Chapter Five - Characterisation of extensive gene duplication discovered within human Xq22-q23

5.1 Introduction

In the process of annotating genes on the sequence of human Xq22-q23 (Chapter 3), sequence similarities were noted between different loci within the region. Further investigation of these sequences revealed a large number of paralogous loci, many of which appear to be expressed genes. Fourteen sets of paralogous loci were found, with numbers of paralogues ranging from two to ten. The large number of paralogous loci discovered prompted further investigation of the extent of paralogy within the region, and of the genes involved, in order to better understand the evolution of Xq22 and relationships between the different loci. These studies are presented in this chapter, for each of the gene families discovered.

The gene families identified were as follows: NADE-like, NB-thymosins, ALEX-like, GASP-like, pp21-like, Rab-like, COL4A5/COL4A6, TEX13A/TEX13B , NXF-like, TCP11-like, PRO0082 pseudogenes, Histone H2B, cU116E7.CX.2-like and cU116E7.CX.3-like. The positions of the genes within Xq22 are described in Chapter 3 and shown in Figure 5-1.

The Xq22 region appears to be a mosaic of paralogous loci (formed by various sequence rearrangements), and contains an unusually high level of paralogues with respect to neighbouring regions of the X chromosome (data not shown and Gareth Howell, PhD thesis, Open University) with several loci showing very high levels of nucleotide sequence similarity. This suggests that some of the duplication events generating highly-related paralogues either occurred relatively recently in evolution, or alternatively gene conversion could be maintaining homology. This chapter presents detailed comparisons of the human and mouse paralogues (described in Chapter 4) within the Xq22 and X E3-F2 regions respectively, in an attempt to further understand the evolution of Xq22.

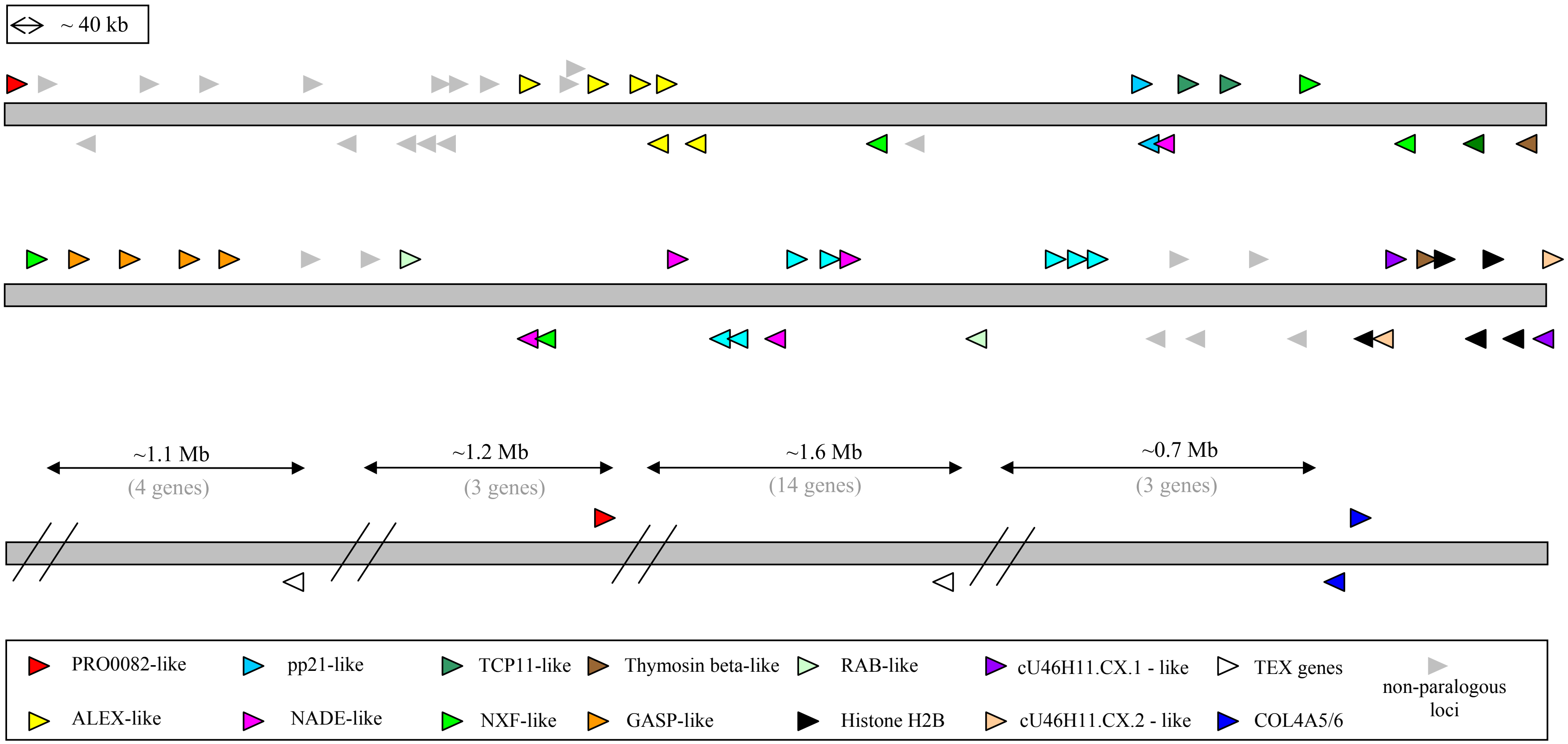


Figure 5-1 Diagram illustrating positions of duplicated loci within Xq22. The first two solid bars depict the ~ 3.2 Mb region bounded by genomic clones dJ341D10 (Z97985) and cU46H11 (Z82254). Genes and directions of transcription are denoted by the arrow direction. Paralogue families and non-duplicated loci are shown as in the key above. The third solid bar represents the ~ 4.6 Mb region (not to scale) bounded by dJ233G16 (AL135959) and dA191P20 (AL034399). Sizes of intervals are shown and numbers of non-duplicated genes between the paralogous loci are given in grey in backets. Pseudogenes are omitted for clarity. Locus names are given in Chapter 3, Table 3-2.

The paralogy noted within the region also raised the question of how the gene functions of paralogues may have diverged. This may have a bearing on studies aimed at identifying genes for genetic disorders mapped within the region (see Chapters 1 and 3). Experiments were designed to assess the expression patterns of the genes, and to explore whether results were consistent with the DDC hypothesis (Force *et al.*, 1999)(see Chapter 1) which states that expression patterns of duplicated genes diverge more quickly than their protein functions.

Some of the gene families contained genes with a level of functional annotation and had been previously described in the literature:

5.1.1 Thymosin-beta genes

Two thymosin-beta genes were annotated within human Xq22. The beta-thymosin proteins have a role in the sequestering of actin monomers, thus modulating actin polymerisation processes (Huff *et al.*, 2001). Members of this family are well conserved, and have been cloned from multiple species. Transcription of beta-thymosins is differentially regulated, suggesting that different family members may play subtly different roles, in accordance with the DDC hypothesis. Extracellular beta-thymosin has also been noted, and there is speculation that there may be a signalling role, similar to cytokines, for beta-thymosins.

5.1.2 NADE family genes

Five NADE family genes were annotated within human Xq22. Members of the NADE (NGFRAP) family have been previously described under several different names. An mRNA sequence pHGR74 was initially characterised in 1990 as a gene expressed in human ovarian granulosa cells (Rapp *et al.*, 1990). Subsequently, three genes named as Bex1, Bex2 and Bex3 were described, from studies aiming to identify imprinted loci in mouse (Brown and Kay 1999). A cDNA described as Rex-3 was also published that was found to be identical to Bex1 (Faria *et al.*, 1998). In 2000, a protein that associates with the p75 neurotrophin receptor (p75NTR) was described, termed NADE (p75NTR-associated cell death executor), and found to be involved in p75NTR-mediated apoptosis in response to nerve growth factor (NGF) (Mukai *et al.*, 2000). NADE (BEX3, pHGR74) has subsequently been named nerve growth factor receptor (TNFRSF16) associated protein 1 (NGFRAP1). Subsequent to studies conducted for this thesis, four further isoforms (NADE2-5) have been noted, confirming the results presented here (Mukai *et al.*, 2003).

5.1.3 NXF family and TCP11-like genes

Five NXF genes, two TCP11-like genes and a TCP11-like pseudogene were annotated within human Xq22. The NXF family of genes were discovered early in analyses performed for this thesis, and have subsequently been well-described in the literature as a family of genes encoding proteins involved in export of mRNA from the nucleus. In particular, NXF2 has been shown to bind RNA and facilitate its export, whilst NXF3 appears to lack this functionality (Herold *et al.*, 2000). The NXF proteins have also been described as forming hetero-dimers with NXT proteins, and intriguingly one of these was also annotated within Xq22-q23 (see Chapter 3). The NXF genes have been the subject of intensive study in several species, and NXF orthologues have also been described in *Drosophila* and *C. elegans*.

T-complex protein 11 (Tcp11) was initially described as a gene residing in the mouse t-complex and expressed in testis (Mazarakis *et al.*, 1991). Subsequently, a human testis-specific homologue, TCP11, was cloned and mapped to 6p21 (Ma *et al.*, 2002). The protein product of *TCP11* is a receptor for fertilisation-promoting peptide (FPP) with a proposed role in sperm function. The NXF and TCP11-like loci are discussed together in this chapter as transcripts were discovered that suggest that these loci may reflect a gene fusion event (see Chapter 3).

5.1.4 ALEX family genes

Five ALEX-like genes and a probable pseudogene were annotated in human Xq22. Early in the studies presented in this chapter, three ALEX family genes were reported (Kurochkin *et al.*, 2001), one of which corresponded to the mRNA for KIAA0512. These genes were reported as ARM (ARMadillo) repeat containing genes and were mapped to Xq21.33-q22.2. The same authors suggested a reduction or loss of expression of ALEX1 and ALEX2 in carcinoma samples, compared to widespread expression in the normal tissues studied.

5.1.5 GASP family genes

Four GASP-like genes were annotated within human Xq22. Initially, the KIAA0443 gene was annotated within Xq22. As will be discussed later, neighbouring genomic sequence was found to contain three further paralogues. A role has been proposed for the KIAA0443 gene product in lysosomal sorting of G-protein coupled receptors (GPCRs) following endocytosis (Whistler *et al.*, 2002); the same work demonstrated binding of the protein to the cytoplasmic tail region of the delta-opioid receptor. On this basis, the gene was renamed *GASP* (GPCR-associated sorting protein). For reasons discussed later, this information may shed light on the functions of many proteins within Xq22.

5.1.6 pp21/TCEAL1 family genes

Nine pp21/TCEAL1 family genes were annotated within human Xq22. During this study, the TCEAL1 (Pillutla *et al.*, 1999) gene was annotated within human Xq22, and subsequently similar genes within the region were identified and annotated. TCEAL1, or p21/SIIR, is a nuclear phosphoprotein involved in regulation of transcription. Although its role and mode of action remain only partially understood, it is not thought to bind DNA directly but to exert its action via interaction with other factors.

5.1.7 Rab-like genes

Two RAB-like genes were annotated within human Xq22. Rab genes have been implicated in vesicle trafficking (Takai *et al.*, 2001).

5.1.8 Histones and cU46H11.CX.1/cU116E7.CX.1 genes

Five histone genes were annotated within human Xq22. Histones play a key role in chromatin formation. Two pairs of genes of unknown function were also annotated. One pair of genes is similar to a paraneoplastic cancer-testis antigen (Rosenfeld *et al.*, 2001), the other, to a mitochondrial carrier protein.

5.1.9 Tex genes and COL4A5/COL4A6

Two Tex genes, and COL4A5 and COL4A6 were annotated within human Xq22. The Tex genes were discovered by studies looking for genes preferentially expressed in spermatogonia, and may play roles in sperm function and fertility (Wang *et* *al.*, 2001). The collagen genes are key components of connective tissue, and defects in COL4A5 cause Fabry disease (OMIM:301500).

This chapter presents the results of analyses of the groups of paralogous genes within the region. The relationships of the paralogue sequences to one another and the paralogy seen within the equivalent region in the mouse were assessed in order to more fully understand the evolution of the genes and the region. The expression patterns of the genes were also assessed to investigate whether the patterns of expression of paralogues are consistent with the DDC hypothesis.

5.2 Duplicated genes within Xq22-q23: sequence analysis, comparative analysis, phylogenetic analysis and RT-PCR expression profiles

Initial analyses focussed on determining which genes within the region formed paralogous groups. This was achieved through BLAST analyses of loci annotated within Xace during the construction of the Xq22-q23 transcript map described in Chapter 3, and examination of gene structures looking for similarities. A combination of BLASTN and TBLASTX (for more distantly related loci) was used to identify related genes in the region.

These data were collated and examined to determine which gene families were represented. Examination of the literature was performed to shed light on the potential functions of some of the genes.

For further sequence analyses (and primer design), genomic sequences were generated from the structures annotated where possible, as these were less likely than cDNA sources to contain sequencing errors that might affect subsequent analyses in instances where sequence similarity was very high.

In order to assess the relationships between genes within families, phylogenetic analyses were performed using sequence data from the loci. The intention was to use this information to make predictions about the evolution of the gene families and the region. Phylogenetic analyses using the coding nucleotide or peptide sequences were performed to determine the relationships of paralogous loci to each other. For distantly related loci, predicted peptide sequences were used to allow optimal alignment of sequences for phylogenetic analysis. For more closely related loci, coding nucleotide sequences were used as these would provide more information due to the increased rate of nucleotide substitutions compared to amino acid substitutions in coding sequences.

Protein or nucleotide sequences were aligned using ClustalX (see Chapter 2). This produced initial alignments that were manually inspected for accuracy. As all phylogenetic analyses assume that residues or nucleotides within a column of an alignment represent the same ancestral position, poorly aligned regions of the alignment were removed, and any columns within the alignments that contained gaps were deleted in an attempt to reduce violation of this assumption and improve the accuracy of subsequent analyses. Alignments edited in this manner were used for subsequent phylogenetic analysis.

The relationships between the sequences were assessed using both distance and maximum-likelihood phylogenetic analysis (see Chapter 2). Two different methods were used in order to improve confidence in the results, as ideally a consensus would be seen. For distance-based analyses of nucleotide sequences, the Kimura 2-parameter model of sequence evolution was employed, which attempts to take into account different transition and transversion rates (Kimura 1980). For protein sequences, a Poisson-corrected model of sequence evolution was used, which attempts to control for multiple substitutions occurring at a site during longer periods of evolution. Whilst still an approximation, if this possibility is not taken into account, the observed differences between sequences within an alignment may under-estimate the number of substitutions occurring within the respective sequences over time.

For both distance and maximum-likelihood analysis, bootstrap replicates were performed. This re-sampling technique randomly takes columns from the original alignment with replacement (i.e. when a column is sampled, it is also available for sampling in the next round), the same number as were in the original alignment, and performs the phylogenetic analysis each time on this “pseudo” alignment. It can then be assessed from the number of bootstrap replicates that were performed, how many times the same tree was produced as that seen in analysis of the original alignment. This provides a measure of confidence in the resulting tree.

As the DDC hypothesis regarding duplicated genes is the subject of debate at present, the expression patterns of many of the paralogues were assessed by RT-PCR using a panel of cDNAs generated from RNA from 20 different tissues. A rigorous approach to primer design was employed to minimise chances of generating "composite" expression patterns due to cross-hybridisation of primers to paralogous loci. Nucleotide sequence alignments of paralogous loci were generated and the exon boundary positions annotated. Results of primer-design for each sequence were then combined with information from the alignment to ensure that primers chosen differed in their 3' regions as much as possible. The 3' UTR regions were particularly targeted for primer design, as paralogous loci showed greatest divergence there.

Primer pairs that did not span exons 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 $T_{0.1}E$ (10 mM Tris-HCl (pH 8.0), 0.1 mM EDTA). Pre-screens were performed using three different primer annealing temperatures (55°C, 60°C and 65°C) to determine the one that gives a visible and specific DNA product (see Chapter 2).

Where possible, primer specificity was tested by performing colony PCR on genomic clones known to contain the specific locus, as well as others containing its paralogues. In several instances, however, this was not possible due to the close proximity of the paralogous loci. When this was the case, only primers designed to relatively divergent sequences were employed in the studies shown. Primers for four of the nine pp21-like family STSs and the NXF/TCP11-like STSs were not tested in this manner as they were designed subsequent to these experiments. The colony PCR protocol is described in Chapter 2. Briefly, genomic clones were grown on LB agar plates, then single colonies were picked into $T_{0.1}E$, and 5 μl of the resulting suspension were used in a PCR with the different STS primers. The results are presented in Table 5-1.

The data presented in Table 5-1 show that the primers designed to two of the most highly-related genes, the NB-thymosin beta paralogues, discriminated between the two loci. For other loci, the conclusions are indistinct as some loci were contained within a single clone (e.g. three of the GASP genes) or lie in adjacent clones which could overlap (as submitted sequence does not necessarily reflect the true insert size of the clone). For the Rab-like loci, no conclusion could be drawn regarding specificity for stSG158869 and stSG158870, as the identity of clone cU237H1 could not be verified (stSG158870, designed to a locus within cU237H1, failed to give a positive result for the clone). No expression data were generated for these loci as a result.

In some cases however, discrimination between paralogues was confirmed. In other instances, a degree of confidence in discrimination is obtained from the fact that the primers were designed to divergent regions of sequence. Comparative analyses in the mouse were facilitated by the generation of a BAC map of the orthologous region and annotation of the genomic sequence as described in Chapter 4.

The results of the analyses described for gene position relationships, orthology/paralogy in the mouse, expression patterns, sequence alignments and phylogenetic analysis are presented for each gene family in the following sections.

stSG No.	Locus	positive clone(s)	negative clone(s)	Family
158852	dJ77O19.CX.1	dJ77O19	cV362H12	thymosin-beta
158853	cV362H12.CX.1	cV362H12	dJ77019	thymosin-beta
158860	NGFRAP1	bB349O2O	dJ198P4/dJ79P11/dJ635G19/cV351F8	NADE
158855	dJ198P4.CX.1	dJ198P4	bB349O20/dJ79P11/dJ635G19/cV351F8	NADE
158856	dJ79P11.1	bB349020/dJ79P11	dJ198P4/dJ635G19/cV351F8	NADE
158857	dJ635G19.2	dJ79P11/dJ635G19	bB349O20/dJ198P4/cV351F8	NADE
158858	cV351F8.CX.2	cV351F8	bB349O20/dJ198P4/dJ793P1/dJ635G19	NADE
158865	dJ1100E15.CX.3	dJ1100E15	dJ769N13	GASP
158862	dJ769N13.1	dJ769N13	dJ1100E15	GASP
158866	dJ769N13.CX.1	dJ769N13	dJ1100E15	GASP
158864	dJ769N13.CX.2	dJ769N13	dJ1100E15	GASP
158907	cU209G1.CX.1	cU209G1	cU61B11/dJ454K15/bA269L6	ALEX
158908	cU61B11.CX.1	cU61B11	cU61B11/dJ454K15/bA269L6	ALEX
158909	dJ545K15.1	bA269L6	cU209G1/cU61B11/dJ454K15	ALEX
158910	dJ545K15.2	bA269L6	cU209G1/cU61B11/dJ454K15	ALEX
158911	cV602D8.CX.1	bA269L6	cU209G1/cU61B11/dJ454K15	ALEX
158870	cU237H1.1		cU237H1/cU250H12	Rab-like
158869	cU250H12.CX.1	cU250H12	cU237H1	Rab-like
158871	cU237H1.1	cU250H12	cU237H1	Rab-like
158900	dJ122O23.CX.1	dJ122O23	cU177E8/cV857G6/dJ1055C14/cU105G4/cV351F8	pp21-like
158901	cV351F8.CX.1	dJ122023/cV351F8	cU177E8/cV857G6/dJ1055C14/cU105G4	pp21-like
158904	cV857G6.CX.1	cV857G6	dJ122O23/cU177E8/dJ1055C14/cU105G4/cV351F8	pp21-like
158913	cU105G4.1.1	cU177E8/cV857G6/dJ1055C14	dJ122O23/cU105G4/cV351F8	pp21-like
158914	TCEAL1	dJ1055C14	dJ122O23/cU177E8/cV857G6/cU105G4/cV351F8	pp21-like

Table 5-1 Results of colony screens using expression profiling STS primers against genomic clones from the Xq22 tiling path. The STS (stSG number given), locus for which it was designed, and clones positive or negative for the STS are given. Clone names in red indicate clones whose identity could not be verified from the STS data. The gene families for the loci are also given.

5.2.1 *Thymosin-beta genes*

The mouse beta-thymosin Tmsb4x (Ptmb4) gene was cloned and it was reported that a single locus existed from Southern Blot analysis (Li *et al.*, 1996). However, the results presented here demonstrate the presence of an additional three homologues of Tmsb4x in the mouse genome. The Tmsb4x gene was shown from mapping in *Mus spretus* to be linked to Btk, and human TMSB4X is reported in locuslink (NCBI) to lie in Xq22 (containing BTK). However, as shown below, the presence of paralogues appears to have misled location assignments.

In LocusLink (NCBI), TMSB4X is erroneously mapped to Xq21.3-q22, and in Ensembl 17.33.1 to HSA4. BLAST analysis against the NCBI nr database using the coding sequence showed a hit to a chromosome 4 BAC, RP11-309H6, with one mismatch, as well as other BACs from different chromosomes. Using the HTGS data however, a perfect hit was seen to BAC RP11-102M2 (accession AC023098) which is mapped distal to RAB9 and GPM6B and just proximal to PRPS2 on Xp22.22.

The mouse orthologue of TMSB4X, Ptmb4 (MGD), is also mapped to a region (X F5) equivalent to Xp22. This example illustrates the difficulties in mapping highly related sequences, and the benefits of the availability of extensive genomic sequence information in the elucidation of gene location and orthology.

Whilst human Xq22 contains two TMSNB paralogues, the orthologous mouse region contains three. This was perhaps surprising, as the high level of similarity seen between the human Xq22 TMSNB genes, even within the introns, suggested a recent duplication event. It appears that as the two human genes and bM250F8.MX.4 and the two genes in bM389M3 are similarly separated, a gene duplication may have occurred prior to the human-mouse divergence, with a subsequent extra duplication in the mouse (there are also shared orthologues between human and mouse at the locations represented by bM250F8 and bM389M3 (see Chapter 4). An alternative is that an extra duplication took place in a common ancestor, followed by loss from human, but this is a less parsimonious explanation). Gene conversion may then be responsible for maintenance of sequence homology between loci. An alternative explanation would invoke independent duplications within each species.

