Gene identification and characterisation on the human X chromosome

Ian Barrett

Darwin College, University of Cambridge

January 2004

This dissertation is submitted for the degree of Doctor of Philosophy.

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

This dissertation does not exceed the size limit for the Biology Degree Committee.

Abstract

This thesis investigates the gene composition and evolution of regions of the human X chromosome, including data from comparative genome analysis of other organisms.

Chapter Three presents studies undertaken to annotate genes within the Xq22-q23 region of the human X chromosome. Selected features of the region are discussed, including investigation of alternative polyadenylation site usage, an insertion of the mitochondrial genome into the nuclear genome, and an inverted duplication and potential gene fusion event involving the NXF2 and TCP11-like genes.

As a result of the annotation described in Chapter Three, extensive paralogy within the Xq22 region was discovered, along with additional examples of paralogy between Xp and Xq22-q23. Work in subsequent Chapters attempted to characterise these aspects further and provide information on the evolution of the regions.

Chapter Four describes work undertaken to map and sequence the region of the mouse genome corresponding to human Xq22-q23, in order to investigate the evolution of the Xq22 paralogues. Annotation of genes within the region of the mouse X chromosome and the orthology of the human and mouse regions is described. Features of the mouse region, such as the presence of two large repeat families, are also discussed.

Chapter Five presents phylogenetic and expression profile analysis of the Xq22 paralogues, and examines orthology in the corresponding region of the mouse genome.

Chapter Six includes a discussion of Xp/Xq paralogy, and presents studies providing evidence for a segmental duplication leading to this paralogy. In addition, orthologues of the genes involved in the paralogy are identified in the marsupial mouse, *Sminthopsis macroura,* and their genomic localisations determined. Evidence suggesting a minimum age of the duplication is presented. Comparative analysis of human, mouse, *Fugu rubripes* and *Sminthopsis macroura* genomic sequence is described.

Acknowledgements

I don't really know how to begin thanking everyone. I've been extremely fortunate to have benefited from the kindness, help, advice and expertise of so many people during these studies, and have had great times along the way (as well as numerous genome T-shirts). The selfless help I've received from so many people really has been very touching and humbling.

Firstly, thanks to Tim, David and Andy for your support and understanding over the last months. Thank you also to the BGS and Education Committee for their support. Financial support from the MRC and The Wellcome Trust is gratefully acknowledged.

A huge thanks to all the Sanger Institute cloning, sequencing and finishing teams for all their work spanning sequence from human, mouse and that furry little fella, the marsupial mouse – without you none of this work would have been possible. Thanks especially to Kirsten McLay, Darren Grafham, Christine Bird, Siobhan Whitehead and Lucy Matthews for answering so many of my queries. Thanks as well to Sean Humphray and Glen Threadgold for all the help with the mouse mapping. A big thanks to Paul Hunt for putting up with me and for all your gridding and being generally very helpful, and to Carol Carder for showing me the ropes, and letting me upset the carefully balanced workings of your lab! Thanks as well to the now HAVANA group, for the sequence analysis, especially Stephen Keenan, whom I badgered on many occasions.

I'd also like to thank Jess Tyson, Karen Woodward and Willem Rens for our collaborations, which I enjoyed very much, and from which I learned a great deal. A particular thanks to Willem for the gifts of resources and advice, and your infectious enthusiasm!

Thank you Jackie Bye and Susan Rhodes for all the work you put into the cDNA resources, and especially to you Jackie for all the advice and encouragement, and our discussions! I'm watching out for that next invention… Thank you to John Collins and David Beare for all the advice on the transcript mapping, and for paving the way.

A very big thanks to all the FISH gang and Nigel Carter. I had a great time as an "imposter" in the lab. Thanks very much to you all for your help in times of need in getting things done. I won't forget my last day at the Sanger, checking out some FISH images (nor my special FISH-cake Louisa….!). Thanks particularly to Ruby Banerjee, Susan Gribble and especially to Deb Burford for really coming through when things were tight (hope I can repay you in some way in your PhD studies).

Thanks to Christina Hedberg-Delouka and Christine Rees for keeping me in check (as much as you could…) and also to Joan Green and Andrew King for help and advice on all things library-related (and sorry for the often late returns..!). Thanks to Richard Summers as well for all the reprographics support and photos for my leaving gifts.

Thanks to Kate Rice and David Bentley for the chats and advice, I hope HumGen continues to go from strength to strength. I feel very lucky to have been part of it. Thank you to David Vetrie, Stephan Beck and Nabeel Affara for helpful advice. A big thanks to Cords, Matt, Rob and Em for being good mates and also to Paw and Vicki for a lot of laughs and chats. Thanks to "Tammar" Tamsin as well for all the encouragement and chats, and congratulations on your rowing achievements. Can ya tell wot it is yit? "Frantastic" Fran Lovell and Christine Burrows, I especially owe you guys a great deal in all the help you gave me when times were tight (and even when they weren't). Thanks also to Adam Whittaker for all the hyb jokes, and Jen Conquer, Helen Steingruber and Barbara Gorick for all the office banter! Thanks very much to “Pod’s Angels” Hazel Arbery and Ruth Bennett as well, for your real efforts in the last push. You’re real stars.

Carol Scott, I bet you’ll be glad I’ve finished! I can’t thank you enough for all the remote-control help you’ve given me, and thanks also to you and Sarah Hunt for all the help while I was actually a genuine member of the Sanger Institute.

Gareth “Gazza” Howell, thanks for all you’ve taught me and for all the clowning around, although I’m glad to be able to walk the corridors without apprehension…. On a more serious note, thanks for all the work you put into Xq22, which formed the basis for this work. Thanks as well to Simon Gregory for the help, advice and laughs, just a shame you weren’t around for the rugby this year…! Thanks both for all your advice and friendship, and for all the good times (plenty more to come).

Pod (a.k.a. Alison Coffey), you really have been a star. I’ve learned so much from you and you really have been an inspiration in more ways than one. Thanks for everything, and especially for all the help in the final push – definitely couldn’t have done it without you.

Mark Ross, what can I say? You’ve taught me a lot over the last years, more than you’ll probably realise and not just good science. You’re an excellent scientist to learn from and a great mentor. Thanks very much from both of us for all you’ve done.

People often say they couldn’t have done it without you etc. but in this case especially I really couldn’t. I’ve been helped by, and learned from, more people than I’ve been able to mention here. I’ll always be grateful to everyone (you and I know who you are) for all your help, and for all the fantastic memories I’ve got from my time at the Sanger, a very special place.

And now thanks to all our family and friends for all their support over the years and getting me this far. Thanks to Andy for the chats and to Mark for being a good mate.

Thank you especially to my long-suffering wife Gill (well, not long yet but pretty suffering) for all your support and love over good times and bad and for all your help, especially your help with the final printing and formatting. You’ve really held things together over the last year, which has truly tested us both. Can’t wait to get on with the rest of our married life, and see what it holds for us.

Thank you to Dad for being understanding when I didn’t feel like giving an update on progress and for all your love, support, interest and encouragement. You really are someone to look up to.

And finally, a very special thanks to Mum for all that you’ve given and taught me. There just isn’t enough space to put it all here. You’re an incredible inspiration and very much loved. I’d like to dedicate this work to you and Dad. Sleep well.

Table of Contents

List of Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

List of Tables

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

Abbreviations

μg	microgram
μl	microlitre
μM	micromolar
μm	micrometre
^{0}C	degrees celsius
ACeDB	A *C. elegans* Database
AT	annealing temperature
BAC	bacterial artificial chromosome
BLAST	basic local alignment search tool
Bp	base pair
cDNA	complementary deoxyribonucleic acid
CEN	centromere
cm	centimetre
cM	centimorgan
CpG	cytidyl phosphoguanosine dinucleotide
Ctg	contig
DDC	duplication degeneration complementation
DDW	double distilled water
DMD	duchenne muscular dystrophy
DNA	deoxyribonucleic acid
EBI	European Bioinformatics Institute
EST	expressed sequence tag
FISH	fluorescence *in situ* hybridisation
g	gram or force of gravity
HAVANA	Human And Vertebrate ANalysis and Annotation
HTGS	high throughput genomic sequence
HUGO	Human Genome Organisation
kb	kilobase pairs
LCR	low copy repeat
LD	linkage disequilibrium
LINE	long interspersed nuclear element
LTR	long terminal repeat
M	molar
Mb	megabase pairs
MHC	major histocompatability complex
MIR	medium interspersed repeat
mL	millilitre
mM	millimolar
mm	millimetre
mRNA	messenger ribonucleic acid
Mya	million years ago
NCBI	National Centre for Biotechnology Information
ng	nanogram
nm	nanometre
numts	nuclear mitochondrial insertions
OMIM	Online Mendelian Inheritance in Man
ORF	open reading frame
PAC	P1-derived artificial chromosome

PAR	pseudoautosomal region
PBS	phosphate-buffered saline
PC	personal computer
PCR	polymerase chain reaction
RACE	rapid amplification of cDNA ends
RFLP	restriction fragment length polymorphisms
RH	radiation hybrid
RNA	ribonucleic acid
rpm	revolutions per minute
RT	reverse transcriptase
RT-PCR	reverse transcriptase-polymerase chain reaction
SINE	short interspersed nuclear element
SNP	single nucleotide polymorphism
SSPCR	single-sided specificity polymerase chain reaction
STS	sequence tagged site
TEL	telomere
U	units
UTR	untranslated region
UV	ultra-violet
V	volts
VEGA	VErtebrate Genome Annotation
WGS	whole-genome shotgun
XAR	X added region
XCR	X conserved region
Xic	X inactivation centre
XIST	X inactivation specific transcript
YAC	yeast artificial chromosome

Chapter One - Introduction

1 Introduction

1.1 Genome mapping and sequencing

1.1.1 Genomes

The genome of a cell contains all of the information necessary for the cell's function, although that function is shaped and facilitated by its environment. The genome remains the ultimate determinant of a cell or organism, because no matter how the information encoded within it is interpreted, only the genomic sequence contains all of the instructive components. Following the discovery of DNA as the inherited material within a cell, the discovery of the structure of DNA in 1953 by Watson and Crick, and the subsequent efforts of many scientists, led to rapid advances in our understanding of the molecular basis of genetic inheritance. Whilst retroviruses, for example, utilise RNA rather than DNA as their genomic material, the central concept of the genome containing all of the information required for a cell's function remains intact.

Although the concept of a genome is well established, the diversity of genome structure and organisation displayed by different organisms is remarkable. Genomes can show dramatic variation in sizes between organisms, as well as in their gene complement. For example, the genome of the SV40 virus is ~ 5 kb, and contains ~ 6 genes, but the human genome is ~3000 Mb and contains ~30-40,000 genes. *Mycoplasma genitalium*, a prokaryote with one of the smallest genomes known, is thought to approach a "minimal genome" of ~0.6 Mb, containing ~503 genes. A striking 85-90% of its genome is coding. In contrast, the genomes of some plants and amphibians can be of the order of 10^5 Mb in size.

Genome organisation also shows large variations between organisms. Some viruses for example, utilise overlapping genes to increase the coding capacity of their genomes, and the use of introns in eukaryotic but not prokaryotic genes is a clear distinction. Prokaryotic genomes comprise circular double-stranded DNA, compacted into a nucleoid structure; in comparison eukaryotic genomes are linear double-stranded DNA molecules, and are tightly packed into chromatin within a nucleus. The ciliated protozoa, an example of which is *Tetrahymena thermophila*, are remarkable in dividing their genome between two distinct nuclei to form a macronuclear and micronuclear genome. The micronuclear genome undergoes extensive rearrangements in producing the macronuclear genome, from which transcription occurs. These examples of complex genome dynamics exemplify our need to further explore the variety of genome dynamics in different organisms.

Within phyla however, there can be significant variation in genome size. Gene number does not seem to increase dramatically with organismal complexity. For example, the human genome is expected to contain only approximately twice the number of genes found in the nematode *Caenorhabditis elegans*. The lack of understanding of relationships between genome size and content and organismal complexity reflects our limited knowledge of gene organisation and regulation, genome dynamics and an often anthropocentric view of evolutionary relationships. The availability of genome sequences for many different organisms will undoubtedly shed light on these aspects of biology.

1.1.2 Genome mapping and sequencing

The term "genome" reached iconic status during the latter part of the 20th Century. From the elucidation of the first genome sequence, that of the bacteriophage *phiX*174 virus in 1977 (Sanger *et al.*, 1978), to the completion of the finished human genome sequence by a consortium of research centres in 2003, the field of genome mapping and sequencing has advanced substantially.

The availability of genomic sequence for an organism is useful for many studies, such as gene identification, genetic trait mapping and the study of genome evolution. This has been illustrated by research involving organisms for which genome sequence has been available for some time, such as *Saccharomyces cerevisae*, *Drosophila melanogaster* and *Caenorhabditis elegans*, as well as many microbial organisms and viruses.

Before genome sequencing was considered practical, mapping and sequencing of the genome was on a smaller scale. These approaches involved isolating clones containing segments of genomic sequence, usually through hybridisation-based approaches. These studies were often motivated by evidence suggesting the location of a gene of interest in a particular region of the genome, based on linkage of traits or diseases to genetic markers placed on genetic maps.

Genetic maps are more readily generated in model organisms such as yeast, flies, worms and even mice, due to the ability to generate genetic crosses and follow genes. In comparison, in humans genetic marker availability at a reasonably high resolution was not available until relatively recently. Also, in model organisms selective breeding can be used to produce informative genetic crosses. Early markers included protein isozymes and restriction fragment length polymorphism (RFLP) markers. These provided limited resolution but were nevertheless useful, and led to the production in 1987 of the first genetic map of the whole human genome (Donis-Keller *et al.*, 1987).

The subsequent availability of highly polymorphic microsatellite markers greatly increased genetic mapping power in humans, and in 1992 a second-generation genetic map of the human genome was produced using this type of marker (Weissenbach *et al.*, 1992). Further improvements in screening techniques have facilitated further refinement of genetic maps (Kong *et al.*, 2002).

These maps were developed in concert with refinements in physical mapping approaches. Initial restriction mapping, RH-mapping and cytogenetic approaches were improved by the advent of the polymerase chain reaction (PCR) which enabled clone contig maps to be produced using PCR-based STS-content approaches. In addition, restriction fingerprinting approaches aided the production of mapped clone substrates for genomic sequencing.

Early whole-genome mapping approaches used cosmid and phage genomic clones but their small size rendered them impractical for the mapping of larger genomes, driving improvements in cloning technologies, initially resulting in the advent of yeast artificial chromosomes (YACs). The first human genome YAC map was produced in 1993 (Cohen *et al.*, 1993). YAC clones are however limited by their propensity for chimaerism and unsuitability for sequencing. Although not suitable for providing sequencing substrates, the STS-content of YAC maps nevertheless provided ordered sets of markers for the production of bacterial clone maps (e.g. the YAC map of chromosome 22 (Collins *et al.*, 1995)).

Subsequent clone-based approaches used the smaller, more stable, cosmid (Collins and Bruning, 1978) and P1 artificial chromosome (PAC) genomic clones propagated in bacteria (Sternberg, 1992). Later, larger insert bacterial artificial chromosome (BAC) clones increased mapping efficiency further and were shown to be remarkably stable, presumably on account of their low copy number (Shizuya *et al.*, 1992). BACs were ultimately adopted as the substrate of choice for genome sequencing. Developments in restriction fingerprinting techniques and increasing availability of STS markers further aided progress.

As sequencing methodologies improved and the density of markers placed on genetic and physical maps increased, the potential to sequence large genomes was explored. The first microbial genome to be sequenced was that of *Haemophilus influenzae* in 1995. The genome of this organism was sequenced using a whole-genome random sequencing strategy (Fleischmann *et al.,* 1995), also called a whole-genome shotgun (WGS) strategy. The first eukaryotic genome to be sequenced was that of *Saccharomyces cerevisae*, using a hierarchical shotgun strategy (Bussey *et al.*, 1997). The same strategy was also applied to the sequencing of the genome of *Caenorhabditis elegans*, the first multicellular organism to have its genome sequence determined (The *C. elegans* Sequencing Consortium, 1998). Publication of the genome sequence of *Drosophila melanogaster* in 2000 demonstrated the use of the whole-genome shotgun approach in the sequencing of complex genomes (Adams *et al.*, 2000).

1.1.3 The human and mouse genome projects

In 2001, two draft versions of the human genome were published, as well as BAC maps covering the different chromosomes (Lander *et al.*, 2001; Venter *et al.*, 2001). The physical maps covered more than 96% of the euchromatic genome and the draft sequence covered approximately 94% of the genome. At the time of writing in 2003, the finished sequence of the human genome was announced.

Debate over the strategies used to sequence the human genome, map-based clone sequencing versus whole-genome shotgun has been much reported, with the publicly-funded effort adopting the map -based clone sequencing approach and a private company, Celera Genomics formed some time later to adopt a whole-genome shotgun approach. Whilst offering the potential to provide information on the value of each approach in sequencing large, complex and repetitive genomes, the ability to make these comparisons was somewhat hampered by the commercial nature of the privately-funded approach.

A clone-based (hierarchical shotgun) sequencing strategy provides information regarding the positioning of the individual clones, information which can prove valuable in studies of the genome. In addition, in instances where short but very highly homologous repeats occur, clone-based mapping can resolve these regions where the clones also contain some unique sequence, unlike a WGS approach in which the repeats would collapse. Sequencing of large genomic clones also allows the process of finishing to be performed, where ambiguities can be resolved. This is impractical with WGS-only approaches.

Whole-genome shotgun approaches can however provide data where there may be gaps in coverage in a clone library, and for regions that prove difficult to clone in larger segments. They also do not require investment in preliminary mapping studies and so can provide rapid sequence coverage of an organism's genome. They are also generally quicker to generate large amounts of sequence coverage. The WGS approach does need careful thought however as to how finishing of the sequence will progress, in the absence of a physical map. As with many areas, a combination of clone-based mapping and WGS strategies (hybrid approach) may prove to be the optimal approach.

A combination of clone-based mapping and WGS sequencing is currently being employed in the sequencing of the mouse genome. Early assemblies using WGS sequence have been released, constituting a "draft" version of the *Mus musculus* genome. Currently, approximately 6x genome coverage of WGS sequence has been generated, and the latest assembly (NCBI 30 assembly) comprises ~ 2,500 Mb of sequence in 136483 contigs (statistics taken from Ensembl Jan 2004). This approach provided an opportunity to test two independently produced assembly programs, Phusion (Mullikin and Ning, 2003) and Arachne (Batzoglou *et al.,* 2002), which were developed simultaneously. This WGS approach has provided a useful sequence resource at an early stage, which will be combined with physically-mapped, clone-based sequence to provide a "finished" genome. This illustrates the utility of a combinatorial approach.

The effort to assemble a BAC map of the mouse genome (Gregory *et al.*, 2002) was greatly facilitated by the availability of the human genome sequence, which provided a framework to allow the mouse map to be generated remarkably quickly, as will be illustrated in Chapter 4. BLAST analyses were used to match mouse BAC-end sequences to human chromosomes. This allowed mouse BAC contigs, assembled by restriction fingerprinting, to be ordered and oriented in regions of conserved synteny between the two genomes. Neighbouring mouse contigs could then be analysed for possible fingerprint overlaps at lower stringency. The availability of comprehensive data from large-scale BAC–end sequencing and whole BAC library restriction-fingerprinting efforts was crucial in this regard.

This combined approach provides an efficient model for the mapping and sequencing of genomes for which related genome data are available, and provides early sequence data as well as a high-quality finished genome sequence and mapped clone resources. This hybrid approach is also currently being applied to the zebrafish genome.

1.1.4 Future directions and related studies

A WGS approach has also been utilised in the sequencing of the genomes of *Fugu rubripes* (Aparicio *et al.,* 2002) and *Ciona intestinalis* (Dehal *et al.,* 2002). In the case of *Ciona intestinalis,* high levels of haplotype diversity have complicated early assemblies of the sequence data. This illustrates the point that WGS approaches alone may not be optimal, and that assembly approaches will need to be continually refined.

For organisms with less well-developed genomic resources, techniques such as HAPPY mapping have shown utility in generating physical mapping data, as was displayed in the *Dictostylium* genome project (Williams and Firtel, 2000). This relatively simple technique assays pools of sheared genomic DNA for shared markers (Dear and Cook, 1989). This will be particularly important for those organisms for which research funding is limited, such as where examination of the genomes of organisms at key evolutionary positions may shed light on genome evolution.

Generation of reference genomic sequence data is complemented by efforts to understand the differences between individuals within a species, by studies of nucleotide variation such as single-nucleotide polymorphisms (SNPs) where different alleles of single base-pairs are found, small insertions and deletions (indels) and larger DNA segment copy number differences. This information provides markers for linkage and association studies, and information regarding the differences between individuals that may lead to differences in gene expression or protein function. In parallel with the release of the human genome draft sequence came efforts to generate large amounts of SNP data. These efforts are underway in many centres, both publicly and privately funded. At the time of writing, ~2.17 million SNPs had been mapped to the human genome assembly (statistics from dbSNP at NCBI).

The utility of SNP data for association studies aimed at discovering genes involved in complex disease has been debated. The success of such approaches will depend on the degree to which the polymorphism observed is causative or confers susceptibility (the “common-variant, common-disease” argument), and on the underlying haplotype structures (conservation of allelic variations over a genomic region) reflecting population stratification.

Efforts are underway to generate a haplotype map of the human genome to determine the extent of linkage disequilibrium (LD) within the genome (Gibbs *et al.*, 2003), and early studies have suggested that relatively large segments of the genome appear to exist as “haplotype blocks” (Wall and Pritchard, 2003). This may reduce the number of SNPs needed to perform genome-wide screens, but at the expense of resolution. Nevertheless, these studies will doubtless shed light on the evolution of the human genome and will further our knowledge of the differences that contribute to our individual traits, susceptibility to disease and response to therapies.

1.2 Gene identification

1.2.1 Genes

Studies of genes, greatly facilitated by the availability of large amounts of genomic sequence, have shown us that gene structures can display great diversity. At the most fundamental level, the difference between eukaryotic and prokaryotic genes is in the use of introns. In prokaryotes, transcription of genes occurs and the mRNA is translated directly. In eukaryotes, the mRNA is subjected to processing including the addition of a 5’ CAP structure, addition of a 3’ polyA tract and the splicing of introns. Only then is the mature mRNA translated. However, the discovery of trans-splicing in *C. elegans* (Blumenthal, 1995) hints at further undiscovered complexities in the expression and composition of genes.

The word “gene” has also proved difficult to define. Alternative splicing of transcribed mRNA, the use of alternative promoters and different polyadenylation sites (as explored in Chapter 3) can all result in further complexity, leading to the production of mature mRNA species encoding different proteins, in different tissues and with different physiological half-lives. Alternative splicing is seen for many genes (see also Chapter 3) and is currently an area of intensive research.

One of the most extreme examples of alternative splicing seen to date is the *Drosophila melanogaster* Dscam gene, reported potentially to encode greater than 38,000 different mRNA species (Schmucker *et al.,* 2000). Verifying which variants are actually produced *in vivo*, and have physiological relevance rather than being just products of inefficient or aberrant splicing, is technically difficult due to incomplete representations of different tissues and different developmental stages, especially in human. In this regard, model organisms often provide a more comprehensive array of resources. The development of microarrays is also maturing, and in the future may allow such high-resolution studies on a larger scale. A striking example of the physiological relevance of alternative splicing is the sex-lethal (sxl) gene in *Drosophila melanogaster*. Skipping of an exon containing a premature termination codon in females allows production of functional sxl protein, leading to further regulation of splicing in other genes to result in sexual differentiation effects.

The use of alternative promoters and of other regulatory elements is another poorly understood facet of genes. Utilisation of these elements allows a gene's expression to be controlled, both temporally and spatially within the tissues throughout the organism. For example, the dystrophin gene uses at least eight different promoters to generate cell-type specific transcripts. One of the clear benefits of the availability of genomic sequence, as opposed to studying only mRNA, is that promoters and other regulatory elements will be contained within the sequence. One of the challenges in this area is the identification of such regions. Experimental techniques for studying these elements are low-throughput and labour-intensive, and whilst various computational approaches have been used to try and identify promoters, they are limited and suffer from high over-prediction (and potentially under-prediction) rates. However, an approach demonstrating utility in this regard is comparative analysis of genomic sequence from multiple species, as will be discussed further in Section 1.4.

Utilisation of different polyadenylation sites has been noted for many genes, and the study of these has benefited from the production of large numbers of 3' region Expressed Sequence Tags (ESTs). By noting when 3' ESTs cluster together and share a common polyadenylation site, and by aligning these against genomic sequence, it can be determined if different EST clusters are present at different regions of a gene's 3' UTR. As regions in the UTRs of genes have been shown to regulate their rate of decay, and so indirectly the level of protein produced from them, utilisation of alternative 3'

UTR sites could allow a gene to produce mRNA species with different physiological properties. Signals present within the 3' UTR controlling the localisation of mRNA within the cell, and hence localisation of resultant protein, have also been described (Veyrune *et al.*, 1996) and could be controlled by alternative polyadenylation site usage.

Other complexities of gene structure include the complex VDJ segment joining of immunoglobin genes, in which gene structures become rearranged with the potential for a great variety of structures to be created. Genetic aberrations can also occur, resulting in fusions of genes in some instances. Fused genes can result from chromosomal translocations bringing different regions of different genes within proximity of one another, resulting in production of hybrid mRNA transcripts. An example of this is the translocation that can occur between chromosome 22 and 9, generating "Philadelphia" chromosomes in chronic myelogenous leukaemia (Nowell and Hungerford, 1960). This translocation generates a fusion between the BCR and ABL genes, with the resulting BCR-ABL gene causing activation of transforming pathways.

The availability of large amounts of genomic and mRNA sequence, and both computational analysis and manual curation (as seen in Chapters 3 and 4) have provided data on the physical dimensions and sequence compositions of a large number of gene structures in complex organisms. Great heterogeneity is apparent, with genes ranging in size from approximately 0.1 kb to 2.4 Mb in human (Zhang, 1998). The large variation in size is largely attributed to differing intron sizes, as the average exon size is relatively uniform at approximately 200 bp. For the largest genes, a very small proportion of the transcribed RNA becomes mature mRNA. This would appear wasteful in terms of energy requirements of the cell, but it has been proposed that this may constitute a form of control due to the length of time taken to transcribe these genes with respect to other events occurring within the cell (e.g. ~ 16 hrs for the transcription of the 2.4 Mb dystrophin gene (Tennyson *et al.*, 1995)).

Depending on the region of the genome in which a gene resides, the repeat and GC content of different loci is seen to vary. In different organisms, orthologous genes can show marked differences in their size and repeat content. For example, the genome of the pufferfish *Fugu rubripes* is more compact than that of *Homo sapiens*. It contains fewer repeat sequences, and its genes generally have shorter introns (Aparicio *et al.*, 2002).

As gene structures and their interpretation by the transcriptional apparatus of the cell are so complex, defining a gene is non-trivial. It is often in practice taken to mean the boundaries of the genome containing the promoter and exons (although not necessarily all regulatory regions) used to produce a range of related transcripts. That is the definition that will be used in this thesis, and the word "locus" will be used to describe defined regions of the genome, which may often be synonymous with a gene.

1.2.2 cDNA-based gene identification methods (direct selection) and exon trapping

Before large amounts of genomic sequence were available, gene discovery techniques focussed on the cloning and sequencing of cDNA sequences generated from mRNA, and on methods such as direct selection.

Direct selection (Lovett *et al.*, 1991), uses the ability of mRNA to be converted to cDNA, and the ability of genomic DNA and cDNA strands to form duplexes via hybridisation. Genomic DNA clones are digested by restriction enzymes, and linkers are attached to the free ends facilitating attachment to magnetic beads. The clone fragments are incubated with cDNA clones amplified from a cDNA library by PCR, and hybridisation can occur if exons are present in the genomic fragment and the corresponding mRNA is represented in the cDNA source library.

Hybridised cDNAs are captured using a magnetic column, the cDNA fragments eluted and re-amplified by PCR, and the process repeated using this refined pool of cDNAs. Successive rounds of this procedure result in an enrichment of cDNA species, which are eventually sequenced. This approach is limited by the representation of genes in the cDNA library chosen, however, and by the low hybridisation efficiency of small exons.

Some of the benefits of this method are that information is generated regarding the expression pattern of the transcript, based on the sources of the mRNAs, and also on the splicing of the transcript as intron splicing has occurred by the stage when processed (polyA+) mRNAs are isolated. In this way, different splice variants can be identified, and their expression patterns determined in a variety of tissues.

Exon-trapping (Duyk *et al.*, 1990), involves the insertion of a genomic DNA fragment into a "mini-gene" vector, containing two exons separated by an intron, which contains a multiple cloning site. If the genomic DNA fragment contains an exon, it can be spliced in-between the existing vector exons following transient expression in mammalian cells, and its presence or absence is revealed by PCR analysis of the mRNA produced. This approach is however limited to small-insert clones, and can also be limited by slow or cryptic splicing.

The cDNA sequence alone, however, does not allow the structure of the gene to be elucidated and does not contain information on promoters or regulatory sequences, other than those which may be present in exons. In addition, mRNA is technically difficult to handle, as it is very susceptible to degradation, and tissue availability as a source of mRNA can be an issue in some species.

1.2.3 Sequence-based gene and regulatory-sequence identification methods

The genomic sequence contains all of the information needed to transcribe functional genes, including exons, promoters and other regulatory sequences such as enhancers. As such, it provides a powerful resource for gene discovery. The problem then becomes one of determining where all the different features of the genes lie. To address this, two main approaches are adopted – *ab initio* gene prediction and annotation using mRNA sequence.

Various programs have been developed that predict exons or gene structures in genomic DNA of different organisms. Such prediction is more straightforward in bacteria and yeast as open reading frames are more easily discernible and the genes lack introns. More sophisticated approaches are required in higher organisms. The prediction algorithms generally all make use of the different composition of coding regions compared to the rest of the genome, as they are constrained by codon usage, to predict coding regions.

Exon prediction programs such as GRAIL (Xu *et al.,* 1994) are used to predict single exons. Gene prediction programs such as FGENES (Solovyev and Salamov, 1997) and GENSCAN (Burge and Karlin, 1997) make further use of conserved splicing signals in introns to attempt to predict exon/intron structure. Whilst all programs suffer to varying degrees from lack of specificity and sensitivity (Guigo *et al.,* 2000), they have nevertheless proved invaluable in the annotation of genomic sequence and can attain high levels of accuracy in some instances (>90% for Genscan (Guigo *et al.,* 2000)). Use of multiple programs can increase sensitivity and confidence in prediction to some extent.

The initiation of large scale cDNA sequencing projects and release of the sequence information into the public domain has had a dramatic effect on genome-scale gene identification in a variety of species, including mouse and human. Some of these projects are listed in Table 1-1.

The cDNA sequences produced can be mapped and aligned onto the genomic sequence by programs such as EST2GENOME, allowing gene structures to be annotated. These programs are not perfect though, and sometimes find difficulty defining splice sites correctly and can miss very small exons.

Project / Centre	Comments	Statistics to date	Link
NEDO – University of Tokyo, Helix Research Institute, Kazusa DNA Research Institute; Japan.	Three-centre human cDNA library generation and sequencing project. Some centres utilise the oligo-capping methodology. Non-Kasuza clones prefixed "FLJ".	Total of 29,314 clones registered with DDBJ.	http://www.nedo.go.jp/bio-e/index.html
Kazusa DNA Research Institute, Japan.	Kazusa cDNA sequencing project. Part of the NEDO project. Utilise size-fractionated human cDNA clones to select larger clones. Clones prefixed "KIAA" (or "FLJ" for NEDO clones).	2,031 KIAA clones in HUGE database. 362 FLJ NEDO database clones (adult spleen, not oligo-capped).	http://www.kazusa.or.jp/en/
RIKEN – Genomics Sciences Centre, Japan	Mouse cDNA library production, sequencing and functional annotation. Developed many methodologies for high-throughput cDNA library preparation and sequencing, including the CAP-trapper method.	60,770 mouse clones in FANTOM2 dataset.	http://www.gsc.riken.go.jp/
MGC – NIH Mammalian Gene Collection. Multi-Institute trans-NIH initiative.	Human and mouse cDNA library production and sequencing of clones containing full-length ORFs.	Human 14,878 full ORF clones (11,061 nr genes). Mouse 10,947 full ORF clones (9,019 nr genes).	http://mgc.nci.nih.gov/
DKFZ – Heidelberg, Germany.	Consortium of eight centres to produce (at the DKFZ) and sequence human cDNA clones.	>3200 FL cDNA clones (tissue/development specific)	http://www.dkfz-heidelberg.de/mga/GCC/

Table 1-1 Major human and mouse cDNA sequencing projects in progress at time of writing.

Key references:

HUGE – (Kikuno *et al.,* 2002)

RIKEN – (Shibata *et al.,* 2000), (Carninci *et al.,* 2000)

NEDO – (Yudate *et al.,* 2001)

MGC – (Strausberg *et al.,* 1999)

There are also technical limitations to be considered in large-scale cDNA sequencing projects. One major problem is avoiding clone size bias due to the relative inefficiencies of ligations for larger insert clones. This is often overcome by conducting a size fractionation step with the source mRNA, with the different size fractions then sub-cloned separately to avoid smaller inserts out-competing larger ones (e.g. KIAA cDNA clones – Kazusa DNA Research Institute, Chiba, Japan). Another problem is ensuring that the clones represent the full 5' UTR of the mRNA transcript. Here, various different approaches have been applied such as oligo-capping, which targets the 5' CAP structure of mature mRNA, and hence selects for full-length mRNA species (e.g. RIKEN cap-trapper methodology (Carninci *et al.,* 2000)).

Although improvements have been made in this area, it is still an issue that impacts on the study of promoters and transcription, as knowledge of the true 5' end of a mature mRNA species can determine where in the genome the transcription start site and core promoter lie. Conversely, mRNA that is incomplete at the 5' end can lead to erroneous conclusions, and other methods for determining the true 5' end such as primer extension are laborious. Additional technologies, such as subtractive hybridisation are also used to enrich the diversity of transcripts represented in the cDNA libraries chosen, in order to reduce redundant sequencing (Konietzko and Kuhl, 1998).

cDNA sequences provide direct information about transcription and RNA processing. However, even the application of the measures described does not provide access to all mRNAs (for example those expressed in tissues that cannot be obtained). Conversely, the genome sequence contains information on all of the exons, but these are small signals hidden in a mass of other sequences. An additional, powerful method of genomic sequence-based gene identification utilises genomic sequence from different species, and the observation that functionally important regions of the genome, such as exons, are usually more conserved between species than non-coding DNA. This approach will be discussed later in Section 1.4.

Genomic sequence alone cannot provide information on which splice variants are produced or where they are expressed. As is often the case, the optimal strategy combines two approaches – use of genomic sequence and gene prediction algorithms combined with mRNA sequence information. The approach to gene identification described in Chapter 3 and 4 utilises mRNA sequence aligned against the genome sequence to facilitate manual annotation of gene structures. Additionally in Chapter 3, results from gene prediction algorithms are used to target the design of primers to putative exons, to interrogate cDNA sources and generate sequence which can be used to elucidate gene structure.

A difficult problem in genomic sequence analysis is the issue of promoter prediction. Unlike gene prediction, promoter prediction must contend with the problems of short transcription factor binding site consensus sequences, which occur by chance many times within the genome. As a result, many *ab initio* promoter prediction approaches suffer from very high false positive rates. In addition, promoter function is poorly understood due to the technical difficulties involved, and there are relatively few verified promoters that can be used for training the computational methods. Promoter function can also involve elements that are at a considerable distance upstream or downstream of a gene (presumably involving looping of DNA to bring these elements into proximity with the gene). These effects present a formidable or insurmountable challenge from a genomic analysis perspective.

Progress has been made however, and *ab initio* promoter prediction packages (e.g. PROMOTERINSPECTOR (Scherf *et al.,* 2000), FirstEF (Davuluri *et al.,* 2001)) and programs that predict transcription start sites (e.g. EPONINE (Down and Hubbard, 2002)) are available, alongside a database of experimentally determined transcription factor binding sites (TRANSFAC (Wingender *et al.,* 2001)). Increasing representation of the true 5' ends of mRNA transcripts in libraries used for large-scale sequencing projects (e.g. RIKEN cDNA sequencing project) and from Rapid Amplification of cDNA Ends (RACE (Frohman *et al.,* 1988)) technologies also lead to information regarding where the core promoter lies for more refined analysis. Further effort is needed in this area though, as neither of these methodologies is completely effective.

1.2.4 Genomic sequence analysis

The availability of the sequence of an organism's genome also allows the study of the composition of genomes as a whole. It allows comparisons between the genomes of different species (see later) and studies of how the composition of a genome varies between species and between different chromosomal regions within a species. Earlier studies of mammalian genome composition utilised approaches based on differential staining of genomic regions by different dyes, such as Giemsa. The banding pattern produced by this stain reflects differences in GC composition and degree of condensation, with dark "G" bands having a relatively low % GC (and also lower gene content) and the lighter "R" bands a higher % GC (and raised gene content). The availability of large amounts of genomic sequence allows questions of genome composition to be addressed directly.

The study of isochores is one such application. Bernardi and others first demonstrated that genomes appear to contain regions with differing GC compositions, termed isochores, from gradient ultacentrifugation of nuclear DNA producing fractions with differing GC content (Bernardi *et al.*, 1985). These isochores comprise regions between 100 kb and many megabases in size. The human genome is thought to contain five classes of isochore – two light (AT-rich) classes; L1 and L2, and three heavy classes (GC-rich); H1, H2 and H3. The differing GC composition of isochores appears to correlate with features such as repeat density, gene content, replication timing and recombination frequency. Building understanding of these correlations and possible mechanisms of genome composition evolution is an area of current research, and benefits greatly from the framework provided by the genomic sequence.

GC content also varies on a finer scale. Within the vertebrate genome, so-called CpG islands are found, and in humans are associated with approximately 56% of genes (Antequera and Bird, 1993). CpG di-nucleotides are depleted within the genome, due to methylation of cytosine pre-disposing it to undergo deamination to form thymine. Due to active promoter regions being methylation-free, CpG di-nucleotides within these regions are preserved, forming CpG "islands". CpG islands are found within the promoter regions or first exons of housekeeping genes, and a proportion of genes with restricted expression. This fact was used in early transcript mapping approaches and estimation of human gene number using restriction digestion of DNA using methylation-sensitive restriction enzymes such as *HpaII* (which cut unmethylated CpG regions). Programs are available to find CpG regions within genomic sequence (e.g. as part of the GRAIL program, and Gos Micklem personal communication), and these can be correlated with gene structures as annotation continues.

The overall GC content of genomes has also been shown to vary substantially. Between orders large variations in overall GC content can be seen. For example, the genome of the Archael extremophile *Methanococcus jannaschii* which occupies an environment with very high pressures (>200 atmospheres) and temperatures (~85^{0}C) (TIGR website) is ~ 31% GC, compared to an average of ~ 41% GC for the human genome (Bult *et al.,* 1996).

Another feature of genomes that can be studied using sequence analysis approaches is the repeat landscape. The genomes of different organisms contain differing numbers and types of repeats. The genomes of the pufferfish *Fugu rubripes* and *Tetraodon nigroviridis* appear to be relatively repeat-poor, and so are considered compact genomes. In contrast, mammalian genomes contain a variety of different interspersed repeats. In human, interspersed repeats constitute approximately 45% of the genome. The predominant repeats are LINEs (Long Interspersed Nuclear Elements) and SINEs (Short Interspersed Nuclear Elements), with smaller contributions from LTR elements and DNA transposons. Their genomic distribution appears to correlate with factors such as GC content and gene density.

Availability of genomic sequence allows repeat composition to be studied in context with other features of the genomic landscape, and these analyses are facilitated by the availability of repeat element databases (e.g. REPBASE (Jurka, 2000)) and search tools (e.g. REPEATMASKER, A. F. A. Smit and P. Green unpublished). Some aspects of genome structure and function are very likely to remain intractable using a pure sequence analysis approach. However, in many areas a combination of sequence analysis and experimental approaches is proving highly productive in building understanding about the composition of genomes, how they function and how their structure and dynamics are related to these functions.

1.3 Gene duplications and evolution of genomes

1.3.1 General evolution framework

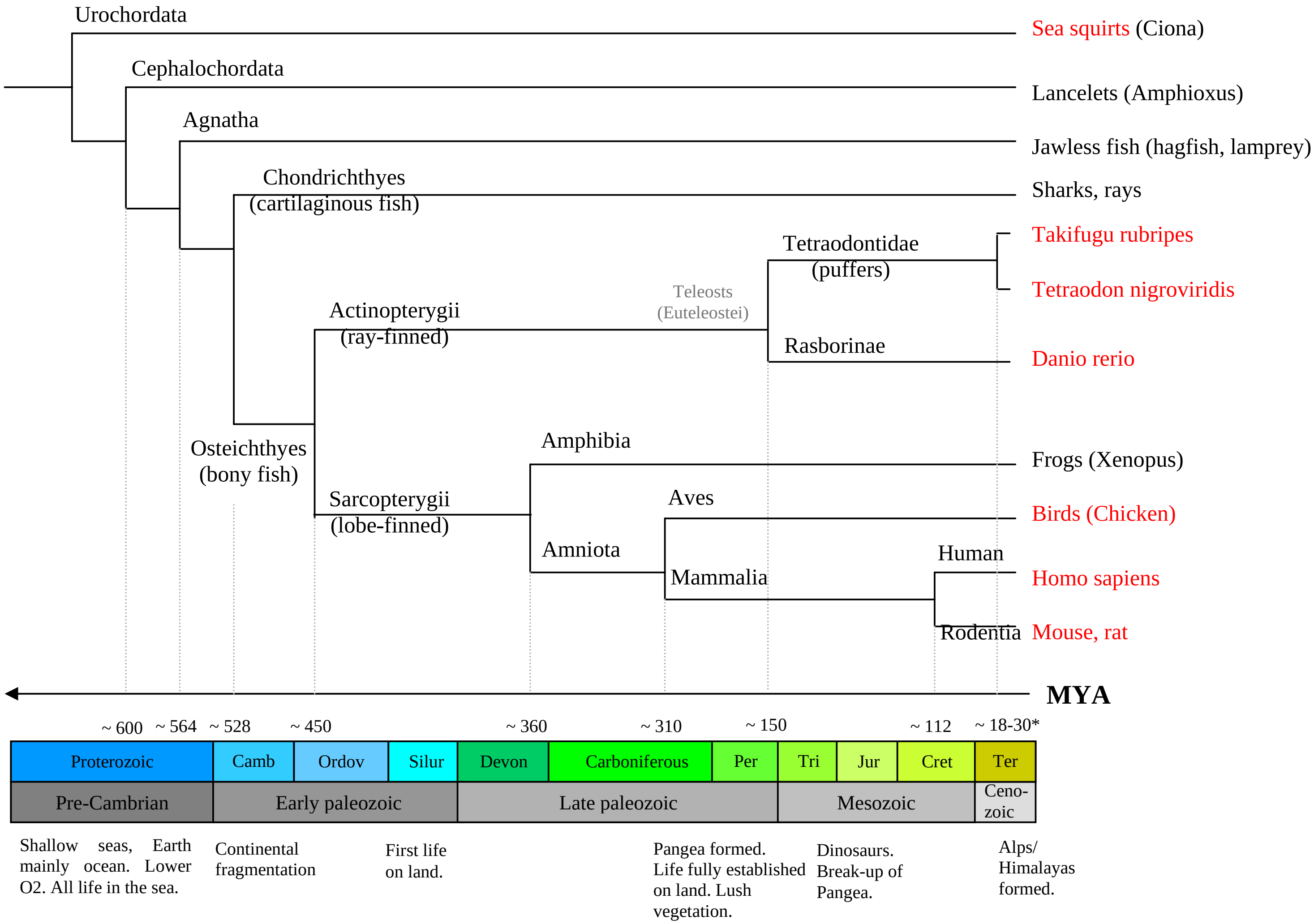
The age of the Earth is debatable, but one estimate is that it is approximately 4.55 billion years old and life is thought to have become established on Earth approximately 3.5 billion years ago. Since then, evolution has resulted in an extraordinarily diverse range of species. Studies of fossils and of the morphology and behaviour of extant species have developed our understanding of the relationships between different orders and species. The more recent application of increasingly sophisticated molecular techniques, to compare protein and nucleic acid sequences among species, has improved our understanding further.

A much-studied aspect of life's evolutionary history is the so-called "Cambrian explosion". During the early Cambrian period approximately 545 million years ago (Mya), there appears to have been a massive diversification of life forms, the reasons for which are the subject of study and debate. The availability of genome sequences of different species will allow us to gain a deeper understanding of the relationships between species, the evolution of molecular functions and the evolution of genomes.

Our current understanding of broad organismal relationships is still expanding (for a recent overview, see (Pennisi, 2003)). Twenty-five metazoan organisms for which representative genome sequences are available, are in the process of being sequenced, or are selected for sequencing (Ureta-Vidal *et al.,* 2003). Further to this, many organisms from earlier evolutionary branches are having their genomes sequenced and smaller selected genomic regions from other organisms can be sequenced with modern technologies and improving genomic clone resources.

Of particular interest with regards to studies presented in this thesis is the evolution of vertebrates. An overview of our current understanding of vertebrate evolution is shown in Figure 1-1. Also shown are key geological features of the Earth during these periods, which help to put this evolutionary period in perspective. These include developments such as the establishment of life on land, continental fragmentation and formation, formation of mountain ranges and changes in atmospheric conditions. Several species at key points in this evolutionary tree have been the subject of studies of particular molecules which have provided insight into species relationships and molecular evolution. Mitochondrial sequences have proved particularly useful in this regard.

A central theme of biological research in the last two centuries is the theory that selective pressures act upon the organism to shape the evolution of functions at the molecular level. This is ultimately reflected in the genome of an organism, the genes it contains and its composition. The following sections discuss key theories under active investigation in this area.



(*Crnogorac-Jurcevic *et al.*, 1997)

Figure 1-1 Schematic cladogram of selected branches of chordate/vertebrate evolutionary relationships. Estimates of divergence times are from Kumar and Hedges (Kumar and Hedges, 1998), unless otherwise indicated (Camb – Cambrian, Ordov – Ordovician, Silur – Silurian, Devon – Devonian, Per – Permian, Tri – Triassic, Jur – Jurassic, Cret – Cretacious, Ter – Tertiary). Organisms for which large amounts of genomic sequence are being generated are in red font.

1.3.2 Whole genome duplication hypothesis

It had been noted previously that for several genes, vertebrates have several copies of genes which are represented by a single copy in invertebrates. The most notable examples are the *Hox* genes. These genes, present in paralogous clusters, have key roles in developmental regulation and have been the target of intensive study (Averof, 2002).

In 1970, Susumu Ohno proposed that, during the course of evolution, the genomes of organisms giving rise to the vertebrate lineage underwent many gene duplication events, and that this contributed to an increase in organism complexity (Ohno, 1999). This has been proposed to have incorporated rounds of genome duplication. This theory could explain why some organisms contain multiple copies of a particular gene, depending on when a genome-duplication occurred. Differences in the exact ratios seen could be explained by gene loss or tandem duplications in different species. Figure 1-2 summarises currently proposed genome-duplication events.

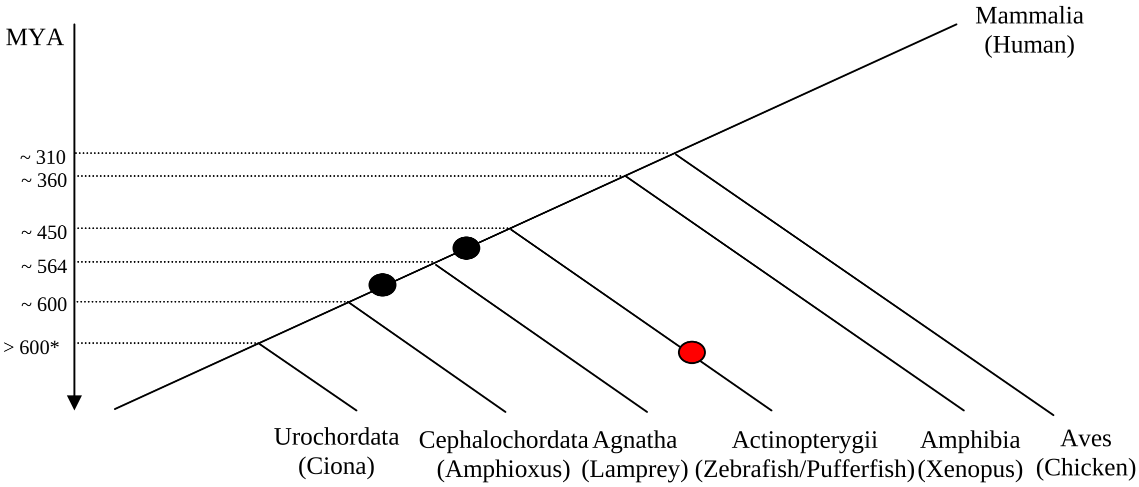


Figure 1-2 Schematic representation of proposed genome duplication events (black circles) during vertebrate evolution (including the genome duplication occurring in the teleost lineage – red circle). Divergence timings are from Kumar and Hedges (1998) (except *, Makalowski, 2001)

Additional evidence supporting this theory comes from the observation of tetraploidy in an amphibian, *Xenopus laevis*, which could represent an intermediate stage in whole-genome duplication. It has also been shown that the yeast genome has undergone whole-genome duplication (Wolfe and Shields, 1997), as well as the

Arabadopsis genome. Proponents of the theory suggest that two rounds of genome duplication occurred early in the vertebrate lineage (“2R”), with a further round of genome duplication occurring in the lineage producing the teleost fish, for example zebrafish (Figure 1-2).

This theory is still the subject of intensive investigation and vigorous debate, with no clear decision yet between the whole-genome duplication model or an alternative in which genomes were shaped by extensive segmental and tandem duplications, as discussed in the following section. It should be noted that the theories are not necessarily mutually exclusive, and that both mechanisms may have played a role in shaping the vertebrate genomes.

Evidence for the theory of whole-genome duplications has come from studies of DNA content and gene/cluster number in a wide variety of organisms, and has been reviewed extensively. Evidence against it has come particularly from the viewpoint that genes duplicated in this manner should maintain a symmetrical ((A,B)(C,D)) phylogeny where genes generated by a genome duplication are more similar to one another than counterparts prior to the duplication. However, it has been noted that this assumption may be violated if an allotetraploidy scenario is involved, whereby mating of closely related species results in an increase in ploidy. In addition, some plants and animals (e.g. salmonid fish) contain a mixture of tetraploid and diploid loci, indicating that genome duplication may not necessarily occur in its entirety, but may instead only involve a number of chromosomes.

The central theme of these scenarios, whichever proves to be correct, is that by increasing its gene complement through duplication of genetic material, an organism increases the repertoire of molecules on which selective pressures can act, and potentially increases its ability to evolve to meet changing evolutionary pressures.

1.3.3 Segmental duplications and tandem duplications

Apart from whole-genome duplication, an organism can increase its gene complement through the processes of segmental duplications and tandem gene duplications, segmental duplication being the duplication of complete region of genomic DNA and tandem duplication being the duplication of a gene within the same region of the genome.

Segmental duplications can lead to duplication of multiple genes in a single event. They are characterised by common order and transcriptional direction of the paralogous genes. These may be tolerated by the organism to different degrees, depending on the copy-number effects of the genes involved. This is non-trivial, as there are examples where an increase in copy number of a gene has deleterious effects, such as in Pelizaeus-Merzbacher disease, which can be caused by duplication of the PLP gene region (Inoue *et al.*, 1996). These effects will likely depend on the complexity of the organism and the functions of the genes involved. Large numbers of segmental duplication would also lead to greater homogeneity of the genome's composition, depending on the size and numbers of duplications, as before the sequences diverge through mutation, they will be equivalent.

In humans, segmental duplications have been implicated in disease, such as DiGeorge and velocardiofacial syndromes, due to their promotion of genomic rearrangements by a variety of mechanisms including unequal crossover (Emanuel and Shaikh, 2001). These rearrangements can lead to duplication or deletion of genes. Such rearrangements have also been attributed to other low-copy repeat (LCR) regions which are not necessarily part of segmental duplications (Mazzarella and Schlessinger, 1998; Stankiewicz and Lupski, 2002).

Tandem duplications of genes also increase the gene complement of an organism and are again constrained by copy-number effects, but as fewer genes are involved in the duplication event, the functional impact is expected to be lessened. This is however still dependent on the gene in question and the context in which it operates. There are many well studied examples of tandemly duplicated genes, including the hox and globin genes.

The generation of genome-scale sequence data for many organisms such as *Saccharomyces cerevisae*, *Drosophila melanogaster*, *Caenorhabditis elegans* and others has allowed studies on the processes of segmental and tandem duplications in these organisms, allowing statistical techniques to be refined and shedding light on the contribution of these processes to genome evolution (Wolfe and Shields, 1997), Friedman and Hughes, 2001), Achaz *et al.*, 2001).

The relevant contributions of these processes to the shaping of the human genome have recently been the subject of two gene-centric genome-scale studies (Gu, *et* *al.,* 2002); McLysaght *et al.,* 2002), and were made possible by the completion of the draft version of the human genome (Lander, Linton *et al.*, 2001). These studies utilised different methodologies, based on molecular-clock analysis of duplicated genes. Both concluded that extensive gene duplication, or at least one round of polyploidy, occurred early in chordate evolution, consistent with widespread segmental or chromosomal duplications. Gu *et al.*, (2000) also noted a wave of duplications subsequent to the mammalian radiation, which were attributed to tandem or segmental duplications.

Further studies of segmental duplications have been performed using the human genome sequence. Bailey *et al.*, (Bailey *et al.,* 2002) made use of the whole-genome shotgun data from Celera to identify regions of the genome that had high-levels of sequence identity to one another, thus representing relatively recent duplication events. In this approach, over-representation of a region with WGS reads was used as evidence of a duplication. The authors concluded that approximately 5% of the human genome consists of highly related duplications, and their study illustrates another application of combined genome-sequencing strategies.

All duplication event studies are faced with two main problems: the first is the loss of genes subsequent to duplication; and the second is the inability to detect similarities over deep evolutionary time, due to rearrangement events disrupting synteny and gene order, and accumulation of separate mutations causing sequences to diverge. The study of whole genomes from different species provides important assistance here. These studies can sometimes help to establish whether a gene has been lost in one species or gained in another. Here a consensus approach is used, as duplication of a gene independently in more than one closely-related species (or gene loss) is less likely than an ancestral event having given rise to the situation seen in both species.

In cases where sequences have diverged substantially and paralogy is less certain, genomic sequence can be extremely powerful in supporting the case for paralogy. Conservation of gene order, transcriptional orientation with respect to neighbouring genes and conserved exon structure (including intron phases) are indicative of paralogy in these circumstances. An example of this is seen in Chapter 6.

Much work has been conducted on paralogous genes, to study how they have diverged in function. Immediately following the duplication event, the paralogues would presumably share a common function and expression pattern, depending on whether all relevant regulatory sequences were also duplicated. Over time, the paralogues will diverge and one might acquire a novel function or alternatively might become a non-functional pseudogene. To some extent, it might be possible to infer the function of one paralogue from information on the functional role of the other. The extent to which this is appropriate will depend on the extent and nature of the divergence at the protein-encoding level and of their expression patterns, both temporally and spatially.

An obvious question relating to paralogues is that immediately after two paralogues are created by a duplication event, why should an organism keep both copies? One possibility is that after the duplication the organism may not have any mechanism for removing the second copy. If such a mechanism exists, the answer will depend on what selective advantages are conferred by divergence of expression pattern and the proteins encoded, balanced against the rate of gene loss and the additional DNA synthesis and transcriptional burden placed on the organism.

Many studies have focussed on the divergent functions of proteins encoded by paralogous genes, such as the *Hox* genes. Recently attention has focussed on the hypothesis that mutations that affect protein function substantially are likely to occur at a lower rate than those in regulatory sequences bound by transcriptional factors, which are generally short motifs. This is thought to be because mutations within the short motif may be more likely to affect transcription factor binding than a mutation occurring within a larger protein coding region, much of which may not be functionally critical, causing a change in function.

In this way, divergence of transcription patterns could occur much earlier with loss of tissue-specific transcription factor binding sites (or reduced efficiency) in different paralogues. If the transcription pattern of the ancestral gene is now encompassed by transcription from the two separate genes, there will be selective pressure to keep both paralogues. This sub-functionalization hypothesis has recently been formalised by Force *et. al.* (Force *et al.,* 1999) as the DDC (Duplication, Degeneration and Complementation) model.

In order to test this hypothesis, expression patterns of paralogous genes need to be determined. Studies of expression patterns in large complex organisms are made more difficult by the problems of generating expression data at the sub-tissue or temporal level. In this way, information regarding subtle differences of expression, which may be crucial functionally, is missed. Studies to date have still been informative though, and are also explored in Chapter 5 of this thesis. Recent advances in microarray technology, imaging techniques and improved tissue collections and probes promise further understanding in this area.

Studies of paralogues are also aided by the availability of data from model organisms such as the mouse and zebrafish. Both organisms are amenable to genetic analysis and to high-resolution gene expression studies. Genes studied in zebrafish in support of the DDC hypothesis include the *engrailed1* genes (Force *et al.,* 1999), and Na+/K+ ATPase α genes (Serluca *et al.*, 2001) and illustrate the utility of studies in model organisms. Other techniques such as gene knock-out/knock-in approaches and RNAi methodologies may also shed light on paralogous gene sub-functionalization.

In summary, the increase in availability of genomic sequence for organisms representing different evolutionary lineages and improvements in experimental approaches to studying gene expression and protein function offers an unprecedented opportunity to gain understanding of the evolutionary processes which have shaped genomes, and led to the diversity of life seen today.

1.4 Comparative genomic analysis

1.4.1 Identification of functionally important sequences

The limitations in using experimental and computational approaches in one species to identify functional elements in its genome have been described above. Comparing the genome sequences of different species provides an enormously powerful tool for identifying these functional elements, as these regions of the genome should diverge more slowly than other, non-functional regions.

With the recent availability of genomic sequence from a variety of organisms has come the development of tools for aligning these much larger segments of sequence and visualising sequence conservation. The choice of method is important for analysing orthologous genomic regions. Some methods, such as PIPMAKER (Schwartz *et al.,* 2000), use a local alignment approach to align different genomic sequences. This is particularly useful for detecting homology between regions that may have been rearranged with respect to one another. Other methods, such as Vista (Mayor *et al.,* 2000), employ a global alignment approach which retains information regarding overall arrangements of the regions, but will miss rearranged segments.

Combinations of these approaches yield optimal analyses of similarities between genomic sequences. In common with both of these broad approaches is the need for databases of repeat elements contained within the genomes of different organisms, to allow masking of these regions in the sequences under study. The RepBase database provided by Arian Smit is an excellent example of such a database, which has proved indispensable in the studies of genomic sequences (Jurka, 2000).

An example of the utility of these comparative sequence analysis approaches is shown by studies of the SCL locus, in which sequences of the loci in five different vertebrates were compared, revealing conserved elements important for function (Gottgens *et al.*, 2002). Promoter motifs were found that were conserved across human, mouse, chicken, pufferfish and zebrafish.

A key element in comparative genomic analyses for gene and other functional element identification is the choice of organisms to compare. Ideally, a balance must be reached whereby the chosen organisms are sufficiently closely related to retain significant sequence similarity in functionally conserved regions, and yet be sufficiently divergent that non-functional sequences do not appear as “noise” in alignments. In practice no single species will satisfy these requirements for all types of functional element.

Pilot studies using regions of genomic sequence from a variety of different organisms have been instructive in this regard, such as the comparative genomic sequencing being conducted for a variety of mammals and chicken at the NIH intramural sequencing centre (http:\\www.nisc.nih.gov) (Thomas *et al.*, 2003). BAC libraries are also being generated for a wide variety of species (http://bacpac.chori.org/) which will facilitate sequence-ready contig generation for defined regions in the absence of any larger scale sequencing initiative.

For gene identification purposes, comparison of mouse and human sequences displays considerable sequence conservation in non-coding sequences, and so whilst useful for detecting exons, potential false-positives are also seen. This non-coding conservation was used in the mapping of the mouse genome. Some groups have attempted to circumvent the problem of noise in the human-mouse alignments by using a gene prediction algorithm which also takes sequence alignment data into consideration (e.g. TWINSCAN (Korf *et al.*, 2001) and SGP2 (Parra *et al.*, 2003)). Decreased false-positives can be achieved by using more distantly related organisms. The marsupial genome sequence represents a potentially useful resource in this regard, as it is more distantly related to humans than the mouse, and yet is perhaps not so distant that functionally important regions would not be recognised. An example is seen in Chapter 6 and in the study by Chapman *et al.* (2003). Chicken genomic sequence is also useful for identifying coding regions (Thomas *et al.*, 2003).

The true power of these approaches will perhaps be seen in the study of regulatory sequences. Difficulties in their computational prediction and experimental study may be partially overcome by the use of comparative sequence analyses to highlight regions for further study and reduce downstream effort. The complexities of gene expression regulation are formidable, as illustrated by the studies of regulation of the globin loci. Understanding of the regulation of expression of these genes has been gained over many years of careful study. Improvements in our ability to predict these regions would aid our understanding of this process.

1.4.2 Evolutionary studies

Comparative genome analyses on the whole-genome scale can also provide information on evolutionary relationships between species, how different chromosomes have evolved, and which rearrangements have occurred in different evolutionary lineages. Knowledge of the relationships between genomic regions of different organisms, shared synteny, allows inferences of likely orthology to be drawn from studies in different organisms. For example, knowing the syntenic relationships between human and mouse allows knowledge from genetic studies in mice to be applied in the search for genes implicated in a particular trait.

Many approaches have been used to generate data on shared synteny, such as the use of FISH, radiation-hybrid or HAPPY mapping techniques. The generation of genomic sequence data provides the highest possible level of resolution for comparative mapping. This is important when small-scale rearrangements have occurred within regions of shared synteny, which may not be detected by other comparative mapping approaches. Examples of this are seen in Chapter 4.

Comparative mapping studies have been particularly useful in attempts to elucidate the events involved in the evolution of the mammalian genome. With karyotypes ranging from three pairs of chromosomes in the Indian muntjac deer, to 67 pairs in the black rhinoceros (O'Brien, *et al.*, 1999), it is clear that many genomic rearrangements have occurred within the mammalian lineage. Techniques such as Zoo-FISH have been used to elucidate patterns of large-scale conserved synteny in some mammals (Chowdhary *et al.*, 1998). As mapping data are produced for mammalian species, from sequencing, cytogenetic, genetic and physical mapping approaches, they can now be related back to a high resolution human, mouse and rat genomic framework (Murphy *et al.*, 2001).

1.4.3 Value of comparative genomic analysis for functional studies

Comparative genomic analyses also aids in the establishment of orthology for genes in different species. Establishment of orthology between genes is helpful in studies aiming to understand the function of a gene product in one species, if functional information is available from studies in a related species. In the absence of comparative mapping data, establishing orthology is difficult when the representation of mRNA sequences is much lower for one of the organisms, and the gene in question may show high levels of similarity to multiple genes.

The ability to see genes in their genomic context allows the investigator to collate additional evidence to support orthology, such as the presence of neighbouring genes that are also shared between the different species, and conservation of transcription direction. When taken over a large region and supported by lower-resolution studies of chromosome conservation (such as by cross-species chromosome painting), this is persuasive evidence that the regions represent either orthologous segments or, less likely, a segmental duplication, and that the sequence similarities seen do not refer to paralogous sequences. As will be seen in Chapter 5, this can be beneficial when sequence similarity data alone may be misleading, due to processes such as gene conversion causing sequence homogeneity between genes within a species. In these cases, phylogenetic analyses suggest different relationships to those seen with the benefit of gene context information.

Genome sequence resources are now well developed for many important model organisms as discussed earlier. These will aid functional studies of genes in these organisms directly, as well as shed light on the relationships between orthologues and paralogues. In these organisms, knowledge of the genomic sequence allows the researcher to search actively for different splice variants suggested by gene identification approaches (as discussed earlier) and can provide additional genetic markers (including SNPs) to refine linkage analysis and association approaches. It also provides further information for "knock-out" or "knock-in" approaches and facilitates identification of genes in gene-trap experimental approaches.

In all these cases, knowledge of the relationships between genes in different species is required to fully understand how the functions of those genes differ between species in different physiological settings.

1.4.4 Current and future prospects/projects

Future prospects in the field of comparative genomics include the expansion of number of species for which high-quality, comprehensive genomic sequence data are provided. Currently, organisms have been chosen depending on their prominence as important model organisms, or on their positions in the evolutionary tree. The production of genomic sequences from organisms with key, intermediary evolutionary relationships will further our knowledge of genome evolution. An example is *amphioxus*, which represents an early chordate lineage.

Another key future area for comparative genomics is that of within-species comparisons. It has already been noted that people differ in the copy number of certain genes, some of which may have important functional or medical consequences. Examples include the X-linked cone pigment genes responsible for red-green colour vision (Neitz and Neitz, 1995), and the cytochrome P450 CYP2D6 genes involved in the metabolism of drugs and other compounds (Agundez *et al.*, 2001). A project to sequence the ~4.6 Mb MHC regions of at least 8 different haplotypes is also already underway (Allcock *et al.*, 2002). As well as gross differences between individuals or closely related species, knowledge of the SNP differences between individuals is expected to facilitate refined association studies for complex diseases, and move further towards the goal of personalised medicine, through pharmacogenomic approaches. The development of a haplotype map has been proposed to aid these approaches. In each of these cases, studies are being made of how the genomes of individuals differ at a high resolution.

Epigenetic studies will also complement these approaches, by providing information about how individual genomes are "programmed". These forms of comparative genomic analyses between individuals will further our understanding on how differences in how the genome is "interpreted", affects phenotype.

Common to all these approaches is that by comparing genomic information between species or individuals, we can gain further understanding of how genomes and the genes within them have evolved, and on the functions of those genes. This is a more powerful approach than the study of a single genome, and allows useful transfer of information between studies. A more complete understanding will result.

1.5 The X chromosome

1.5.1 Human X chromosome overview

Comprising approximately 153 Mb of euchromatic DNA (Ensembl v16.33.1), the human X chromosome represents ~ 5% of the human genome. A sub-metacentric chromosome, almost its entire genomic sequence is now known. The GC content of the X chromosome, at 39%, is slightly lower than the genome average of 41%. It is also notably richer in interspersed repeats, containing approximately 57% repeats compared to a genome average of 45%. Much of this increase is attributed to an abundance of LINE1 elements, constituting 30% of the chromosome compared to a genome average of 17%. It appears to be relatively gene-poor compared to the genome average, containing an estimated 800-1000 genes (~ 3% of the human gene complement). Initially estimated from CpG island and EST RH-mapping approaches, this has been supported by transcript mapping approaches, as described in Chapter 3.

