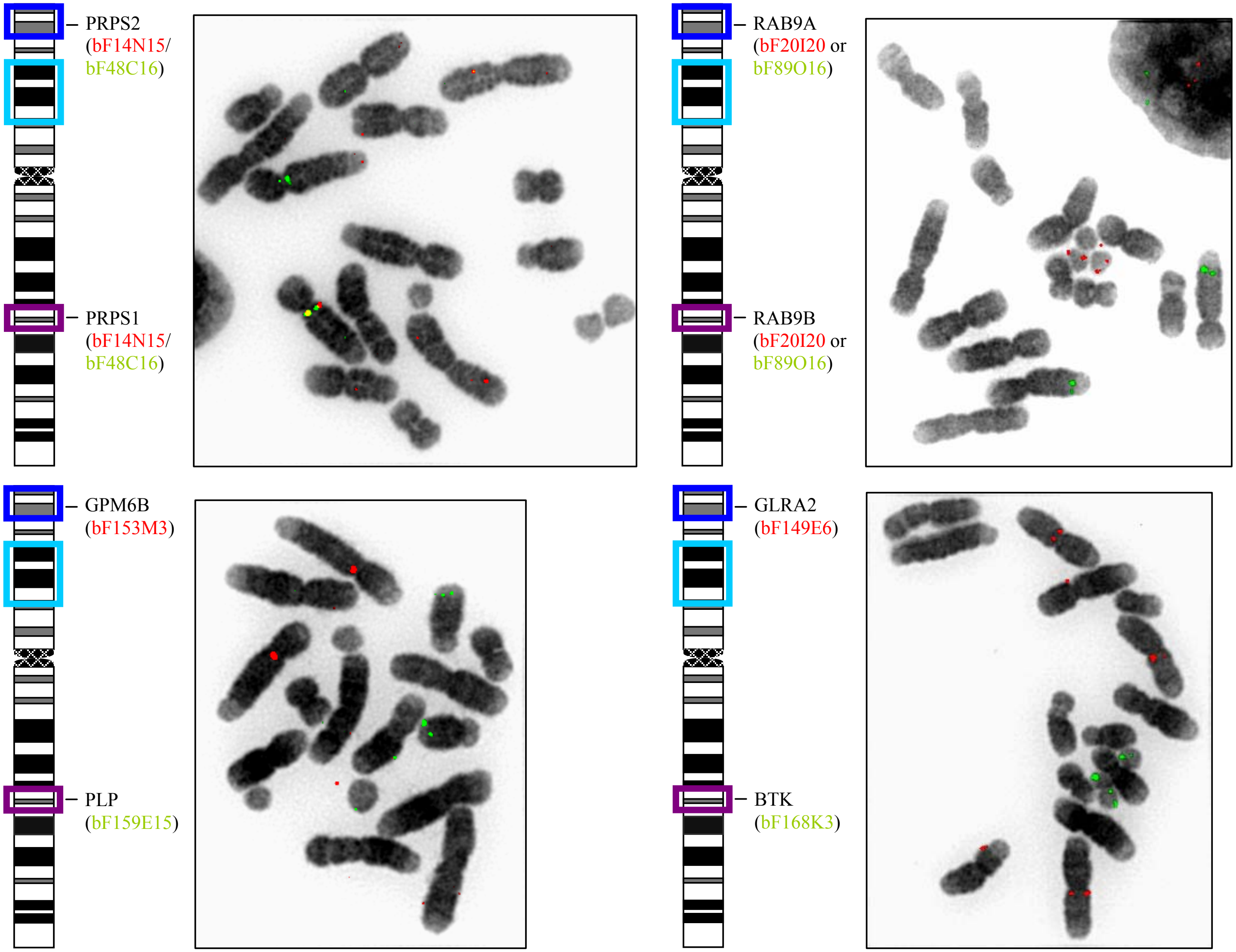


Figure 6-13 Legend as for Figure 6-12.

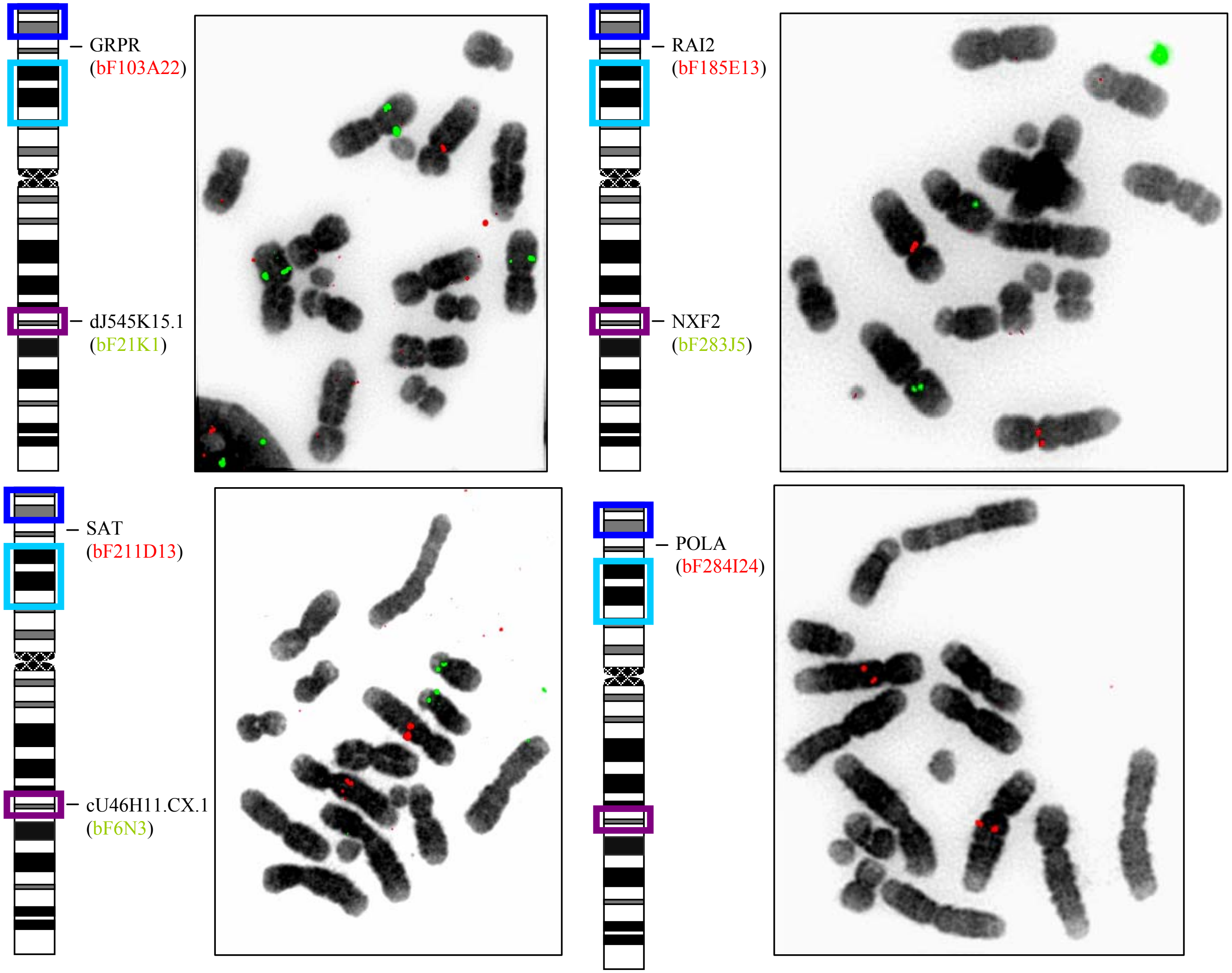


Figure 6-14 Legend as for Figure 6-12.

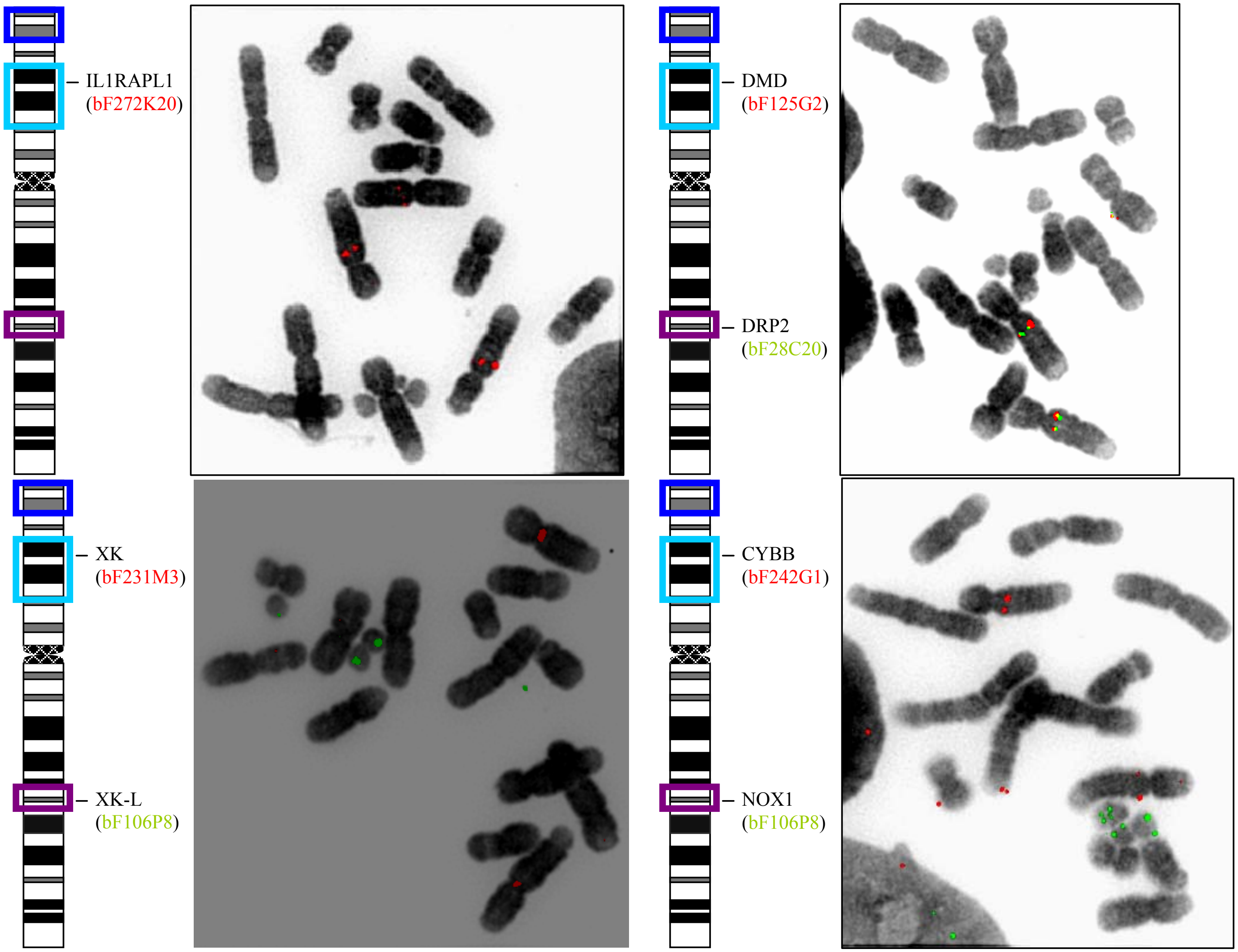


Figure 6-15 Legend as for Figure 6-12.

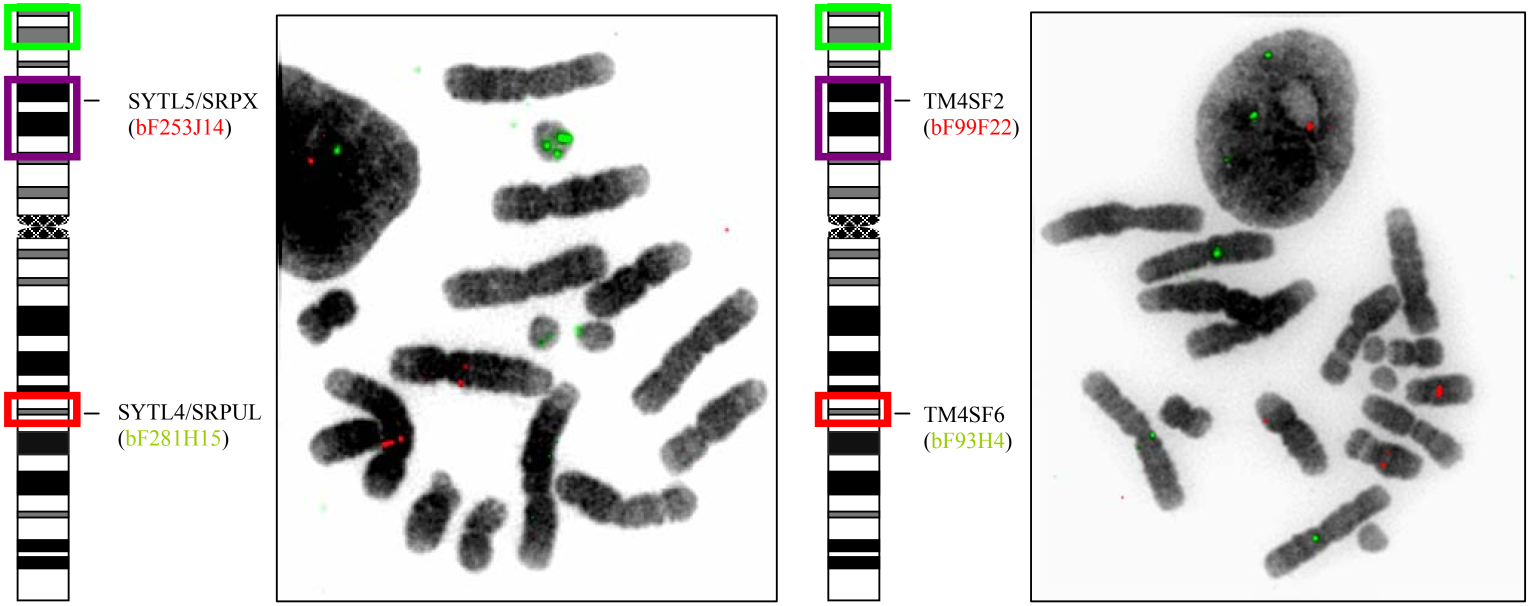


Figure 6-16 Legend as for Figure 6-12.