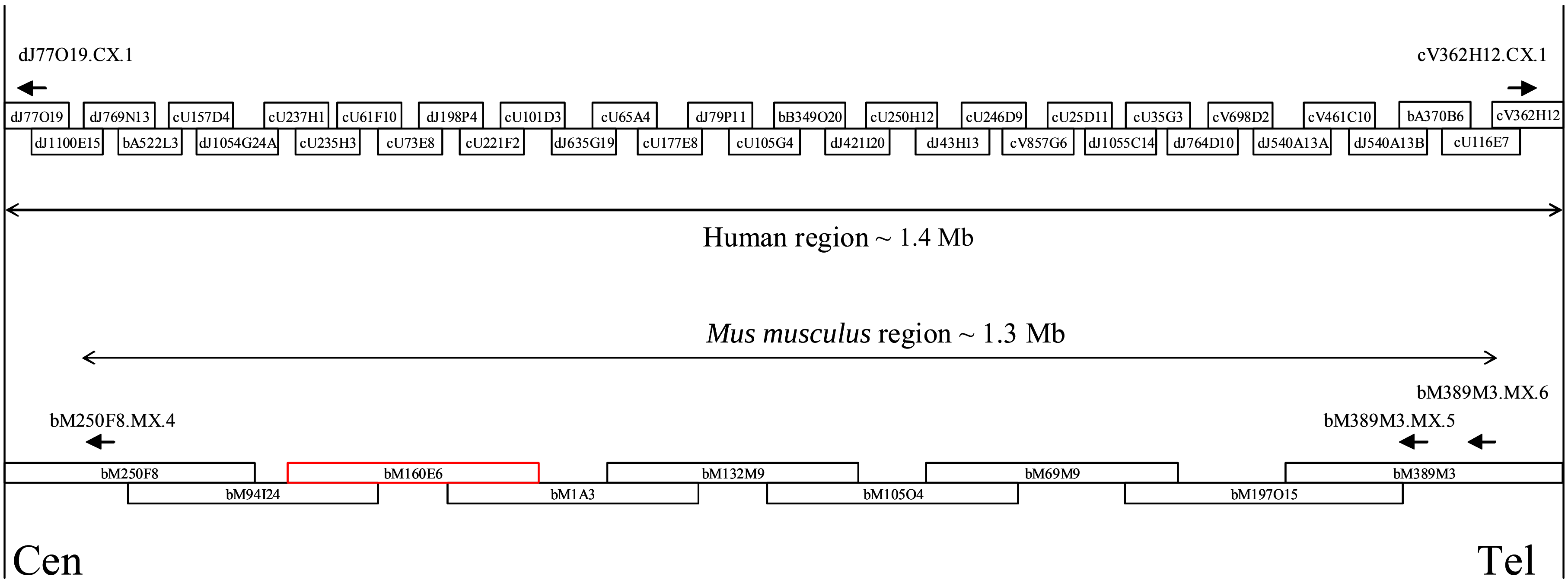
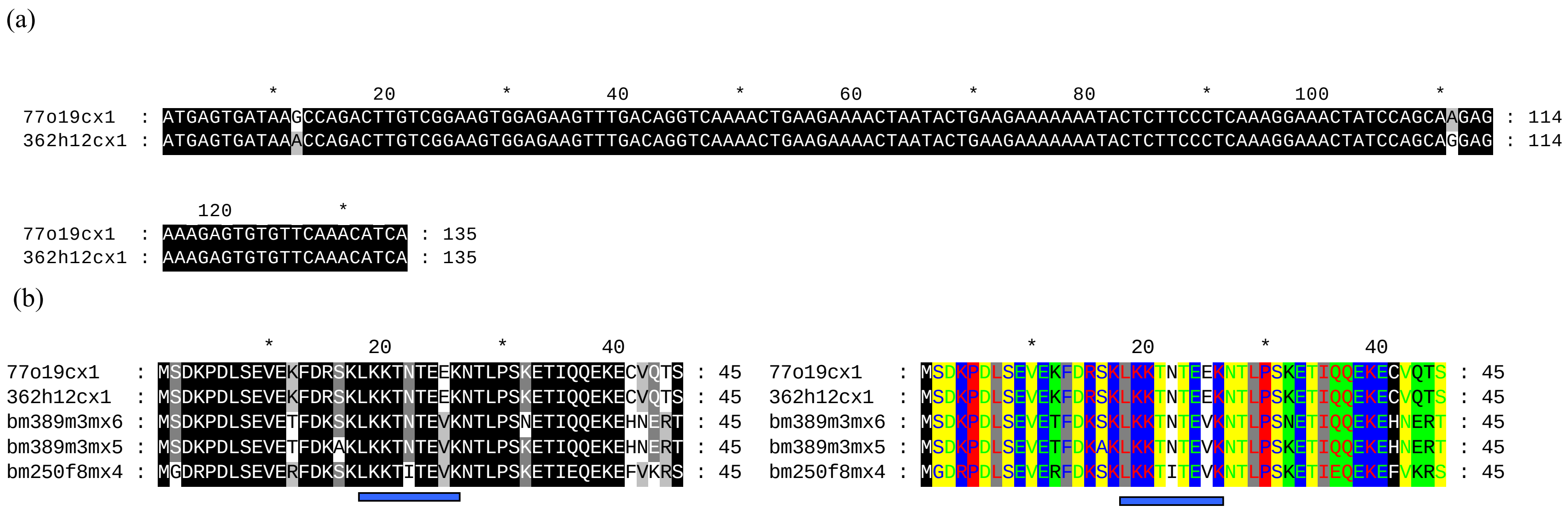


Figure 5-2 Figure showing a schematic representation (not to scale) of thymosin-beta paralogue gene order and orientation within human Xq22, and the corresponding orthologous region in *Mus musculus* (mouse). Large double-headed arrows depict approximate sizes of the regions, small arrows depict gene positions (names above) and direction of transcription. The genomic clone sequence-ready tile-paths for human and mouse are represented with clone names (not to scale). Clone in red indicates unfinished sequence, and hence a gap in annotation.



(a)

```
                        *        20         *        40         *        60         *        80         *       100         *
77o19cx1  : ATGAGTGATAAGCCAGACTTGTCGGAAGTGGAGAAGTTTGACAGGTCAAAACTGAAGAAAACTAATACTGAAGAAAAAAATACTCTTCCCTCAAAGGAAACTATCCAGCAAGAG : 114
362h12cx1 : ATGAGTGATAAACCAGACTTGTCGGAAGTGGAGAAGTTTGACAGGTCAAAACTGAAGAAAACTAATACTGAAGAAAAAAATACTCTTCCCTCAAAGGAAACTATCCAGCAGGAG : 114

                 120         *
77o19cx1  : AAAGAGTGTGTTCAAACATCA : 135
362h12cx1 : AAAGAGTGTGTTCAAACATCA : 135
```

(b)

```
                        *        20         *        40                                  *        20         *        40
77o19cx1   : MSDKPDLSEVEKFDRSKLKKTNTEEKNTLPSKETIQQEKECVQTS : 45   77o19cx1   : MSDKPDLSEVEKFDRSKLKKTNTEEKNTLPSKETIQQEKECVQTS : 45
362h12cx1  : MSDKPDLSEVEKFDRSKLKKTNTEEKNTLPSKETIQQEKECVQTS : 45   362h12cx1  : MSDKPDLSEVEKFDRSKLKKTNTEEKNTLPSKETIQQEKECVQTS : 45
bm389m3mx6 : MSDKPDLSEVETFDKSKLKKTNTEVKNTLPSNETIQQEKEHNERT : 45   bm389m3mx6 : MSDKPDLSEVETFDKSKLKKTNTEVKNTLPSNETIQQEKEHNERT : 45
bm389m3mx5 : MSDKPDLSEVETFDKAKLKKTNTEVKNTLPSKETIQQEKEHNERT : 45   bm389m3mx5 : MSDKPDLSEVETFDKAKLKKTNTEVKNTLPSKETIQQEKEHNERT : 45
bm250f8mx4 : MGDRPDLSEVERFDKSKLKKTITEVKNTLPSKETIEQEKEFVKRS : 45   bm250f8mx4 : MGDRPDLSEVERFDKSKLKKTITEVKNTLPSKETIEQEKEFVKRS : 45
```

Figure 5-3 Figure illustrating alignments of thymosin-beta paralogues. (a) alignment of the coding sequences of human genes dJ77019.CX.1 (labelled 77o19cx1) and cV362H12.CX.1 (labelled 362h12cx1). (b) two versions of an alignment of the predicted peptides from the two human thymosin-beta like genes and the three homologous mouse genes bM250F8.MX.4 (labelled bm250f8mx4), bM389M3.MX.5 (labelled bm389m3mx5) and bM389M3.MX.6 (labelled bm389m3mx6). One alignment (left) is shaded to illustrate residue conservation, the other (right) is shaded to illustrate conservation of physiochemical properties of residues. Blue bars represent actin-binding regions.

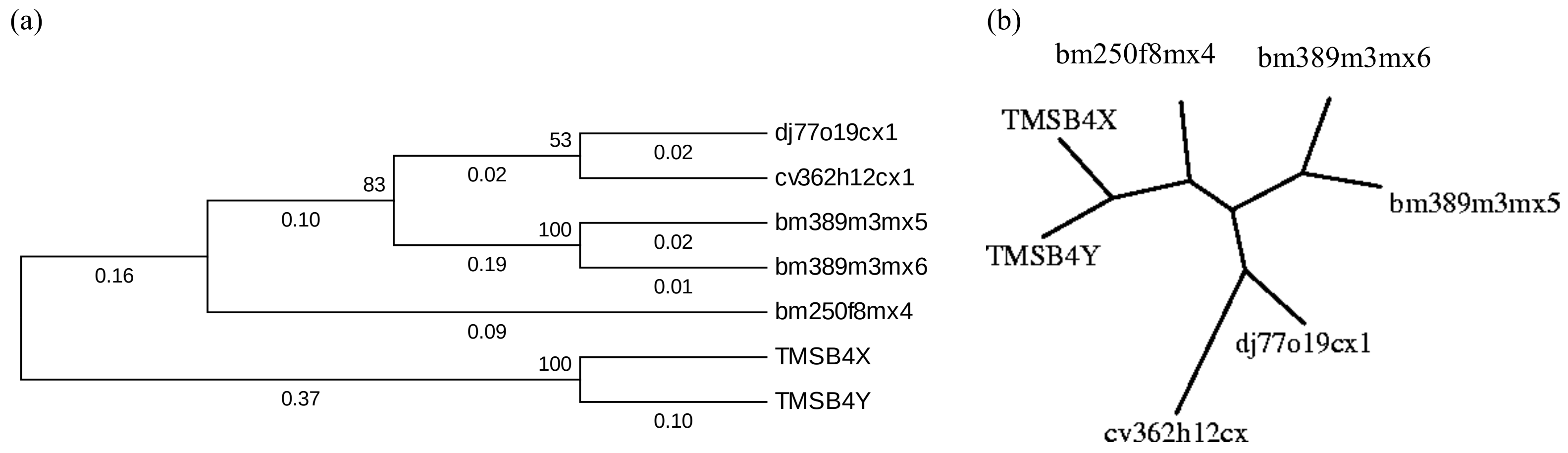
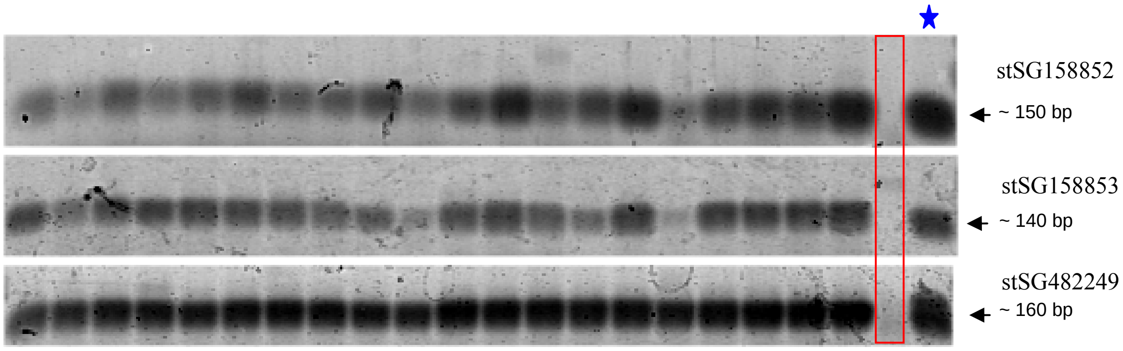
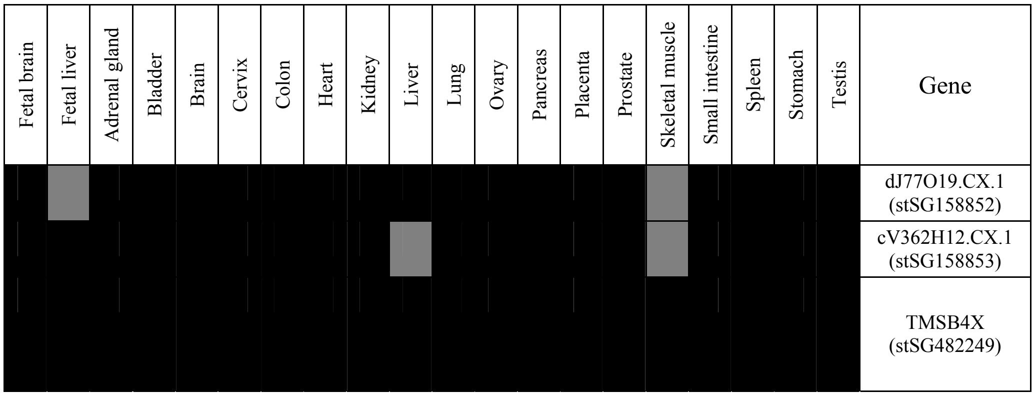


Figure 5-4 Phylogenetic analysis of human and mouse thymosin-beta genes. (a) distance-based cladogram computed from human and mouse gene coding nucleotide sequence. Numbers adjacent to nodes represent the number of bootstrap replicates supporting the adjacent node. Numbers below branches indicate distance. For clarity, only the tree topology is shown and branches are not scaled. (b) maximum-likelihood analysis performed using the same data as in (a). For technical reasons, gene names have had "." separators removed in the figures.

(a)



(b)



Fetal brain	Fetal liver	Adrenal gland	Bladder	Brain	Cervix	Colon	Heart	Kidney	Liver	Lung	Ovary	Pancreas	Placenta	Prostate	Skeletal muscle	Small intestine	Spleen	Stomach	Testis	Gene
■	▨	■	■	■	■	■	■	■	■	■	■	■	■	■	▨	■	■	■	■	dJ77O19.CX.1 (stSG158852)
■	■	■	■	■	■	■	■	■	▨	■	■	■	■	■	▨	■	■	■	■	cV362H12.CX.1 (stSG158853)
■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	TMSB4X (stSG482249)

Figure 5-5 RT-PCR expression profiling of human thymosin-beta family genes. (a) Vistra green stained gel images of RT-PCR products for primers designed to thymosin-beta family genes, after 35 cycles of PCR. Experiments performed using 30 and 40 cycles gave similar results. Approximate band sizes are arrowed. The red box in the gel images denotes the negative control lane. The lane denoted with a blue asterisk denotes the positive control lane. (b) a schematic representation of the data from figure (a). Black filled cells denote medium to strong PCR product bands detected, and grey cells denote weak product bands detected.

As seen in Figure 5-3 (a), the human Xq22 sequences are very closely related, with only two synonymous differences seen. In contrast, the mouse predicted peptides all differ slightly from each other (seen in Figure 5-3 (b)). These changes appear functionally constrained however, as the amino acids mainly share conserved physiochemical properties.

Both phylogenetic analyses suggest that the human Xq22 thymosin-beta gene sequences are more related to each another than to their mouse counterparts. This is supported by manual inspection of the peptide alignments shown in Figure 5-3, and from the extensive sequence similarity seen between introns of the two human Xq22 thymosin-beta genes (data not shown). TMSB4X and TMSB4Y also appear to be more closely related to one another than to either of the Xq22 paralogues.

The involvement of TMSB4X in an additional duplication event involving Xq22 (see Chapter 6) suggests that TMSB4X and cV362H12.CX.1 shared a common ancestor (based on gene order in relation to other paralogous genes). Based on the phylogenetic and gene order data, the following hypothesis can be suggested: following an initial duplication producing TMSB4X and cV362H12.CX.1, a further duplication of cV362H12.CX.1 produced dJ77O19.CX.1 prior to divergence of mouse and human. Subsequent to mouse-human divergence, the mouse bM389M3 locus underwent a further duplication, and gene conversion is acting on the human Xq22 loci. Subsequent to the marsupial-human divergence but prior to the eutherian radiation, the region containing TMSB4X was translocated to the ancestral X and Y PAR, and became TMSB4X and TMSB4Y.

The thymosin beta predicted peptides are highly conserved, and although the mouse peptides appear more divergent many of the differences still maintain similar physiochemical properties. No discernible differences were seen in expression patterns for TMSB4X and the two Xq22 genes, and all appeared to be ubiquitously expressed. As the predicted peptides of the Xq22 genes are identical, functional selection at the protein level would not discriminate between the loci. A more exhaustive approach would be required to determine potential temporal and spatial differences in expression pattern, as predicted by the DDC hypothesis.

5.2.2 NADE family genes

Within human Xq22 there are five genes belonging to the NADE family. In the mouse, four NADE-like genes have been annotated (see Chapter 4) but a fifth gene could reside in the sequence gap illustrated in Figure 5-6. The orientations of the four annotated mouse NADE genes reflect those of the four most distal Xq22 genes. The prediction of orthology would appear straightforward from this information, and from neighbouring genes (see Chapter 4). The phylogenetic analyses support the prediction of orthology for NGFRAP1/bM105O4.MX.4 and dJ635G19.2/bM1A3.MX.7; but, as was seen for the TMSNB genes, the other two mouse genes appear more closely related to each other than to the human genes dJ198P4.CX.1 and dJ79P11.1, and *vice versa*.

Both phylogenetic analyses suggest that two of the NADE-like genes are more closely related within the species than between the species. This observation is also supported by manual inspection of alignments of the respective sequence. These human and mouse genes share similar positioning and transcription directions, and so again it may be indicative of gene conversion events causing sequence homogenisation rather than independent gene duplications.

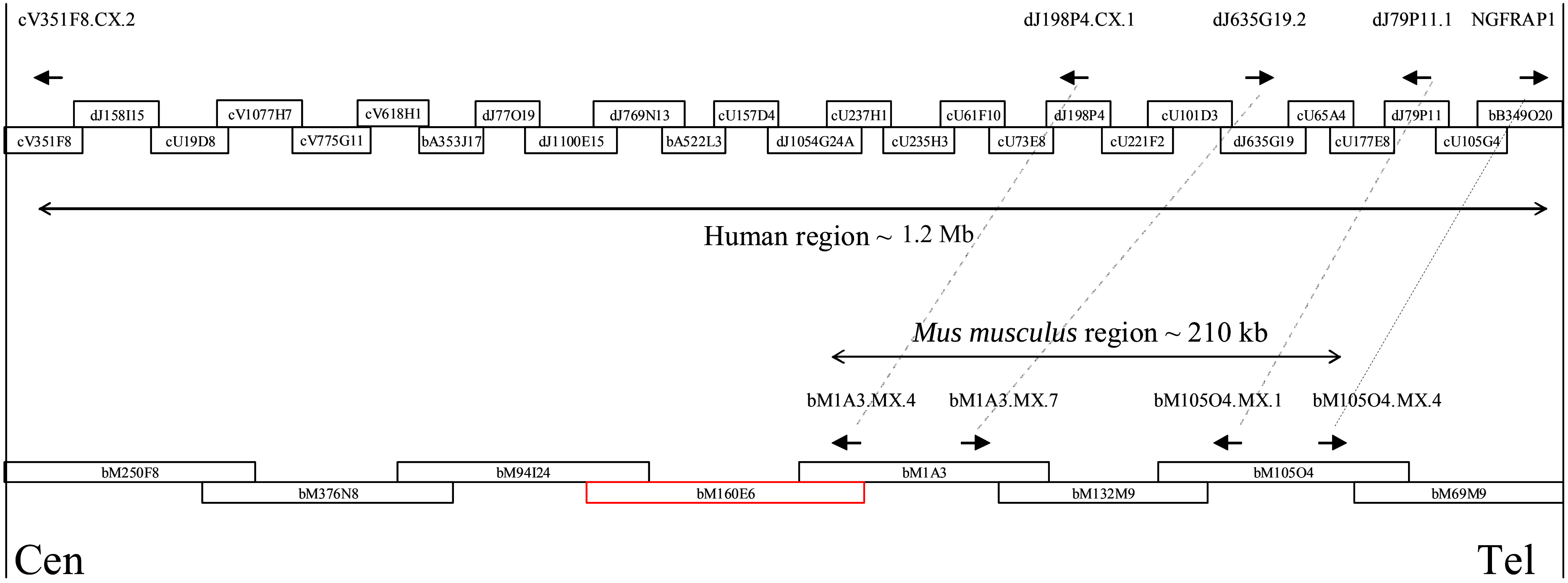
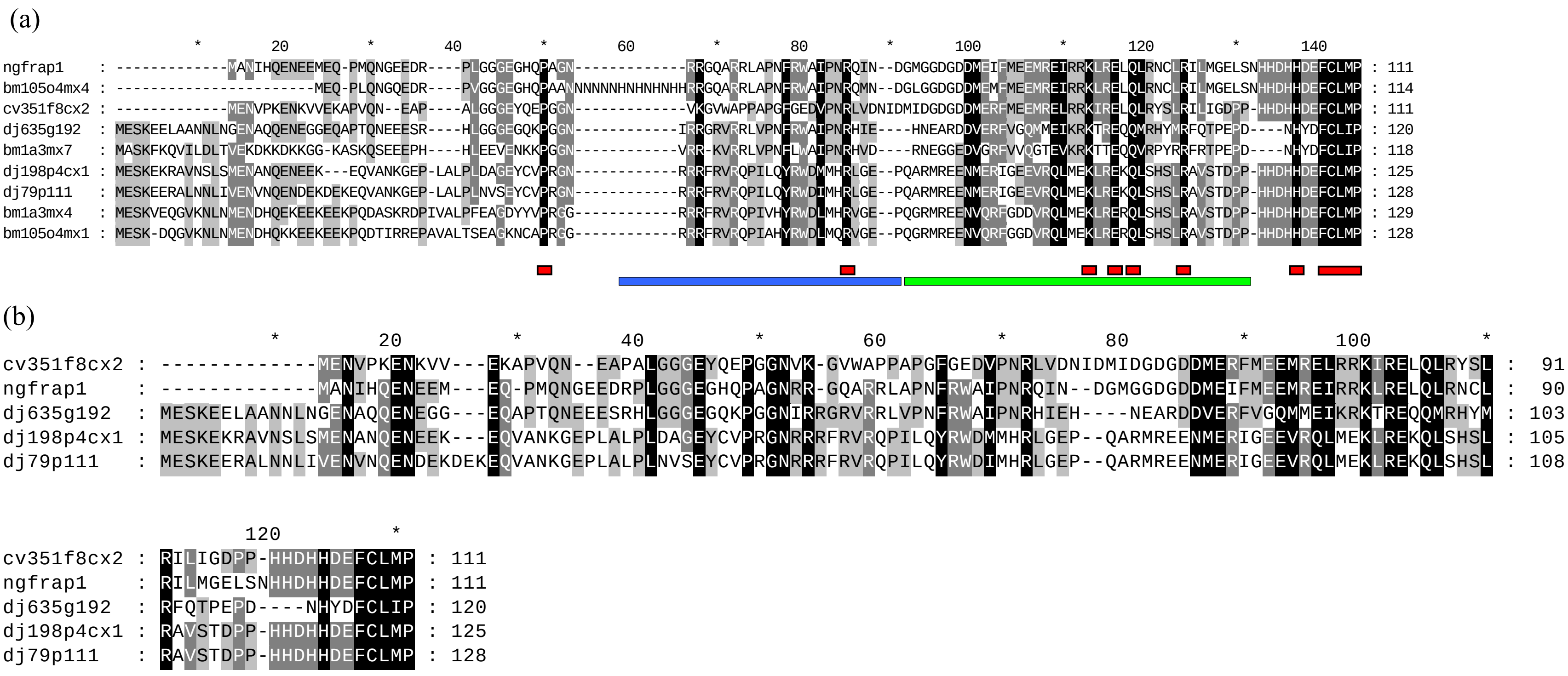


Figure 5-6 A schematic representation (not to scale) of NADE family gene order and orientation within human Xq22 and the corresponding orthologous region in mouse. Large double-headed arrows depict approximate sizes of the regions, small arrows depict gene positions (names above) and direction of transcription. The genomic clone sequence-ready tile-paths for human and mouse are represented with clone names. Clone in red indicates unfinished sequence, and hence a gap in the annotation. Grey dotted lines illustrate likely orthologous relationships.