The X chromosome has been mapped and sequenced by a consortium of centres, led by the Wellcome Trust Sanger Institute (Figure 1-3), which has generated three-quarters of the finished sequence. The other centres involved include the Baylor College of Medicine (Texas, USA), the Washington University Genome Sequencing Center (St. Louis, USA), the Max Planck Institute for Molecular Genetics (Berlin, Germany), and the Institute of Molecular Biotechnology (Jena, Germany). The chromosome was mapped in bacterial clones using a combination of STS-content mapping and restriction-fingerprinting approaches, to identify BAC and PAC clones for sequencing. In the final phases of the project, YAC clones have also been utilised to close remaining gaps.

The human X chromosome has long been the subject of intensive study due to its unique biology as one of the sex chromosomes. Many disease genes have been mapped to the X chromosome, partly facilitated by the manifestation of many recessive conditions in males giving rise to a characteristic inheritance pattern.

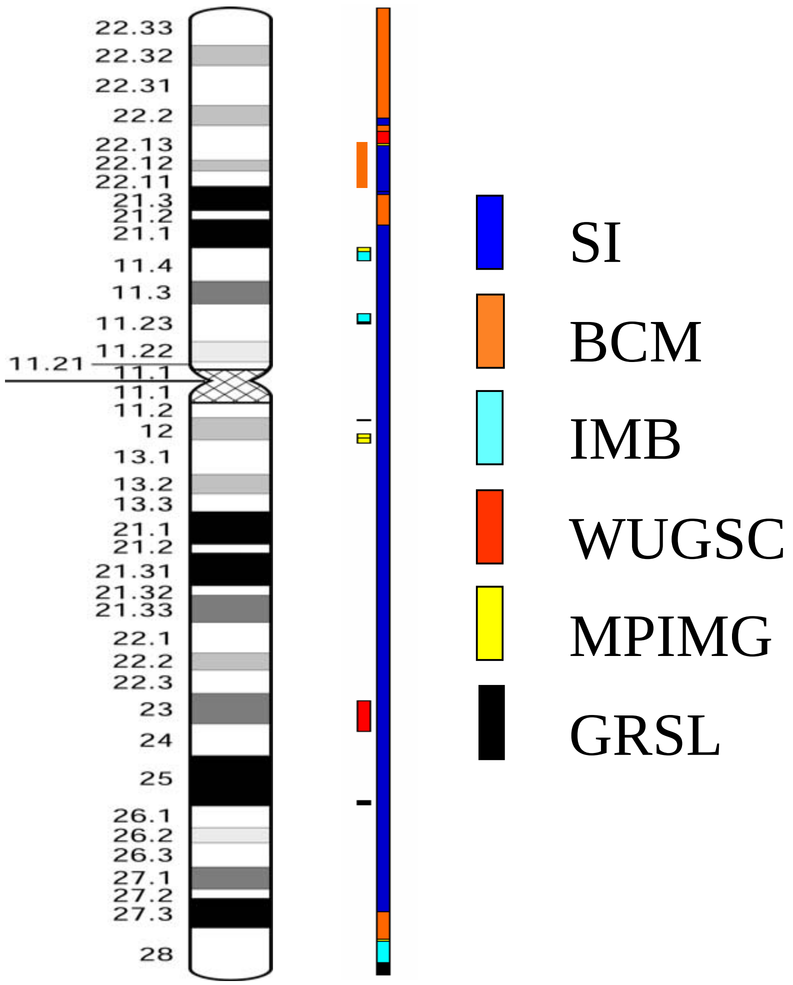


Figure 1-3 Overview of the regions being mapped and sequenced by each of the major sequencing centres. Illustration kindly provided by Dr. Mark Ross, Wellcome Trust Sanger Institute.

At the time of writing, 208 disorders showing mendelian inheritance have been mapped to the X chromosome (Ensembl v16.33.1 – data derived from OMIM), including well known disorders such as haemophilia A and colour-blindness. In particular, there appears to be an abundance of genes implicated in mental retardation located on the X chromosome, many of which have now been characterised (Frints *et al*., 2002).

One distinctive facet of X chromosome biology is X inactivation (Lyon, 1998). As female cells possess two copies of the X chromosome, and males only one copy, a dosage-compensation mechanism has evolved whereby one of the female's X chromosomes is inactivated (Lyon, 1999). Our current understanding of the mechanism is that early in female embryonic development, the X chromosomes are somehow "counted", and all but one of the X chromosomes are subjected to inactivation (Avner and Heard, 2001). The utilisation of a counting mechanism is illustrated by the observation that in XXX cells, two of the X chromosomes are inactivated.

The inactivation mechanism involves coating of the inactive X by the non-coding RNA transcript of the *XIST* locus, which is expressed from the inactive X chromosome (Brown *et al*., 1991). The *XIST* gene is located at the X inactivation centre (Xic) at Xq13. As the choice of X chromosomes for inactivation is usually a random process, either the paternal or maternal chromosome may become inactivated. All descendants of the cell in which the decision was made inactivate the same X chromosome. As the embryo matures, this leads to mosaicism with patches of cells containing either an active paternally or maternally inherited X chromosome. This in turn can lead to mosaic phenotypes, the most striking example of which is the coat patterning of piebald cats. Male spermatogonia also inactivate their single X chromosome, but this becomes reactivated during the production of sperm.

Whilst X inactivation appears to have evolved as a way of maintaining X chromosome gene dosage between sexes, there are a number of genes that escape inactivation (Brown and Greally, 2003). In some instances these are genes with functional homologues on the Y chromosome; for others it may simply be the case that dosage of the gene involved is not important or that the gene has not yet been drawn into the process of inactivation.

1.5.2 Sex chromosome evolution

A key question regarding the X and Y sex chromosomes is - how did they evolve? The X and Y chromosomes are morphologically distinct, and are very different in their gene and repeat composition. However, the human X and Y chromosomes share various regions of homology. Two of these - the major and minor pseudoautosomal regions (PARs)(Rappold, 1993) - enable the sex chromosomes to pair during male meiosis (Figure 1-4). There is an obligatory recombination during meiosis in the ~ 2.5

Mb major pseudoautosomal region (PAR1). This equates to an exceedingly high recombination rate of 20 centiMorgans (cM) per Mb. Recombination within the smaller (~320 kb) minor pseudoautosomal region (PAR2) is not obligatory but is still elevated above the genome average. Between PAR1 and PAR2, which reside at the ends of the chromosome arms, there is normally no recombination in male meiosis.

In addition to these two regions, there are other regions of homology shared between the sex chromosomes. For example there is an XY homology block located at Xq21 and Yp (Sargent *et al.*, 2001). Many X chromosome genes also have homologous counterparts on the Y chromosome.

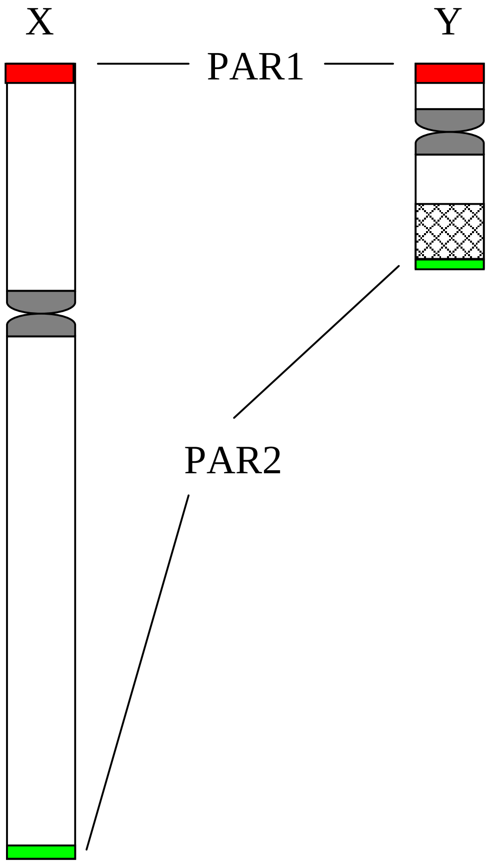


Figure 1-4 Schematic diagram of the human X and Y chromosomes. PAR1 and PAR2 are illustrated. Heterochromatic region of the Y is displayed as a hatched box.

The current theory of mammalian sex chromosome evolution suggests that the X and Y chromosomes were initially a pair of homologous autosomes. When one chromosome, the Y, gained a major sex-determining region (thought to be SRY), it would lead to a need for reduction in recombination between the two chromosomes in order to maintain sex differences.

As the pair of chromosomes evolved, lack of recombination between them would lead to the genetic isolation of most of the Y chromosome, whilst the X chromosome was still able to pair and recombine in females. This isolation of the Y led to its gradual degradation, a process termed "Muller's ratchet", as genes accumulated mutations and became inactive and genetic material was lost. In concert, the evolution of the X inactivation mechanism described earlier would ensure that dosage was largely conserved for X-linked genes between the sexes. Indeed, it has been proposed that the Y is heading for extinction, although this could overlook unknown or poorly understood aspects of Y chromosome biology (including the recent suggestion that the Y chromosome recombines with itself in palindromic regions instead of with a homologous chromosome).

Studies of the sex chromosomes in monotremes (prototheria) and marsupials (metatheria) have shed further light on the evolution of the mammalian sex chromosomes. Mammalian evolutionary relationships are summarised in Figure 1-5.

The X chromosome in marsupials is considerably smaller than that of eutherian mammals, accounting for approximately 3% of the genome, and the Y chromosome is tiny (~ 10 Mb) (Toder *et al.*, 2000). The marsupial X and Y chromosomes do not appear to pair at meiosis and there is no evidence that they possess pseudoautosomal regions. Monotremes differ, in that both the X and Y chromosomes are large, and pairing takes place between the entire short arm of the X and the long arm of the Y. In addition, monotremes appear unique amongst animals studied to date, due to the presence of a number of small chromosomes which form a paired end-to-end chain, pairing to the Y chromosome short arm, at meiosis.

X chromosome inactivation is known to occur in marsupials, but inactivation is imprinted, always affecting the paternal X chromosome. This imprinted inactivation has been also been observed in the extra-embryonic tissues of the mouse. It remains to be established whether or not X chromosome inactivation occurs in monotremes, owing to the lack of a chromatin body in female cells.

Comparative mapping studies between monotreme, marsupial and eutherian X chromosomes has revealed that many genes show conserved synteny, constituting an X Conserved Region (XCR). However, many of the genes located on human Xp are found to be autosomal in monotremes and marsupials (Spencer *et al.*, 1991). This was further demonstrated by a striking experiment in which a marsupial X chromosome 'paint' probe was hybridised to human metaphase chromosomes (Glas *et al.*, 1999). This clearly demonstrated that homology was restricted to Xq and a small proximal region of Xp. This is depicted in Figure 1-6. PAR1 is also likely to have been added to the X during eutherian evolution.

In contrast, the synteny for the X chromosome (if not gene order) is remarkably conserved within the eutherian lineage, possibly due to the strong evolutionary pressures to maintain synteny once genes were recruited into the X inactivation system (Ohno's Law).

The most parsimonious explanation for these observations is that during the eutherian radiation, there was at least one addition to the X and Y chromosomes from autosomes. These studies demonstrate the utility of comparative mapping approaches in the elucidation of events during evolution of the mammalian sex chromosomes.

1.5.3 Human Xq22-q23

When studies for this thesis began, Xq22-q23 was the region with the most comprehensive coverage of finished genomic sequence and presented an ideal region for sequence-based gene identification and annotation approaches. Comprising approximately 15 Mb of DNA, the region spans light, dark and intermediate Giemsa-staining bands. As such, it could be expected to display a heterogenous gene density, isochore content and repeat content. These features combined to make the region ideal for preliminary studies as human X chromosome sequencing progressed.

The Xq22-q23 region has also been shown to contain genes responsible for a variety of Mendelian, inherited disorders, some of which have been cloned. At least 13 mendelian-inherited conditions are associated with Xq22-q23 (OMIM). Gene identification and annotation studies of the region would therefore be of potential benefit for studies aiming to identify the remaining, uncloned disease genes.

A first generation transcript map of the region was published by Srivastava *et al.*, (1999). In addition, other genes had been localised to the region in other studies including the Gene Map '99 RH mapping effort (Deloukas *et al.*, 1998). It was reasoned that a genomic-sequence based approach would be complementary to these studies.

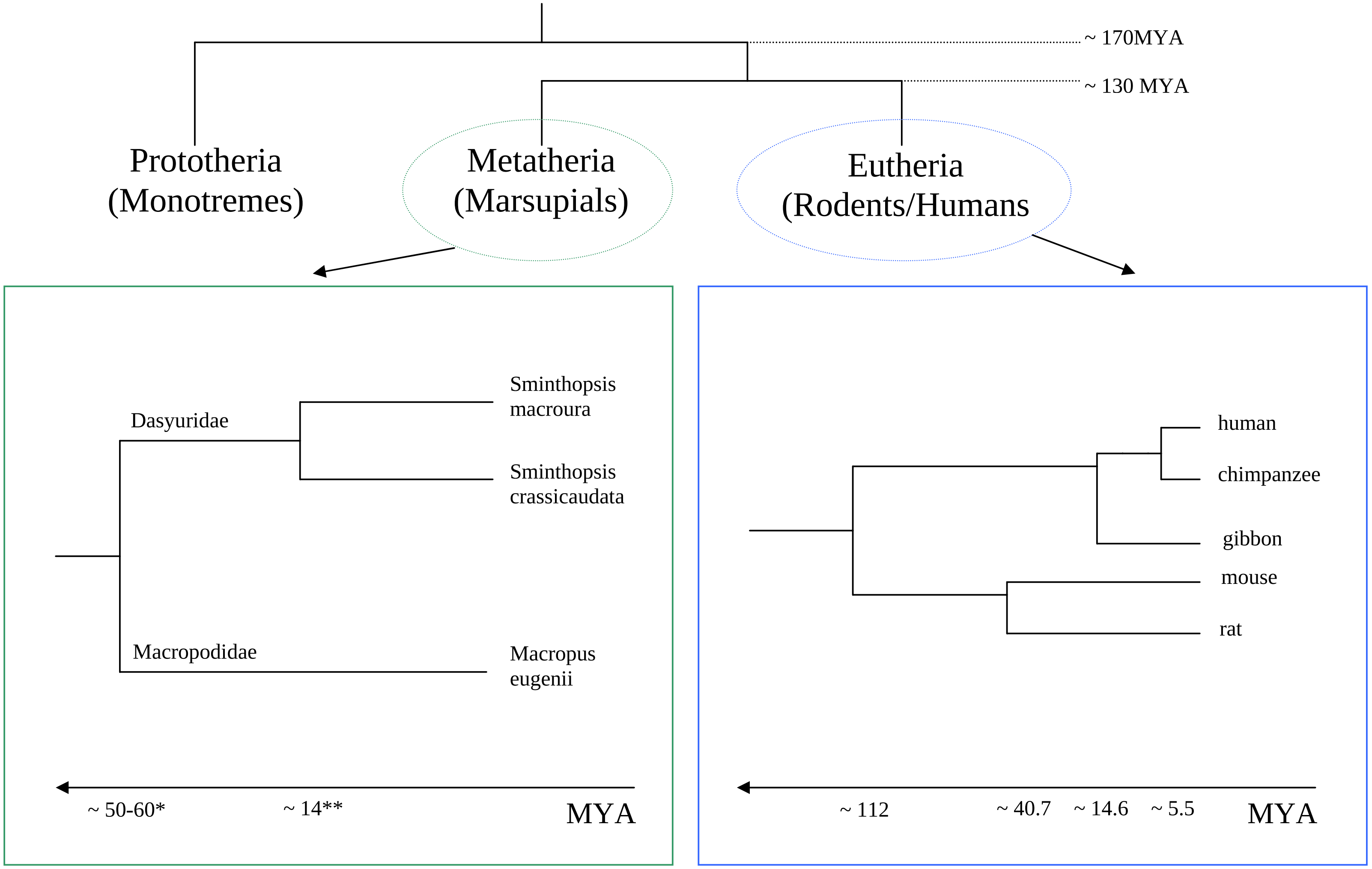


Figure 1-5 Figure depicting selected aspects of mammalian phylogenies. Timings of divergence (except therian divergences) are according to Kumar and Hedges, (1998) except *(Graves and Westerman, 2002) and ** (Blacket *et al.*, 1999).

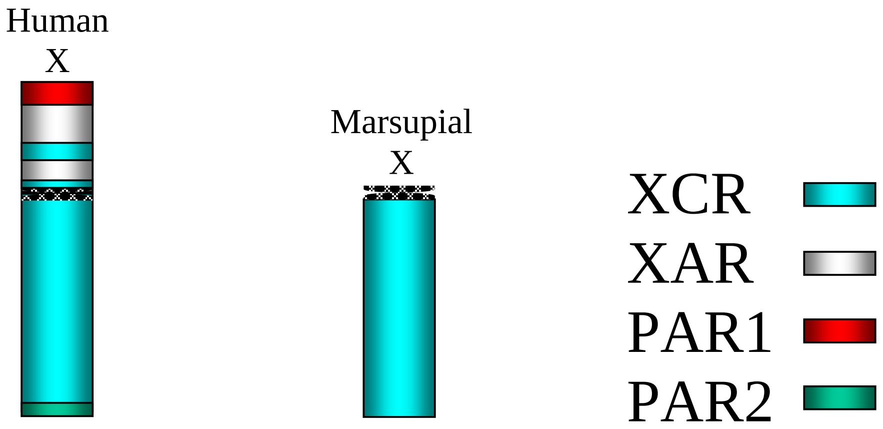


Figure 1-6 Diagram illustrating conservation between human and marsupial X chromosomes. X-conserved regions (XCR), regions added to the X during eutherian evolution (XAR) and pseudo-autosomal regions (PAR1 and 2) are indicated.

The X chromosome is well conserved in gene content between human and mouse, and indeed in all eutherian mammals studied to date, for reasons discussed in the previous section. Gene order is not always conserved on a large scale however, particularly between human and mouse, due to multiple inversion events in the different lineages. Human Xq22-q23 has shared synteny with the E3-F2 region in mouse, and this region has been shown to contain genes leading to mutant phenotypes. An example is the "jimpy" mutant, caused by defects in the Plp gene (Sidman *et al.*, 1964). Investigation of the shared synteny between human and mouse in this region would reveal detail about the fine-structure of conservation present. It would also prove useful in the transfer of information from observed phenotypes in mouse in searches for possible causative genes for disorders mapped to human Xq22-q23, particularly from imminent or future mouse saturation-mutagenesis studies.

In summary, Xq22-q23 provided an ideal region for pilot sequence-based studies of gene identification and annotation on the human X chromosome, which could then be applied on a chromosome-wide basis as sequencing of the chromosome progressed. It would also provide important information for searches for genes involved in disorders mapped to the region, as well as gene structure information for those genes already shown to be causative in disorders, to facilitate mutation screening and studies of the biology of the genes.

Chapter Two - Materials and Methods

2 Materials and Methods

Materials

2.1 Chemical reagents

All common chemicals were purchased from Sigma Chemical Co., BDH Chemical Ltd., and Difco Laboratories unless specified below or in the text.

Bio-Rad Laboratories	β-mercaptoethanol
Gibco BRL Life Technologies	ultraPURE^{TM} Ammonium sulphate, enzyme grade, ultraPURE^{TM} agarose
Amresco	EZ Squeeze Gene-PAGE PLUS acrylamide
Amersham Pharmacia Biotech	Dextran sulphate
VWR	Acetic acid
Merck	Acetone
Fluka	Formamide
PE Applied Biosystems	TET, HEX and NED fluorescent nucleotide dyes
Amersham Biosciences	Vistra Green stain
Boehringer	Blocking reagent

2.2 Enzymes and commercially prepared kits

Amersham Pharmacia Biotech	*Sau 3*A1
PE Applied Biosystems	Amplitaq^{TM} AmplitaqFS
Qiagen	Genomic DNA and DNA gel purification
Sigma	Ribonuclease A Deoxyribonuclease I DNA polymerase I (10 U/µl)
Ambion	DNA-free DNase treatment kit
Invitrogen Life Technologies	Superscript II cDNA synthesis kit
NEB Biolabs	*Hin*d III (20 U/µl)

2.3 Nucleotides

Amersham Pharmacia Biotech	Redivue™ [α-^{32}P]-dCTP (AA 005) aqueous solution (370 Mbq/ml, 10 mCi/ml)
Amersham Pharmacia Biotech	2'-deoxynucleoside 5'-triphosphates (dATP, dTTP, dGTP, dCTP)
Boehringer	biotin-16-dUTP digoxigenin-11-dUTP

2.4 Solutions

Solutions used in the present study are listed below, alphabetically within each section. Final concentrations of reagents are given for most solutions. Amounts and/or volumes used in preparing solutions are given in some cases. Unless otherwise specified, solutions were made up in nanopure water.

2.4.1 Buffers

10x PCR buffer	670 mM Tris-HCl (pH 8.8) 166 mM $(NH_4)_2SO_4$ 67 mM $MgCl_2$
1x $T_{0.1}E$	10 mM Tris-HCl (pH 8.0) 0.1 mM EDTA

2.4.2 Electrophoresis solutions

6x Glycerol dyes	30% v/v glycerol 0.1% w/v bromophenol blue 0.1% w/v xylene cyanol 5 mM EDTA (pH 7.5)
20x SSC	3 M NaCl 0.3 M Trisodium citrate
10x TBE	890 mM Tris base 890 mM Borate 20 mM EDTA (pH 8.0)

2.4.3 Media

All media were made up in nanopure water and either autoclaved or filter-sterilised prior to use. For agar used for bacterial growth 15 mg/ml bacto-agar were added to the appropriate media. Antibiotics were added to media as appropriate (see Table 2.1) to the following final concentrations: Kanamycin (purchased as a solution, stored at 4°C), 30 μg/ml; Chloramphenicol (stored at 4°C), 12.5 μg/ml, both supplied by Sigma).

Table 2-1 Clones and appropriate antibiotics

Clone type	Library	Antibiotic
Cosmid	LL0XNC01	Kanamycin
PAC	RPCI1,3,4,5, 6	Kanamycin
BAC	RPCI-11, 13, 23, 24. RZPD668 Sminthopsis *macroura* library.	Chloramphenicol

LB	10 mg/ml bacto-tryptone
	5 mg/ml yeast extract
	10 mg/ml NaCl
	(pH 7.4)
2 x TY	15 mg/ml bacto-tryptone
	10 mg/ml yeast extract
	5 mg/ml NaCl
	(pH 7.4)

2.4.4 DNA labelling and hybridisation solutions

Hybridisation buffer	6x SSC
	1% w/v N-lauroyl-sarcosine
	10x Denhardt's
	50 mM Tris-HCl (pH 7.4)
	10% w/v dextran sulphate

2.4.5 General DNA preparation solutions

GTE	50 mM glucose
	1 mM EDTA
	25 mM Tris-HCl (pH 8.0)
3 M K^+/5 M Ac^-	60 ml 5 M potassium acetate (pH 4.8)
	11.5 ml glacial acetic acid
	28.5 ml H_2O
Lysis buffer	50mM glucose
	10mM EDTA
	25mM Tris pH 8.0

2.4.6 FISH solutions

10x nick translation buffer	0.5 M Tris-HCl (pH 7.5)
	0.1 M $MgSO_4$,
	1 mM dithiothreitol,
	500 µg/ml bovine serum albumin
Lysis solution	5 parts 70mM NaOH
DNA fibre preparation	2 parts absolute ethanol
	(made up in distilled H_20)
Hybridisation buffer	50% deionised formamide
	2x SSC
	10% dextran sulphate
	0.1% Tween 20
	10 mM Tris (pH 7.4)
Cadenza wash solution	0.05% Tween 20
	4x SSC
Cadenza blocking buffer	1% w/v blocking reagent
	0.001% sodium azide (in Cadenza wash solution)
	50% formamide
	50% 2x SSC v/v
Cadenza Layer 1 solution	4 µg/ml avidin Texas Red DCS (Vector) (diluted in Cadenza blocking buffer)

Cadenza Layer 2 solution	4 µg/ml biotinylated anti-avidin D plus 1 µg/ml mouse anti-digoxigenin (Boehringer) (diluted in Cadenza blocking buffer)
Cadenza Layer 3 solution	4 µg/ml avidin Texas Red DCS plus 10µg/ml goat anti-mouse FITC conjugate (Sigma) (diluted in Cadenza blocking buffer)

2.5 Size markers

1 kb ladder (1 mg/ml) (Gibco BRL Life Technologies)
Contains 1 to 12 repeats of a 1,018 bp fragment and vector fragments from 75 to 1,636 bp to produce the following sized fragments in bp: 75, 142, 154, 200, 220, 298, 344, 394, 516/506, 1,018, 1,635, 2,036, 3,054, 4,072, 5,090, 6,108, 7,125, 8,144, 9,162, 10,180, 11,198, 12,216.

2.6 Hybridisation membranes and X-ray and photographic film

Amersham	Hybond-N™ Nylon (78 mm x 119 mm)
Polaroid	Polaroid 667 Professional film
Autoradiographs	Fuji RX medical X-ray film

2.7 Sources of genomic DNA

Human placental DNA for pre-reassociation (ready-sheared) was purchased from Sigma Chemical Co.. Human placental DNA for PCR was purchased from Sigma Chemical Co.. DNA from hybrid Clone 2D (Cl2D) that contains the entire X chromosome was kindly provided by Adam Whittacker. Mouse genomic DNA was purchased from Sigma Chemical Co.. Hamster genomic DNA was kindly provided by Frances Lovell and Christine Burrows, and mouse CotI DNA was kindly provided by Ruby Banerjee.

2.8 Sources of RNA

Total RNA was obtained from Ambion and Clontech. Tissue origins are given below in table 2-2.

Table 2-2 Sources of total RNA used for RT-PCR experiments

Supplier	Tissue panel number	Tissue	Supplier	Tissue panel number	Tissue
Clontech	1	Adrenal gland	Clontech	21	Fetal brain
Clontech	2	Bone marrow	Clontech	22	Fetal liver
Clontech	3	Brain (cerebellum)	Ambion	23	Adrenal gland
Clontech	4	Brain (whole)	Ambion	24	Bladder
Clontech	5	Fetal brain	Ambion	25	Brain
Clontech	6	Fetal liver	Ambion	26	Cervix
Clontech	7	Heart	Ambion	27	Colon
Clontech	8	Kidney	Ambion	28	Heart
Clontech	9	Liver	Ambion	29	Kidney
Clontech	10	Lung	Ambion	30	Liver
Clontech	11	Placenta	Ambion	31	Lung
Clontech	12	Prostate	Ambion	32	Ovary
Clontech	13	Salivary gland	Ambion	33	Pancreas
Clontech	14	Skeletal muscle	Ambion	34	Placenta
Clontech	15	Spleen	Ambion	35	Prostate
Clontech	16	Testis	Ambion	36	Skeletal muscle
Clontech	17	Thymus	Ambion	37	Small intestine
Clontech	18	Thyroid gland	Ambion	38	Spleen
Clontech	19	Trachea	Ambion	39	Stomach
Clontech	20	Uterus	Ambion	40	Testis

2.9 Sources of cells for *Sminthopsis macroura* and mouse FISH

Metaphase preparations of a *Sminthopsis macroura* cell line were a kind gift from Dr. Willem Rens (University of Cambridge). Mouse cells from a primary culture derived from the spleen were a kind gift from Dr. Ruby Banerjee (Wellcome Trust Sanger Institute).

2.10 Bacterial clone libraries

2.10.1 Cosmid libraries

Cosmids from the Lawrence Livermore flow-sorted X chromosome cosmid library (LL0XNC01) (prefixed 'cU') were kindly provided by Dave Vetrie and Elaine Kendall. Cosmids from a library constructed from a male with 5 X chromosomes (Holland *et al.*, 1993)(prefixed 'cV') were also kindly provided by Dave Vetrie and Elaine Kendall.

2.10.2 PAC and BAC libraries

The RPCI-1, RPCI-3, RPCI-4, RPCI-5 (prefixed 'dJ'), and RPCI-6 (prefixed 'dA') PAC libraries, and the RPCI-11 (prefixed 'bA') and RPCI-13 (prefixed 'bB') BAC libraries were used as a source of human derived PAC clones and BAC clones respectively in this thesis. Mouse-derived BAC clones were obtained from the RPCI-23 (prefixed 'bM') female C57BL/6J and RPCI-24 (prefixed 'bN') male C57BL/6J libraries, and *Sminthopsis macroura*-derived BAC clones were obtained from the RZPD668 library (prefixed 'bF'). The human and mouse libraries were all kindly provided by Pieter de Jong and Joe Catanese (see http://bacpac.chori.org/), and the marsupial library by MA Chapman. Libraries were imported and maintained by the Sanger Institute Clone Resources Group.

2.10.3 cDNA libraries

A range of up to 19 different cDNA libraries were used in this study (see Table 2-3). cDNA libraries were imported and maintained by Jacqueline Bye. Each library contains 500,000 cDNA clones, divided into 25 pools of 25,000 clones. Five pools were combined to form a superpool containing 100,000 clones. Prior to their use in PCR, each superpool was diluted 1:100 and 1:1000 in $T_{0.1}E$.

Table 2-3 cDNA libraries used for SSPCR

cDNA library code	cDNA library description	Vector	Source/ Reference
1. U	(Monocyte NOT activated-from a patient with promonocytic leukaemia) (U937+)	pCDM8	Simmons (1993)
2. H*	Placental, full term normal pregnancy (H9)	pH3M	Simmons (1993)
3. P	Adult brain	pCDNA1	Pfizer
4. DAU	B lymphoma (Daudi)	pH3M	Simmons (1993)
5. FB	Fetal brain	pCDNA1	Invitrogen
6. FL	Fetal liver	pcDNA1	Invitrogen
7. HL	Peripheral blood (HL60)	pCDNA1	Invitrogen
8. SK	Neuroblastoma cells	pCDNA1	Invitrogen
9. T	Testis	pCDM8	Clontech
10. FLU	Fetal lung	pCDNA1	Invitrogen
11. AL	Adult lung	pCDNA1	Clontech
12. UACT*	(Monocyte PMA activated – from a patient with promonocytic leukaemia) (U937act)	pCDM8	Simmons (1993)
13. YT*	HTLV-1+ve adult leukaemia T cell	pH3M	Simmons (1993)
14. NK*	Natural killer cell	pH3M	Simmons (1993)
15. HPB*	T cell from a patient with acute lymphocytic leukaemia (HPBALL)	pH3M	Simmons (1993)
16. BM*	Bone Marrow	pH3M	Simmons (1993)
17. DX3*	Melanoma	pH3M	Simmons (1993)
18. AH	Adult Heart	pcDNA3-Uni	Invitrogen
19. SI **	Small Intestine	pcDNA3	Stammers

* Generously provided by Dr Simmons, Oxford **(**Simmons *et al.*, 1993).

** Generously provided by Dr Stammers (Sanger Institute)

2.11 Primer sequences

All primers were synthesised in house by Dave Fraser or externally by Genset. Table 2-4 lists the vector-specific primers and sequences used in SSPCR. Appendices A to D lists the STSs used in this thesis, the sequence of each primer and the expected size in base pairs (bp) of each product, and the optimal annealing temperature (AT – given in °C). Where appropriate, the clones, or genes from which the STSs were derived are also listed.

Table 2-4 Vector-specific primer sequences for primers used in SSPCR

Primer Name	Primer Sequence	Vector
SP6PAC*	ATTTAGGTGACACTATAG	pcDNA3
pH3M1FP	CTTCTAGAGATCCCTCGA	pCDM8, pH3M
pH3M2FP	GCTCGGATCCACTAGTAA	pCDM8, pH3M
pH3M1RP	CTCTAGATGCATGCTCGA	pCDM8, pH3M
pH3M2RP	CGACCTGCAGGCGCAGAA	pCDM8, pH3M
pCDM8.RP	TAAGGTTCCTTCAGAAAG	pcDNA1, pCDM8
T7.2FP	AATACGACTCACTATAG	pCDM8, pcDNA1, pcDNA3

* designed by John Collins (Sanger Institute)

2.12 Key World Wide Web addresses

Baylor College of Medicine Search Launcher	http://searchlauncher.bcm.tmc.edu/
Baylor College of Medicine Sequencing Center	http://www.hgsc.bcm.tmc.edu/
British Columbia Genome Sequence Centre	http://www.bcgsc.bc.ca/
DOTTER	http://www.cgr.ki.se/cgr/groups/sonhammer/Dotter.html
ENSEMBL	http://www.ensembl.org

Genome Sequencing Center, St Louis	http://www.ibc.wustl.edu/ cgm/jcgm.html
Genome Sequencing Center, Jena	http://genome.imb-jena.de/
Genome Sequencing Center, Naples	http://hpced.area.na.cnr.it/grsl/
INTERPRO	http://www.ebi.ac.uk/interpro/scan.html
MPIMG, Berlin (X sequencing)	http://www.mpimg-berlin- dahlem.mpg.de/~xteam/
National Centre for Biotechnology Information	http://www.ncbi.nlm.nih.gov/
OMIM	http://www3.ncbi.nlm.nih.gov/Omim/
RepeatMasker	http://ftp.genome.washington.edu/ RM/webrepeatmaskerhelp.html
The Institute for Genome Research	http://www.tigr.org/
The Wellcome Trust Sanger Institute	http://www.sanger.ac.uk/
X Chromosome Mapping Project at the Sanger Institute	http://www.sanger.ac.uk/ HGP/ChrX/
European Bioinformatics Institute	http://www.ebi.ac.uk/
Human Genome Mapping Project- Resource Centre (HGMP-RC)	http://www.hgmp.mrc.ac.uk/
Mouse Genome Informatics	http://www.informatics.jax.org/
UCSC genome browser	http://genome.cse.ucsc.edu/

Methods

2.13 Isolation of bacterial clone DNA

2.13.1 Miniprep of BAC DNA

1. Ten ml of 2 x TY media, supplemented with 10 µl of (12.5 mg/ml) chloramphenicol, were inoculated with a scraping from a frozen bacterial glycerol stock. The culture was incubated overnight at 37°C with shaking.
2. A bacterial pellet was formed from the culture by centrifugation at 2000 g (Sorvall RT7, Du Pont Company Sorvall, Delaware US) at 4°C for 10 minutes. The supernatant was discarded into hycolin and the tube inverted on tissue to drain.
3. 200 µl lysis buffer were added to the pellet and pipetted gently to resuspend the pellet. The resulting suspension was transferred to a 1.5 ml microfuge tube and left to stand at room temperature for 10 minutes.
4. 400 µl of fresh 0.2 M NaOH/1% SDS were added to the cells and the lysate mixed gently by inversion. The sample was then incubated on ice for 5 minutes.
5. 300 µl of 3 M sodium acetate (pH 5.2) were added and the lysate mixed gently by inversion. The sample was then incubated on ice for 10 minutes.
6. The tube contents were pelleted by centrifugation in a microfuge at 13,000 rpm for 5 minutes in an Eppendorf microfuge. The clear supernatant was transferred to a fresh microfuge tube. This procedure was repeated twice (if the supernatant still was not clear, the sample was incubated on ice for another 10-30 minutes, and the centrifugation repeated).
7. 600 µl of isopropanol were added to the supernatant, the tube contents were mixed gently, and incubated at -70°C for 10 minutes or longer.
8. The sample was then subjected to centrifugation at 13,000 rpm in an Eppendorf microfuge for 5 minutes, the supernatant removed and the tube inverted on tissue to drain.
9. The pellet was resuspended in 200 µl 0.3 M sodium acetate (pH 7).
10. 200 µl of 50/50 v/v phenol/chloroform were added to the sample. The sample was then vortexed, and subjected to centrifugation for 3 minutes at 13,000 rpm in an Eppendorf microfuge. 150 µl of the aqueous phase (upper layer) were then transferred to a new tube.

11. 50 µl of 0.3 M sodium acetate (pH 7) were added to the remaining organic layer, the sample was vortexed and subjected to centrifugation in an Eppendorf microfuge for 2 minutes at 13,000 rpm. 50 µl of the aqueous phase were pooled with the aqueous layer from step 10.
12. 200 µl of isopropanol were added to the pooled aqueous phases, and mixed by inversion. The sample was then incubated at -70°C for 10 minutes.
13. The sample was subjected to centrifugation in an Eppendorf microfuge for 8 minutes at 13,000 rpm, the supernatant was removed, and the tube inverted on tissue to drain.
14. 500 µl of ice-cold, 70% ethanol were added, taking care not to disturb the pellet. The sample was subjected to centrifugation in an Eppendorf microfuge for 5 minutes at 13,000 rpm, the supernatant removed by aspiration, and the pellet dried at 37°C for approximately 30 minutes.
15. 50 µl of $T_{0.1}E$ containing 200 µg/ml RNAseA were added, and the tube “flicked” to resuspend the DNA pellet. The sample was then incubated at 55°C for 15 minutes.
16. Two µl of the sample were analysed by electophoresis on a 1% agarose gel. The remainder of the sample was stored at -20°C until required.

2.13.2 Microprep of BAC DNA for restriction digest fingerprinting

1. 500 µl of 2 x TY containing chloramphenicol (see Table 2-1) were added to a 96-well deep-well microtitre plate (COSTAR).
2. Each well was inoculated from a glycerol stock with either a 96-well inoculating tool, or a sterile cocktail stick. A plate sealer (Dynax) was placed on top of the plate to seal the wells, and the culture grown for 18 hours at 37°C with gentle shaking.
3. For each well, 250 µl of the overnight growth were transferred to a clean microtitre plate. The cells were collected by centrifugation (Sorvall RT7, Du Pont Company Sorvall, Delaware US) at 1550 g for 4 minutes.
4. For each well, the supernatant was removed and the pellet resuspended in 25 µl of GTE, by vortexing gently (a cocktail stick was used for resuspending pellets still attached to the plate).
5. 25 µl of GTE were added to each well and gently mixed. 25 µl of freshly prepared 0.2 M NaOH/1% SDS were added, mixed and left to stand for 5 minutes at room temperature.

6. 25 µl of 3 M K^+/5 M Ac^- (pH 5.0) were added, mixed and left at room temperature for 5 minutes. A plate sealer was placed on top of the plate and the plate was vortexed gently for 10 seconds.
7. A microtitre plate containing 100 µl of isopropanol was taped to the bottom of 2 µm filter-bottomed plate (Millipore cat. no. MAGVN2250). The total well volume of the sample was transferred to the filter-bottomed plate and the sample was filtered by centrifugation at 1550 g for 2 minutes at 20°C.
8. The filter-bottomed plate was removed and the microtitre plate was left at room temperature for 30 minutes, before being centrifuged at 1500 g for 20 minutes at 20°C.
9. The supernatant was removed and the DNA was dried by inverting the plate on clean tissue paper, ensuring no disruption of the pellet.
10. 100 µl of 70% ethanol were added to the dried DNA, mixed gently, and DNA precipitated by centrifugation at 1500 g for 10 minutes at 20°C. The wash was repeated for restriction digest fingerprinting. The supernatant was removed and the DNA dried as described in step 9.
11. 5 µl of freshly prepared $T_{0.1}E$ / 1 µg/ml RNase were added and mixed gently to resuspend the DNA. Samples were stored at -20°C.

2.14 Bacterial clone *Hin*d III/*Sau* 3A 1 fluorescent fingerprinting

1. For one 96-well microtitre plate of sample DNAs, three digest premixes were prepared, one for each fluorescent label, in three 1.5 ml microfuge tubes labelled TET, HEX and NED. Each premix contained 25.5 µl $T_{0.1}E$, 24.5 µl NEB2 buffer (as supplied by the manufacturer), 5.0 µl *Hin*d III (20 U/µL), 8.0 µl *Taq* FS (32 U/µl) and 3.0 µl *Sau* 3A I (30 U/µl), 4.0 µl of the appropriate ddA-dye. Each premix was mixed prior to being aliquotted.
2. 2 µl of the TET premix were added to wells A1-H4 of the microtitre plate containing sample DNAs using a Hamilton repeat dispenser. Similarly, 2 µl of the HEX premix were added to wells A5-H8, and 2 µl of the NED premix were added to wells A9-H12. The plate was covered with a plate sealer and the reaction mixed by gentle agitation on a vortex. In order to ensure the sample was in the bottom of the wells

the plate was centrifuged at 150 g for 10 seconds (Sorvall RT7, Du Pont Company Sorvall, Delaware US).

3. The reaction was incubated for 1 hour at 37°C.
4. To precipitate the DNA, 7 µl 0.3 M sodium acetate and 40 µl 96% ethanol were added to each well. For multiplexing the samples, rows 5 and 9 were added to row 1, rows 6 and 10 were added to row 2, rows 7 and 11 were added to row 3, and rows 8 and 12 were added to row 4 respectively, using a multichannel pipette.
5. The samples were incubated at room temperature for 30 minutes in the dark.
6. The samples were subjected to centrifugation at 1550 g for 20 minutes at 20°C to pellet the DNA.
7. The supernatants were discarded and the pellets dried by tapping the plate face down onto tissue paper.
8. The pellets were washed by adding 100 µl of 70% ethanol to each well, mixed gently by tapping the plate. The samples were subjected to centrifugation at 1550 g for 10 minutes at 20°C.
9. The supernatants were discarded and the pellet dried as described in step 7.
10. The DNAs were resuspended in 5 µl $T_{0.1}E$.
11. Prior to loading, 2 µl of the marker DNA (kind gift from Frances Lovell, Wellcome Trust Sanger Institute, see Section 2.5) were added to each sample using a Hamilton repeat dispenser. The samples were denatured for 10 minutes at 80°C. 1.00 µl of each sample were loaded on a 5% denaturing acrylamide gel and resolved on an ABI377 Automated DNA sequencer. Data were collected using the ABI Prism Collection Software v1.1.
12. After data collection, the gel image was transferred to a UNIX workstation for entry into IMAGE.

2.15 Agarose gel preparation and electrophoresis

1. Agarose gels were prepared in 1x TBE containing 250 ng/µl ethidium bromide and the appropriate percentage of agarose according to the size of fragments being separated: 2.5 % agarose gels were used for electrophoresis of fragments below 1 kb; 1.0 % agarose gels were used for analysis of larger fragments.

2. Electrophoresis was performed at 50 - 90 V for 15 - 45 minutes depending on the separation required.
3. Products were visualised by UV illumination.

2.16 Applications using the polymerase chain reaction

2.16.1 General primer design

Primers were designed using Primer3 (Rozen and Skaletsky, 2000), or manually using the following guidelines:

As far as possible, sequences chosen were 18 - 25 bp in length.

Sequences were chosen to avoid areas of simple sequence and obvious repetitive sequence i.e., runs of single nucleotide (e.g. TTTT) or double nucleotide (CGCGC) motifs.

Sequences were chosen to exclude palindromes which will form inhibitory secondary structure, especially at the 3' ends (e.g. GACGTC).

As far as possible, sequences were chosen with a GC content of at least 50%.

Sequences were chosen to avoid complementarity between pairs of primers, especially at the 3' end, which could result in primers annealing to each other and forming primer dimers.

If possible, sequences were chosen which would generate products of at least 100 bp in length.

2.16.2 Oligonucleotide preparation

All oligonucleotides used were synthesised in house by David Fraser or supplied as working dilutions from Genset. The concentration of the primer in ng/µl was determined by measuring the absorbance at 260 nm (Abs260) and multiplying this by 33 and any necessary dilutionfactor.

2.16.3 Amplification of genomic DNA by PCR

1. 1-10 ng/µl of genomic DNA were amplified in a reaction volume of 15 to 50 µl as required. Reactions contained approximately 1.3 µM of each oligonucleotide primer, 67 mM Tris-HCl (pH 8.8), 16.6 mM $(NH_4)_2S0_4$, 6.7 mM $MgCl_2$, 0.5 mM of each deoxyribonucleoside triphosphate (dATP, dCTP, dGTP, dTTP), 0.6 U of AmplitaqTM (Cetus Inc.). 10 mM β-mercaptoethanol and 170 µg/ml of BSA (Sigma Chemical Co., A-4628) were added to the reactions from freshly made stock solutions as the reactions were set up.

2. Unless specified otherwise, cycling conditions were as follows: all reactions were preceeded by an initial denaturing step of 5 minutes at 94°C, followed by 35 cycles of: 93°C for 30 seconds, [primer-specific annealing temperature] for 30 seconds, and 72°C for 30 seconds; followed by a final extension step of 5 minutes at 72°C. Primer-specific annealing temperatures are given for each primer pair in the text or in Appendices A-D.

3. PCR products were separated on 2.5% agarose minigels as described in Section 2.15 and visualised by ethidium bromide staining.

2.16.4 Colony PCR of STSs from bacterial clones

1. Colony PCR on bacterial clones was performed by touching a sterile toothpick onto the surface of a colony and stirring this into 200 µl of $T_{0.1}E$, and using 5 µl of the resulting suspension in a 15 µl final volume PCR (as described in Section 2.15.3).

2. PCR products were separated on 2.5% agarose minigels as described in Section 2.15 and visualised by ethidium bromide staining.

2.17 Radiolabelling of DNA probes by direct incorporation

PCR products were radiolabelled essentially as described in (Bentley *et al.*, 1992).

1. 10-15 µl of PCR product were separated on a 2.5% agarose minigel and visualised by ethidium bromide staining.
2. The gel was rinsed in deionised water to remove excess buffer. The desired band was excised from the gel and placed in 100 µl of $T_{0.1}E$ at 4°C overnight.
3. 5 µl of the $T_{0.1}E$ were used as template in the 15 µl PCR-labelling reaction containing 1.3 µM of each primer, 1.5 µl of 10x PCR buffer, 0.5 µl of [α-32P]-dCTP (3,000 Ci/mmol), 0.12 U of *Taq* polymerase (Cetus) and 0.5 mM each of dATP, dTTP and dGTP. Reactions were performed in a 0.5 ml microfuge tube and overlaid with mineral oil (Sigma) in a DNA thermal cycler (Perkin Elmer, USA).

4. PCR cycling conditions were as follows: 94°C for 5 minutes; followed by 20 cycles of: 93°C for 30 seconds, 55°C for 30 seconds, and 72°C for 30 seconds; followed by 72°C for 5 minutes.
5. For marsupial BAC screening, unincorporated nucleotides were then removed by elution on a Sephadex G50 column (Pharmacia Biotech). Probe was diluted to 400 μl in $T_{0.1}E$ and applied to the column. Five further additions of 400 μl $T_{0.1}E$ were made to the column, and the corresponding fractions were collected. Each fraction was counted using a scintillation counter (Easicount4000, Scotlab, UK). The labelled probe was eluted in fraction 2 which was used for subsequent hybridisations. The percentage incorporation of [α-32P]-dCTP into the labelled probe was calculated (count of fraction2/total counts for all fractions).
6. All probes were boiled for 5 minutes and snap-chilled on ice prior to use.

2.18 Hybridisation of radiolabelled DNA probes

2.18.1 Hybridisation of DNA probes derived from STSs to same-species BAC filters

1. Filters were pre-hybridised tightly rolled in 15 ml Sterilin tubes or flat in sandwich boxes for 3 hours in sufficient hybridisation buffer to cover the filters at 65°C with gentle shaking.
2. Radiolabelled probe was added and hybridised to the filters for greater than 16 hrs at 65°C in a shaking incubator.
3. Filters were washed twice at room temperature in 2x SSC for 5 minutes, twice at 65°C in 0.5 x SSC, 1% sarkosyl for 30 minutes. Filters were rinsed at room temperature in 0.2x SSC prior to draining the excess liquid, wrapping in Saran wrap (Dow Chemical Co.) and exposing to autoradiograph film.

2.18.2 Hybridisation of DNA probes derived from human STSs to Sminthopsis macroura BAC filters

1. Filters were prehybridised tightly rolled in 15 ml Sterilin tubes or flat in sandwich boxes for 3 hours in sufficient hybridisation buffer to cover the filters at 58°C with gentle shaking.
2. Radiolabelled probe was added and hybridised to the filters for greater than 16 hours at 58°C with gentle shaking.

3. Filters were washed twice at room temperature in 2x SSC for 5 minutes, once at 58°C in 1.5 x SSC, 1% sarkosyl for 30 minutes, and once at 58°C in 1 x SSC, 1% sarkosyl for 30 minutes (unless as directed in the text). Filters were rinsed at room temperature in 0.2x SSC prior to draining the excess liquid, wrapping in Saran wrap (Dow Chemical Co.) and exposing to autoradiograph film.

2.18.3 Stripping radiolabelled probes from hybridisation filters

Filters were washed in 0.4 M NaOH for 30 minutes at 42°C followed by 30 minutes in 0.2 M Tris-HCl (pH 7.4), 0.1x SSC, and 0.1% w/v SDS at 42°C with gentle shaking. Successful removal of radiolabelled probe was assessed by autoradiography.

2.19 Clone library screening

2.19.1 cDNA library screening by PCR

1. Nineteen different cDNA libraries were subdivided into 25 subpools of 20,000 clones, which were then combined to produce 5 superpools of 100,000 clones by J. Bye and S. Rhodes. Details of the cDNA libraries are given in Table 2.3.
2. Aliquots of the superpools of each library were arranged in a microtitre plate to facilitate subsequent manipulations and gel-loading post PCR with a multi-channel pipetting device.
3. In the primary screen, 5 µl of each superpool were used as template in a 15 µl final volume PCR using buffer and PCR conditions as described in Section 2.16.3.
4. PCR products were loaded on 20 cm x 20 cm 2.5% agarose horizontal slab gels using an 8-way multi-channel pipetting device, separated by electrophoresis and visualised by ethidium bromide staining.
5. In the secondary screen, 5 µl of each of the 5 subpools of 20,000 clones corresponding to the superpool that were positive in the first round, were screened by PCR with the same primer pair as used in step 2. PCR products were separated by electrophoresis through 2.5% agarose gels and visualised by ethidium bromide staining.

2.19.2 Single-sided specificity PCR (SSPCR) of cDNA

The principle of SSPCR (Huang *et al.* 1993) is illustrated in Figure 2-1.

1. SSPCR was performed on the subpools of the cDNA libraries, each containing 20,000 clones. Prior to their use in PCR, the subpools were diluted 1:10 in $T_{0.1}$E and boiled. Dilutions were stored at -20°C until required. On removing from -20°C, tubes were centrifuged briefly in a microfuge to settle the contents and then mixed carefully when thawed.
2. In the first round, PCR was performed using 1 µl of the diluted subpools as template in a 15 µl final volume using buffer conditions as described in Section 2.16.3. The primer combinations used are given in Table 2-5.
3. PCR was performed in microtitre plates in a DNA thermocycler (Omnigene) using hot-start. Cycling conditions were as follows: an initial denaturation step of 5 minutes at 95°C, followed by 25 cycles of: 94°C for 30 seconds, 60°C for 30 seconds, and 72°C for 3 minutes; followed by a final step of 10 minutes at 72°C.
4. For the second round of PCR, products from the first round were diluted 1 in 50 and 1 in 500 in $T_{0.1}$E. 5 µl of each dilution were used as template in 15 µl final volume PCR using buffer conditions as described in Section 2.16.3. Cycling conditions were as described in step 3. The primer combinations used are given in Table 2-5.

Table 2-5 Primer combinations used in SSPCR*

First round SSPCR	Second round SSPCR
Specific primer A and FP vector primer	Specific primer B and FP vector primer
Specific primer A and RP vector primer	Specific primer B and RP vector primer
Specific primer C and FP vector primer	Specific primer D and FP vector primer
Specific primer C and RP vector primer	Specific primer D and RP vector primer

*Primer sequences are given in Table 2-4.

5. 5 µl of the second-round PCR products were separated by electrophoresis through either 1% or 2.5% agarose gels depending on product size and visualised by ethidium bromide staining. Products were gel purified using the Qiaquick gel extraction kit (Qiagen™) prior to sequencing directly.

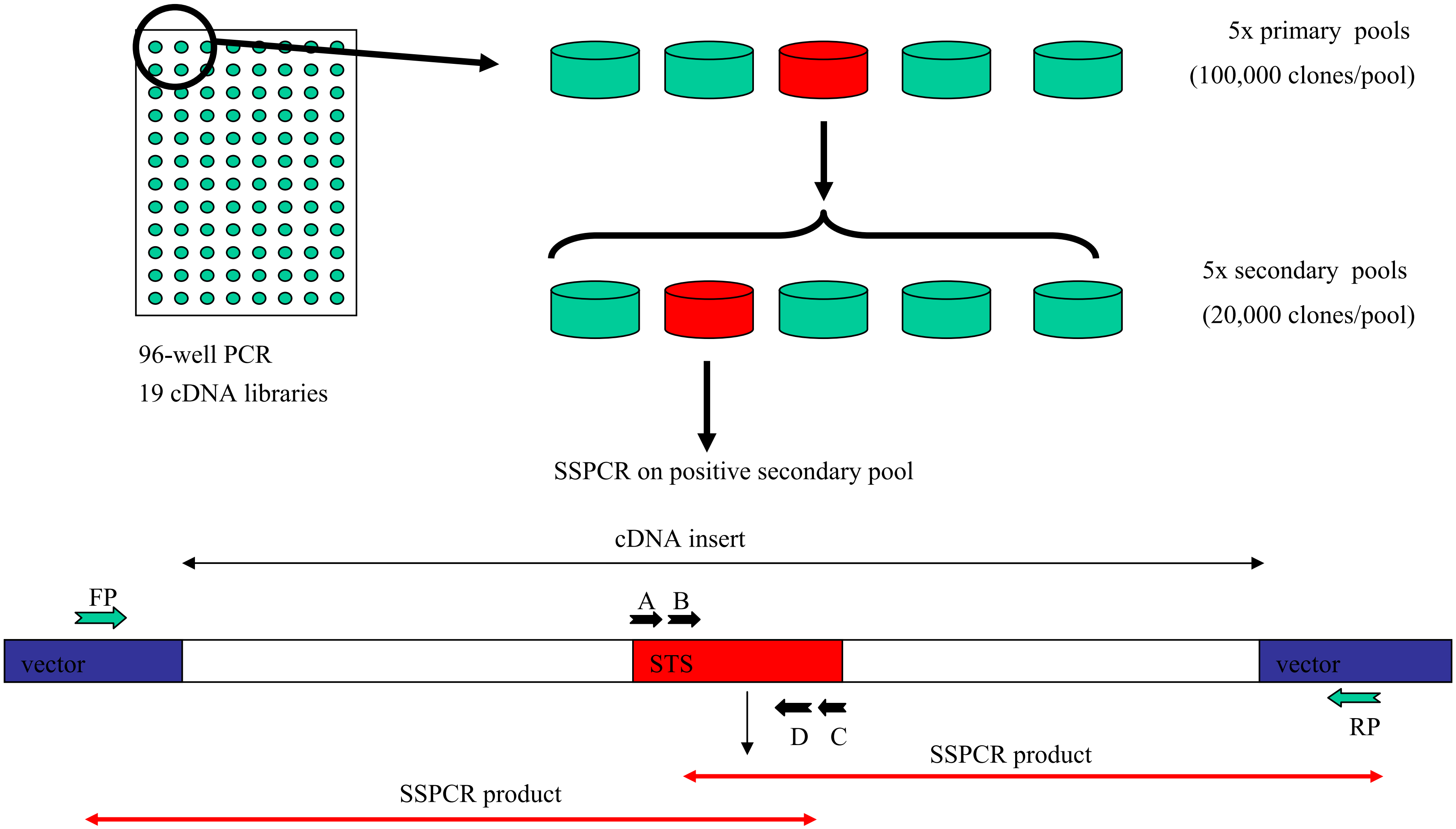


Figure 2-1 Diagram illustrating the SSPCR procedure described in Section 2.19.2.

2.20 RT-PCR expression profiling using total RNA samples

2.20.1 Generation of cDNA from total RNA

2.20.1.1 RNA treatment

1. RNA obtained from Clontech was subjected to DNase treatment as described in Section 2.20.1.2 (Ambion RNA was supplied pre-treated with DNase). RNA was stored at -70°C until required.
2. cDNA synthesis reaction omitting RT was performed as described in Section 2.20.1.2 for RNA samples 1-20 and tested by PCR using various non intron-spanning STSs.

2.20.1.2 DNase treatment of total RNA samples

1. Total RNA samples were treated with DNase to remove contaminating genomic DNA as required, using a DNA-Free kit (Ambion), as per the manufacturers' instructions.

2.20.1.3 Generation of single-stranded cDNA from total RNA

1. Single-stranded cDNA was generated from 0.5 μg–2 μg of total RNA using a Superscript II cDNA synthesis kit (Invitrogen Life Technologies) as per the manufacturers' instructions.
2. The resultant cDNA was diluted in $T_{0.1}E$ to give 200 μl/ug, and stored at -20°C. 5μl of the cDNA were then used per reaction in subsequent PCR.

2.20.2 Screening of cDNA samples by PCR

1. PCR was performed in a total volume of 15 μl using 5 μl of single-stranded cDNA as a template (Section 2.20.1.2) in a 96-well format essentially as described in Section 2.16.3. Primer annealing temperatures used are given in Appendix C. Unless otherwise noted in the text, 35 cycles of PCR were performed.

2.20.2.1 Visualisation of PCR products by agarose gel electrophoresis and Vistra Green post-staining

1. PCR products were separated by agarose gel electrophoresis as described in Section 2.15, except ethidium bromide was omitted and electrophoresis tanks were rinsed with distilled H_20 prior to use.
2. Following electrophoresis, gels were stained in Vistra Green (50 µl Vistra Green, 5 ml 1M Tris-HCl, 0.5 mM 0.1 M EDTA made up to 500 ml in doubly-distilled H_20) for ~30-45 minutes with gentle rocking, rinsed with distilled H_20 and scanned on a FluorImager SI.

2.21 FISH of BAC-derived probes to Sminthopsis macroura metaphase preparations

2.21.1 Metaphase slide preparation

1. A coplin jar containing freshly prepared 3:1 methanol/acetic acid fixative was prepared. Microscope slides were soaked in 96% ethanol and dried using lint-free tissue then lain horizontally on a humid-tray (containing tissue soaked in DDW).
2. The *Sminthopsis macroura* cell line suspension was mixed and the appropriate numbers of drops were added to the slide using a pasteur pipette (most slides were prepared using two single drops placed side-by-side).
3. The slides were examined under a phase-contrast microscope and the limits of the cell-spot noted on the slide using a pencil. The slides were placed in a coplin jar containing 3:1 methanol/acetic acid fixative for 30-60 minutes, then air-dried.
4. Slides were then passed through a 70%, 70%, 90%, 90%, 100% ethanol series (one minute in each) by dipping in a series of coplin jars and then air-dried.
5. After fixing in acetone for 10 minutes, slides were kept at room temperature in a sealed box containing dessicant for at least 24 hours before use.

2.21.2 Nick-translation labelling of BAC clones

1. 25 µl reactions were prepared for each clone. Approximately 1 µg of BAC clone DNA, was used per reaction.

2. The following reagents were added to a 1.5ml microfuge tube on ice: approximately 1 μg BAC clone DNA, 2.5 μl 10x nick translation buffer, 1.9 μl 0.5 mM dNTPs, 0.7 μl 1 mM hapten-conjugated dUTP, 1 μl DNAseI working solution (titrated to give fragment smears of 200-700 bp under appropriate incubation times), 0.5 μl DNA polymerase (10 U/μl) and sterile distilled water to give a final volume of 25μl.
3. The tube contents were gently mixed by flicking, briefly centrifuged in a microfuge and then incubated at 14°C for the appropriate length of time determined by DNAseI titration.
4. 2.5 μl of 0.5 M EDTA were added to the reactions to inactivate the enzymes. 2.5 μl of 3M sodium acetate (pH 7) and 1 ml ice-cold absolute ethanol were added and mixed. The reactions were incubated at -70°C for 30 minutes to precipitate the DNA.
5. The reactions were subjected to centrifugation at 13,000 rpm for 10 minutes in a microfuge, and the supernatant removed. 500 μl ice-cold 70% ethanol were added to the pellet, and the tubes were then centrifuged at 13,000 rpm in a microfuge for 2 minutes.
6. The supernatant was removed carefully, and the pellets air-dried. 10 μl of $T_{0.1}E$ were added to each pellet, and incubated on ice for 10 minutes. The tube was then flicked to resuspend the pellet, and 2 μl were removed for analysis by electrophoresis on a 1 % agarose gel. Samples were stored at -20°C until required.

2.21.3 Hybridisation of FISH probes to Sminthopsis macroura metaphase slides

1. For each hybridisation, the following were added to a 0.5 ml microfuge tube: 0.5 μl each nick-translated probe (prepared as described in Section 2.21.2), competing DNA as required, FISH hybridisation buffer to give a final volume of 13 μl. The contents of the tubes were then mixed by vortexing and briefly subjected to pulsed centrifugation in a microfuge.
2. Probes were denatured by incubating the reactions at 65°C for 10 minutes. Pre-annealing was then performed by incubating the reactions at 37°C for 40 minutes.

3. Metaphase slides, prepared as described in Section 2.21.1, were denatured by incubation in 70% formamide at 65°C for the appropriate length of time (1 minute and 5 seconds unless directed otherwise in the text).
4. Slides were immediately transferred to ice-cold 70% ethanol, then passed through a 70%, 70%, 90%, 90%, 100% ethanol series (one minute in each) by dipping in a series of coplin jars and air-dried.
5. Each probe mix was pipetted onto a 22 x 22 mm coverslip (one per cell spot contained on the slide as described in Section 2.21.1), and placed onto a slide. The coverslip was then sealed in place with rubber cement and the slide incubated at 37°C overnight in a humidified incubator.

2.21.4 Washing of Sminthopsis macroura metaphase slides and detection of signal

1. After removal of rubber cement, the slides were soaked in 2x SSC for 15 minute to remove the coverslips.
2. The slides were then washed as follows: 2x SSC at 42°C for 5minutes, 50% formamide at 42°C for 5minutes, a second wash of 50% formamide at 42°C for 5minutes, a second wash of 2x SSC at 42°C for 5minutes.
3. Slides were then loaded into a Cadenza rack for staining (See Section 2.21.5).
4. Following staining, slides were washed briefly with 2x SSC, then stained in 0.08 μg/ml DAPI solution for 3 minutes. Slides were rinsed in 2x SSC, and passed through a 70%, 70%, 90%, 90%, 100% ethanol series (one minute in each) by dipping in a series of coplin jars and air-dried.
5. For each slide, 3 drops of antifade (Citifluor AF1) were applied to a 22 x 50 mm coverslip, which was then overlaid onto the slide and sealed in place using nail varnish. Slides were left at 4°C for at least 30 minutes before image capture. Slides were stored at 4°C until required.
6. Images were obtained using a CoolSnap HQ camera (Photometrics) on an Axioplan 2 microscope (Zeiss), and analysed using SmartCapture 2 (Digital Scientific).

2.21.5 Two-colour detection using a Cadenza instrument

1. Rubber cement was removed from slides and coverslips soaked off at room temperature in 2x SSC for 15 minutes.
2. Slides were incubated in 2x SSC at 42^{0}C for 5 minutes in a Coplin jar.
3. Slides were incubated in 50% formamide (in a Coplin jar) at 42^{0}C for 5 minutes.
4. Slides were incubated in 50% formamide (in a second Coplin jar) at 42^{0}C for 5 minutes.
5. Slides were washed in 2x SSC at 42^{0}C for 5 minutes.
6. The Cadenza unit reservoir was topped up with Cadenza wash solution, the slides were mounted to Cadenza coverplates, and 2xSSC was pipetted into each the top of each Cadenza coverplate/slide reservoir to check mounting.
7. The reagent carousel was loaded: Position 1 – layer 1 solution, position 2 – layer 2 solution, position 3 – layer 3 solution, position 4 – blocking buffer (100 µl solution per slide plus 500 µl excess was prepared for each solution).
8. The detection run was then performed using the Cadenza 3-layer detection protocol.

2.22 Fibre-FISH of BAC-derived probes to Mus musculus DNA fibres

2.22.1 Mus musculus DNA fibre slide preparation

1. 2-3 mls cell suspension were taken by disposable pasteur pipette and subjected to centrifugation at 1200 rpm for 5 minutes.
2. Supernatant was removed and the pellet resuspended in PBS, then subjected to centrifugation at 1200 rpm for 5 minutes. This step was repeated twice.
3. Supernatant was removed and the pellet resuspended in 1 ml PBS.
4. An aliquot of cells were counted using a haemocytometer, and the cell suspension was diluted or concentrated accordingly to give a concentration of approximately 2-3x10^6 cells/ml.
5. 10 µl cell suspension were spread over a 1 cm area on the upper part of a clean microscope slide.
6. The slides were left to air dry at room temperature for 30 minutes.

7. The slides were fitted to Cadenza coverplates (cell spot uppermost) and clamped in a vertical position in a rack. 150 µl lysis solution were then added to the top of the slide.
8. Once the level of lysis solution dropped below the upper frosted region of the slide, 150 µl 96 % ethanol were added to the top of the slide, and allowed to drain.
9. Once the meniscus stopped falling, the slide and Cadenza unit were removed from the rack. The top of the slide was pulled gently away from the Cadenza coverplate, allowing the meniscus to move down the slide.
10. Slides were air-dried, and then fixed in acetone in a Coplin jar for 10 minutes.
11. Limits of the fibres were marked on the slide by pencil, and slides were left at room temperature for at least 24 hours before hybridising.

2.22.2 Hybridisation of FISH probes to Mus musculus DNA fibre slides

1. Probe master mix was prepared for each probe by mixing 4 µl labelled *Mus musculus* BAC clone (see section 2.21.2), 6 µl mouse CotI DNA and 40 µl FISH hybridisation buffer.
2. 5 µl of each probe master mix were then combined and the volume adjusted to 15 µl with FISH hybridisation buffer. Hybridisation was then performed as described in Section 2.21.3, step 2. Washing and two-colour detection were performed as described in Sections 2.21.4 and 2.21.5.

2.23 Mapping and sequence analysis software and databases

2.23.1 IMAGE

All processing of fingerprinting gels was carried out using IMAGE. IMAGE processed gels from fluorescent fingerprinting and extracted a normalised band pattern for each lane on a gel. Several procedures were run on each gel in turn:

Lane tracking – a grid was superimposed on the gel image and the grid manually edited to ensure it exactly matched the lanes on the gel.

Band calling – an analysis module traced the band pattern along the lanes and tried to identify the bands. Manual editing ensured the correct bands are chosen.

Marker locking – in order to compare band patterns from one gel to another all band positions were normalised to one master gel. A set of DNA fragments of known length or migration distance was loaded as a marker lane. Manual editing ensured the standard pattern matched to the pattern from the master gel.

Normalisation – once the marker lane patterns were locked onto the standard lane, the band positions of the sample lanes were normalised so that each lane appeared to have been run on the master gel with all distortions cancelled out. IMAGE finally generated a 'Bands' file for each gel containing normalised migration distances for all selected bands in each clone lane.

2.23.2 FPC

All contig construction and visualisation described in this thesis was performed in FPC (Soderlund *et al.*, 1997). FPC took as the input a set of clones and their restriction fragments (called Bands) from IMAGE. Each fingerprint pattern for each clone is compared to the fingerprint patterns of all other clones in the database. The relationship between two clones was reported as a probability of coincidence, i.e. the probability that two clones overlap by chance. Two variables can be set to filter the reported overlaps:

Cut off – a match between two clones will only be reported if the probability of coincidence is less than or equal to the cut off. When analysing matches between *Sminthopsis macroura* BAC clones, the tolerance was set to 1e –04, and when analysing larger insert *Mus musculus* BAC clones, the tolerance was decreased to 1e –14.