As noted above, it was observed that the majority of the BAC clones predicted to contain orthologues of the human Xp genes appeared to be localising to the same autosome, potentially chromosome 3 or chromosome 1, in the same region of the long-arm close to the centromere. As seen in Figure 6-11, assigning autosomes was difficult due to poor chromosome morphology, but acrocentric and metacentric chromosomes could be discerned, thus reducing the possibilities. The prediction would be that this is actually chromosome 3. This is based on previous studies showing that *Sminthopsis crassicaudata* chromosome 3 corresponds to *Macropus Eugenii* (Tammar Wallaby) chromosome 5 (Rens *et al.*, 2001), to which several genes orthologous to human Xp genes have been mapped (Spencer *et al.*, 1991).

Experiments were performed using selected pairs of BAC clones which had been localised to an autosome to confirm or refute co-localisations. The results are shown in Table 6-10 and Figure 6-17 (some of these experiments were performed by Deborah Burford, Molecular Cytogenetics Group, Wellcome Trust Sanger Institute – these experiments are indicated in the table and figures showing the results).

Clone pair	Genes	Same autosome?
bF232B10 and bF211D13	KIAA0316 and SAT	yes
bF125G2 and bF211D13	DMD and SAT	yes
bF231M3 and bF211D13	XK and SAT	no
bF283J5 and bF21K1	NXF2 and dJ545K15.1	no
bF6N3 and bF283J5	cU46H11.CX.1 and NXF2	no
bF211D13 and bF253J14 *	SAT and SYTL5/SRPX	yes
bF242G1 and bF211D13 *	CYBB and SAT	yes
bF284I24 and bF211D13 *	POLA and SAT	yes
bF103A22 and bF211D13 *	GRPR and SAT	yes
bF272K20 and bF125G2 *	IL1RAPL1 and DMD	yes

Table 6-10 Results from co-localisation experiments by FISH of *Sminthopsis macroura* BAC clones against spreads of metaphase chromosomes. Experiments performed by Deborah Burford are denoted with an asterisk.

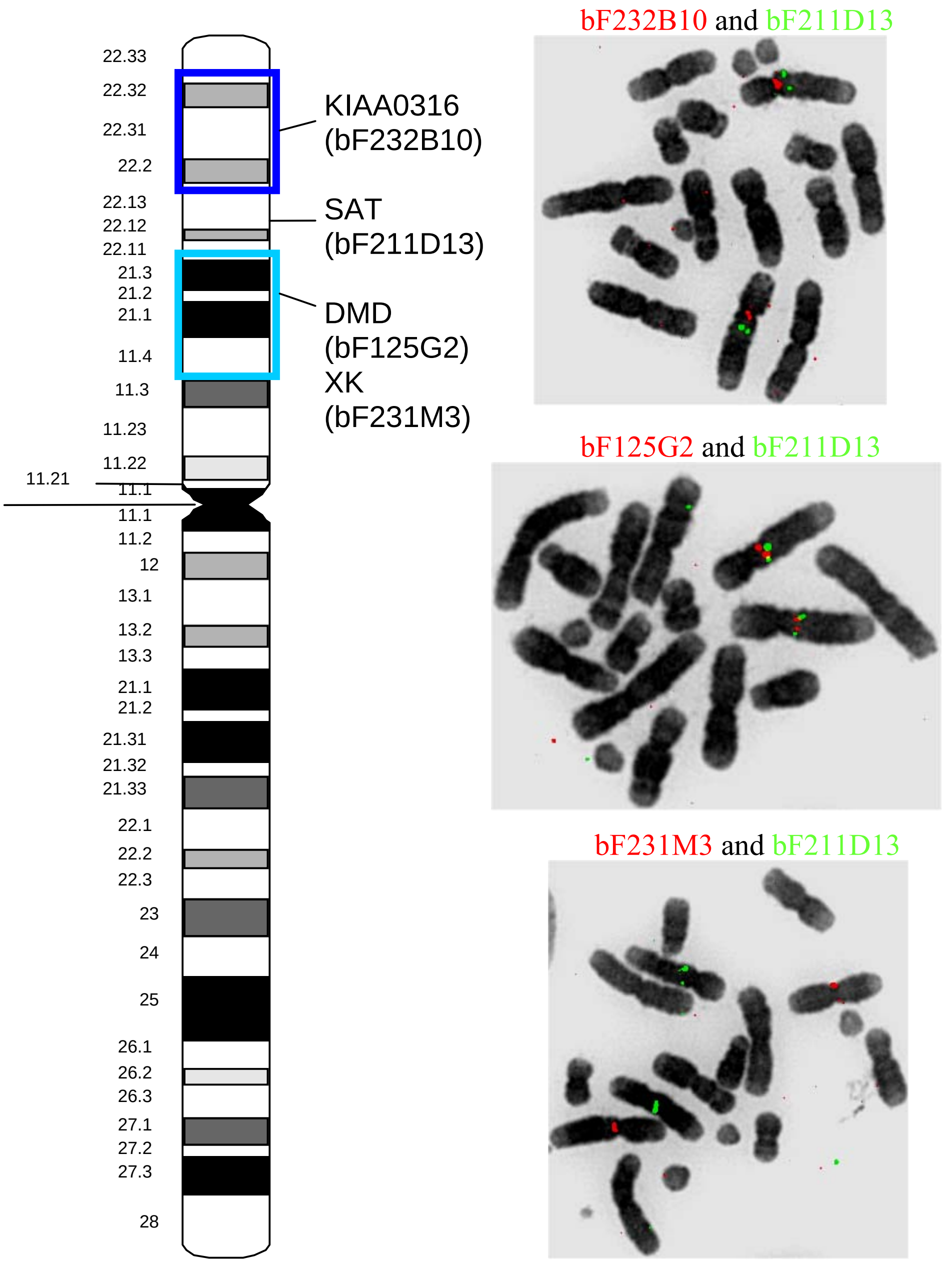


Figure 6-17 Figure showing example of results from co-localisation experiments using FISH of *Sminthopsis macroura* BAC clones against spreads of metaphase chromosomes.

These results confirmed observations that some of the clones were mapping to autosomes that appeared to be the same as one another. Of the nine clones tested (10 Xp genes), only bF231M3 (thought to contain the XK orthologue) failed to co-localise. This confirms that the orthologues of KIAA0316, SAT, DMD, SYTL5, SRPX, CYBB, POLA, GRPR and IL1RAPL1 localise to the same autosome.

Of the Xq22 orthologues tested, NXF2 failed to co-localise with dJ545K15.1 or cU46H11.CX.1. As mentioned earlier, the NXF2 BAC was relatively weakly hybridising and may represent a false positive or another paralogue. Further work would be needed to explore the Xq22 gene relationships using additional clones.

In summary, seven orthologues of Xq22 genes were localised to the marsupial X as expected. These data also confirmed co-localisation of many of the orthologues of human Xp genes, including those without paralogues on Xq22, to the same autosome in the *Sminthopsis macroura* genome. The results provide evidence that supports the hypothesis that the duplication leading to Xp/Xq paralogy did not occur as an intra-chromosomal event within the eutherian mammal lineage, and, that the region corresponding to the portion of human Xp with MID1 (Tel) to SRPX (Cen) marking the minimal boundaries was translocated to an ancestral X chromosome as one block in a single event during the time between the divergence of metatherian mammals and eutherian mammals (~130Mya) and the radiation of eutherian mammals (~90Mya). The alternative explanation, that the block was acquired by an autosome from the X is less likely, given reports from the literature.

The data also suggest that the Xp paralogues and possibly the intervening region separating the two blocks of paralogues (containing POLA) were duplicated in a single event. If so, the genes from the intervening region must have been lost from the ancestral X. The alternative is that the region including POLA was inserted into the autosomal paralogous region subsequent to the duplication. Further studies in more evolutionary distant organisms may shed light on these alternate hypotheses.

6.5 Dating the Xq22-q23/Xp regional duplication

The completion of the draft human genome sequence has enabled studies of gene duplication events to be studied on an unprecedented scale. Whilst the theory of whole genome duplications remains an area of active debate, recent studies utilising whole-genome approaches suggest a combination of segmental duplications and smaller tandem duplications leading to paralogous regions. Utilising molecular clock methodology, these studies were also able to provide data on the temporal sequence of events. Although these methods are subject to large errors, these studies suggest that there was a wave of segmental duplications ~550 Mya (Gu *et al.*, 2002), (McLysaght *et al.*, 2002), with a wide distribution of tandem duplications throughout evolution. In light of these studies, attempts were made to date the Xp/Xq segmental duplication to put it in context with these studies.

6.5.1 Gene-based evidence from the scientific literature

Several of the genes involved in the Xp/Xq segmental duplication have been the focus of intensive study, due to their involvement in human disease. In some cases, review of the literature revealed information on evolutionary studies of protein families to which these genes belong. These genes include the lipophilin family (GPM6B/PLP) and the dystrophins (DMD/PLP). For each of these families, the literature was reviewed and information regarding the evolution of the families is given below.

6.5.1.1 Lipophilins

The lipophilin family of proteins have been the subject of intensive study, particularly motivated by the fact that defects of one of the members, PLP (Proteolipid Protein) are involved in Pelizaeus-Merzbacher disease. Kitagawa *et. al.* reported cloning of homologues of three lipophilin members DMα, DMβ and DMγ from two elasmobranches, *Squalus acanthias* and *Torpedo marmorata* (Kitagawa *et al.*, 1993). Subsequent studies have referred to these as representing homologues of PLP/DM20 (DMα), GPM6A (DMβ) and GPM6B (DMγ) (Gow 1997). If these genes do in fact represent orthologues of the human genes, it would imply that any duplication event generating PLP and GPM6B would have had to have occurred before the cartilaginous/bony fish divergence approximately 528 Mya. In addition, Yoshida *et. al.* (Yoshida *et al.*, 1999) cloned representatives of these genes from an amphibian, *Xenopus laevis*, which would again imply a duplication event before the amphibians diverged from the lineage leading to mammals. An alternative explanation is that the gene duplications occurred independently in the separate lineages. Whilst certainly a possibility, it seems a more complex explanation of the data and so a less attractive hypothesis.

6.5.1.2 Dystrophins

The dystrophins have also been the subject of intensive study, again largely motivated because defects in the dystrophin gene can cause a range of abnormalities. The evolutionary origins of the dystrophins have been extensively studied and reviewed (Roberts 2001). These studies indicate that an ancestral dystrophin-like gene was present before invertebrates and vertebrates diverged (from identification of a gene similar to the dystrophin gene in *Caenorhabditis elegans* (Segalat 2002), *Drosophila melanogaster* and a sea urchin (Neuman *et al.*, 2001), and that subsequently the ancestral dystrophin gene was partially duplicated to generate DRP2. Subsequently the ancestral dystrophin gene underwent a further complete duplication to generate Utrophin and Dystrophin.

As with the lipophilins (see above), homologues of dystrophin and DRP2 have been found in dogfish and a ray (Roberts *et al.*, 1996), indicating that the duplication event generating dystrophin and DRP2 occurred prior to the divergence of cartilaginous and bony fish.

The dystrophin duplications are particularly intriguing, as authors have speculated that DRP2 was generated by a partial duplication of the ancestral gene, as is consistent with the presence of a larger dystrophin-like gene structure in invertebrates. However, if the DRP2 and dystrophin/utrophin precursor genes were generated as part of a larger segmental duplication as presented in this Chapter, it is perhaps more likely that the truncated gene structure of DRP2 is the result of a subsequent deletion/rearrangement. For DRP2 to be found widely amongst other vertebrates, such a truncation may have occurred relatively soon after the segmental duplication occurred. This explanation would predict that there may be evolutionary distant vertebrate lineages that preserve a larger DRP2 gene structure.