(a)

```
                          *        20         *        40         *        60         *        80         *       100         *       120         *       140
ngfrap1    : -------------MANIHQENEEMEQ-PMQNGEEDR----PLGGGEGHQPAGN------------RRGQARRLAPNFRWAIPNRQIN--DGMGGDGDDMEIFMEEMREIRRKLRELQLRNCLRILMGELSNHHDHHDEFCLMP : 111
bm105o4mx4 : ----------------------MEQ-PLQNGQEDR----PVGGGEGHQPAANNNNNNHNHNHNHHRRGQARRLAPNFRWAIPNRQMN--DGLGGDGDDMEMFMEEMREIRRKLRELQLRNCLRILMGELSNHHDHHDEFCLMP : 114
cv351f8cx2 : -------------MENVPKENKVVEKAPVQN--EAP----ALGGGEYQEPGGN------------VKGVWAPPAPGFGEDVPNRLVDNIDMIDGDGDDMERFMEEMRELRRKIRELQLRYSLRILIGDPP-HHDHHDEFCLMP : 111
dj635g192  : MESKEELAANNLNGENAQQENEGGEQAPTQNEEESR----HLGGGEGQKPGGN-----------IRRGRVRRLVPNFRWAIPNRHIE----HNEARDDVERFVGQMMEIKRKTREQQMRHYMRFQTPEPD----NHYDFCLIP : 120
bm1a3mx7   : MASKFKQVILDLTVEKDKKDKKGG-KASKQSEEEPH----HLEEVENKKPGGN-----------VRR-KVRRLVPNFLWAIPNRHVD----RNEGGEDVGRFVVQGTEVKRKTTEQQVRPYRRFRTPEPD----NHYDFCLIP : 118
dj198p4cx1 : MESKEKRAVNSLSMENANQENEEK---EQVANKGEP-LALPLDAGEYCVPRGN-----------RRRFRVRQPILQYRWDMMHRLGE--PQARMREENMERIGEEVRQLMEKLREKQLSHSLRAVSTDPP-HHDHHDEFCLMP : 125
dj79p111   : MESKEERALNNLIVENVNQENDEKDEKEQVANKGEP-LALPLNVSEYCVPRGN-----------RRRFRVRQPILQYRWDIMHRLGE--PQARMREENMERIGEEVRQLMEKLREKQLSHSLRAVSTDPP-HHDHHDEFCLMP : 128
bm1a3mx4   : MESKVEQGVKNLNMENDHQEKEEKEEKPQDASKRDPIVALPFEAGDYYVPRGG-----------RRRFRVRQPIVHYRWDLMHRVGE--PQGRMREENVQRFGDDVRQLMEKLRERQLSHSLRAVSTDPP-HHDHHDEFCLMP : 129
bm105o4mx1 : MESK-DQGVKNLNMENDHQKKEEKEEKPQDTIRREPAVALTSEAGKNCAPRGG-----------RRRFRVRQPIAHYRWDLMQRVGE--PQGRMREENVQRFGGDVRQLMEKLRERQLSHSLRAVSTDPP-HHDHHDEFCLMP : 128
```

(b)

```
                         *        20         *        40         *        60         *        80         *       100         *
cv351f8cx2 : -------------MENVPKENKVV---EKAPVQN--EAPALGGGEYQEPGGNVK-GVWAPPAPGFGEDVPNRLVDNIDMIDGDGDDMERFMEEMRELRRKIRELQLRYSL :  91
ngfrap1    : -------------MANIHQENEEM---EQ-PMQNGEEDRPLGGGEGHQPAGNRR-GQARRLAPNFRWAIPNRQIN--DGMGGDGDDMEIFMEEMREIRRKLRELQLRNCL :  90
dj635g192  : MESKEELAANNLNGENAQQENEGG---EQAPTQNEEESRHLGGGEGQKPGGNIRRGRVRRLVPNFRWAIPNRHIEH----NEARDDVERFVGQMMEIKRKTREQQMRHYM : 103
dj198p4cx1 : MESKEKRAVNSLSMENANQENEEK---EQVANKGEPLALPLDAGEYCVPRGNRRRFRVRQPILQYRWDMMHRLGEP--QARMREENMERIGEEVRQLMEKLREKQLSHSL : 105
dj79p111   : MESKEERALNNLIVENVNQENDEKDEKEQVANKGEPLALPLNVSEYCVPRGNRRRFRVRQPILQYRWDIMHRLGEP--QARMREENMERIGEEVRQLMEKLREKQLSHSL : 108

                  120        *
cv351f8cx2 : RILIGDPP-HHDHHDEFCLMP : 111
ngfrap1    : RILMGELSNHHDHHDEFCLMP : 111
dj635g192  : RFQTPEPD----NHYDFCLIP : 120
dj198p4cx1 : RAVSTDPP-HHDHHDEFCLMP : 125
dj79p111   : RAVSTDPP-HHDHHDEFCLMP : 128
```

Figure 5-7 Alignments of NADE family predicted peptides. (a) alignment of human and mouse sequences. Red boxes indicate invariant residues. The blue bar represents a pro-apoptotic region and the green bar a regulatory region (Mukai *et al.*, 2002) (b) alignment of only the human sequences. For technical reasons, gene names have had "." separators removed in the figures.

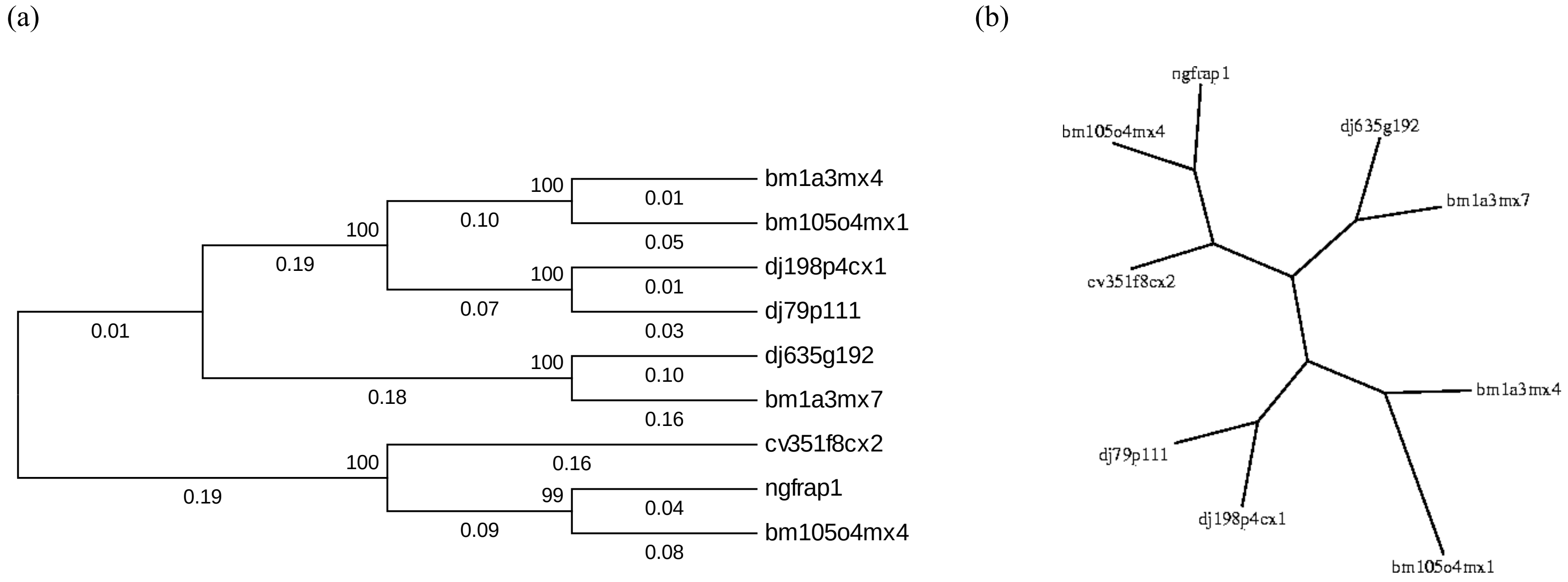
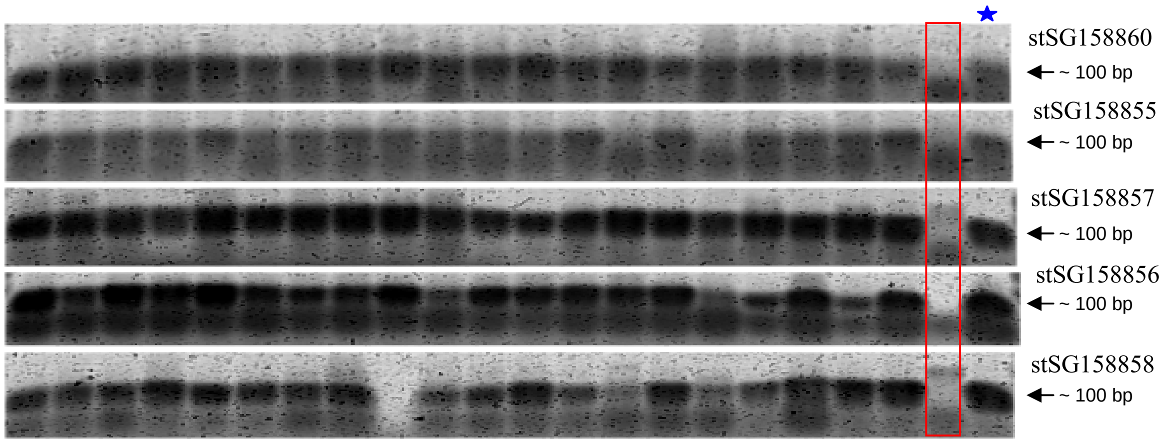
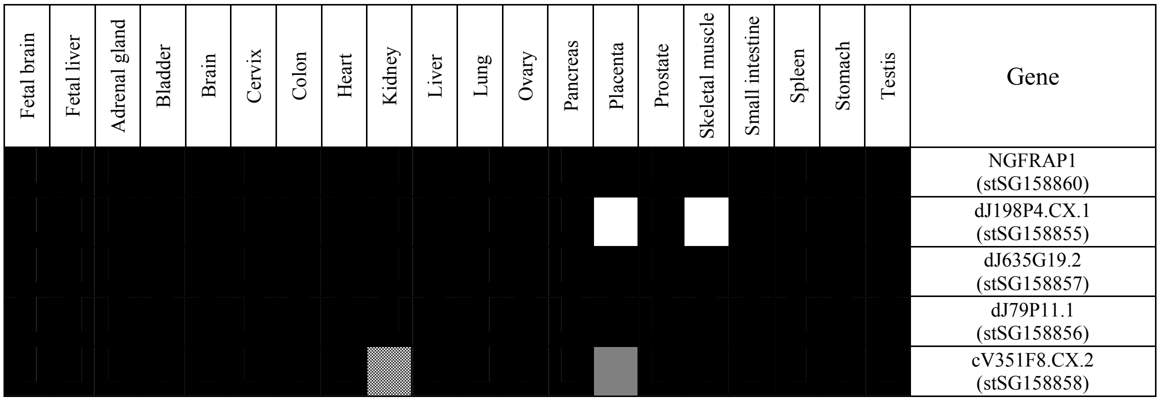


Figure 5-8 Phylogenetic analysis of NADE Xq22 paralogues. (a) a distance-based cladogram constructed using an alignment of human and mouse NADE-like genes coding nucleotide sequences. Numbers adjacent to nodes represent the number of bootstrap replicates supporting the adjacent node. Numbers below branches indicate distance. For clarity, only the tree topology is shown and branches are not scaled. (b) maximum-likelihood analysis of the same data used for (a). For technical reasons, gene names have had "." separators removed in the figures.

(a)



(b)



Fetal brain	Fetal liver	Adrenal gland	Bladder	Brain	Cervix	Colon	Heart	Kidney	Liver	Lung	Ovary	Pancreas	Placenta	Prostate	Skeletal muscle	Small intestine	Spleen	Stomach	Testis	Gene
■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	NGFRAP1 (stSG158860)
■	■	■	■	■	■	■	■	■	■	■	■	■	□	■	□	■	■	■	■	dJ198P4.CX.1 (stSG158855)
■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	dJ635G19.2 (stSG158857)
■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	dJ79P11.1 (stSG158856)
■	■	■	■	■	■	■	■	▨	■	■	■	■	▒	■	■	■	■	■	■	cV351F8.CX.2 (stSG158858)

Figure 5-9 RT-PCR expression profiles of NADE family genes. Legend as for Figure 5-5. Hatched cells denote uninformative assay points (where a RT-PCR reaction was omitted), prohibiting conclusions regarding expression in that tissue.

Apparent from Figure 5-7 (a) is the strong conservation of the "CLMP" motif at the carboxyl terminus. This corresponds to a known prenylation motif (Prosite PS00294) and suggests that this region may be important for prenylation of the NADE-like genes. Also shown in Figure 5-7 (a) by red boxes are several residues that appear to be invariant and may be important for function. The carboxyl terminal region also appears histidine-rich. The blue bar in Figure 5-7 (a) represents a pro-apoptotic region, and the green bar is a regulatory region including the NES and p75NTR-binding regions.

There appear to be several key conserved residues shared by the human and mouse NADE family members. Studies on mouse NADE/Bex3 (bM105O4.MX.4) have demonstrated which regions of the protein appear to be responsible for different functions. These have been annotated on Figure 5-7. This information can be transferred to other members of the NADE family due to the level of homology seen in these regions, and experiments designed to test the validity of the predicted functions. The authors reporting structure/function studies of mouse NADE also report that NADE undergoes self-association (Mukai *et al.*, 2002). This raises the possibility that different members of the NADE family may form hetero-dimers, potentially altering their functional properties.

The expression data indicate that the NADE family genes are widely expressed, and the only gene that showed a slight difference in expression pattern was dJ198P4.CX.1, for which no transcript was detected in placenta or skeletal muscle. The STS for this gene was demonstrated to be specific by colony PCR (see Table 5-1). As for the TMSNB genes, more exhaustive analyses may reveal subtleties in expression patterns not revealed at this gross level. Alternatively, given the probable role of these proteins in mediating apoptosis via protein-protein interaction, slight differences in affinities between the different family members may confer selective advantages sufficient to drive retention of paralogues, and as such there may be no need or less drive to diverge in expression pattern.

5.2.3 NXF family and TCP11-like genes

Five NXF-family genes were annotated within the human Xq22 region (the sixth member, NXF1, is mapped to 11q12-q13), and two NXF-like genes were annotated within the orthologous region in mouse. Three TCP11-like loci were annotated in human Xq22, of which one appears to be a pseudogene (dJ158I15.1), and one TCP11-like locus was annotated within the orthologous region in mouse. It is possible that two further NXF-family loci may reside in the sequence gaps in the mouse region. The presence of only one TCP11-like locus in mouse is consistent with the apparently recent NXF2/TCP11-like duplication in the human lineage (see Chapter 3) subsequent to human/mouse divergence; this may have been accompanied by a further duplication in human Xq22 generating dJ158I15.1.

Both phylogenetic analyses for the NXF genes largely agree on the topology of the gene relationships, and agree with relationships described by Herold *et al.*, (Herold *et al.*, 2000). NXF1, an autosomal NXF family gene (11q12-q13), appears to be more related to NXF2, NXF5 and dJ1100E15.2 (NXF4) than to NXF3. As expected, NXF2 and bA353J17.1 (NXF2a) cluster together, as their nucleotide sequences are almost identical, differing by only 1bp in ~2.3 kb.

The phylogenetic analyses for the TCP11-like genes suggest that the human Xq22 sequences are more related to one another than to their autosomal paralogue, TCP11 (6p21).

The gene positional and orientation information presented here suggest that bM250F8.MX.1 and bM250F8.MX.3 are the mouse orthologues of bA353J17.1 (NXF2a) and bA353J17.2 respectively, as their orientations are the same and no additional TCP11 pseudogene is noted. This would suggest that bA353J17.1 represents the ancestral NXF2 locus. Similarly, bM1A3.MX.5 would appear to be orthologous to NXF3. This is supported by the phylogenetic analyses (see Figure 5-13), which also indicate that NXF5 and dJ1100E15.2 (NXF4) may be more related to each other and the NXF2 loci than NXF2 is to its potential mouse orthologue. Completion of sequencing of the region will confirm whether mouse lacks orthologues of NXF5, NXF4 and possibly NXF2 as suggested here.

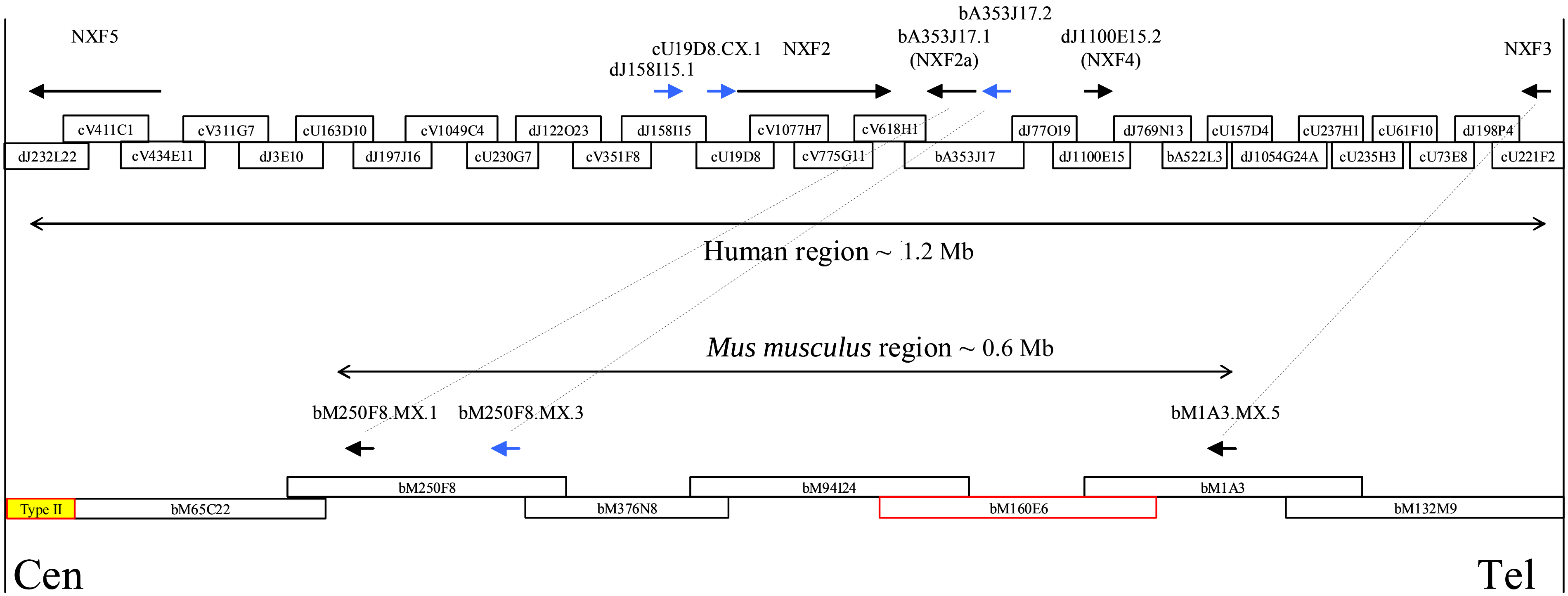
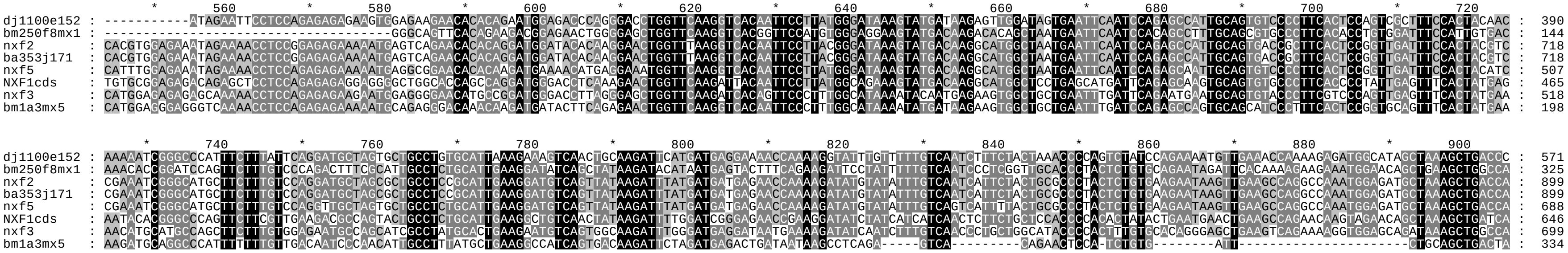


Figure 5-10 Figure showing a schematic representation (not to scale) of NXF and TCP11-like paralogue gene order and orientation within human Xq22, and the corresponding orthologous region in mouse. Gene names reflect locus names in Chapter 3 and 4, with alternative names given in parentheses. Large double-headed arrows depict approximate sizes of the regions, small arrows depict gene positions (names above) and direction of transcription. The genomic clone sequence-ready tile-paths for human and mouse are represented with clone names. Clone in red indicates unfinished sequence, and hence a gap in annotation. Blue arrows represent TCP11-like genes/pseudogenes, black arrows represent NXF family genes. Grey dotted lines illustrate likely orthologous relationships. The yellow box represents a Type II gap (region of the contig where there is no sequence coverage, but clones span the region).

(a)



(b)

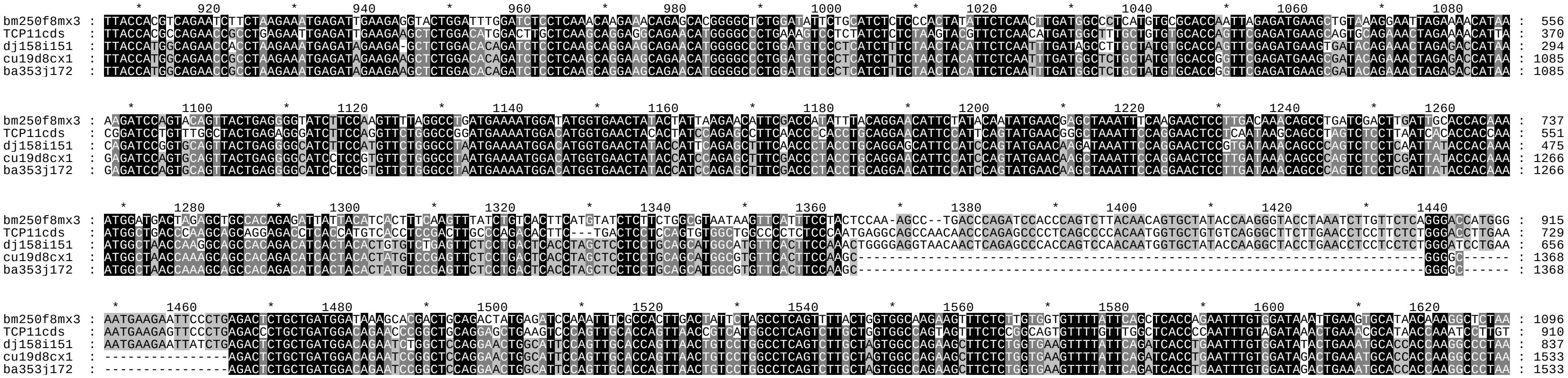


Figure 5-11 Alignments of NXF family genes and TCP11-like genes. (a) depicts part of an alignment of NXF-family nucleotide sequences to illustrate the level of homology seen. (b) depicts part of an alignment of TCP11-like nucleotide sequences. For technical reasons, gene names have had "." separators removed in the figures.