Tolerance – two bands are considered as their migration distances differ by less than tolerance. For the analysis carried out in this thesis the tolerance was set to 7.

Overlapping clones were identified automatically and contigs were constructed manually using the available editing tools provided by FPC. A minimum set of clones for sequencing was chosen where appropriate based on a combination of shared bands and shared marker data.

2.23.3 Xace and other custom mapping databases

Human X chromosome data and annotation generated in this thesis were stored in Xace, a chromosome-specific implementation of ACeDB. Other ACeDB implementations were used to store mouse and marsupial mapping and annotation data, as described in Chapter 4 and 6. ACeDB was originally developed for the *C. elega*ns genome project (Durbin and Thierry-Mieg, 1991). Documentation code and data available from anonymous FTP servers at lirmm.lirmm.fr,cele.mrc-lmb.cam.ac.uk and ncbi.nlm.nih.gov).

ACeDB works using a system of windows and presents data in different types of windows according to the type of data. All windows are linked in a hypertext fashion, so that clicking on an object will display further information about that object. For example, clicking on a region of a chromosome map will highlight landmarks mapping to that part of the chromosome; clicking on a landmark will display information about that landmark including landmark-clone associations, etc.

All PAC, BAC and cosmid library filters and polygrids are represented graphically in Xace where each square on the grid represents an individual clone. Hybridisation data were entered directly via the grid. Data were then saved in the database establishing landmark-to-clone associations that can be displayed as text windows relating to either the landmark or the clone. Data can also be entered via text windows or via an internal web page. PCR library pool screening and colony PCR results were entered via the text windows.

In addition to the data generated by the X chromosome mapping group, Xace also contains displays of published X chromosome maps. Genomic sequence data is also displayed in ACeDB along with the collated results from the computational sequence analysis performed by the Sanger Institute Human Sequence Analysis Group.

Xace can be accessed by following the instructions at:
http://www.sanger.ac.uk/HGP/ChrX.

2.23.4 BLIXEM

Individual matches identified as a result of similarity searches using the BLAST algorithm, or matches between sequences of cDNA clones or PCR products amplified from genomic DNA generated as part of the project, were viewed in more detail using

BLIXEM. BLIXEM, (Blast matches In an X-windows Embedded Multiple alignment) is an interactive browser of pairwise Blast matches displayed as a multiple alignment. Either protein or DNA matches can be viewed in this way at either the amino acid or nucleotide level respectively. BLIXEM contains two main displays: the bottom display panel shows the actual alignment of the matches to the genomic DNA sequence, and the top display shows the relative position of the sequence being viewed within the context of the larger region of genomic DNA. A program "EFETCH" retrieves the record from an external database (e.g. EMBL, SWISSPROT).

2.23.5 RepeatMasker

Human repeat sequences were masked using RepeatMasker, a program that screens DNA sequence for interspersed repeats and low complexity DNA sequence (Smit & Green RepeatMasker at http://ftp.genome.washington.edu/RM/RepeatMasker.html). The output of the program is a detailed annotation of the repeats that are present in the query sequence and a version of the sequence with repeats masked by "N" characters. Sequence comparisons are performed by the program cross_match, an implementation of the Smith-Waterman-Gotoh algorithm developed by P. Green. The interspersed repeat databases screened by RepeatMasker are based on the repeat databases (Repbase Update (Jurka, 2000)) copyrighted by the Genetic Information Research Institute.

2.24 Alignment of nucleic acid and protein sequences and phylogenetic analysis

2.24.1 Alignment of nucleic acid and protein sequences

Nucleic acid and protein sequences were aligned using the program ClustalW (Thompson *et al.*, 1994) via a web-based server at the EBI (http://www.ebi.ac.uk/), or ClustalX installed locally on a PC unless otherwise noted in the text (Thompson *et al.*, 1997). User-defined parameters were left at their default settings unless directed otherwise in the text. Alignments were then manually edited and presented using the program GeneDoc (Nicholas, 1997).

2.24.2 Calculation of sequence identities and similarities

Nucleic acid and protein sequences were aligned as described in section 2.24.1. The "statistics report" function of GeneDoc was then used to calculate and display sequence identities and similarities.

2.24.3 Phylogenetic analysis of nucleic acid and protein sequences

Nucleic acid and protein sequence alignments were subjected to various phylogenetic analyses to estimate their order of relationship. In each case, alignments produced as described in Section 2.24.1 were manually edited as necessary to minimise the number of gaps, and the most reliably aligned region of the alignment was then used for the respective phylogenetic analyses. Any columns within the alignment containing gaps were removed prior to phylogenetic analysis.

2.24.3.1 Neighbour-joining distance analysis

Phylogenetic analyses were performed using the program MEGA2 (Kumar *et al*., 2001) installed locally on a PC. This package combines various phylogenetic analysis methodologies in a straightforward interface. For distance-based phylogenetic analysis, a Neighbour-Joining method was employed (Saitou and Nei, 1987). The distance measure used for peptide sequences was the poisson-corrected measure, to attempt to account for multiple substitutions at sites when distantly related sequences were compared. For nucleotide comparisons, the Kimura 2-paramater measure was used, considering both transitions and transversions. In all analyses, 1000 bootstrap replicates were selected to assess robustness of the tree produced.

2.24.3.2 Maximum likelihood analysis

Maximum-likelihood analysis was performed using the package PIE via a web-server at the MRC Rosalind Franklin Centre for Genomics Research. For protein sequences, the program PROTML was used, using the JTT substitution model. For nucleotide sequences, the program DNAML was used. For each type of analysis, a search for the best tree was performed, 100 bootstrap replicates were chosen, and sites were not weighted.

2.25 Comparative sequence analysis

2.25.1 PIPmaker

PIP plots were generated using PIPmaker as per the authors instructions (Schwartz *et al.*, 2000). Text files were generated containing relevant sequences in fasta format, and an annotation file was generated as per the authors instructions. The annotation file was also used to generate an underlay file as per the authors instructions. The base sequence (human unless otherwise specified) was masked for repeats using RepeatMasker. Most program parameters were as default, except sequences were searched on both strands, and chaining was employed. Chaining reports only those matches occurring in the same order in the different species, and avoids build-up of matches due to repetitive sequence occurring throughout the sequences. Chaining assumes that the order of matches should be conserved.

The "high sensitivity" setting was also employed.

Key to PIP annotations (reproduced from authors web-site):

White pointed box – L1 repeat, Light grey triangle – SINEs, Black triangle – MIR, Black pointed box – LINE2, Dark grey triangle and pointed boxes – other interspersed repeats (e.g. LTRs, DNA transposons), Short dark grey box – CpG island with CpG/GpC > 0.75, Short white box - CpG island with CpG/GpC between 0.6-0.75.

2.25.2 VISTA

VISTA plots were generated as per the authors instructions (Mayor *et al.*, 2000). Input files were the same as those generated for PIPmaker as described above. The Fugu and Human Xp sequences were reverse-complemented. Unless otherwise specified in the text, a window level of 100 bp was used with a conservation level of 75%, minimum conservation level of 50% and maximum difference was left as default. Percentage identity was plotted on the y-axis.

Chapter Three - Gene annotation of the human Xq22-q23 region

3.1 Introduction

As this project began, the sequencing of chromosome 22 was nearing completion. This was the first human chromosome sequence to be completed (Dunham *et al.,* 1999). At this time, the human X chromosome sequence was in a relatively unfinished state (~ 48 % finished sequence), and spanned by many sequence-ready contigs. There were regions however with large segments of contiguous finished sequence, and one of these was selected for studies utilising the genomic sequence to guide identification of genes. These efforts are described in this Chapter.

The region chosen for study was the human Xq22-q23 region. Comprising of approximately 15 Mb of euchromatic DNA, the region begins and ends in dark-staining Giemsa bands (G-band) but is predominantly a light G-band, containing within it a "grey" G-band. From studies of the composition of the sequence within Xq21.3-q22.2 (G. R. Howell, PhD thesis, Open University), the GC content of Xq22.1 remains above 38% (consistently higher than the genome average of 41% (Lander *et al.*, 2001), with a variable LINE and SINE content. In general, Xq22.1 appeared to show a higher % GC and SINE and lower LINE content compared to Xq21.3 and Xq22.2. From these characteristics, it was expected that the region would be relatively gene-rich, and that differences in gene size and density may be observed in the dark/grey/light G-band transitions. Initially, the region was spanned by three bacterial clone contigs (see Figure 3-1), including the largest contig on the chromosome (G.R. Howell, PhD thesis, Open University). Within this study, efforts were undertaken to close the gap between contigs Xctg200 and Xctg18. An STS designed to a PAC clone (dJ19N1) in Xctg18 identified clones in a small unassigned contig (Xctg1057) following hybridisation to filters of X chromosome allocated clones (polygrids). Xctg1057 was then found to share fingerprint bands with GSCX Ctg17241, which in turn shared bands with contig Xctg200 (fpc analysis performed by Adam Whittaker, Wellcome Trust Sanger Institute). This closed the gap between contigs Xctg18 and Xctg200.

In addition to containing large regions of contiguous sequence, which are ideal for large-scale genome-based gene identification, many disease genes had been mapped to the region. The genes for several of these conditions remained un-cloned. These include DFN2 (OMIM:304500), X-linked megalocornea (OMIM:309300), EFMR (OMIM:300088), MRX53 (OMIM:300324) and an X-linked mental retardation syndrome with seizures, hypogammaglobulinemia and progressive gait disturbance (Chudley *et al.*, 1999).

The most comprehensive transcript map of the region to that point had identified 30 genes and 56 additional expressed sequences, from STS (derived from genes and ESTs) screening of YACs mapped to the region (Srivastava, *et al.,* 1999). A comprehensive set of annotated gene structures would thus provide useful data for disease gene mapping and mutation screening projects. An example of this is where genes were assessed as candidates for the hereditary deafness disorder, DFN2, as part of a collaboration with Dr. Jess Tyson (Institute of Child Health, London).

During the gene identification studies presented here, various loci within the region provided illustrations of elements of genomic organisation. Some examples are presented here, including an example of an insertion of an almost complete copy of the mitochondrial genome into the nuclear genome, examples of alternative polyadenylation sites, a novel, inverted repeat containing a well-studied gene (NXF2), and evidence for a gene fusion event involving this gene.

Landmark-based mapping and restriction fingerprinting had been used to generate bacterial clone contigs, which were positioned on the physical and genetic X chromosome maps (Bentley *et al.,* 2001). In Xq22-q23, these clones included cosmids, P1-Artificial Chromosomes (PACs) and Bacterial Artificial Chromosomes (BACs). A set of minimally overlapping clones (*tiling path*) had been picked for sequencing at the Sanger Institute, using the following approach: clones are sheared and shotgun-cloned into a sequencing vector; these sub-clones are sequenced, and their sequences are assembled into contigs using the alignment program PHRAP (P. Green, University of Washington); finally, remaining sequence gaps or ambiguities are resolved by directed sequencing (“finishing”) of genomic templates.

Following the closure of the contig gap described above, gene identification efforts focussed on the Xq22-q23 region spanned by markers *DXS*1510 and *DXS*8088, now spanned by two sequence-ready contigs.

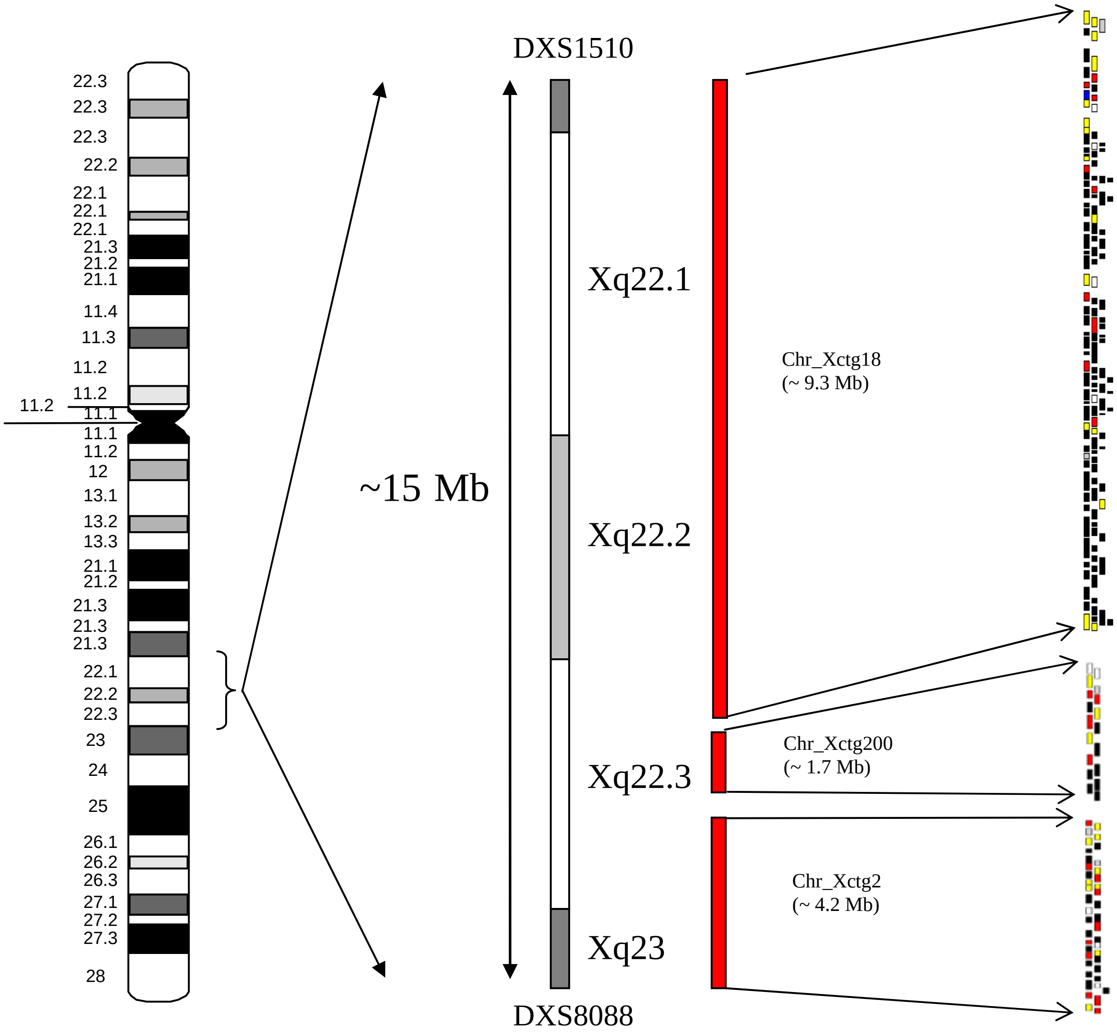


Figure 3-1 G-banded ideogram of the human X chromosome (Francke), illustrating the region Xq22-q23 chosen for this study. The ~15 Mb region bounded by markers *DXS*1510 (Cen) and *DXS*8088 (Tel) is shown. The three sequence-ready contigs spanning the region are shown (red bars) and a depiction of the tiling paths given at the far right (black – finished and submitted clones, red – finished clones, other colours – unfinished clones).

3.2 Generation of an annotated gene map of human Xq22-q23

Finished sequences of genomic clones from the region were analysed on a clone-by-clone basis for protein and mRNA homologies (using BLAST with repeat-masked sequence genomic sequence) to sequences in EMBL, TrEMBL and SwissProt. The sequence was also analysed for repeats (using RepeatMasker to search RepBase (Jurka, 2000)) and GC content (using unmasked sequence). Gene prediction programs (GENSCAN, FGENESH) and exon prediction programs (GRAIL) were also used to analyse the sequence (unmasked sequence). This analysis was performed by the

Informatics Group, Wellcome Trust Sanger Institute. Sequences from 230 finished clones were analysed by this approach.

All sequence analysis results were collated in an ACeDB database, Xace (Human Genetics Informatics group, Wellcome Trust Sanger Institute). The 230 finished clone sequences, comprising approximately 14.8 Mb of finished sequence, were systematically manually analysed in the Xace viewer for features indicating potential genes. These features included: overlapping gene predictions from GENSCAN and FGENESH (indicating an increased confidence in the prediction being a true positive), mRNA/EST sequences matching to the genomic sequence and indicating splicing, or protein homologies to the genomic sequence. An example of a typical sequence view is shown in Figure 3-2.

When matching mRNA sequences were found representing genes, the gene structure was annotated using annotation tools within Xace. Protein and mRNA matches were visualised using BLIXEM, a BLAST result visualisation tool within Xace. An example of this visualisation is shown in Figure 3-3. If a gene could be annotated from a single mRNA, the gene was termed a "gene" and the locus designated "GD_mRNA" in Xace. Where homologies were found to proteins that included frameshifts or stop codons in the genomic sequence homologous region, these sequences were annotated as "pseudogenes", and termed "pseudogene" in Xace.

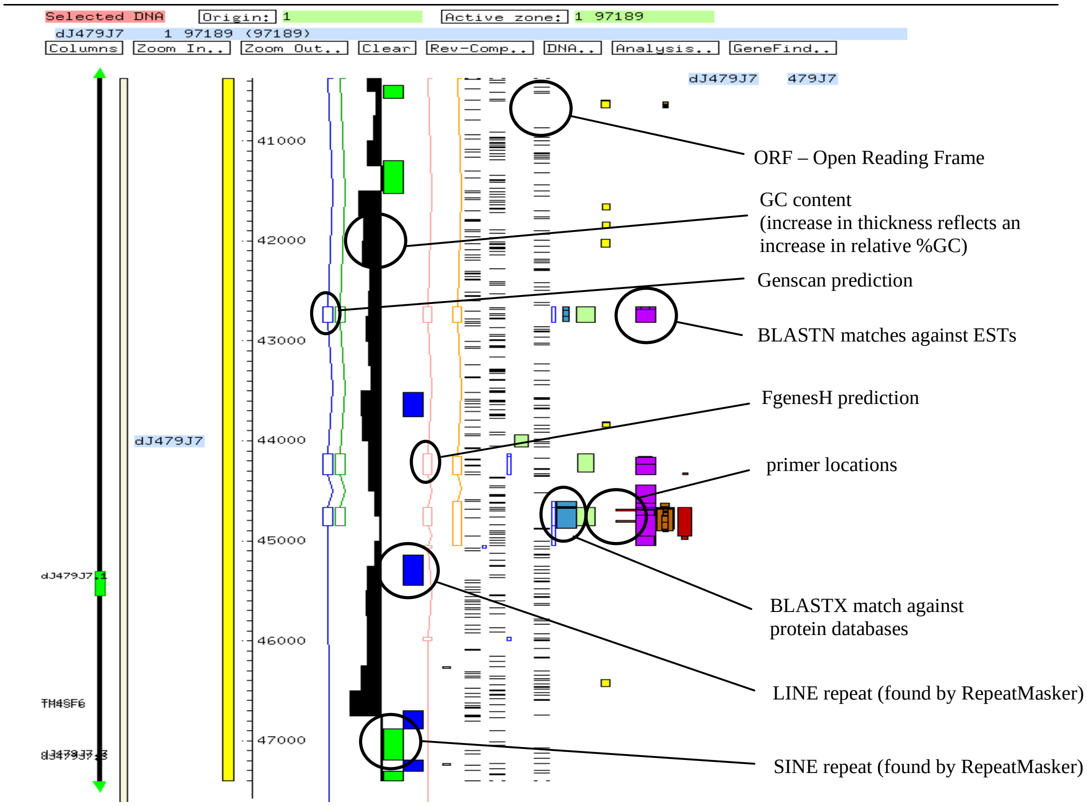
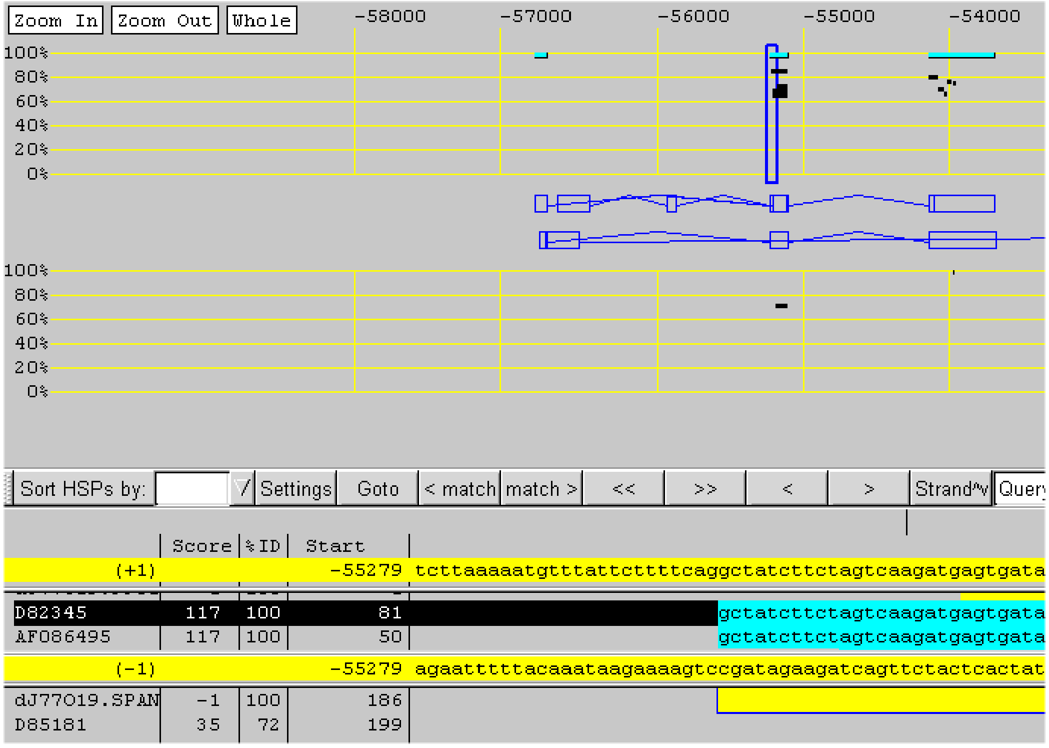


Figure 3-2 The image above illustrates how results of sequence analyses were collated and viewed within Xace. The yellow bar to the left of the image represents a section of the genomic clone's sequence.

When a gene was annotated through matches to EST sequences, SCCD sequence (see below) or mRNA/protein homologies, it was designated a "predicted gene" and assigned "GD_composite" in Xace. Most of these genes have evidence of expression, and the assignment of predicted gene reflects inherent limitations of accuracies of annotation when not annotating from a single contiguous mRNA sequence.

Individual loci for all three types of gene were assigned a locus identifier following the syntax – clone name.CX.number or clone name.number. In some instances where a well known HUGO identifier was available, the locus was named as such.



```
                  Score  %ID  Start
           (+1)               -55279  tcttaaaaatgtttattcttttcaggctatcttctagtcaagatgagtgata
D82345              117  100      81                              gctatcttctagtcaagatgagtgata
AF086495            117  100      50                              gctatcttctagtcaagatgagtgata
           (-1)               -55279  agaattttttacaaataagaaaagtccgatagaagatcagttctactcactat
dJ77019.SPAN         -1  100     186
D85181               35   72     199
```

Figure 3-3 Example of an mRNA match viewed using BLIXEM. The diagram illustrates BLASTN matches to mRNA accession D82345 to genomic clone AL035609. The vertical blue box represents the position of the region of alignment highlighted in the lower section in context with other matches to the highlighted mRNA (black) sequence. In this case, the intronic “ag” splice site can be seen preceeding the mRNA match in the lower section. The forward and reverse genomic sequences are highlighted in yellow.

In this manner, 74 genes, 51 predicted genes and 46 pseudogenes were annotated within the region. Some of these structures had been annotated previously by the Human Informatics group (Welcome Trust Sanger Institute) and in these cases where the gene structure did not need updating it was left as the representative annotation. A total of 26 loci could not be fully annotated or updated due to database limitations or their occurrence in recently-finished sequence – in these instances their locus type was determined from examination of the supporting evidence. An example of a gene, a predicted gene and a pseudogene are shown in Figures 3-4, 3-5 and 3-6 respectively.

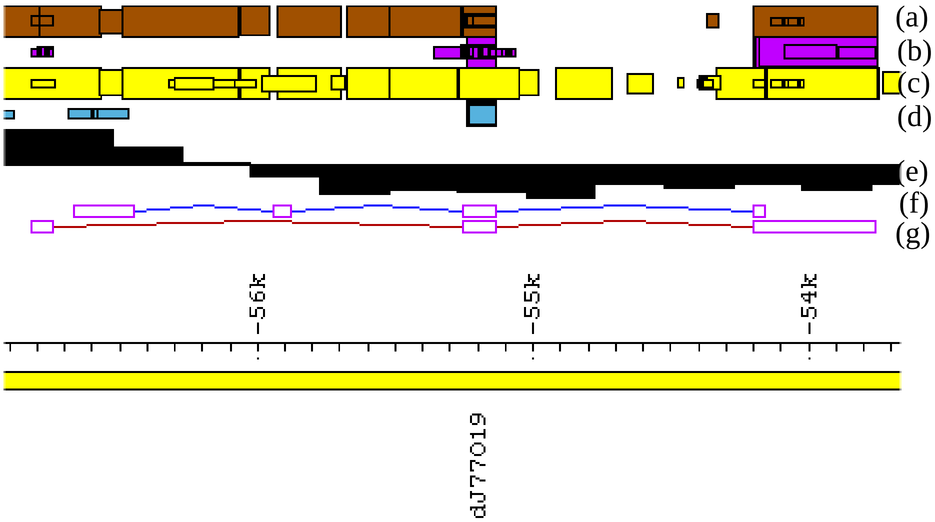


Figure 3-4 Example of a "gene" (GD_mRNA) structure, for locus dJ77O19.CX.1. In this case, the gene was annotated from mRNA accession D82345. The diagram shows an ACeDB representation of the gene structure. Key – (a) mRNA BLASTN matches, (b) EST BLASTN matches, (c) genomic sequence BLASTN matches (d) protein BLASTX matches, (e) GC content (increasing upward thickness of bars represents increased %GC relative to adjacent sequence, downwards a decrease), (f) FGENESH gene prediction, (g) annotated gene structure. The yellow bar represents the clone sequence with scale (in bp) noted. Exons are depicted as coloured boxes, with introns represented as coloured lines connecting the exons.

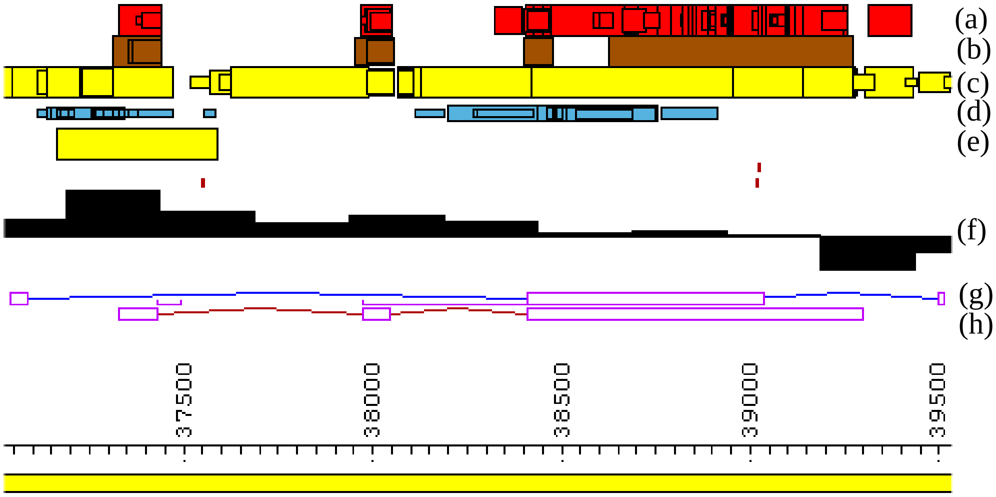


Figure 3-5 Example of a "predicted gene" (GD_composite) structure, for locus cV857G6.CX.2. This locus was annotated from overlapping EST sequences. The diagram shows an ACeDB representation of the gene structure. Key – (a) EST BLASTN matches, (b) mRNA BLASTN matches, (c) genomic sequence BLASTN matches (d) protein BLASTX matches, (e) CpG island, (f) GC content (increasing upward thickness of bars represents increased %GC relative to adjacent sequence, downwards a decrease), (g) GENSCAN and FGENESH gene predictions, (h) annotated gene structure. The yellow bar represents the clone sequence with scale (in bp) noted. Exons are depicted as coloured boxes, with introns represented as coloured lines connecting the exons.

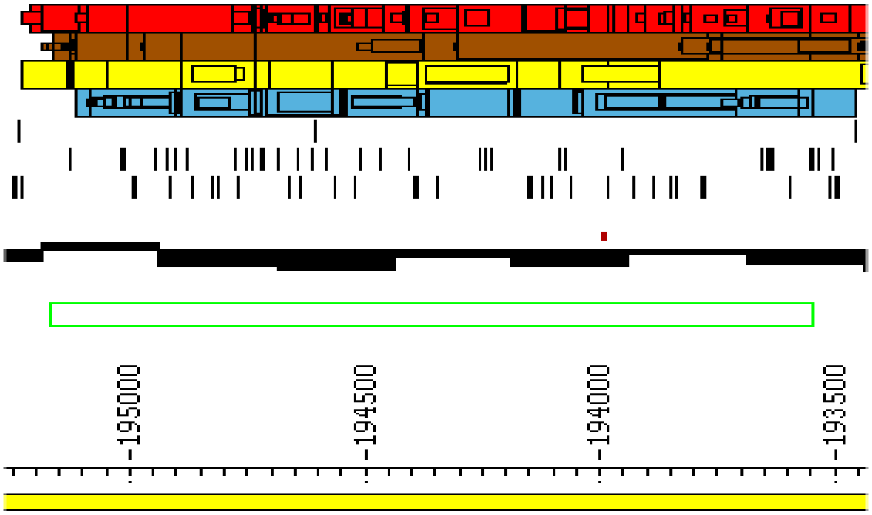


Figure 3-6 Diagram illustrating a “pseudogene” (pseudogene structure) structure, for the pseudogene locus dJ232L22.CX.2. The diagram shows an ACeDB representation of the gene structure. Key: mRNA/protein homologies as in Figure 3-5 above, vertical lines – boundaries of open reading frames (one row for each forward strand reading frame). In this case, an intronless BLASTX match to testis-specific glycerol kinase (accession Q14410) has an in-frame stop codon.

In addition to annotating gene structures on the basis of matching mRNA or splicing EST sequences, when features indicative of potential genes were found (as described above), and no (or partial) human mRNA sequence for the locus was available, an STS was designed within a putative exon. Primers were designed to the putative exonic sequence (multiple STSs were designed in instances where a large gene structure was expected) and were used to screen pools of clones from cDNA libraries by PCR.

Primer pairs designed to putative 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 $T_{0.1}E$. Pre-screens were performed using three different primer annealing temperatures (55°C, 60°C and 65°C) to determine the cycling parameters that give a visible and specific DNA product.

Screening was performed first on primary pools, and STSs positive in these pools were taken forward for screening of secondary pools of lower complexity. Nested primers were then designed, and SSPCR performed on up to three, positive secondary pools from different tissues where appropriate. The libraries screened and the protocols used are as described in Chapter 2. Sequence from resulting SSPCR products (termed sccd sequence) was then viewed in Xace following BLAST of the sequence against Xace clone sequences, and used to extend gene structures.

A total of 161 STSs were designed, to 127 putative genes. Of these, 13 failed pre-screening (as described in Chapter 2) and a further 16 were not taken further due to updated mRNA BLASTN information rendering them redundant. Of the 132 STSs screened against the primary cDNA pools, 6 experiments failed and 77 STSs gave positive results against one or more pools. In addition, for the NRK gene, due to the large predicted structure of this gene, four additional STSs and 5' RACE primers were designed to give product from direct PCR from an additional placental cDNA RACE library (kindly supplied by Jackie Bye). Sequence of products derived from PCR using these reagents were also used in annotation of the NRK gene. Results of pre-screening and pool-screens are given in table 3-1.

A total of 142 SSPCR products were sequenced (Wellcome Trust Sanger Institute, R&D group), and resulting sequence was entered into Xace to display matches to genomic sequence and used for gene annotation. An example of an experiment illustrating steps from pre-screening of primers to generation of PCR product by SSPCR is shown in Figure 3-7. An example of sccd sequence being used to annotate a gene structure is shown in Figure 3-8.

A striking feature of this study was the redundancy of the approach described caused by release of large amounts of mRNA sequence from large-scale cDNA sequencing projects (see Chapter 1). The majority of initially novel predicted genes gained mRNA coverage from these sources. As the genomic sequence analysis was "static", this redundancy did not become apparent before many of the STSs had been screened. Nevertheless, the directed approach demonstrated that, when combined with such large-scale mRNA data, comprehensive gene identification and annotation can be achieved as not all genes gained mRNA coverage from publicly available sources.

A complete list of genes, predicted genes and pseudogenes is given in table 3-2, with brief descriptions of their functions (derived from LocusLink, NCBI). In cases, where no information was available from LocusLink, information regarding similarity to known genes or domains found within the predicted protein (from analysis using InterPro at the EBI) is shown. A schematic representation of the genes within the Xq22-q23 region is given in Figure 3-9.

Gene	type	stSG no.	anneal temp (^{0}C)	primary screen	secondary screen	sccd
bA524D16A.2	predicted gene	84336	60	P, ALU	P, ALU	4954/4955
		84337	60	NONE		
NOX1	gene	84338	60	NONE		
NOX1	gene	84339	60			
dJ479J7.1	gene	84340	60	AH, T	AH, T	4956/4960
dJ479J7.1	gene	84341	60	AH, T	AH, T	
		87849	60	NONE		
cU209G1.CX.1	predicted gene	87850	60	P	P	4957/6450/6451
		87852	60	FAIL		
dJ164F3.CX.2	predicted gene	87853	60	H, P, HPB, ALU	H, P, ALU	4958/4961/4958
		87854	60	NONE		
		87855	65	FAIL		
		87856	60	NONE		
cU105G4.2	gene	88169	60	H, YT, HP, DX3, FB, FL, SK, T, FLU, AH		
cU116E7.CX.1	ps	88170	60	T		
cU144A10.CX.1	ps	88171	60	NONE		
		88172	FAIL			
cU177E8.CX.3	predicted gene	88173	FAIL			
cU177E8.CX.1	gene	88174	60	U, H, YT, DAU, HPB, DX3, FB, FL, HL, SK, T, FLU, ALU, AH		
		88175	60	SK		
cU19D8.CX.1	gene	88176	60	T	T	6455
bA370B6.1	predicted gene	88327	60	NONE		
dJ19N1.1	predicted gene	88328	55	U, NK, HPB, UACT, SK, FLU, AH, HSI, WEAK FB	NK 7. WEAK HPB 2. HSI 3	8592/8593
IL1RAPL2	gene	88329	55			
		88330	60	NONE		
NXF3	gene	88331	55	FAIL		
IL1RAPL2	gene	88332	60			

		88333	60	NONE		
		88334	60	NONE		
cU240C2.1	predicted gene	88336	60	NONE		
cU250H12.CX.1	predicted gene	88337	60	DX3, FB, FL, WEAK FLU, AH	DX3 2,4,5. FB 9. FL 20	4962/4963
dJ1055C14.3	predicted gene	88338	60	NONE		
IL1RAPL2	gene	88339	60			
cU42H12.CX.1/TEX13A)	gene	88340	60	NONE		
cU46H11.CX.1	predicted gene	88341	60	YT, FB, WEAK AH	YT 2. FB 22	4968/4969
cU46H11.CX.2	predicted gene	88342	60	FB	FB 9	4970/4971/6388/6389/8590
cU50F11.CX.1	predicted gene	88343	60	SK, T, AH	SK 2. T 2. AH 17	6382/6383/8576/8589
IL1RAPL2	gene	88344	60			
		88345	60	FAIL		
IL1RAPL2	gene	88346	60			
dJ3E10.CX.1	predicted gene	88347	60	FB, T, ALU, AH , HSI	FB 7,8. T 1. HSI 21,24	6384/6385/6386/6452/6453/8566/8596
cV351F8.CX.1	predicted gene	88348	60	WEAK P, FB, T, FLU, WEAK ALU, AH, HSI	T 3 (WEAK), 4.	4972/4973
cU46H11.CX.1	predicted gene	88349	60			
NXF2	gene	88350	60	T		
(genomic clone moved)		88351	55	AH, SK, T, U, NK, DAU, HPB, BM, UACT, FB, HL, SK	AH 7. SK 11. T 3	4974/4975/4976
cV857G6.CX.2	predicted gene	88352	FAIL			
cV857G6.CX.1	gene	88353	60	P, HPB, BM, DX3, FB, FL, SK, T, FLU, ALU, AH, HSI	P 11. HPB 1,2. FB 1,3,5.	6377/6378
(genomic clone redundant)		88354	60			
dA141H5.1(NEURALIN)	gene	88355	60	FB, ALU	FB 16. ALU 9	
dA141H5.1(NEURALIN)	gene	88356	60	H, FB, ALU, AH	H 13. FB 16. ALU 9	

dA141H5.1(NEURALIN)	gene	88357	FAIL			
dA149D17.CX.1	pseudo	88358	60	HSI	HSI 3,5.	6379/6387
		88359	FAIL			
dA191P20.1	pseudo	88360	60	FAIL		
dJ1055C14.3	predicted gene	88361	60	NONE		
dJ1100E15.2	pr/pseudo	88362	60	NONE		
dJ1100E15.CX.3	gene	88363	60	FB, AH,	FB 24, 25. AH 22, 23, 24.	6380/6381/8597/8598
dJ115K14.CX.1	predicted gene	88364	60	U, H, YT, HPB, FB	U 13. H 2, (WEAK 3). YT 1	6390/6403/6404
dJ122O23.CX.1	predicted gene	88365	60	NONE		
dJ197J16.CX.1	pseudo	88366	60	FAIL		
dJ198P4.CX.1	gene	88367	60	U, FB, HL, WEAK DX3, T, FLU, HSI		
		88368	60	NONE		
dJ364I1.1	predicted gene	88369	60	U, H, YT, NK, DAU, HPB, BM, UACT, DX3, HL, SK, T, ALU, AH		
		88370	60	NONE		
		88371	60	NONE		
dJ1070B1.1	predicted gene	88372	60	all, but larger band in HSI		
dJ1070B1.1	predicted gene	88373	60	DAU	DAU 1.	4966/4967/6391/6392
		88374	60	FAIL		
dJ3E10.CX.2	pseudo	88375	60			
dJ513M9.1	gene	88376	60	T	T 3.	4964/4965/4054
dJ519P24.CX.1	pseudo	88377	60	U,H, YT, NK, DAU, HPB, BM, UACT, DX3, FB, FL, HL, SK, T, FLU, ALU, AH, HSI		
dJ596C15.1	gene	88378	60	DAU, HPB, UACT, HL, SK, T		
dJ635G19.2	gene	88379	FAIL			
dJ298J18.CX.2	predicted gene	88380	60	YT, NK, HPB, DX3, FB, HL, FLU	YT 22. NK 6,9. HPB 5	6393/6394/8591
dJ298J18.CX.2	predicted gene	88381	60	WEAK U, YT, NK, HPB, DX3, HL, FL, SK, T, FLU, ALU, AH, HSI	YT 8. NK 14. AH 2,4.	6395/6396
dJ769N13.1	gene	88382	FAIL			
dJ769N13.CX.1	predicted	88383	60	U, P, NK, HPB, FB, ALU, AH	U 5. P 8. NK 21, 23.	

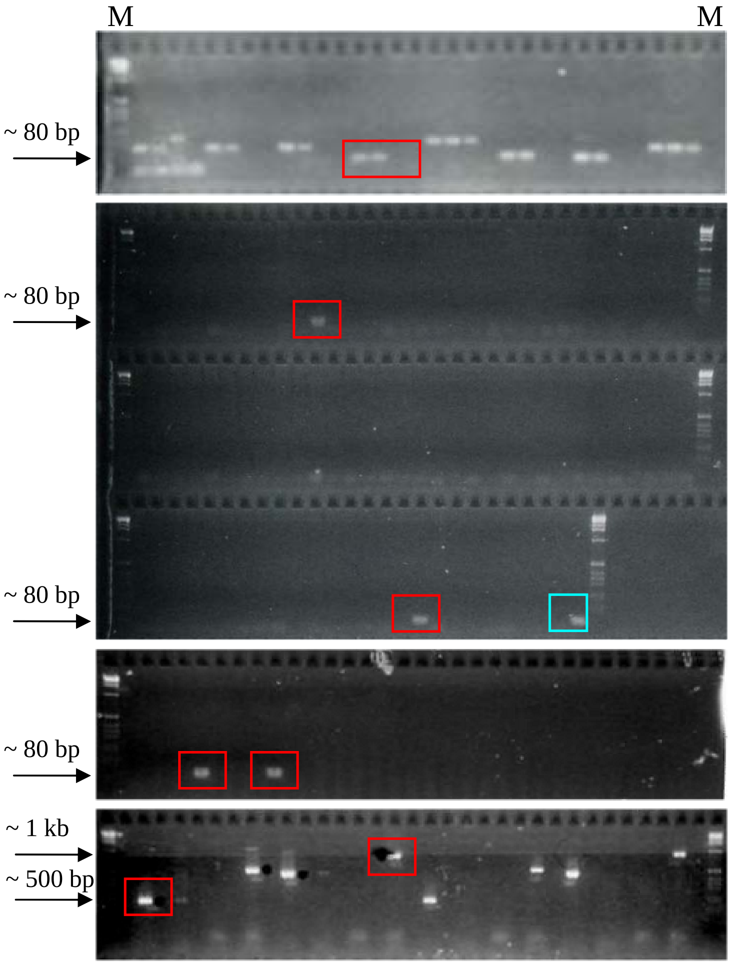
	gene					
		88384	60	NONE		
dJ769N13.CX.2	gene	88385	FAIL			
dJ839M11.1	predicted gene	88386	60	DAU, HPB, FB, SK, FLU	DAU 1. HPB 24	
dJ839M11.2	predicted gene	88387	FAIL			
dJ889N15.1	predicted gene	88388	60	T	T 9.	6397/6398
dJ889N15.1	predicted gene	88389	60	T	T 9.	6399/6400/6401/6402
dJ889N15.CX.1	predicted gene	88390	60	H, UACT, DX3, HL, SK, ALU, AH, HSI	H 14. UACT 2. DX3 5	6405/6406/6407/8585
dJ889N15.CX.1	predicted gene	88391	60	H, P, YT, DAU, UACT, DX3, HL, SK, FLU, ALU, AH, HSI, WEAK FB	H 9. P 8. YT 6	6408/6409/8586
dJ914P14.1	gene	88392	60	NONE		
		95448	60	NONE		
cU131B10.CX.1	predicted gene	95449	60	NONE		
		95450	60	NONE		
		95451	FAIL			
		95452	60	NONE		
dJ233G16.CX.1	predicted gene	95453	60	H, DAU, UACT, DX3, AH	H 10. DX3 13. AH 7.	6412/6413/8587
dJ233G16.1	predicted gene	95454	60	Very weak T	T 5.	6414/6415
		95455	60	NONE		
		95456	60	Very weak AH	NONE	
		95457	60	Very weak HPB, Very weak FL	NONE	
		95458	60	NONE		
dJ302C5.CX.1	gene	95459	60	T, AH	T 8,10. AH 23.	6419/6420
IL1RAPL2	gene	95460	60	NONE		
		95461	60	NONE		
dJ44L15.CX.1	gene	95462	60	BM, SK, T		
dJ44L15.CX.1	gene	95463	60	BM		
dJ545K15.CX.1	predicted gene	95464	60	AH, ALU, FLU, HPB, FB	AH 2. ALU 2. FB 6,9.	6410/6411
dJ664K17.CX.1	gene	95465	60	HSI, FB, FLU	HSI 1, 5. FB 1,2,4,5. FLU 2,3,4.	6416/6417/6418

dJ738A13.1	predicted gene	95466	60	NONE		
dJ820B18.1	predicted gene	95467	60	NONE		
		95468	60	NONE		
cU240C2.2	predicted gene	88335	60	DAU, HPB, FB, SK, FLU	DAU 1,5. WEAK HPB 2. FB 25	
cU46H11.CX.1	predicted gene	88349	60	YT, FB	YT 2	
cU250H12.CX.1	predicted gene	101533	65	FB	FB 11	
cU237H1.1	predicted gene	101443	55	FL	FL 20	
		118977	65	T	Very faint T 7	8582/8583
		118978	65	NONE		
		118979	60	NONE		
		118968	FAIL			
		118969	65	T	T 6.	6447/6448
		118970	60	NONE		
		118971	65	AH	AH 10	8579
		118972	65	FB, AH	FB 2. AH 5	8573/8595
KCNE1L	predicted gene	118973	65	NONE		
dJ164F3.CX.2	predicted gene	118974	60	NONE		
dJ269O5.CX.2	gene	118975	65	T	T 3,4.	8567/8568
		118976	65	NONE		
		118961	60	NONE		
		118962	65	NONE		
		118963	65	NONE		
		118964	65	NONE		
cU250H12.CX.1	predicted gene	118965	65	NONE		
		118966	65	FB, FL, T, FLU, AH, HSI	FB 22. T 17. HSI 2,5.	8571/8572
		118967	65	NONE		
dA170F5.CX.1	predicted gene	46776	60	T	T 11	8580
dJ122O23.CX.1	predicted gene	119002	65	Weak NK, T, HSI	NONE	

cV351F8.CX.2	predicted gene	119015	60	FB, T, FLU, ALU, HSI	FB 2,3. T 8. FLU 4.	8562/8563/8588
dJ769N13.CX.1	predicted gene	119016	65	FB, SK, T, FLU, ALU, AH, HSI	FB 7,8. SK 5. AH 1,3 (WEAK 5)	8564/8565
dJ341D10.2	predicted gene	119008	65	P, T	P 14. T 3,4.	8560/8561
dJ341D10.3	gene	119009	65	U, NK, HPB, BM, UACT, SK, FLU, ALU, AH	U 1,3,4,5. NK 6,8. BM 17.	
dJ514P16.CX.1	gene	119010	65	P	P 5	6449
dJ545K15.CX.1	predicted gene	119011	60	NONE		
cU19D8.CX.1	pr/gene	119013	65	NONE		
cV351F8.CX.1	predicted gene	119014	60	Weak HSI	HSI 6 (WEAK), 7.	6454/8581
		118980	n/a	NONE		
cU46H11.CX.1	predicted gene	119018	65	YT, DAU	YT 6. DAU 6.	8574/8575
		119022	65	H	H 6	4055/4056
FLJ22679	gene	119023	65	U, DAU, BM, UACT, AH	DAU 9,10. BM 17. AH 6,7.	8584/8594
dJ1070B1.1	predicted gene	119024	60	NONE		
dJ1070B1.1	predicted gene	119025	60	YT, DAU (DOUBLET), FLU	YT 6,9. FLU 11. DAU 17 (DOUBLET).	8577/8578
bA524D16A.2	predicted gene	119027	60	HD, T, ALU, AH	H 17. T 3. ALU 1.	8558/8559
dJ820B18.1	predicted gene	119075	65	NONE		
dJ769N13.CX.2	gene	119017	65	FB, FL, SK, FLU, ALU	FB 3. FLU 5. ALU 14.	8569/8570
dJ889N15.CX.1	predicted gene	119098	65	NONE		
dJ1055C14.CX.1	predicted gene	119121	60	HPB, T	HPB 2	6456/4053
MYCL2	predicted gene	118040	65	NONE		
		119076		NONE		
dJ3E10.CX.1	predicted gene	119120		T	T 1.	
cU84B10.CX.1	predicted gene	84327	60	H, AH , Weak T	H 2. AH 6.	4838/4839/4840/4841
cU84B10.CX.1	predicted gene	84328	60	H	H 14	4842/4843

cU84B10.CX.1	predicted gene	84329	60	H	H 3	4844/4845
cU84B10.CX.1	predicted gene	88115	RACE	not applicable - PCR directly on placenta 5' RACE-ready library		4947
cU84B10.CX.1	predicted gene	88151	88152	not applicable - PCR directly on placenta 5' RACE-ready library		4948
cU84B10.CX.1	predicted gene	88153	88154	not applicable - PCR directly on placenta 5' RACE-ready library		4949
cU84B10.CX.1	predicted gene	88155	88156	not applicable - PCR directly on placenta 5' RACE-ready library		4950
cU84B10.CX.1	predicted gene	88162	88163	not applicable - PCR directly on placenta 5' RACE-ready library		4951/4952
cU84B10.CX.1	predicted gene	99719	60	H, T, AH	H 11,14. T 2. AH 25.	6421/6422
cU84B10.CX.1	predicted gene	99720	60	H, T,FLU, AH	FLU 15. H 5. AH 1.	6423/6424/6425
clone no longer in path	n/a	95469	FAIL			

Table 3-1 Results of STS pre-screens and pool-screens. Locus names and gene type are given where features that STSs were designed to became annotated. The stSG numbers and optimal pre-screen annealing temperatures are given. FAIL denotes an unclear pre-screen result. Positive cDNA libraries are denoted by their letter codes (see Chapter 2). Sccd numbers assigned to SSPCR products sent for sequencing are given where appropriate.



Prescreen:
Red box shows genomic DNA, clone2D, hamster DNA and $T_{0.1}E$ lanes for stSG84336

Primary poolscreen;
Red boxes show positive pool A in library “P” (top) and positive pool C in library “ALU” (bottom).
Blue box denotes $T_{0.1}E$ and genomic DNA lanes

Secondary prescreen:
Red boxes denote positive sub-pools P A5 and ALU C4

SSPCR:
Red boxes denote products generated with primer combinations 87976S/2FP and 87976A/1RP, designated sccd4954 and sccd4955 respectively

Figure 3-7 Prescreening, poolscreening and SSPCR for STS stSG84336. Prescreening results at annealing temperature of 55^{0}C are shown. M denotes 1kb ladder.

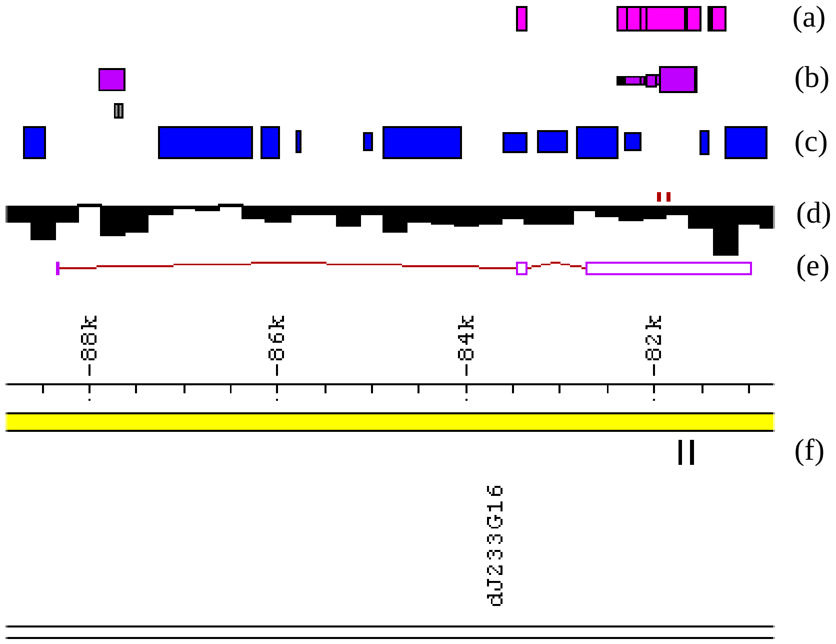


Figure 3-8 Diagram illustrating annotation of a gene structure using SCCD sequence. The diagram shows an ACeDB representation of locus dJ233G16.CX.1. Key – (a) SCCD sequence BLASTN matches, (b) EST BLASTN matches, (c) LINE repeats, (d) GC content (increasing upward thickness of bars represents increased %GC relative to adjacent sequence, downwards a decrease), (e) annotated gene structure, (f) vertical lines depicting positions of the primers used for the initial primary poolscreen. The yellow bar represents the clone sequence with scale (in bp) noted. Exons are depicted as coloured boxes, with introns represented as coloured lines connecting the exons. In this case, the short length of the 5' exon meant BLASTN failed to locate a match to the SCCD sequence (also occurring if masked by repeats), but the splicing of the SCCD sequence could be verified on manual inspection of the sequence.

Gene (locus name)	HUGO name	Other name(s)	Similarity	Locus type	Function/predicted function
dJ377O6.1			Ku70	pseudo	n/a
bA368G3.CX.1			NAGtransferase	pseudo	n/a
bA402K9.CX.1			FLJ10523	pseudo	n/a
bA402K9.CX.2			LAMR1	pseudo	n/a
bA99E24.CX.1	PCDH19	KIAA1313	protocadherin	gene	cell-cell adhesion
dJ479J7.1		myodulin/TNMD	chondromodulin	gene	cell surface glycoprotein, possible regulatory role
TM4SF6,	TM4SF6	T245	tetraspanin	gene	cell surface glycoprotein, signal transduction
dJ479J7.3		SRPUL	sushi-repeat	gene	Contains SUSHI repeat. These have been noted in several complement system proteins
bA524D16A.2	SYTL4	Granuphilin A	synaptotagmin	predicted gene	protein transport/exocytosis
dJ347M6.1			cyclophilin A	pseudo	n/a
dJ347M6.2			galactosyltransferase	pseudo	n/a
CSTF2	CSTF2		cleavage stimulation factor	gene	part of CSTF, involved in polyadenylation and 3'-end cleavage of pre-mRNAs
NOX1	NOX1	MOX1	NADPH oxidase subunit	gene	Voltage-gated proton channel
dJ146H21.3,			hnRNP A1	pseudo	n/a
dJ146H21.CX.1			hnRNP A1	pseudo	n/a
cU131B10.CX.1			XK (KX antigen)	predicted gene	potential membrane transport protein
dJ341D10.1			PR00082/GALNACT-2	predicted gene/ possible pseudo	GALNACT-2 has a role in synthesis of chondroitin sulphate
dJ341D10.2			ADP-ribosylation factor (GTP-binding)	predicted gene	ARF GTP-binding proteins involved in vesicular transport processes
dJ341D10.3		FLJ12687	HTF9c	gene	Contains SAM-dependent methyl-transferase domain
dJ664K17.CX.1		FLJ14084		gene	
FSHPRH1	FSHPRH1	LRPR1		gene	Possibly involved in response to FSH
cV210E9.CX.1			14.3.3 protein	pseudo	n/a
cV210E9.CX.2			14.3.3 protein	pseudo	n/a
DRP2	DRP2		-	gene	Possible role in maintenance of membrane-associated complexes
dJ738A13.1	TAF7L	TAF2Q/FLJ23157	TAFII55	predicted gene	Possible TATA box binding protein associated factor (similar to mouse testis-specific gene)
dJ164F3.CX.1			RPL21	pseudo	n/a
TIMM8A	TIMM8A	DFN1/DDP		gene	inner mitochondrial membrane translocase
BTK	BTK	ATK		gene	protein tyrosine kinase. Defects cause Agammaglobulinaemia

RPL44	RPL36A	RPL44		gene	Ribosomal protein, component of 60S subunit
GLA	GLA		alpha-galactosidase	gene	Defects cause Fabry disease
HNRPH2		HNRPH2		predicted gene	HnRNPs are RNA binding proteins. Associate with pre-mRNA in nucleus.
dJ164F3.CX.2				predicted gene	
cU209G1.CX.1			ALEX	predicted gene	Weakly similar to GASP. Possible role in receptor sorting
cU209G1.CX.2				predicted gene	
dJ514P16.CX.1				gene	
dJ514P16.1			PRKCI	pseudo	n/a
cU61B11.CX.1		ALEX1	ALEX	gene	Weakly similar to GASP. Possible role in receptor sorting
dJ545K15.CX.1		FLJ20811	ALEX	predicted gene	Weakly similar to GASP. Possible role in receptor sorting
dJ545K15.1		FLJ20811	ALEX	predicted gene	Weakly similar to GASP. Possible role in receptor sorting
dJ545K15.2		ALEX3	ALEX	gene	Weakly similar to GASP. Possible role in receptor sorting
cV602D8.CX.1		ALEX2/KIAA0512	ALEX	gene	Weakly similar to GASP. Possible role in receptor sorting
dJ232L22.CX.1			Zn-finger	pseudo	n/a
dJ232L22.CX.2			GK	pseudo	n/a
NXF5	NXF5		NXF	gene	Possible role in mRNA export from nucleus
dJ3E10.CX.1			-	predicted gene	Contains Zn-finger domain, a domain found in nucleic acid-binding proteins
dJ3E10.CX.2			ZNF135	pseudo	n/a
dJ197J16.CX.1			ND6 (mitochondrial)	pseudo	n/a
dJ122O23.CX.1			Myo48/pp21	predicted gene	Similar to pp21/TCEAL1. Possible transcriptional regulator
cV351F8.CX.1			Myo48/pp21	predicted gene	Similar to pp21/TCEAL1. Possible transcriptional regulator
cV351F8.CX.2			NADE	predicted gene	Similar to NADE. Possibly involved in signal transduction/apoptosis
dJ158I15.1			TCP11	pseudo	n/a
cU19D8.CX.1			TCP11	predicted gene/gene	TCP11 is the receptor for FPP, may have role in sperm function and fertility
NXF2	NXF2		NXF	gene	Possible role in mRNA export from nucleus
bA353J17.1			NXF	gene	Possible role in mRNA export from nucleus
bA353J17.2			TCP11	gene	TCP11 is the receptor for FPP, may have role in sperm function and fertility

dJ77O19.CX.1		NB thymosin beta/TMSNB	thymosin-beta	gene	Beta-thymosins involved in regulation of actin polymerisation
dJ1100E15.1			checkpoint suppressor 1	pseudo	n/a
dJ1100E15.2	NXF4		NXF (includes partial duplication)	predicted gene/pseudo	Possible role in mRNA export from nucleus
dJ1100E15.CX.3		FLJ12969/FLJ13382	GASP	gene	Similar to GASP, possible role in receptor sorting
dJ1100E15.CX.4			Histone H3	pseudo	n/a
dJ769N13.1		GASP/KIAA0443	GASP	gene	GPCR-associated sorting protein
dJ769N13.CX.1			GASP	predicted gene	Similar to GASP, possible role in receptor sorting
dJ769N13.CX.2		KIAA1701	GASP	gene	Similar to GASP, possible role in receptor sorting
dJ769N13.CX.3				predicted gene	
cU157D4.CX.1				predicted gene	
cU237H1.1			Rab	predicted gene	part of ras family of GTP-ases. Possible role in vesicular trafficking
cU73E8.CX.1			mouse RP2	pseudo	n/a
dJ198P4.CX.1			NADE	gene	Similar to NADE. Possibly involved in signal transduction/apoptosis
NXF3	NXF3		NXF	gene	Possible role in mRNA export from nucleus
cU221F2.CX.1			ZN-finger proteins	pseudo	n/a
dJ635G19.1			LAMR1	pseudo	n/a
dJ635G19.2		FLJ10097	NADE	gene	Similar to NADE. Possibly involved in signal transduction/apoptosis
cU177E8.CX.1		FLJ22696	pp21	gene	Similar to pp21/TCEAL1. Possible transcriptional regulator
cU177E8.CX.2			GMP reductase	pseudo	n/a
cU177E8.CX.3			pp21	predicted gene	Similar to pp21/TCEAL1. Possible transcriptional regulator
dJ79P11.1			NADE	gene	Similar to NADE. Possibly involved in signal transduction/apoptosis
cU105G4.1			pp21	gene	Similar to pp21/TCEAL1. Possible transcriptional regulator
cU105G4.2			pp21	gene	Similar to pp21/TCEAL1. Possible transcriptional regulator
NGFRAP1	NGFRAP1	NADE/BEX3/HGR74/DXS6984E	NADE	gene	p75NTR-associated cell-death executor

cU250H12.CX.1			rab	predicted gene	part of ras family of GTP-ases. Possible role in vesicular trafficking
cU246D9.1			histone	pseudo	n/a
cV857G6.CX.1		FLJ21174		gene	Similar to pp21/TCEAL1. Possible transcriptional regulator
cV857G6.CX.2				predicted gene	Similar to pp21/TCEAL1. Possible transcriptional regulator
TCEAL1	TCEAL1	pp21	pp21	gene	Potential transcription modulator
dJ1055C14.2	MORF4L2	MRGX/KIAA0026		gene	Contains MRG domain. Possible role in regulation of transcription, cell proliferation
dJ1055C14.CX.1				predicted gene	
dJ1055C14.3			GLRA4	predicted gene	Potential glycine receptor - could be pseudogene
PLP	PLP1	PLP/PMD	lipophilins	gene	Membrane protein, constituent of myelin. Implicated in Pelizaeus-Merzbacher disease
dJ540A13A.CX.1	RAB9B	RAB9L	rab	gene	part of ras family of GTP-ases. Possible role in vesicular trafficking
bA370B6.1			histone H2B	predicted gene	Histone H2B
cU116E7.CX.1			NERF-1	pseudo	n/a
cU116E7.CX.2		FLJ22859		gene	
cU116E7.CX.3				predicted gene	Similar to mitochondrial carrier protein (but probable frameshift)
cV362H12.CX.1			beta-thymosin	predicted gene	Beta-thymosins involved in regulation of actin polymerisation
dJ839M11.1			histone H2B	predicted gene	Histone H2B
dJ839M11.2			histone H2B	predicted gene	Histone H2B
cU240C2.1			histone H2B	predicted gene	Histone H2B
cU240C2.2			histone H2B	predicted gene	Histone H2B
cU46H11.CX.1			-	predicted gene	Similar to mitochondrial carrier proteins
cU46H11.CX.2			-	predicted gene	Similar to paraneoplastic cancer-testis-brain antigen MA3 (PNMA3)
dJ233G16.CX.1				predicted gene	In LINE. ?
dJ233G16.1			eg of 2 genes become 1 thro sccd	predicted gene	
dJ513M9.1			mouse Esx1	gene	Homeodomain family, probable transcription factor
IL1RAPL2	IL1RAPL2	IL1R9/TIGIRR-1		gene	Possibly involved in receptor signal transduction
dJ519P24.CX.1			prohibitin	pseudo	n/a

cU144A10.CX.1			RPL18a	pseudo	n/a
stcU42H12.2	TEX13A		Tex	gene	Contains Zn-coordinating RNA binding domain. Possible role in spermatogenesis
cU84B10.CX.1		NRK	nik-related kinase	predicted gene	Similarity to GCK family Ser/Thr protein kinases. Mus NESK ativates JNK pathway
TBG	SERPINA7	TBG		gene	Serine (or cysteine) proteinase inhibitor
bA560L11.1			HSPC129	pseudo	n/a
cU50F11.CX.1		FLJ31916	-	predicted gene	Contains PWWP domain, found in nuclear proteins. Similar to Mus UBE-1C2
dJ19N1.1			PR00082/GALNACT-2	predicted gene/possible pseudo	GALNACT-2 has a role in synthesis of chondroitin sulphate
bA565G2.1			NAP1L4	pseudo	n/a
bB483F6.1			serpin/TBG	pseudo	n/a
FLJ10178/14191		FLJ10178/FLJ14191		gene	Predicted to contain nucleic-acid binding OB-fold.
FLJ23516	RNF128	FLJ23516/GRAIL		gene	Transmembrane protein with RING Zn-finger motif. Can function as E3 ubiquitin ligase
FLJ20298		FLJ20298		gene	Predicted to contain GRAM domain, found in glucosyltransferases, myotubularins and other putative membrane-associated proteins. Also predicted to contain TBC domain (found in GTPase activator proteins of Rab-like GTPases) and Ca2+ binding EF-hand
bA321G1.2			IMAGE:3609599	gene	
CLDN2	CLDN2	CLAUDIN 2		gene	Tight-junction protein
dJ75H8.2	ZCWCC2	FLJ31673/FLJ11565	KIAA0136 (Mm MORC-nuc spermatogenesis)	gene	Zn-finger, CW-type with coiled coil domain 2.
dJ75H8.3			EEF1A2	pseudo	n/a
bB383K5.1		FLJ11016/FLJ13670		gene	Contains an RNA recognition motif
bB383K5.2		FLJ20130		gene	Similar to a region of a nuclear pore glycoprotein
dJ1126E12.1				predicted gene	
MYCL2	MYCL2			predicted gene	Processed gene related to L-MYC
dJ320J15.CX.1			DNAJ	pseudo	n/a
dJ3D11.1			cytokeratin 18	pseudo	n/a
dJ1070B1.1			KIAA0316/KIAA0967	predicted gene	Similar to KIAA0316, which contains a PDZ domain and is a Band 4.1 homologue

PRPS1	PRPS1		PRPS	gene	Phosphoribosyl pyrophosphate synthetase, involved in PRPS-related gout
DSIPI	DSIPI	GILZ		gene	Similar to leu-zipper proteins that function as transcriptional regulators
dJ820B18.1			-	predicted gene	Similar to CBP20, nuclear CAP-binding protein (but intronless, possible pseudogene)
MID2	MID2	FXY2	Midline	gene	Member of TRIM family, localises to microtubules in cytoplasm
dA191P20.1			AGTRII	pseudo	n/a
dA191P20.CX.1	TEX13B		TEX	gene	Possible role in spermatogenesis
dJ889N15.1			CTX	predicted gene	Similar to Xenopus cortical thymocyte receptor. Member of Ig superfamily.
dJ889N15.2	PSMD10	26Sp28	PSMD10	gene	Part of 26S proteasome
dJ889N15.CX.1	AUTL2			predicted gene	Member of autophagin family, involved in autophagy. A cysteine protease
COL4A6	COL4A6		Collagen	gene	Subunit of type IV collagen, major component of basement membrane
COL4A5	COL4A5		Collagen	predicted gene	Subunit of type IV collagen, major component of basement membrane. Involved in Alport syndrome
dA149D17.CX.1			PLRP2	pseudo	n/a
dA24A23.2	IRS4			gene	Cytoplasmic protein with multiple phosphorylation site. Signal transduction
dJ31B8.CX.1			GAPDH	pseudo	n/a
GUCY2F	GUCY2F			gene	Guanylate cyclase
dJ596C15.1	NXT2	P15-2		gene	Binds NXF genes, possible role in mRNA export
dJ136J15.3			FZO (putative GTPase)	pseudo	n/a
KCNE1L	KCNE1L	AMMECR2		predicted gene	Similar to a voltage-gated potassium channel
FACL4	FACL4		long chain fatty acid coenzyme A ligase	gene	Involved in fatty acid degradation and lipid synthesis. Mutated in nonspecific X-linked mental retardation
dJ205E24.CX.1			ribosomal protein S5	pseudo	n/a
FLJ22679		FLJ22679		gene	
bB360B22.2			intronless, in AMMECR1 3' utr	gene	
AMMECR1	AMMECR1			gene	Predicted to contain AMMECR1/DUF51 domain
dJ364I1.1			GNG5 - pseudo??	predicted gene	Similar to GNG5, a G-protein. Intronless compared to GNG5, but uninterrupted ORF.

dJ302C5.CX.1		KIAA1318		gene	
TDGF3	TDGF3	CRIPTO-3		gene	Probable retrotransposed gene
bA441A11.CX.1			mannose-6-phosphate receptor	pseudo	n/a
dA141H5.1		Neuralin 1/NRLN1/Ventroptin		gene	Potential secretory protein involved in development
PAK3	PAK3		PAK	gene	Ser/Thr kinase. Mutated in nonsyndromic X-linked mental retardation
dJ914P14.CX.1			GLUD-1	pseudo	n/a
dJ914P14.1	CAPN6	CANPX	Calpain-like protease	gene	Similar to calpain cysteine proteases
DCX	DCX	LISX		gene	Involved in microtubule organisation. Mutations cause X-linked lissencephaly
dA170F5.1			HMG1	pseudo	n/a
dA170F5.CX.1				predicted gene	
bA111F16.1			EIF-4B	pseudo	n/a
dJ298J18.1			RPL18A	pseudo	n/a
dJ298J18.CX.2		FLJ23018		predicted gene	
dJ269O5.CX.1			Fau	pseudo	n/a
dJ269O5.CX.2		Image:4822062		gene	
TRPC5	TRPC5			gene	Cation channel
bB266I11.1			FLJ13646	pseudo	n/a
dJ115K14.CX.1			HMG	predicted gene	HMG (High mobility Group) proteins involved in nucleoprotein structure assembly
dJ44L15.CX.1	AMOT	Angiomotin/KIAA1071		gene	Possible role in cell motility
dJ1170D6.1			USA-cyclophilin	pseudo	n/a

Table 3-2 Genes, predicted genes and pseudogenes annotated within Xq22-q23. Pseudogene function/predicted functions are not given, and are denoted n/a.