Together, studies of the dystrophins and lipophilins suggest that duplications generating PLP/GPM6B and DMD/DRP2 occurred before the divergence of cartilaginous and bony fish approximately 528 Mya. If we accept the hypothesis that has been argued in this Chapter, that PLP/GPM6B and DMD/DRP2 were generated as part of a segmental duplication, these observations suggest that the duplication occurred at least 528 Mya, but most likely after the divergence of protochordates and chordates. These data must be viewed with caution, as duplications within different lineages can confound predictions of orthology, and such duplications are known to have occurred. They do however provide a working hypothesis to investigate using sequence data from other organisms and phylogenetic analysis, as presented in the next section.

6.5.2 Comparative analysis of the Fugu rubripes genome

As work for this Chapter was in progress, completion of a draft whole-genome shotgun assembly of the *Fugu rubripes* genome was announced (Aparicio et al., 2002). This provided an opportunity to search the *Fugu* genome for orthologues of the Xp/Xq paralogues. If the segmental duplication occurred at least 528 Mya as suggested by the literature reviewed above, orthologues for each of the Xp/Xq paralogues should be present in *Fugu*, which diverged from the lineage giving rise to tetrapods some 450 Mya.

Initial work employed TBLASTN analysis of the *Fugu* genome, using human Xp/Xq paralogue protein sequences as queries via the Ensembl web server. This approach was designed to provide sensitivity given the long evolutionary period separating *Homo sapiens* and *Fugu rubripes*. Subsequently, further releases of the *Fugu rubripes* draft assembly via Ensembl provided data on *Homo sapiens*-*Fugu rubripes* orthology from reciprocal BLAST analyses. At this point, the approach switched to collating the orthology data for each of the Human Xp/Xq paralogues via Ensembl. The collated data are presented in Table 6-11. From Table 6-11, some of the Xp/Xq paralogues are also duplicated in *Fugu*, and some of these genes co-localise on the same genome scaffolds. The property of shared synteny is an indicator of orthology. If the orientations of *Fugu* genes and proximities to non-paralogous genes were conserved with respect to their human counterparts, this would provide strong support for the *Fugu* genes being true orthologues of human Xp/Xq paralogues. In addition, conservation of exon size would provide further evidence that the genes shared a common ancestor and are not similar via convergent evolution. To ascertain this information, the *Fugu* scaffolds and the transcript exon details were examined via the Ensembl (*Fugu*) web server for selected genes with shared synteny. Gene order and transcription direction are presented schematically in Figure 6-18, and transcript exon sizes are provided in Table 6-12 and Table 6-13 in comparison to human Xp/Xq paralogues.

The gene structure information shows good agreement in many cases between the human Xp/Xq genes and their potential *Fugu* orthologues, providing supporting evidence that they arose from a shared ancestral gene. From Figure 6-18, we see that for the strongest indication of true orthology for Xp/q paralogue pairs is provided for XK/XK-L, SYTL5/SYTL4 and SRPX/SRPUL. For each member of these pairs, a *Fugu* gene is noted with a similar transcriptional direction with respect to its neighbours (allowing for a small inversion in the case of SRPUL and SYTL4), and positioning reflecting that of its human orthologue.

Whilst limited, the genomic data from *Fugu* appear to demonstrate strong evidence of orthology for some of the Xp/q paralogues. The presence of each member of an Xp/q paralogue pair in the *Fugu* genome would indicate that each member of the pair was generated in a duplication occurring before the divergence of *Fugu rubripes* and *Homo sapiens*, approximately 450 Mya.

As it has been demonstrated earlier in this chapter that the Xp/q paralogues appear to have been generated at the same time as part of a segmental duplication, the indication of orthology in *Fugu* for a limited number of Xp/q paralogues may be extrapolated to indicate that the age of the complete segmental duplication occurred ~450 Mya.

Gene Name	Human Ensembl gene identifier	*Fugu* Ensembl Gene identifier	*Fugu* scaffold sequence
TM4SF2	ENSG00000156298	SINFRUG00000126322	Chr_scaffold_368
		SINFRUG00000139047	
SRPX	ENSG00000101955	SINFRUG00000147882	Chr_scaffold_1498
SYTL5	ENSG00000147041	SINFRUG00000147873	Chr_scaffold_1498
CYBB	ENSG00000165168	SINFRUG00000153805	Chr_scaffold_69
XK	ENSG00000047597	SINFRUG00000147861	Chr_scaffold_1498
DMD	ENSG00000132438	SINFRUG00000144800	Chr_scaffold_35
		SINFRUG00000144805	
IL1RAPL1	ENSG00000169306	SINFRUG00000138032	Chr_scaffold_1433
BMX	ENSG00000102010	None noted	
GLRA2	ENSG00000101958	SINFRUG00000136562	Chr_scaffold_811
		SINFRUG00000147089	
		SINFRUG00000147091	
GPM6B	ENSG00000046653	SINFRUG00000127596	Chr_scaffold_1534
RAB9A	ENSG00000123595	SINFRUG00000127608	Chr_scaffold_1534
TMSB4X	Not located		
PRPS2	ENSG00000101911	None noted	
KIAA0316	ENSG00000169933	SINFRUG00000153014	Chr_scaffold_280
MID1	ENSG00000101871	SINFRUG00000137619	Chr_scaffold_642
TM4SF6	ENSG00000000003	SINFRUG00000125878	Chr_scaffold_347
SRPUL	ENSG00000102359	SINFRUG00000125883	Chr_scaffold_347
SYTL4	ENSG00000102362	SINFRUG00000125885	Chr_scaffold_347
NOX1	ENSG00000007952	SINFRUG00000125864	Chr_scaffold_347
XK-like	Not located	SINFRUG00000125861	Chr_scaffold_347
DRP2	ENSG00000102385	SINFRUG00000139028	Chr_scaffold_3836
IL1RAPL2	ENSG00000182513	None noted	
BTK	ENSG00000010671	SINFRUG00000147533	Chr_scaffold_191
GLRA4	Not located		
PLP	ENSG00000123560	SINFRUG00000130567	Chr_scaffold_594
RAB9B	ENSG00000123570	SINFRUG00000130565	Chr_scaffold_594
cV362H12.CX.1	Not located		
PRPS1	ENSG00000147224	SINFRUG00000122961	Chr_scaffold_432
KIAA0316-L	Not located		
MID2	ENSG00000080561	SINFRUG00000134118	Chr_scaffold_57

Table 6-11 *Fugu rubripes* orthologues (as determined by reciprocal BLAST analysis) collated from Ensembl (*Fugu*) release 15.2.1 and Ensembl (Human) release 15.33.1. The Ensembl gene identifiers are given for each species' orthologue, as well as the genome sequence scaffold that the *Fugu* gene maps to. Scaffolds common to different genes are denoted in the same coloured type. The human genes are listed in order from XpCen - XpTel, then XqCen - XqTel.

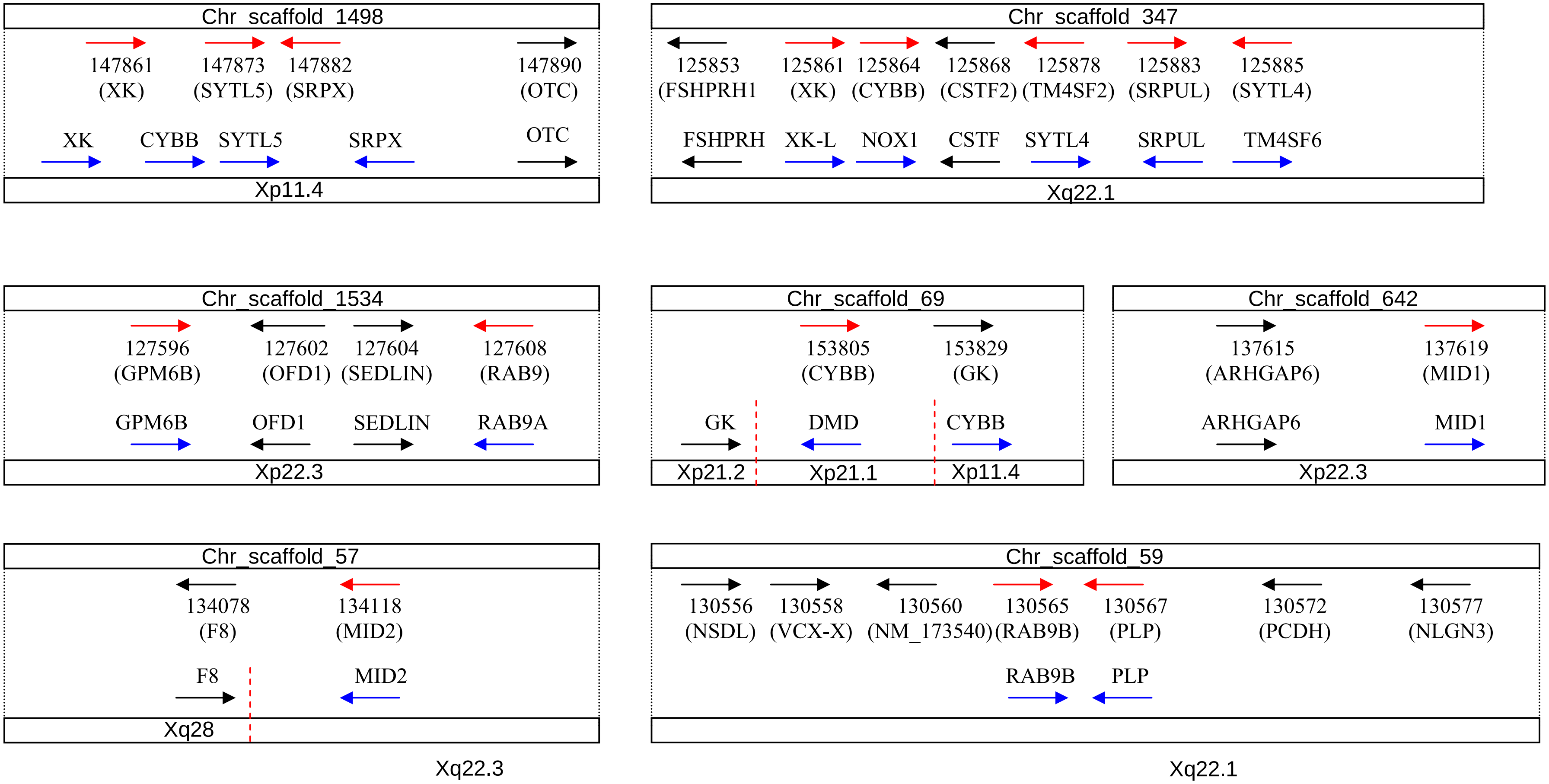


Figure 6-18 Figure showing a schematic representation of selected *Fugu rubripes* WGS sequence scaffolds with information regarding putative Fugu orthologue gene order, transcription direction and shared synteny with human Xp/Xq paralogue and non-Xp/Xq paralogue orthologues. Dotted lines join the Fugu scaffold representations to a representation of the putative orthologous human genomic region. Red arrows denote transcriptional direction of Fugu genes, blue arrows that of their potential human orthologue. Black arrows denote transcriptional direction and positioning of non-Xp/Xq paralogue genes and their potential Fugu orthologues.

Gene	No. exons	Exon sizes (bp)																				
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
MID1	10	130	**716**	96	108	149	128	144	162	208	**1609**											
MID2	10	201	**716**	96	108	149	128	240	162	208	**521**											
FrMID1	7				111	149	128	144	162	208	328											
FrMID2	9	353	307	96	202	159	144	240	▨	208	328											
KIAA0316	16	212	117	161	103	46	105	108	132	120	137	127	90	198	139	1065	1289					
KIAA0316-L																						
FrKIAA0316	16			106	103	46	105	108	132	120	137	127	90	117	57	215	845	108	332			
	15						105	108	132	120	137	127	▨	117	48	215	845	108	457	692	1133	315
PRPS2	7	**209**	184	99	125	174	160	**1514**														
PRPS1	7	**244**	184	99	125	174	160	**1089**														
FrPRPS1	7	119	184	99	125	174	160	90														
RAB9A	1	**940**																				
RAB9B	3	169	74	**806**																		
FrRAB9A	1	603																				
FrRAB9B	1	606																				
GPM6B	7	**191**	187	157	172	74	66	**671**														
PLP	7	**125**	187	262	169	74	66	**2054**														
FrGPM6B	6		188	157	169	74	66	147														
FrPLP	5		188	157	169	74	66															
GLRA2	9	**598**	134	68	224	83	138	215	150	**1606**												
GLRA4	9	**71**	131	68	224	83	141	215	150	**282**												
FrGLRA2	7			127	72	121	138	215	154	269												

Table 6-12 Table showing human gene structure information obtained from Ensembl v15.33.1 (based on the NCBI 33 assembly) and the Xq22-q23 transcript map described in Chapter 3, and *Fugu* gene structure information obtained from Ensembl (Fugu) v15.2.1. Dark row borders separate different Xp/Xq gene pairs and their potential Fugu orthologues. Exon sizes in red type are of equal size in each paralogue/orthologue. Exon sizes in blue type differ by a multiple of 3 (preserving coding frame) between genes. Exons in bold type denote the codons containing the translation start and stop codons. *Fugu rubripes* gene names are pre-fixed “Fr”. Hatched cells represent instances where the following exons in the row have been right-shifted to match the human exons.