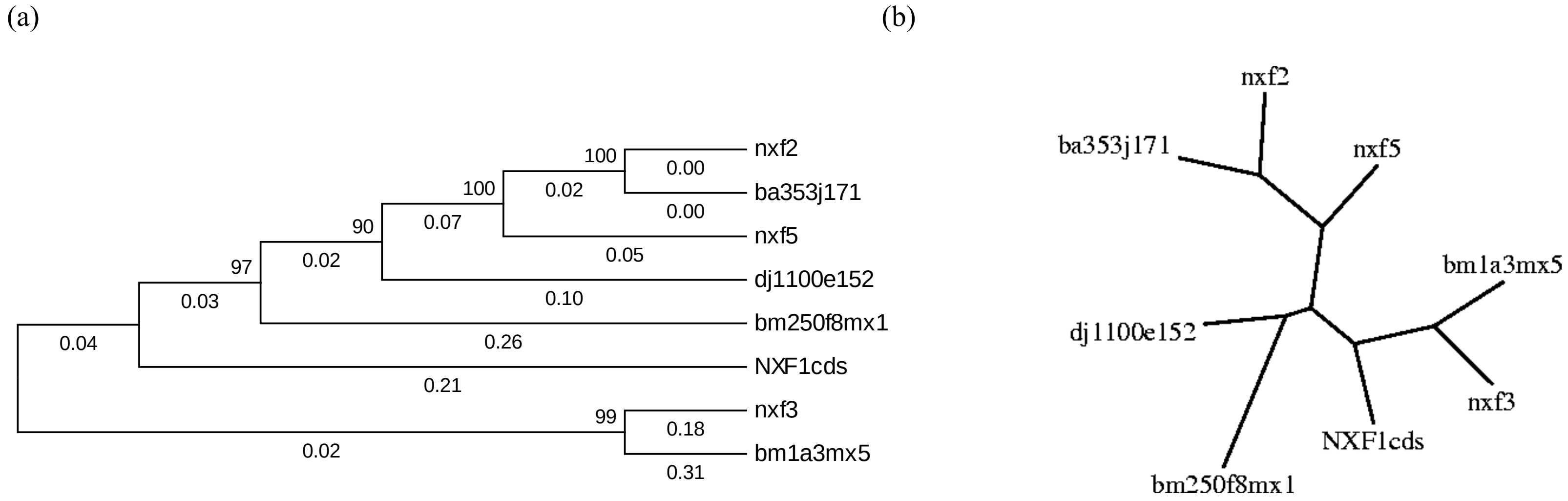


Figure 5-12 Figure illustrating phylogenetic analysis of human and mouse NXF genes. (a) a distance-based cladogram constructed using an alignment of human and mouse NXF-like genes coding nucleotide sequence. Numbers adjacent to nodes represent the number of bootstrap replicates supporting the adjacent node. Numbers below branches indicate distance. For clarity, only the tree topology is shown and branches are not scaled. (b) maximum-likelihood analysis of the same data used for (a). For technical reasons, gene names have had "." separators removed in the figures.

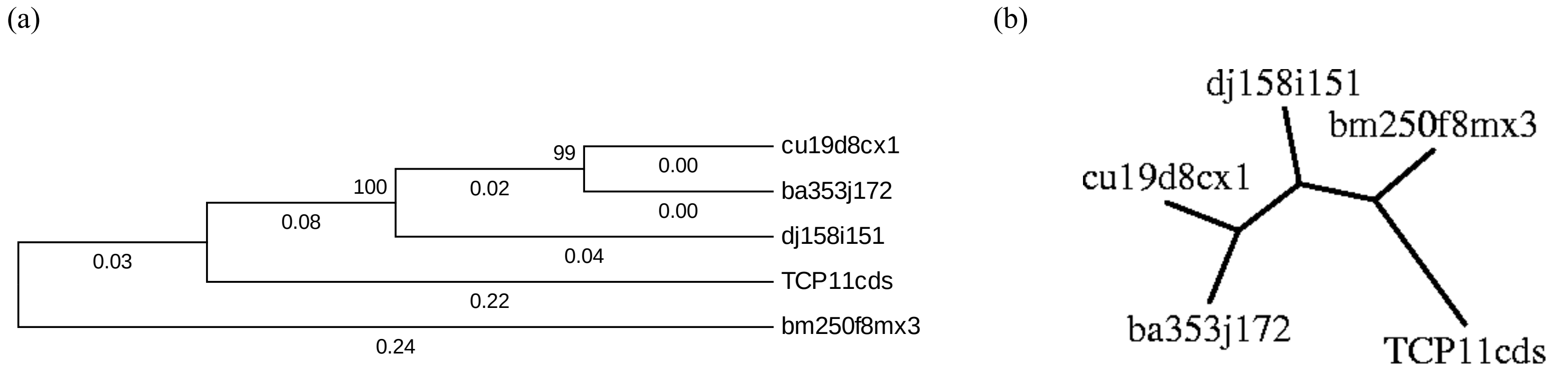
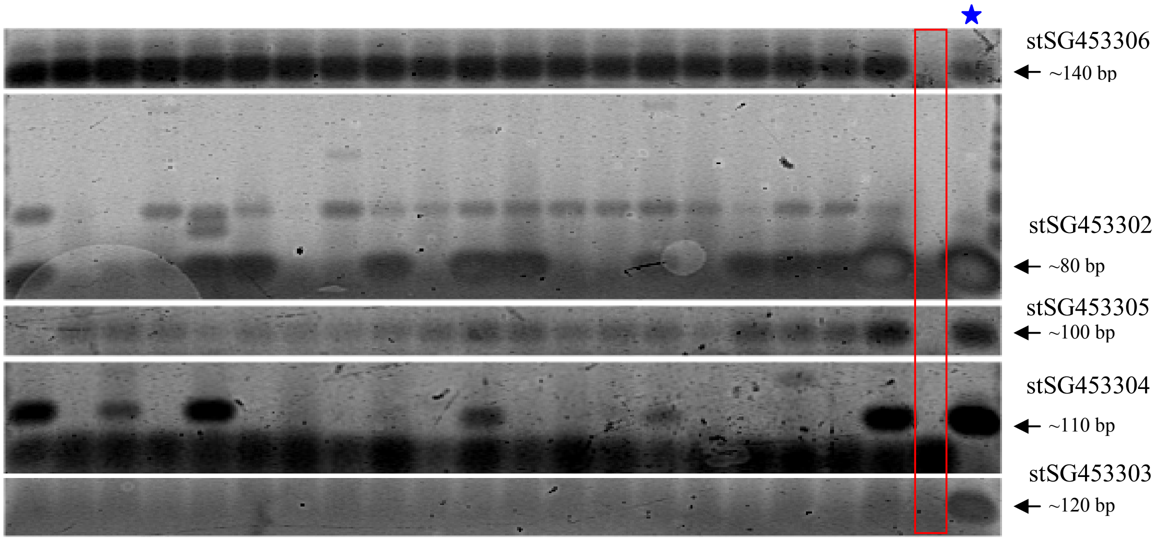
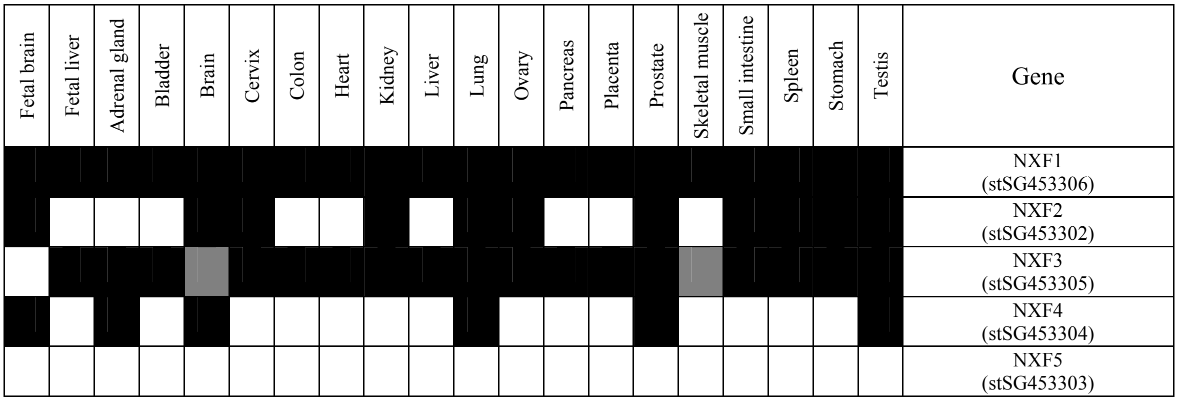


Figure 5-13 Figure illustrating phylogenetic analysis of human and mouse TCP11-like genes (a) a distance-based cladogram constructed using an alignment of human and mouse TCP-like genes coding nucleotide sequence. Numbers adjacent to nodes represent the number of bootstrap replicates supporting the adjacent node. Numbers below branches indicate distance. For clarity, only the tree topology is shown and branches are not scaled. (b) maximum-likelihood analysis of the same data used for (a). For technical reasons, gene names have had "." separators removed in the figures.

(a)



(b)



Fetal brain	Fetal liver	Adrenal gland	Bladder	Brain	Cervix	Colon	Heart	Kidney	Liver	Lung	Ovary	Pancreas	Placenta	Prostate	Skeletal muscle	Small intestine	Spleen	Stomach	Testis	Gene
■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	NXF1 (stSG453306)
■	□	□	□	■	■	□	□	■	□	■	■	□	□	■	□	■	■	■	■	NXF2 (stSG453302)
□	■	■	■	▨	■	■	■	■	■	■	■	■	■	■	▨	■	■	■	■	NXF3 (stSG453305)
■	□	■	□	■	□	□	□	□	□	■	□	□	□	■	□	□	□	□	■	NXF4 (stSG453304)
□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	NXF5 (stSG453303)

Figure 5-14 RT-PCR expression profiling of NXF family genes. Legend as for Figure 5-5.

For the TCP11-like loci, the phylogenetic analyses strongly suggest that the human Xq22 loci were duplicated subsequent to the human-mouse divergence, as noted earlier from gene orientation information, and subsequent to an initial TCP11 duplication creating an X-linked locus (or loci).

The NXF genes show quite different expression patterns, with two main patterns dominant. The patterns of NXF1 and NXF3 appear similar to one another with almost ubiquitous expression, whilst NXF2 and NXF4 show tissue differences. No expression was detected for NXF5.

It should be noted that these similarities in expression pattern reflect the phylogenetic relationship of the loci, and could reflect early divergence of expression patterns within the family, with subsequent gene duplications maintaining more similar patterns. This is likely to be an over-simplification however due to the extensive alternative splicing noted for the NXF-family genes (Herold *et al.*, 2000), which was seen during annotation of the loci.

5.2.4 ALEX family genes

In human Xq22 six ALEX-related sequences were annotated. One of these, dJ545K15.CX.1, appears to be truncated and probably represents a pseudogene. From annotation of the corresponding region of mouse, six ALEX-related sequences were also found which shared orientation and similar positioning with their human counterparts. For all of these genes, orthologue assignment appears straightforward due to their positional information, but only four are supported by phylogenetic analyses of their sequence.

For human genes dJ545K15.CX.1 and dJ545K15.1, and mouse genes bM316A19.MX.2 and bM316A19.MX.3, relationships appear to be closer within than between species. This includes the gene which appears truncated in human. The murine counterpart, based on positional information, bM316A19.MX.2, also appears to be truncated at a similar position, suggesting they are also orthologous. It may be that these genes are in fact functional, but are more highly diverged from the other ALEX-like genes in their more 5' regions hampering annotation. Again, gene conversion or independent gene duplication events driven by shared sequence features leading to similar duplication outcomes (with respect to locations) may underlie evolution of these loci.

The ALEX genes appear to be widely expressed, and no discernible differences were apparent in their expression patterns (Figure 5-18). This is in general agreement with the finding of widespread expression of ALEX1 and ALEX2 in tissues studied by Kurochkin *et. al*. (Kurochkin *et al.*, 2001). The dJ545K15.CX.1 locus was so highly similar in sequence to dJ545K15.1 that an STS could not be designed to discriminate these loci, thus the STS for dJ545K15.1 may detect transcripts from both loci.

The ARM repeat shown in Figure 5-16 represents a multi-helical fold found in a variety of proteins, and may be involved in protein-protein interactions (Peifer *et al.*, 1994). The alignment indicates that this region is the most conserved across the ALEX-family proteins, whilst other regions of the proteins are highly divergent in both length and composition. This indicates an important functional role for this region in the ALEX genes.

The predicted ALEX-family peptide sequences are highly divergent across much of their N-terminal segments, displaying great variation in sequence length and composition. This is most striking in the case of cU209G1.CX.1, which whilst annotated as a predicted structure, appears to be much longer than the other sequences. The predicted mouse orthologue bM26D22.MX.5 appears to share this gene structure. Tandem repeats were noted within the human locus, and it is possible that these have contributed to an expanded gene structure, although further analysis is required to confirm this.

An intriguing observation is the homology seen at the peptide level to KIAA0443 (GASP), as also noted by the authors describing ALEX1-3 (Kurochkin *et al.*, 2001). This suggests that the ALEX and GASP family (see next section) may in fact constitute a larger family comprising 10 members within human Xq22, with 10 members also seen in mouse. Further comments relating to this are made in the next section describing the GASP family.

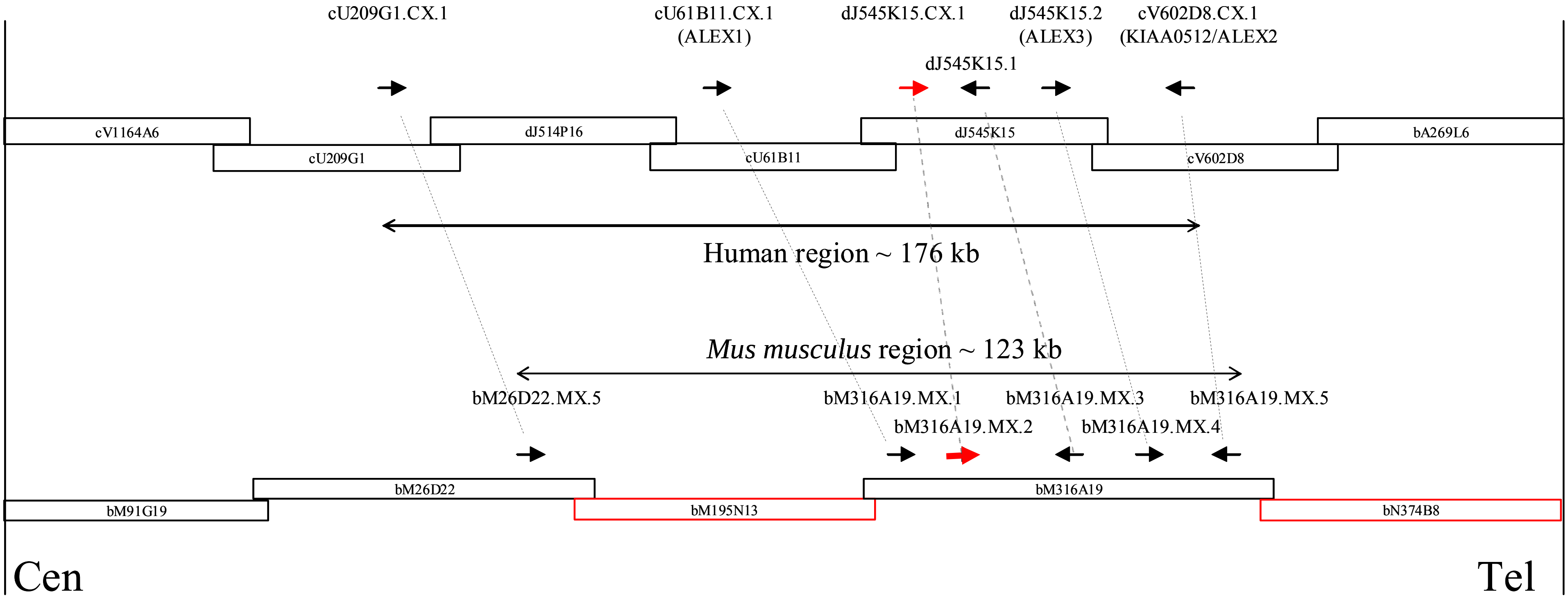
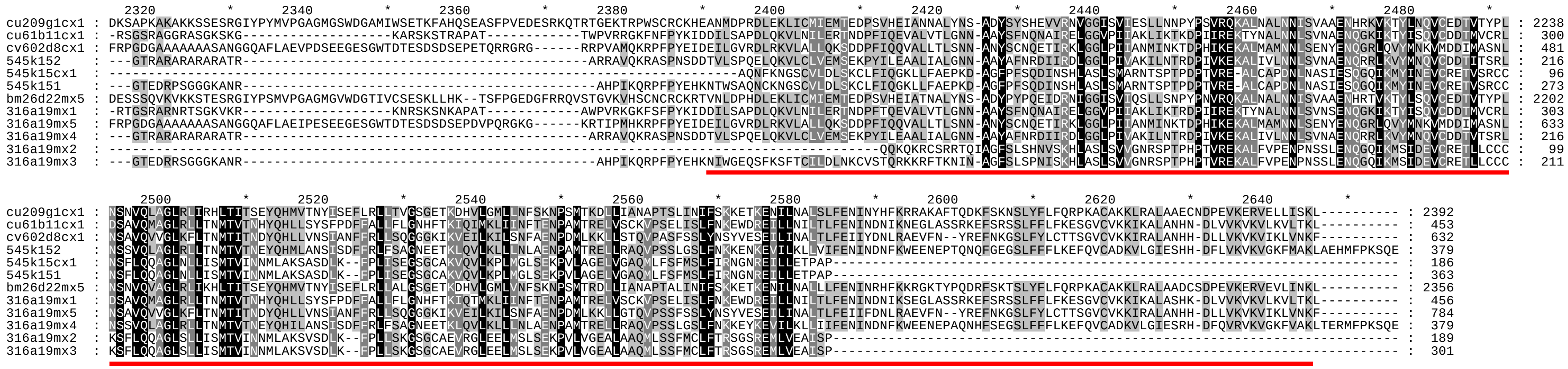


Figure 5-15 Figure showing a schematic representation (Not to scale) of ALEX paralogue gene order and orientation within human Xq22, and the corresponding orthologous region in mouse. Large double-headed arrows depict approximate sizes of the regions, small arrows depict gene positions (names above) and direction of transcription. The genomic clone sequence-ready tile-paths for human and mouse are represented with clone names (N.B. not to scale). Clones in red indicate unfinished sequences, and hence annotation gaps. The red arrow represents the truncated ALEX-like gene that is likely to represent a pseudogene. Grey dotted lines illustrate likely orthologous relationships.



```
               2320         *      2340         *      2360         *      2380         *      2400         *      2420         *      2440         *      2460         *      2480         *
cu209g1cx1 : DKSAPKAKAKKSSESRGIYPYMVPGAGMGSWDGAMIWSETKFAHQSEASFPVEDESRKQTRTGEKTRPWSCRCKHEANMDPRDLEKLICMIEMTEDPSVHEIANNALYNS-ADYSYSHEVVRNVGGISVIESLLNNPYPSVRQKALNALNNISVAAENHRKVKTYLNQVCEDTVTYPL : 2238
cu61b11cx1 : -RSGSRAGGRASGKSKG-------------------KARSKSTRAPAT------------TWPVRRGKFNFPYKIDDILSAPDLQKVLNILERTNDPFIQEVALVTLGNN-AAYSFNQNAIRELGGVPIIAKLIKTKDPIIREKTYNALNNLSVNAENQGKIKTYISQVCDDTMVCRL :  300
cv602d8cx1 : FRPGDGAAAAAAASANGGQAFLAEVPDSEEGESGWTDTESDSDSEPETQRRGRG------RRPVAMQKRPFPYEIDEILGVRDLRKVLALLQKSDDPFIQQVALLTLSNN-ANYSCNQETIRKLGGLPIIANMINKTDPHIKEKALMAMNNLSENYENQGRLQVYMNKVMDDIMASNL :  481
545k152    : ---GTRARARARARATR--------------------------------------------ARRAVQKRASPNSDDTVLSPQELQKVLCLVEMSEKPYILEAALIALGNN-AAYAFNRDIIRDLGGLPIVAKILNTRDPIVKEKALIVLNNLSVNAENQRRLKVYMNQVCDDTITSRL :  216
545k15cx1  : --------------------------------------------------------------------------------AQNFKNGSCVLDLSKCLFIQGKLLFAEPKD-AGFPFSQDINSHLASLSMARNTSPTPDPTVRE-ALCAPDNLNASIESQGQIKMYINEVCRETVSRCC :   96
545k151    : ---GTEDRPSGGGKANR---------------------------------------------AHPIKQRPFPYEHKNTWSAQNCKNGSCVLDLSKCLFIQGKLLFAEPKD-AGFPFSQDINSHLASLSMARNTSPTPDPTVRE-ALCAPDNLNASIESQGQIKMYINEVCRETVSRCC :  273
bm26d22mx5 : DESSSQVKVKKSTESRGIYPSMVPGAGMGVWDGTIVCSESKLLHK--TSFPGEDGFRRQVSTGVKVHSCNCRCKRTVNLDPHDLEKLICMIEMTEDPSVHEIATNALYNS-ADYPYPQEIDRNIGGISVIQSLLSNPYPNVRQKALNALNNISVAAENHRTVKTYLSQVCEDTVTYPL : 2202
316a19mx1  : -RTGSRARNRTSGKVKR-------------------KNRSKSNKAPAT------------AWPVRKGKFSFPYKIDDILSAPDLQKVLNILERTNDPFTQEVALVTLGNN-AAYSFNQNAIRELGGVPIIAKLIKTRDPIIREKTYNALNNLSVNSENQGKIKTYISQVCDDTMVCRL :  303
316a19mx5  : FRPGDGAAAAAAASANGGQAFLAEIPESEEGESGWTDTESDSDSEPDVPQRGKG------KRTIPMHKRPFPYEIDEILGVRDLRKVLALLQKSDDPFIQQVALLTLSNN-ANYSCNQETIRKLGGLPIIANMINKTDPHIKEKALMAMNNLSENYENQGRLQVYMNKVMDDIMASNL :  633
316a19mx4  : ---GTRARARARARATR--------------------------------------------ARRAVQKRASPNSDDTVLSPQELQKVLCLVEMSEKPYILEAALIALGNN-AAYAFNRDIIRDLGGLPIVAKILNTRDPIVKEKALIVLNNLSVNAENQRRLKVYMNQVCDDTVTSRL :  216
316a19mx2  : --------------------------------------------------------------------------------------------------QQKQKRCSRRTQIAGFSLSHNVSKHLASLSVVGNRSPTPHPTVREKALFVPENPNSSLENQGQIKMSIDEVCRETLLCCC :   99
316a19mx3  : ---GTEDRRSGGGKANR---------------------------------------------AHPIKQRPFPYEHKNIWGEQSFKSFTCILDLNKCVSTQRKKRFTKNIN-AGFSLSPNISKHLASLSVVGNRSPTPHPTVREKALFVPENPNSSLENQGQIKMSIDEVCRETLLCCC :  211

                 2500         *      2520         *      2540         *      2560         *      2580         *      2600         *      2620         *      2640         *
cu209g1cx1 : NSNVQLAGLRLIRHLTITSEYQHMVTNYISEFLRLLTVGSGETKDHVLGMLLNFSKNPSMTKDLLIANAPTSLINIFSKKETKENILNALSLFENINYHFKRRAKAFTQDKFSKNSLYFLFQRPKACAKKLRALAAECNDPEVKERVELLISKL---------- : 2392
cu61b11cx1 : DSAVQMAGLRLLTNMTVTNHYQHLLSYSFPDFFALLFLGNHFTKIQIMKLIINFTENPAMTRELVSCKVPSELISLFNKEWDREILLNILTLFENINDNIKNEGLASSRKEFSRSSLFFLFKESGVCVKKIKALANHN-DLVVKVKVLKVLTKL---------- :  453
cv602d8cx1 : NSAVQVVGLKFLTNMTITNDYQHLLVNSIANFFRLLSQGGGKIKVEILKILSNFAENPDMLKKLLSTQVPASFSSLYNSYVESEILINALTLFEIIYDNLRAEVFN--YREFNKGSLFYLCTTSGVCVKKIRALANHH-DLLVKVKVIKLVNKF---------- :  632
545k152    : NSSVQLAGLRLLTNMTVTNEYQHMLANSISDFFRLFSAGNEETKLQVLKLLLNLAENPAMTRELLRAQVPSSLGSLFNKKENKEVILKLLVIFENINDNFKWEENEPTQNQFGEGSLFFFLKEFQVCADKVLGIESHH-DFLVKVKVGKFMAKLAEHMFPKSQE :  379
545k15cx1  : NSFLQQAGLNLLISMTVINNMLAKSASDLK--FPLISEGSGCAKVQVLKPLMGLSEKPVLAGELVGAQMLFSFMSLFIRNGNREILLETPAP------------------------------------------------------------------------ :  186
545k151    : NSFLQQAGLNLLISMTVINNMLAKSASDLK--FPLISEGSGCAKVQVLKPLMGLSEKPVLAGELVGAQMLFSFMSLFIRNGNREILLETPAP------------------------------------------------------------------------ :  363
bm26d22mx5 : NSNVQVAGLRLIKHLTITSEYQHMVTNYISEFLRLLALGSGETKDHVLGMLVNFSKNPSMTRDLLIANAPTALINIFSKKETKENILNALLLFENINRHFKKRGKTYPQDRFSKTSLYFLFQRPKACAKKLRALAADCSDPEVKERVEVLINKL---------- : 2356
316a19mx1  : DSAVQMAGLRLLTNMTVTNHYQHLLSYSFPDFFALLFLGNHFTKIQTMKLIINFTENPAMTRELVSCKVPSELISLFNKEWDREILLNILTLFENINDNIKSEGLASSRKEFSRSSLFFLFKESGVCVKKIKALASHK-DLVVKVKVLKVLTKL---------- :  456
316a19mx5  : NSAVQVVGLKFLTNMTITNDYQHLLVNSIANFFRLLSQGGGKIKVEILKILSNFAENPDMLKKLLGTQVPSSFSSLYNSYVESEILINALTLFEIIFDNLRAEVFN--YREFNKGSLFYLCTTSGVCVKKIRALANHH-DLLVKVKVIKLVNKF---------- :  784
316a19mx4  : NSSVQLAGLRLLTNMTVTNEYQHILANSISDFFRLFSAGNEETKLQVLKLLLNLAENPAMTRELLRAQVPSSLGSLFNKKEYKEVILKLLIIFENINDNFKWEENEPAQNHFSEGSLFFFLKEFQVCADKVLGIESRH-DFQVRVKVGKFVAKLTERMFPKSQE :  379
316a19mx2  : KSFLQQAGLSLLISMTVINNMLAKSVSDLK--FPLLSKGSGCAEVRGLEELMSLSEKPVLVGEALAAQMLSSFMCLFTRSGSREMLVEAISP------------------------------------------------------------------------ :  189
316a19mx3  : KSFLQQAGLSLLISMTVINNMLAKSVSDLK--FPLLSKGSGCAEVRGLEELMSLSEKPVLVGEALAAQMLSSFMCLFTRSGSREMLVEAISP------------------------------------------------------------------------ :  301
```