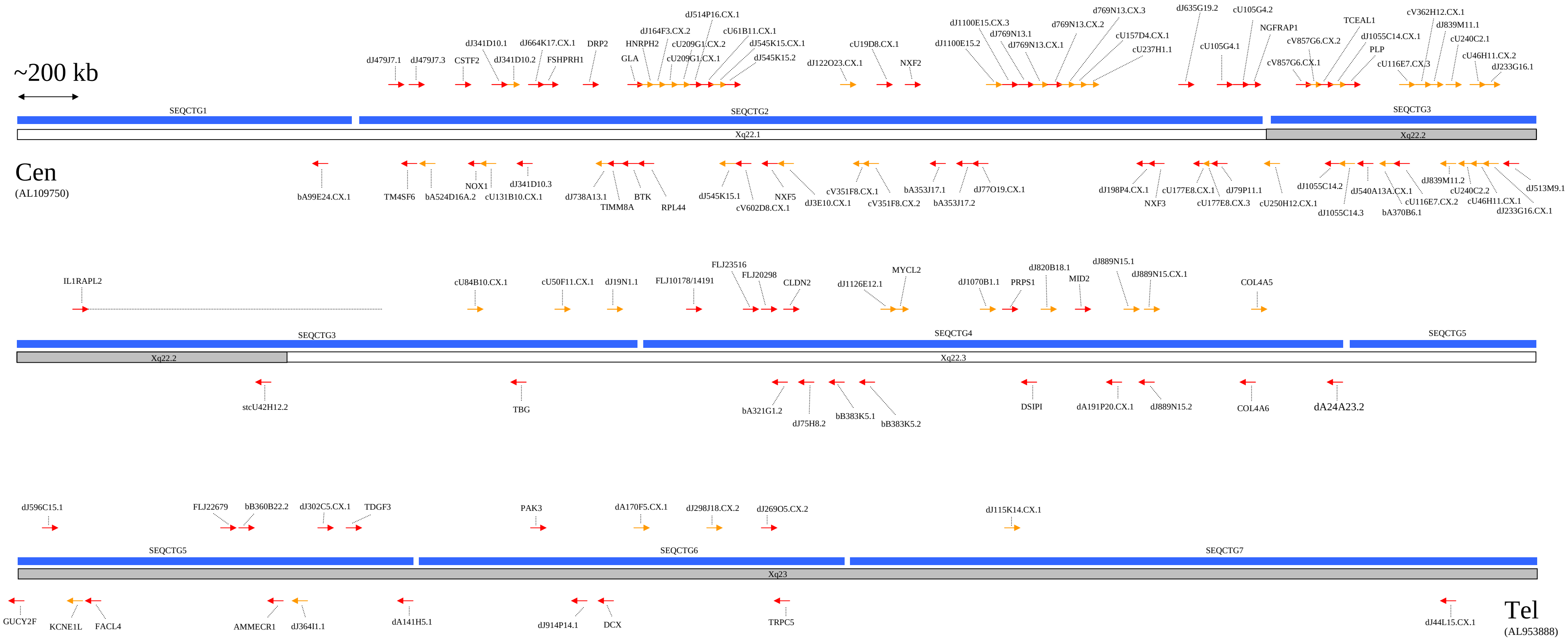


Figure 3-9 Genes annotated on finished sequence of the human Xq22-q23 region from clone dJ902O5 (AL109750) to dJ137P21 (AL953888), annotated as described in this Chapter. The region beginning is at top left, continuing onto the lower sections of the diagram. "Cen" denotes the centromeric end, "Tel" the telomeric end. Arrows represent annotated genes, direction indicating transcription direction. Red arrows represent "gene" loci, orange arrows "predicted gene" loci. Pseudogenes are omitted for clarity. Sequence contigs are represented by blue bars. The order of the clones (and their accession numbers) within the sequence contigs is given in Appendix A.1. A dotted grey line extends from the IL1RAPL2 gene to illustrate the length of this very large gene. Approximate boundaries of cytogenetic bands are indicated beneath the blue bars (from Ensembl human v19.34a.1).

3.3 Selected features of the region

3.3.1 Discovery of extensive paralogy within human Xq22 and between Xp and Xq22-q23

In the process of annotating genes on the sequence of human Xq22-q23, 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 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 extent of paralogy within the corresponding region of the mouse genome is explored in Chapter 4 and a full description of these genes is given in Chapter 5.

During the annotation of genes in Xq22-q23, it was also noted that several genes had similarly named counterparts mapping to Xp, such as MID1 (Xp22.3) and MID2 (Xq22). In addition, during BLAST analyses using certain Xq22 genes as queries, genomic sequences from Xp were registered as hits. Perry *et al.* (Perry *et al.,* 1999) also noted paralogy between Xp and Xq, and suggested an intra-chromosomal duplication involving the Xq22 region. As the Xq22 transcript map developed, a systematic search was made for genes mapping to Xp with paralogues within Xq22-q23. This search involved both literature review and BLAST analyses utilising Xq22 genes from the transcript map against genomic and mRNA/protein sequences.

In this way, a total of 15 pairs of paralogues shared between Xp and Xq were discovered. These include 11 novel observations of Xp/Xq22 paralogue pairs. The remaining four gene pairs were noted by Perry *et al.;* at present PHKA1 and PHKA2 are not included in this description of the putative segmental duplication due to their relative distances from other Xp/Xq22 paralogues, although their involvement in the event cannot be discounted. For a diagram illustrating the Xp/Xq paralogue pairs, see Figure 3-10. These Xp/Xq paralogues will be described in further detail in Chapter 6.

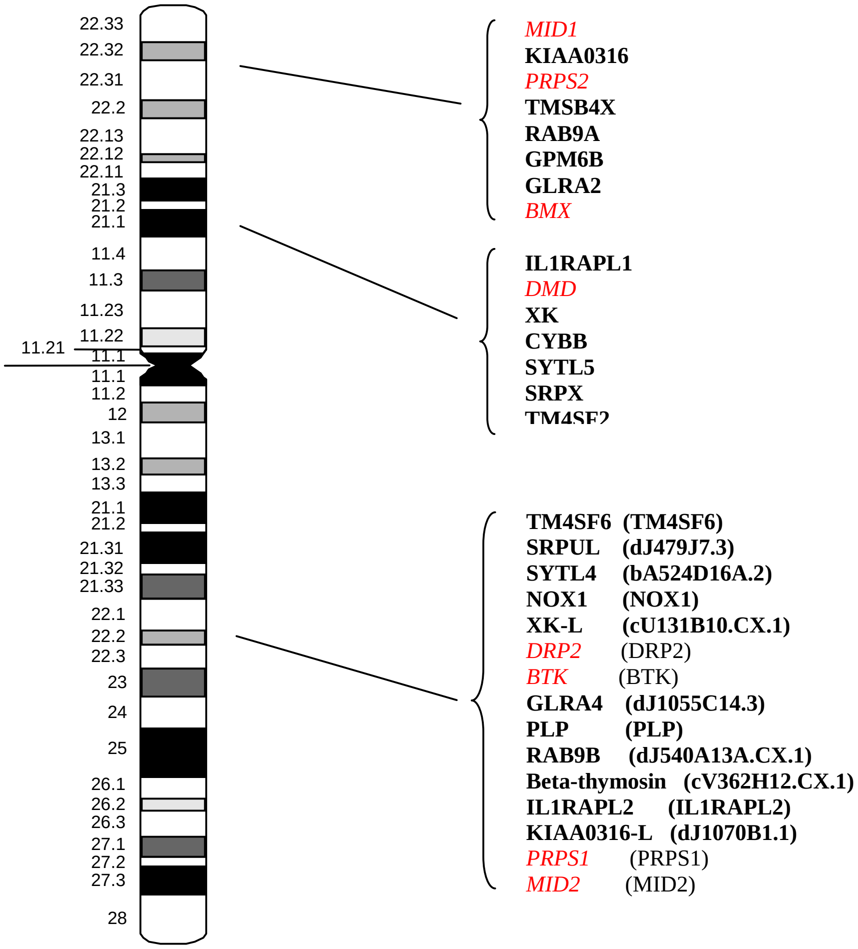


Figure 3-10 Observations of Xp/Xq paralogues. Xp/q paralogues noted by Perry *et al.***,** (1999) are in red italic type, new observations are in bold type. Locus names assigned during annotation of Xq22 are given in parentheses.

3.3.2 NXF2 inverted repeat and gene fusion

During annotation of Xq22, the NXF2 gene was found to reside in an inverted repeat of approximately 140 kb with extremely high sequence conservation (see Figure 3-11). The NXF2 family of genes have been the subject of intensive study in the last few years since the discovery of their role in mRNA export (Herold *et al.*, 2000). That there are two copies of the NXF2 gene would have escaped notice previously, as there is only a single nucleotide difference between their predicted mRNAs, encoding a silent mutation within an alanine codon towards the C-terminus of the predicted protein.

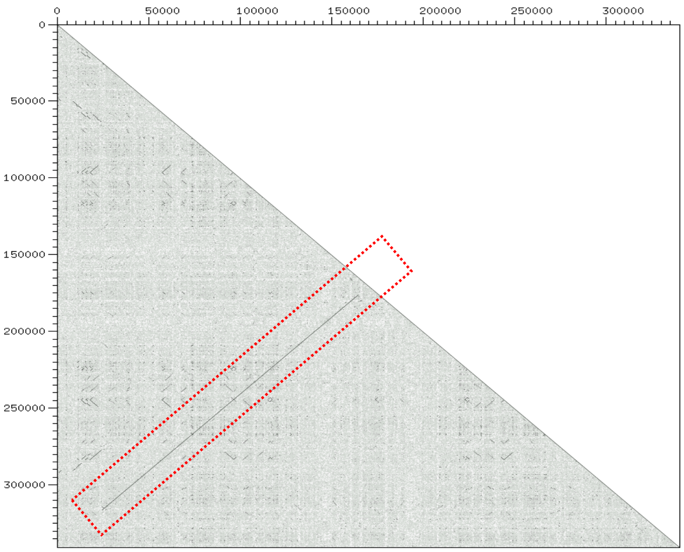


Figure 3-11 Diagram showing results of Dotter analysis of the genomic sequence flanking the two NXF2 loci (against itself). The red box highlights the inverted repeat.

Additionally, a TCP11-like gene upstream of the NXF2 locus was found to be included in the same duplication (Figure 3-12). The TCP11 gene (located on human 6p21.3-p21.2) encodes a receptor for fertilisation-promoting peptide, thought to play a role in fertility and sperm function (Ma *et al.*, 2002). Two transcripts were observed (represented by EMBL sequences AK057385 and AJ277659) that spanned the TCP11-like and NXF2 genes, linking their structures. This suggests that the two loci are a part of the same gene and potentially represent a gene-fusion event, as the other NXF genes are not linked to TCP11-like loci. Each locus appears to also give rise to separate transcripts also (represented by EMBL sequences AK005772 and AJ277526). An alternative explanation is that the mRNA transcript is an example of aberrant transcription. Without further study it is difficult to reconcile these alternate hypotheses. It is interesting to note in this regard that the NXF2 gene has been suggested to play a role in spermatogenesis (Wang *et al.*, 2001).

This provides a striking example of genomic sequence analysis revealing previously unknown complexity in gene organisation, and further studies could now be directed to elucidate roles of different TCP11-like and NXF2 transcripts. Any studies on NXF2 must now address the issue of two almost identical genes and transcripts complicating interpretation of results.

The occurrence of a conserved *Alu*Y repeat within the inverted repeat (Figure 3-12) provides evidence for the duplication having occurred subsequent to the divergence of the human and mouse lineages and approximately 15 Mya, when the *Alu*Y family is thought to have dispersed throughout the genome. An alternative explanation for the conserved *Alu*Y - that two copies integrated independently at similar positions - is highly unlikely.

However, another possibility is that gene conversion between the two loci has resulted in the propagation of an initial *Alu* insertion, and could account for the high level of sequence similarity seen. A relatively young age of the duplication would also be consistent with the very high level of sequence similarity seen.

At the time of writing, the presence of a sequence gap near the Nxf2 locus in mouse precluded confirmation of a single locus in the mouse, which would discriminate between these two alternate hypotheses but annotation of the mouse region did provide an indication that there may only be one locus (see Chapter 4).

The presence of two highly-related NXF2 loci in humans, with complex gene structures including transcripts spanning a TCP11-like gene, means that any studies aimed at elucidating the function of NXF2 in humans using information from mouse models must be interpreted with caution.

An RT-PCR experiment provided some information on NXF2 and autosomal TCP11 transcript tissue distribution (Figure 3-13). Primers were designed to autosomal TCP11, and to different TCP11-like and NXF2 variants shown in Figure 3-12. Attempts were made to design primers that would discriminate between some of the TCP11-like and NXF2 variants, and the positions of these primers are indicated in Figure 3-12. PCR was performed on cDNA from twenty different tissue RNA samples as described in Chapter 2.

A striking feature was the strong expression seen in testis for most of the NXF2/TCP11-like locus transcript variants and for autosomal TCP11, in accordance with what has been noted in the literature. Further detailed experiments would be required to investigate the patterns of different variant and TCP11-like/NXF2 fusion transcripts comprehensively.

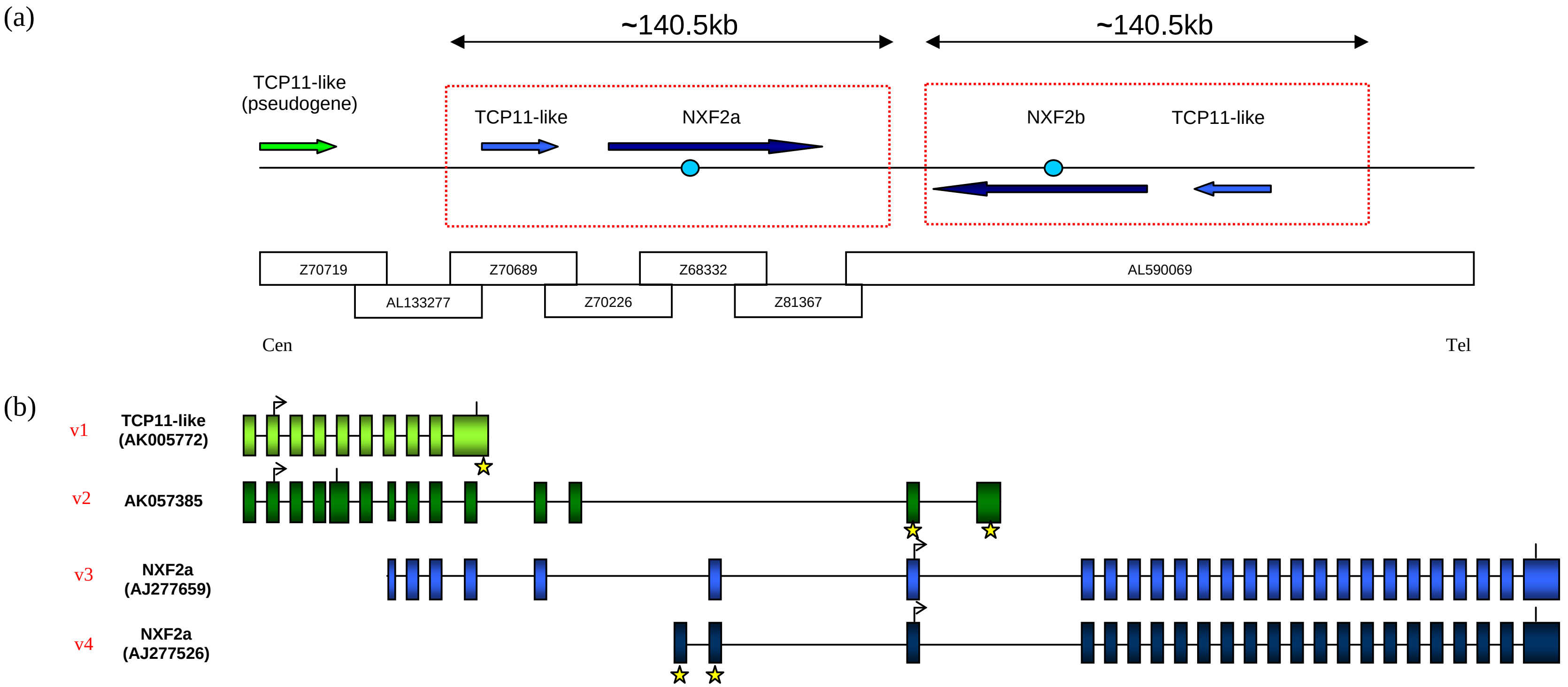
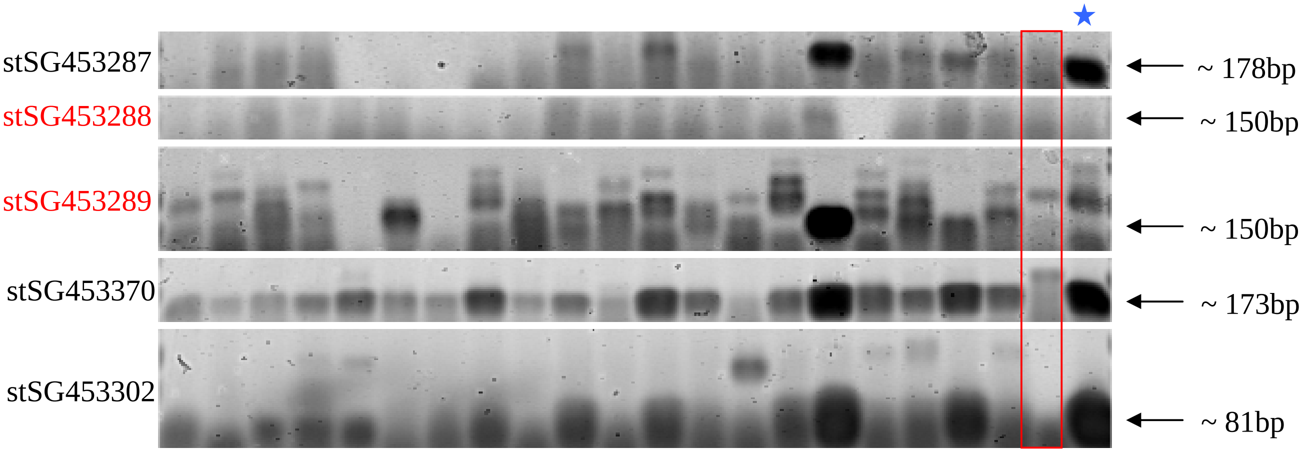


Figure 3-12 (a) A schematic representation of the inverted repeat containing the TCP11-like and NXF2 loci, and an adjacent TCP11-like pseudogene. The repeat boundaries are denoted by red boxes, and the location and transcriptional directions of relevant genes denoted by arrows. The blue circles represent an *Alu*Y repeat. Genomic clones forming the tiling path of the region are depicted as open boxes. (b) a schematic representation of the TCP11-like and NXF2 genes, and mRNAs linking the genes. Transcripts v1 and v4 represent the separate loci transcripts, and v2 and v3 the transcripts linking the loci. Boxes denote exons, connected by black lines representing introns. The angled arrows and upright lines represent putative start and stop codons respectively for predicted protein products. Asterisks denote positions of primers used for expression profiling, as shown in Figure 3-13. A single asterisk for a transcript denotes closely paired primers, two asterisks the individual primers.

(a)



(b)

	Gene	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
stSG453287	TCP11Lv1										▒		▒				■		▒	▒	
stSG453288	NXF2/TCP11Lv2																■	▨			
stSG453289	NXF2/TCP11Lv4						■										■			■	
stSG453370	TCP11	▒	▒	▒	■	■	▒	▒	■	▒	▒		■	■		▒	■	▒	▒	■	▒
stSG453302	NXF2	▨	▨	▨	▨	▨	▨	▨	▨	▨	■	▨	■	▨	▨	■	■	▨	▨	■	▨

Figure 3-13 (a) Images of Vistra Green stained 2.5% agarose gels containing RT-PCR products for primers designed to NXF2 and TCP11-like (TCP11L) variants and TCP11. The expected products are arrowed and their expected sizes shown. The red box in the gel images is the negative control lane, which whilst showing a faint product in some cases, is not the same size as that for specific product. The lane with a blue asterisk is the genomic DNA positive control. STS names in red denote primer pairs which span an intron. (b) a summary of the RT-PCR results, with tissues tabulated according to the images shown in (a). Black filled cells denote medium to strong PCR product bands detected, grey cells denote weaker bands and white cells denote no PCR product detected. Hatched cells denote uninformative tissues, where a RT-PCR reaction was omitted or product is difficult to discern prohibiting conclusions regarding expression in that tissue. The NXF2 variant used to design the primers is indicated, and relates to figure 3-12. STS stSG453302 was designed to the 3' exon of NXF2.

3.3.3 Alternative 3’-UTR usage

From annotation of gene structures within Xq22-q23, several instances were noted where 3’ ESTs were found to cluster at several positions in the 3’ UTR of a gene. These appeared to represent evidence of different polyadenylation (polyA) site usage. Three such genes, ALEX3, TBG and CSTF2, were chosen for RT-PCR studies to assess expression of the different 3’ UTR variants in different tissues.

BLAST matches of ESTs to the genes were studied, and primers were designed to regions of the 3’ UTR just upstream of the different polyadenylation sites (indicated by common start sites of 3’ EST matches). These primers were then used to screen twenty human cDNA samples from tissue total RNAs by PCR (see Chapter 2).

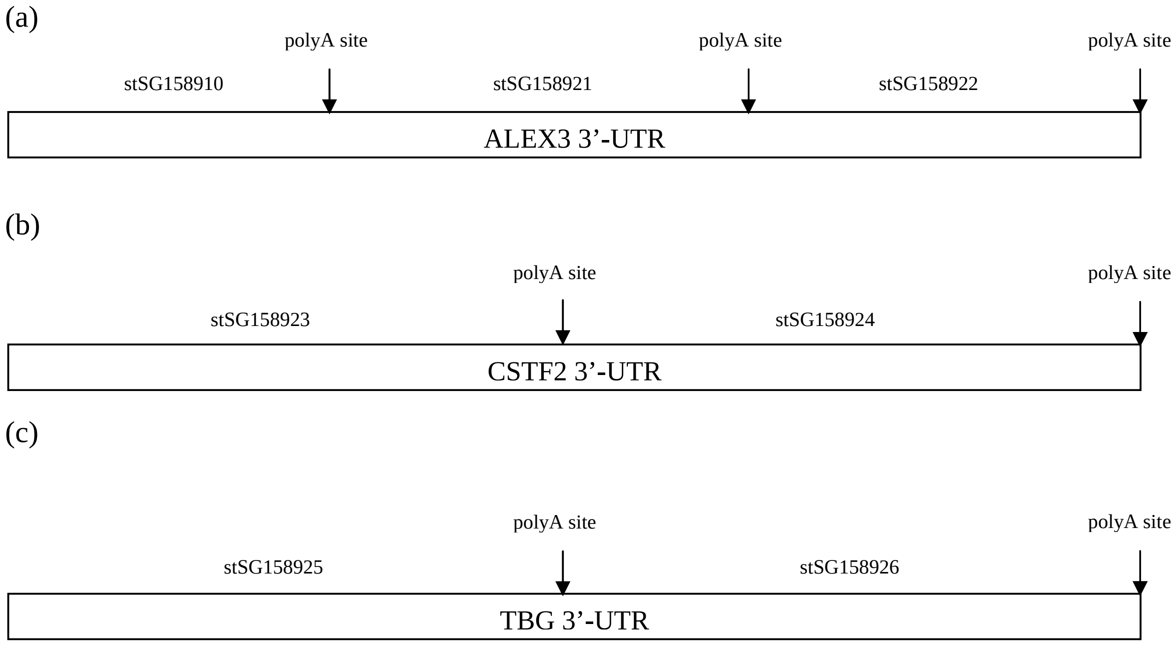


Figure 3-14 a schematic representation of the primer positions within the 3’ UTRs of (a) ALEX3, (b) CSTF2 and (c) TBG.

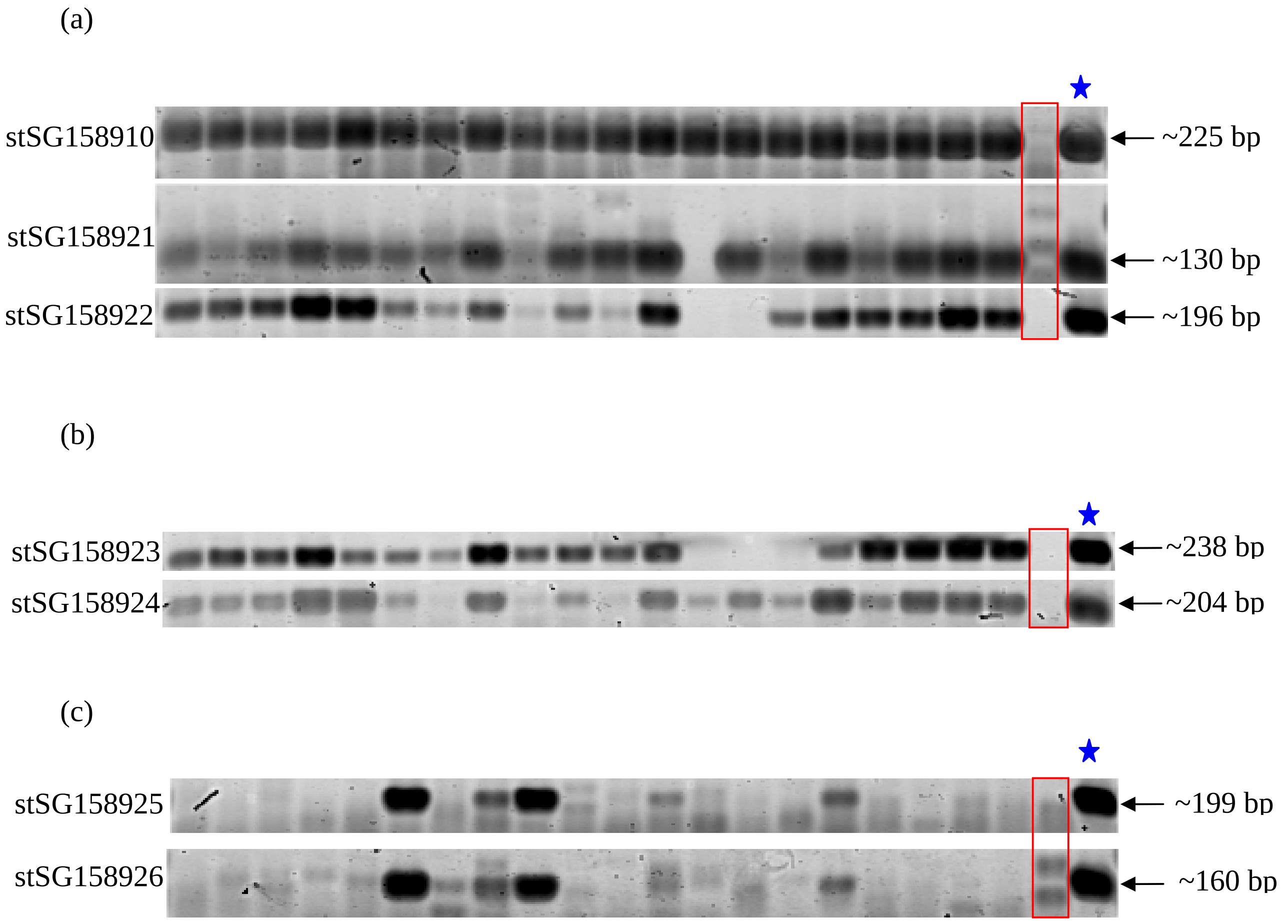


Figure 3-15 Images of 2.5% agarose gels containing RT-PCR products for primers designed to 3’-UTR variants, for (a) ALEX3, (b) CSTF2 and (c) TBG. Expected product band sizes are shown with an arrow. The red box in the gel images is the negative control lane, which whilst showing a faint product in some cases, is not the same size as that for specific product. The lane denoted with a blue asterisk is the genomic DNA positive control lane.

For CSTF2 and TBG, the most 5’ STS would detect both UTR transcripts, with the more 3’ STSs detecting the longer transcript. For ALEX3, the most 5’ STS would detect all three transcripts, with the next most 3’ STS detecting two longer transcripts and the furthest 3’ STS detecting the longest transcript.

	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
stSG158910	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
stSG158921	■	▒	■	■	■	■	■	■		■	■	■	▨	■	▒	■	▒	■	■	■
stSG158922	■	■	■	■	■	■	■	■	▒	■	▒	■			■	■	■	■	■	■

(a)

	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
stSG158923	■	■	■	■	■	■	■	■	■	■	■	■	▨	▨		■	■	■	■	■
stSG158924	■	■	■	■	■	■		■	▒	■	▒	■	▒	■	▒	■	■	■	■	■

(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
stSG158925						■		■	■			▒				■				
stSG158926		▒			▒	■	▒	■	■			▒				■				

(c)

Figure 3-16 A tabulated summary of the RT-PCR results, with tissues tabulated according to the images shown in Figure 3-15, for (a) ALEX3, (b) CSTF2 and (c) TBG. Black filled cells denote medium to strong PCR product bands detected, grey cells denote weaker product bands detected and white cells denote no PCR product band detected. Hatched cells denote uninformative tissues, where a RT-PCR reaction was omitted or product is obscured, prohibiting conclusions regarding expression in that tissue.

The results of these experiments are shown in Figure 3-15 and Figure 3-16. For ALEX3, some differences in expression were seen for different 3'-UTR variants. The longest UTR variant was not detected in salivary gland or skeletal muscle, indicating that in these tissues the first and second polyadenylation sites are preferentially utilised.

For CSTF2, the longer UTR variant was not detected in heart, indicating that in this tissue the first polyadenylation site is preferred. No distinct differences in transcript detection were noted for TBG.

For other tissues, in some cases longer UTR variants were detected in tissues where a more 5’ STS (which should detect both shorter and longer variants) had failed to detect product. This illustrates limitations of the RT-PCR approach. The results noted above for ALEX3 and CSTF2 are more clear however, and whilst further work would be needed to confirm this preliminary data, some differences in expression patterns of different 3’ UTR variants have been suggested.

3.3.4 Mitochondrial insertion into the nuclear genome at Xq22

Annotation of Xq22 revealed a sequence (bA522L3; accession AL590407) with BLASTX matches to all the proteins encoded by mitochondrial genome. The matches lie within an intron of the gene dJ769N13.CX.3, as depicted in Figure 3-17.

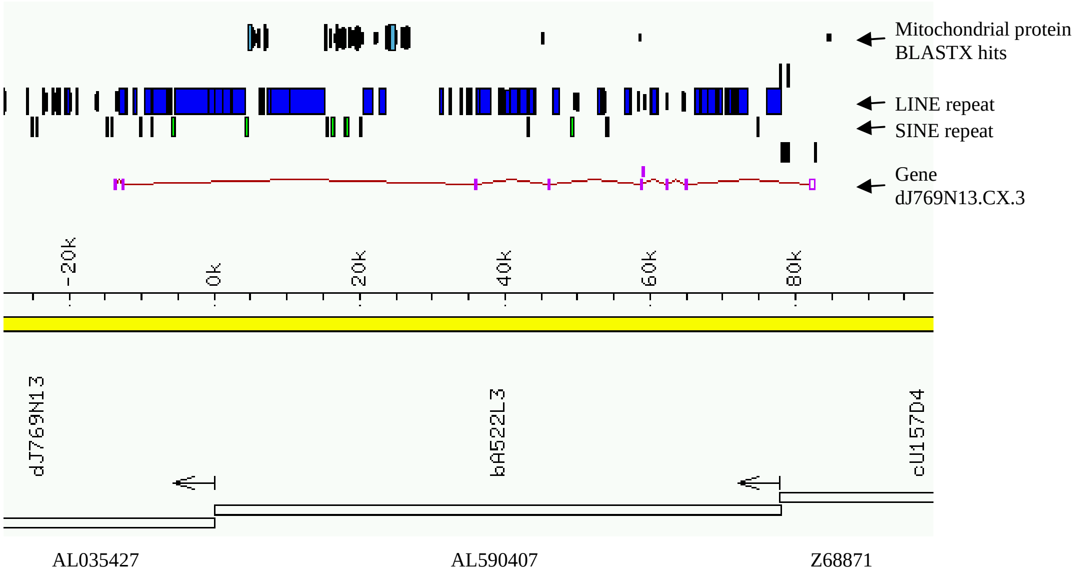


Figure 3-17 Xace representation of mitochondrial genome-encoded protein matches within clone bA522L3. Vertical magenta boxes represent exons of dJ769N13.CX.3, connected by horizontal lines representing introns. The positions of the mitochondrial protein homologies are shown. Genomic clones are depicted and their accession numbers shown at the base of the figure.

Nuclear genome sequences related to mitochondrial sequences have been observed before, but rarely to this extent (Tourmen *et al.*, 2002). Further investigation of these homologies revealed well conserved order and orientation of the homologies with respect to the mitochondrial genome, and a BLASTN comparison of the Xq22 sequence against the mitochondrial genome (performed by Dr. Julian Parkhill) confirmed that the matches appeared to represent an almost complete insertion of the ~16.6 kb mitochondrial genome into the nuclear genome at Xq22 (see Figure 3-18). BLAST matches from in order from the 12S RNA gene to the CYTB gene indicate insertion of approximately bases 650-15882 of the 16.571 kb mitochondrial genome, (approximately 92%). This is an approximation as some segments do not show high BLAST matches.

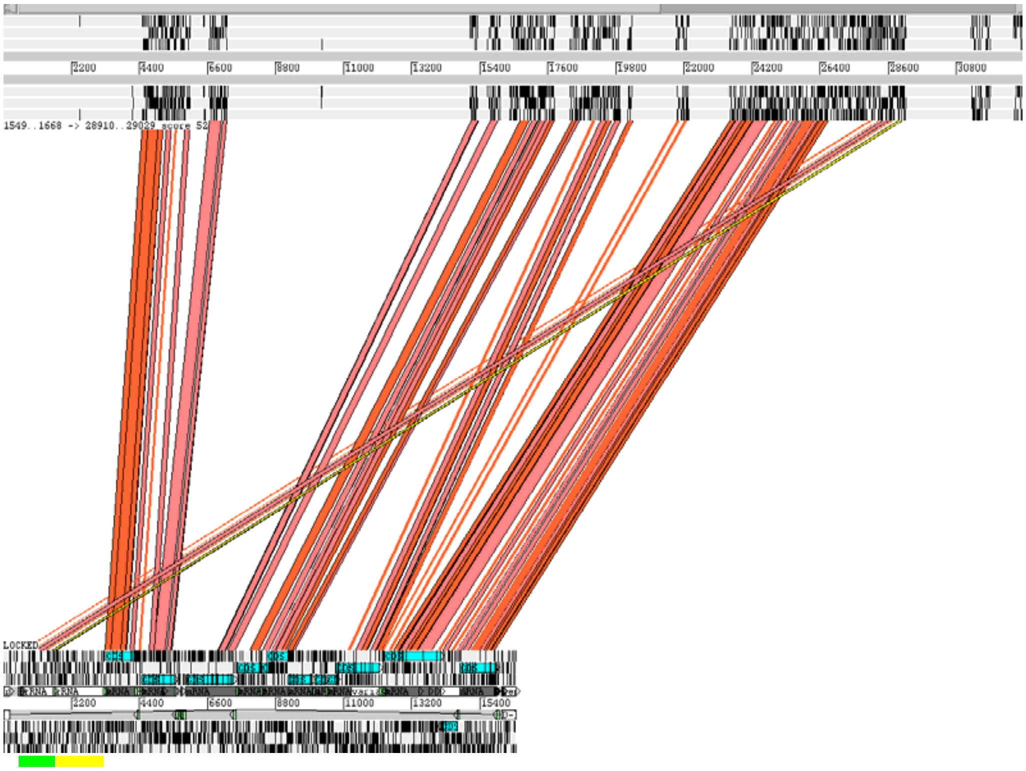


Figure 3-18 Diagram illustrating BLASTN matches between human Xq22 sequence (from clone bA522L3) and the mitochondrial genome. The matches were visualised using ACT (thanks to Dr. Julian Parkhill, Microbial Sequencing Unit, Wellcome Trust Sanger Institute). The red lines represent BLASTN matches. The upper section represents bA522L3 sequence (masked for repeats), and the lower section the mitochondrial genome. The green and yellow bars underline approximate positions of the 12S and 16S rRNA genes respectively. The blue annotations of the mitochondrial genome depict positions of protein-coding genes.

Furthermore, the pattern of BLAST matches seen in Figure 3-18 suggests a mechanism for the insertion event. Between the 12S and 16S rRNA genes, a break in the order of the BLASTN matches is seen, whereby the 12S matches are seen distal to the Cytochrome b gene in the nuclear genome. This suggests that a breakage occurred in the mitochondrial genome sequence between the 12S and 16S rRNA genes, and that the linearised mitochondrial genome then integrated into the nuclear genome via a DNA-mediated mechanism. The integration could also have occurred via recombination between the two genomes, with the recombination site located between the 12S and 16S genes (Figure 3-19). The other alternative, that the integration occurred via an mRNA transcript, is much less likely: the mitochondrial promoter lies upstream of the 12S rRNA gene, and insertion via the transcript should result in completely co-linear homologies between the two genomes.

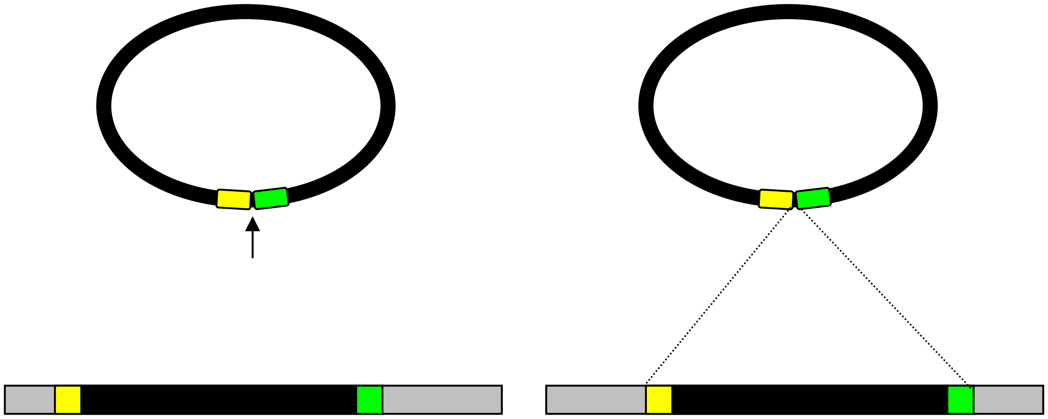


Figure 3-19 Schematic diagram of a model for integration of the mitochondrial genome in sequence accession AL590407 (bA522L3). Grey bars represent the nuclear genome, black bars the mitochondrial genome. The yellow boxes represents the 16S rRNA gene, the green boxes the 12S rRNA gene. The left section represents a linearization-based model, the right section represents a recombination-based model.

3.4 Discussion

The studies presented in this chapter have demonstrated the utility of genomic sequence information, when combined with the availability of large-scale mRNA sequence data, in the identification and description of genes. When this study began, 30 genes were noted within the region; when the study was completed 74 genes, 51 predicted genes and 46 pseudogenes had been manually annotated within the region. A feature of note was the annotation of a GK pseudogene, which probably accounts for the mis-assignment of the GK gene to Xq22 (Grutzner *et al.*, 2002), illustrating the benefit of manual annotation.

Initially, many novel genes were identified, often as partial structures. A more complete description of these structures required targeted screening of cDNA resources. However, as the study progressed, the release of large amounts of mRNA sequence information superseded these efforts, and illustrated the utility of that resource in gene identification.

During construction of the transcript map, the manual analysis and annotation of 15 Mb of human genomic sequence revealed several unusual aspects of gene organisation. The NXF2 locus provided a good example of how genomic sequence information combined with annotation can reveal subtleties in gene structures that are unlikely to be identified from mRNA-based approaches alone - in this case the presence of an almost identical copy of the gene, which could potentially be under different transcriptional regulation. It also highlighted a previously unobserved fusion of the NXF2 gene structure with that of a TCP11-like gene, an intriguing observation given that NXF2 and TCP11 have been implicated in male fertility.

The observation of alternative polyadenylation site usage by several genes within the region, a small sample compared to the genome as a whole, highlights that alternate polyadenylation site usage is a widespread occurrence. Some differences in expression patterns were seen for different 3’ UTR variants for ALEX3 and CSTF2, but these studies were limited in scope and did not address any temporal aspects of differences between variants. Alternate polyadenylation site usage could be used to control the incorporation of elements conferring different mRNA stability or localisation properties. The presence of functional sequences within the 3’ UTR of genes suggests that further studies of alternative polyadenylation of genes will aid in the understanding of their transcriptional and translation control, and will need to be taken into account in completing annotation of the genome.

The discovery of an almost complete (approximately 92%) insertion of the mitochondrial genome into the nuclear genome not only demonstrated utility of the genomic sequence in uncovering events in genome evolution, but also provided information which allowed a DNA-mediated mechanism of insertion to be inferred. The presence of various nuclear mitochondrial insertions (“numts”) has been noted, and the example presented here is unusual in its completeness. Early BLAST analysis of the draft human genome sequence identified 1105 sequences homologous to mitochondrial DNA, representing 286 pseudogenes (Tourmen *et al.*, 2002). From this study, only seven numts greater than 10 kb in length were found. Insertion of the mitochondrial genome or fragments thereof into the nuclear genome, presumably occurring over a period of time, highlights the dynamic nature of the genome and the potential for interaction of cellular material normally segregated within the cell.

The generation of a transcript map of the Xq22-q23 region will prove valuable in studies aimed at screening genes for mutations in hereditary disorders, and was utilised in one such approach attempting to identify the DFN2 gene (collaboration with Dr. Jess Tyson, Institute of Child Health, London).

Most importantly, the gene annotation map provided evidence of an unusually high number of duplicated genes within the region, as well as a set of paralogues that appear part of a larger segmental duplication resulting in paralogy between Xp and Xq22. Sequence repeats within Xq22 had been noted previously (G.R. Howell, personal communication) and the studies presented in this chapter revealed the striking degree of gene duplication present.

These observations provided the impetus for the studies presented in Chapter 4 where the region equivalent to Xq22 was investigated in order to ascertain the level of duplication within the mouse region. The gene duplications within Xq22 and the larger segmental duplication are described in detail in Chapter 5 and 6 respectively.

During this study, genomic sequencing and automated annotation of the genome (Ensembl, UCSC genome browser and NCBI map viewer) progressed rapidly. Whilst invaluable in genomic studies and interpreting the genome, automated approaches alone may miss subtleties of gene structure and genomic organisation, and should be combined with careful manual annotation. This is indeed now being adopted, by the HAVANA group (Wellcome Trust Sanger Institute) and VEGA initiative (Wellcome Trust Sanger Institute).

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

4.1 Introduction

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

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

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

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

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

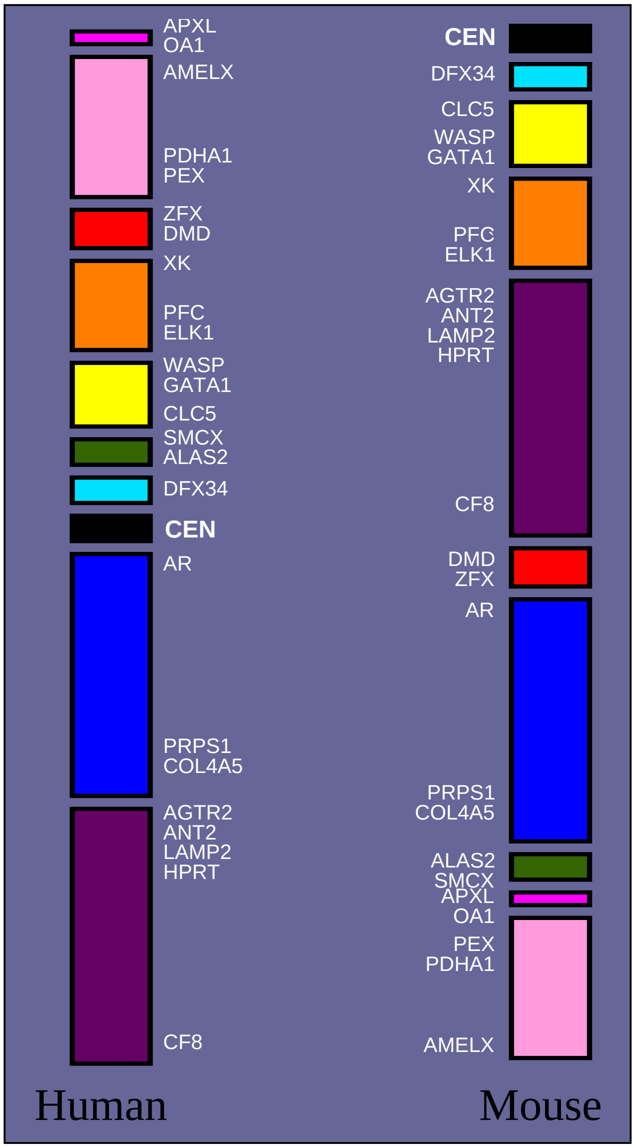


Figure 4-1 Figure illustrating the mouse X chromosome showing blocks with conserved synteny to the human X chromosome (reproduced from MRC Harwell website).

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

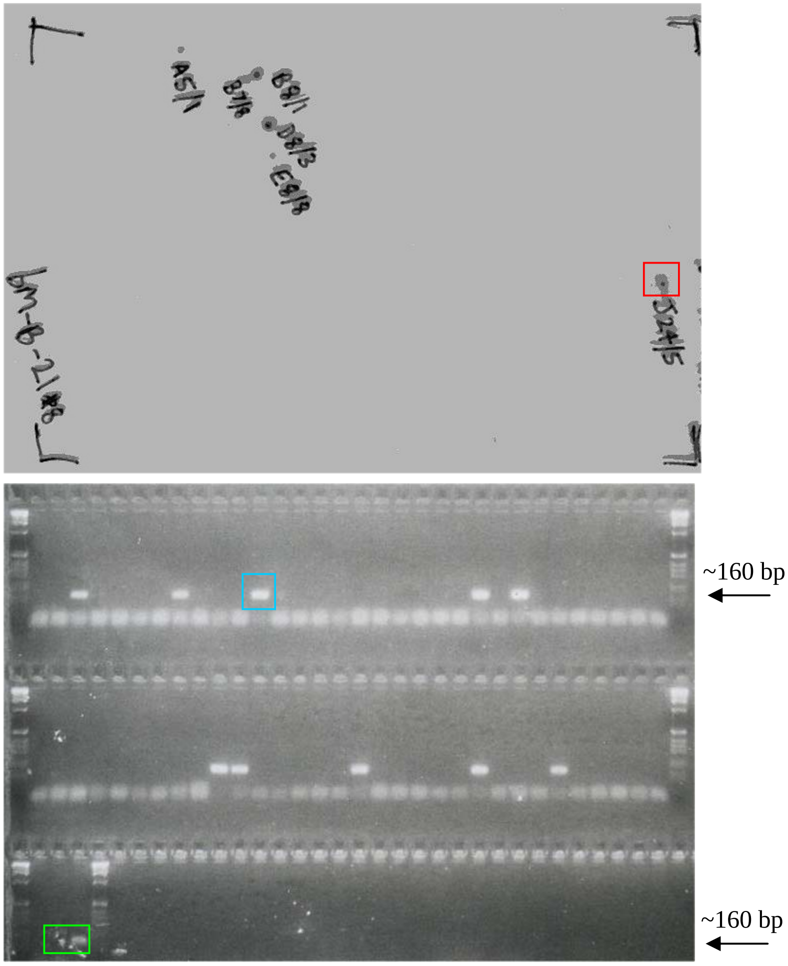


Figure 4-2 Diagram illustrating the STS-based hybridisation strategy used to isolate mouse RPCI-23 BAC clones. The upper image shows an autoradiograph of a mouse BAC filter following hybridisation of pooled radiolabelled STS products and washing. The red box highlights a positive signal for BAC bM325J24.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

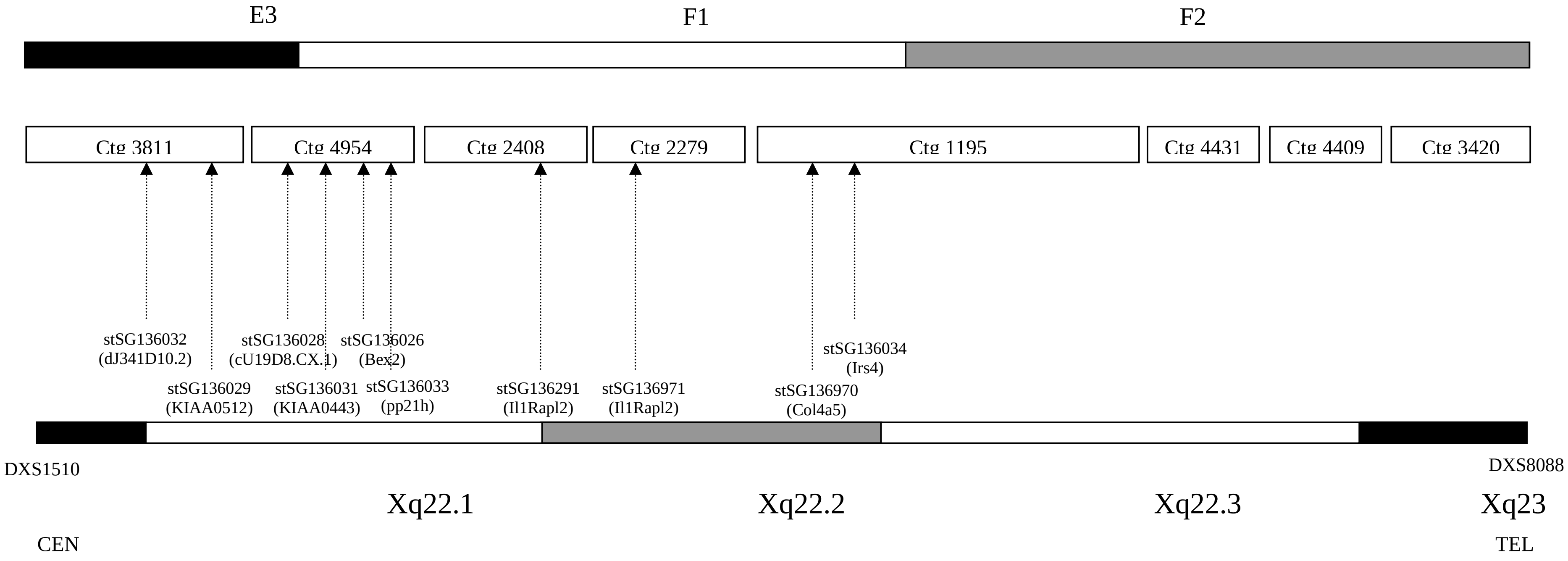


Figure 4-3 A first-generation BAC contig map of the mouse X E3-F2 region. The mouse G-banding pattern of the region is shown at the top of the figure. Initial BAC contigs are shown as open boxes. The approximate positions of BACs positive with mouse gene probes are arrowed. The STS and gene names are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

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

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

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

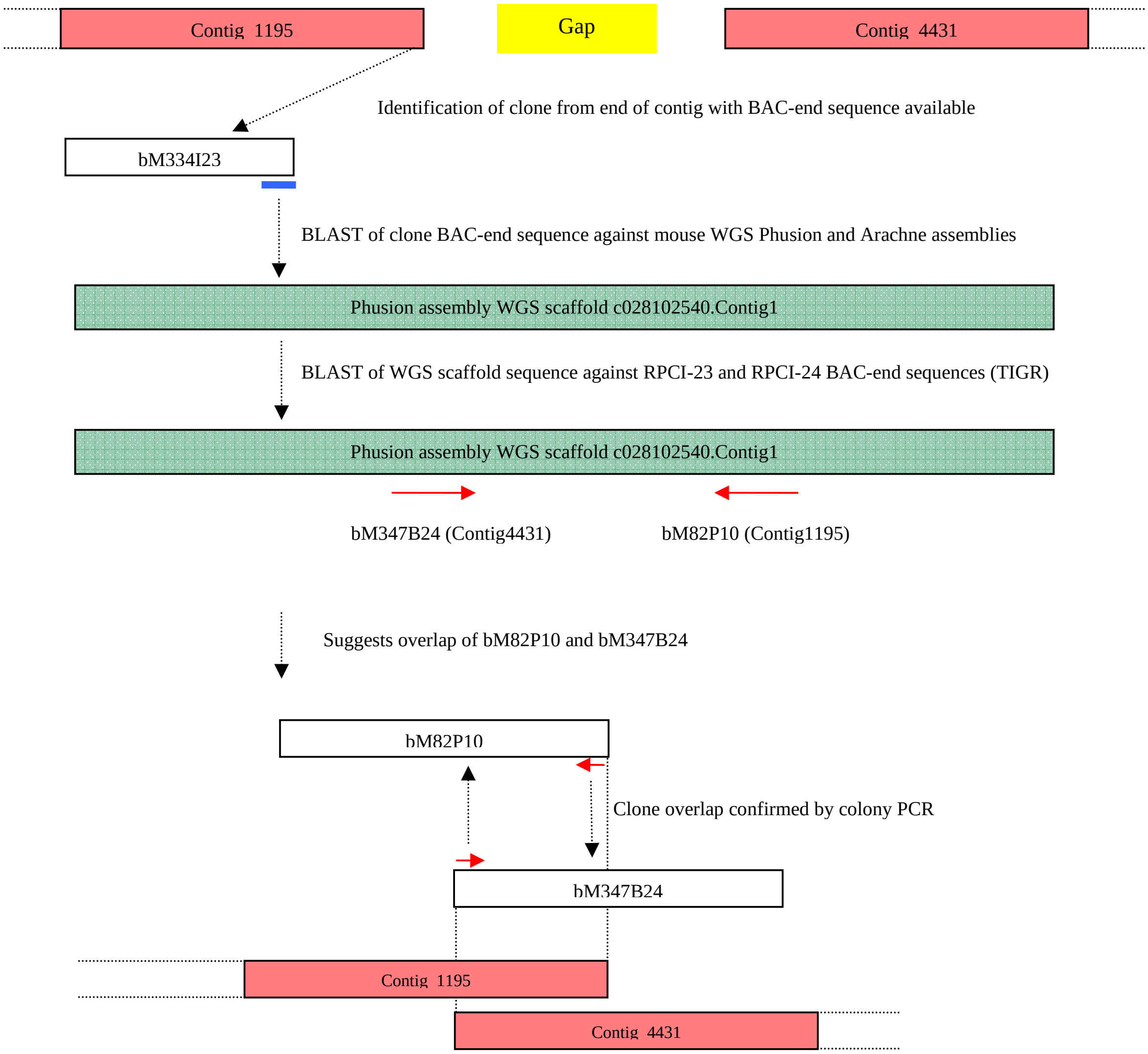

Contig 1195
Gap
Contig 4431
Identification of clone from end of contig with BAC-end sequence available
bM334I23
BLAST of clone BAC-end sequence against mouse WGS Phusion and Arachne assemblies
Phusion assembly WGS scaffold c028102540.Contig1
BLAST of WGS scaffold sequence against RPCI-23 and RPCI-24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

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

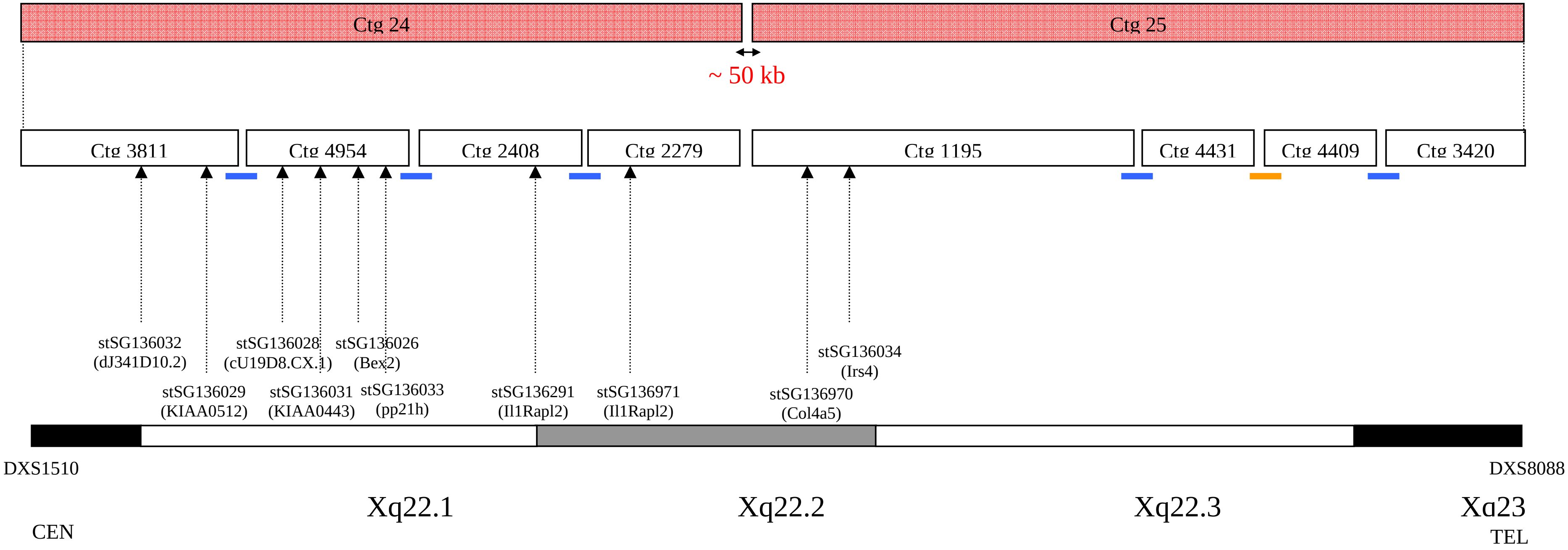


Figure 4-5 Finalised BAC contig map of the mouse X E3-F2 region. Initial BAC contigs are shown as open boxes. The approximate positions of BACs positive with mouse gene probes are arrowed. The STS and gene names are drawn beside the G-banded ideogram of the human Xq22 region to indicate the locations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

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

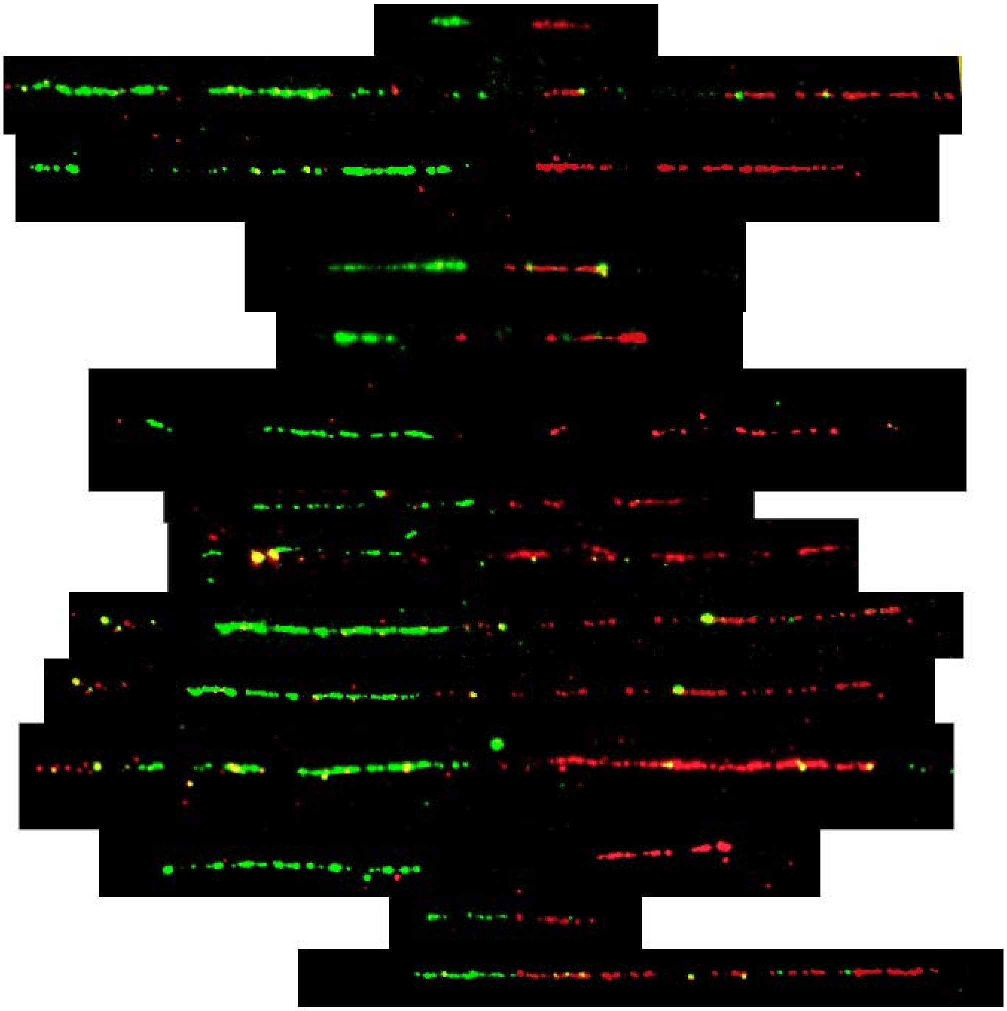
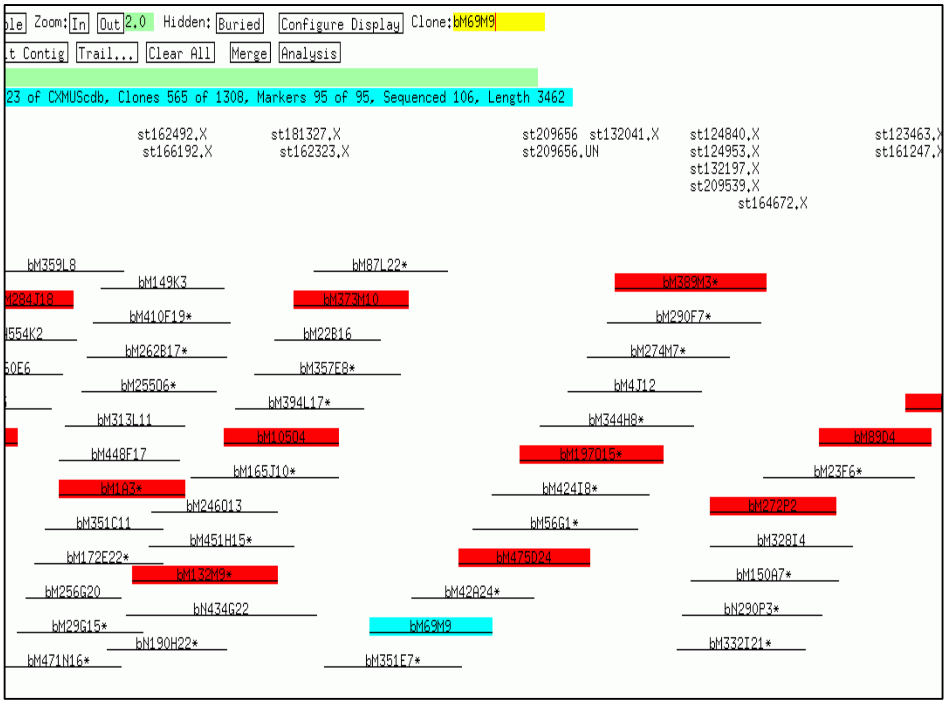


Figure 4-6 Results of fibre-FISH of bM149O3 (Ctg3811/2279 - red) and bM62F12 (Ctg1195 – green). A composite of images captured from 14 separate fibres is shown. A gap of ~ 50kb can be estimated (assuming ~ 150 kb per clone).