Gene	No. exons	Exon sizes (bp)																							
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
BMX	18	138	105	82	120	65	242	78	54	55	80	128	75	172	217	65	119	162	68						
BTK	18	141	99	69	82	129	68	188	63	55	80	128	75	172	217	65	119	158	**500**						
FrBTK	17	141	105	82	126	62	▨	188	63	55	80	128	72	172	220	65	119	158	66						
IL1RAPL1	10	82	280	187	154	75	133	146	144	171	719														
IL1RAPL2	10	82	274	187	154	75	130	146	144	171	698														
FrIL1RAPL1	5						134	146	144	171	725														
DMD	78	190	173	157	121	269	147	79	61	62	75	202	86	158	167	112	137	39	66	66	159	244	124	93	32
DRP2	22	108	164	157	121	269	147	79	61	62	75	202	86	158	167	112	137	66	66	144	238	121	125		
FrDRP2	5	162	121	112	157	150																			
XK	3	**327**	263	**4495**																					
XK-L	3	239	269	**1639**																					
FrXK	3	245	263	704																					
CYBB	13	**81**	96	111	85	146	191	130	93	254	163	147	125	**2671**											
NOX1	13	**251**	96	111	85	152	182	133	93	236	163	147	125	**187**											
FrCYBB	11			108	85	149	182	133	93	254	163	147	125	115											
FrNOX1	12		96	111	85	145	4	173	124	93	242	163	147	123											
SYTL5	16	119	210	116	109	135	142	130	101	93	179	100	162	109	136	209	143								
SYTL4	16	110	216	110	103	102	76	91	104	93	179	103	162	109	100	209	**1683**								
FrSYTL5	7										209	103	162	109	139	209	134								
FrSYTL4	7										182	103	162	103	109	209	134								
SRPX	10		97	60	192	177	127	122	180	134	122	**556**													
SRPUL	11	288	**212**	81	192	177	127	122	180	134	122	**493**													
FrSRPX	8				190	177	127	122	180	134	122	181													
FrSRPUL	8				184	177	124	122	180	134	122	175													
TM4SF2	7	**150**	189	75	96	156	84	69																	
TM4SF6	8	**190**	189	75	99	135	84	**108**	1189																
FrTM4SF2	6	81	189	75	96	156	87																		
FrTM4SF6	5		189	75	96	156	87																		

Table 6-13 Table showing human gene structure information obtained from Ensembl v15.33.1 (based on the NCBI 33 assembly) and the Xq22-q23 transcript map described in Chapter 3, and *Fugu* gene structure information obtained from Ensembl (Fugu) v15.2.1. Dark row borders separate different Xp/Xq gene pairs and their potential Fugu orthologues. Exon sizes in red type are of equal size in each paralogue/orthologue. Exon sizes in blue type differ by a multiple of 3 (preserving coding frame) between genes. Exons in bold type denote the codons containing the translation start and stop codons. *Fugu rubripes* gene names are pre-fixed “Fr”. Hatched cells represent instances where the following exons in the row have been right-shifted to match the human exons.

A different interpretation of the results could be that the *Fugu rubripes* orthologues could in fact be paralogues themselves, generated in a segmental duplication occurring after the divergence of *Fugu* and Human. Such duplications can confound prediction of orthology. This is less likely, given the presence of other non-Xp/q paralogue potential orthologues within the respective regions (e.g. OTC and CSTF2). In order to assess this alternative hypothesis however, phylogenetic analysis was performed using selected *Fugu rubripes* and *Homo sapiens* protein sequences (for genes which appear to have strong orthology support), including sequences from other selected species where available. If the genes were generated as part of a duplication occurring within the *Fugu* lineage, the sequences should be closer to one another than to their potential human orthologues.

In combination with this approach, searches were made for other homologous sequences in other species for phylogenetic analyses. TBLASTN analyses were performed using human Xp/Xq paralogue protein sequences as queries against the non-redundant mRNA database via the NCBI web server. The results were separated according to taxonomy, and the top 2 hits recorded for each species.

The phylogenetic analysis techniques utilised are described in detail in Chapter 2. Briefly, protein sequences were obtained from links to mRNA sequences found by TBLASTN analysis of Genbank at the NCBI as mentioned earlier, in addition to direct download from Ensembl v15.33.1. Alignments were performed and edited, and phylogenetic trees were constructed using both distance and maximum-likelihood methods and are presented in Figure 6-19 – Figure 6-23. Protein sequences were utilised to increase the quality of the alignments and to minimise error due to multiple replacements at sites, due to the long evolutionary period hypothesised.

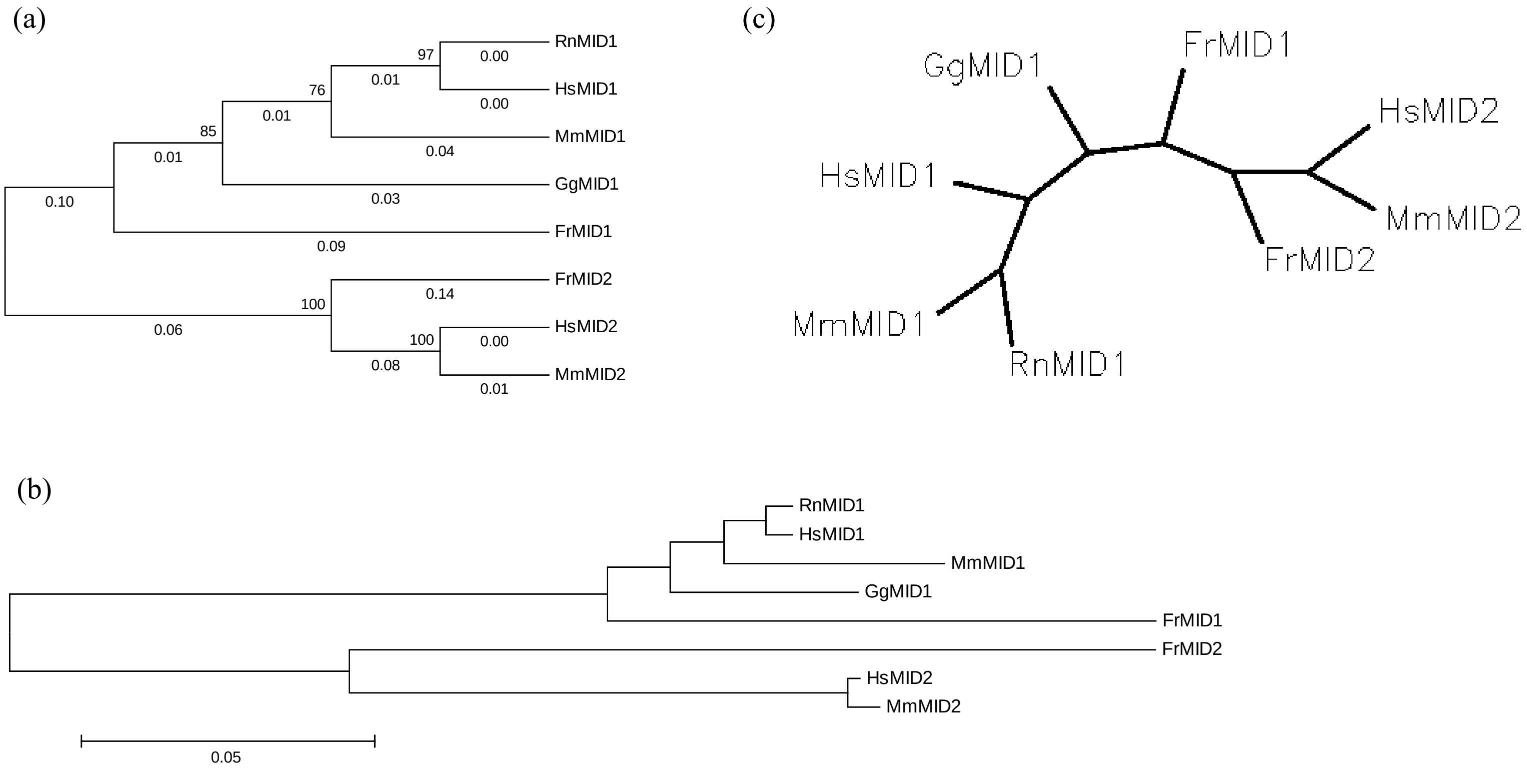


Figure 6-19 The figure shows phylogenetic trees constructed for the MID genes. (a) shows an un-rooted tree constructed using distance measurements. For clarity, only the topology is shown, with distance measurements for each branch shown below the branch and bootstrap support (% agreement from 1000 replicates) shown above the branch. (b) shows the same tree but with branch lengths proportional to distance. (c) shows an un-rooted maximum-likelihood tree constructed from the same alignment. The different organism sequences are denoted by the following pre-fixes: Hs - *Homo Sapiens*, Mm - *Mus musculus*, Fr - *Fugu rubripes*, Rn - *Rattus norvegicus*, Gg – *Gallus gallus*.

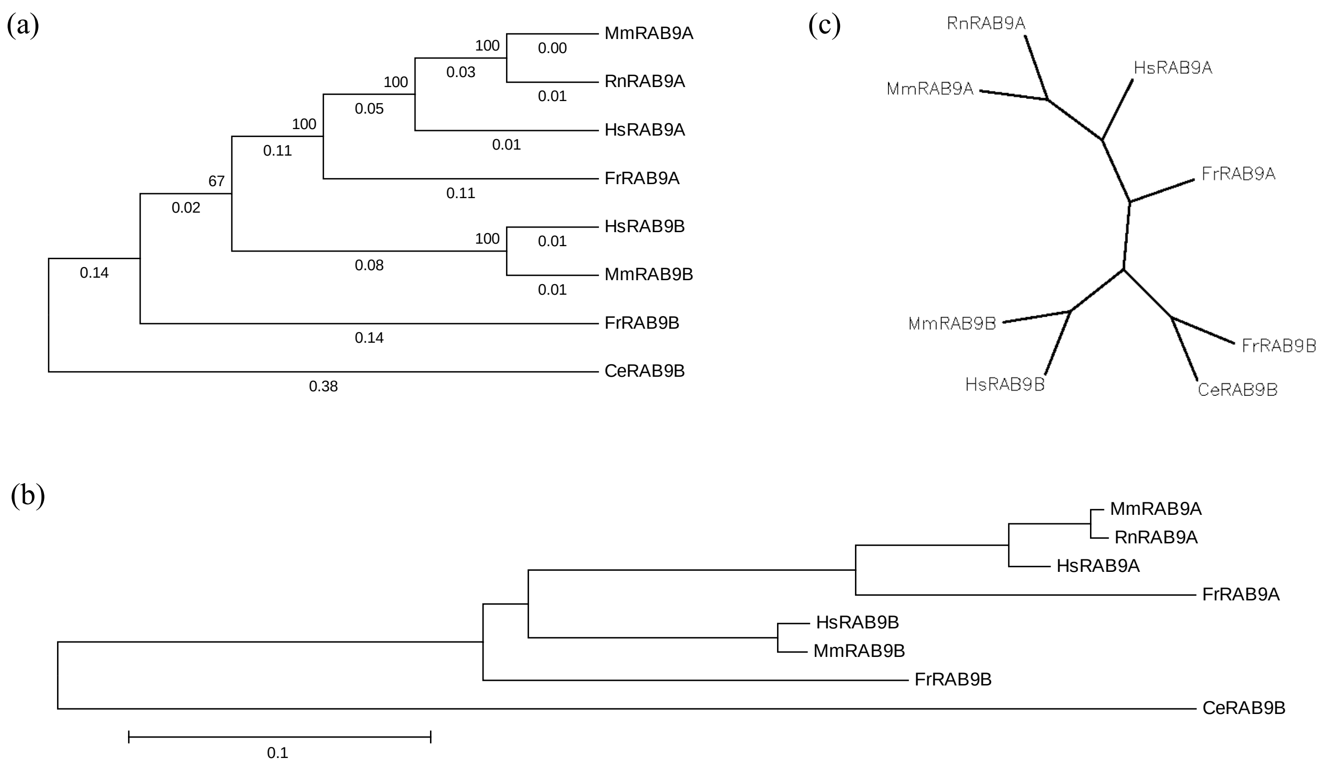


Figure 6-20 The figure shows phylogenetic trees constructed for the RAB genes. (a) shows an un-rooted tree constructed using distance measurements. For clarity, only the topology is shown, with distance measurements for each branch shown below the branch and bootstrap support (% agreement from 1000 replicates) shown above the branch. (b) shows the same tree but with branch lengths proportional to distance. (c) shows an un-rooted maximum-likelihood tree constructed from the same alignment. The different organism sequences are denoted by the following pre-fixes: Hs - *Homo Sapiens*, Mm - *Mus musculus*, Fr - *Fugu rubripes*, Rn - *Rattus norvegicus*, Ce – *Caenorhabditis elegans*.