Figure 5-16 Figure illustrating alignment of ALEX-like predicted peptides, for human and mouse. For clarity, only the most highly conserved carboxyl terminus regions are shown. For technical reasons, gene names have had "." separators removed in the figures, and some genes have had their clone-based prefix ("cV", "dJ", "bM" etc.) removed. The red box under the aligned residues represents an ARM repeat region (IPR008938), predicted from InterPro analysis of ALEX2 peptide sequence.

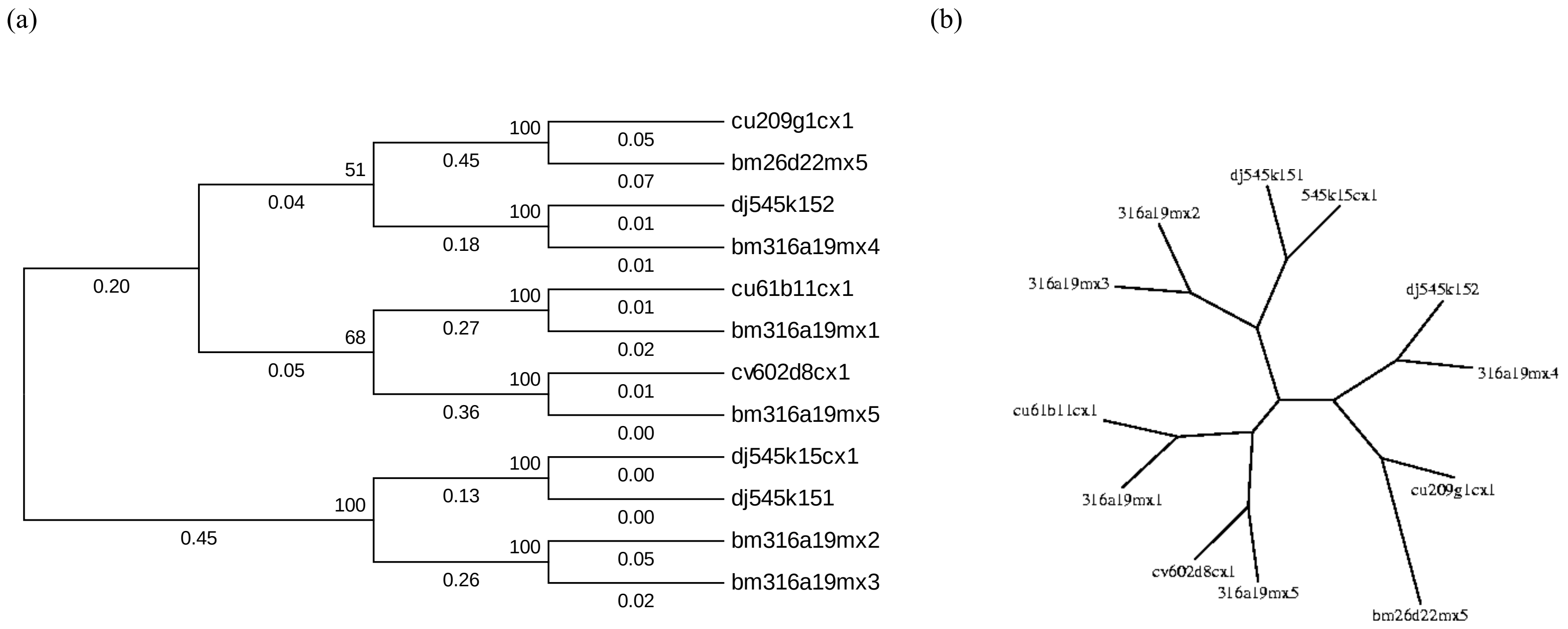
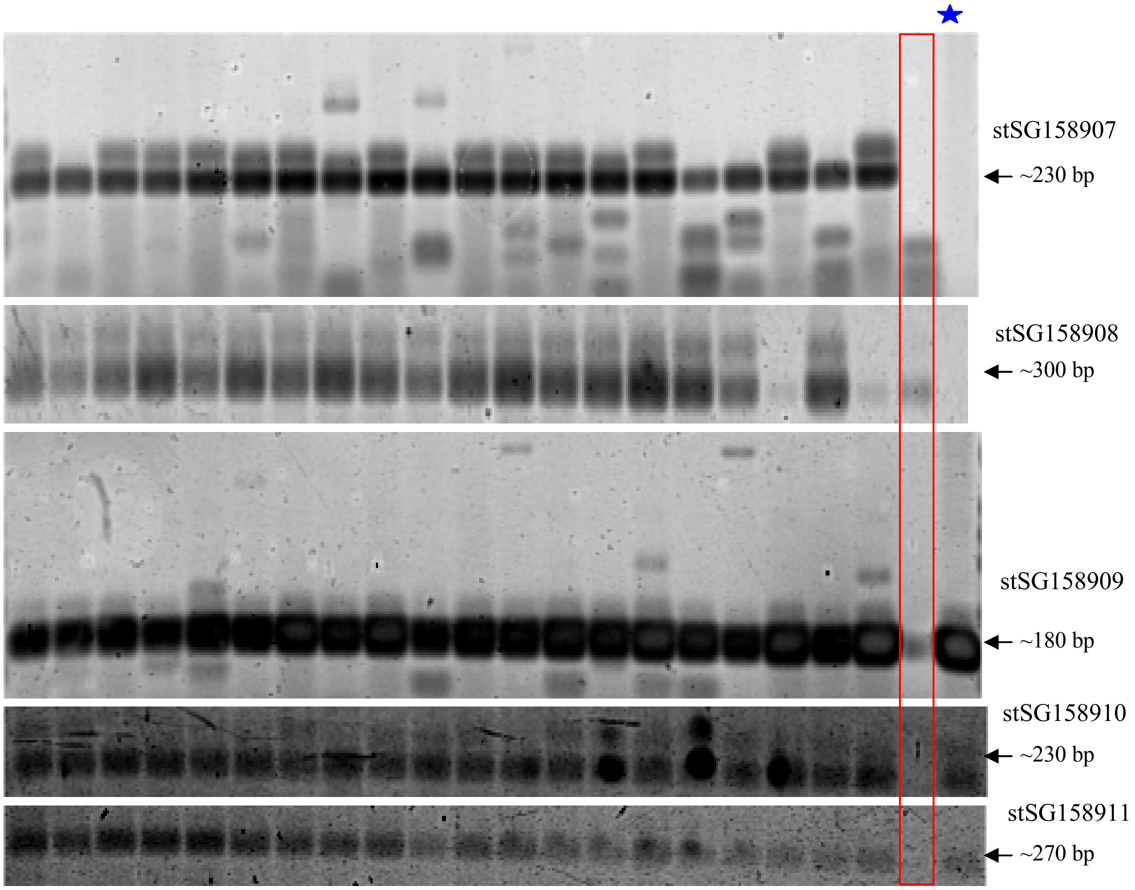
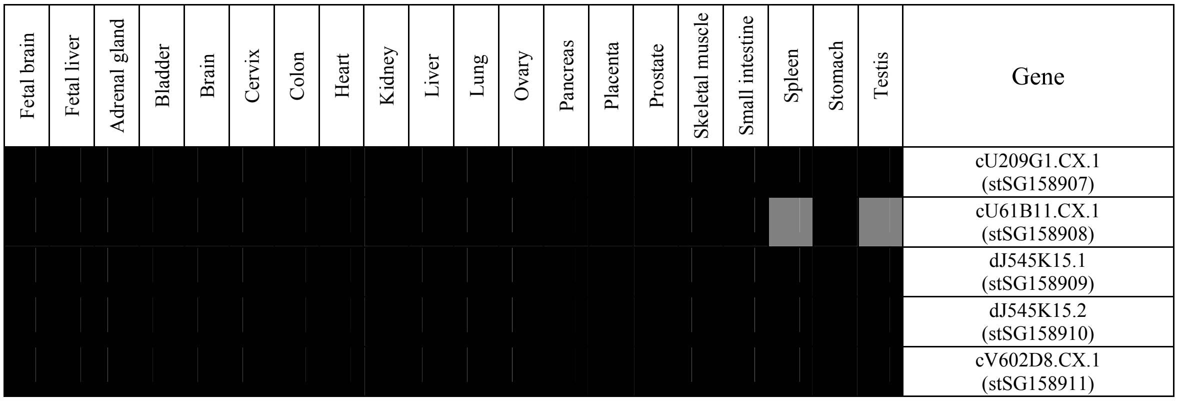


Figure 5-17 Phylogenetic analysis of human and mouse ALEX-like genes (a) a distance-based cladogram computed using an alignment of human and mouse ALEX-like genes predicted peptide sequences. Numbers adjacent to nodes represent the number of bootstrap replicates supporting the adjacent node. Numbers below branches indicate distance. For clarity, only the tree topology is shown and branches are not scaled. (b) maximum-likelihood analysis of the same data used in (a). For technical reasons, gene names have had "." separators removed in the figures, and some genes have had their clone-based prefix ("cV", "dJ", "bM" etc.) removed.

(a)



(b)



Fetal brain	Fetal liver	Adrenal gland	Bladder	Brain	Cervix	Colon	Heart	Kidney	Liver	Lung	Ovary	Pancreas	Placenta	Prostate	Skeletal muscle	Small intestine	Spleen	Stomach	Testis	Gene
■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	cU209G1.CX.1 (stSG158907)
■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	▨	■	▨	cU61B11.CX.1 (stSG158908)
■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	dJ545K15.1 (stSG158909)
■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	dJ545K15.2 (stSG158910)
■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	cV602D8.CX.1 (stSG158911)

Figure 5-18 RT-PCR expression profiling of ALEX family genes. Legend as for Figure 5-5. Failure of the genomic DNA control was noted for stSG158907 and stSG158908, although the tissue reactions gave product, most likely indicating an experimental error attributed to the genomic control only.

5.2.5 GASP family genes

Within human Xq22 four GASP-like sequences have been annotated and the same number of GASP-like loci was annotated in the mouse. Directions of transcription and gene positioning appear to be shared between the two species, with the genes quite tightly clustered compared to some of the other Xq22 duplicated genes. The phylogenetic analyses further support straightforward assignments of orthology, as indicated in Figure 5-21. The phylogenetic analyses shown in Figure 5-21 suggest clear relationships between each of the GASP-like genes and a mouse gene from maximum-likelihood based analysis, but less clear for dJ769N13.1 and bM94I24.MX.1 from distance-based analysis.

The GASP genes appear to be widely expressed, and expression of all four members of the family was detected in all tissues examined.

The GASP family genes appear divergent, and vary greatly in the N-terminal regions of their predicted peptides. Higher conservation is seen in the C-terminal regions. This corresponds to the region reported to be involved in internalization of GPCRs by GASP (Whistler *et al.*, 2002), implying a shared functional role for this region across the family.

The red bars in Figure 5-20 underline the region of GASP (dJ769N13.1/KIAA0443) shown to be involved in GPCR binding (Whistler *et al.*, 2002). The raised level of conservation in this region across the GASP family predicted peptides, which differ markedly in length and composition in other areas of the peptides, may be indicative that this region is important for function in other GASP-related genes.

The homology seen between Xq22 ALEX family proteins and GASP family proteins suggests that these two families may represent a 10-gene family which have diverged greatly. Three strands of reasoning support this hypothesis. One is that the homology seen is found in the C-terminal regions of the predicted peptides (see Figure 5-23), consistent with a functional role for this region in GASP and the prediction of an ARM repeat domain within the ALEX family. Furthermore, domain analysis of GASP peptide predicts the presence of a DUF634 domain in the C-terminal region. This domain is also predicted within ALEX2, overlapping the ARM repeat. The second is that similarities are seen between the ALEX and GASP gene structures, with several small 5’ exons preceding a larger exon. The third is that the two gene families are both within Xq22, and could have been produced as part of events shaping the evolution of the region that appears to have created extensive paralogy.

If the ALEX and GASP genes constitute a larger gene family as suggested here, this would have important implications for functional studies of these proteins. As there is partial functional information available for GASP, this could indicate a potential role for the other 3 GASP family genes and ALEX-family genes in the sorting of GPCRs within cells. Further studies could be designed to test this hypothesis. As many X-linked non-syndromic mental retardation (MRX) loci have been mapped to the Xq22-q23 region, a potential role in GPCR sorting would make the GASP and ALEX loci worthwhile candidates for mutation screening in these disorders, due to the role of GPCRs in neural signalling (Hescheler and Schultz 1993). It must be stressed though that, even if the GASP and ALEX genes share a common ancestor, they have diverged substantially and may have adopted quite different roles. Their widespread expression may indicate that the different gene products have undergone sub-functionalisation in order to retain selective advantage.

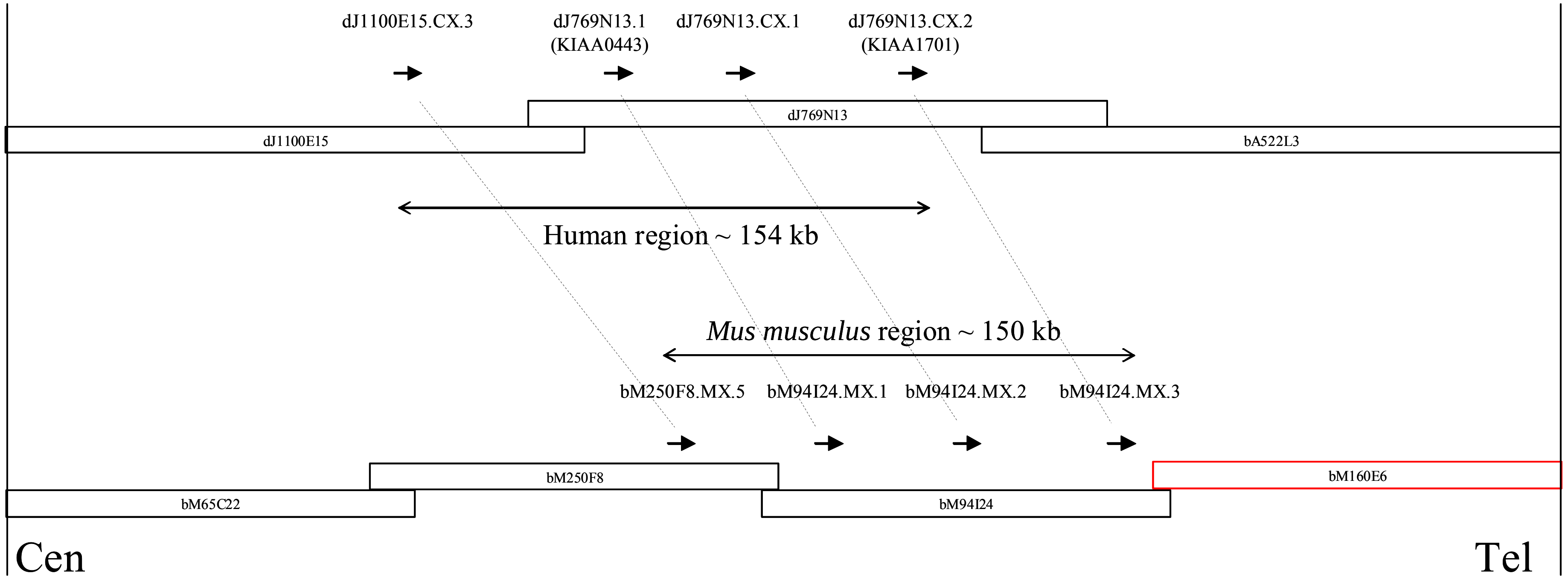
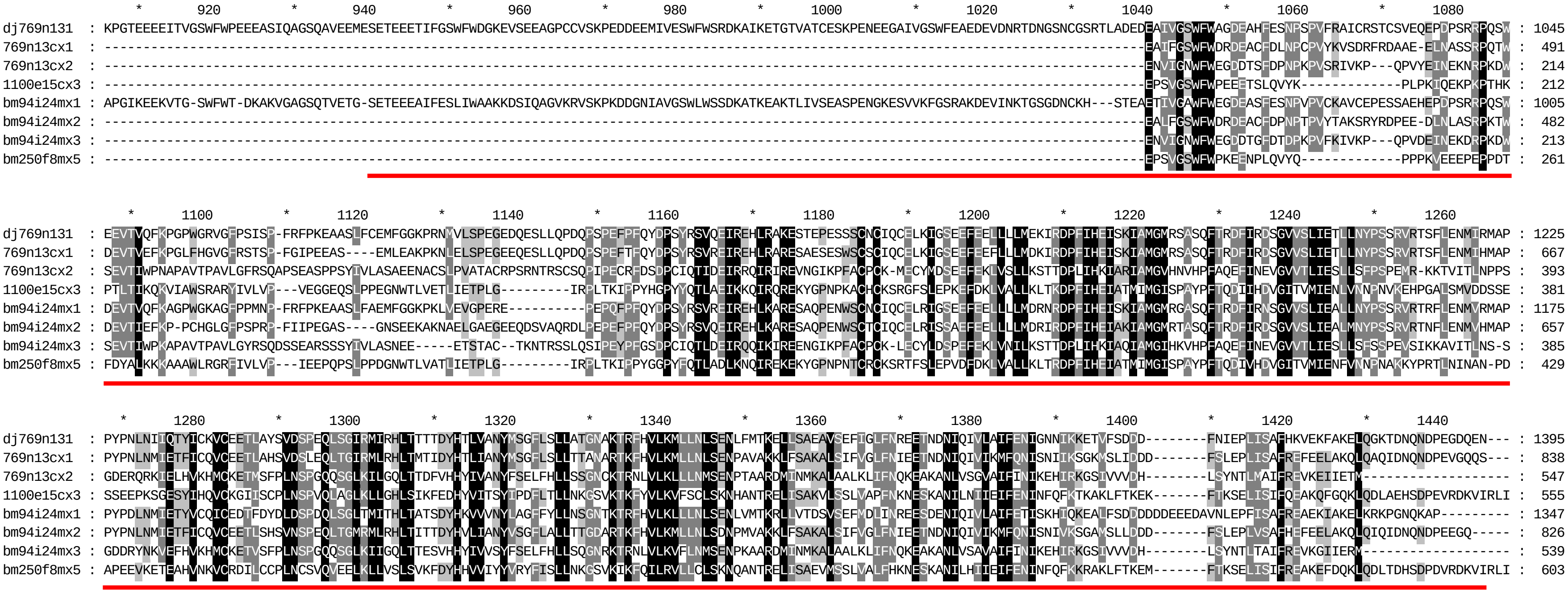