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

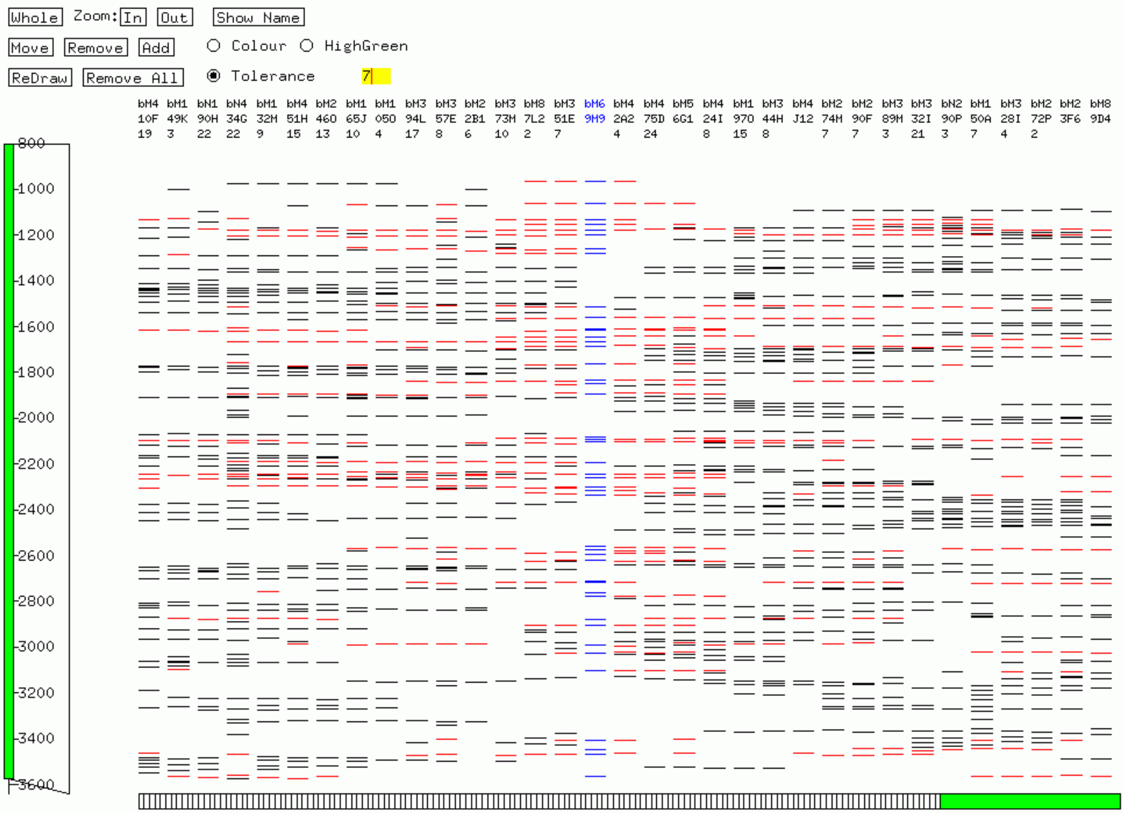
Gap closures resulted in two sequence-ready BAC contigs covering the mouse X chromosome E3-F2 region, renamed Contig 24 and Contig 25. A tiling path of BAC clones was chosen based on shared fingerprint bands – 68 clones were chosen for Contig 24 and 31 clones for Contig 25. These 99 BACs were picked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



Clone:bM69M9

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

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



● Tolerance 7

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

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

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

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

4.3 Identification of genes and their structures using sequence analysis

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

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

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

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

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

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

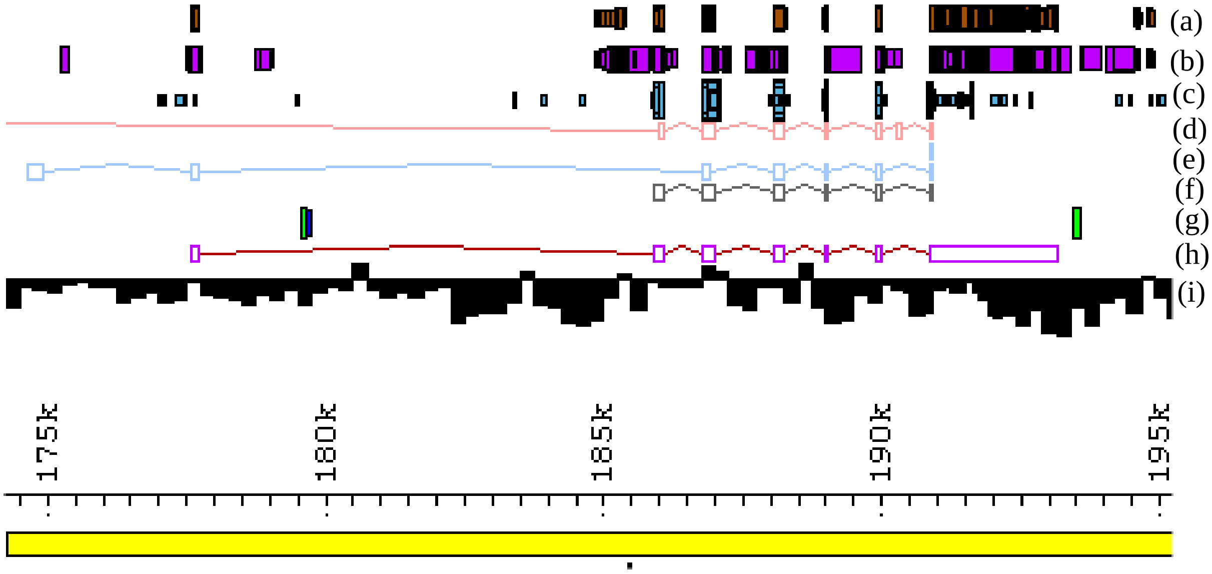


Figure 4-9 Diagram illustrating a “gene” (GD_mRNA structure) structure, for the Plp gene (locus bM197O15.MX.3). The diagram shows an ACeDb representation of the gene structure. Key – (a) mRNA BLASTN matches, (b) EST BLASTN matches, (c) protein BLASTX matches, (d) FGENESH gene prediction, (e) GENSCAN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

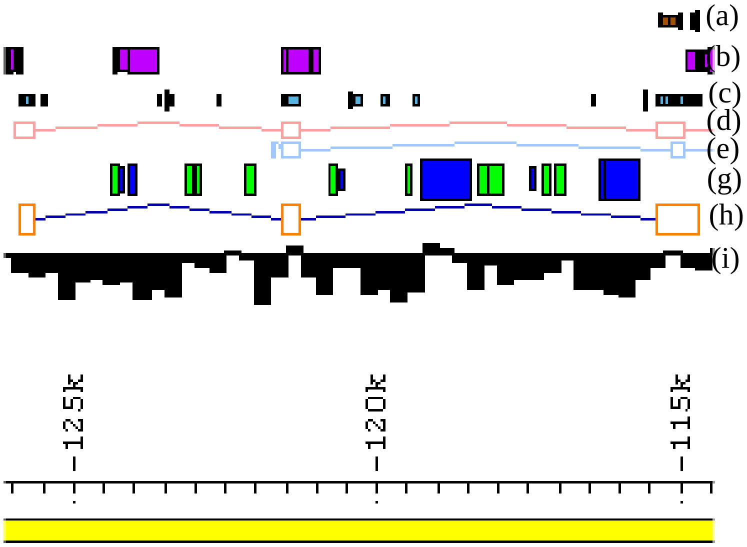


Figure 4-10 Diagram illustrating a "predicted gene" (GD_supported structure) structure, for locus bM182N4.MX.3. The diagram shows an ACeDb representation of the gene structure. Key – as for Figure 4-9. In this case, the gene structure was annotated from BLASTX matches to human XK protein (accession P51811).

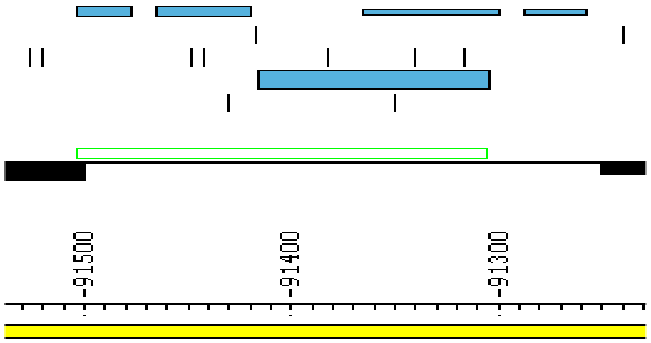


Figure 4-11 Diagram illustrating a "pseudogene" (pseudogene structure) structure, for the pseudogene locus bM389M3.MX.4. The diagram shows an ACeDb representation of the gene structure. Key: green box – annotated pseudogene, blue boxes – BLASTX protein matches, vertical lines – boundaries of open reading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

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

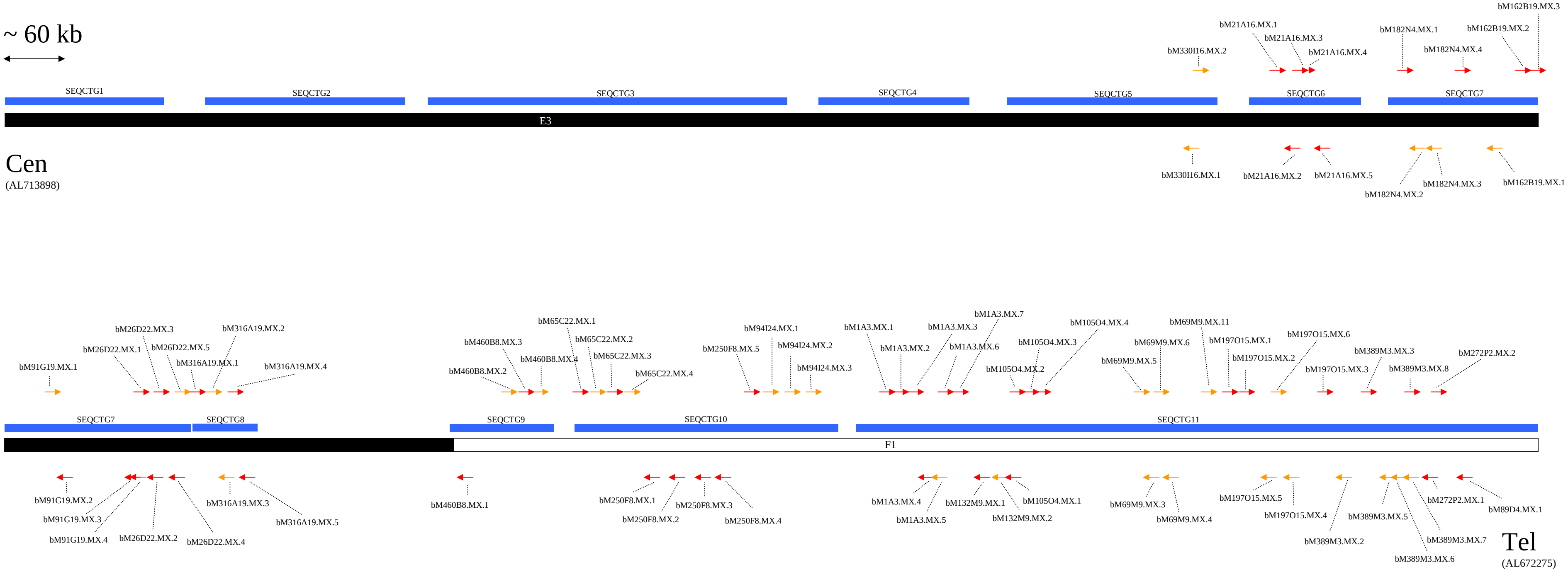


Figure 4-12 Genes annotated on finished sequence of the mouse X E3-F2 region from clone bM253D13 (AL713898) to bM89D4 (AL672275) (in contig 24), annotated as described in this Chapter. The region beginning is at top left, continuing onto the lower section of the diagram. "Cen" denotes the centromeric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

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

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

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

4.4.1 *Orthologues of human Xq22 genes*

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

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

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

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

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

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

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

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

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

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

4.4.2 PARL repeats

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

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

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

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

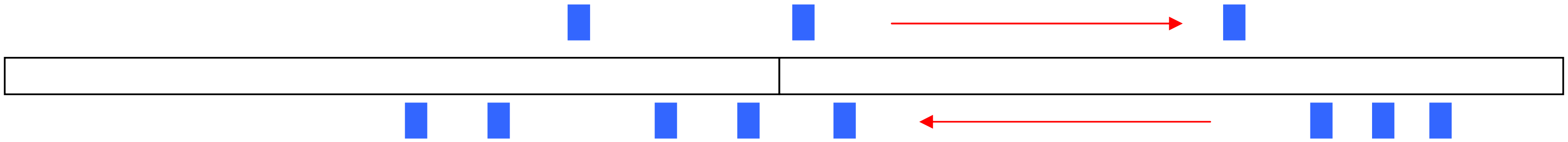
The sequence of the region containing the Parl-like loci was also analysed using Dotter (Sonnhammer and Durbin, 1995) to identify repeats in context with the gene loci (Figure 4-14). The program aligns two sequences against each other, and nucleotide identities are plotted as points to scale along the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

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

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

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

(a)



(b)

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

                *      1100         *      1120         *      1140         *      1160         *      1180         *      1200         *      1220         *      1240         *      1260
bm105o4mx8 : TAAAACTATTT-CTTCAGCAAAAT----GTACATTTATAAAGGTACCGACGAAGAATAACCTTGAAATC-AGTCAGTGTCATGGGACAAGCATACAG---AAGTGCCTTGATCCAGTCCCTGAAGGAACCA---GTCTTTTGTCCTCCCCCTAGCTCATAAGAACTCCTGTGAAGCCT :  332
bm105o4mx9 : TAAAACTATTT-CTTCAGCAAAAT----GTACATTTATAAAGGTACCGACGAAGAATAACCTTGAAATC-AGTCAGTGTCATGGGACAAGCATACAG---AAGTGCCTTGATCCAGTCCCTGAAGGAACCA---GTCTTTTGTCCTCCCCCTAGCTCATAAGAACTCCTGTGAAGCCT :  332
bm105o4mx7 : TAAAATTATTT-CTTCAGTAAAAT----GTACATTTATAAAGGTACCGACGAAGAATAACCTTGAAATC-AGTCAGTGTCATGGGACAAGCATACAG---AAGTGCCTTGATCCGGTCCCTGAAGGAACCA---GTCTTTTGTCCTCCCCCTAGCTCATAAGAACTCCTGTGAAGCCT :  332
bm105o4mx6 : TAAAATTTTTTCTTCAGCAATAT----TTACATTTATAAAAGTACCAAGGAAAATAAACCTTGAATATC-AGATAGTGTTAAGGGACAAGCATACAG---AGGCGCCTTGATCCCACGCCTGAAGGAAACA---GTCTTTTGTCCTCCCC-TAGCTTATAAGAAGTATAGTGAAGCGG :  332
bm69m9mx10 : TAAAATTATTT-CTTTAACAAAAAAAGAGTACATTTATAAAGGTACCGAGGAAGAATAACCCTTGAAATC-AGTCAGTGTCATGGGACAAGCATACAG---AAGTGCCTTGATCTGGCCCCTGAAGGAAACA---GTCTTTTAACCTCCACCTAGCTCATAAGAACTCCTGTGAA--CT :  237
bm69m9mx1  : TTAACTTATTT-CTTCAGCAAACT----ATGGATTTAGAAAAGTACTGAGGAAGGCTGAACCTTGAAATTTAAACAGTGTCAAGCTGAGAAGCATACAAGACAAGTGCCTTGATCTTGTCCCTGGAGGAAACGCTTGTGTGCTTTCGTCCCCCTAGCTTATAGGATCTCTCGAGAAGCCT :  287
bm69m9mx8  : TTAACTTATTT-CTTCAGCAAACT----ATGGATTTAGAAAAGTACTGAGGAAGGCTGAACCTTGAAATTTAAACAGTGTCAAGCTGAGAAGCATACAAGACAAGTGCCTTGATCTTGTCCCTGGAGGAAACGCTTGTGTGCTTTCGTCCCCCTAGCTTATAGGATCTCTCGAGAAGCCT :  193
bm105o4mx5 : TGAAATTATTT----CTTCAGAAAAAA-AGTACATTTATAAAGGTACCGAGGAAGGATAAACTGTGAAGATCAGACAGTGTTATGGCGACAAGCATACAG---ATGTGCCTTGATCCCACCCCTAGAGGAAACA--GTCTTTTATCCTTGCCCTTAGCTTATAAGAACTGCAGTGAAGCCC :  335
bm69m9mx9  : ---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------- :    -
bm69m9mx2  : ATTTGGAAGAA----CAGGGAGCCTCT-AGTGAAAGTCTGGTATAAAATAAGGACTAACAACCCCAAGAATGTAGATGGTTTAAGTGAAGCAAAGTAAGT---CACTATGGTGGTGTGTCTGAACCTTGCTGCC--TG---------------------------------------- :  421
bm69m9mx7  : ATTTGGAAGAA----CAGGGAGCCTCT-AGTGAAAGTCTGGTATAAAATAAGGACTAACAAACCCAAGAATGTAGATGGTTTAAGTGAAGCAAAGTAAGT---CACTATGGTGGTGCATCTGAACCTTGCTGCC--TG---------------------------------------- : 1213
```

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

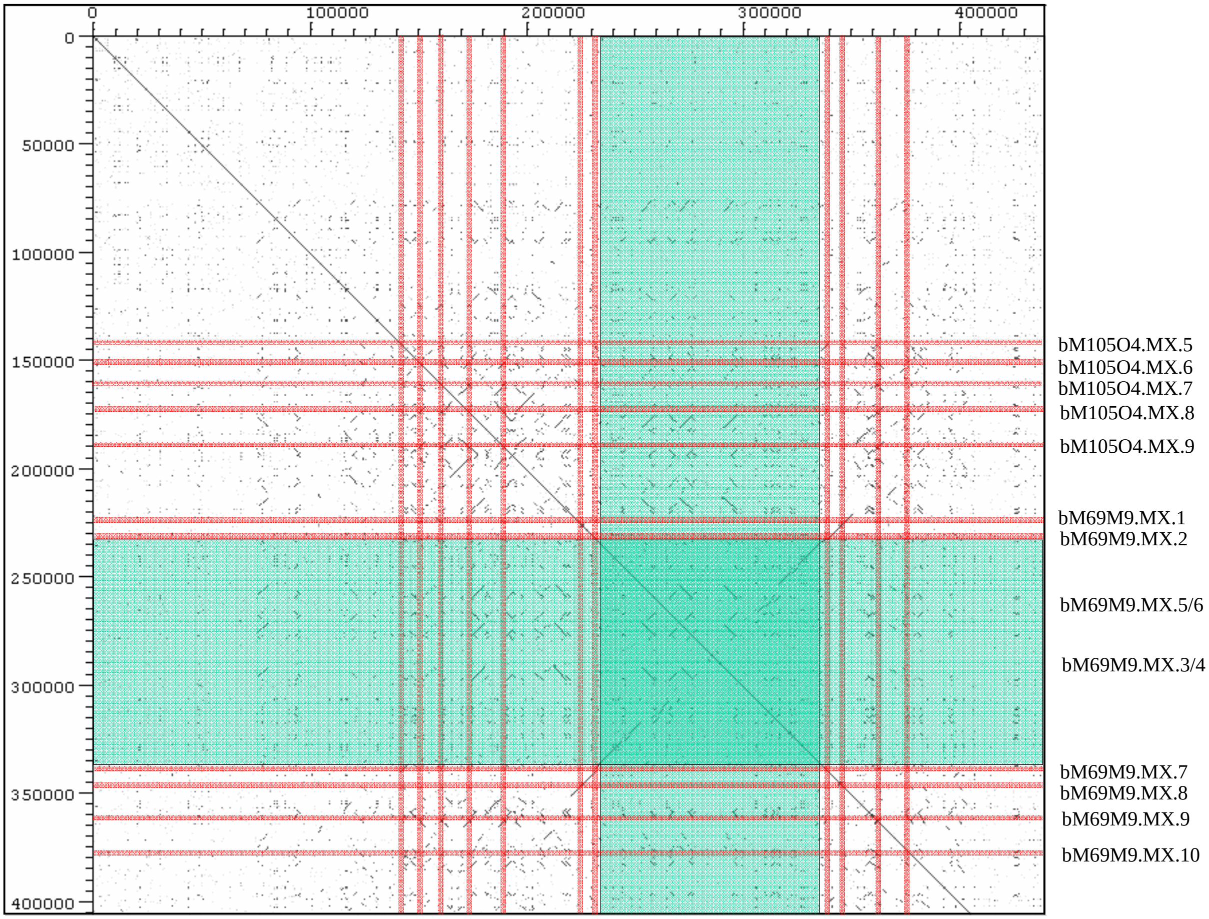


Figure 4-14 Diagram illustrating Dotter analysis of the *Mus musculus* X E3-F2 region containing PARL-like repeats. Approximate positions of PARL-like loci (illustrated in Figure 4-13) are shown by red bars located on both axes. The approximate position of the region containing the Kir3dl1 gene is shown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

4.4.3 Pramel3L repeats

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

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

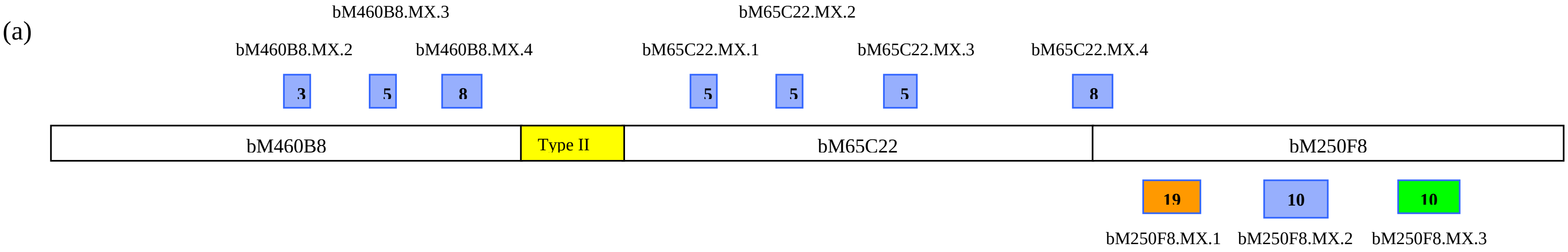
Gene	bM460B8. MX.3	bM460B8. MX.2	bM460B8. MX.4	bM65C22. MX.1	bM65C22. MX.2	bM65C22. MX.3	bM65C22. MX.4	bM250F8. MX.2
% ID	100	71	69	99	69	99	70	73

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

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

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

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



(b)

```
                   1240         *      1260         *      1280         *      1300         *      1320         *      1340         *
bm460b8mx4 : GGATCTTTTCCTTGATGGCTCCTTAATTGAAAAGGATTTTTTGATTTTGCTTATGCATAAAATTGAAGAGAGTTTAGGGTTTTTGCATGTGTGCTGTCGAGATTTGCAAATTGATAAACCGTG : 1140
bm65c22mx4 : GGATCTTTCCCTCAATGGCTCCTTAATTGAAAAGGATTTTTTGATTTTGCTTATGCATAAAATTGAAGAGAGTTTAGGGTTTTTGCATGTGTGCTGTCGATATTTGCAAATTTATAAGTTGTG : 1140
bm65c22mx2 : GGATCTTTTCCTTGATGGCTCCTTAATTGAAAAGGATTTTTTGATTTTGCTTATGCATAAAATTGAAGAGAGTTTAGGGTTTTTGCATGTGTGCTGTCGAGATTTGCAAATTGATAAACCGAG :  856
bm460b8mx3 : GGACATTTCCCTTGATGGCACTTTGAGGGAAAGGAATTTTTTTGCTTTGCTTCAGAATAAAGTAGAGCAGAGCCTAGGCTCTCTGCACCTGTGCTGCAGAGATTTGCAAATTAATAACTTGTG :  878
bm65c22mx1 : GGACATTTCCCTTGATGGCACTTTGAGGGAAAGGAATTTTTTTGCTTTGCTTCAGAATAAAGTAGAGCAGAGCCTAGGCTCTCTGCACCTGTGCTGCAGAGATTTGCAAATTAATAACTTGTG :  875
bm65c22mx3 : GGACATTTCCCTTGATGGCACTTTGAGGGAAAGGAATTTTTTTGCTTTGCTTCAGAATAAAGTAGAGCAGAGCCTAGGCTCTCTGCACCTGTGCTGCAGAGATTTGCAAATTAATAACTTGTG :  875
bm460b8mx2 : GGGCCTTTCCCTTAACAGCAACTTGAGAACAAGGACATTTTTGTCTTTCCTTCTGAGTAAAGTTGAACAGAGCTCAGGGTCCTTACATCTCTGCTGTCGTGATTTGCAAATTTATAAACTGTC :  628
bm250f8mx2 : AGACCTTTCCCTTGATGGTACCTTGAGAGAAAGGGAATTTTTTGCTTTGCTTCTGAATAAAGTACAGCAGAGCTCAGGGTCTTTGCACCTCTGCTGCCGAGATCTACAAATTGATAGATTTC : 1350