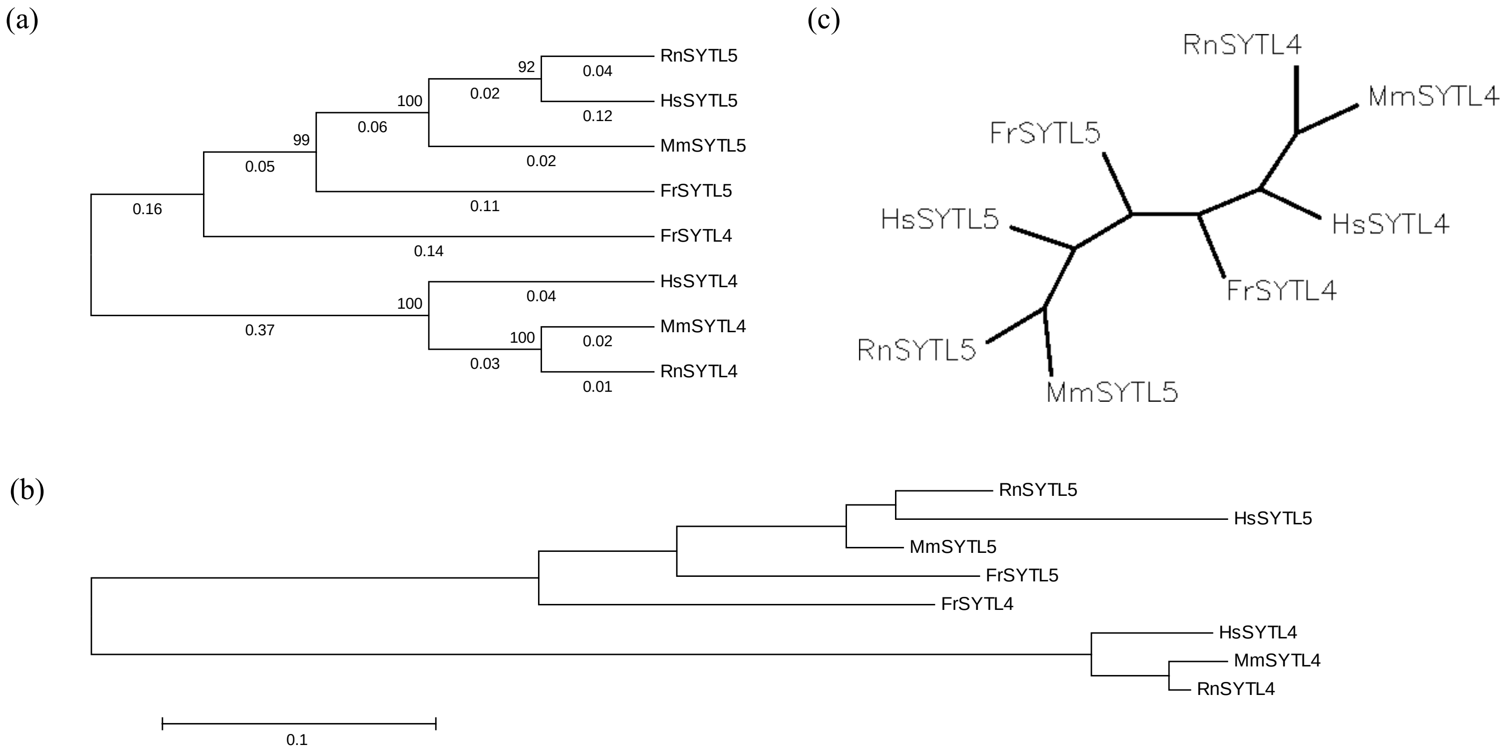


Figure 6-21 The figure shows phylogenetic trees constructed for the SYTL genes. (a) shows an un-rooted tree constructed using distance measurements. For clarity, only the topology is shown, with distance measurements for each branch shown below the branch and bootstrap support (% agreement from 1000 replicates) shown above the branch. (b) shows the same tree but with branch lengths proportional to distance. (c) shows an un-rooted maximum-likelihood tree constructed from the same alignment. The different organism sequences are denoted by the following pre-fixes: Hs - *Homo Sapiens*, Mm - *Mus musculus*, Fr - *Fugu rubripes*, Rn - *Rattus norvegicus*.

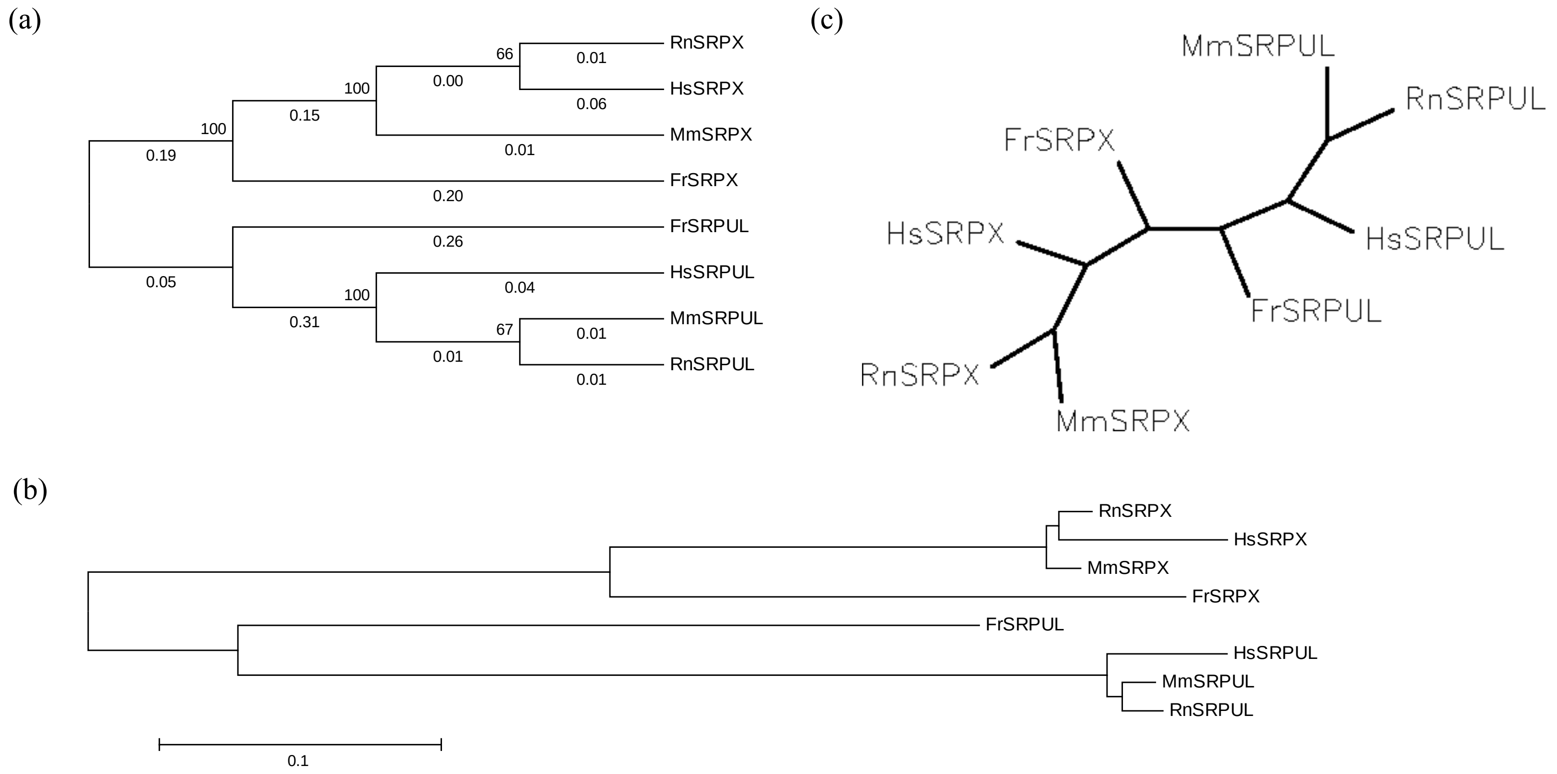


Figure 6-22 The figure shows phylogenetic trees constructed for the Sushi-repeat genes. (a) shows an un-rooted tree constructed using distance measurements. For clarity, only the topology is shown, with distance measurements for each branch shown below the branch and bootstrap support (% agreement from 1000 replicates) shown above the branch. (b) shows the same tree but with branch lengths proportional to distance. (c) shows an un-rooted maximum-likelihood tree constructed from the same alignment. The different organism sequences are denoted by the following pre-fixes: Hs - *Homo Sapiens*, Mm - *Mus musculus*, Fr - *Fugu rubripes*, Rn - *Rattus norvegicus*.

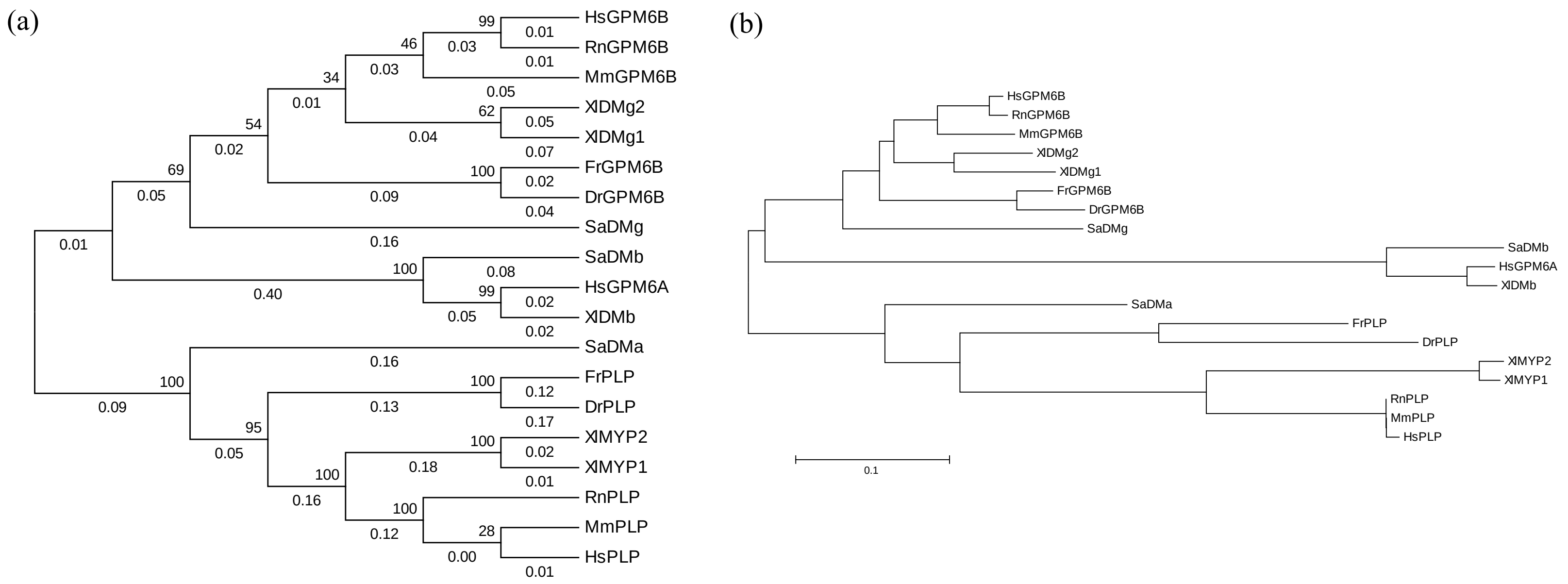


Figure 6-23 The figure shows phylogenetic trees constructed for the lipophilin genes. (a) shows an un-rooted tree constructed using distance measurements. For clarity, only the topology is shown, with distance measurements for each branch shown below the branch and bootstrap support (% agreement from 1000 replicates) shown above the branch. (b) shows the same tree but with branch lengths proportional to distance. No maximum-likelihood tree was computed due to the high number of sequences used increasing the computational intensity. The different organism sequences are denoted by the following pre-fixes: Hs - *Homo Sapiens*, Mm - *Mus musculus*, Fr - *Fugu rubripes*, Rn - *Rattus norvegicus*, Dr – *Danio rerio*, Xl – *Xenopus laevis*, Sa – *Squalus acanthias*.

The phylogenetic analysis data shown above are consistent with the hypothesis that the paralogous genes in *Fugu rubripes* are the true orthologues of the paralogous genes on human Xp/Xq. In this case, it can be predicted that the paralogous pairs were generated by a segmental duplication that occurred greater than 450 Mya. Whilst the RAB and SYTL *Fugu* orthologues do not cluster tightly with their human counterparts, they do not seem to cluster together either as would be predicted if they had arisen from independent duplications within the *Fugu* lineage. In four of the cases shown, tree topology is generally in agreement when calculated by both distance and maximum-likelihood methods. In addition, whilst phylogenetic analyses can be affected by mutation rate heterogeneity amongst sites, due to different parts of the molecules being under different selective pressures, these genes presented appear to have different functions and so no systematic bias should be present.