Figure 5-19 Figure showing a schematic representation (not to scale) of GASP paralogue gene order and orientation within human Xq22, and the corresponding orthologous region in mouse. Large double-headed arrows depict approximate sizes of the regions, small arrows depict gene positions (names above) and direction of transcription. The genomic clone sequence-ready tile-paths for human and mouse are represented with clone names (N.B. not to scale). Clones in red indicate unfinished sequence, and hence a sequence gap. Grey dotted lines illustrate likely orthologous relationships.



```
                 *       920         *       940         *       960         *       980         *      1000         *      1020         *      1040         *      1060         *      1080
dj769n131  : KPGTEEEEITVGSWFWPEEEASIQAGSQAVEEMESETEEETIFGSWFWDGKEVSEEAGPCCVSKPEDDEEMIVESWFWSRDKAIKETGTVATCESKPENEEGAIVGSWFEAEDEVDNRTDNGSNCGSRTLADEDEAIVGSWFWAGDEAHFESNPSPVFRAICRSTCSVEQEPDPSRRPQSW : 1045
769n13cx1  : --------------------------------------------------------------------------------------------------------------------------------------EAIFGSWFWDRDEACFDLNPCPVYKVSDRFRDAAE-ELNASSRPQTW :  491
769n13cx2  : --------------------------------------------------------------------------------------------------------------------------------------ENVIGNWFWEGDDTSFDPNPKPVSRIVKP---QPVYEINEKNRPKDW :  214
1100e15cx3 : --------------------------------------------------------------------------------------------------------------------------------------EPSVGSWFWPEEETSLQVYK-------------PLPKIQEKPKPTHK :  212
bm94i24mx1 : APGIKEEKVTG-SWFWT-DKAKVGAGSQTVETG-SETEEEAIFESLIWAAKKDSIQAGVKRVSKPKDDGNIAVGSWLWSSDKATKEAKTLIVSEASPENGKESVVKFGSRAKDEVINKTGSGDNCKH---STEAETIVGAWFWEGDEASFESNPVPVCKAVCEPESSAEHEPDPSRRPQSW : 1005
bm94i24mx2 : --------------------------------------------------------------------------------------------------------------------------------------EALFGSWFWDRDEACFDPNPTPVYTAKSRYRDPEE-DLNLASRPKTW :  482
bm94i24mx3 : --------------------------------------------------------------------------------------------------------------------------------------ENVIGNWFWEGDDTGFDTDPKPVFKIVKP---QPVDEINEKDRPKDW :  213
bm250f8mx5 : --------------------------------------------------------------------------------------------------------------------------------------EPSVGSWFWPKEENPLQVYQ-------------PPPKVEEEPEPPDT :  261

                 *      1100         *      1120         *      1140         *      1160         *      1180         *      1200         *      1220         *      1240         *      1260
dj769n131  : EEVTVQFKPGPWGRVGFPSISP-FRFPKEAASLFCEMFGGKPRNMVLSPEGEDQESLLQPDQPSPEFPFQYDPSYRSVQEIREHLRAKESTEPESSSCNCIQCELKIGSEEFEELLLLMEKIRDPFIHEISKIAMGMRSASQFTRDFIRDSGVVSLIETLLNYPSSRVRTSFLENMIRMAP : 1225
769n13cx1  : DEVTVEFKPGLFHGVGFRSTSP-FGIPEEAS----EMLEAKPKNLELSPEGEEQESLLQPDQPSPEFTFQYDPSYRSVREIREHLRARESAESESWSCSCIQCELKIGSEEFEEFLLLMDKIRDPFIHEISKIAMGMRSASQFTRDFIRDSGVVSLIETLLNYPSSRVRTSFLENMIHMAP :  667
769n13cx2  : SEVTIWPNAPAVTPAVLGFRSQAPSEASPPSYIVLASAEENACSLPVATACRPSRNTRSCSQPIPECRFDSDPCIQTIDEIRRQIRIREVNGIKPFACPCK-MECYMDSEEFEKLVSLLKSTTDPLIHKIARIAMGVHNVHPFAQEFINEVGVVTLIESLLSFPSPEMR-KKTVITLNPPS :  393
1100e15cx3 : PTLTIKQKVIAWSRARYIVLVP---VEGGEQSLPPEGNWTLVETLIETPLG---------IRPLTKIPPYHGPYYQTLAEIKKQIRQREKYGPNPKACHCKSRGFSLEPKEFDKLVALLKLTKDPFIHEIATMIMGISPAYPFTQDIIHDVGITVMIENLVNNPNVKEHPGALSMVDDSSE :  381
bm94i24mx1 : DEVTVQFKAGPWGKAGFPPMNP-FRFPKEAASLFAEMFGGKPKLVEVGPERE----------PEPQFPFQYDPSYRSVREIREHLKARESAQPENWSCNCIQCELRIGSEEFEELLLLMDRNRDPFIHEISKIAMGMRGASQFTRDFIRNSGVVSLIEALLNYPSSRVRTRFLENMVRMAP : 1175
bm94i24mx2 : DEVTIEFKP-PCHGLGFPSPRP-FIIPEGAS----GNSEEKAKNAELGAEGEEQDSVAQRDLPEPEFPFQYDPSYRSVQEIREHLKARESAQPENWSCTCIQCELRISSAEFEELLLLMDRIRDPFIHEIAKIAMGMRTASQFTRDFIRDSGVVSLIEALMNYPSSRVRTNFLENMVHMAP :  657
bm94i24mx3 : SEVTIWPKAPAVTPAVLGYRSQDSSEARSSSYIVLASNEE-----ETSTAC--TKNTRSSLQSIPEYPFGSDPCIQTLDEIRQQIKIREENGIKPFACPCK-LECYLDSPEFEKLVNILKSTTDPLIHKIAQIAMGIHKVHPFAQEFINEVGVVTLIESLLSFSSPEVSIKKAVITLNS-S :  385
bm250f8mx5 : FDYALKKKAAAWLRGRFIVLVP---IEEPQPSLPPDGNWTLVATLIETPLG---------IRPLTKIPPYGGPYFQTLADLKNQIREKEKYGPNPNTCRCKSRTFSLEPVDFDKLVALLKLTRDPFIHEIATMIMGISPAYPFTQDIVHDVGITVMIENFVNNPNAKKYPRTLNINAN-PD :  429

                 *      1280         *      1300         *      1320         *      1340         *      1360         *      1380         *      1400         *      1420         *      1440
dj769n131  : PYPNLNIIQTYICKVCEETLAYSVDSPEQLSGIRMIRHLTTTTDYHTLVANYMSGFLSLLATGNAKTRFHVLKMLLNLSENLFMTKELLSAEAVSEFIGLFNREETNDNIQIVLAIFENIGNNIKKETVFSDDD--------FNIEPLISAFHKVEKFAKELQGKTDNQNDPEGDQEN--- : 1395
769n13cx1  : PYPNLNMIETFICQVCEETLAHSVDSLEQLTGIRMLRHLTMTIDYHTLIANYMSGFLSLLTTANARTKFHVLKMLLNLSENPAVAKKLFSAKALSIFVGLFNIEETNDNIQIVIKMFQNISNIIKSGKMSLIDDD-------FSLEPLISAFREFEELAKQLQAQIDNQNDPEVGQQS--- :  838
769n13cx2  : GDERQRKIELHVKHMCKETMSFPLNSPGQQSGLKILGQLTTDFVHHYIVANYFSELFHLLSSGNCKTRNLVLKLLLNMSENPTAARDMINMKALAALKLIFNQKEAKANLVSGVAIFINIKEHIRKGSIVVVDH--------LSYNTLMAIFREVKEIIETM------------------- :  547
1100e15cx3 : SSEEPKSGESYIHQVCKGIISCPLNSPVQLAGLKLLGHLSIKFEDHYVITSYIPDFLTLLNKGSVKTKFYVLKVFSCLSKNHANTRELISAKVLSSLVAPFNKNESKANILNIIEIFENINFQFKTKAKLFTKEK-------FTKSELISIFQEAKQFGQKLQDLAEHSDPEVRDKVIRLI :  555
bm94i24mx1 : PYPDLNMIETYVCQICEDTFDYDLDSPDQLSGLTMITHLTATSDYHKVVVNYLAGFFYLLNSGNTKTRFHVLKLLLNLSENLVMTKRLLVTDSVSEFMDLINREESDENIQIVLAIFETISKHIQKEALFSDDDDDDEEEDAVNLEPFISAFREAEKIAKELKRKPGNQKAP--------- : 1347
bm94i24mx2 : PYPNLNMIETFICQVCEETLSHSVNSPEQLTGMRMLRHLTITTDYHVLIANYVSGFLALLTTGDARTKFHVLKMLLNLSDNPMVAKKLFSAKALSIFVGLFNIEETNDNIQIVIKMFQNISNIVKSGAMSLLDDD-------FSLEPLVSAFHEFEELAKQLQIQIDNQNDPEEGQ----- :  826
bm94i24mx3 : GDDRYNKVEFHVKHMCKETVSFPLNSPGQQSGLKIIGQLTTESVHHYIVVSYFSELFHLLSQGNRKTRNLVLKVFLNMSENPKAARDMINMKALAALKLIFNQKEAKANLVSAVAIFINIKEHIRKGSIVVVDH--------LSYNTLTAIFREVKGIIERM------------------- :  539
bm250f8mx5 : APEEVKETEAHVNKVCRDILCCPLNCSVQVEELKLLVSLSVKFDYHHVVIYYVRYFISLLNKGSVKIKFQILRVLLCLSKNQANTRELISAEVMSSLVALFHKNESKANILHIIEIFENINFQFKKRAKLFTKEM-------FTKSELISIFREAKEFDQKLQDLTDHSDPDVRDKVIRLI :  603
```

Figure 5-20 Figure illustrating alignments of GASP paralogue predicted peptides. For clarity, only selected regions of the carboxyl terminus regions are shown. For technical reasons, gene names have had "." separators removed in the figures, and some genes have had their clone-based prefix ("cV", "dJ", "bM" etc.) removed. The red bars underline the region of GASP (dJ769N13.1/KIAA0443) shown to be involved in GPCR binding (Whistler *et al.*, 2002).

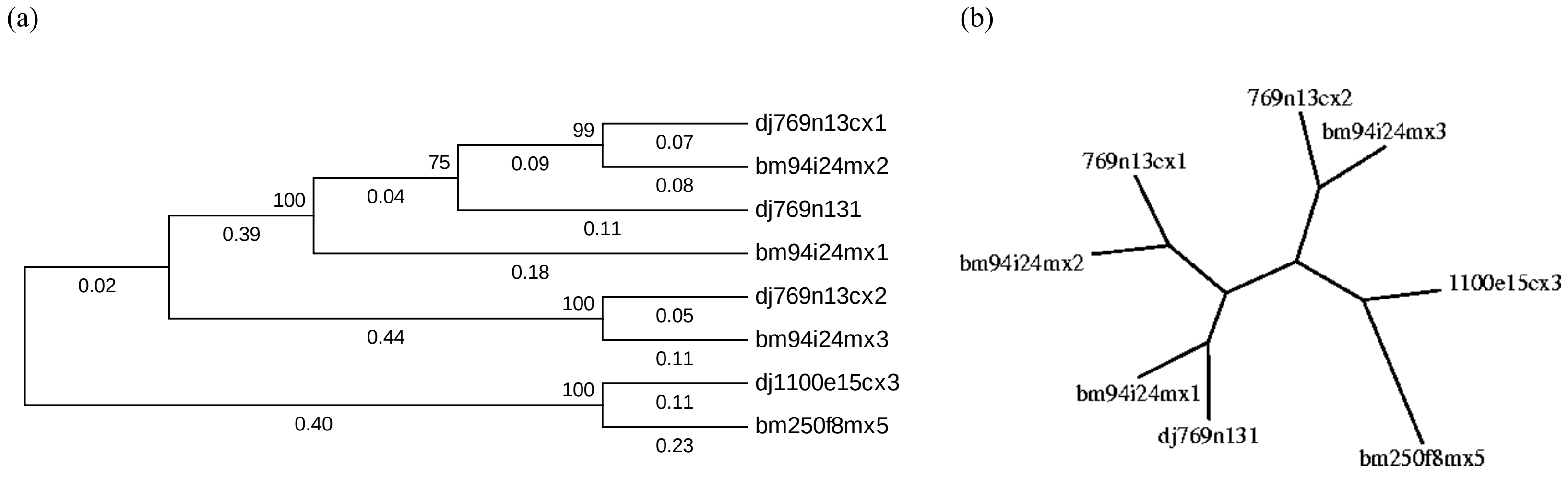
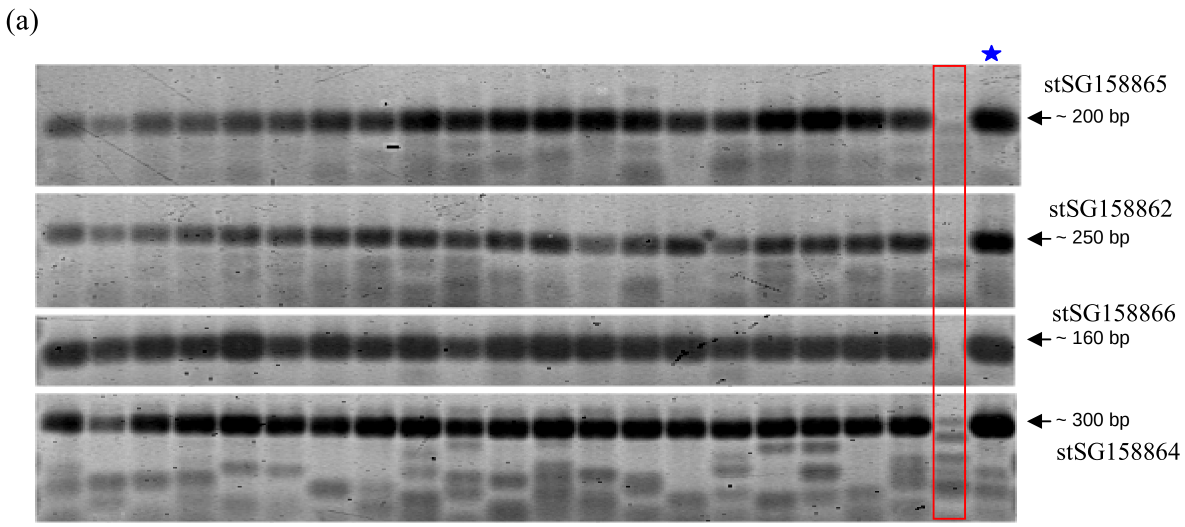


Figure 5-21 Figure illustrating phylogenetic analysis of GASP-like genes. (a) a distance-based cladogram computed using an alignment of human and mouse GASP-like genes predicted peptide sequences. Numbers adjacent to nodes represent the number of bootstrap replicates supporting the adjacent node. Numbers below branches indicate distance. For clarity, only the tree topology is shown and branches are not scaled. (b) maximum-likelihood analysis of the same data used in (a). For technical reasons, gene names have had "." separators removed in the figures, and some genes have had their clone-based prefix ("cV", "dJ", "bM" etc.) removed.



(b)

Fetal brain	Fetal liver	Adrenal gland	Bladder	Brain	Cervix	Colon	Heart	Kidney	Liver	Lung	Ovary	Pancreas	Placenta	Prostate	Skeletal muscle	Small intestine	Spleen	Stomach	Testis	Gene
■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	dJ1100E15.CX.3 (stSG158865)
■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	dJ769N13.1 (stSG158862)
■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	dJ769N13.CX.1 (stSG158866)
■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	dJ769N13.CX.2 (stSG158864)

Figure 5-22 RT-PCR expression profiling of GASP family genes. Legend as for Figure 5-5.

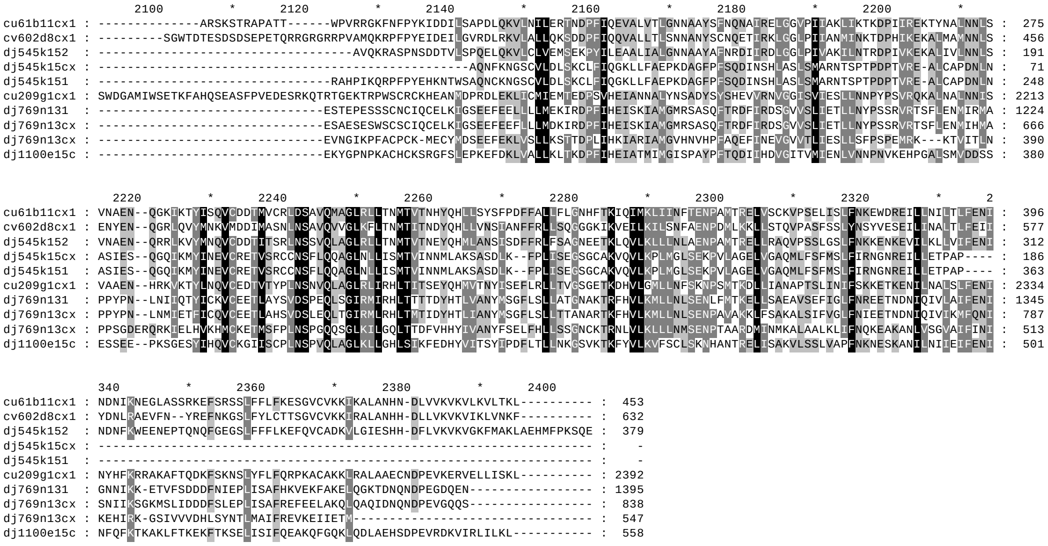


Figure 5-23 Alignment of GASP and ALEX family predicted peptides showing conservation between the sequences. For clarity, only the C-terminal regions are displayed.

5.2.6 *pp21/TCEAL genes*

After the hypothesised ALEX/GASP family, the pp21 family is the second largest family within Xq22 with nine members annotated. Only seven pp21-like genes are found in the mouse orthologous region, although an extra gene may reside in sequence gaps proximal to clone bM460B8 (see Chapter 4). For four of the genes, orthologous relationships appear to be clear, based on gene position and orientation (Figure 5-24), and these are supported by phylogenetic analyses (Figure 5-26). For the remaining genes, the following orthologous relationships are suggested by relative positioning and orientation: cV351F8.CX.1 and bM460B8.MX.1, cU177E8.CX.3 and bM132M9.MX.2, and, finally, either cV857G6.CX.1 or cV857G6.CX.2 and bM197O15.MX.1. However, as has been seen in previous sections, phylogenetic analyses suggest that these sequences are more related to their counterparts within the same species than to their potential orthologues.

As before, either gene conversion is occurring in these cases or there have been multiple independent duplications in each lineage, driven by sequence features shared within the human and mouse genomic regions. As this has been seen now for three other paralogous families within Xq22, independent gene duplications seem less likely to have generated all three families and gene conversion may be a factor in sequence homogenisation within these gene families.

For the gene subset shown in the upper part of Figure 5-26, many genes appear more homologous within the species. For the other subset, clear relationships are seen between species.

Whilst homology is seen across the pp21-like family, the sequences are quite divergent overall, and retain most of their sequence conservation within two sub-groups as seen in Figure 5-25. Several key residues appear to be well conserved across the whole family (Figure 5-25), and may be important for the functions of these proteins. A clue as to the potential function of the pp21-like proteins comes from the analysis of TCEAL1, which suggests that proteins of this family may be involved in regulation of transcription (Pillutla *et al.*, 1999).

The expression patterns of these genes are more informative than for the families seen in previous sections. Whilst these genes also appear to be widely expressed, differences are seen between the paralogues. Of particular note are the patterns of TCEAL1 and cU105G4.2 (pp21 homologue) which appear similar to one another. These genes also appear most related to one another by phylogenetic analyses. The patterns of dJ122O23.CX.1 and cV351F8.CX.1 also appear similar, in this instance the genes are not the most related from phylogenetic analyses, but are adjacent to one another. For these STSs the colony PCR assay did not provide clear evidence that the primers discriminated between these loci however. Further exploration of the expression patterns of the pp21-like gene family and their mouse orthologues may shed further light on their evolution in support of the DDC hypothesis.

It should also be noted that the expression pattern of TCEAL1 from Northern Blot analysis has been reported (Pillutla *et al.*, 1999), and the authors detected expression in all of the tissues for which RT-PCR results were negative in this study (see Figure 5-27). This could be due to RNA source differences or methodological differences, and illustrates the difficulties in maintaining consistency in expression profiling approaches from which to draw inferences regarding differences in paralogue expression patterns.

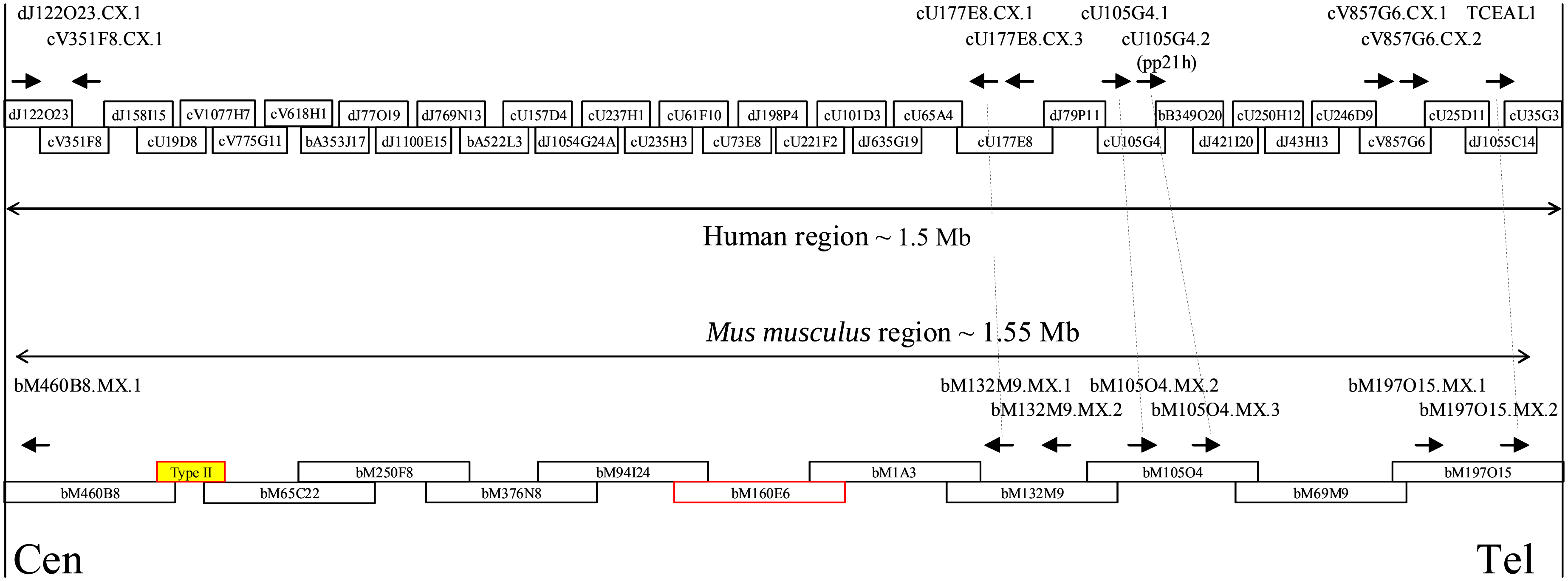


Figure 5-24 Figure showing a schematic representation (not to scale) of pp21/TCEAL1 paralogue gene order and orientation within human Xq22, and the corresponding orthologous region in mouse. Large double-headed arrows depict approximate sizes of the regions, small arrows depict gene positions (names above) and direction of transcription. The genomic clone sequence-ready tile-paths for human and mouse are represented with clone names (not to scale). Clones in red indicate unfinished sequence, and hence a sequence gap. The yellow box represents a type -II gap (no clone selected for sequencing at this position). Grey dotted lines illustrate clear likely orthologous relationships.