                1360         *      1380         *      1400         *      1420         *      1440         *      1460         *
bm460b8mx4 : TAAATGCAAATGCACCCTGAAATTTCTTGATCTCCAATGTGTTGATCAGTTGTCAGTTGATAGAGGCTCACTGAGTGATATCACCAGCATCCTGGGCCAGATGGGCCACCTAGAAAGCCTGAC : 1263
bm65c22mx4 : TAAATGCAAATGCACACTGAGATTTCTTGATCTCCAATGTGTTAATCAGTTGTCAGTTGATAGAGGCTCACTGAGTGATATCACCAGCATCCTGTGCCAAATGGGCCACCTAGAAAGCCTGAG : 1263
bm65c22mx2 : TAAATGCAAATGCACCCTAAAATTTCTTGATCTCCAATGTGTTGATCAGTTGTCAGTTGATAGAGGCTCACTGAGTGATATCACCAGCATCCTGGGACAGATGGGCCACCTAGAAAGCCTGAC :  979
bm460b8mx3 : TGAATGCAGACATGCATTAAGCCATCTGGATCTGAAATGTGTTGATCACCTTGCAGTTGATGAGTCTCCTCTTACTGAAGTCACCAAACTTTTATCTCATACAATACAGCTGGACAGTCTTAG : 1001
bm65c22mx1 : TGAATGCAGACATGCATTAAGCCATCTGGATCTGAAATGTGTTGATCACCTTGCAGTTGATGAGTCTCCTCTTACTGAAGTCACCAAACTTTTATCTCATACAATACAGCTGGACAGTCTTAG :  998
bm65c22mx3 : TGAATGCAGACATGCATTAAGCCATCTGGATCTGAAATGTGTTGATCACCTTGCAGTTGATGAGTCTCCTCTTACTGAAGTTACCAAACTTTTATCTCATACAATACAGCTGGACAGTCTTAG :  998
bm460b8mx2 : TGACTATAACGATACCTTGAAACTCCTGAATCTGTTATGTACTGATCACCTGGCAGTTGATAAGGCTTCCCTGAATGATATCAACACCCTTTTGTCTCAGATGGTCCACTTGGATAGTCTTAG :  751
bm250f8mx2 : TTATGCCAAAAACGCTCTGAAGTTCCTCGATCTAACTTGCATTCAGAACCTGACAGTTGATCAGGCTTCACTGAGTGAAGTCACCACTCTTCTGGCTCGCATGATCTATCTGGACAGCCTGAG : 1473
```

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

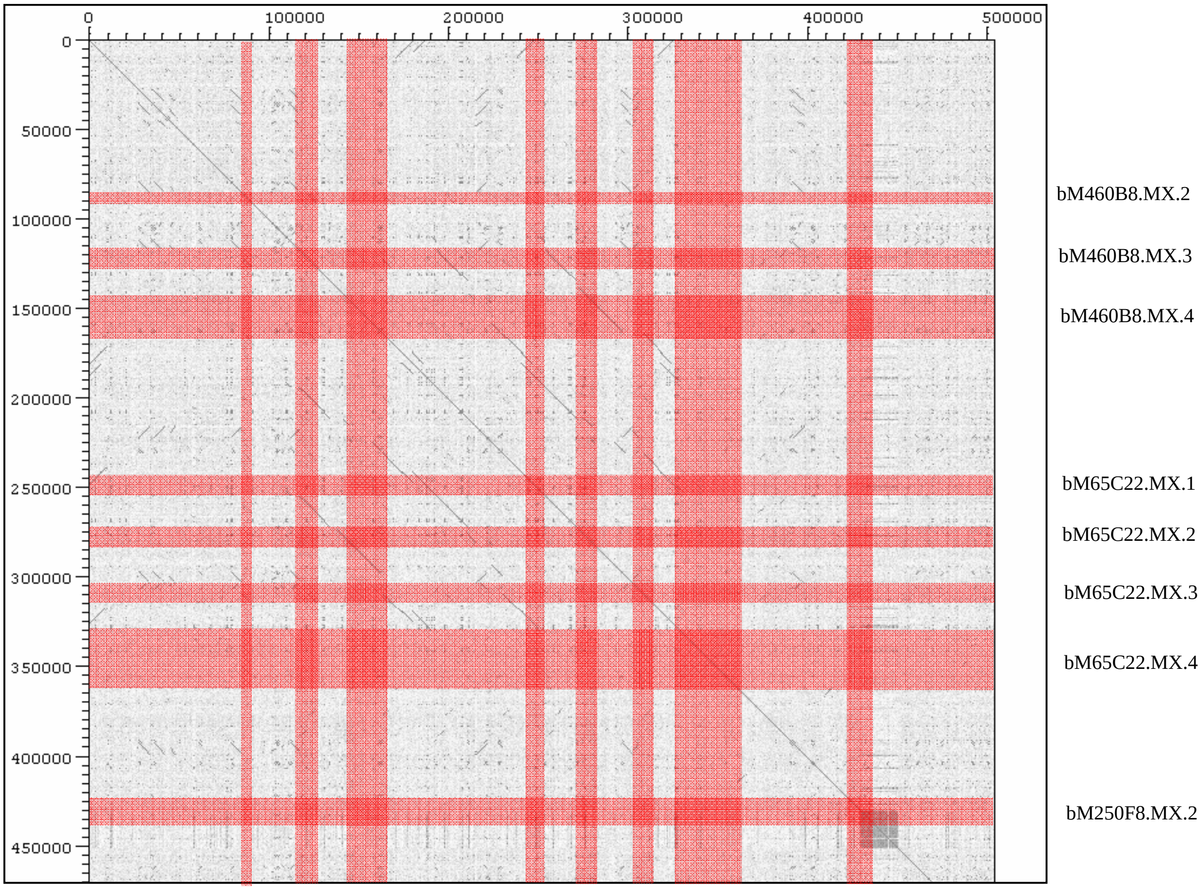


Figure 4-16 Diagram illustrating Dotter analysis of the *Mus musculus* X E3-F2 region containing Pramel3L repeats. Approximate positions of annotated loci (illustrated in Figure 4-15) are shown by red bars located on both axes. The sequence analysed comprised of linked sequences of clones bM460B8, (gap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

4.4.4 The mouse Nxf2 locus

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

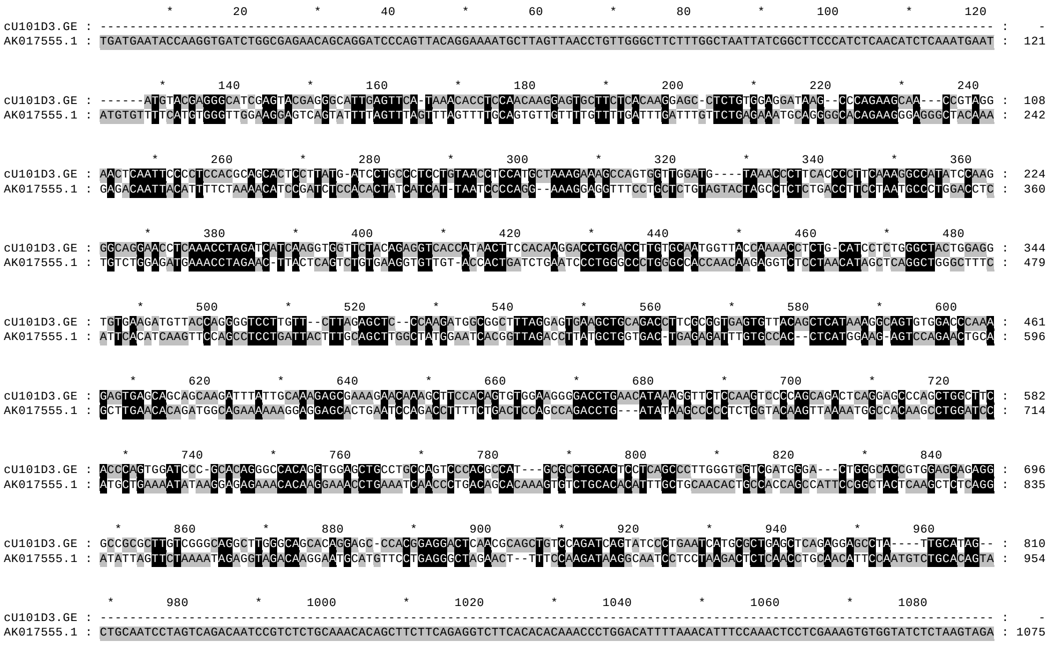
Unlike the human situation however, a Pramel3L gene (named for its similarity to the mouse Prame-like 3 gene) was annotated between the Nxf2 and Tcp11-like loci that appears functional, from the identity to the mRNA sequence used to annotate the gene. As discussed earlier, the human PRAMEL3L loci in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

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

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

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

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



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

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

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

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

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

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

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

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

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

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

4.5 Discussion

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

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

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

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

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.

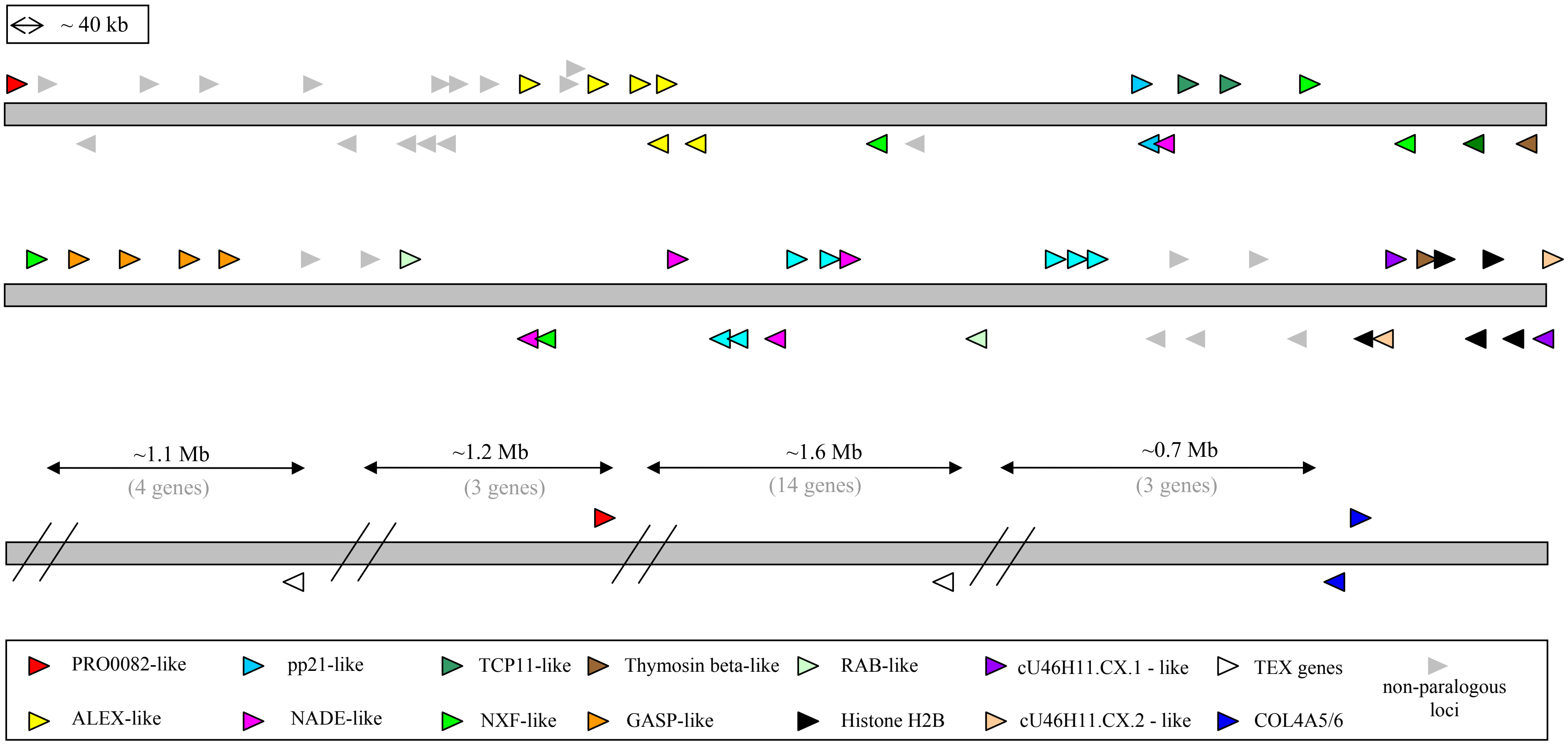


Figure 5-1 Diagram illustrating positions of duplicated loci within Xq22. The first two solid bars depict the ~ 3.2 Mb region bounded by genomic clones dJ341D10 (Z97985) and cU46H11 (Z82254). Genes and directions of transcription are denoted by the arrow direction. Paralogue families and non-duplicated 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.

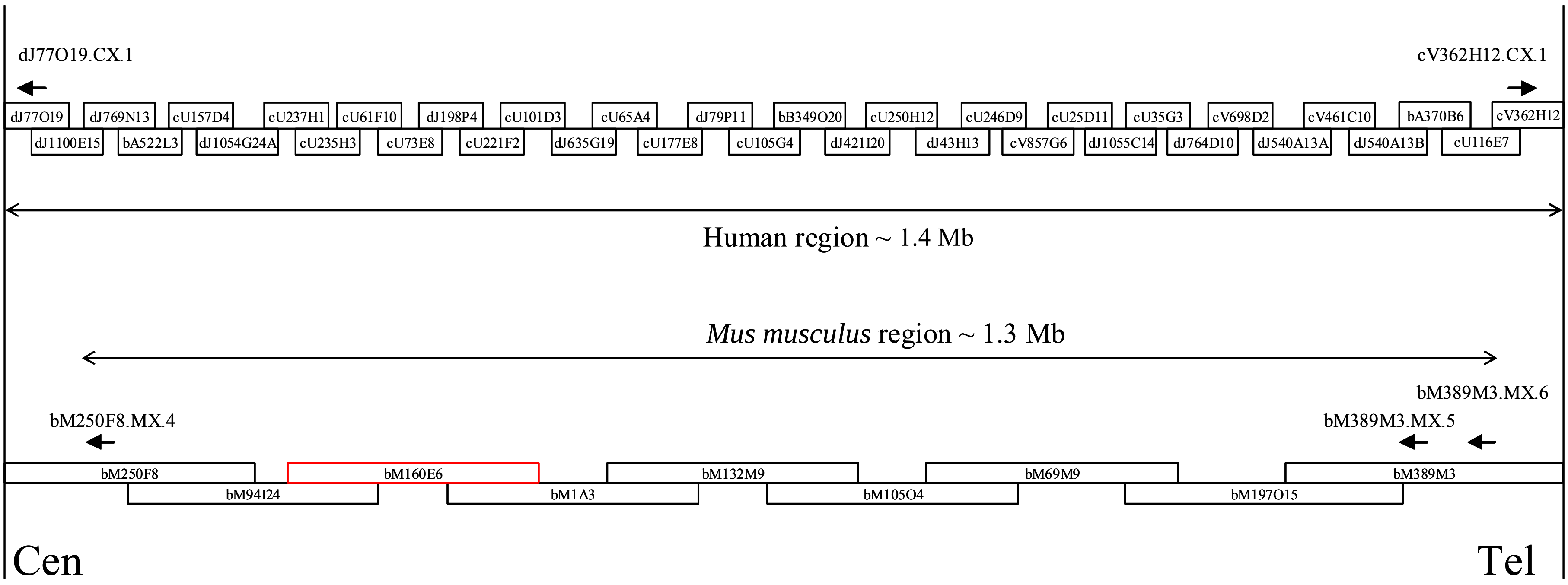


Figure 5-2 Figure showing a schematic representation (not to scale) of thymosin-beta paralogue gene order and orientation within human Xq22, and the corresponding orthologous region in *Mus musculus* (mouse). Large double-headed arrows depict approximate sizes of the regions, small arrows depict gene 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.

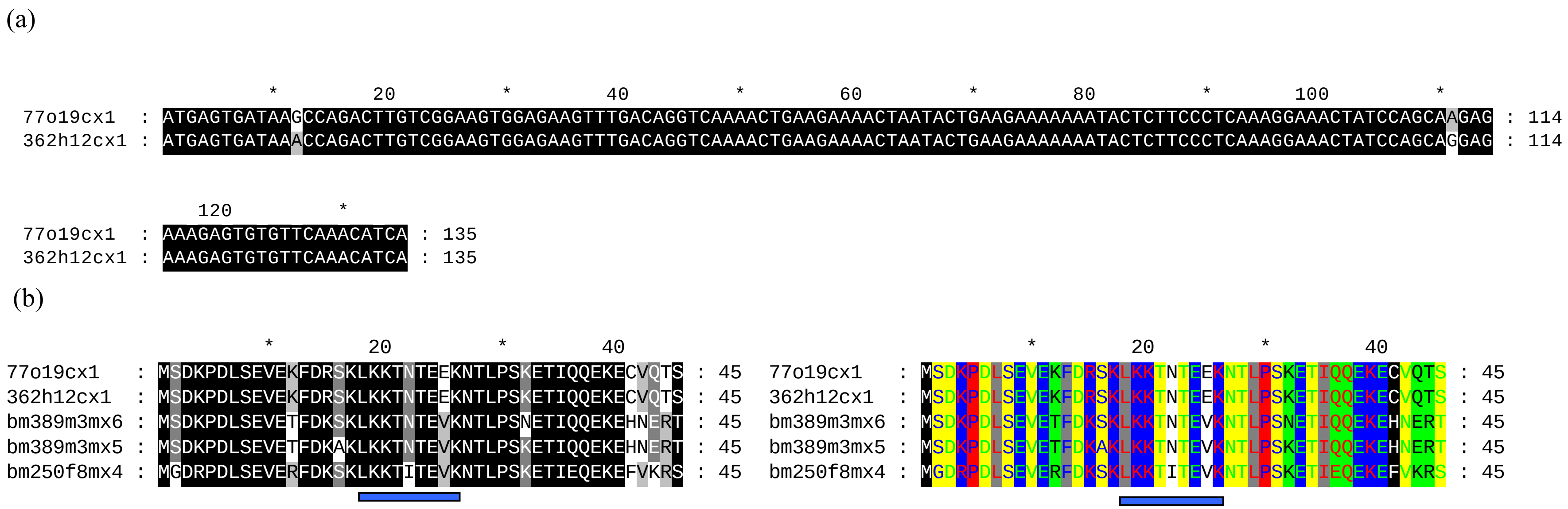

(a)
* 20 * 40 * 60 * 80 * 100 *
77o19cx1 : ATGAGTGATAAGCCAGACTTGTCGGAAGTGGAGAAGTTTGACAGGTCAAAACTGAAGAAAACTAATACTGAAGAAAAAAATACTCTTCCCTCAAAGGAAACTATCCAGCAAGAG : 114
362h12cx1 : ATGAGTGATAAACCAGACTTGTCGGAAGTGGAGAAGTTTGACAGGTCAAAACTGAAGAAAACTAATACTGAAGAAAAAAATACTCTTCCCTCAAAGGAAACTATCCAGCAGGAG : 114
120 *
77o19cx1 : AAAGAGTGTGTTCAAACATCA : 135
362h12cx1 : AAAGAGTGTGTTCAAACATCA : 135
(b)
* 20 * 40
77o19cx1 : MSDKPDLSEVEKFDRSKLKKTNTEEKNTLPSKETIQQEKECVQTS : 45
362h12cx1 : MSDKPDLSEVEKFDRSKLKKTNTEEKNTLPSKETIQQEKECVQTS : 45
bm389m3mx6 : MSDKPDLSEVETFDKSKLKKTNTEVKNTLPSNETIQQEKEHNERT : 45
bm389m3mx5 : MSDKPDLSEVETFDKAKLKKTNTEVKNTLPSKETIQQEKEHNERT : 45
bm250f8mx4 : MGDRPDLSEVERFDKSKLKKTITEVKNTLPSKETIEQEKEFVKRS : 45
* 20 * 40
77o19cx1 : MSDKPDLSEVEKFDRSKLKKTNTEEKNTLPSKETIQQEKECVQTS : 45
362h12cx1 : MSDKPDLSEVEKFDRSKLKKTNTEEKNTLPSKETIQQEKECVQTS : 45
bm389m3mx6 : MSDKPDLSEVETFDKSKLKKTNTEVKNTLPSNETIQQEKEHNERT : 45
bm389m3mx5 : MSDKPDLSEVETFDKAKLKKTNTEVKNTLPSKETIQQEKEHNERT : 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.

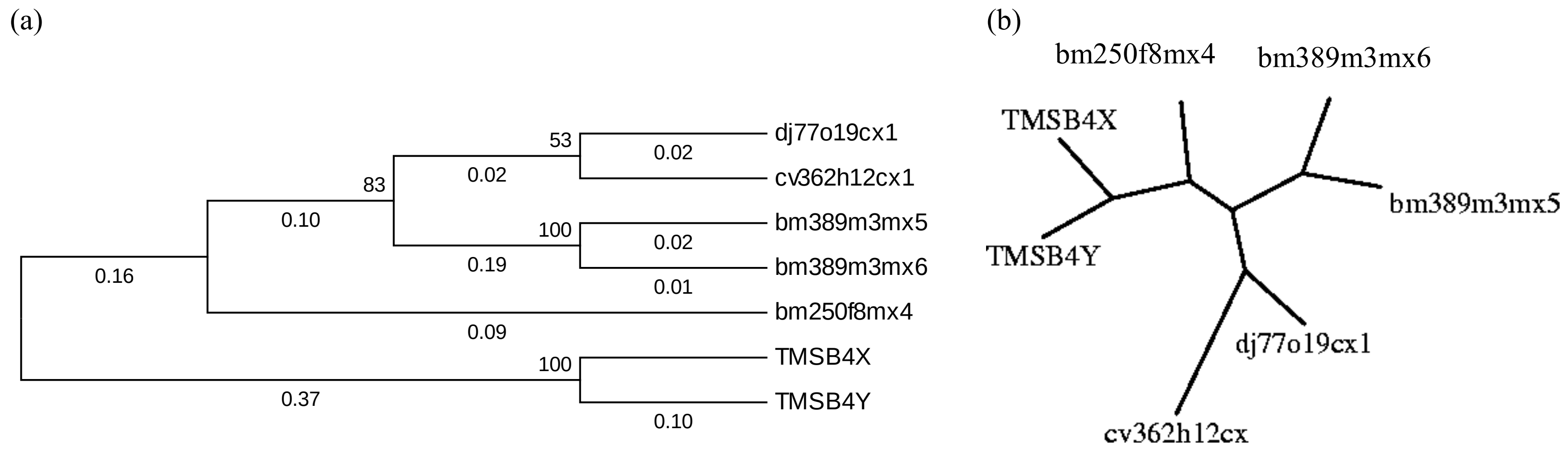
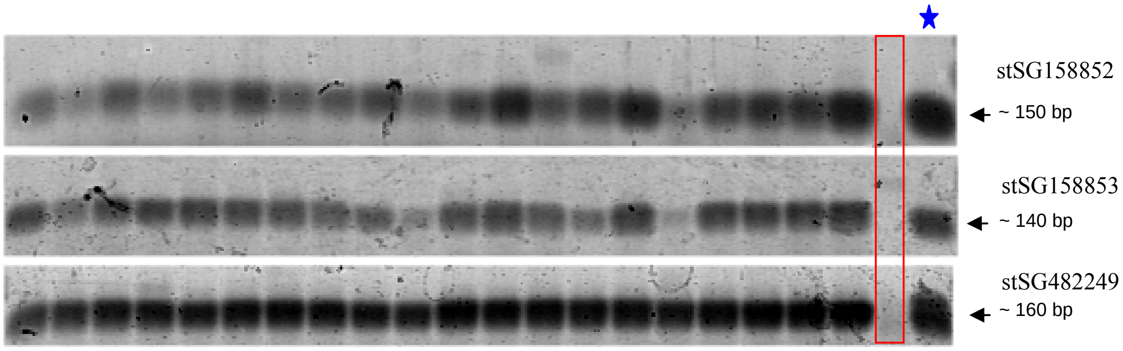


Figure 5-4 Phylogenetic analysis of human and mouse thymosin-beta genes. (a) distance-based cladogram computed from human and mouse gene coding nucleotide sequence. Numbers adjacent to nodes represent the number of bootstrap replicates supporting the adjacent node. Numbers below branches indicate distance. 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.

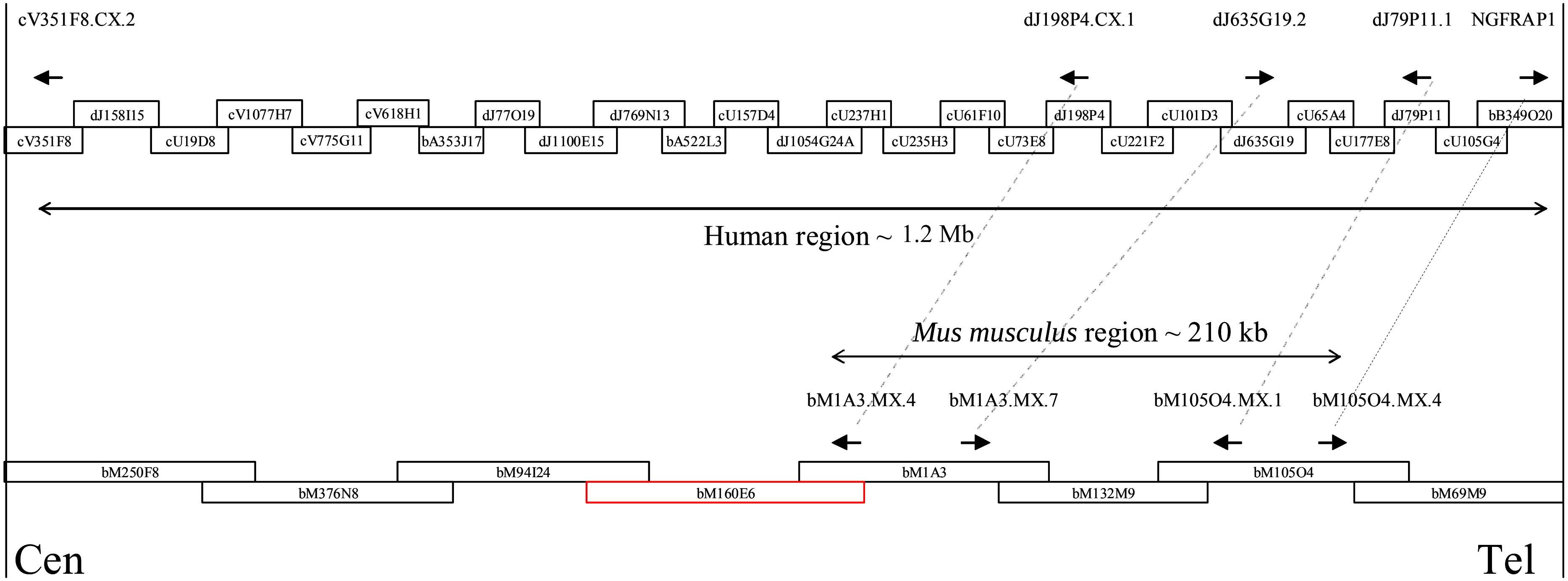
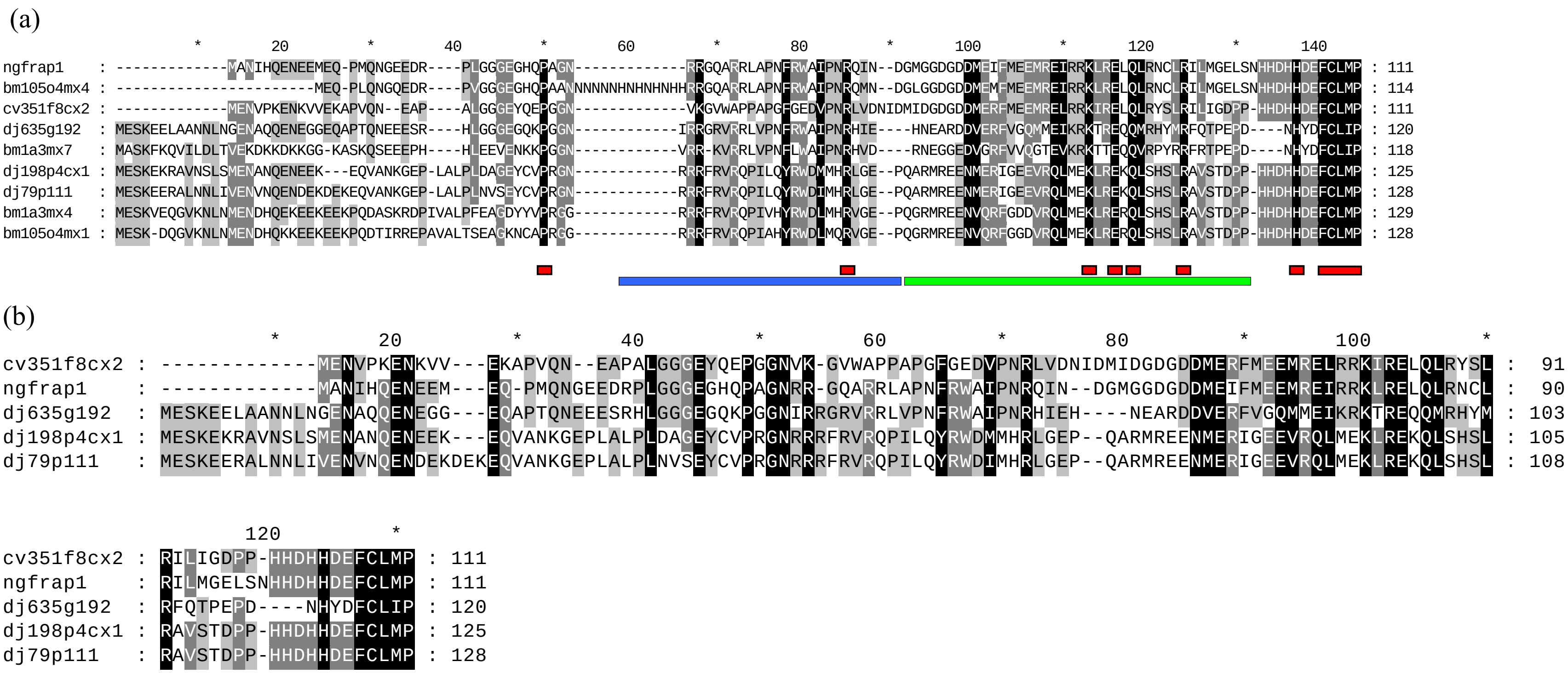


Figure 5-6 A schematic representation (not to scale) of NADE family gene order and orientation within human Xq22 and the corresponding orthologous region in mouse. Large double-headed arrows depict approximate sizes of the regions, small arrows depict gene positions (names above) and direction of transcription. 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.

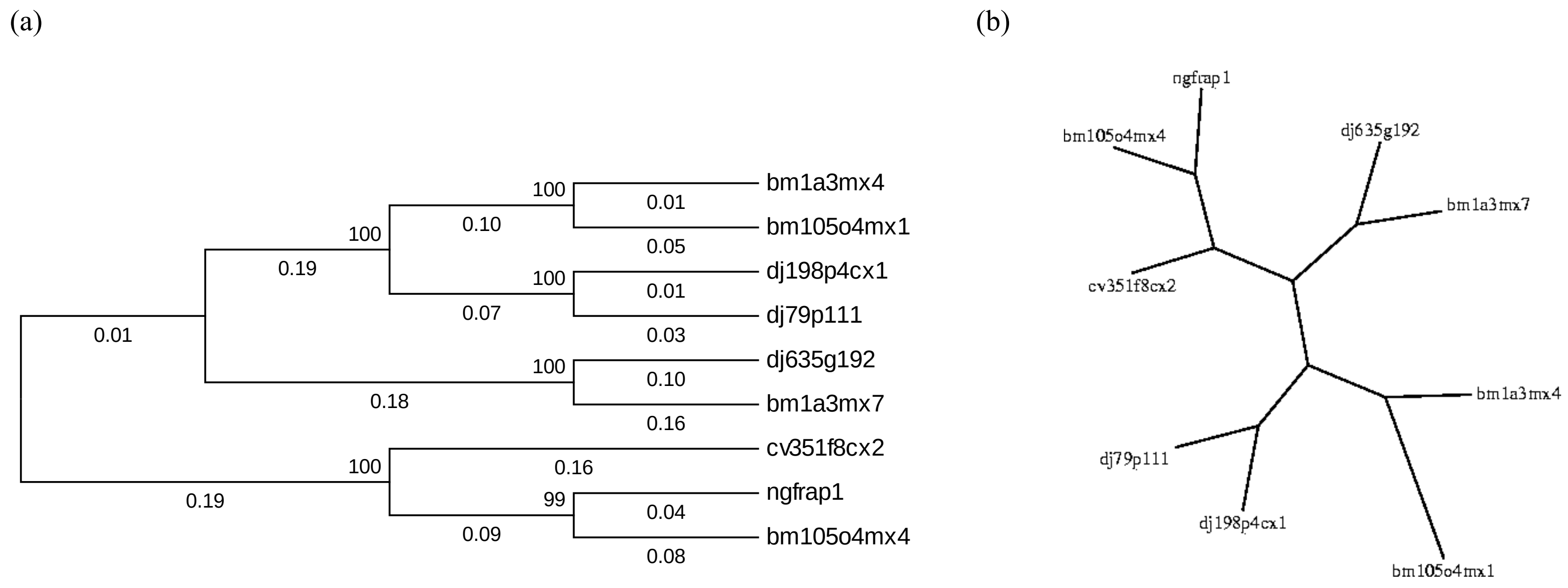
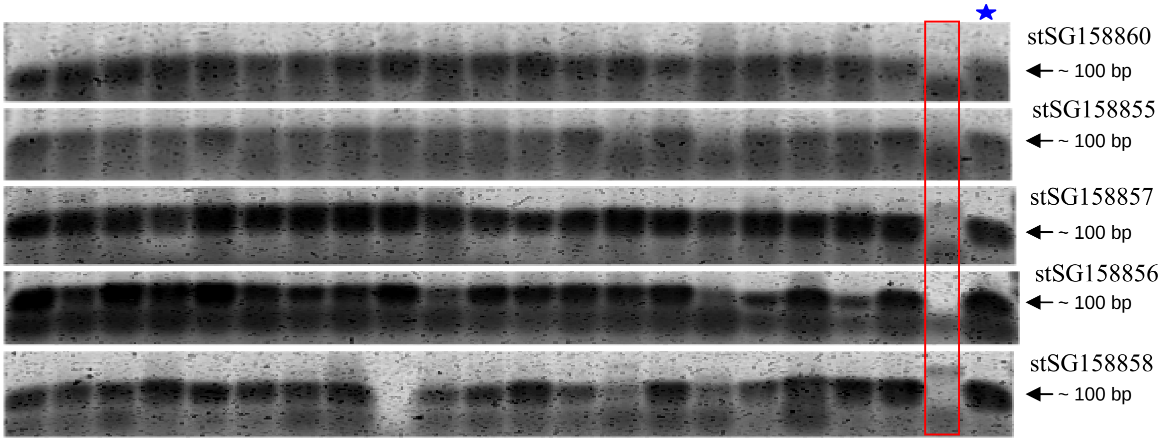


Figure 5-8 Phylogenetic analysis of NADE Xq22 paralogues. (a) a distance-based cladogram constructed using an alignment of human and mouse NADE-like genes coding nucleotide sequences. Numbers adjacent to nodes represent the number of bootstrap replicates supporting the adjacent node. Numbers below branches 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.

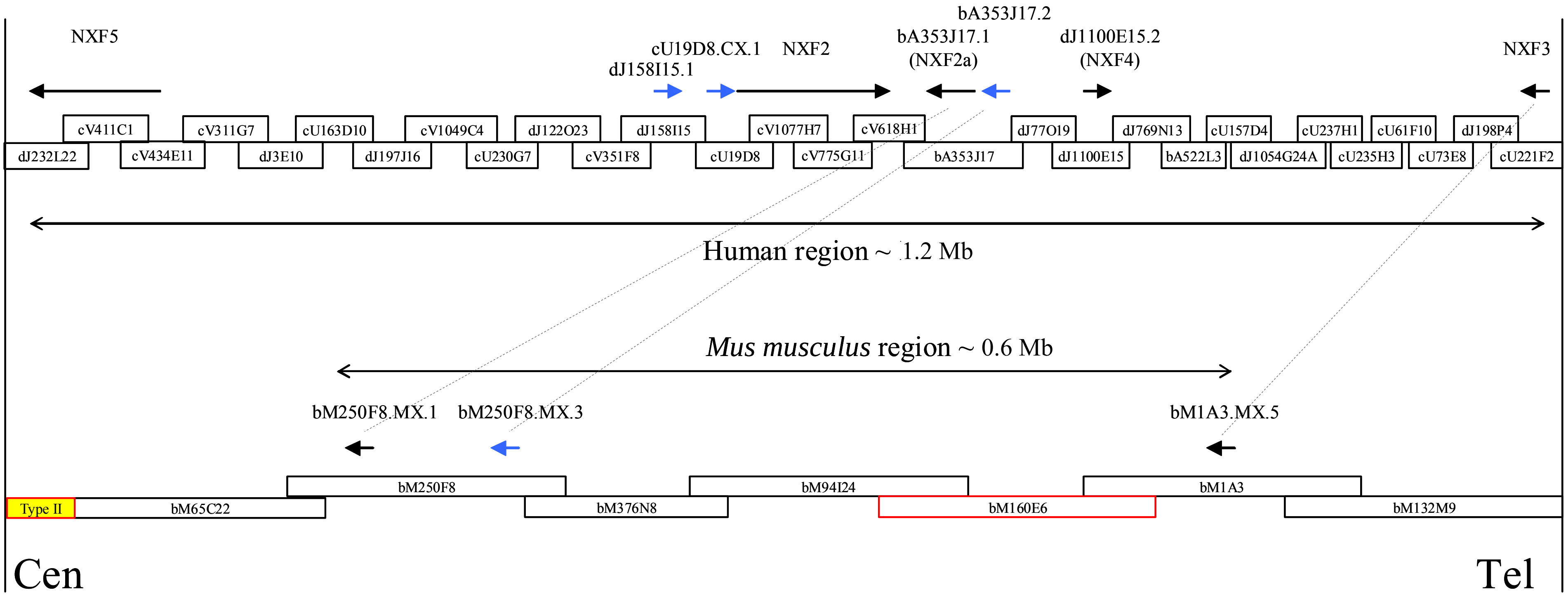


Figure 5-10 Figure showing a schematic representation (not to scale) of NXF and TCP11-like paralogue gene order and orientation within human Xq22, and the corresponding orthologous region in mouse. Gene names reflect locus names in Chapter 3 and 4, with alternative names given in parentheses. Large double-headed 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).

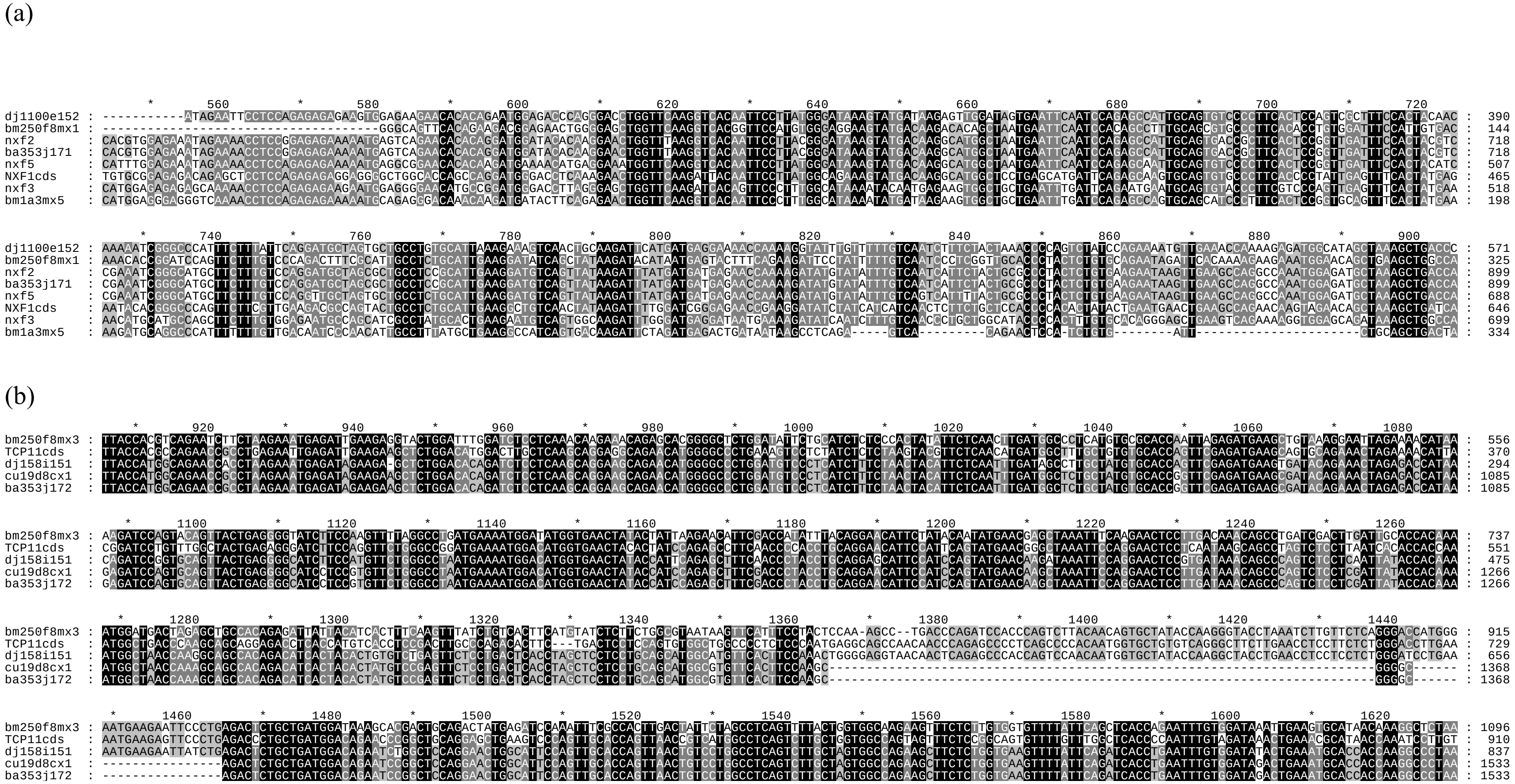


Figure 5-11 Alignments of NXF family genes and TCP11-like genes. (a) depicts part of an alignment of NXF-family nucleotide sequences to illustrate the level of homology seen. (b) depicts part of an alignment of TCP11-like nucleotide sequences. For technical reasons, gene names have had "." separators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

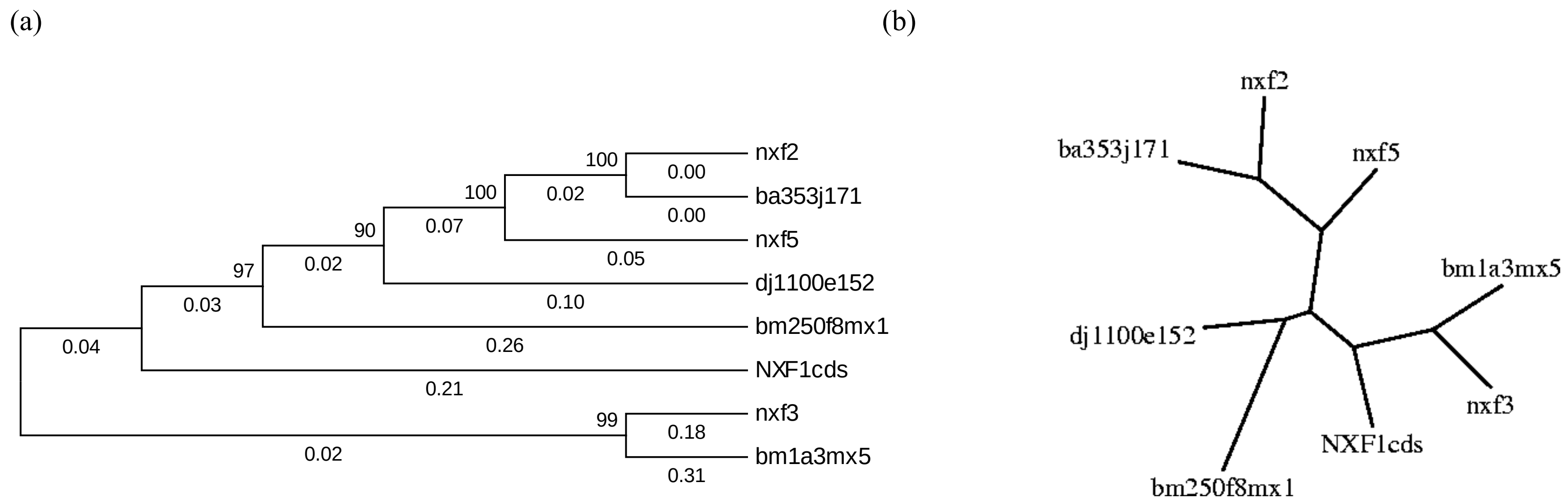


Figure 5-12 Figure illustrating phylogenetic analysis of human and mouse NXF genes. (a) a distance-based cladogram constructed using an alignment of human and mouse NXF-like genes coding nucleotide sequence. Numbers adjacent to nodes represent the number of bootstrap replicates supporting the adjacent 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.

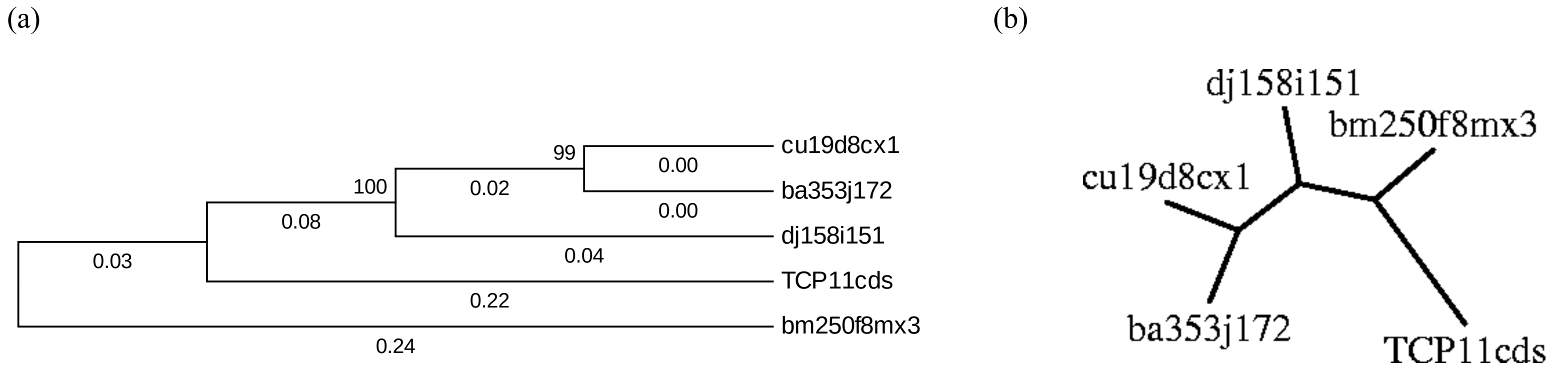
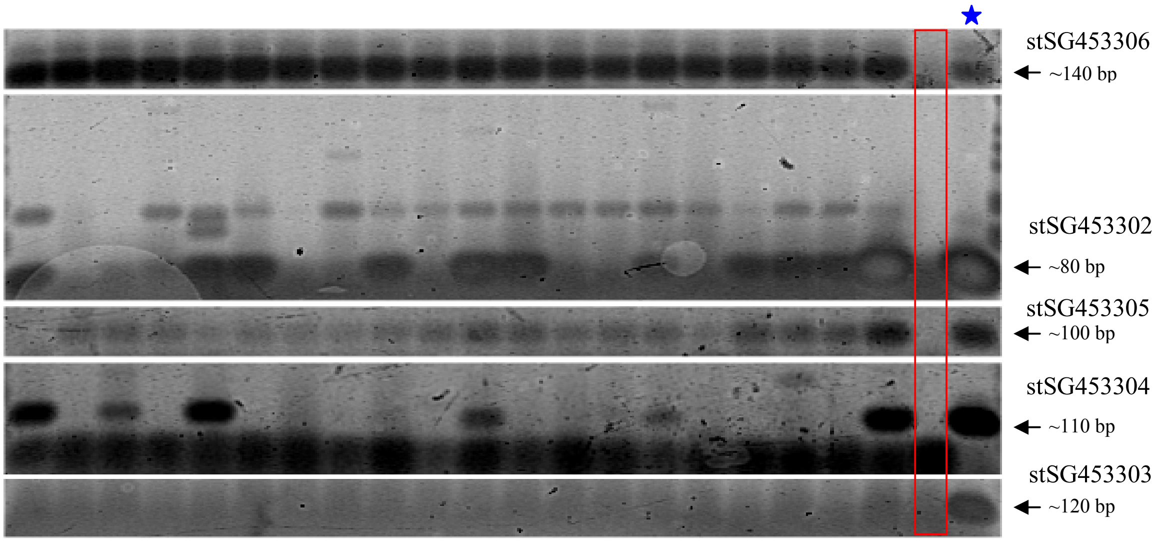


Figure 5-13 Figure illustrating phylogenetic analysis of human and mouse TCP11-like genes (a) a distance-based cladogram constructed using an alignment of human and mouse TCP-like genes coding nucleotide sequence. Numbers adjacent to nodes represent the number of bootstrap replicates supporting the adjacent 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.

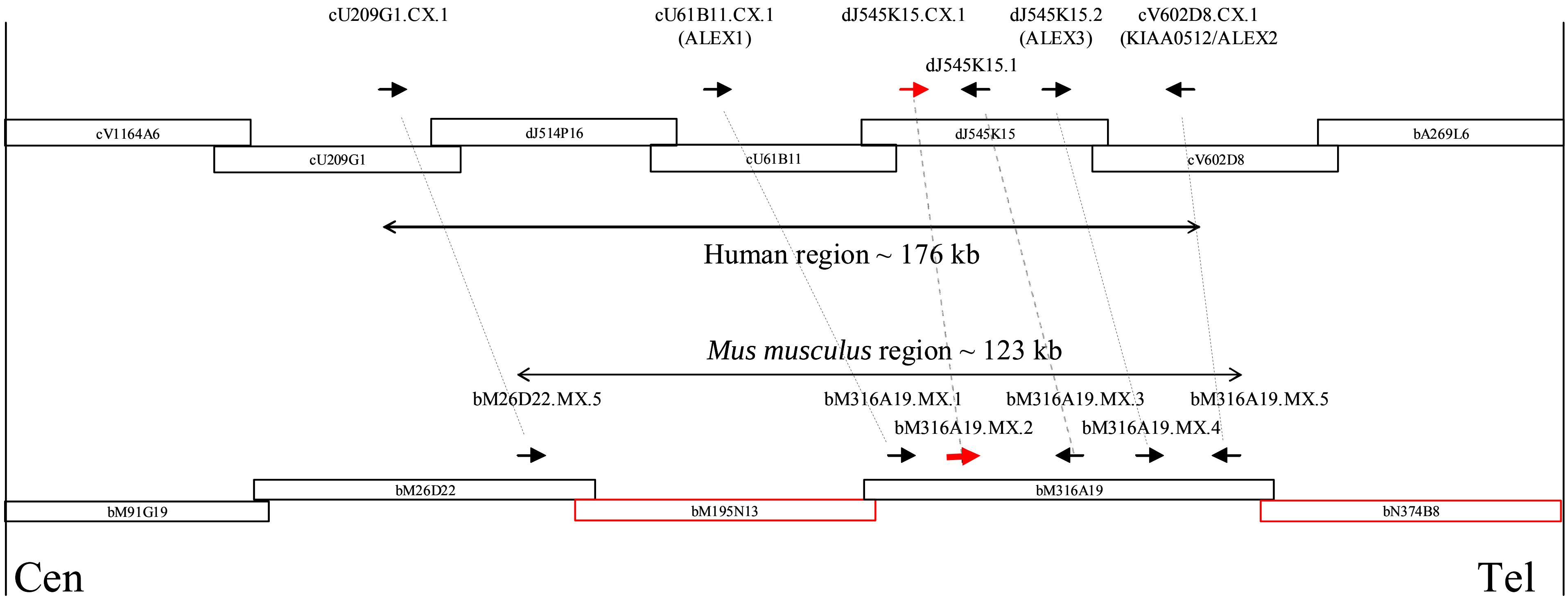
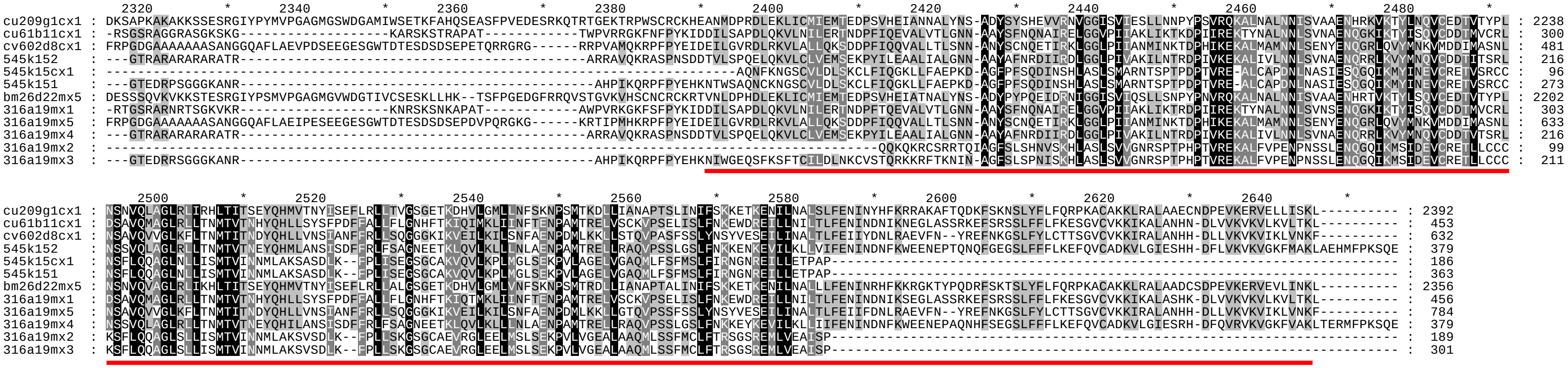


Figure 5-15 Figure showing a schematic representation (Not to scale) of ALEX paralogue gene order and orientation within human Xq22, and the corresponding orthologous region in mouse. Large double-headed arrows depict approximate sizes of the regions, small arrows depict gene positions (names above) 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.

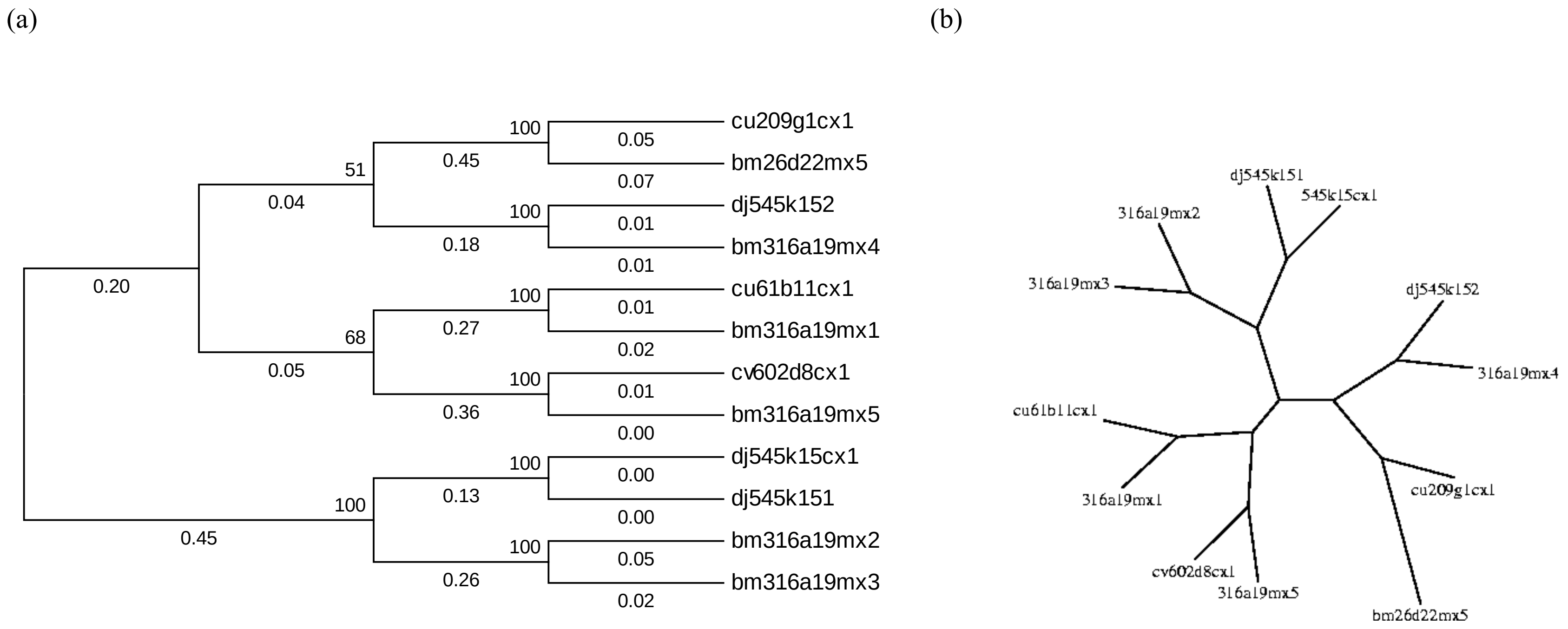
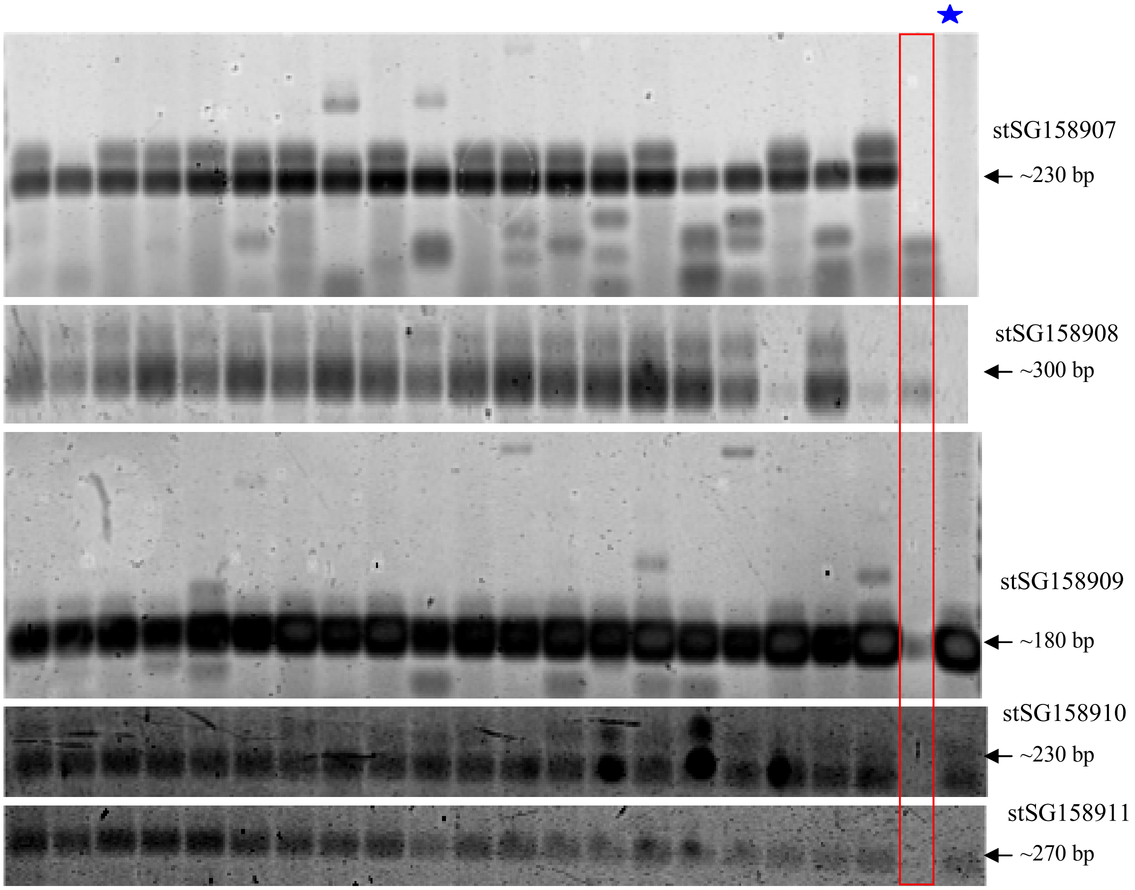


Figure 5-17 Phylogenetic analysis of human and mouse ALEX-like genes (a) a distance-based cladogram computed using an alignment of human and mouse ALEX-like genes predicted peptide sequences. Numbers adjacent to nodes represent the number of bootstrap replicates supporting the adjacent node. Numbers 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.

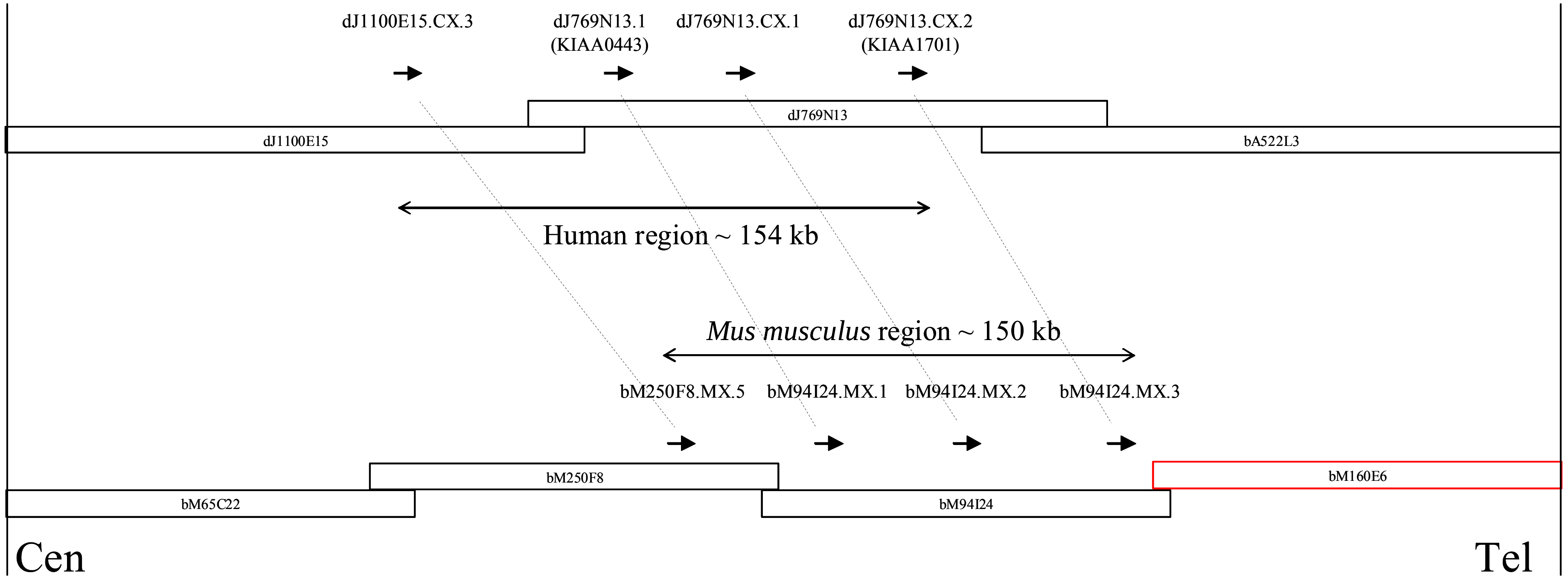
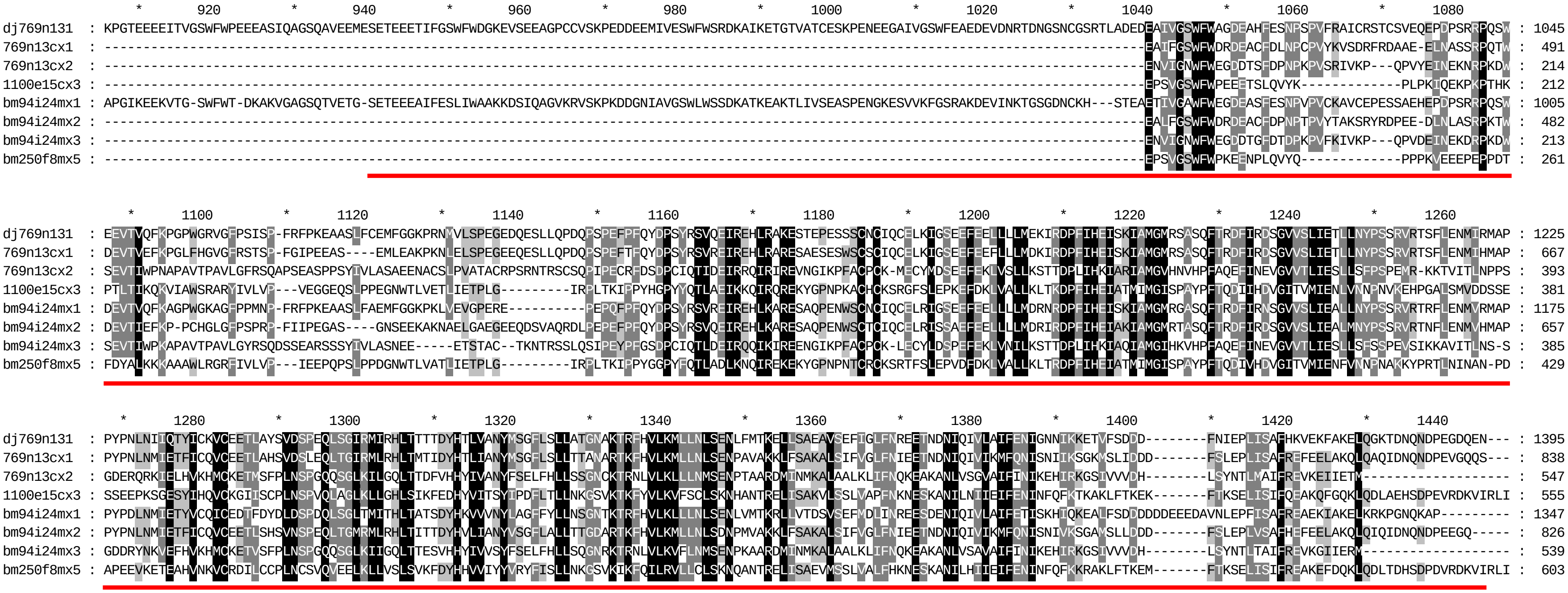


Figure 5-19 Figure showing a schematic representation (not to scale) of GASP paralogue gene order and orientation within human Xq22, and the corresponding orthologous region in mouse. Large double-headed arrows depict approximate sizes of the regions, small arrows depict gene positions (names above) 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).

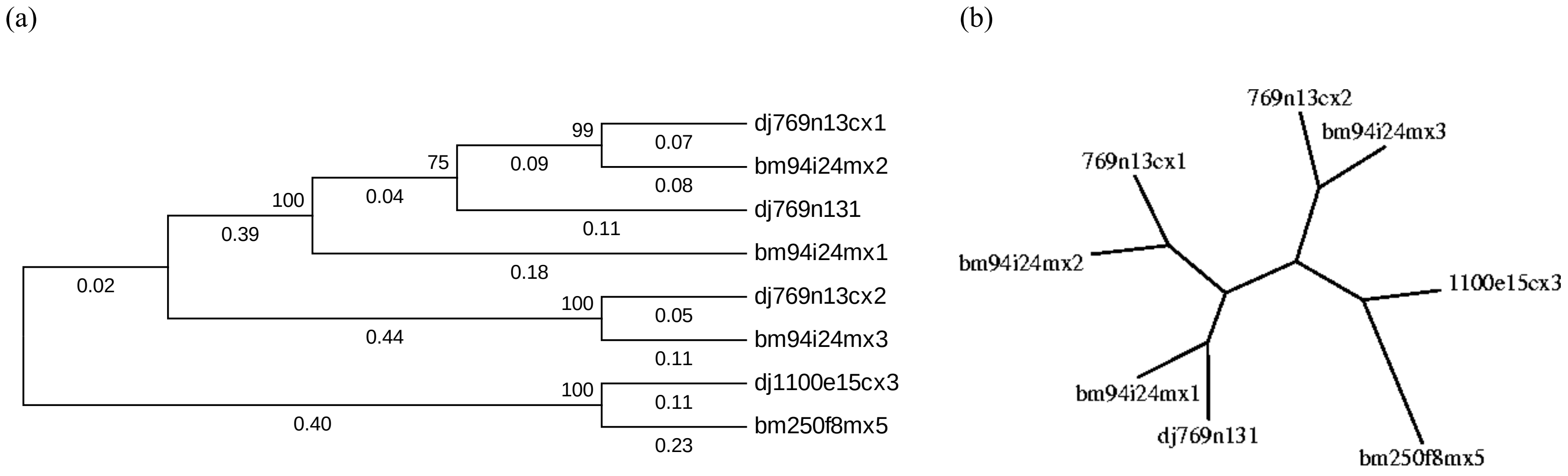
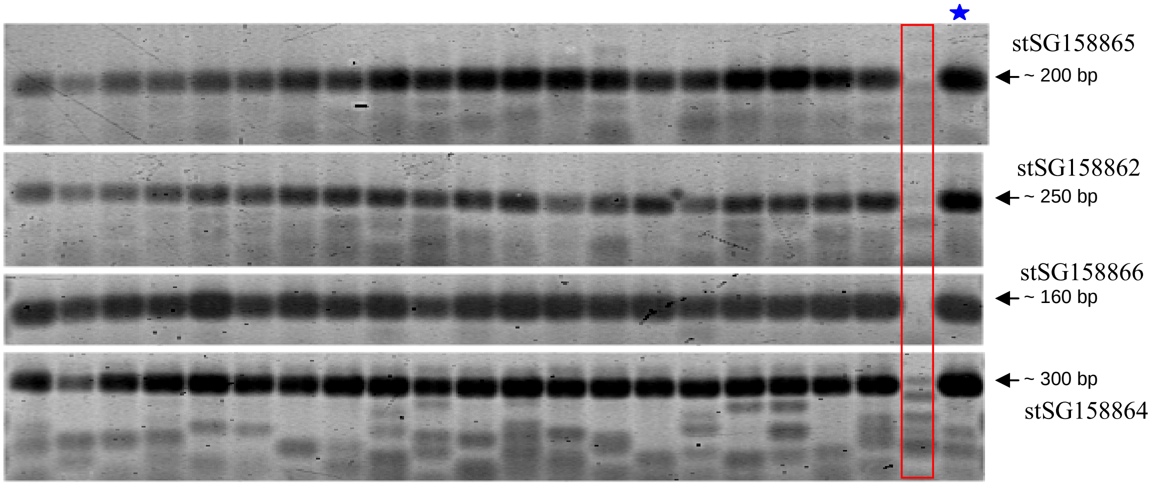


Figure 5-21 Figure illustrating phylogenetic analysis of GASP-like genes. (a) a distance-based cladogram computed using an alignment of human and mouse GASP-like genes predicted peptide sequences. Numbers adjacent to nodes represent the number of bootstrap replicates supporting the adjacent node. Numbers 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.

(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
■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	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.

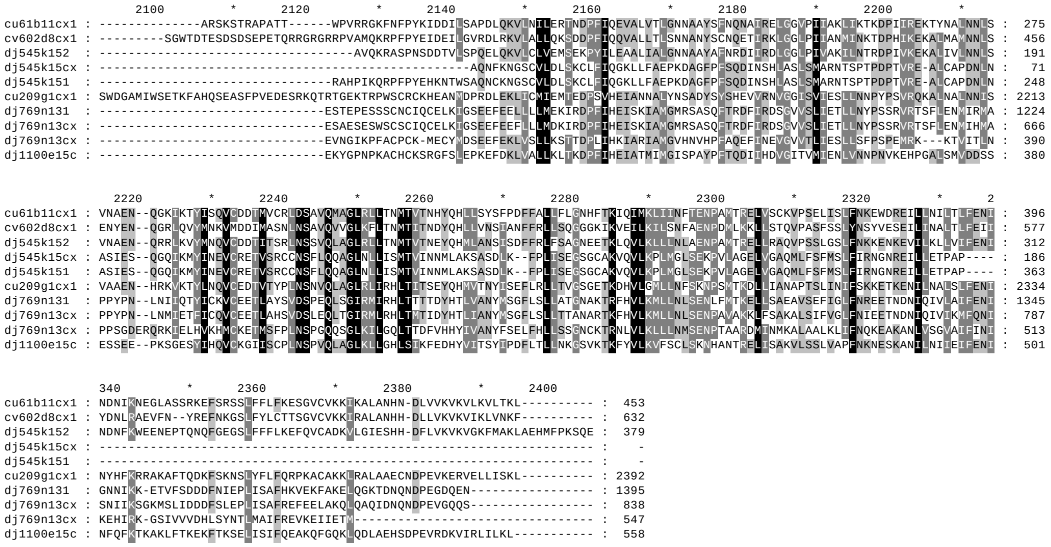


Figure 5-23 Alignment of GASP and ALEX family predicted peptides showing conservation between the sequences. For clarity, only the C-terminal regions are displayed.

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.

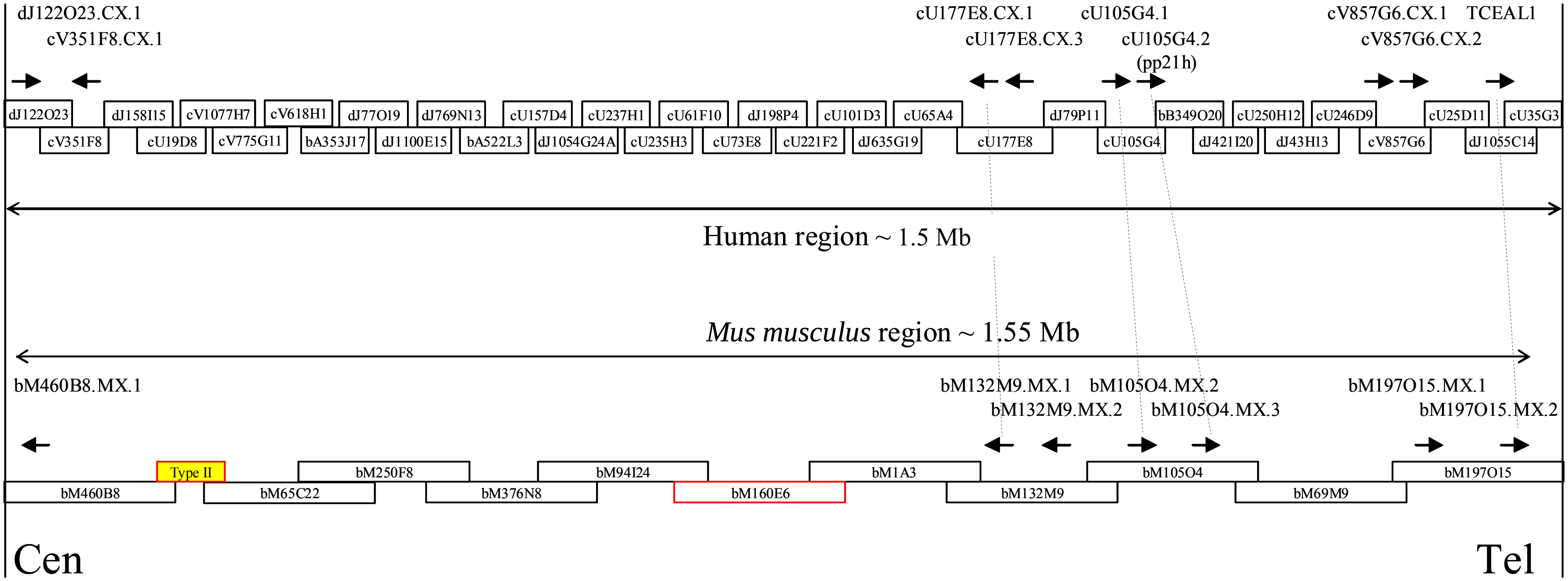


Figure 5-24 Figure showing a schematic representation (not to scale) of pp21/TCEAL1 paralogue gene order and orientation within human Xq22, and the corresponding orthologous region in mouse. Large double-headed arrows depict approximate sizes of the regions, small arrows depict gene positions (names 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.

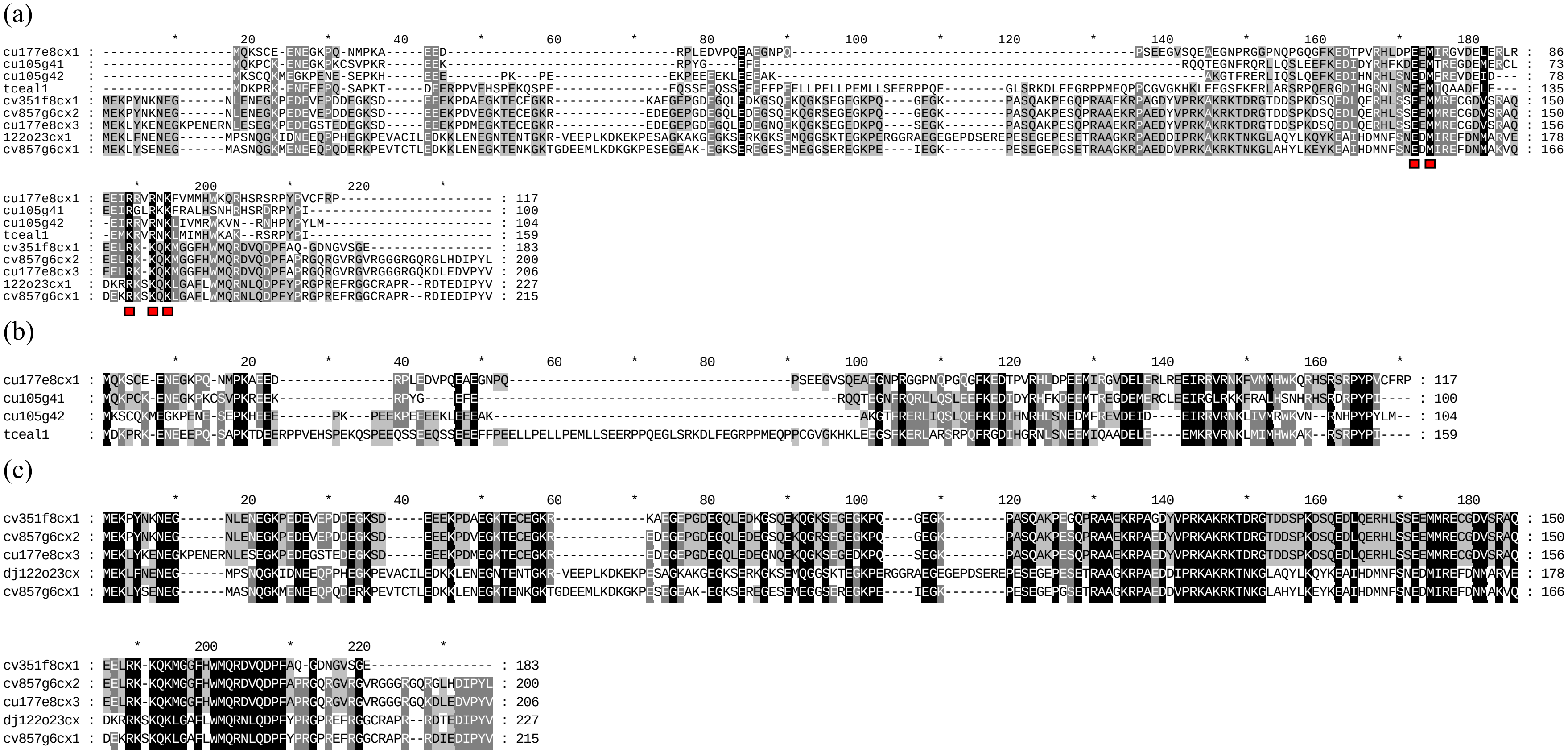


Figure 5-25 Figure illustrating alignments of pp21/TCEAL human paralogue predicted peptides. For technical reasons, gene names have had "." separators removed in the figures, and some genes have had their clone-based prefix ("cV", "dJ", "bM" etc.) removed. The alignment in (a) illustrates varied homology 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.

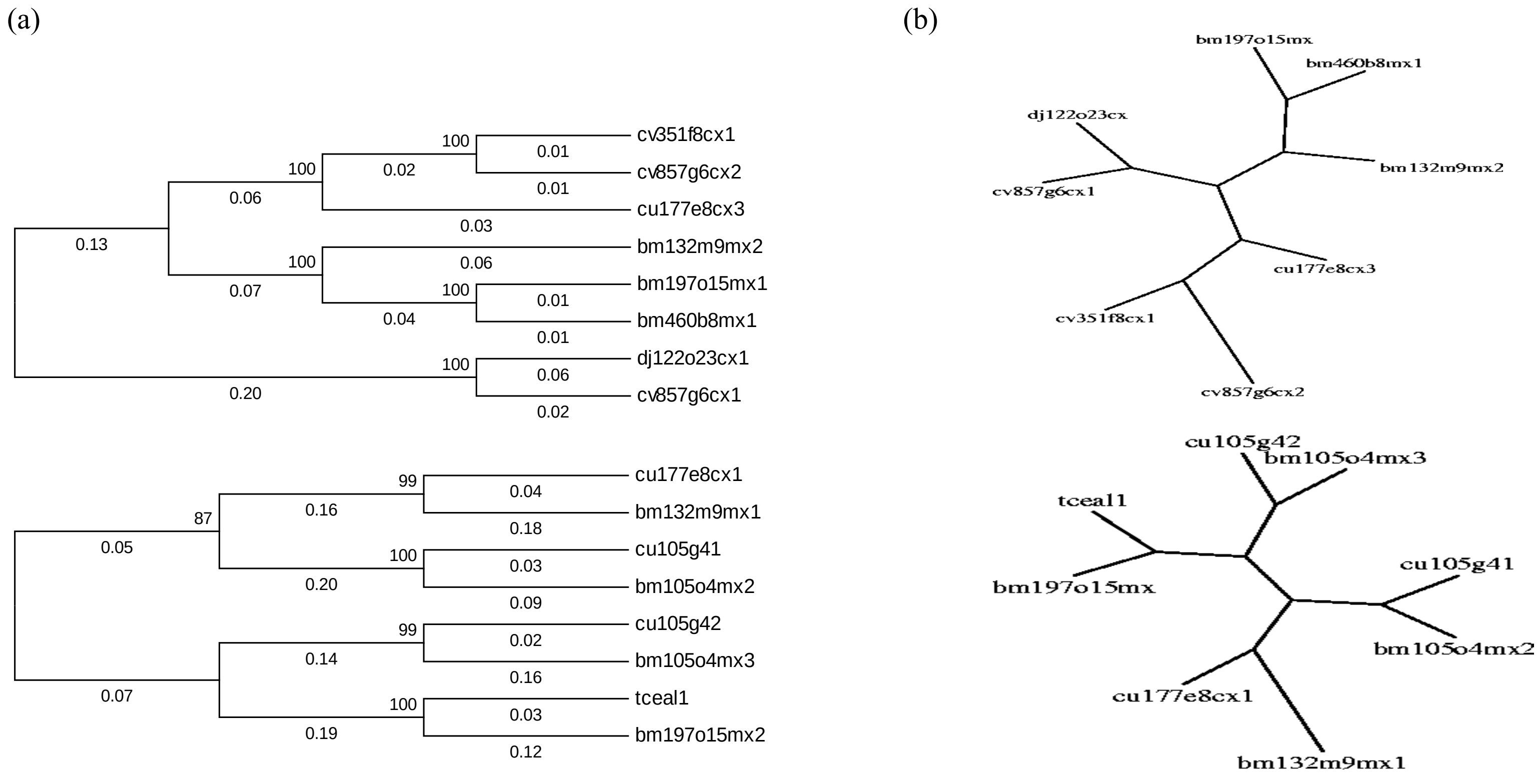
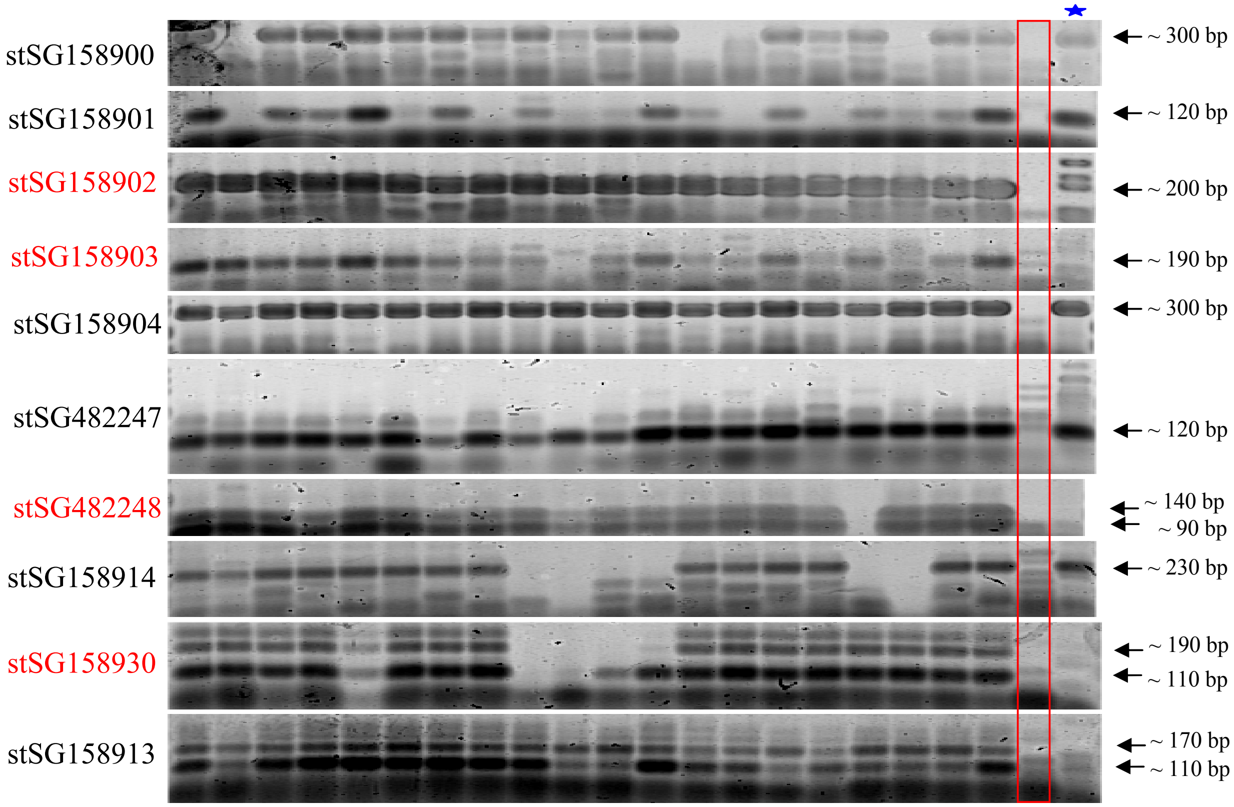


Figure 5-26 Figure illustrating phylogenetic analysis of pp21/TCEAL genes (a) distance-based cladograms constructed using alignments of human and mouse pp21-like genes coding nucleotide sequence. Numbers adjacent to nodes represent the number of bootstrap replicates supporting the adjacent node. Numbers 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.

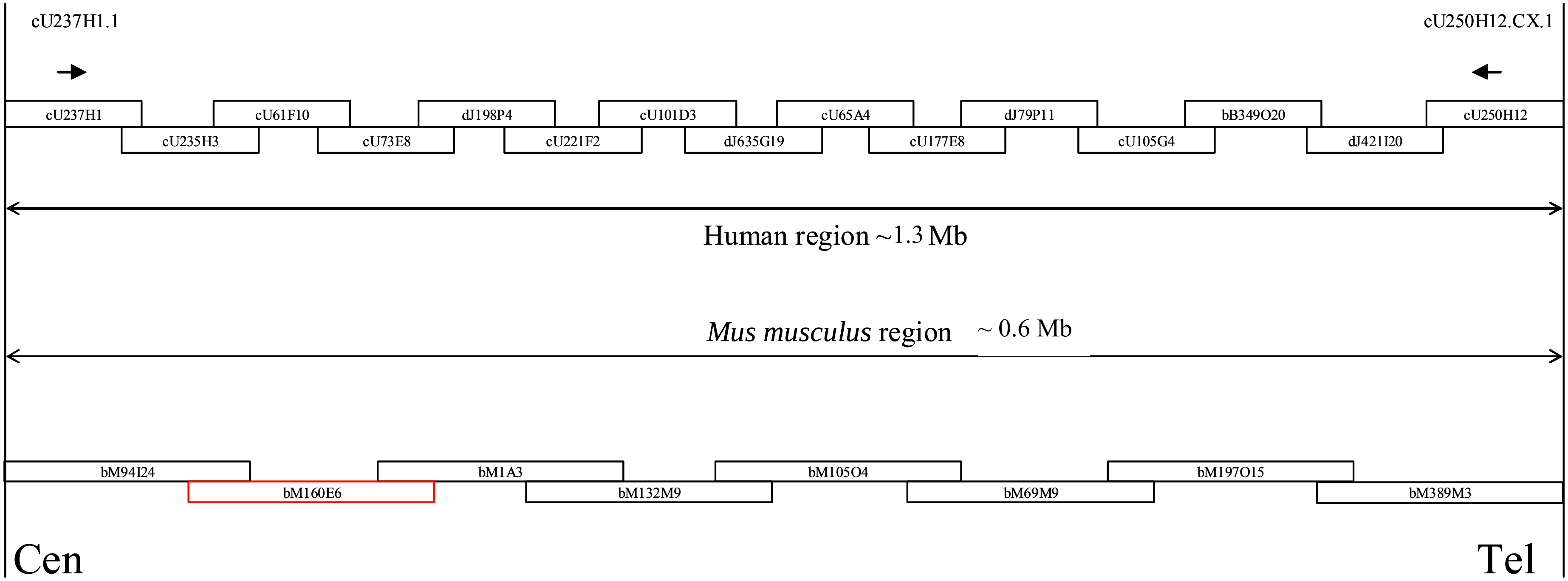
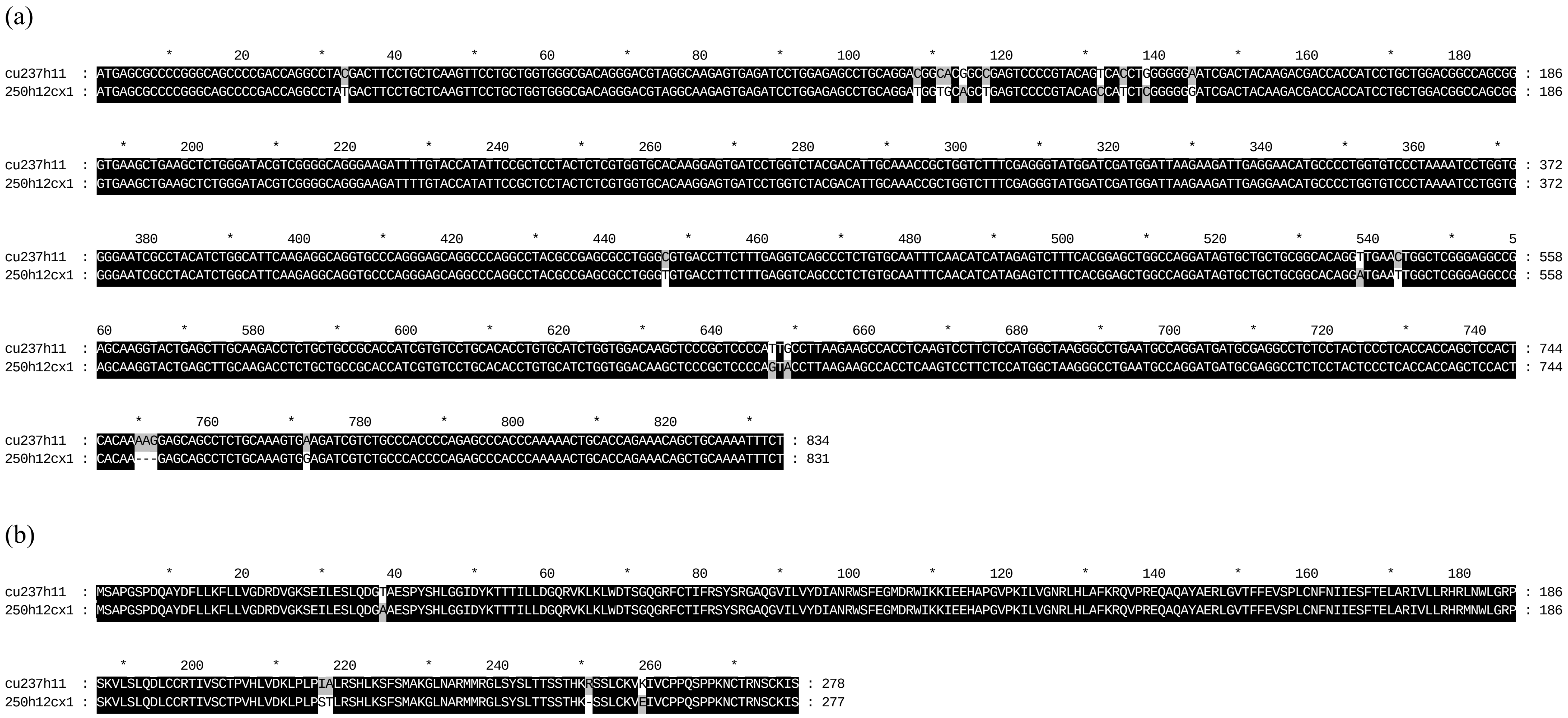


Figure 5-28 Figure showing a schematic representation (not to scale) of RAB-like paralogue gene order and orientation within human Xq22, and the corresponding orthologous region in mouse. Large double-headed arrows depict approximate sizes of the regions, small arrows depict gene positions (names above) 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)

```
                     *        20         *        40         *        60         *        80         *       100         *       120         *       140         *       160         *       180
cu237h11  : ATGAGCGCCCCGGGCAGCCCCGACCAGGCCTACGACTTCCTGCTCAAGTTCCTGCTGGTGGGCGACAGGGACGTAGGCAAGAGTGAGATCCTGGAGAGCCTGCAGGACGGCACGGCCGAGTCCCCGTACAGTCACCTGGGGGAATCGACTACAAGACGACCACCATCCTGCTGGACGGCCAGCGG : 186
250h12cx1 : ATGAGCGCCCCGGGCAGCCCCGACCAGGCCTATGACTTCCTGCTCAAGTTCCTGCTGGTGGGCGACAGGGACGTAGGCAAGAGTGAGATCCTGGAGAGCCTGCAGGATGGTGCAGCTGAGTCCCCGTACAGCCATCTCGGGGGGATCGACTACAAGACGACCACCATCCTGCTGGACGGCCAGCGG : 186

               *       200         *       220         *       240         *       260         *       280         *       300         *       320         *       340         *       360         *
cu237h11  : GTGAAGCTGAAGCTCTGGGATACGTCGGGGCAGGGAAGATTTTGTACCATATTCCGCTCCTACTCTCGTGGTGCACAAGGAGTGATCCTGGTCTACGACATTGCAAACCGCTGGTCTTTCGAGGGTATGGATCGATGGATTAAGAAGATTGAGGAACATGCCCCTGGTGTCCCTAAAATCCTGGTG : 372
250h12cx1 : GTGAAGCTGAAGCTCTGGGATACGTCGGGGCAGGGAAGATTTTGTACCATATTCCGCTCCTACTCTCGTGGTGCACAAGGAGTGATCCTGGTCTACGACATTGCAAACCGCTGGTCTTTCGAGGGTATGGATCGATGGATTAAGAAGATTGAGGAACATGCCCCTGGTGTCCCTAAAATCCTGGTG : 372

                 380         *       400         *       420         *       440         *       460         *       480         *       500         *       520         *       540         *       5
cu237h11  : GGGAATCGCCTACATCTGGCATTCAAGAGGCAGGTGCCCAGGGAGCAGGCCCAGGCCTACGCCGAGCGCCTGGGCGTGACCTTCTTTGAGGTCAGCCCTCTGTGCAATTTCAACATCATAGAGTCTTTCACGGAGCTGGCCAGGATAGTGCTGCTGCGGCACAGGTTGAACTGGCTCGGGAGGCCG : 558
250h12cx1 : GGGAATCGCCTACATCTGGCATTCAAGAGGCAGGTGCCCAGGGAGCAGGCCCAGGCCTACGCCGAGCGCCTGGGTGTGACCTTCTTTGAGGTCAGCCCTCTGTGCAATTTCAACATCATAGAGTCTTTCACGGAGCTGGCCAGGATAGTGCTGCTGCGGCACAGGATGAATTGGCTCGGGAGGCCG : 558

            60         *       580         *       600         *       620         *       640         *       660         *       680         *       700         *       720         *       740
cu237h11  : AGCAAGGTACTGAGCTTGCAAGACCTCTGCTGCCGCACCATCGTGTCCTGCACACCTGTGCATCTGGTGGACAAGCTCCCGCTCCCCATTGCCTTAAGAAGCCACCTCAAGTCCTTCTCCATGGCTAAGGGCCTGAATGCCAGGATGATGCGAGGCCTCTCCTACTCCCTCACCACCAGCTCCACT : 744
250h12cx1 : AGCAAGGTACTGAGCTTGCAAGACCTCTGCTGCCGCACCATCGTGTCCTGCACACCTGTGCATCTGGTGGACAAGCTCCCGCTCCCCAGTACCTTAAGAAGCCACCTCAAGTCCTTCTCCATGGCTAAGGGCCTGAATGCCAGGATGATGCGAGGCCTCTCCTACTCCCTCACCACCAGCTCCACT : 744

                 *       760         *       780         *       800         *       820         *
cu237h11  : CACAAAAGGAGCAGCCTCTGCAAAGTGAAGATCGTCTGCCCACCCCAGAGCCCACCCAAAAACTGCACCAGAAACAGCTGCAAAATTTCT : 834
250h12cx1 : CACAA---GAGCAGCCTCTGCAAAGTGGAGATCGTCTGCCCACCCCAGAGCCCACCCAAAAACTGCACCAGAAACAGCTGCAAAATTTCT : 831
```

(b)

```
                     *        20         *        40         *        60         *        80         *       100         *       120         *       140         *       160         *       180
cu237h11  : MSAPGSPDQAYDFLLKFLLVGDRDVGKSEILESLQDGTAESPYSHLGGIDYKTTTILLDGQRVKLKLWDTSGQGRFCTIFRSYSRGAQGVILVYDIANRWSFEGMDRWIKKIEEHAPGVPKILVGNRLHLAFKRQVPREQAQAYAERLGVTFFEVSPLCNFNIIESFTELARIVLLRHRLNWLGRP : 186
250h12cx1 : MSAPGSPDQAYDFLLKFLLVGDRDVGKSEILESLQDGAAESPYSHLGGIDYKTTTILLDGQRVKLKLWDTSGQGRFCTIFRSYSRGAQGVILVYDIANRWSFEGMDRWIKKIEEHAPGVPKILVGNRLHLAFKRQVPREQAQAYAERLGVTFFEVSPLCNFNIIESFTELARIVLLRHRMNWLGRP : 186

               *       200         *       220         *       240         *       260         *
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).

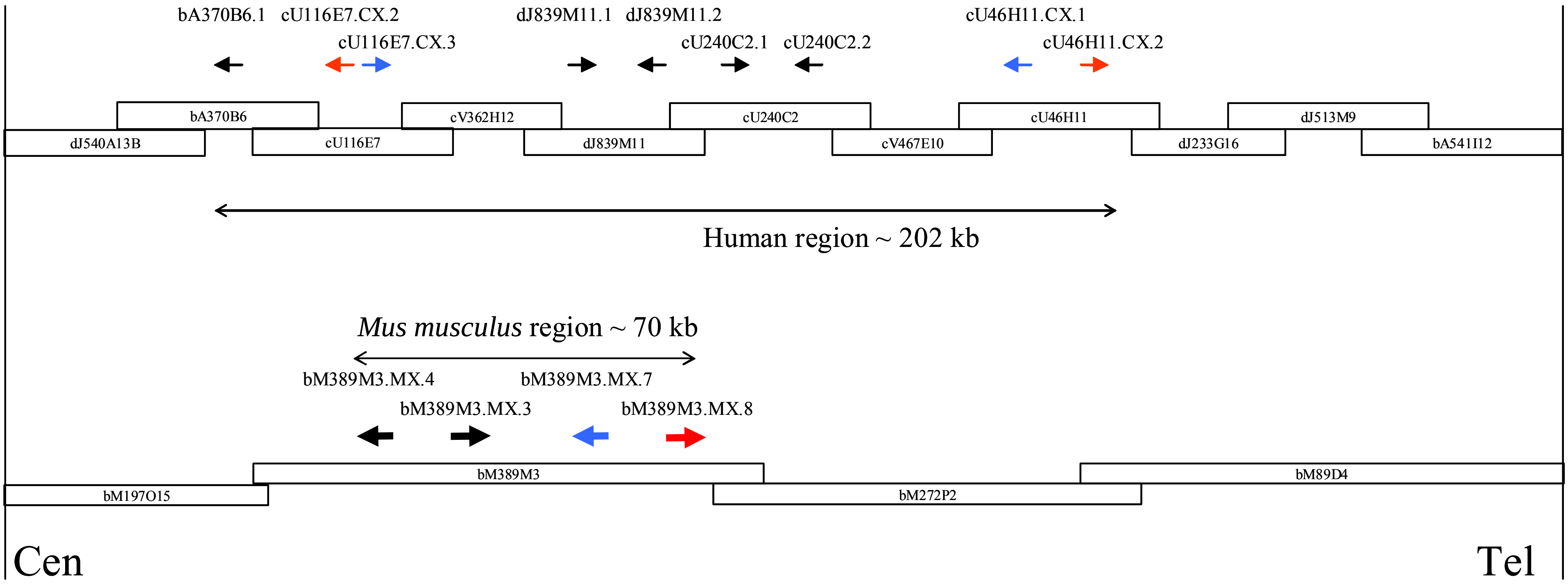
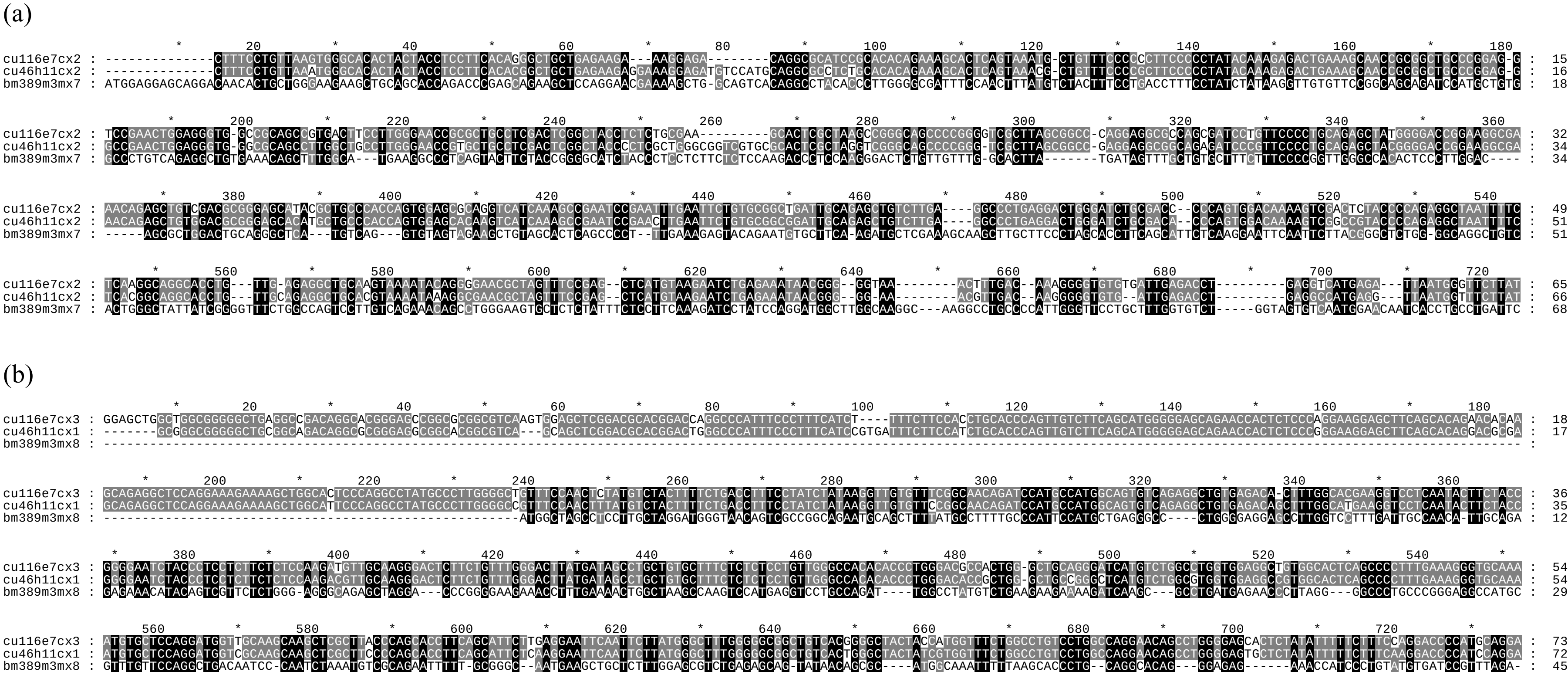


Figure 5-30 Figure showing a schematic representation (not to scale) of histone and cU46H11.CX.1/cU116E7.CX.3 paralogue gene order and orientation within human Xq22, and the corresponding orthologous region in mouse. Large double-headed arrows depict approximate sizes of the regions, small arrows depict 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.