Whilst further analysis is needed to expand the evidence and broaden the number of genes analysed phylogenetically, these data in combination with the genomic data and literature evidence described earlier strongly support the hypothesis that the segmental duplication giving rise to Xp/q paralogy occurred at least as long ago as the divergence of *Fugu rubripes* and *Homo Sapiens* (~450 Mya) and possibly as long ago as the divergence of cartilaginous and bony fish (~528 Mya). This would mean that the segmental duplication occurred at a time in evolution when a wave of segmental duplications was thought to have occurred, in agreement with Gu *et. al.* (2002) and McLysaght *et. al.* (2002).

6.6 Comparative analysis of *Sminthopsis macroura* genomic sequence

As described in the previous sections, seven *Sminthopsis* BACs were selected for whole-insert sequencing on the basis of hybridisation and FISH results. This was performed in order to assess gene structures of the expected orthologues and to perform comparative analysis between marsupial genomic sequence and that from other organisms.

Clone bF232B10 (KIAA0316 orthologue) was chosen to represent the telomeric Xp paralogy region, bF284I24 (POLA) the intervening region lacking Xq paralogues and bF231M3 (XK) and bF253J14 (SYTL5/SRPX) the centromeric Xp paralogy region.

Clones bF281H15 (SRPUL/SYTL4), bF106P8 (NOX1, XK-L and SRPUL) and bF13K23 (KIAA0316-L) were chosen to represent the Xq22 paralogy region and also to permit comparison with their autosomal counterparts in *Sminthopsis*. For supporting evidence, see sections 6.3 and 6.4.

Clones were picked from the library, grown and their identity validated by *Hin*d III/*Sau* 3AI fingerprinting (compared to results described in section 6.3) by Frances Lovell (Wellcome Trust Sanger Institute), and were subsequently sequenced by the Wellcome Trust Sanger Institute sub-cloning and sequencing teams. The sequences were submitted to EMBL with accession numbers as follows: bF232B10 (BX649239), bF284I24 (BX649240), bF231M3 (BX649270), bF253J14 (BX649259), bF281H15 (BX649310), bF106P8 (BX649374) and bF13K23 (BX649465).

The sequences were analysed and loaded into an ACeDb database and annotated as described in Chapter 3. The annotated genes are tabulated in Table 6-14. This confirmed the presence of genes expected as mentioned above, with the exception of clone bF231M3 (XK). Clone bF231M3 was strongly hybridising with the XK probe, but failed to co-localise with other Xp orthologues by FISH analysis (Section 6.4). It was thought this may represent a re-arrangement, but the sequencing suggested it was a false-positive. Matches to NOX1 were observed in clone bF106P8, but were not sufficiently comprehensive to allow full annotation. Clone bF106P8 was also found to contain a gene not annotated in the orthologous region in Xq22 (bF106P8.SM.1). This gene was similar to human mRNA BC011713 (FLJ20772). BLASTN of BC011713 against the human genome produced a high-scoring match to chromosome 8, but also a partial match ~4 kb proximal to CSTF2, which is consistent with the picture in the marsupial. In the human genome, L1 repeats and retroviral remnants are found just proximal to CSTF2, and it is possible that their insertion obliterated a paralogue of the locus represented by BC011713 subsequent to the divergence of the metatherian and eutherian lineages. A partial match was also found just proximal to Cstf2 in the mouse genome, suggesting that such an event may have occurred prior to the human-mouse divergence (the highest-scoring match to the mouse genome was to chromosome 15 in a region with shared synteny with human chromosome 8).

Clone	Accession	Annotated locus	No. exons	Human Orthologue
bF231M3	BX649270	none		none
bF232B10	BX649239	bF232B10.SM.1	2	KIAA0316
bF284I24	BX649240	bF284I24.SM.1	14	POLA
bF253J14	BX649259	bF253J14.SM.1	9	SYTL5
		bF253J14.SM.2	3	SRPX
bF281H15	BX649310	bF281H15.SM.1	9	SRPUL
		bF281H15.SM.2	14	SYTL4
bF106P8	BX649374	bF106P8.SM.1	7	Sim. FLJ20772
		bF106P8.SM.2	14	CSTF2
		Homology found		NOX1
		bF106P8.SM.4	3	XK
bF13K23	BX649465	bF13K23.SM.1	10	KIAA0316-L

Table 6-14 Marsupial clone sequences and genes annotated.

6.6.1 Comparative analysis of sequence composition for human, mouse and Sminthopsis macroura

The compositions of the sequences were examined in order to assess how they differed with respect to repeat and GC content. If the duplication leading to the Xp and Xq paralogy blocks was as old as suggested in the previous section, differences in GC and repeat content may be expected. In addition, as the Xp paralogy block remained autosomal until relatively recently, differences in repeat content may distinguish these sequences from those which are on the X chromosome in all the mammals, which since the latter have possibly are more likely to have been recruited into the X inactivation system (based on the hypothesis that LINE repeats may be involved in the inactivation mechanism).

Sequences BX649239, BX649240, BX649259, BX649310, BX649374 and BX649465 were retrieved via NCBI Entrez and subjected to repeat and GC content analysis via the RepeatMasker web-server. The results for each clone were collated from the RepeatMasker analysis reports.

In order to compare the composition of marsupial sequences with that of mouse and human, for each marsupial clone the exons nearest each end of the insert were located and their sequences translated. These sequences were used to identify similar sequences in the human and mouse genomes by TBLASTN analysis (Ensembl Human v19.34a.1, NCBI 34 assembly and Ensembl Mouse v19.30.1, NCBI 30 assembly). The locations of highest matches were noted and extended by the distances between the respective marsupial exons and the end of the corresponding insert. These orthologous human and mouse genomic regions were exported from Ensembl, subjected to repeat and GC content analysis via the RepeatMasker web-server and the results collated.

The results of these sequence composition analyses for marsupial, human and mouse are presented in Table 6-15.

Clone	length	%GC	% interspersed	% simple	% low complexity	% masked	% SINE	% MIR	% LINE	% L1	% L2	% L3	Chromosome	Gene(s)	Organism
bF232B10	40274	36.04	17.46	1.3	1.22	19.99	12.2	4.08	4.04	1.89	2.14	0	A	KIAA0316	Sm
232B10Hs2	41161	39.64	40.78	1.16	0.24	42.45	10.83	2.63	19.28	16.43	2.85	0	Xp		Hs
232B10Mm2	42473	39.18	43.44	1.57	0.56	45.58	7.59	0.78	28.63	28.06	0.57	0	X F5		Mm
bF284I24	42474	33.11	14.25	0.77	1.17	16.21	6.73	3.46	6.87	3.54	2.78	0.56	A	POLA	Sm
284I24Hs2	41009	38.44	38.36	0.4	0.38	39.14	16.18	3.18	17.99	8.57	9.42	0	Xp		Hs
284I24Mm2	51478	36.7	38.05	1.21	0.24	39.5	8.3	0.58	24.48	23.26	1.22	0	X C1		Mm
bF253J14	66719	33.97	23.89	2.81	1.43	28.1	8.82	3.48	14.85	5.66	7.55	1.64	A	SYTL5/SRPX	Sm
253J14Hs2	67910	38.62	39.48	0.82	0.2	40.5	4.98	2.17	17.05	13.61	3.24	0.21	Xp		Hs
253J14Mm2	126772	38.61	34.15	2.66	0.33	37.09	2.72	0.06	25.81	25.3	0.52	0	X A1.2		Mm
bF281H15	67497	45.04	25.28	1.72	1.39	28.89	7.32	3.05	15.33	5.14	9.29	0.9	X	SRPUL/SYTL4	Sm
281H15Hs2	62307	42.23	41.61	0.23	0.66	42.49	14.54	3.74	26.94	15.94	10.22	0.78	Xq		Hs
281H15Mm2	57925	41.75	24.04	2.08	0.53	26.92	8.06	1.57	12.9	10.34	2.2	0.36	X E3		Mm
bF106P8	112071	43.77	20.46	1.07	0.89	22.46	5.41	2.46	14.98	10.28	2.59	2.11	X	NOX1/XK-L/CSTF2	Sm
106P8Hs2	138792	40.48	51.64	0.79	0.43	52.86	16.34	2.57	23.06	21.87	0.79	0.4	Xq		Hs
106P8Mm2	133602	40.92	41.04	1.81	0.35	43.43	8.73	1.21	22.89	22.35	0.39	0.15	X E3		Mm
bF13K23	59918	44.67	17.13	3.52	2.35	22.99	7.72	2.61	8.27	2.05	4.15	2.07	X	KIAA0316L	Sm
13K23Hs2	56593	40.46	30.6	0.58	0.44	31.63	9.1	2.89	12.53	11.96	0	0.57	Xq		Hs
13K23Mm2	67411	43.55	34.87	2.76	0.12	37.76	16.55	0.56	11.8	8.89	2.91	0	X F1		Mm

Table 6-15 Sequence composition data from RepeatMasker analysis of marsupial, human and mouse orthologous regions. Sequences from each organism are grouped for each region, and are listed in order Xpter-Xqter respective to the human X chromosome. Human and mouse sequences are named with the marsupial clone name they are orthologous to, with a suffix "Hs2" for human and "Mm2" for mouse. A = autosome. Paralogous loci are coloured similarly.

The most striking features of the composition data are the differences in GC content seen between the sequences on Xp and Xq in human, which are autosomal and X chromosomal in marsupial respectively. A lower GC content is seen for those sequences which are Xp/autosomal. This feature is much more pronounced in the marsupial sequences than in the human and mouse sequences. Specifically, the marsupial autosomal sequences have a lower GC content than their X chromosome counterparts in human and mouse, and the marsupial X chromosome sequences have a higher GC content than the human or mouse X chromosome sequences.

Another major feature is the increased interspersed repeat content of the human and mouse sequences compared to the marsupial. Examination of the data shows this to be mainly due to LINE, particularly L1 repeats. No major trends in simple repeats, low complexity regions or SINE were noted. The lengths of the genome sequences in the different organisms were also relatively uniform, with the notable exception of the region represented by clone bF253J14, where the mouse sequence was almost double the size of the human and marsupial sequences.

6.6.2 Comparative sequence analysis of the CSTF2/NOX1/XK-L region in human, mouse, Sminthopsis macroura and Fugu rubripes using PIP and VISTA

As marsupial sequence analysis has been suggested as a useful aid to human gene (and other functional element) identification, with a lower background of sequence homology in non-functional regions compared to mouse (Chapman *et al.*, 2003), a study was undertaken to compare a region of sequence between human, mouse, *Sminthopsis macroura* and *Fugu rubripes*. For this study, the region containing the CSTF2, NOX1 and XK-L genes was chosen, because it was the most gene-rich marsupial sequence identified, and the orthologous region in *Fugu* was also available (see Section 6.5).

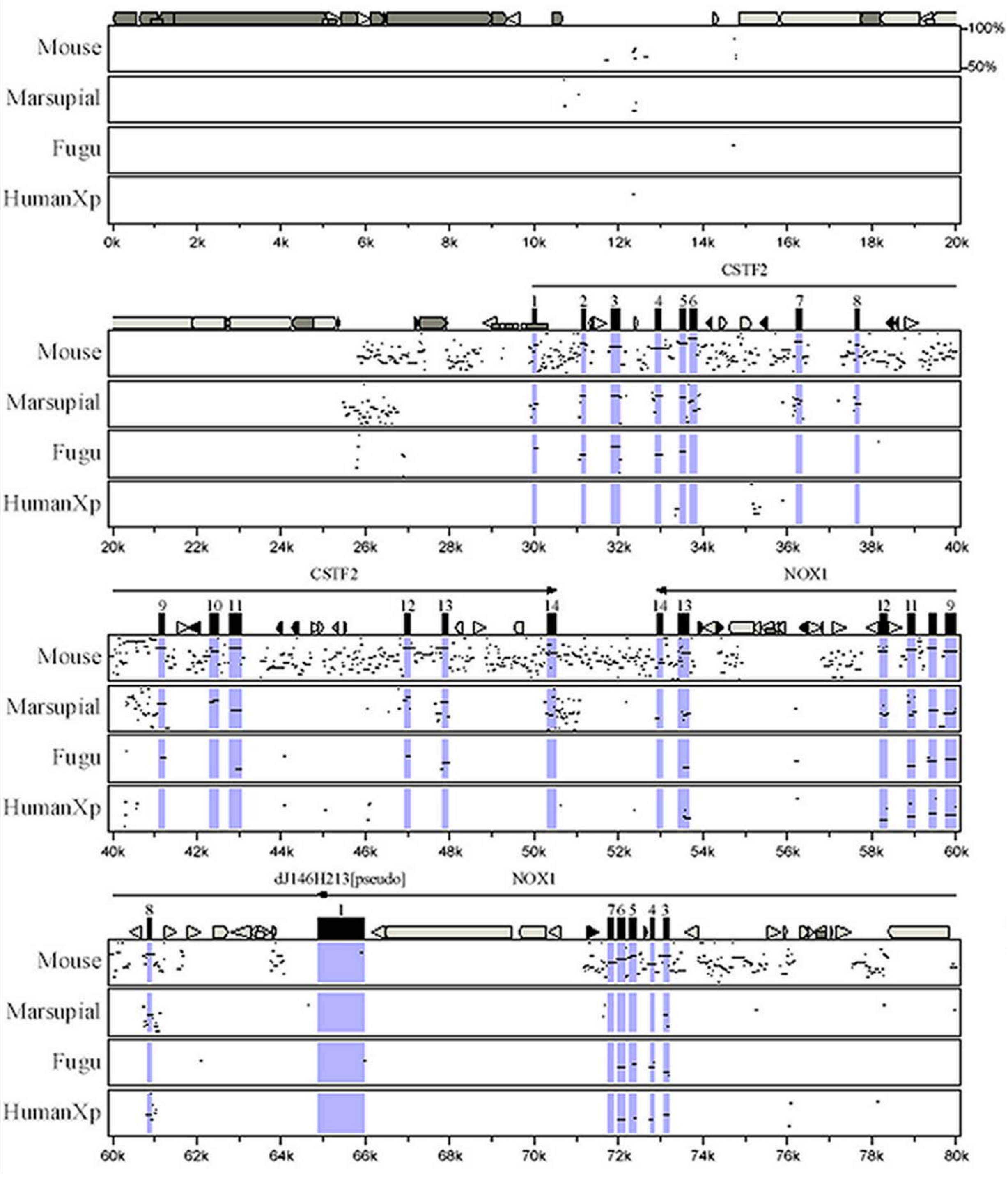
As the studies described in Section 6.5 have argued that the duplication leading to Xp and Xq paralogy occurred prior to human-*Fugu* divergence, and because NOX1 and XK-L were involved in the duplication, the human Xp paralogous region was also included in the comparative analysis. If the duplication was indeed ancient, the results of the human Xp/Xq comparison would be expected to be relatively similar to the human Xq/Fugu comparison, and less similar to the human Xq/mouse and human Xq/marsupial comparisons.