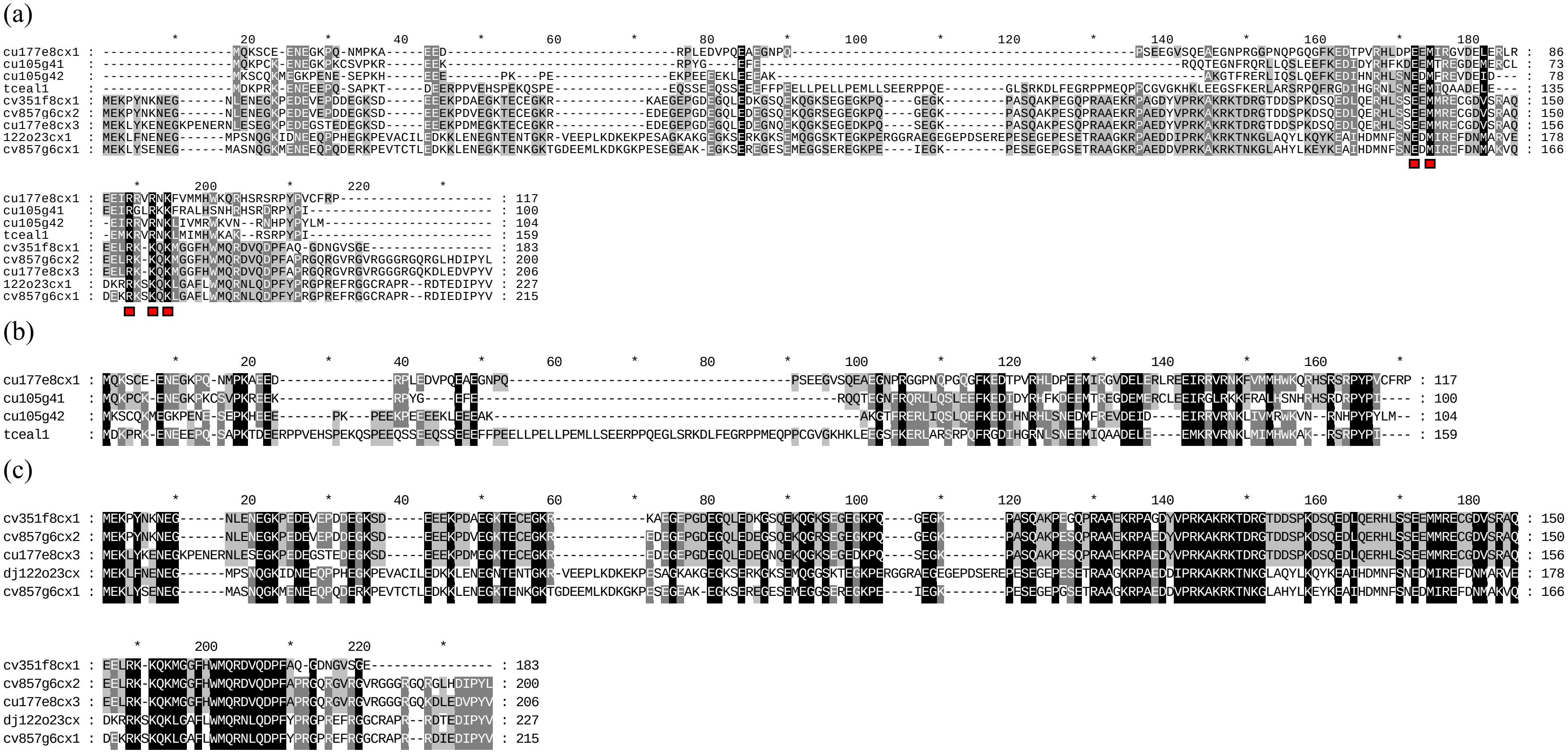


Figure 5-25 Figure illustrating alignments of pp21/TCEAL human paralogue predicted peptides. For technical reasons, gene names have had "." separators removed in the figures, and some genes have had their clone-based prefix ("cV", "dJ", "bM" etc.) removed. The alignment in (a) illustrates varied homology across the family, with some particularly conserved residues that may be important for function (indicated by red boxes underneath). Alignments (b) and (c) show predicted peptide homologies for subsets of the genes that appear most related to one another.

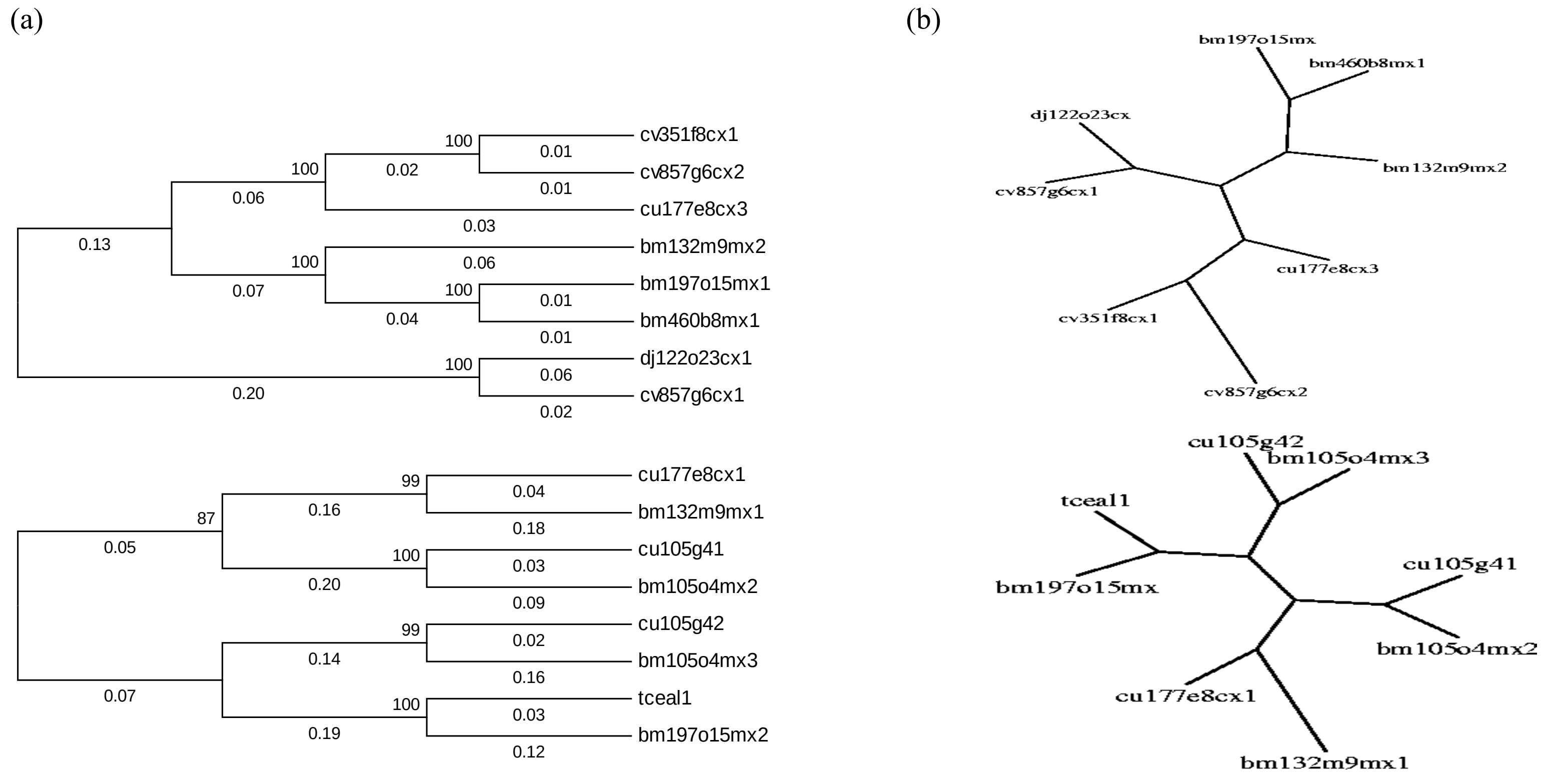
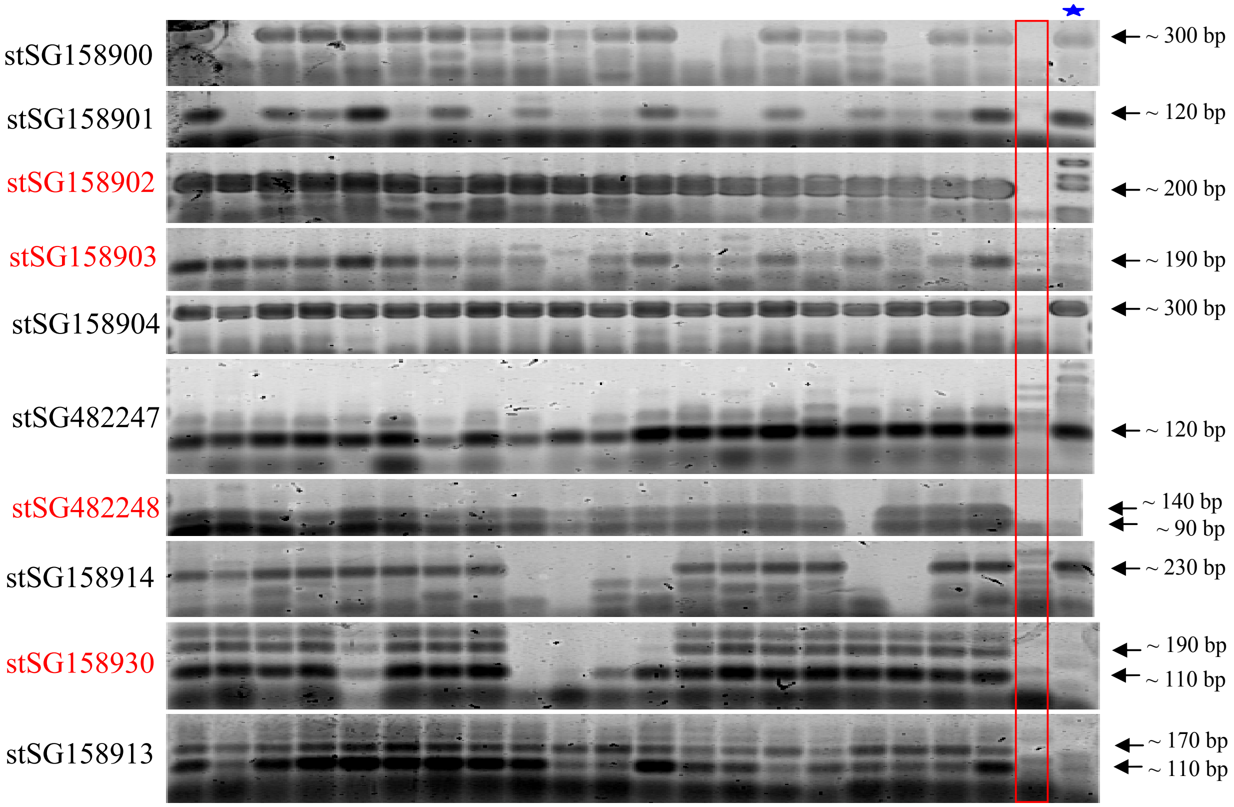
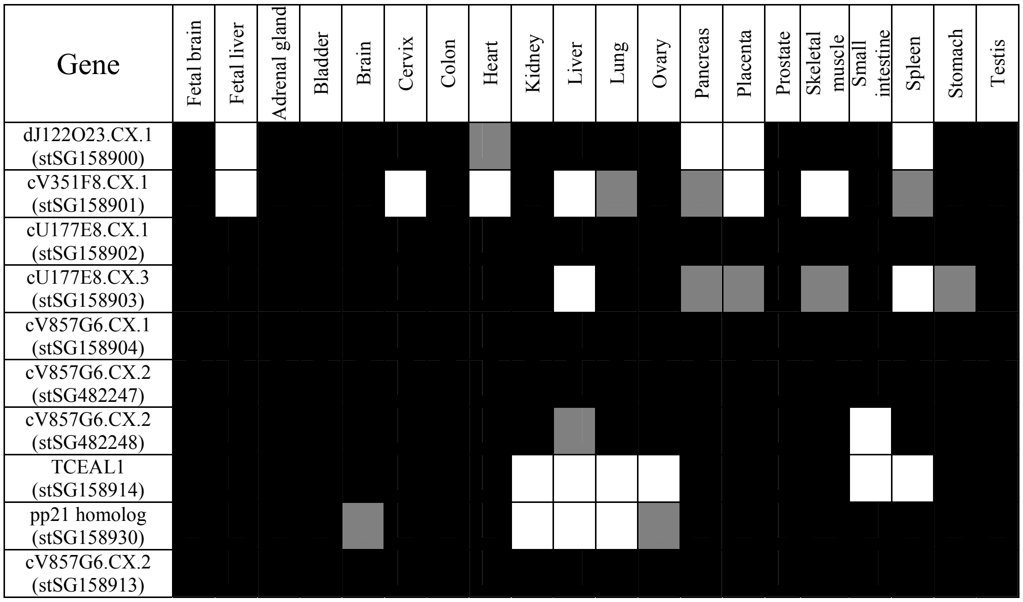


Figure 5-26 Figure illustrating phylogenetic analysis of pp21/TCEAL genes (a) distance-based cladograms constructed using alignments of human and mouse pp21-like genes coding nucleotide sequence. Numbers adjacent to nodes represent the number of bootstrap replicates supporting the adjacent node. Numbers below branches indicate distance. For clarity, only the tree topology is shown and branches are not scaled. (b) maximum-likelihood analysis of the same data used for (a). For technical reasons, gene names have had "." separators removed in the figures, and some genes have had their clone-based prefix ("cV", "dJ", "bM" etc.) removed.

(a)



(b)



Gene	Fetal brain	Fetal liver	Adrenal gland	Bladder	Brain	Cervix	Colon	Heart	Kidney	Liver	Lung	Ovary	Pancreas	Placenta	Prostate	Skeletal muscle	Small intestine	Spleen	Stomach	Testis
dJ122O23.CX.1 (stSG158900)	■	□	■	■	■	■	■	▨	■	■	■	■	□	□	■	■	■	□	■	■
cV351F8.CX.1 (stSG158901)	■	□	■	■	■	□	■	□	■	□	▨	■	▨	□	■	□	■	▨	■	■
cU177E8.CX.1 (stSG158902)	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
cU177E8.CX.3 (stSG158903)	■	■	■	■	■	■	■	■	■	□	■	■	▨	▨	■	▨	■	□	▨	■
cV857G6.CX.1 (stSG158904)	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
cV857G6.CX.2 (stSG482247)	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
cV857G6.CX.2 (stSG482248)	■	■	■	■	■	■	■	■	■	▨	■	■	■	■	■	■	□	■	■	■
TCEAL1 (stSG158914)	■	■	■	■	■	■	■	■	□	□	□	□	■	■	■	■	□	□	■	■
pp21 homolog (stSG158930)	■	■	■	■	▨	■	■	■	□	□	□	▨	■	■	■	■	■	■	■	■
cV857G6.CX.2 (stSG158913)	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■

Figure 5-27 RT-PCR expression profiling of pp21 family genes. Legend as for Figure 5-5. STS names in red type denote instances where intron-spanning primers were used.

5.2.7 RAB-like genes

Within human Xq22, two RAB-like sequences were annotated. No counterparts have been annotated within the orthologous region in *Mus musculus*, however it is possible that Rab-like genes may reside in sequence gaps.

One of the human loci, cU250H12.CX.1 has a three-exon structure drawn from ESTs derived from the same IMAGE cDNA clone, but for cU237H1.1 homology was not sufficiently conserved in the 5' sequence to annotate the two 5' exons. Due to the high level of homology of these genes, attempts to generate cDNA coverage for these genes were not successful using the approaches described in Chapter 3. For loci such as these, targeted approaches must be used (such as exploiting restriction enzyme site differences), unless cDNA coverage is provided from random cDNA sequencing projects. As cU237H1.1 was annotated on the basis of homology in the third exon (containing the ORF in both loci), as no matches were found for the two 5' exons, the possibility remains that the locus may in fact represent a pseudogene.

The high level of sequence homology of these two genes is illustrated in Figure 5-29. This level of homology meant that it was not possible to design locus-specific primers, and so these loci were excluded from expression profile analysis.

Subsequent to studies conducted for this thesis, disruption of a Ras-like gene, termed "RLGP" was described in a patient with Duchenne Muscular Dystrophy (DMD) and mental retardation via an inversion event (Saito-Ohara *et al.*, 2002). RLGP appears to correspond to locus cU237H1.1. In this patient, the breakpoint was 143-145 bp upstream from the putative start codon, and the authors suggest disruption of the gene promoter. Given the three-exon structure of the closely related cU250H12.CX.1, it is possible that instead the inversion disrupts the second intron of the gene. As no RLGP mRNA sequence appears to be reported though (NCBI LocusLink), the possibility remains that involvement of RLGP in this disorder may be erroneous, and that Northern blot analyses reflect cU250H12.CX.1 expression.

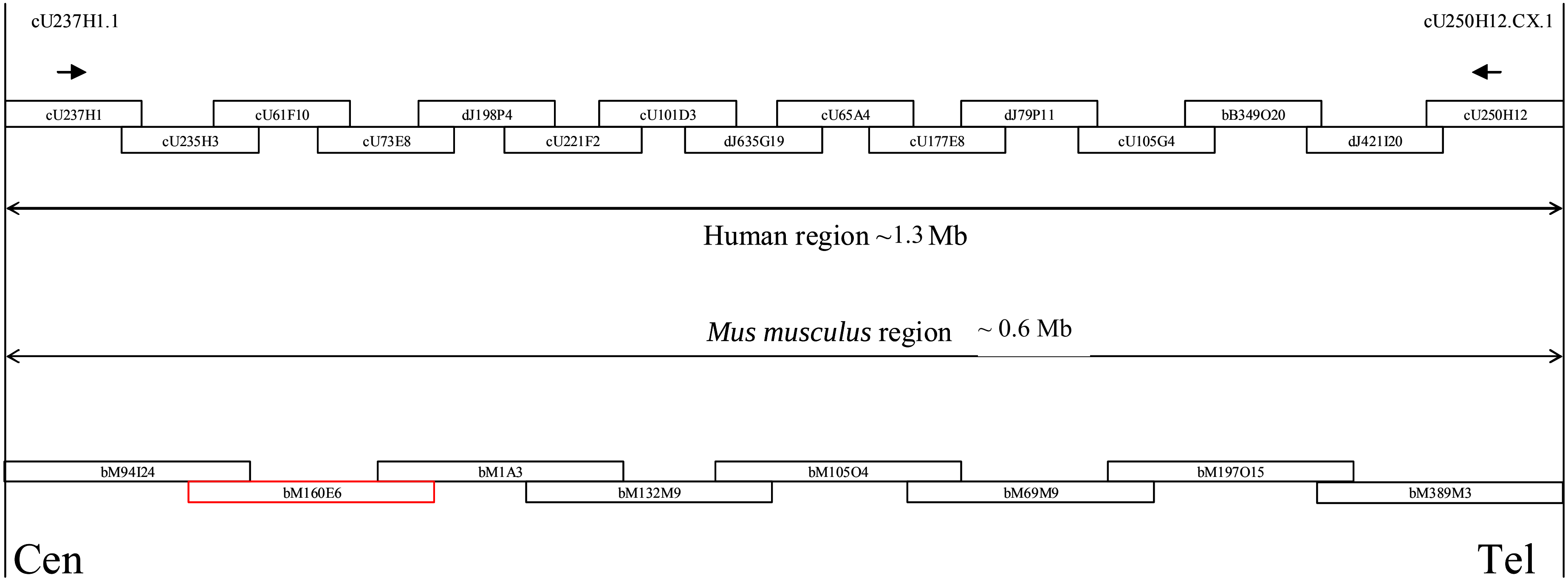


Figure 5-28 Figure showing a schematic representation (not to scale) of RAB-like paralogue gene order and orientation within human Xq22, and the corresponding orthologous region in mouse. Large double-headed arrows depict approximate sizes of the regions, small arrows depict gene positions (names above) and direction of transcription. The genomic clone sequence-ready tile-paths for human and mouse are represented with clone names (not to scale). Clone in red indicates unfinished sequence, and hence a gap in annotation.