(a)

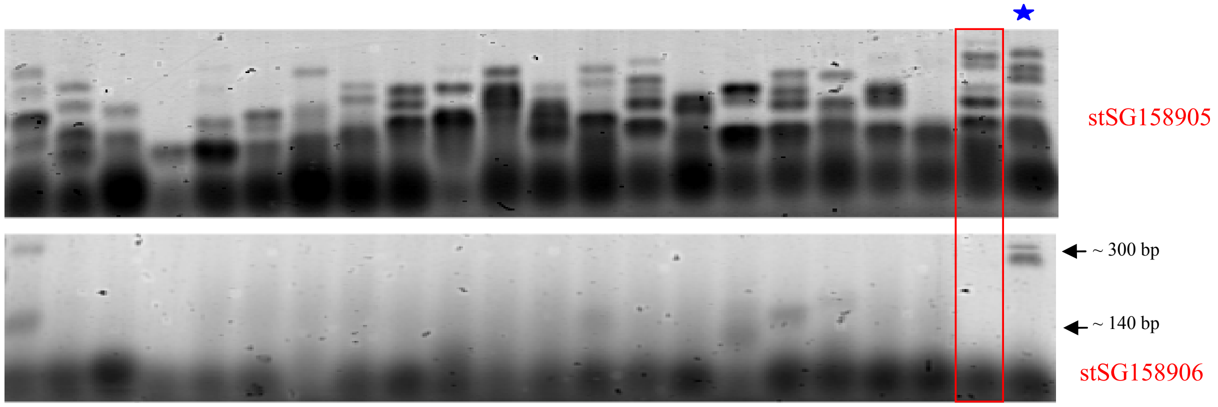
```
                      *        20         *        40         *        60         *        80         *       100         *       120         *       140         *       160         *       180
cu116e7cx2 : --------------CTTTCCTGTTAAGTGGCACACTACTACCTCCTTCACAGGGCTGCTGAGAAGA---AAGGAGA--------CAGGCGCATCCGCACACAGAAAGCACTCAGTAAATG-CTGTTTCCCCCTTCCCCCTATACAAAGAGACTGAAAGCAACCGCGGCTGCCCGGAG-G :  154
cu46h11cx2 : --------------CTTTCCTGTTAAATGGGCACACTACTACCTCCTTCACACGGCTGCTGAGAAGAGGAAAGGAGATGTCCATGCAGGCGCCTCTGCACACAGAAAGCACTCAGTAAACG-CTGTTTCCCCGCTTCCCCCTATACAAAGAGACTGAAAGCAACCGCGGCTGCCCGGAG-G :  165
bm389m3mx7 : ATGGAGGAGCAGGACAACACTGCTGGGAAGAAGCTGCAGCACCAGACCCGAGCAGAAGCTCCAGGAACGAAAAGCTG-GCAGTCACAGGCCTACACCCTTGGGGCGATTTCCAACTTTATGTCTACTTTCCTGACCTTTCCTATCTATAAGGTTGTGTTCCGGCAGCAGATCCATGCTGTG :  180

                     *       200         *       220         *       240         *       260         *       280         *       300         *       320         *       340         *       360
cu116e7cx2 : TCCGAACTGGAGGGTG-GCCGCAGCCGTGACTTCCTTGGGAACCGCGCTGCCTCGACTCGGCTACCTCTCTGCGAA---------GCACTCGCTAAGCCGGGCAGCCCCGGGGTCGCTTAGCGGCC-CAGGAGGCGCCAGCGATCCTGTTCCCCTGCAGAGCTATGGGGACCGGAAGGCGA :  324
cu46h11cx2 : GCCGAACTGGAGGGTG-GGCGCAGCCTTGGCTGCCTTGGGAACCGTGCTGCCTCGACTCGGCTACCCCTCGCTGGGCGGTCGTGCGCACTCGCTAGGTCGGGCAGCCCCGGG-TCGCTTAGCGGCC-GAGGAGGCGGCAGAGATCCCGTTCCCCTGCAGAGCTACGGGGACCGGAAGGCGA :  343
bm389m3mx7 : GCCCTGTCAGAGGCTGTGAAACAGCTTTGGCA---TGAAGGCCCTCAGTACTTCTACCGGGGCATCTACCCTCCTCTTCTCTCCAAGACCCTCCAAGGGACTCTGTTGTTTG-GCACTTA-------TGATAGTTTGCTGTGCTTTCTTTCCCCGGTTGGGCCACACTCCCTTGGAC---- :  346

                    *       380         *       400         *       420         *       440         *       460         *       480         *       500         *       520         *       540
cu116e7cx2 : AACAGAGCTGTCGACGCGGGAGCATACGCTGCCCACCAGTGGAGCGCAGGTCATCAAAGCCGAATCCGAATTTGAATTCTGTGCGGCTGATTGCAGAGCTGTCTTGA----GGCCCTGAGGACTGGGATCTGCGACC--CCCAGTGGACAAAAGTCGACTCTACCCCAGAGGCTAATTTTC :  499
cu46h11cx2 : AACAGAGCTGTGGACGCGGGAGCACATGCTGCCCACCAGTGGAGCACAAGTCATCAAAGCCGAATCCGAACTTGAATTCTGTGCGGCGGATTGCAGAGCTGTCTTGA----GGCCCTGAGGACTGGGATCTGCGACA--CCCAGTGGACAAAAGTCGGCCGTACCCCAGAGGCTAATTTTC :  518
bm389m3mx7 : -----AGCGCTGGACTGCAGGGCTCA---TGTCAG---GTGTAGTAGAAGCTGTAGCACTCAGCCCCT--TTGAAAGAGTACAGAATGTGCTTCA-AGATGCTCGAAAGCAAGCTTGCTTCCCTAGCACCTTCAGCATTCTCAAGGAATTCAATTCTTACGGGCTCTGG-GGCAGGCTGTC :  512

                   *       560         *       580         *       600         *       620         *       640         *       660         *       680         *       700         *       720
cu116e7cx2 : TCAAGGCAGGCACCTG---TTG-AGAGGCTGCAAGTAAAATACAGGGGAACGCTAGTTTCCGAG--CTCATGTAAGAATCTGAGAAATAACGGG--GGTAA--------ACTTTGAC--AAAGGGGTGTGTGATTGAGACCT---------GAGGTCATGAGA---TTAATGGGTTCTTAT :  650
cu46h11cx2 : TCACGGCAGGCACCTG---TTGCAGAGGCTGCACGTAAAATAAAGGCGAACGCTAGTTTCCGAG--CTCATGTAAGAATCTGAGAAATAACGGG--GG-AA--------ACGTTGAC--AAGGGGGTGTG--ATTGAGACCT---------GAGGCCATGAGG---TTAATGGTTTCTTAT :  667
bm389m3mx7 : ACTGGGCTATTATCGGGGTTTCTGGCCAGTCCTTGTCAGAAACAGCCTGGGAAGTGCTCTCTATTTCTCCTTCAAAGATCCTATCCAGGATGGCTTGGCAAGGC---AAGGCCTGCCCCATTGGGTTCCTGCTTTGGTGTCT-----GGTAGTGTCAATGGAACAATCACCTGCCTGATTC :  685
```

(b)

```
                      *        20         *        40         *        60         *        80         *       100         *       120         *       140         *       160         *       180
cu116e7cx3 : GGAGCTGGCTGGCGGGGGCTGAGGCCGACAGGCACGGGAGCCGGCGCGGCGTCAAGTGGAGCTCGGACGCACGGACCAGGCCCATTTCCCTTTCATCT----TTTCTTCCACCTGCACCCAGTTGTCTTCAGCATGGGGGAGCAGAACCACTCTCCCAGGAAGGAGCTTCAGCACAGAACACAA :  180
cu46h11cx1 : -------GCGGGCGGGGGCTGCGGCAGACAGGCGCGGGAGGCGGCACGGCGTCA---GCAGCTCGGACGCACGGACTGGGCCCATTTCCCTTTCATCCGTGATTTCTTCCATCTGCACCCAGTTGTCTTCAGCATGGGGGAGCAGAACCACTCTCCCGGGAAGGAGCTTCAGCACAGGACGCGA :  174
bm389m3mx8 : ---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------- :    -

                  *       200         *       220         *       240         *       260         *       280         *       300         *       320         *       340         *       360
cu116e7cx3 : GCAGAGGCTCCAGGAAAGAAAAGCTGGCACTCCCAGGCCTATGCCCTTGGGGCTGTTTCCAACTCTATGTCTACTTTTCTGACCTTTCCTATCTATAAGGTTGTGTTTCGGCAACAGATCCATGCCATGGCAGTGTCAGAGGCTGTGAGACA-CTTTGGCACGAAGGTCCTCAATACTTCTACC :  363
cu46h11cx1 : GCAGAGGCTCCAGGAAAGAAAAGCTGGCATTCCCAGGCCTATGCCCTTGGGGCCGTTTCCAACTTTATGTCTACTTTTCTGACCTTTCCTATCTATAAGGTTGTGTTCCGGCAACAGATCCATGCCATGGCAGTGTCAGAGGCTGTGAGACAGCTTTGGCATGAAGGTCCTCAATACTTCTACC :  358
bm389m3mx8 : ------------------------------------------------------ATGGCTAGCCTCCTTGCTAGGATGGGTAACAGTCGCCGGCAGAATGCAGCTTTTATGCCTTTTGCCCATTCCATGCTGAGGGCC----CTGGGGAGGAGCCTTGGTCCTTTGATTGCCAACA-TTGCAGA :  125

              *       380         *       400         *       420         *       440         *       460         *       480         *       500         *       520         *       540         *
cu116e7cx3 : GGGGAATCTACCCTCCTCTTCTCTCCAAGATGTTGCAAGGGACTCTTCTGTTTGGGACTTATGATAGCCTGCTGTGCTTTCTCTCTCCTGTTGGGCCACACACCCTGGGACGCCACTGG-GCTGCAGGGATCATGTCTGGCCTGGTGGAGGCTGTGGCACTCAGCCCCTTTGAAAGGGTGCAAA :  546
cu46h11cx1 : GGGGAATCTACCCTCCTCTTCTCTCCAAGACGTTGCAAGGGACTCTTCTGTTTGGGACTTATGATAGCCTGCTGTGCTTTCTCTCTCCTGTTGGGCCACACACCCTGGGACACCGCTGG-GCTGCCGGGCTCATGTCTGGCGTGGTGGAGGCCGTGGCACTCAGCCCCTTTGAAAGGGTGCAAA :  541
bm389m3mx8 : GAGAAACATACAGTCGTTCTCTGGG-AGGGCAGAGCTAGGA---CCCGGGGAAGAAACCTTTGAAAACTGGCTAAGCCAAGTCCATGAGGTCCTGCCAGAT----TGGCCTATGTCTGAAGAAGAAAAGATCAAGC---GCCTGATGAGAACCCTTAGG---GGCCCTGCCCGGGAGGCCATGC :  295

                  560         *       580         *       600         *       620         *       640         *       660         *       680         *       700         *       720         *
cu116e7cx3 : ATGTGCTCCAGGATGGTTGCAAGCAAGCTCGCTTACCCAGCACCTTCAGCATTCTTGAGGAATTCAATTCTTATGGGCTTTGGGGGCGGCTGTCACGGGGCTACTACCATGGTTTCTGGCCTGTCCTGGCCAGGAACAGCCTGGGGAGCACTCTATATTTTTCTTTCCAGGACCCCATGCAGGA :  730
cu46h11cx1 : ATGTGCTCCAGGATGGTCGCAAGCAAGCTCGCTTCCCCAGCACCTTCAGCATTCTCAAGGAATTCAATTCTTATGGGCTTTGGGGGCGGCTGTCACTGGGCTACTATCGTGGTTTCTGGCCTGTCCTGGCCAGGAACAGCCTGGGGAGTGCTCTATATTTTTCTTTCAAGGACCCCATCCAGGA :  725
bm389m3mx8 : GTTTGTTCCAGGCTGACAATCC-CAATCTAAATGTCGCAGAATTTTT-GCGGGC--AATGAAGCTGCTCTTTGGAGCGTCTGAGAGCAG-TATAACAGCGC----ATGGCAAATTTTTAAGCACCTG--CAGGCACAG---GGAGAG------AAACCATCCCTGTATGTGATCCGTTTAGA- :  458
```

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.

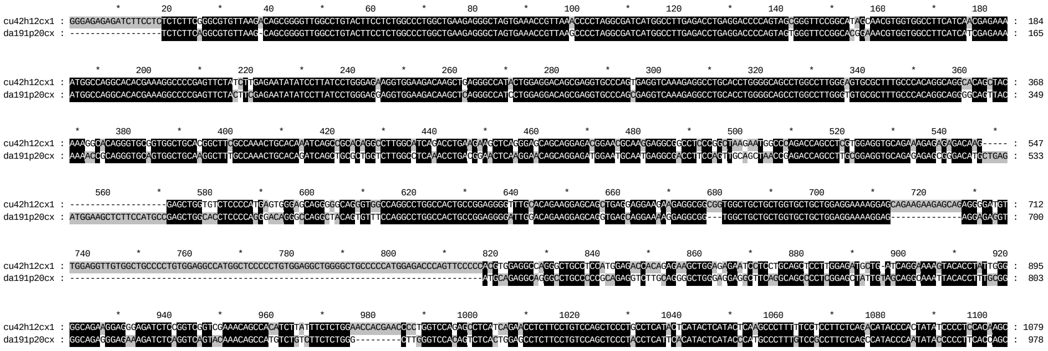


Figure 5-33 Partial alignment of TEX mRNA sequences.

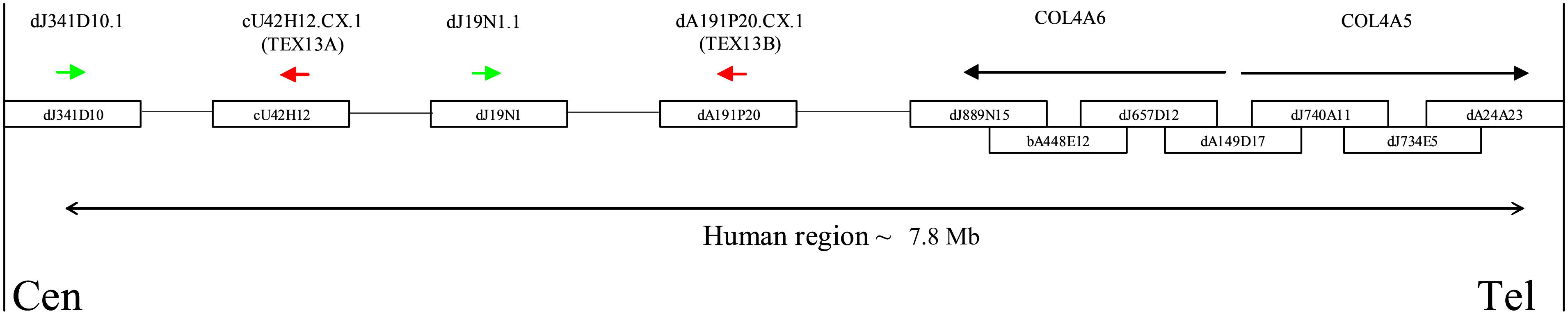


Figure 5-34 Figure showing a schematic representation (not to scale) of pseudogene, TEX and COL4A5/COL4A6 paralogue gene order and orientation within human Xq22. Large double-headed arrows depict approximate sizes of the regions, small arrows depict gene positions (names above) and direction of transcription. 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.

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

6.1 Introduction

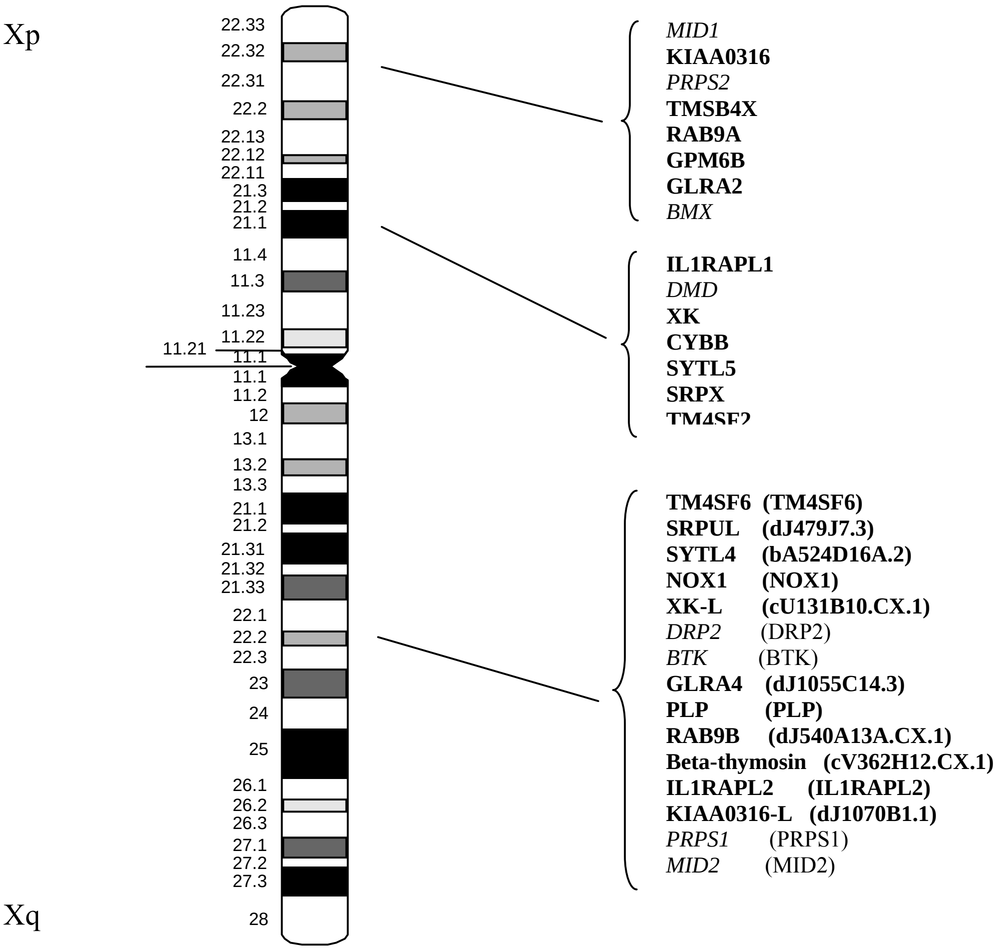
The availability of genomic sequence data has enabled several recent studies of sequence duplications within the human genome (McLysaght *et al.*, 2002), (Gu *et al.*, 2002). These genome-wide studies shed light on the extent of tandem and regional duplications within the human genome, and provide data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

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

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

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

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



6.2 Characterisation of the Xq22-q23/Xp regional duplication

6.2.1 Extent of the duplication and genes involved

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

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

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

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

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

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

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

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

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

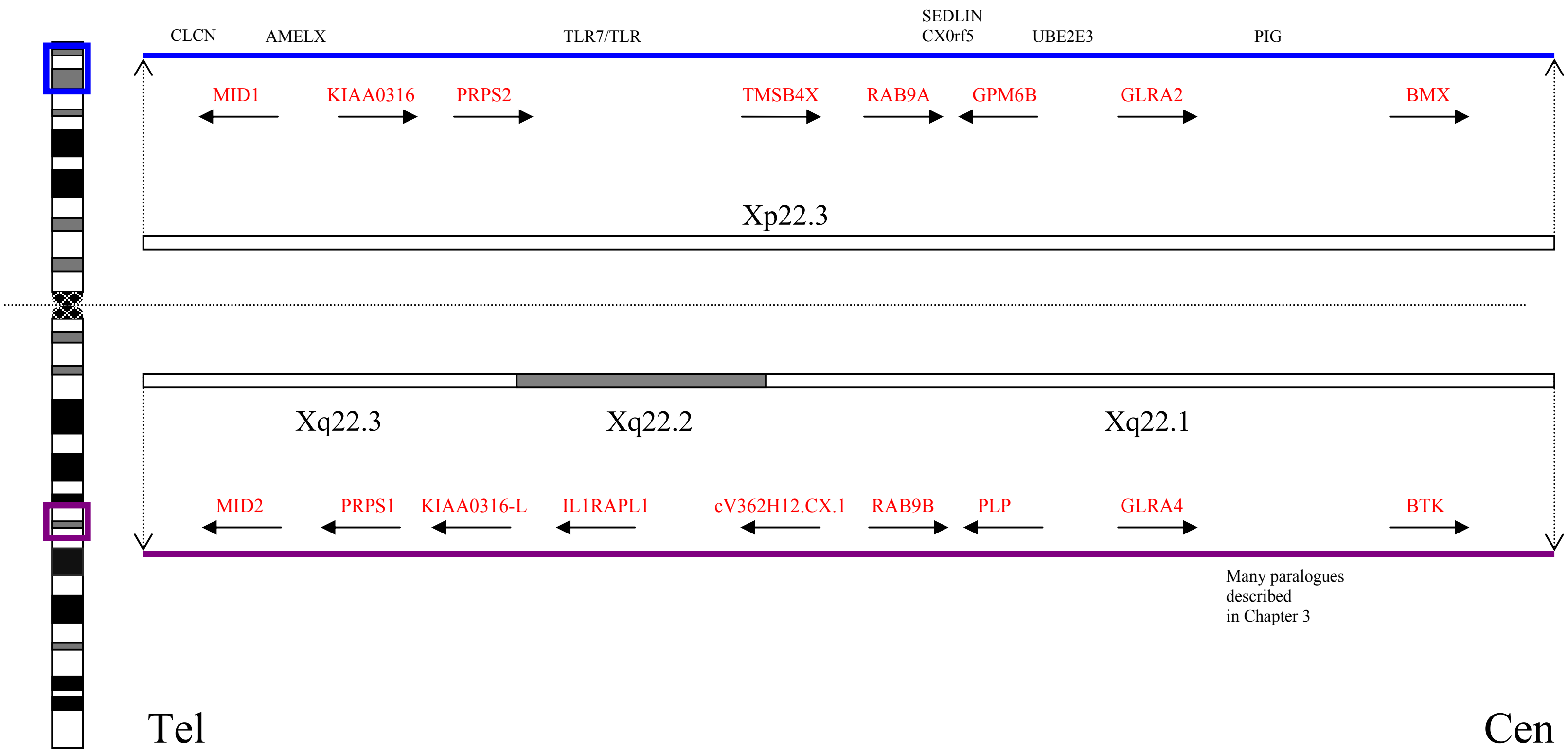


Figure 6- 2 Schematic representation of paralogy between Xp22.3 and Xq22.1-q23 (Block 1). Paralogous genes are represented in red type, with their direction of transcription depicted by a black arrow. Genes are shown in their order along the chromosome (Tel to Cen) relative to one another. Xp genes are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

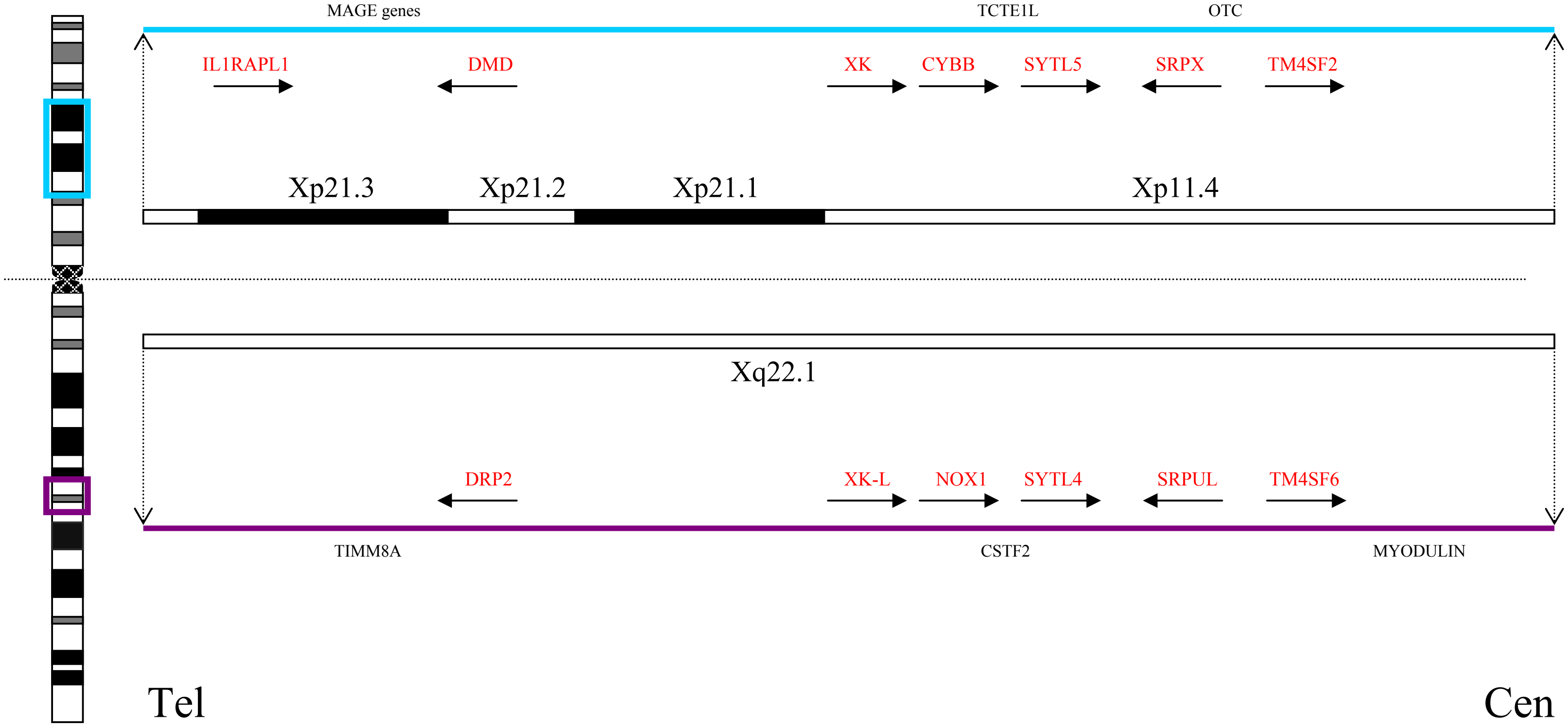


Figure 6-3 Schematic representation of paralogy between Xp21.3-p11.4 and Xq22.1 (Block 2). Paralogous genes are represented in red type, with their direction of transcription depicted by a black arrow. Genes are shown in their order along the chromosome (Tel to Cen) relative to one another. Xp genes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

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

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

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

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

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

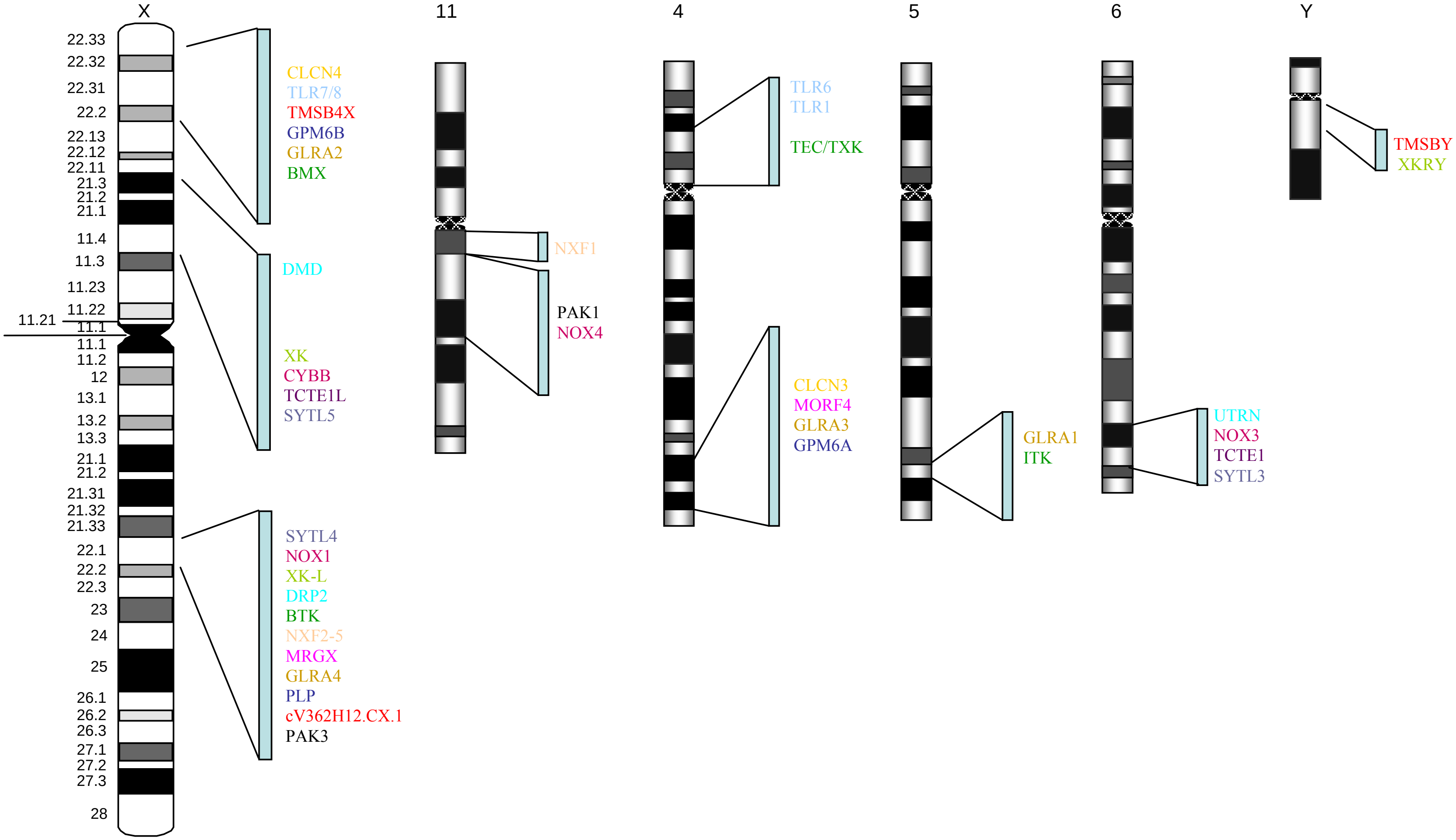


Figure 6-4 Schematic representation of chromosomal locations of autosomal genes with paralogues on the X chromosome, some of which are Xp/Xq paralogues. Names are coloured according to similarity.

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

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

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

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

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

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

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

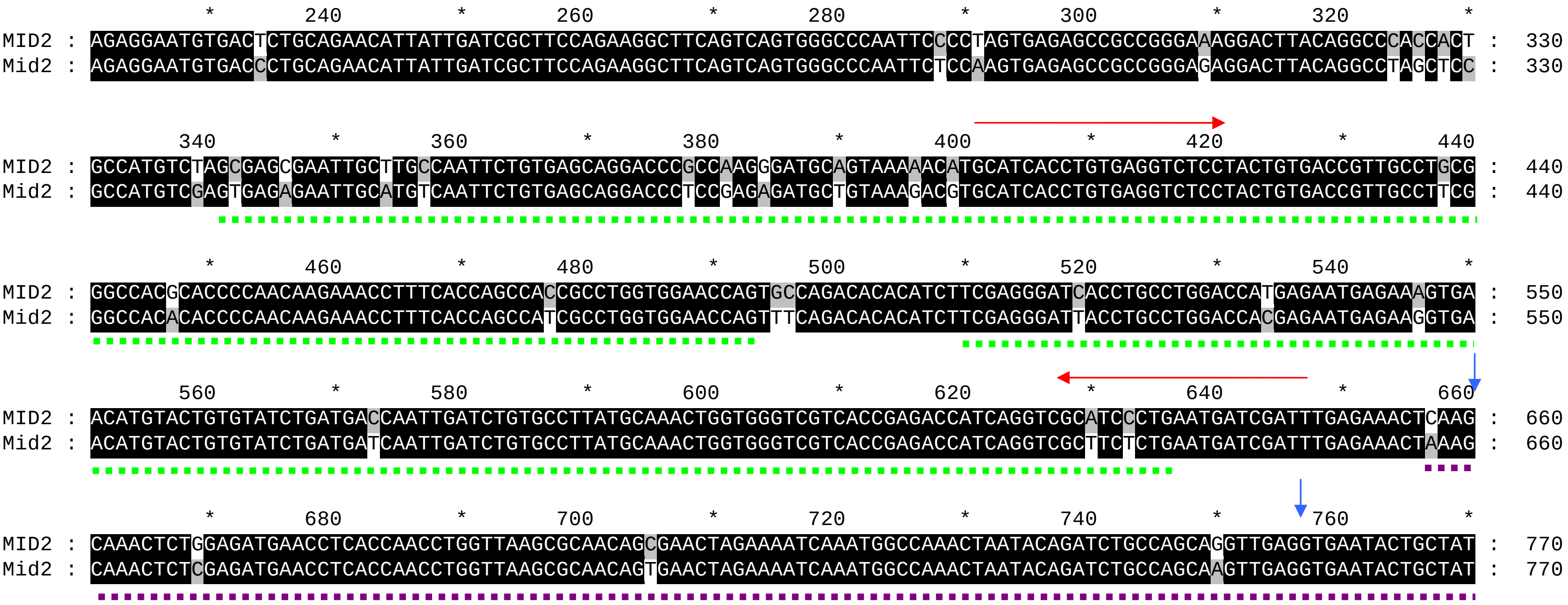
In Figure 6-5, the green line represents domain IPR000315 (Zn-finger B-box, matches 385 proteins) and the purple line domain IPR003649 (Bbox_C, matches 66 proteins). Although encoding protein domains, this region was chosen as further 3', domains with a higher number of protein matches were found. Primers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

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

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

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

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



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

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

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

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

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

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

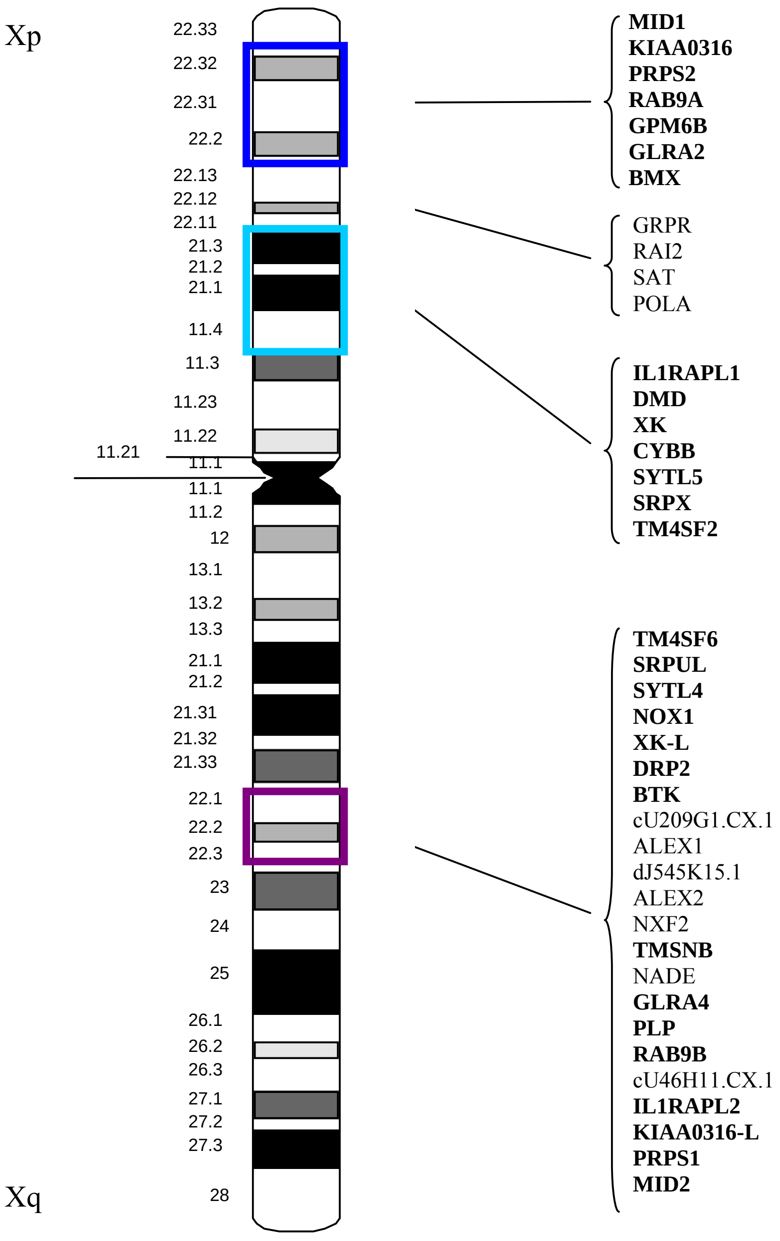


Figure 6-6 Diagram showing the genes for which probes were designed to identify orthologues in *Sminthopsis macroura*, and their positions on the human X chromosome. The genes are listed in order from Xpter to Xqter. The main blocks of Xp/Xq paralogy are denoted by the red, purple and green boxes on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

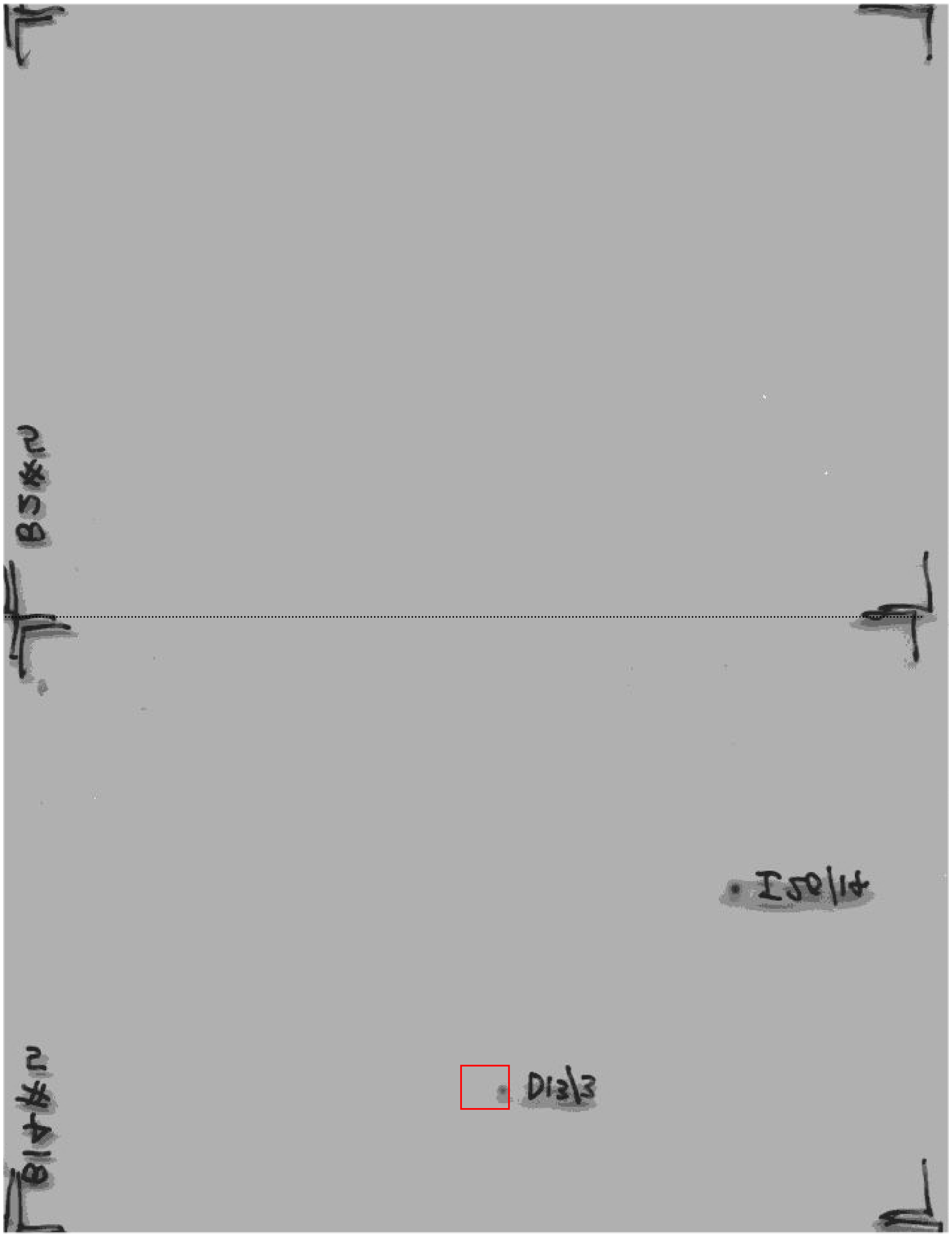


Figure 6-7 An example of a hybridisation of a pool of five probes to filters of the *Sminthopsis macroura* library. The diagram shows two filters of the gridded library (separated by a dotted line) following hybridisation of a pool of five STSs and washing as described in the text. The four corner edge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

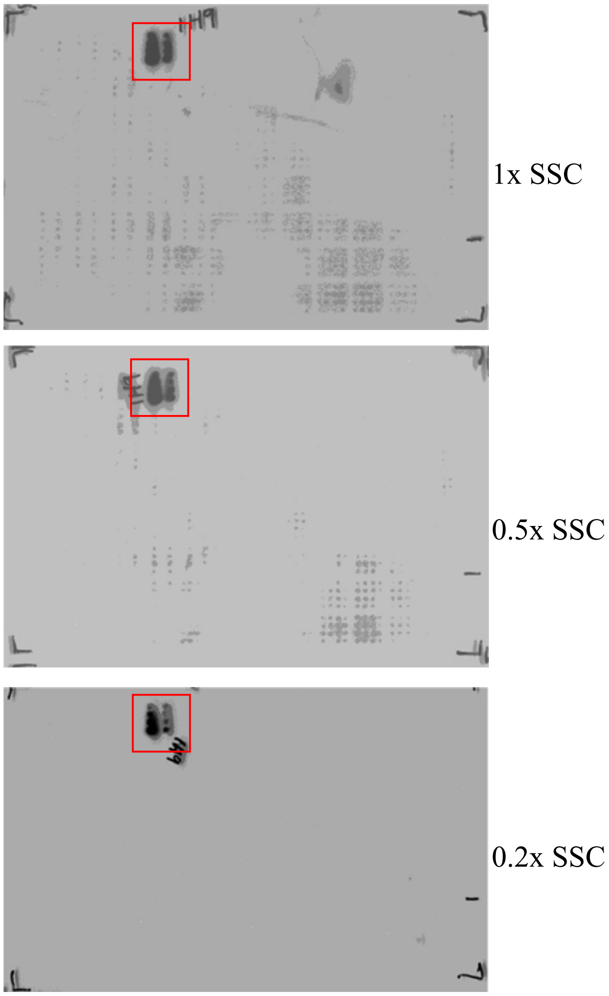


Figure 6-8 An example of the second round of the reduced-stringency hybridisation procedure. Three images of autoradiographs are shown, following hybridisation with a probe generated to the MID1 gene (stSG187894), and filters washed at increasing stringency (1x SSC, 0.5x SSC and 0.2x SSC). The red box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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

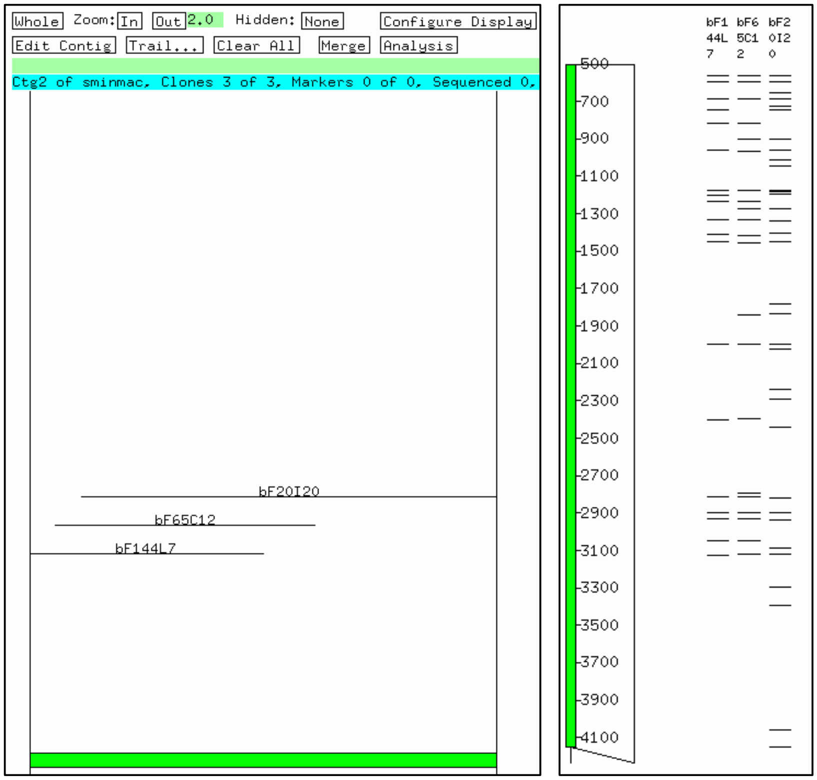
The 157 fingerprints were assembled into contigs in FPC (Chapter 2). Fingerprinting resulted in the incorporation of 37 clones into 11 contigs. It is possible that more contigs may have been generated by lowering the stringency parameters for contig formation, however already one of the contigs, contig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

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

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

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

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



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

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

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

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

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

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

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

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

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

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

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

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

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

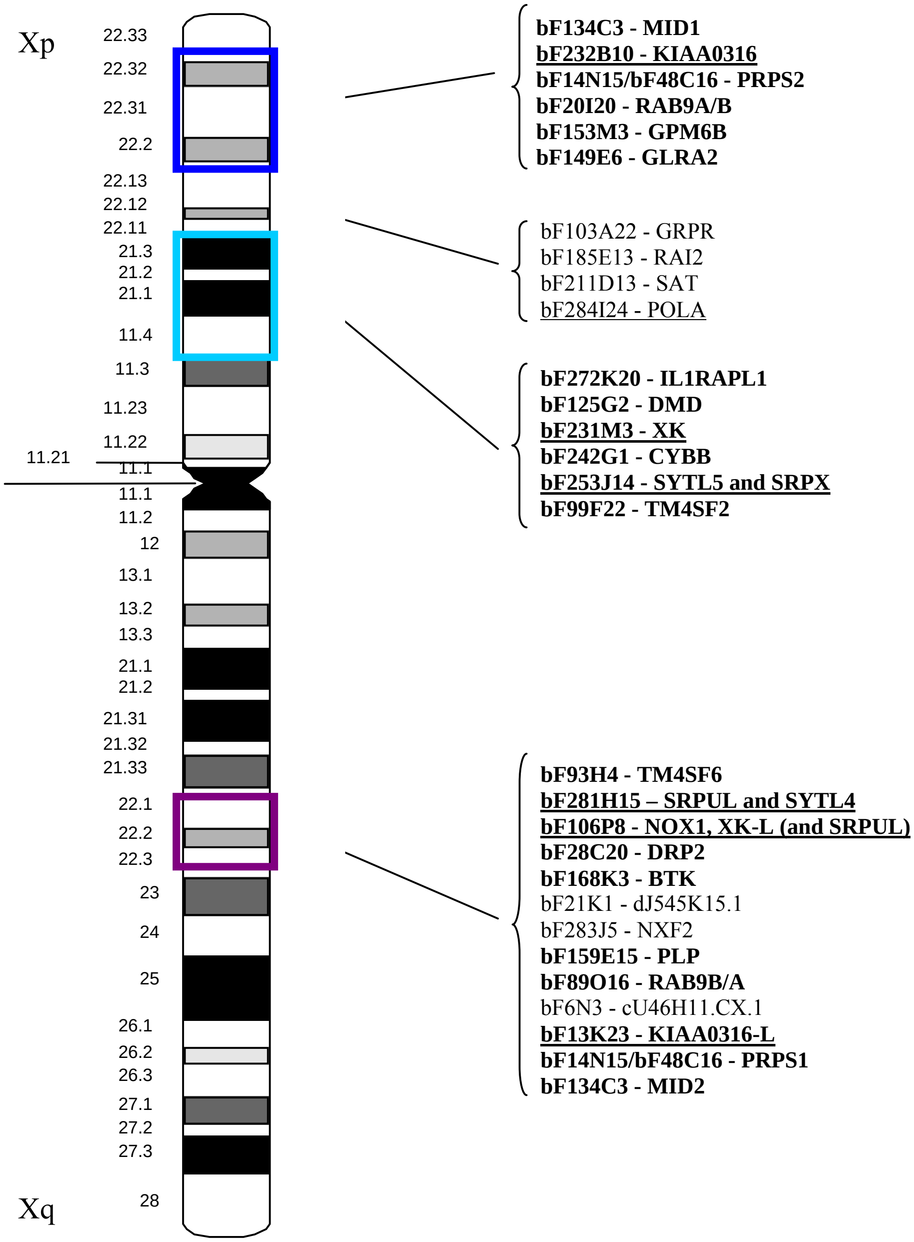


Figure 6-10 Diagram illustrating genes for which *S. macroura* positive BACs were selected for FISH analysis and sequencing. Positions of the human genes relative to the human X chromosome are illustrated, together with their selected BACs. The genes are listed in order from Xpter-Xqter. The main blocks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

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

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

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

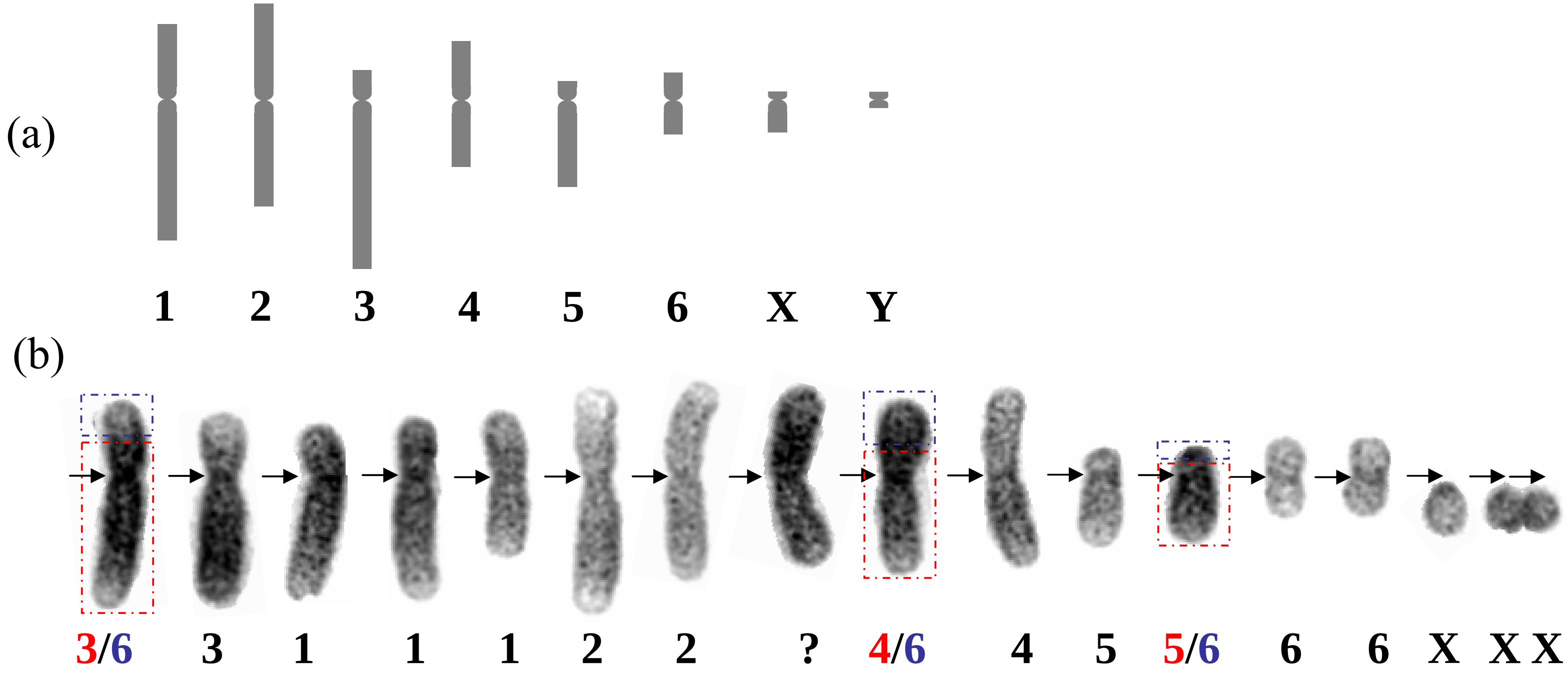


Figure 6-11 Karyogram showing (a) *Sminthopsis macroura* normal karyotype ideogram (from (De Leo *et al.*, 1999)), (b) Representative DAPI-stained chromosomes from metaphase chromosome preparations from a male *Sminthopsis macroura* cell line (2n=18) used for FISH analyses, obtained as a kind gift from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

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

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

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

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

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

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

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

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

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

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

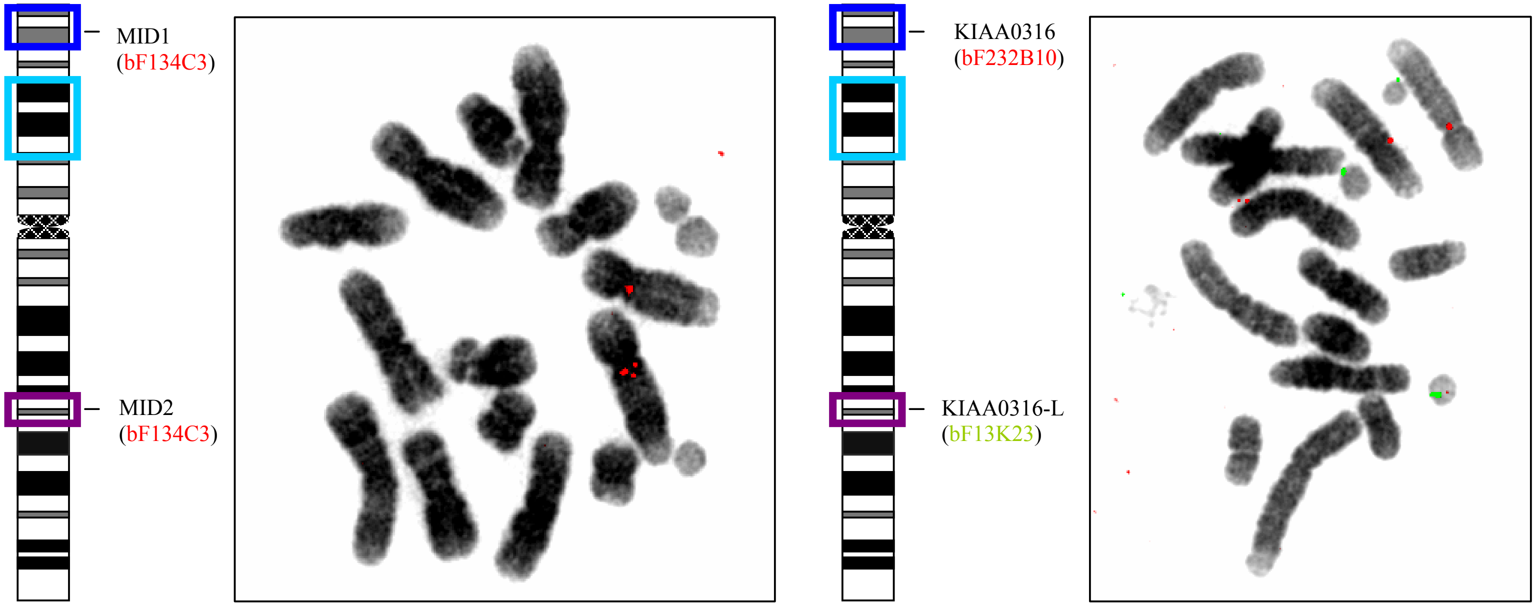


Figure 6-12 FISH of *Sminthopsis macroura* BAC clones against spreads of metaphase chromosomes. The human gene and the hybridisation-positive *Sminthopsis macroura* BAC clone used for FISH are shown against an ideogram of the human X chromosome to illustrate positioning. The colour of the BAC clone name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

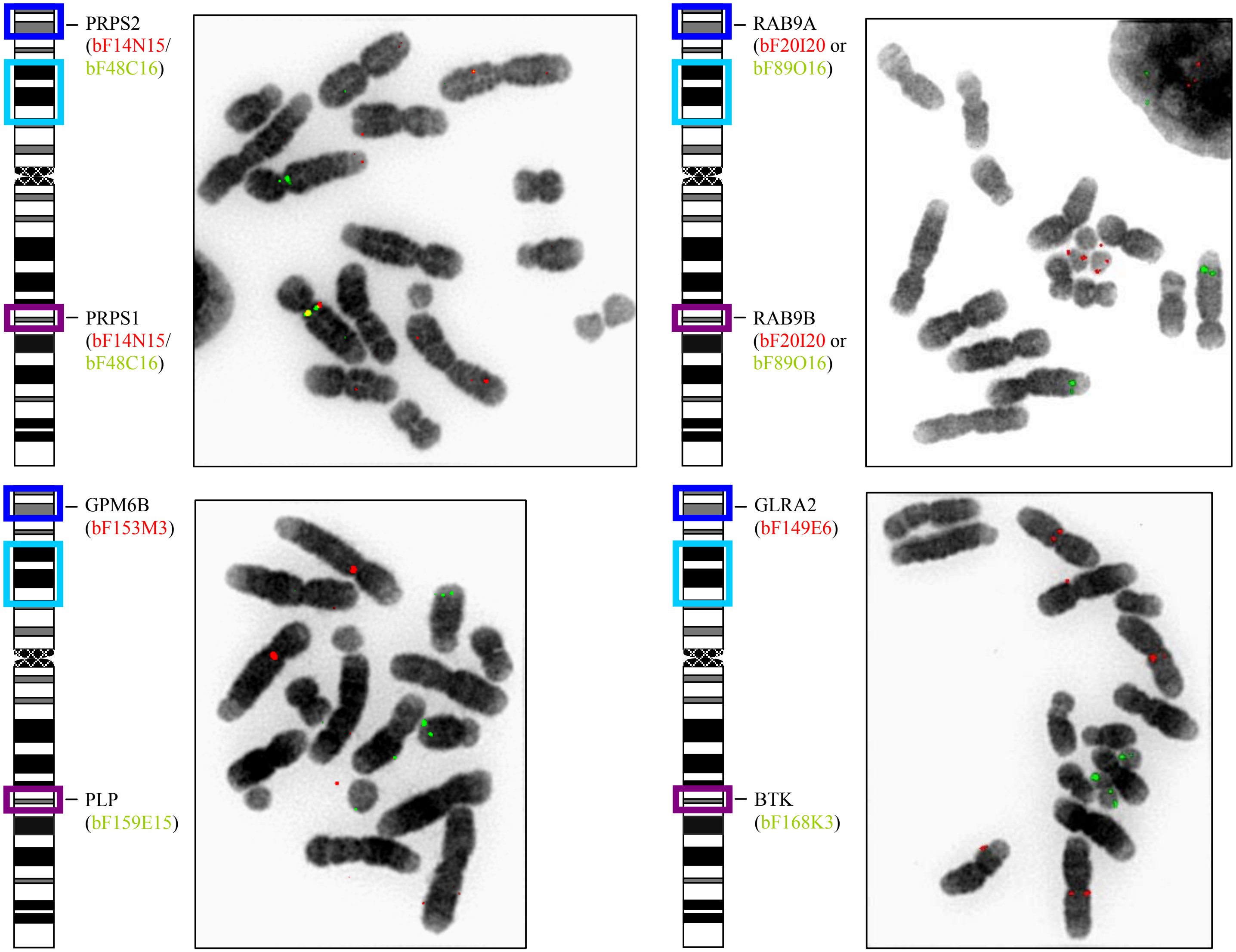


Figure 6-13 Legend as for Figure 6-12.

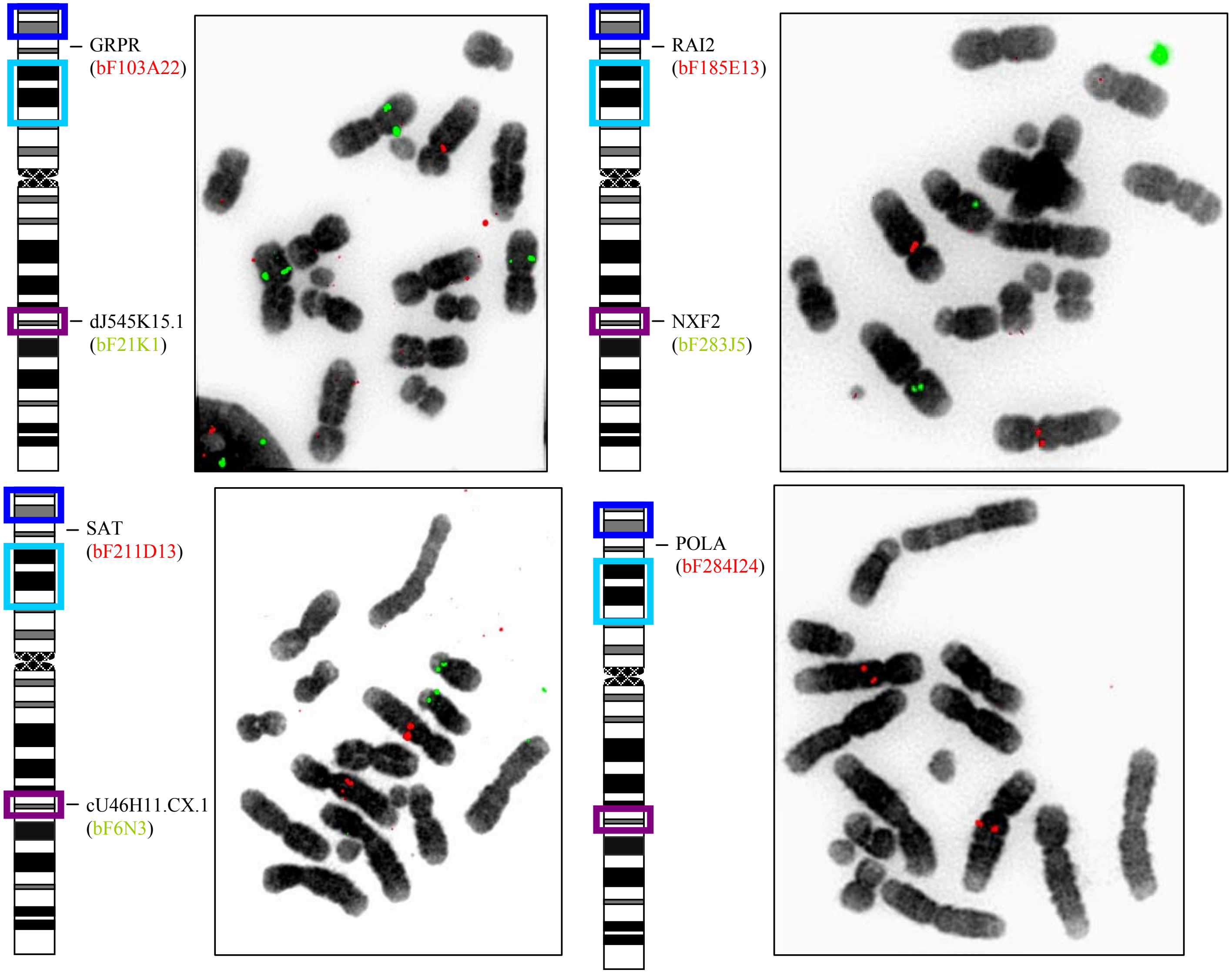


Figure 6-14 Legend as for Figure 6-12.

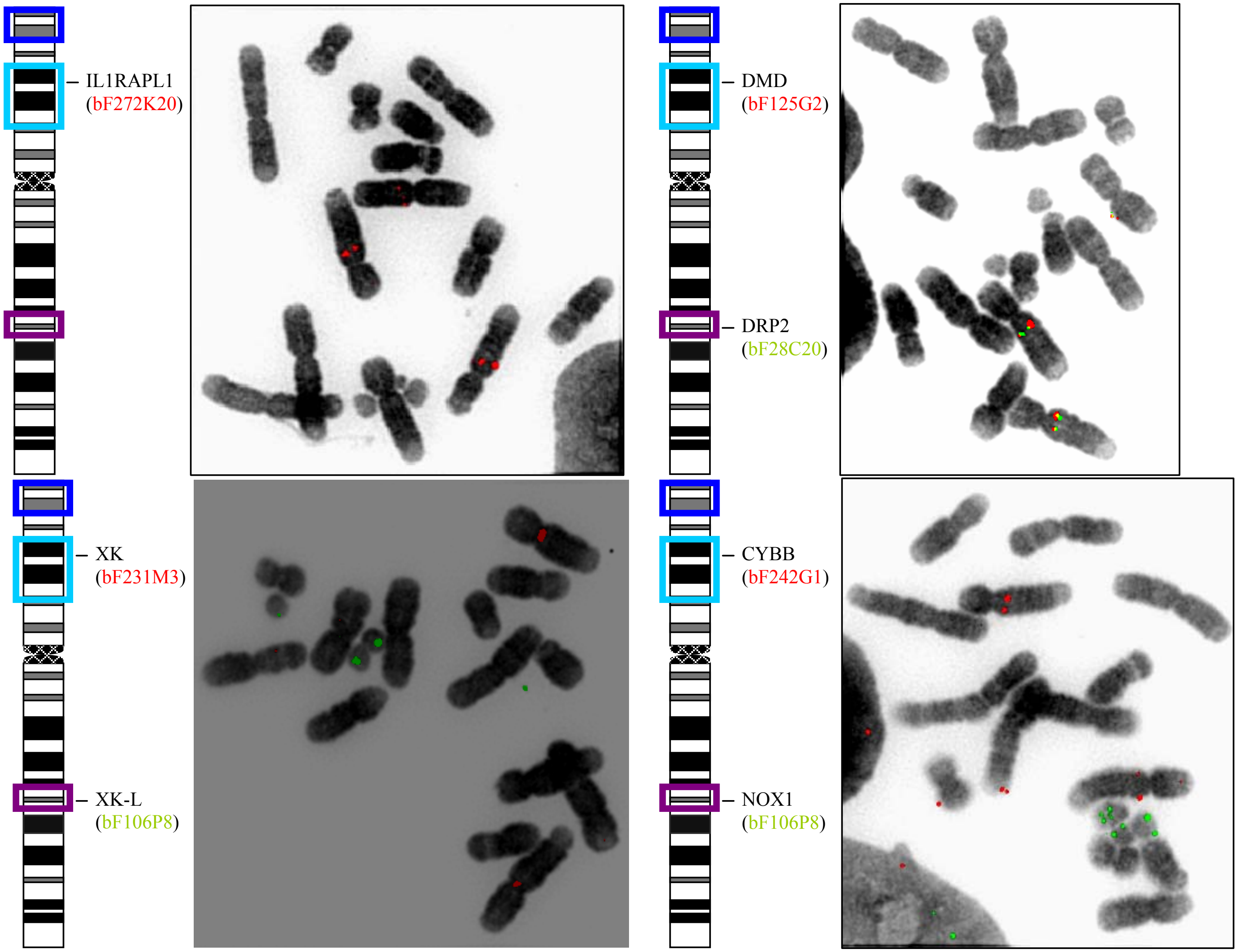


Figure 6-15 Legend as for Figure 6-12.

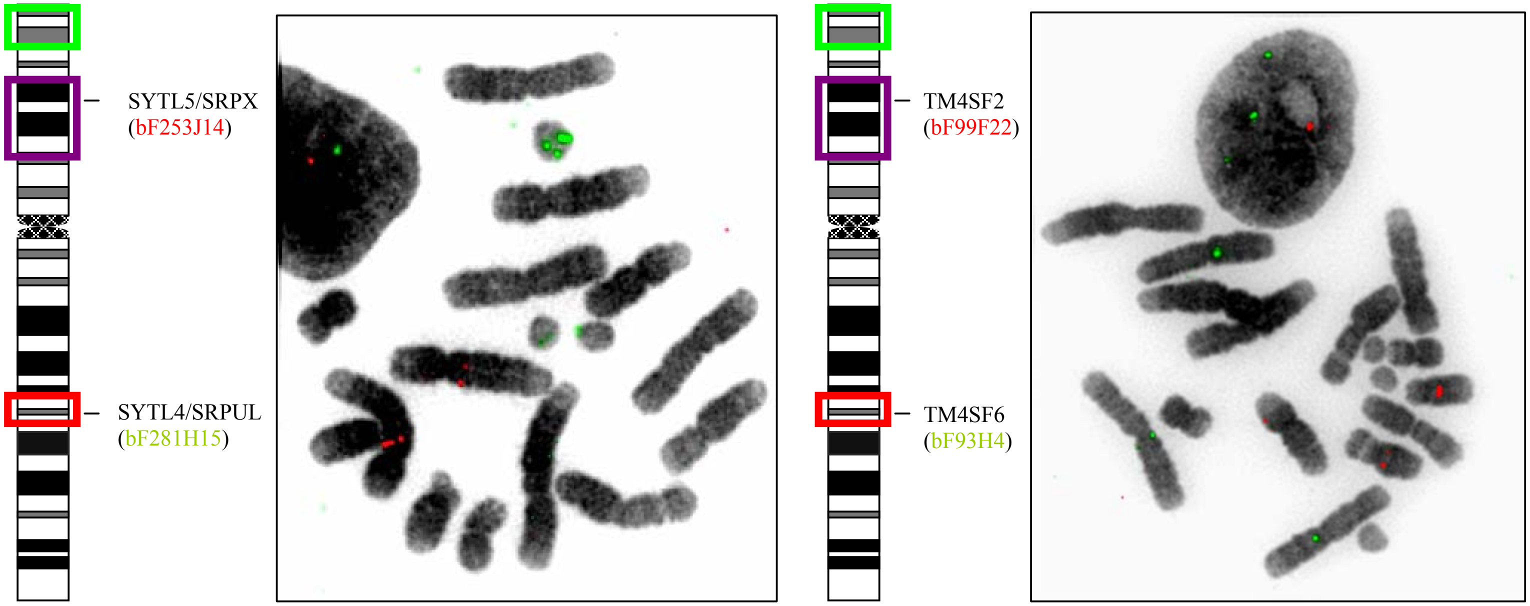


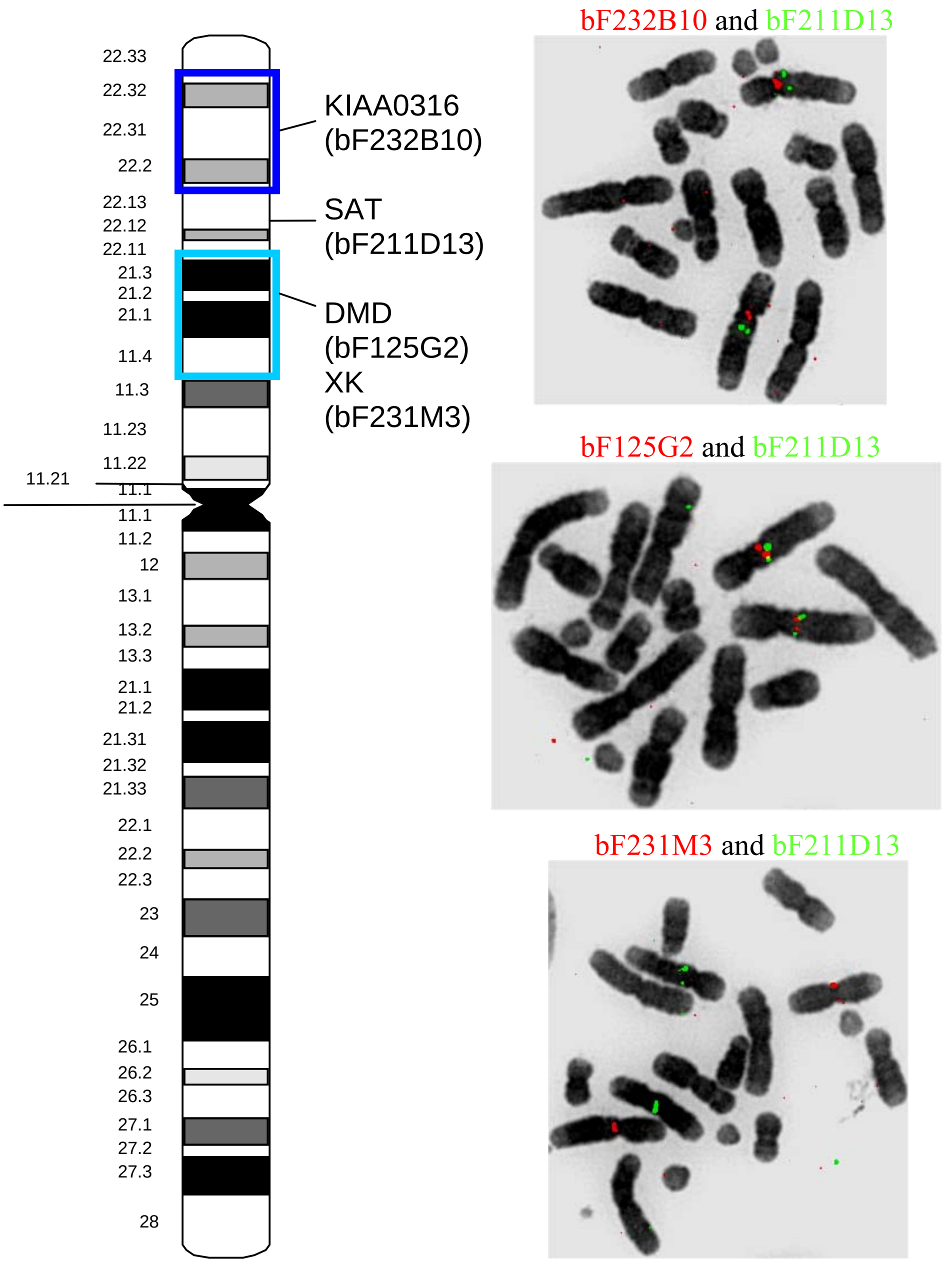
Figure 6-16 Legend as for Figure 6-12.