The sequences used for human Xq, mouse and marsupial were bF106P8, 106P8Hs2, and 106P8Mm2, respectively, as described in the previous section. The Fugu and human Xp region sequences were identified in Ensembl Fugu v19.2.1 and Ensembl Human v19.34a.1 respectively, and the genomic regions encompassing the paralogous genes were exported. The comparative sequence analysis tools PIP and VISTA were both used for the analysis, following instruction given by the authors. Detailed methods are given in Chapter 2. Both methods were used, as they employ different methodologies to perform the comparisons. In each case, the human Xq sequence was used as the base sequence and was masked for repeats (using RepeatMasker). The exon annotations for this sequence were also used. A representative PIP and VISTA plot are shown on the following pages (Figures 6-24 and 6-25 respectively).

From these analyses, PIP appeared to be more sensitive using the parameters described in Chapter 2. PIP identified similarities to cU131B10.CX.1 (XK-L) exons one and two, which were missed by VISTA, in the human Xp sequence. Both programs successfully identified exons for CSTF2 in marsupial, mouse and Fugu, and for NOX1 and cU131B10.CX.1 (although only weakly for Fugu and human Xp using VISTA) in all sequences including the human Xp region. No matches were seen as expected for CSTF2 in the human Xp sequence, as there is no Xp paralogue for CSTF2 noted.

The marsupial sequence showed a reduced background of sequence conservation in non-exonic regions compared to mouse, and yet all fourteen exons of CSTF2 and all three exons of cU131B10.CX.1 could be identified. As noted earlier, NOX1 was not annotated in the marsupial sequence although matches to NOX1 were seen, and this is reflected in the PIP and VISTA plots, where although exons 1,2,3,8 and 9-14 can be detected in marsupial in the PIP plot, exons 4,5,6 and 7 remain undetected. This could reflect differences in gene structure between human and marsupial, and further studies could be aimed at determining if NOX1 is indeed expressed in marsupials.

The levels of sequence conservation seen for the human Xp region are consistent with the studies presented in Section 6.5, with a low background seen in non-exonic regions and exonic sequence identity levels similar to those seen for *Fugu*. This supports the hypothesis that the duplication generating Xp/Xq paralogy is a relatively ancient event.


100%
50%
Mouse
Marsupial
Fugu
HumanXp
0k
2k
4k
6k
8k
10k
12k
14k
16k
18k
20k
CSTF2
1
2
3
4
5
6
7
8
22k
24k
26k
28k
30k
32k
34k
36k
38k
40k
CSTF2
NOX1
9
10
11
12
13
14
14
13
12
11
9
42k
44k
46k
48k
50k
52k
54k
56k
58k
60k
dJ146H213[pseudo]
NOX1
8
1
7
6
5
4
3
62k
64k
66k
68k
70k
72k
74k
76k
78k
80k

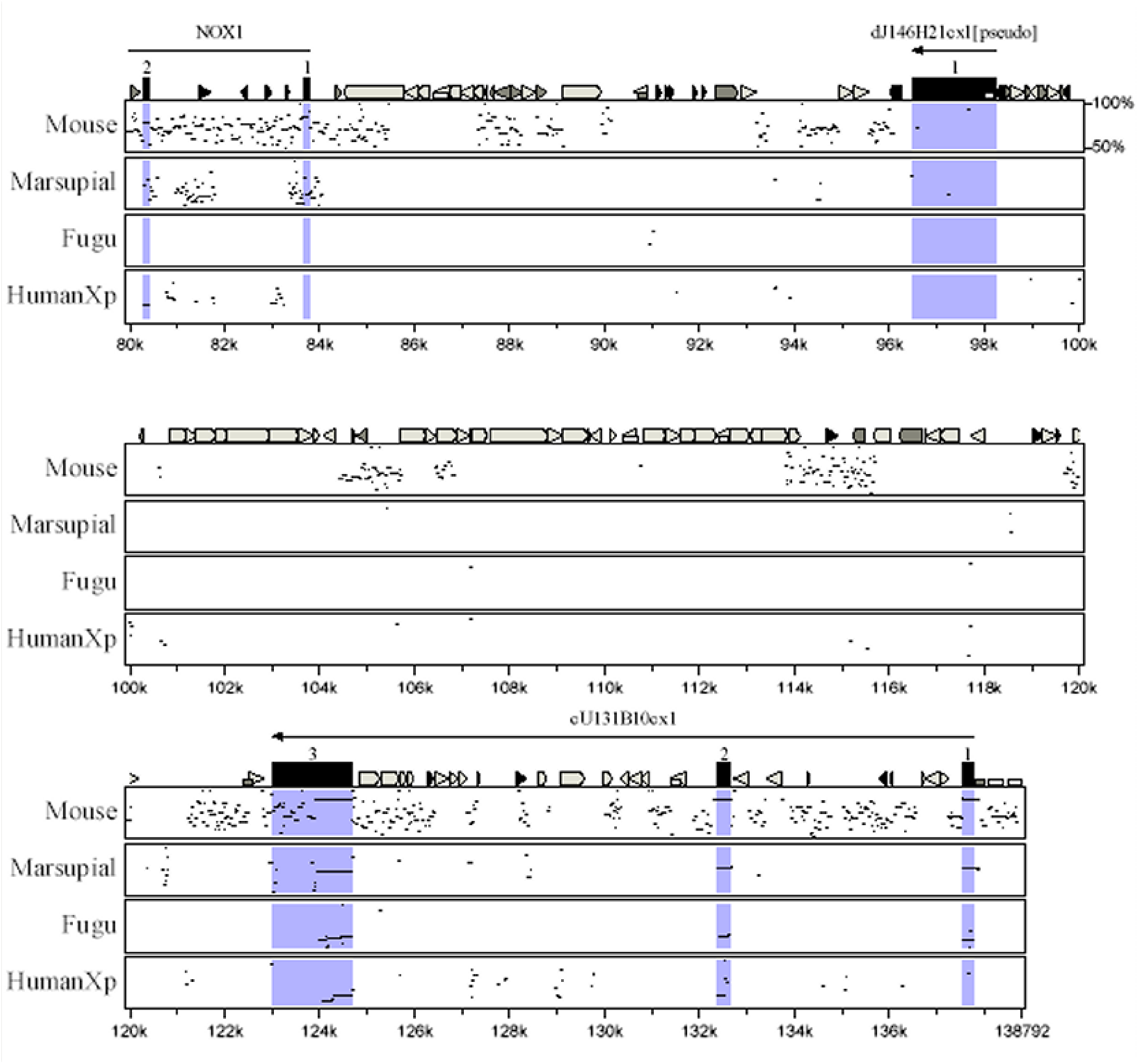
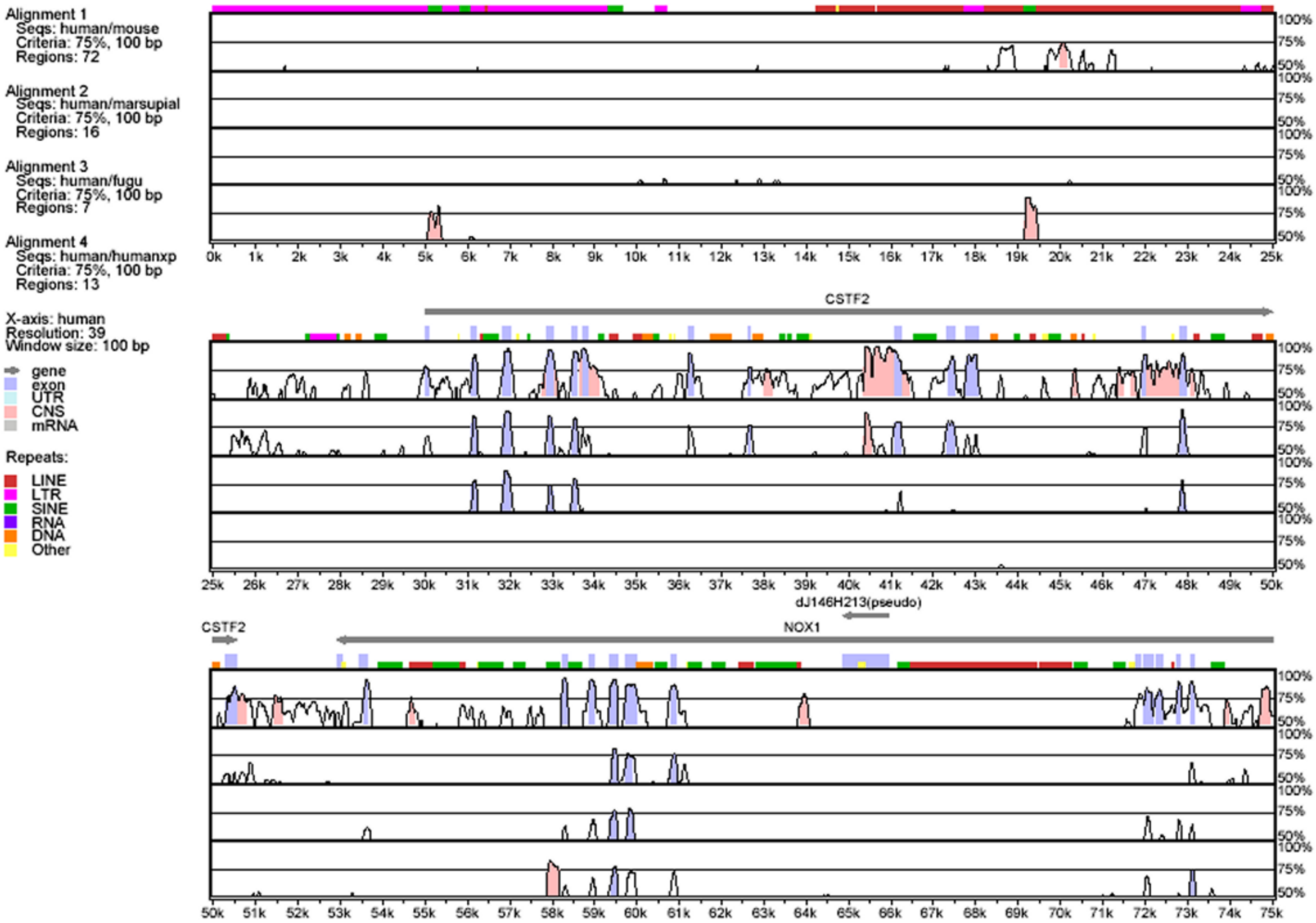


Figure 6-24 PIP plot of the human Xq22 region encompassing genes CSTF2, NOX1 and XK-L. Exonic regions are shaded blue and marked and numbered by vertical black boxes. Regions of high sequence identity to the orthologous mouse, marsupial and *Fugu* regions, and the paralogous Human Xp region, are depicted by horizontal black lines in the PIP. Masked repeats are denoted by boxed arrows.