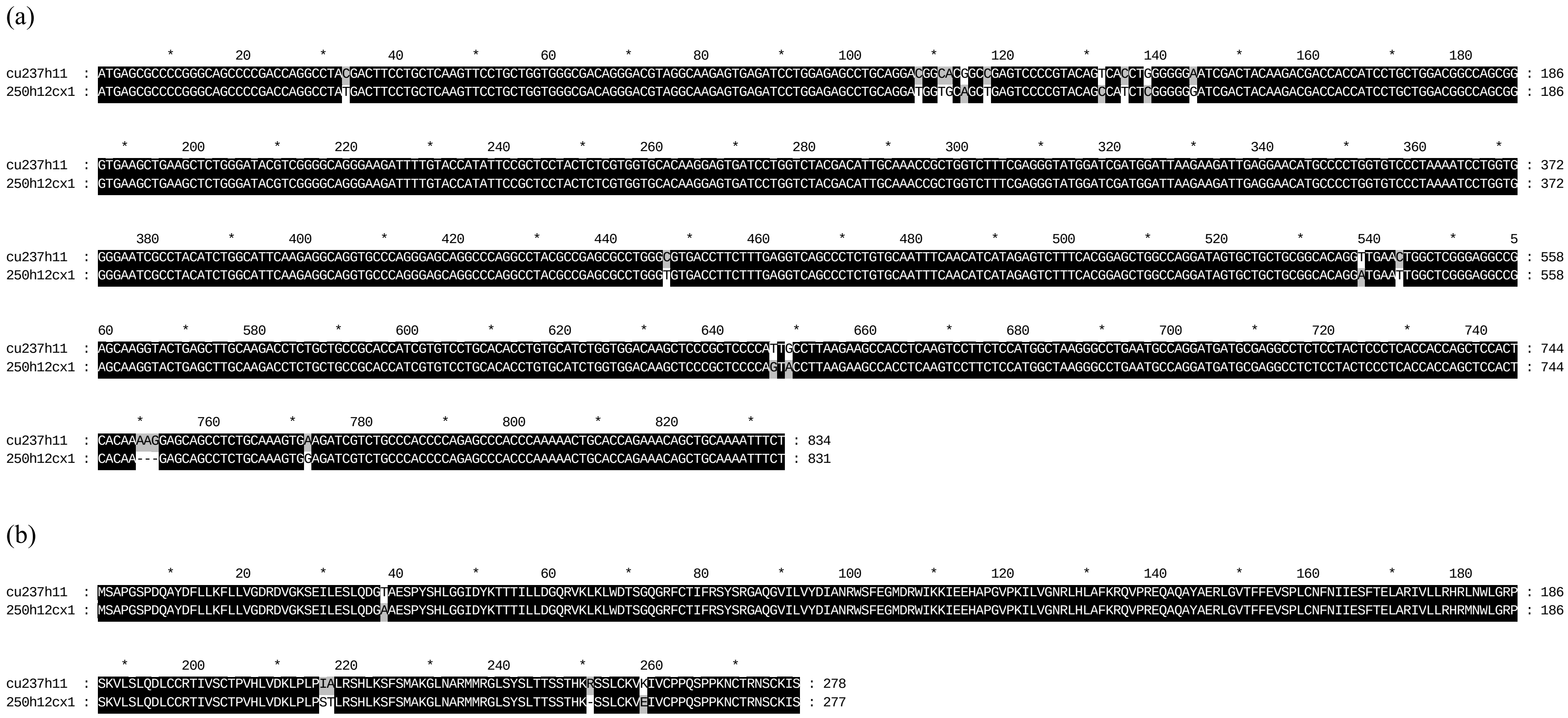

(a)
cu237h11 : ATGAGCGCCCCGGGCAGCCCCGACCAGGCCTACGACTTCCTGCTCAAGTTCCTGCTGGTGGGCGACAGGGACGTAGGCAAGAGTGAGATCCTGGAGAGCCTGCAGGACGGCACGGCCGAGTCCCCGTACAGTCACCTGGGGGAATCGACTACAAGACGACCACCATCCTGCTGGACGGCCAGCGG : 186
250h12cx1 : ATGAGCGCCCCGGGCAGCCCCGACCAGGCCTATGACTTCCTGCTCAAGTTCCTGCTGGTGGGCGACAGGGACGTAGGCAAGAGTGAGATCCTGGAGAGCCTGCAGGATGGTGCAGCTGAGTCCCCGTACAGCCATCTCGGGGGGATCGACTACAAGACGACCACCATCCTGCTGGACGGCCAGCGG : 186
cu237h11 : GTGAAGCTGAAGCTCTGGGATACGTCGGGGCAGGGAAGATTTTGTACCATATTCCGCTCCTACTCTCGTGGTGCACAAGGAGTGATCCTGGTCTACGACATTGCAAACCGCTGGTCTTTCGAGGGTATGGATCGATGGATTAAGAAGATTGAGGAACATGCCCCTGGTGTCCCTAAAATCCTGGTG : 372
250h12cx1 : GTGAAGCTGAAGCTCTGGGATACGTCGGGGCAGGGAAGATTTTGTACCATATTCCGCTCCTACTCTCGTGGTGCACAAGGAGTGATCCTGGTCTACGACATTGCAAACCGCTGGTCTTTCGAGGGTATGGATCGATGGATTAAGAAGATTGAGGAACATGCCCCTGGTGTCCCTAAAATCCTGGTG : 372
cu237h11 : GGGAATCGCCTACATCTGGCATTCAAGAGGCAGGTGCCCAGGGAGCAGGCCCAGGCCTACGCCGAGCGCCTGGGCGTGACCTTCTTTGAGGTCAGCCCTCTGTGCAATTTCAACATCATAGAGTCTTTCACGGAGCTGGCCAGGATAGTGCTGCTGCGGCACAGGTTGAACTGGCTCGGGAGGCCG : 558
250h12cx1 : GGGAATCGCCTACATCTGGCATTCAAGAGGCAGGTGCCCAGGGAGCAGGCCCAGGCCTACGCCGAGCGCCTGGGTGTGACCTTCTTTGAGGTCAGCCCTCTGTGCAATTTCAACATCATAGAGTCTTTCACGGAGCTGGCCAGGATAGTGCTGCTGCGGCACAGGATGAATTGGCTCGGGAGGCCG : 558
cu237h11 : AGCAAGGTACTGAGCTTGCAAGACCTCTGCTGCCGCACCATCGTGTCCTGCACACCTGTGCATCTGGTGGACAAGCTCCCGCTCCCCATTGCCTTAAGAAGCCACCTCAAGTCCTTCTCCATGGCTAAGGGCCTGAATGCCAGGATGATGCGAGGCCTCTCCTACTCCCTCACCACCAGCTCCACT : 744
250h12cx1 : AGCAAGGTACTGAGCTTGCAAGACCTCTGCTGCCGCACCATCGTGTCCTGCACACCTGTGCATCTGGTGGACAAGCTCCCGCTCCCCAGTACCTTAAGAAGCCACCTCAAGTCCTTCTCCATGGCTAAGGGCCTGAATGCCAGGATGATGCGAGGCCTCTCCTACTCCCTCACCACCAGCTCCACT : 744
cu237h11 : CACAAAAGGAGCAGCCTCTGCAAAGTGAAGATCGTCTGCCCACCCCAGAGCCCACCCAAAAACTGCACCAGAAACAGCTGCAAAATTTCT : 834
250h12cx1 : CACAA---GAGCAGCCTCTGCAAAGTGGAGATCGTCTGCCCACCCCAGAGCCCACCCAAAAACTGCACCAGAAACAGCTGCAAAATTTCT : 831
(b)
cu237h11 : MSAPGSPDQAYDFLLKFLLVGDRDVGKSEILESLQDGTAESPYSHLGGIDYKTTTILLDGQRVKLKLWDTSGQGRFCTIFRSYSRGAQGVILVYDIANRWSFEGMDRWIKKIEEHAPGVPKILVGNRLHLAFKRQVPREQAQAYAERLGVTFFEVSPLCNFNIIESFTELARIVLLRHRLNWLGRP : 186
250h12cx1 : MSAPGSPDQAYDFLLKFLLVGDRDVGKSEILESLQDGAAESPYSHLGGIDYKTTTILLDGQRVKLKLWDTSGQGRFCTIFRSYSRGAQGVILVYDIANRWSFEGMDRWIKKIEEHAPGVPKILVGNRLHLAFKRQVPREQAQAYAERLGVTFFEVSPLCNFNIIESFTELARIVLLRHRMNWLGRP : 186
cu237h11 : SKVLSLQDLCCRTIVSCTPVHLVDKLPLPIALRSHLKSFSMAKGLNARMMRGLSYSLTTSSTHKRSSLCKVKIVCPPQSPPKNCTRNSCKIS : 278
250h12cx1 : SKVLSLQDLCCRTIVSCTPVHLVDKLPLPSTLRSHLKSFSMAKGLNARMMRGLSYSLTTSSTHK-SSLCKVEIVCPPQSPPKNCTRNSCKIS : 277

Figure 5-29 Figure illustrating alignments of RAB-like paralogues. Alignment (a) shows the high level of similarity between the nucleotide sequences. Alignment (b) shows that many of the differences seen are synonymous, and that the predicted peptides are very similar. For technical reasons, gene names have had "." separators removed in the figures, and some genes have had their clone-based prefix ("cV", "dJ", "bM" etc.) removed.

5.2.8 Histones and cU46H11.CX.1/cU116E7.CX.1 genes

Five histone H2B-like loci were annotated within human Xq22. In contrast, only two loci were found in the orthologous region in *Mus musculus*. Within the human region containing the histone paralogues, two other pairs of paralogous genes were found. These paralogue pairs (cU116E7.CX.2/cU46H11.CX.2 and cU116E7.CX.3 /cU46H11.CX.1) appear to have been generated from an inverted duplication within the human lineage of a segment containing the two genes in a head-to-head configuration (see Figure 5-30), presumably also containing two histone loci also, represented by dJ839M11.1/dJ839M11.2 and cU240C2.1/cU240C2.2. This is also consistent with the closer sequence similarity between the human loci compared to the mouse loci.

Consistent with a reduced number of histone genes found within the orthologous region in *Mus musculus*, only a single copy each of the non-histone genes were found. Based on orientation evidence, these appear to be orthologous to cU46H11.CX.1 and cU46H11.CX.2. As there are inconsistencies in the positioning and orientation of the histone genes between the two species though, more complex rearrangements (including independent duplications in each species) may have shaped this sub-region.

As can be seen in Figure 5-31, sequence homology between the cU46H11/cU116E7 paralogues is high for both pairs. Both pairs are also more similar to one another within human than to their mouse counterparts. This made design of suitable RT-PCR primers difficult. However, results were generated for cU46H11.CX.1, which suggest that this locus may show restricted expression as it was only detected in testis and adrenal gland.

Locus cU116E7.CX.3 appears to have a frameshift mutation in the predicted ORF, compared to cU46H11.CX.1, whose predicted peptide shows homology to mitochondrial carrier proteins. This may indicate that cU116E7.CX.3 could be an expressed pseudogene. As seen in Chapter 3, cDNA sequence generated by an STS designed to cU46H11.CX.1 mapped to cU116E7.CX.3, and cU46H11.CX.1 is annotated by homology. Loci cU116E7.CX.2 and cU46H11.CX.2 show homology to a paraneoplastic neuronal antigen gene, MA3. An additional paralogue related to these genes is found in human Xq24 (Gareth Howell, PhD thesis, Open University). Several PNMA genes have been reported, and two are mapped to Xq28 (NCBI LocusLink).

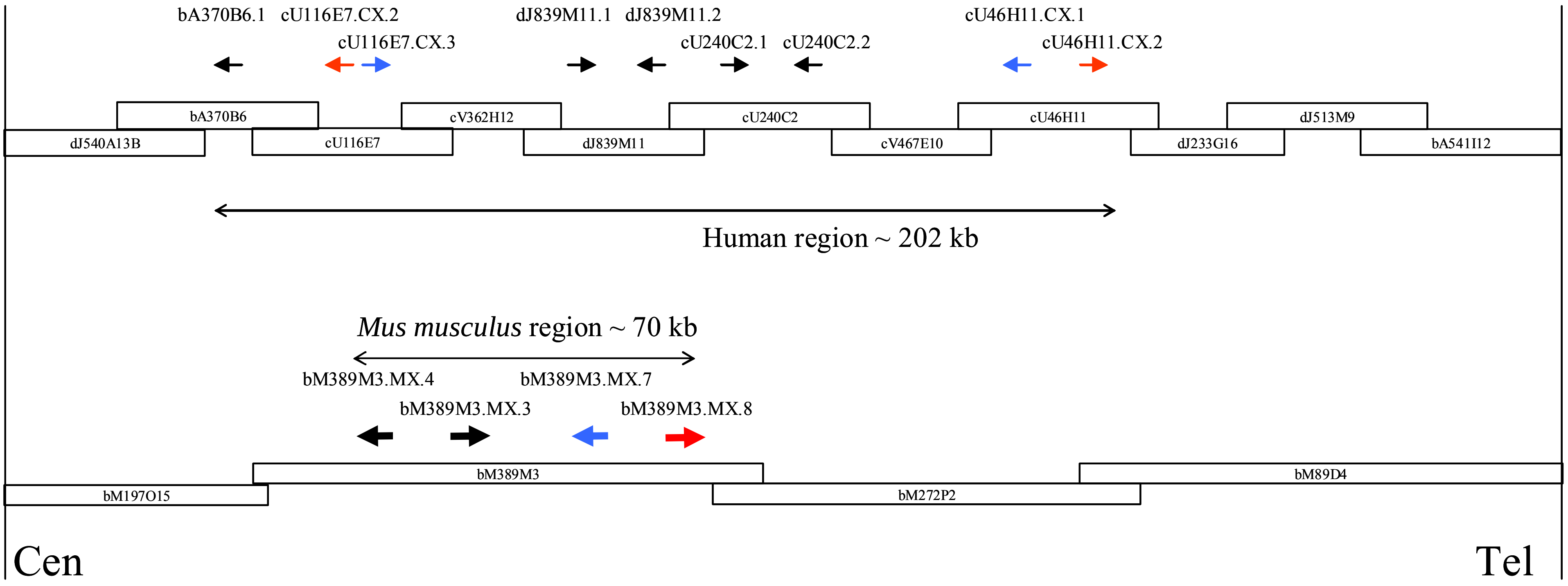


Figure 5-30 Figure showing a schematic representation (not to scale) of histone and cU46H11.CX.1/cU116E7.CX.3 paralogue gene order and orientation within human Xq22, and the corresponding orthologous region in mouse. Large double-headed arrows depict approximate sizes of the regions, small arrows depict gene positions (names above) and direction of transcription. The genomic clone sequence-ready tile-paths for human and mouse are represented with clone names (not to scale). The red and blue arrows indicate genes sharing homology between human and mouse. The black arrows depict histone genes.

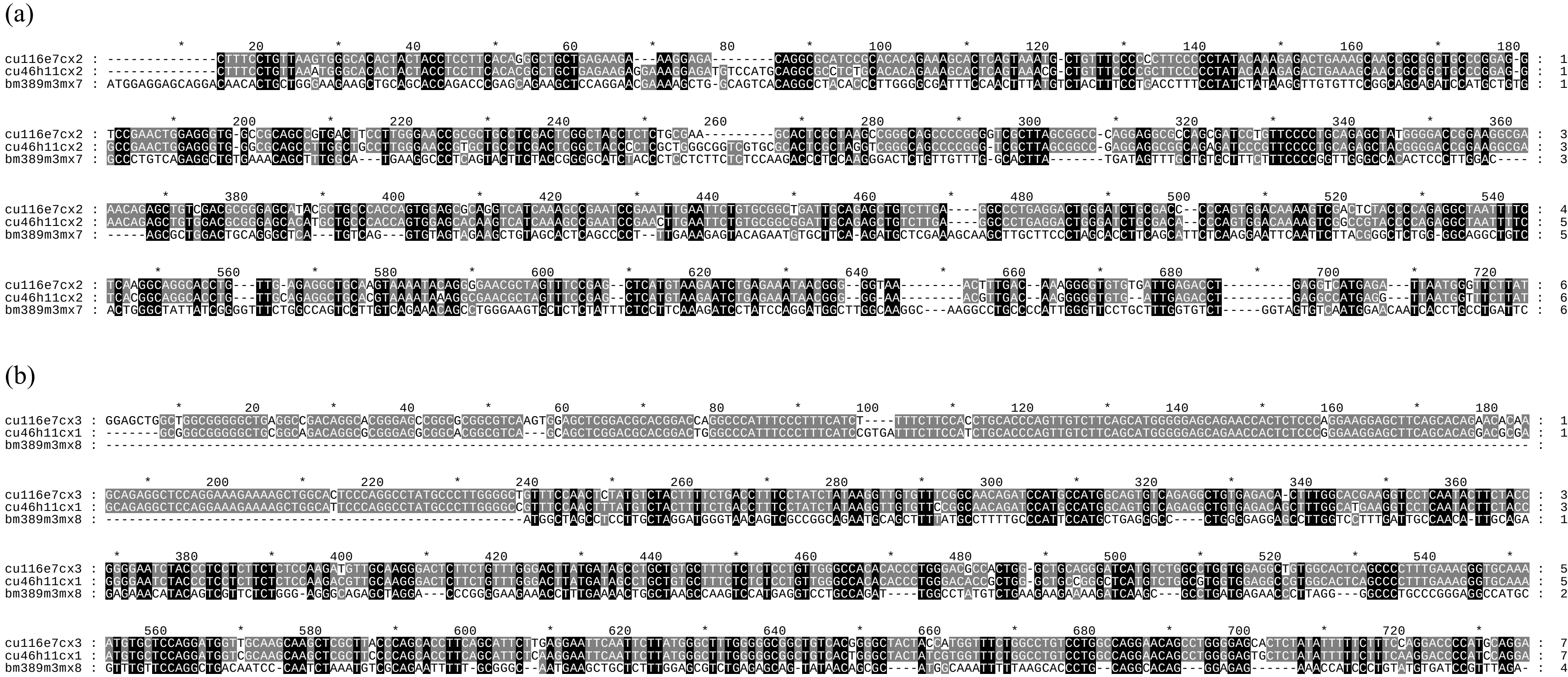
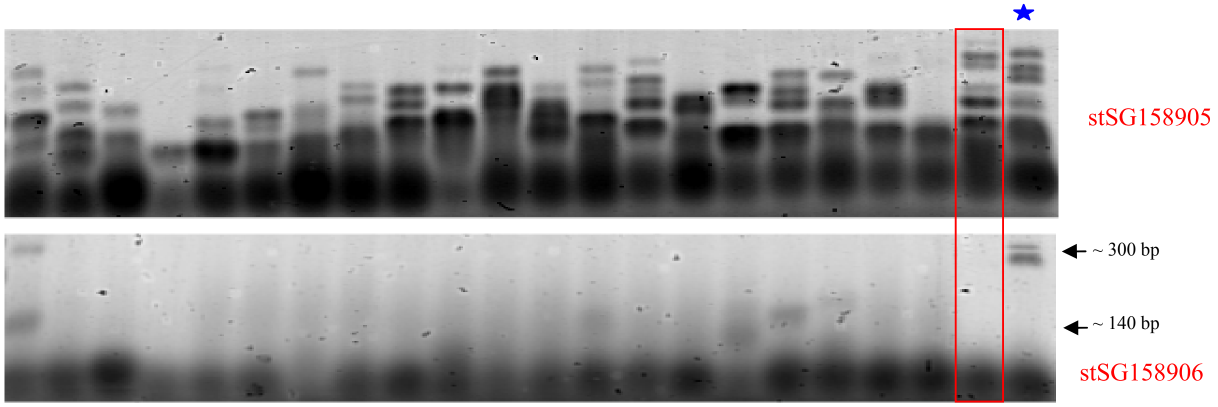


Figure 5-31 Figure illustrating alignments of cU46H11/cU116E7 genes. Alignment (a) shows the high level of nucleotide sequence conservation seen between the cU116E7.CX.2 and cU46H11.CX.2 genes. Alignment (b) shows the high level of nucleotide sequence conservation also seen between the cU116E7.CX.3 and cU46H11.CX.1 genes. For technical reasons, gene names have had "." separators removed in the figures.

(a)



(b)

Adrenal gland	Bone marrow	Brain (cerebellum)	Brain (whole)	Fetal brain	Fetal liver	Heart	Kidney	Liver	Lung	Placenta	Prostae	Salivary gland	Skeletal muscle	Spleen	Testis	Thymus	Thyroid gland	Trachea	Uterus	Gene
▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	cU116E7.CX.3 (stSG158905)
■															■					cU46H11.CX.1 (stSG158906)

Figure 5-32 RT-PCR expression profiling of cU116E7.CX.3 and cU46H11.CX.1. Legend as for figure 5-5. Hatched cells denote instances where sample being omitted or lack of clarity of PCR products leads to uninterpretable results. STS names in red type denote instances where intron-spanning primers were used.

5.2.9 Duplicated pseudogenes, TEX genes and COL4A5/COL4A6

These genes are discussed together as they are more widely distributed across the region with respect to the other paralogues. Within the human Xq22 region, COL4A5 and COL4A6 are well characterised genes that probably arose by gene duplication. Also within the region are two paralogous genes unrelated to the COL4A genes, TEX13A and TEX13B, as well as two paralogous sequences that appear to represent pseudogenes (similar to sequence AF116646 – PRO0082). TEX13A is unusual in that it resides within the first intron of the IL1RAPL2 gene, on the opposite strand.

No systematic search for these genes in the orthologous region in *Mus musculus* was employed, as they lay in regions where some sequence gaps remained.

The COL4A5 and COL4A6 genes are positioned in a head-to-head configuration in Xq22, and have been well characterised. This is particularly so of COL4A5, mutations in which cause Alport's syndrome (OMIM:301050). The PRO0082 gene is mapped to 10q11 (LocusLink) and is also known as GALNACT-2, encoding a protein involved in the synthesis of chondroitin sulphate. The pseudogenes in Xq22 may have arisen from retrotransposition, and subsequent tandem duplication. The TEX genes were discovered during a study to identify genes preferentially expressed in mouse spermatogonia (Wang *et al.*, 2001), and are highly related at the nucleotide level as seen below. The authors found that there was a bias of location of these genes to the Y and, in particular, X chromosomes.

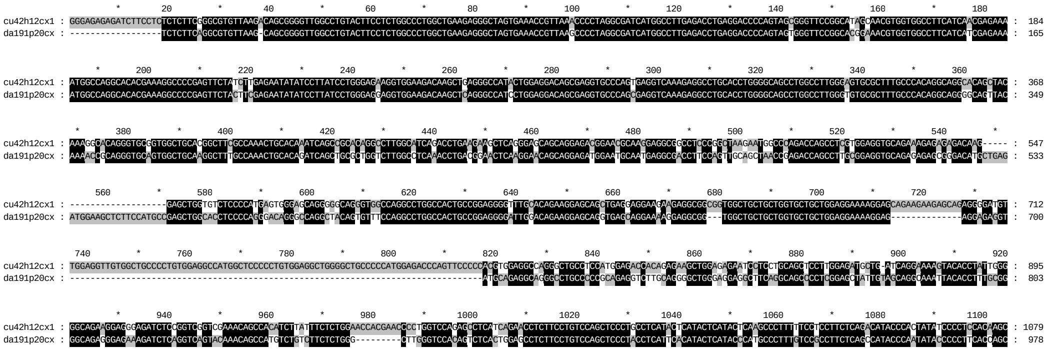


Figure 5-33 Partial alignment of TEX mRNA sequences.

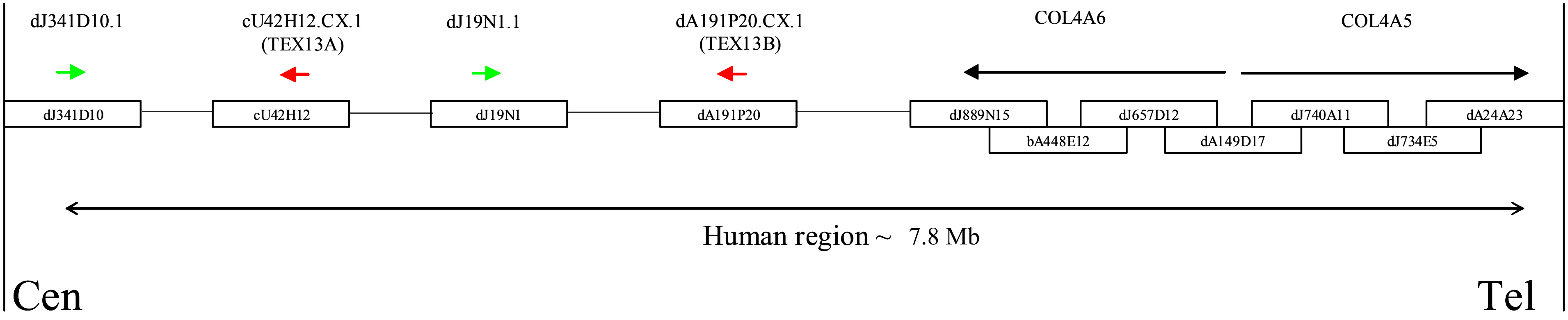


Figure 5-34 Figure showing a schematic representation (not to scale) of pseudogene, TEX and COL4A5/COL4A6 paralogue gene order and orientation within human Xq22. Large double-headed arrows depict approximate sizes of the regions, small arrows depict gene positions (names above) and direction of transcription. The genomic clone sequence-ready tile-path is represented with clone names (not to scale). The lines connecting some of the clones represent large regions of sequence which are not depicted for clarity. Green arrows represent the pseudogenes, red arrows the TEX genes and black arrows the COL4A5/COL4A6 genes.

5.3 Discussion

The studies presented in this Chapter have described in more detail the extensive gene paralogy within Xq22 noted in earlier chapters. Fourteen families of paralogous genes have been described. The number of paralogues contained within each family ranges from two to ten genes. The paralogous loci include both expressed genes predicted to encode peptides and also apparent pseudogenes. The level of paralogy within Xq22.1-q22.3 is higher than that seen in neighbouring regions of the chromosome (Xq21 and Xq23), and may reflect underlying features of the region's sequence that predispose it to duplication and deletion events.

Expression analyses have revealed that many of the Xq22 paralogues appear to be widely expressed. However some differences in paralogue expression have been noted. For genes which were ubiquitously expressed in the tissues studied, a more extensive approach would be required to reveal any quantitative, temporal and spatial differences in expression pattern for these genes.

Whilst providing some useful information regarding the specificity of different primer pairs, the colony PCR-based approached employed should ideally be complemented by other methods. In the expression analyses, the correct discrimination between the loci is of paramount importance. For example, for the STSs which amplified multiple paralogues, restriction enzyme sites were identified (although were not used due to time constraints) that could be used to digest RT-PCR products and thus discriminate between transcripts from different loci. This type of assay may be particularly useful in instances where paralogues are physically close together, and colony PCR of large-insert clones is less likely to be a successful assay. A further alternative could be to sub-clone a large-insert clone and verify the different sub-clones by partial sequencing, followed by colony PCR. A simpler approach would to sub-clone RT-PCR products and analyse them by sequencing. Some loci however, such as the two NXF2 paralogues described in Chapter 3, are refractory to any of these approaches due to extremely high sequence identity and represent instances where expression data generated must be considered a composite of the loci.

Whilst some of the duplicated genes appear to be pseudogenes, most appear to be expressed and presumably functional. The levels of paralogue homology vary between families, with some copies appearing almost identical (e.g. TMSNB family), and others showing great diversity and only displaying conservation in some parts of the protein sequence (e.g. ALEX family). Some of this is likely to be due to differences in ages of the duplication events. However, an intriguing feature of some of the duplicated genes is the higher level of homology for some paralogues within human or mouse as compared to between the species. This could suggest independent duplication events within each lineage. In some instances, though, the similarity of transcription direction and gene order between human and mouse would require very similar duplications, which could be a result of shared sequence features in the regions. However, an alternative explanation in such cases could be the occurrence of gene conversion maintaining homogeneity between the genes within a species.

Finally, availability of functional information for some members of different families has enabled predictions to be made regarding the functions of the rest of the family members. This will provide useful information for the cloning of genes involved in diseases mapped to the region, such as mental retardation and deafness. The mapping information information collated in this chapter also allows inferences to be made of orthologous relationships between human and mouse genes where sequence similarity alone may be misleading, if gene conversion is operating. Some of the species-specific differences shown here can also now be taken into account in any studies aimed at elucidating the functions of the genes involved.