Alignment 1
Seqs: human/mouse
Criteria: 75%, 100 bp
Regions: 72
Alignment 2
Seqs: human/marsupial
Criteria: 75%, 100 bp
Regions: 16
Alignment 3
Seqs: human/fugu
Criteria: 75%, 100 bp
Regions: 7
Alignment 4
Seqs: human/humanxp
Criteria: 75%, 100 bp
Regions: 13
X-axis: human
Resolution: 39
Window size: 100 bp
gene
exon
UTR
CNS
mRNA
Repeats:
LINE
LTR
SINE
RNA
DNA
Other
100%
75%
50%
0k 1k 2k 3k 4k 5k 6k 7k 8k 9k 10k 11k 12k 13k 14k 15k 16k 17k 18k 19k 20k 21k 22k 23k 24k 25k
CSTF2
25k 26k 27k 28k 29k 30k 31k 32k 33k 34k 35k 36k 37k 38k 39k 40k 41k 42k 43k 44k 45k 46k 47k 48k 49k 50k
dJ146H213(pseudo)
CSTF2
NOX1
50k 51k 52k 53k 54k 55k 56k 57k 58k 59k 60k 61k 62k 63k 64k 65k 66k 67k 68k 69k 70k 71k 72k 73k 74k 75k

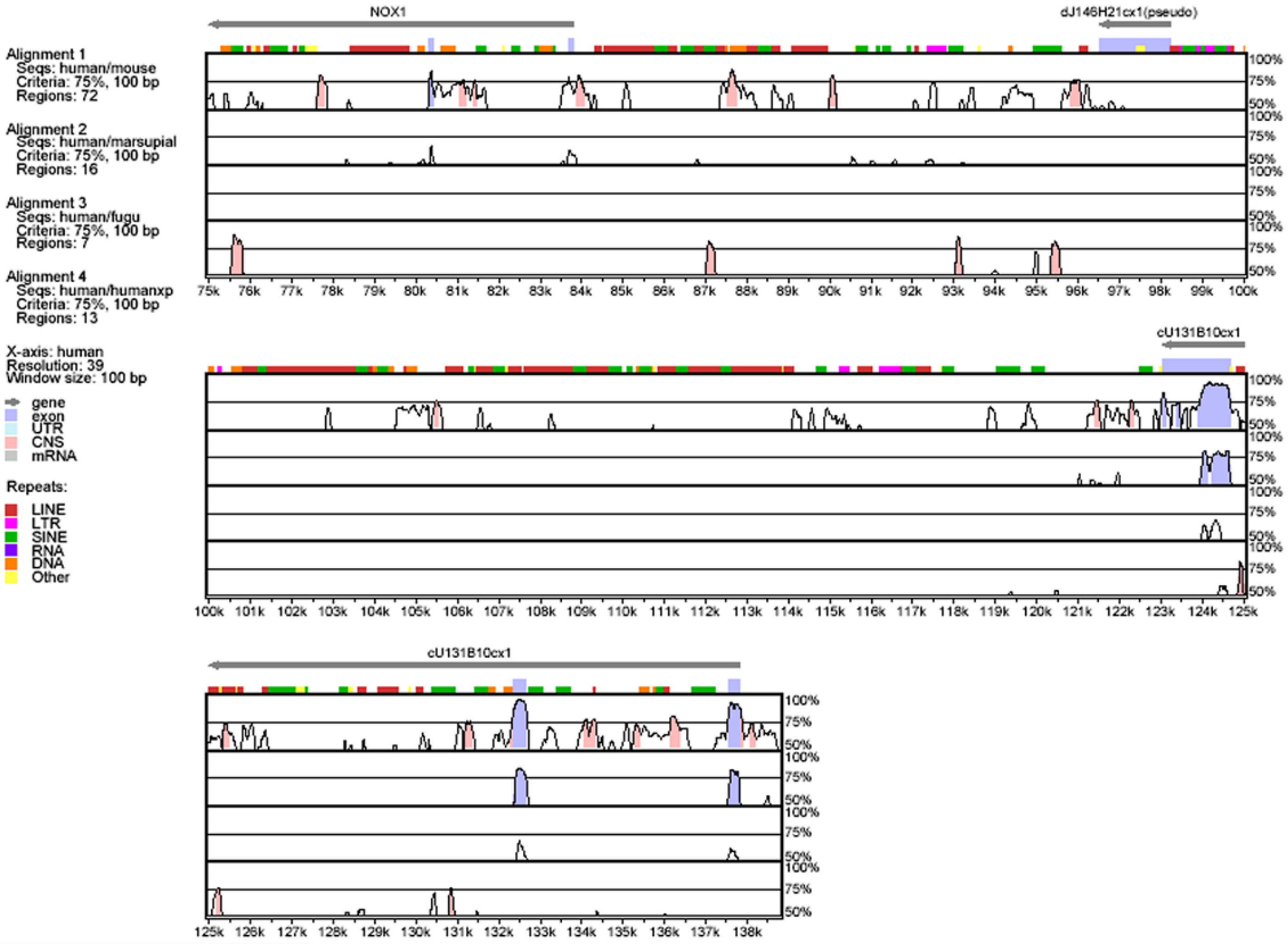


Figure 6-25 VISTA plot of the human Xq22 region encompassing genes CSTF2, NOX1 and XK-L. The figure legend is given in the diagrams. Regions of high sequence identity are depicted by blue peaks in the plot, with other regions of significant similarity shown as light-red peaks.

6.7 Discussion

This Chapter has presented evidence supporting the hypothesis that a segmental duplication was responsible for generating paralogy between human Xp and Xq. The data discussed have expanded the number of genes previously noted as sharing Xp/Xq paralogy from 4 pairs to 15 pairs. Furthermore, it has been demonstrated that the duplication was not a result of an intra-chromosomal duplication within the mammalian X chromosome as previously suggested (Perry *et al.*, 1999) but was instead generated from an ancestral chromosome of unknown origin. Subsequently, the region represented by Xq22-q23 was incorporated into an ancestral X chromosome, whilst the region represented on Xp became incorporated onto the X chromosome subsequent to the metatherian/eutherian mammal divergence.

The marsupial mapping data shown also provide further evidence to support the hypothesis that much of the region now represented by human Xp was localised to the ancestral X chromosome in a single addition from an autosome (Glas *et al.*, 1999). The mapping information and methodologies employed have expanded our knowledge and will allow further analysis of these regions in the marsupial.

Data presented support the hypothesis that the segmental duplication described was a relatively ancient event, occurring at least ~450 Mya. This puts the duplication in context with other genome-wide analyses of segmental and tandem duplications, and suggests that the duplication occurred at a time when a wave of segmental duplications was thought to have occurred.

Analyses assessing the evolution of the regions have been described and a model for the evolution of the regions is illustrated in Figure 6-26 below.

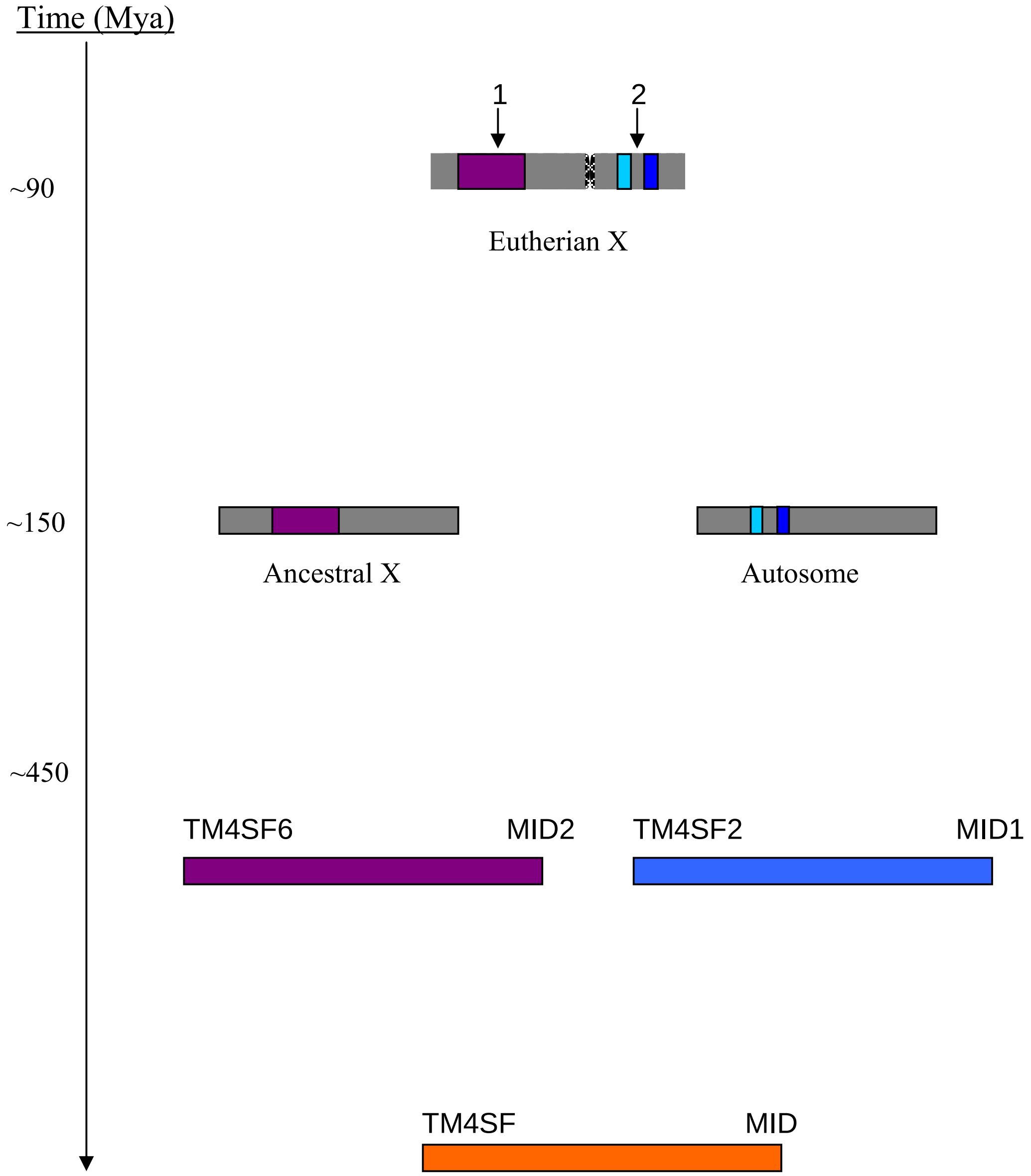


Figure 6-26 Diagram summarising analyses presented in this Chapter and providing a model for the evolution of the Xp/Xq paralogous regions. Duplication of an ancestral genomic segment (orange) generated two paralogous regions (purple and blue). These then diverged in composition, with one segment localising to an ancestral, mammalian X chromosome and one to an autosome. The autosomal region then became localised to the eutherian X chromosome. Arrow 1 denotes the region of paralogy described in Chapter 5. It remains unclear whether this was gained or lost from the other region. Arrow 2 denotes the large non-paralogous block containing SAT and POLA. It is unclear whether this was lost from the other paralogous region or gained here.

The establishment of the genes involved in this duplication and its characterisation allow further information to be brought to bear in evolutionary studies of the 15 genes involved, some of which are of medical importance. As all 15 gene pairs would have been generated at the same time, and have possibly been undergoing different selective pressure for greater than 450 Mya, this information will provide context for studies of divergence of function and the relative selective pressures.

The sources of information employed in the analyses presented reflect the expansion of genomic resources within a short period of time and their utility. This includes availability of marsupial BAC resources, human genomic sequence information and also the generation of WGS assemblies, in this case for *Fugu rubripes*. The availability of even draft quality genomic sequence allows important contextual information to be considered in the generation and testing of hypotheses regarding genome evolution.

Further studies on the Xp/q paralogous regions beyond the scope of this thesis could shed further light on their evolutionary history. Genomic sequence information from other organisms diverging at earlier evolutionary branches would be particularly informative for establishing the date of the segmental duplication. Organisms such as the lamprey and hagfish (agnathans) are currently the focus of such studies for other regions of paralogy such as those involving the MHC region. Further studies examining the relationships between the additional autosomal paralogues of the Xp/q paralogues (e.g. Utrophin) and also of other X chromosome genes potentially involved in the segmental duplication described (e.g. PHKA1/PHKA2) would also be useful.

At this stage several questions regarding the paralogous regions remain. One is the origin of the block of extensive gene duplications seen within Xq22 and described in Chapter 5. Was this block present in the ancestral region before the segmental duplication and lost from the Xp region, or was it instead gained by the Xq22 region? Also, several rearrangements have been noted between the paralogous regions, involving the IL1RAPL genes and the PRPS and KIAA0316 genes. A rearrangement was also presumably responsible for truncating the DRP2 gene, which was thought to have evolved from an ancestral dystrophin-like extended gene structure. The timing and extent of these events is currently unclear. Finally, it is not known from these studies whether the large non-paralogous region represented by SAT and POLA was gained by the Xp region or lost from the Xq22 region.

It is an interesting apparent coincidence that although the segmental duplication described here appears to have occurred at an early stage in vertebrate genome evolution, both regions resulting from the duplication came to reside on what is now the mammalian X chromosome, with one region being added to the X much more recently subsequent to the divergence of marsupials and eutherian mammals. The implications of this, if any, are unclear at present. Studies on X chromosome inactivation for the genes involved may yield interesting information in this regard.

Ultimately, studies of this nature illustrate the utility of genomic sequence information in providing contextual detail that takes us beyond studies of simple gene-to-gene relationships and preserves information regarding genome evolution, in this case from an event which appears to have occurred at a time when all life on earth was believed to be confined to the oceans and selective pressures would have been quite different to those today.