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

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

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

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



bF232B10 and bF211D13

bF125G2 and bF211D13

bF231M3 and bF211D13

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

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

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

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

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

6.5 Dating the Xq22-q23/Xp regional duplication

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

6.5.1 Gene-based evidence from the scientific literature

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

6.5.1.1 Lipophilins

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

6.5.1.2 Dystrophins

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

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

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

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

6.5.2 Comparative analysis of the Fugu rubripes genome

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

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

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

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

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

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

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

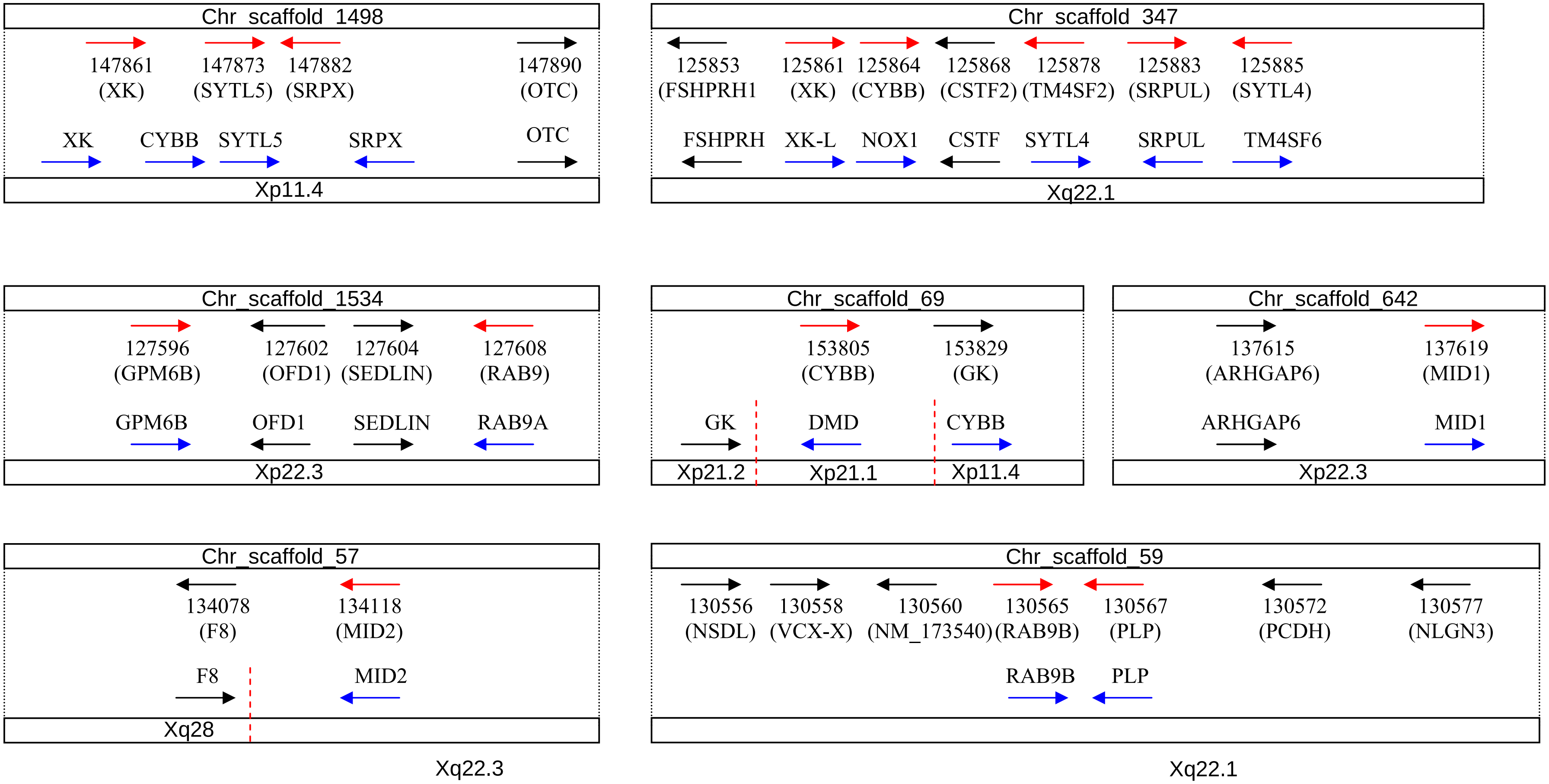


Figure 6-18 Figure showing a schematic representation of selected *Fugu rubripes* WGS sequence scaffolds with information regarding putative Fugu orthologue gene order, transcription direction and shared synteny with human Xp/Xq paralogue and non-Xp/Xq paralogue orthologues. Dotted lines join the Fugu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

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

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

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

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

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

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

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

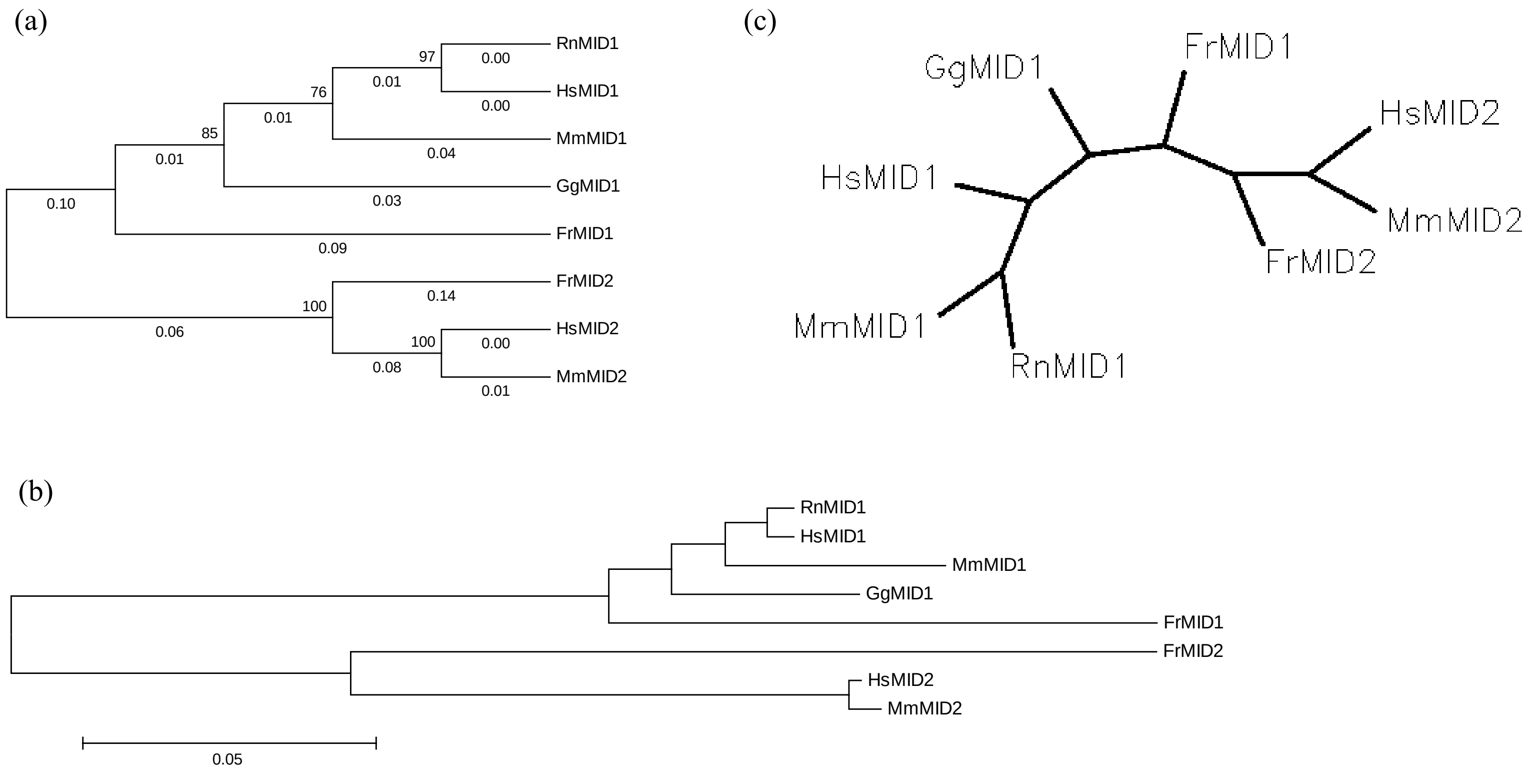


Figure 6-19 The figure shows phylogenetic trees constructed for the MID genes. (a) shows an un-rooted tree constructed using distance measurements. For clarity, only the topology is shown, with distance measurements for each branch shown below the branch and bootstrap support (% agreement from 1000 replicates)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

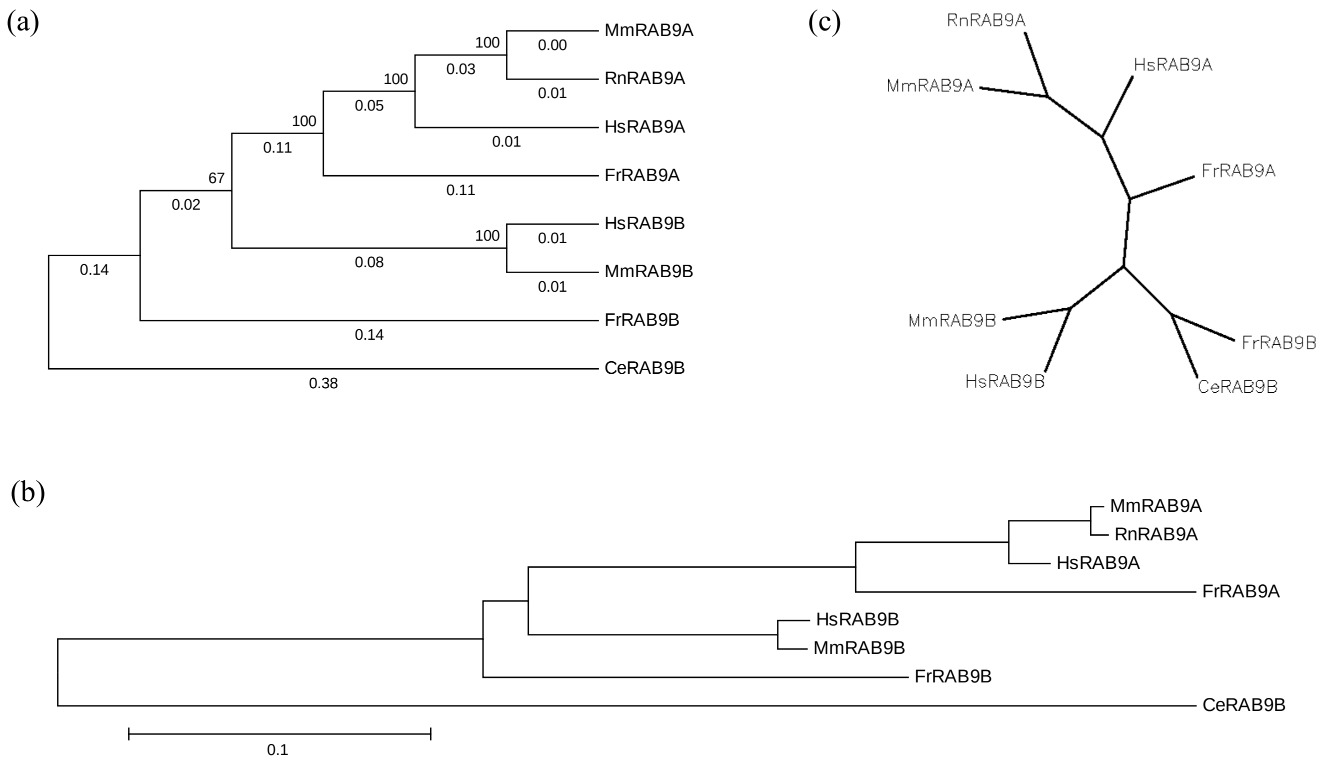


Figure 6-20 The figure shows phylogenetic trees constructed for the RAB genes. (a) shows an un-rooted tree constructed using distance measurements. For clarity, only the topology is shown, with distance measurements for each branch shown below the branch and bootstrap support (% agreement from 1000 replicates)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

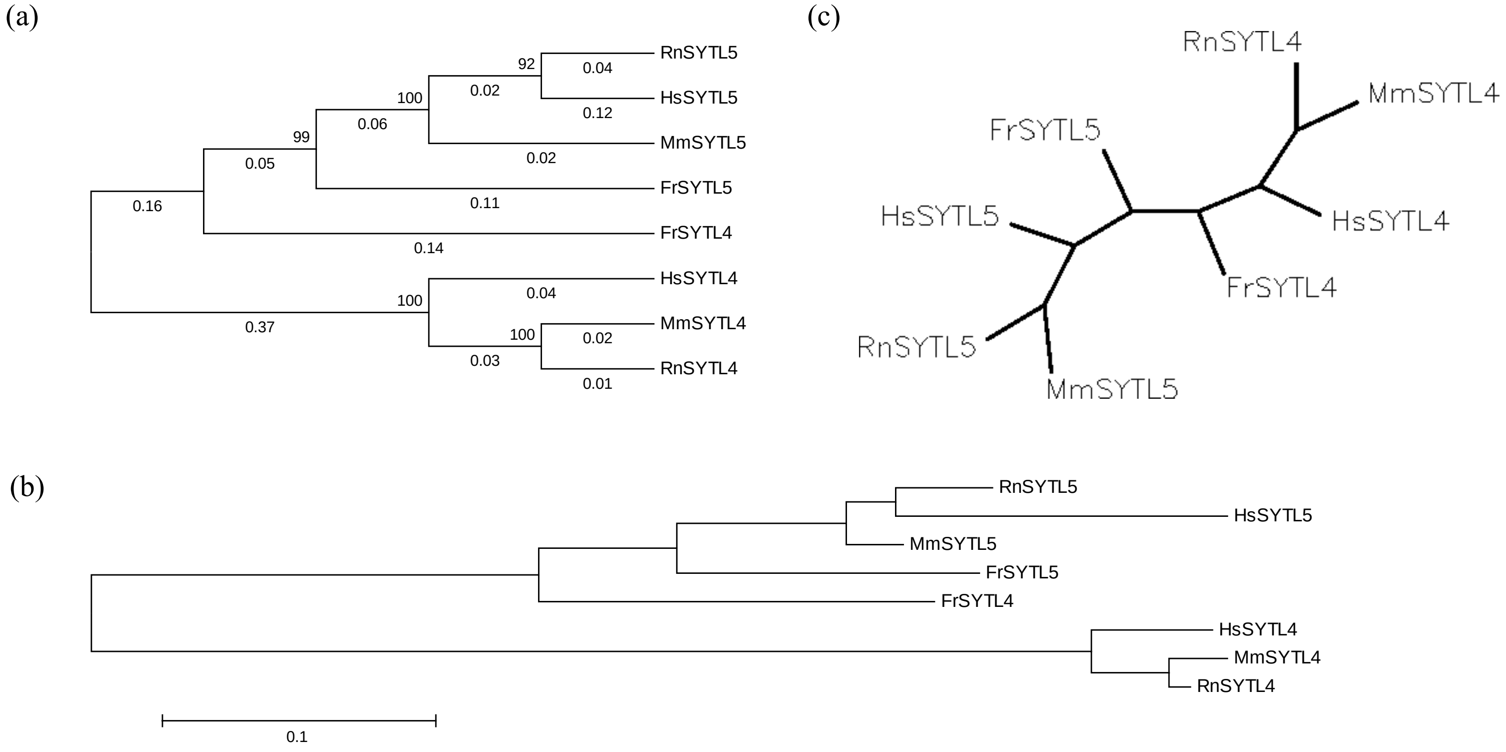


Figure 6-21 The figure shows phylogenetic trees constructed for the SYTL genes. (a) shows an un-rooted tree constructed using distance measurements. For clarity, only the topology is shown, with distance measurements for each branch shown below the branch and bootstrap support (% agreement from 1000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

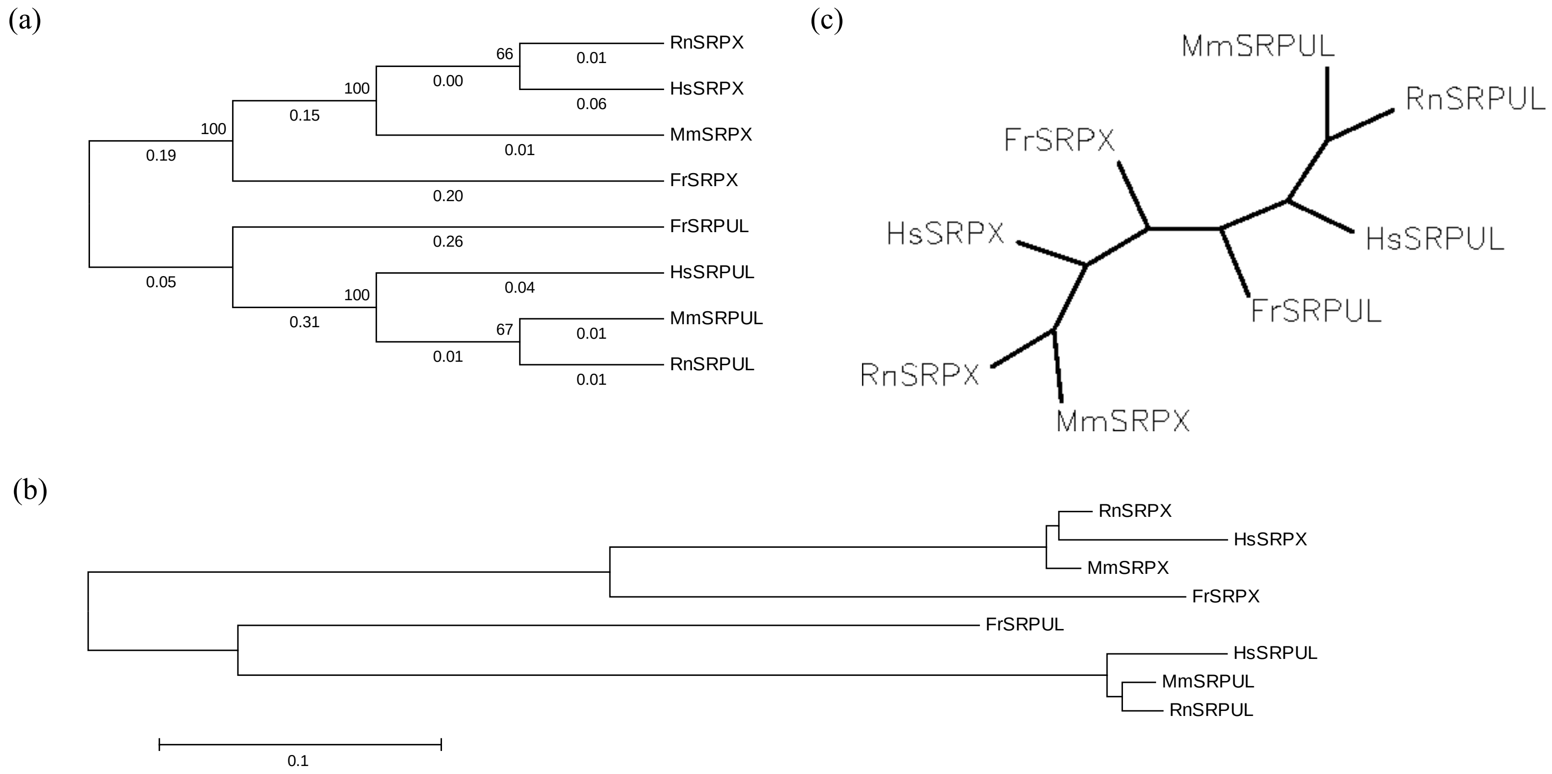


Figure 6-22 The figure shows phylogenetic trees constructed for the Sushi-repeat genes. (a) shows an un-rooted tree constructed using distance measurements. For clarity, only the topology is shown, with distance measurements for each branch shown below the branch and bootstrap support (% agreement from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

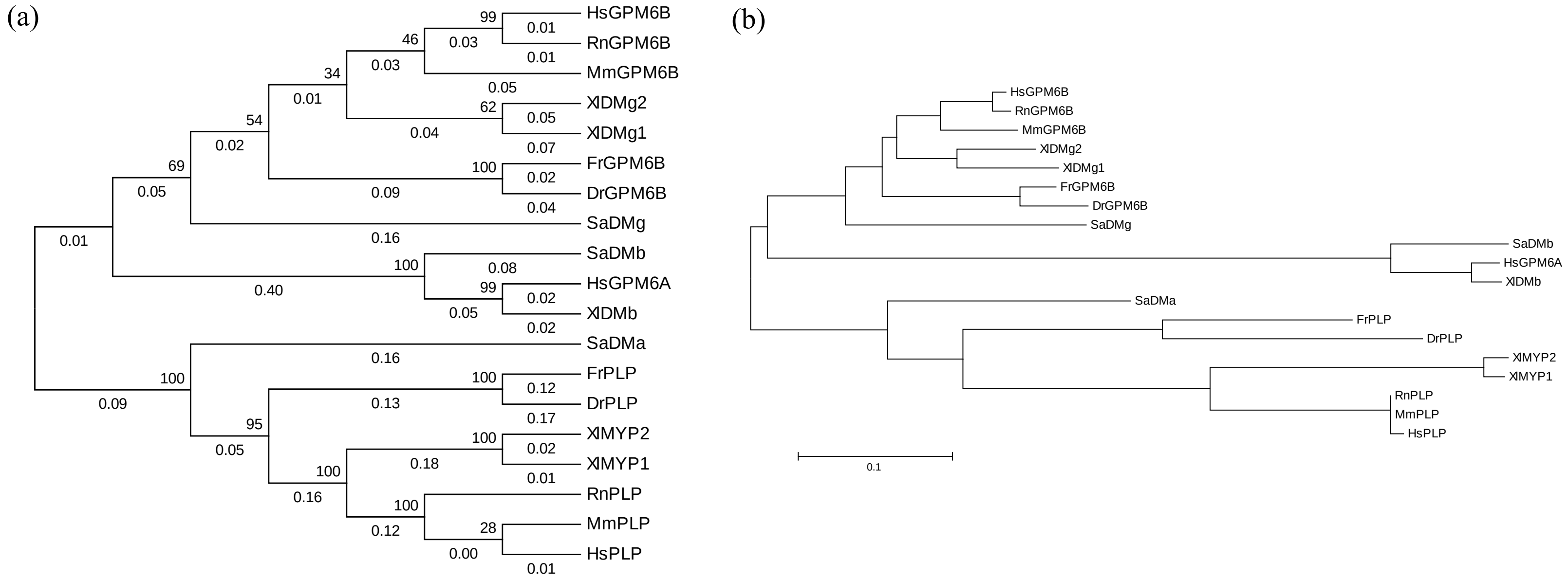


Figure 6-23 The figure shows phylogenetic trees constructed for the lipophilin genes. (a) shows an un-rooted tree constructed using distance measurements. For clarity, only the topology is shown, with distance measurements for each branch shown below the branch and bootstrap support (% agreement from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

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

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

6.6 Comparative analysis of *Sminthopsis macroura* genomic sequence

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

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

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

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

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

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

Table 6-14 Marsupial clone sequences and genes annotated.

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

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

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

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

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

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

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

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

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

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

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

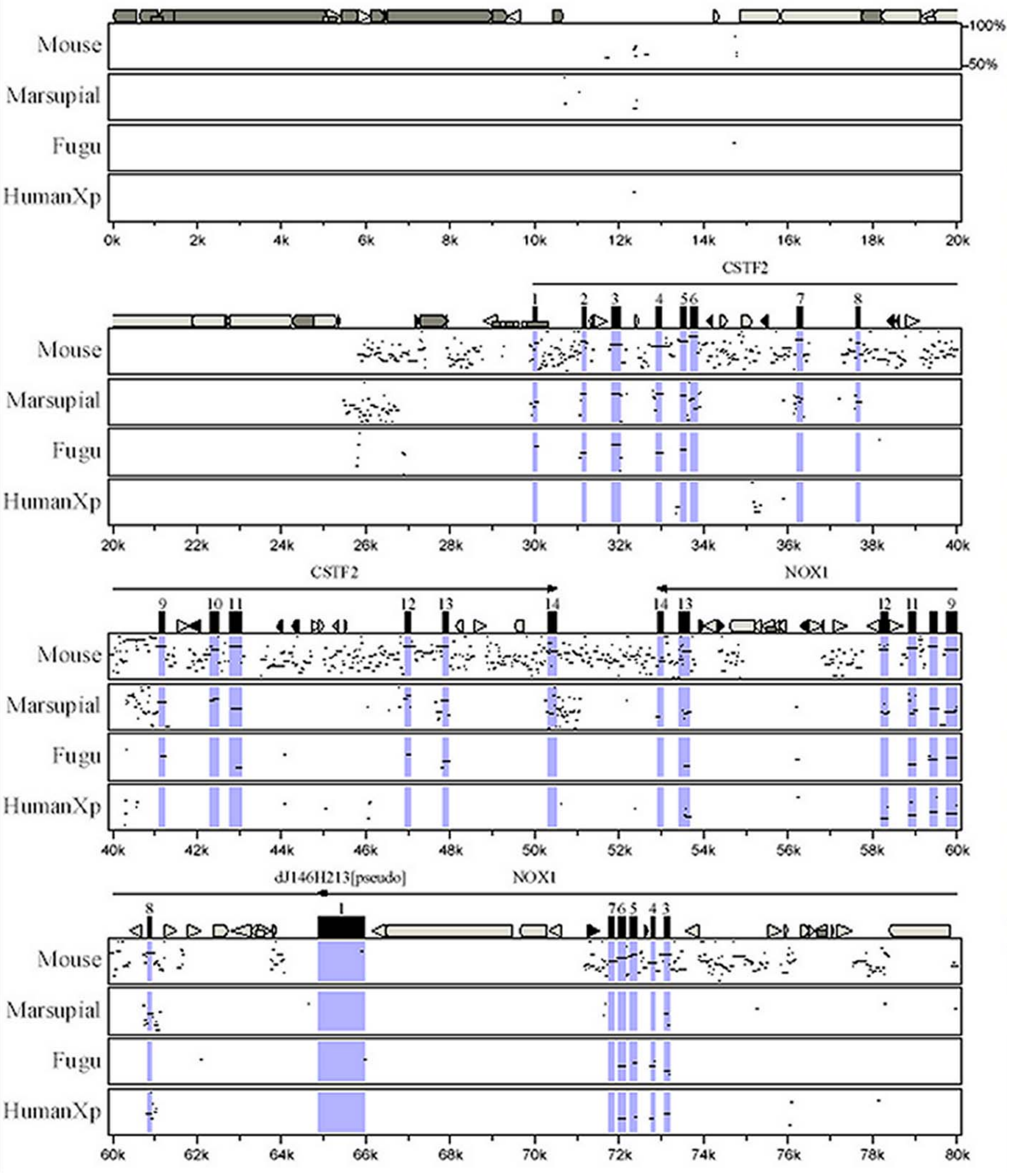
As the studies described in Section 6.5 have argued that the duplication leading to Xp and Xq paralogy occurred prior to human-*Fugu* divergence, and because NOX1 and XK-L were involved in the duplication, the human Xp paralogous region was also included in the comparative analysis. If the duplication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

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

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

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

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


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

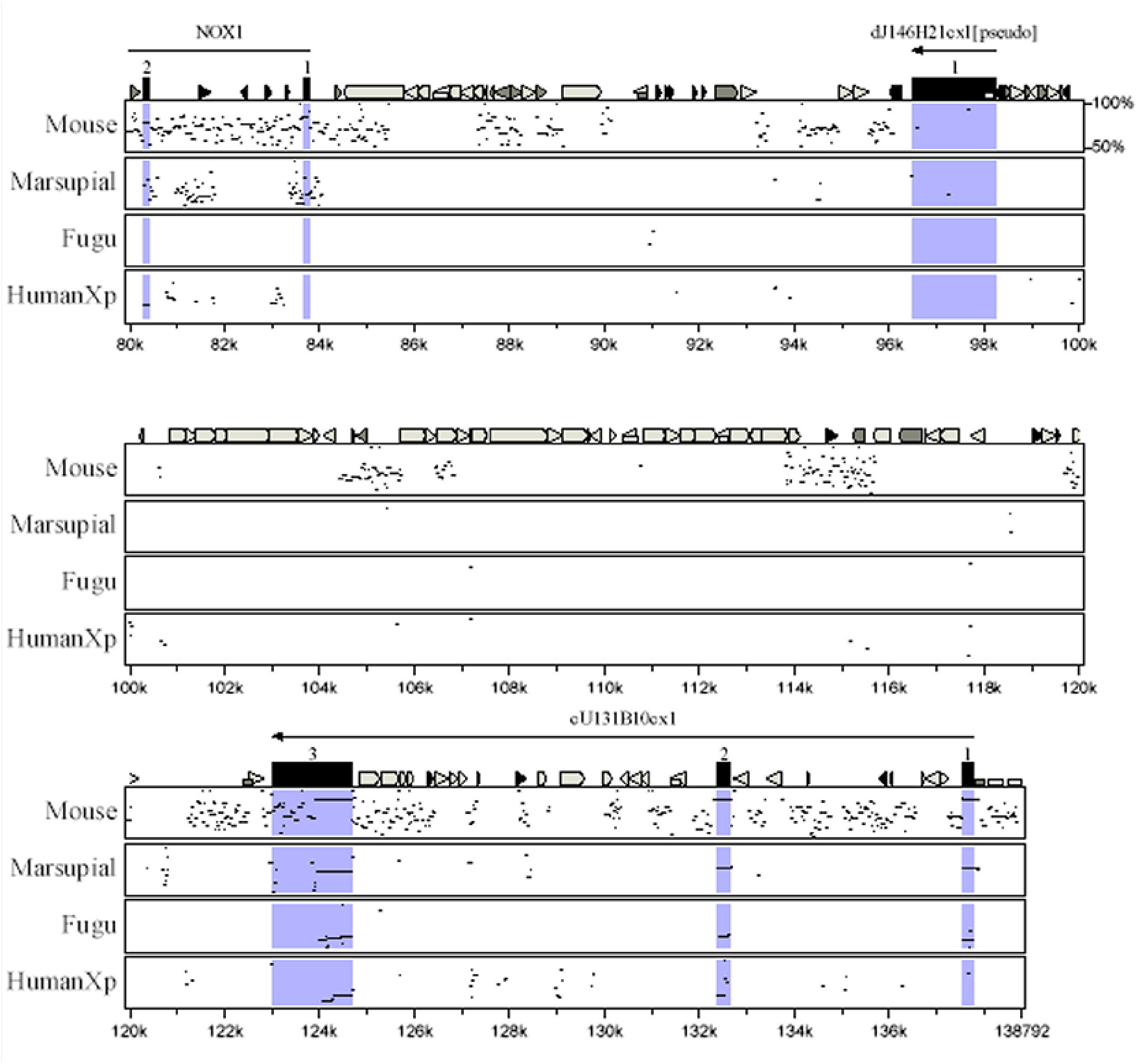
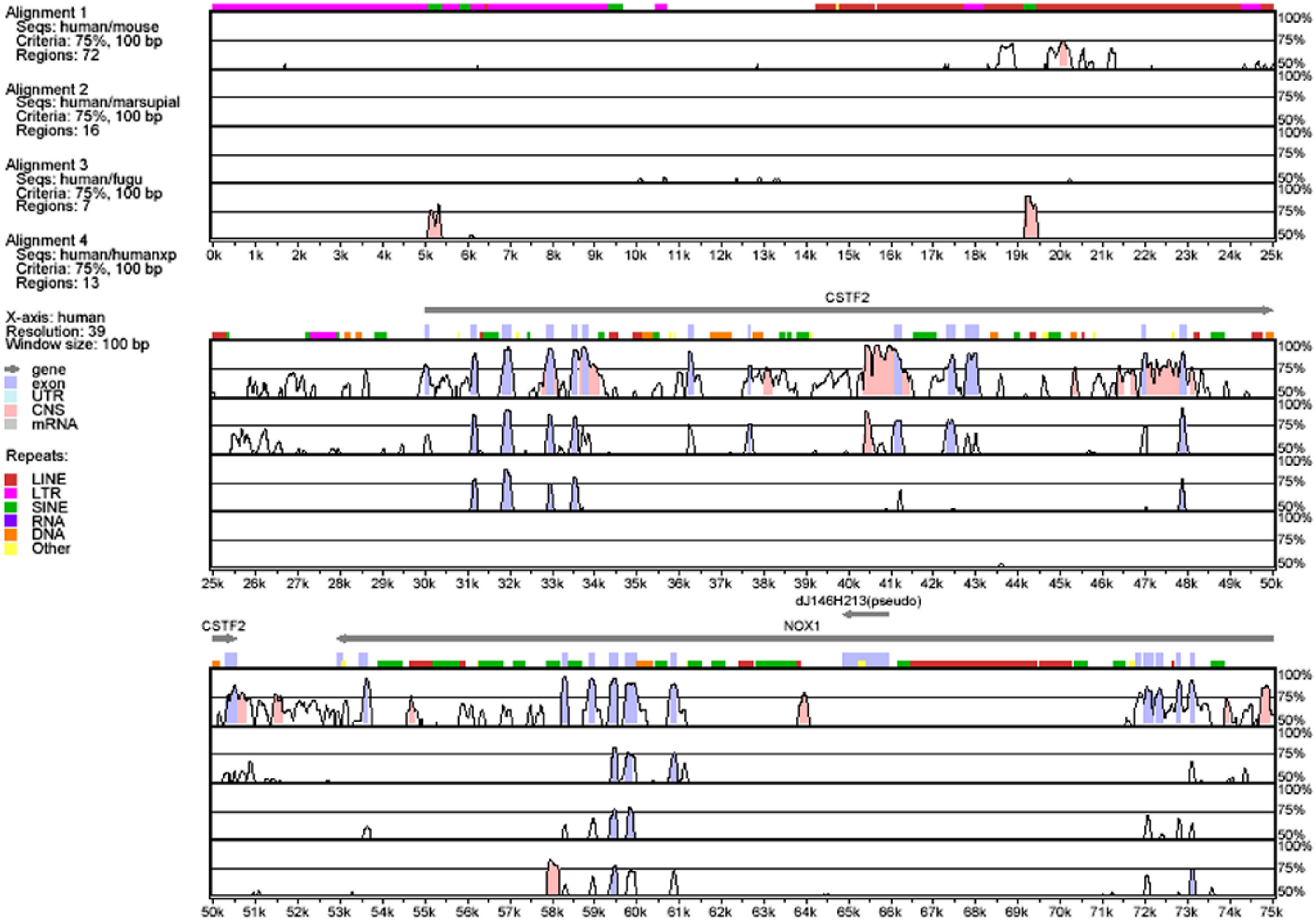


Figure 6-24 PIP plot of the human Xq22 region encompassing genes CSTF2, NOX1 and XK-L. Exonic regions are shaded blue and marked and numbered by vertical black boxes. Regions of high sequence identity to the orthologous mouse, marsupial and *Fugu* regions, and the paralogous Human Xp region, are depicted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


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

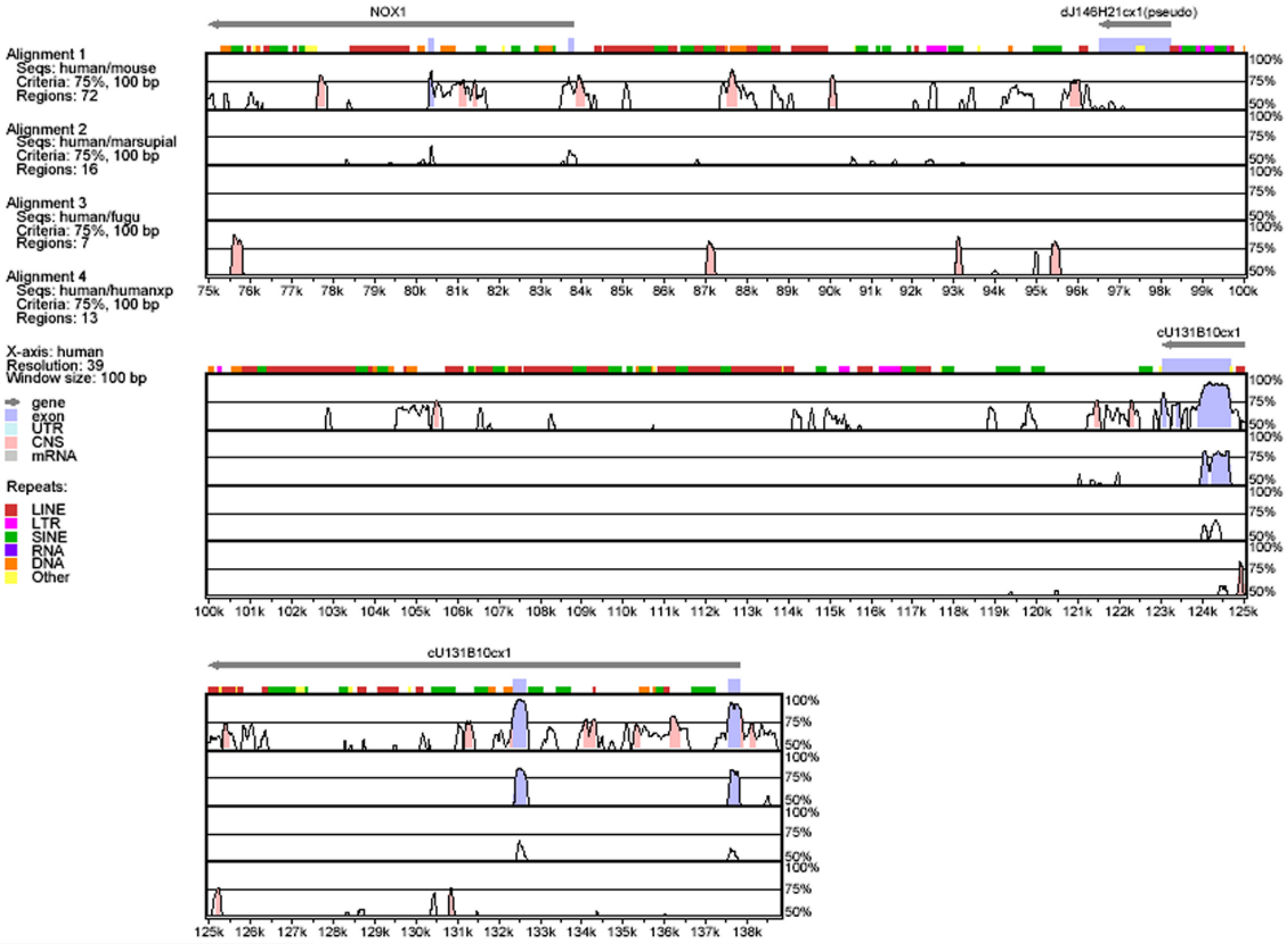


Figure 6-25 VISTA plot of the human Xq22 region encompassing genes CSTF2, NOX1 and XK-L. The figure legend is given in the diagrams. Regions of high sequence identity are depicted by blue peaks in the plot, with other regions of significant similarity shown as light-red peaks.

6.7 Discussion

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

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

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

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

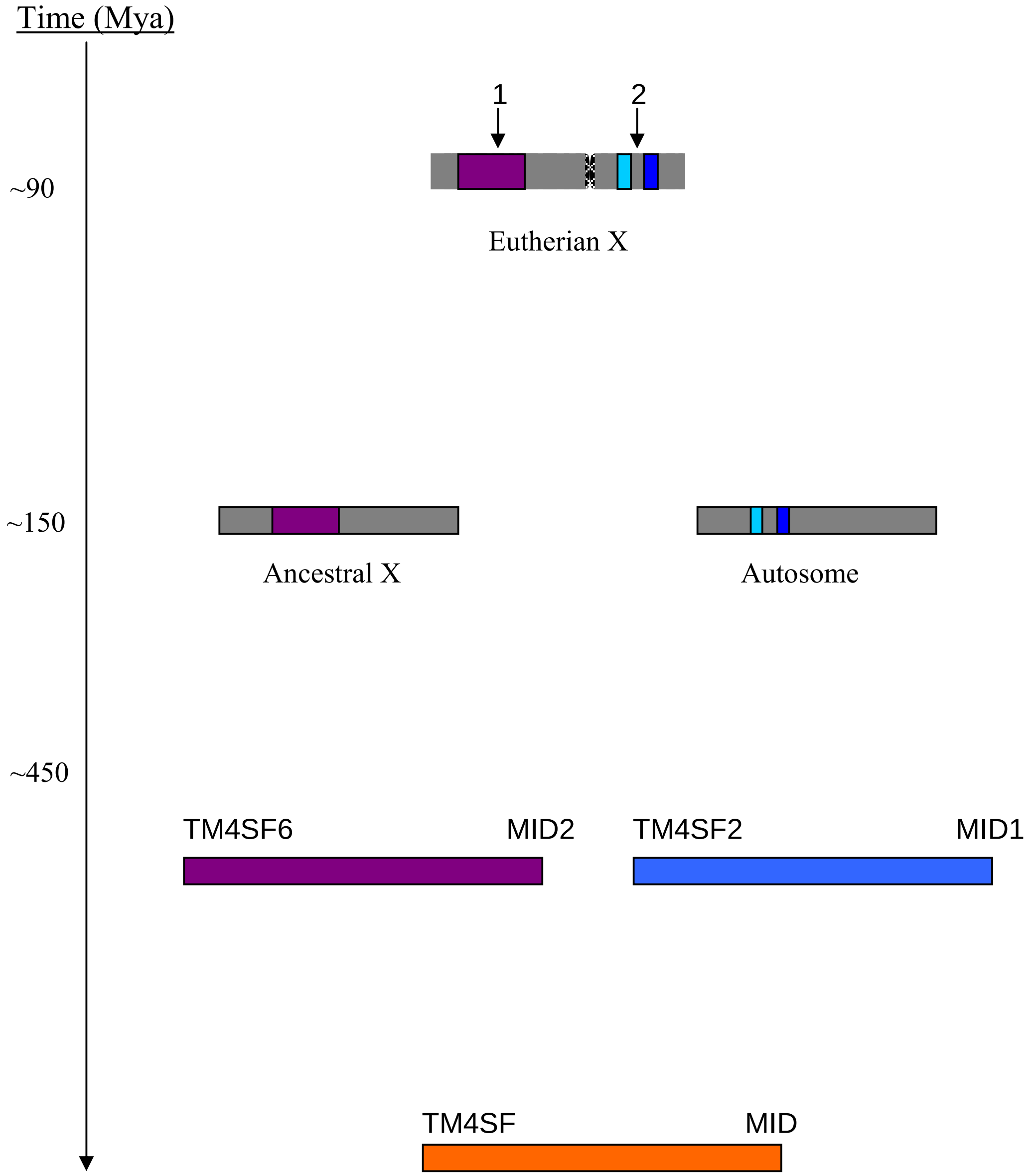


Figure 6-26 Diagram summarising analyses presented in this Chapter and providing a model for the evolution of the Xp/Xq paralogous regions. Duplication of an ancestral genomic segment (orange) generated two paralogous regions (purple and blue). These then diverged in composition, with one segment localising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

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

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

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

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

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

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

Chapter Seven - Discussion

7 Discussion

7.1 Summary of thesis results and discussion of major themes

The studies presented in this thesis were conducted over a period of rapid development in the field of genomics. Shortly after their beginning, the first complete sequence of a human chromosome was published (Dunham *et al.*, 1999). Shortly after their conclusion, the publication of the finished human genome sequence was announced. During that time other developments occurred, such as releases of sequences from large-scale cDNA sequencing projects, production of a physical map and draft sequence of the mouse genome, and the sequencing of genomes of organisms such as *Ciona intestinalis* (Dehal *et al.*, 2002), *Fugu rubripes* (Aparicio *et al.*, 2002) and *Arabidopsis thaliana* (Arabidopsis Genome initiative (2000)).

Chapter 3 described the production of a transcript map of the human Xq22-q23 region. This study provided an illustration of how effective the combination of the genomic sequence and results of large-scale cDNA sequencing projects is for gene-identification. Other aspects of genome structure only revealed through knowledge of the genomic sequence were also outlined. These included an inverted duplication with a very high level of sequence similarity, containing a gene from a family under intensive investigation, as well as an example of how the mitochondrial genome has integrated into the nuclear genome during evolution. Aspects of transcriptional control were briefly explored for genes with alternative polyadenylation sites.

Finally, and most importantly, the process of rigorous study of the genes in their genomic context revealed the presence of high numbers of duplicated genes, with some families contained within the region, some having paralogues on autosomes, and others sharing paralogy with genes on Xp. This finding stimulated the experimental and computational studies described in the subsequent chapters of the thesis, which were aimed at furthering our knowledge of the evolution of these gene families.

In Chapter 4, the generation of a sequence-ready BAC contig was described spanning *Mus musculus* X E3-F2, which is equivalent to human Xq22-q23. This comparative mapping was undertaken in order that the extent of sequence duplication within the corresponding region of the mouse genome could be assessed. It also contributed to efforts underway to sequence the whole mouse X chromosome. From analysis of the sequence generated from this region, it was found that many of the duplicated genes were also present in the mouse genome, indicating events which occurred before the divergence of the human and mouse lineages. Whilst many of the genes were conserved between human and mouse (as described in previous, lower-resolution studies), some differences were described including the KIR3DL1 gene, which is autosomal in human and rat. The mouse region studied also contained non-interspersed sequence repeats not detected in human.

In Chapter 5, the duplicated genes whose paralogues resided within Xq22 were studied in detail. Their degree of relatedness was assessed through sequence-based phylogenetic analysis, and their arrangement in comparison with the mouse X E3-F2 region was considered. In some cases, these analyses indicated orthologous relationships between human and mouse genes. However, in other cases, there were revealed potential examples of gene conversion between loci. This emphasised the benefit of being able to study the genes in their genomic context, revealing the details of paralogue proximity and orientation, and underscored problems with inferring orthology and paralogy based on sequence similarity alone. Studies of the expression patterns of the genes found patterns ranging from paralogues with similar, ubiquitous expression to those with more restricted patterns, with some paralogues showing differences within a family.

Overall, these studies described the striking degree of paralogy found within the Xq22 region and provide further avenues for targeted research into the evolution of the region and divergence of paralogue function and expression. As several of the genes across different families had been functionally characterised to some degree, this information can now also be used to focus studies for the various uncharacterised paralogues.

Chapter 6 presented evidence for a hypothesis suggesting a segmental duplication leading to the generation of paralogues with copies arrayed between the Xq22 region and Xp. This expanded further on previous observations and hypotheses, and the ability to examine the genes in their genomic context again proved its value in supporting the segmental duplication model. To provide further information on the evolutionary history of the Xp/Xq22 paralogous blocks, marsupial orthologues were identified and their genomic localisation determined. This confirmed previous studies which had shown that much of the region constituting human Xp is autosomal in marsupials (Glas et al., 1999), and demonstrated that many of the Xp paralogues were co-localised in the marsupial genome.

In order to try to estimate the minimum age of the duplication event leading to Xp/Xq22 paralogy, use was made of the recently completed draft sequence of *Fugu rubripes* and of the available literature for the paralogous genes. Phylogenetic and genomic organisation evidence suggested that the duplication occurred before the divergence of lineages leading to *Fugu* and humans approximately 430 Mya, and possibly before the divergence of cartilaginous and bony fish approximately 530 Mya. Comparative sequence analysis also supported this hypothesis, and demonstrated differences in sequence composition between the human Xp and Xq regions, and between marsupial, human and mouse.

An underlying theme throughout these studies has been illustration of the benefit of the availability of genomic sequence, particularly long-length finished sequence. Initially providing a basis for a comprehensive transcript map of a genomic region, the ability to see genes in their genomic context revealed aspects of the chromosome's evolution and biology that may not have been otherwise observed. The context information in itself provided evidence supporting a hypothesis for a model of evolution leading to generation of paralogy between Xp and Xq22.

The extent to which a gene identification strategy using mRNA and genomic sequence can be implemented is naturally limited by the availability of both types of data. The human genome sequence is now almost complete, and the efforts of several large-scale cDNA sequencing projects have provided large amounts of useful data, from which many genes have been identified and annotated. These have added to data on specific genes generated over the years by many investigators.

A limitation of any cDNA-based approach is the nature of generation of the cDNA, from RNA derived from tissue samples. The availability of different tissue types can be limiting, especially in humans. For example the cDNA libraries employed in gene identification in Chapter 3 omit different tissues that may express a putative gene. The sensitivity of RNA to degradation and the nature of steps involved in cloning of cDNAs can also result in incomplete cDNA representation of the mRNAs from a tissue. Finally, some mRNAs may have temporally-restricted expression which again may result in incomplete mRNA representation.

These limitations may be partially circumvented by the availability of genomic and mRNA sequence information from other organisms. As discussed in earlier chapters, this information can be used to identify conserved regions of genomic sequence, suggesting functional roles for these sequences (Gottgens *et al.*, 2002). For example, in Chapter 4, comparisons were based on gene annotation conducted independently in human and mouse. A direct comparison of the two regions may reveal further conserved regions indicative of genes. Tools such as TWINSCAN (Korf *et al.*, 2001) could be implemented in such an approach. Tissue availability is less of an issue with model organisms and a more comprehensive array of RNA samples can be accessed, thus leading to a better representation of the transcriptome.

A limiting factor in genomic sequence comparisons for different organisms is the degree of relatedness of the organisms being compared. If the organisms' genomes are not sufficiently divergent, the conserved regions may be hidden by "noise". This is illustrated in the comparative analysis of the CSTF2/NOX1/XK-L loci described in Chapter 6. If the genomes are too divergent, though, identification of conserved sequences will be problematic, especially when employing algorithms designed to analyse large amounts of sequence which may have to trade sensitivity for pragmatic reasons of computational intensity.

Gene annotation has been pioneered on organisms such as yeast, bacteria and worms, and has recently been applied on a genome-wide scale for a variety of higher organisms through initiatives such as the Ensembl project (EBI and Wellcome Trust Sanger Institute), Genome Browser (University of California at Santa Cruz) and genome resources at the NCBI (NIH). These initiatives have addressed the issues of tracking draft and finished genomic sequence records and their subsequent revisions, analysing the sequence using a variety of gene and repeat identification algorithms and combining results of similarity searches of mRNA and protein databases with the genomic sequences. Incorporation of other sources of data such as SNPs is also performed, and all of these data are made available via searchable graphical interfaces.

These initiatives have made great progress in facilitating use of the genomic sequence information by investigators worldwide. Indeed, anybody with access to the internet can view the genomic sequences and their annotation for organisms such as human, mouse, fruitfly and pufferfish, and use was made of this in the studies described in this thesis (for example the *Fugu* analysis in Chapter 6). A more complete annotation of genome sequences combines approaches such as these with manual annotation. For example, algorithms used to align mRNA sequences to genomic sequence can miss very small exons, and can fail to identify splice sites correctly. Assessment of how complete a gene structure appears to be is also best achieved through manual inspection. Finally, features such as unusual gene structures, as seen for the NXF2 gene (Chapter 3), and context-dependent features such as high levels of paralogy (Chapters 3-6) may be missed by purely computational approaches.

A combined approach to human genome annotation is now being employed at different centres, such as by the HAVANA (Human And Vertebrate ANalysis and Annotation) Group at the Wellcome Trust Sanger Institute. Such annotations are being collated in databases such as the VEGA (VErtebrate Genome Annotation) database. Gene annotation is also being complemented by targeted gene identification efforts for human, similarly to those described in Chapter 3, by the EGAG (Experimental Gene Annotation Group) group at the Wellcome Trust Sanger Institute, and by large-scale mRNA sequencing efforts as described in Chapter 1.

Combined with large-scale cDNA sequencing data and comparative analysis of genomes of different organisms, this provides a powerful approach for the comprehensive annotation of genes within a genome.

Knowledge of gene structure is of interest for several reasons. For one, it provides insight into how genes have evolved in different organisms under different evolutionary pressures. A key question in genetics remains as to how genes have evolved and the mechanisms of their transcription. Were introns always a feature of gene structure or have they evolved from precursors? The “intron-early” and “intron-late” hypotheses are still the subject of debate. Elucidation of the gene structures of all genes within the genome of an organism will provide comprehensive information regarding ranges of exon and intron sizes, and how they vary within different regions of the genome with different sequence compositions. Within the human genome for example, introns appear to be longer in Giemsa dark-band regions.

It has also been suggested that intron-length may be related to level of transcription of the gene. This would be logical, as it would take less time and energy to transcribe a shorter gene, secondary structure effects notwithstanding. Indeed, knowledge of gene structure is fundamental to studies aimed at understanding gene transcription, as *cis*-acting elements for transcription control such as promoters and enhancers are not fully represented in the transcriptome (although such elements may occur also in exonic and intronic regions).

It is also apparent that we probably still have much to learn regarding interpretation of genes by the process of transcription. The recently discovered process of trans-splicing in *Caenorhabditis elegans* (Blumenthal, 1995) and the growing research into the process of RNA-editing have served to remind us of this, in addition to active research into transcription control and mRNA localisation elements contained within untranslated portions of RNA transcripts. Gene structures provide a framework on which to base these studies, and the genomic sequence provides the information which contains within it the cis-acting elements involved in transcription.

Knowledge of gene structures is also valuable in studies attempting to identify genes involved in genetic disorders, and genetic differences between individuals giving rise to different traits. In order to focus methods for detection of genetic differences, information regarding exon/intron structure is invaluable in designing assay reagents (for example primers designed to amplify an exonic region for sequencing).

Finally, gene structure information is useful in identifying genes that have shared a common ancestor prior to gene duplication events, particularly in instances where the sequences have diverged sufficiently to be in the “grey-area” of homology. In such instances, aspects of gene structure such as exon sizes and intron phase can provide evidence of common origin. Exon size information provided compelling evidence for the paralogy discussed in Chapter 6.

The main aspect of this thesis has centred on the discovery and characterisation of extensive paralogy within Xq22 and between Xp and Xq. Studies of gene duplication have benefited greatly from the large amounts of genomic sequence now available. Although questions remain as to predominant mechanisms of duplication and functional importance of paralogues, it is clear that gene duplication is an important feature of all organisms whose genomes have been studied to date, from bacteria to humans.

One intriguing feature of gene duplication in the human genome is the apparent heterogeneity of the distribution of gene duplications. To date, the most intensively studied regions of gene duplications have been the Hox gene clusters and the MHC regions. Prior to elucidation of the genome sequence and annotation of genes, it would not be clear if gene duplications were indeed enriched in these regions, or if study of these regions had led to an ascertainment bias.

The gene duplications within Xq22 were striking in that in adjacent regions of the chromosome, multiple duplicated genes were not noted. However, the X chromosome does in fact contain a variety of duplicated loci (e.g. MAGE genes). It is possible that the X chromosome may be enriched for segmental duplications as it is largely unpaired at meiosis in males, and there may arise an opportunity for increased rearrangement. Complete annotation of the human X chromosome and comparative analysis with other organisms may shed light on this hypothesis.

Gene duplications provide challenges regarding assessment of function of gene products. In cases where sequence similarity is high, assay platforms such as hybridisation techniques may provide “composite” information regarding gene expression, for example. Mutation screening of genes may also be affected depending on the strategy used. Knowledge of the existence of paralogues provides useful information to the investigator in this regard. Examples where this could be an issue have been described in this thesis. In particular, the two thymosin-beta genes within Xq22 encode identical proteins, and any differences between the loci conferring selective advantages to the retention of both gene copies would more likely act at the level of transcription of the genes. Alternatively there may be selective advantages in keeping both copies similar, perhaps via gene conversion. Although the mechanisms of gene loss are incompletely understood and are also factors in retention of duplicated loci, these two loci appear to have been conserved over at least 90 million years of evolution.

In summary, sequence duplications are an indication of the dynamic nature of a genome, and how it may adapt to different selective pressures. Availability of genomic sequence allows careful and comprehensive study of these loci, which may otherwise be refractory to investigation, and provides important information regarding their context and distribution within the genome of an organism.

7.2 Future directions

Future directions for gene identification and annotation have been partly discussed in earlier chapters and sections. It is unlikely that there will be major advances in overcoming the inherent difficulties of working with human RNA samples, but large-scale sequencing efforts or SAGE analysis (Sun *et al.*, 2004) of cDNA libraries from a range of different tissues will continue to provide valuable resources. These will also aid in unravelling complexities of transcription, such as alternative splicing of pre-mRNAs. Whilst EST sequences can provide useful evidence of alternative splicing events, for full elucidation of the transcript sequence and hence gene structure they are often too short to allow definitive conclusions to be drawn. The cDNA clones themselves will also provide a resource for functional studies of the gene products, as in the *C. elegans* “ORFeome” effort (Reboul *et al.*, 2003).

One area of gene identification that would particularly benefit from development of methodology is in the representation of the 5’-ends of transcripts. It has proved difficult to ensure faithful representation of the beginning of an mRNA transcript for various technical reasons. However, representation of the 5’ end of a transcript is crucial in the annotation of core promoter regions. Truncated transcripts may result in erroneous annotation of promoters. Some advances have been made in this area, but it still remains a difficult topic and will benefit from continued research.

To fully characterise the genes of an organism, aspects such as alternative splicing, RNA-editing, alternative polyadenylation site usage, alternative promoters, enhancers, mRNA localisation signals and sequence elements affecting mRNA turnover and translation all need to be taken into account. This is in addition to any other protein-related aspects to be considered, such as post-translational modification. It is clear that there is still much to be investigated in these areas. As genes are studied further in different organisms, we will learn more regarding the use of different aspects of genetic organisation and control in different species, and those elements that are shared.

The study of gene duplications will benefit greatly from increasing genome mapping and sequencing of a variety of organisms from different evolutionary lineages and with different degrees of divergence. Gene duplication, via single gene tandem duplications and segmental duplication, and gene loss are balancing factors in shaping the genome of an organism. Extended knowledge of the genes within different organisms will aid in the understanding of the relative balances of these factors at different loci. Species-specific differences can also then be taken into account when attempting to elucidate the molecular evolution of paralogues and their functions, and in interpreting data derived from animal models.

Comprehensive annotation of highly related loci will also provide useful data for the experimental studies addressing functions of these loci. It may avoid confusion due to “composite” data being produced from highly related loci, and will highlight cases of possible functional redundancy. It will also provide a framework on which to base studies of differences in copy number of paralogous loci between individuals. As SNP data become available, this would be complemented by large-scale analysis of the gene complement of individuals, which may contribute to different traits, disorders, susceptibility to disease and response to drug treatment. It would be interesting to assess the composition of the Xq22 region in different individuals as an example.

As annotation and characterisation of paralogues within a genome progresses, studies of why some regions are rich in gene duplications will be aided. For example, why should so many duplicate genes be found within Xq22? What features of the genome have affected the evolutionary dynamics to cause this? How stable is Xq22 (and its mouse counterpart)? In this regard, it is of interest that the proteolipid protein (PLP) gene in Xq22, is often duplicated in Pelizaeus-Merzbacher disease. Whilst this may be unconnected to the presence of multiple repeat families (some of which are non-genic) within the region, the repeats may predispose to genetic rearrangement. In some cases however the relevant sequence features may have become unrecognisable if the gene duplications occurred a long time ago. The process of gene conversion, possible examples of which are also seen within Xq22, also clearly has to be taken into account in these studies. Analysis of the genes in other primates may shed light on whether this is the case (Rozen *et al.*, 2003).

The evidence provided for the segmental duplication generating the paralogues seen between Xp and Xq22 has raised further questions related to X chromosome biology and evolution. We have demonstrated that the segmental duplication appears to be a relatively ancient event, and that marsupial orthologues of the Xp paralogues are autosomal and co-localised (for those studied). Following the duplication, one copy (represented at Xq22) was localised on the autosome that in mammals became the ancestral X chromosome, whilst the second copy (represented at Xp) remained autosomal. It is unclear what the initial arrangement of the regions following the duplication was, and what rearrangements occurred subsequently.

It is also unclear whether the region of extensive paralogy *within* Xq22, which is flanked by genes with paralogues on Xp, was also part of the segmental duplication. A similar question remains as to the origin of the large non-paralogous region separating the two regions of Xp paralogues. Were these regions originally present in the duplication and have subsequently undergone different rates of gene loss or rearrangement, or have they been acquired more recently by the paralogous regions? The large number of non-paralogous genes interspersed throughout the blocks of paralogy are part of this same question. Given the apparent considerable age of the duplication, there have presumably been large numbers of both gene gains and losses which have degraded the original segments. Further study of these regions in the marsupial and more distantly related organisms may provide useful information to address these questions.

Finally, whilst initial translocation of the Xp paralogues to the X chromosome from an autosome would not result in problems of dosage if they were also transferred to the Y (i.e. into an existing pseudoautosomal region), subsequent degradation of their copies on the Y would result in dosage imbalance between sexes unless they became recruited into the X inactivation system. This is an issue relevant to any genes translocated to the sex chromosomes. It is unclear what the inactivation statuses of many of the genes are at present. If inactivated, what features did they gain (or lose?) in order to accomplish this?

As the X chromosome sequence and annotation approaches completion, the extent of paralogy within the chromosome can be fully assessed, and compared to other regions of the genome. It also paves the way for further studies of individual variation, gene functions and chromosome evolution. This thesis has illustrated how the genome sequence can be used to view not only the content of genomic regions, but also their contexts. Both perspectives are required to begin to understand fully the biology of chromosomes and the evolution that has shaped them. It is also clear from the studies presented here that there is much still to be explored in this area, and our understanding is far from complete. The unique biology and structure of the X chromosome ensures that the process of addressing these questions will be a fascinating one.

References

Achaz, G., Netter, P., *et al.* (2001). Study of intrachromosomal duplications among the eukaryote genomes. *Mol Biol Evol* **18:** 2280-8.

Adams, M. D., Celniker, S. E., *et al.* (2000). The genome sequence of Drosophila melanogaster. *Science* **287:** 2185-95.

Agundez, J. A., Gallardo, L., *et al.* (2001). Functionally active duplications of the CYP2D6 gene are more prevalent among larynx and lung cancer patients. *Oncology* **61:** 59-63.

Allcock, R. J., Atrazhev, A. M., *et al.* (2002). The MHC haplotype project: a resource for HLA-linked association studies. *Tissue Antigens* **59:** 520-1.

Antequera, F. and Bird, A. (1993). Number of CpG islands and genes in human and mouse. *Proc Natl Acad Sci U S A* **90:** 11995-9.

Aparicio, S., Chapman, J., *et al.* (2002). Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* **297:** 1301-10.

Arabidopsis Genome Initiative. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408:** 796-815.

Averof, M. (2002). Arthropod Hox genes: insights on the evolutionary forces that shape gene functions. *Curr Opin Genet Dev* **12:** 386-92.

Avner, P., Bruls, T., *et al.* (2001). A radiation hybrid transcript map of the mouse genome. *Nat Genet* **29:** 194-200.

Avner, P., Heard, E. (2001). X-chromosome inactivation: counting, choice and initiation. *Nat. Rev. Genet.* **2**:59-67.

Bailey, J. A., Gu, Z., *et al.* (2002). Recent segmental duplications in the human genome. *Science* **297:** 1003-7.

Batzoglou, S., Jaffe, D. B., *et al.* (2002). ARACHNE: a whole-genome shotgun assembler. *Genome Res.* **12:** 177-89.

Bentley, D. R., Todd, C., *et al.* (1992). The development and application of automated gridding for efficient screening of yeast and bacterial ordered libraries. *Genomics* **12:** 534-41.

Bentley, D. R., Deloukas, P., *et al.* (2001). The physical maps for sequencing human chromosomes 1, 6, 9, 10, 13, 20 and X. *Nature* **409:** 942-3.

Bernardi, G., Olofsson, B., *et al.* (1985). The mosaic genome of warm-blooded vertebrates. *Science* **228:** 953-8.

Blacket, M. J., Krajewski, C., *et al.* (1999). Systematic relationships within the dasyurid marsupial tribe *Sminthopsini*--a multigene approach. *Mol Phylogenet Evol* **12:** 140-55.

Blumenthal, T. (1995). Trans-splicing and polycistronic transcription in *Caenorhabditis elegans*. *Trends Genet* **11:** 132-6.

Brown, A. L. and Kay, G. F. (1999). Bex1, a gene with increased expression in parthenogenetic embryos, is a member of a novel gene family on the mouse X chromosome. *Hum Mol Genet* **8:** 611-9.

Brown, C. J., Ballabio, A., *et al*., (1991). A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. *Nature* **349**: 38-44.

Brown, C. J., Greally, J. M., (2003). A stain upon the silence: genes escaping X inactivation. *Trends Genet.* **19**: 432-438.

Bult, C. J., White, O., *et al.* (1996). Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science* **273:** 1058-73.

Burge, C. and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *J Mol Biol* **268:** 78-94.

Bussey, H., Storms, R. K., *et al.* (1997). The nucleotide sequence of *Saccharomyces cerevisiae* chromosome XVI. *Nature* **387:** 103-5.

Carninci, P., Shibata, Y., *et al.* (2000). Normalization and subtraction of cap-trapper-selected cDNAs to prepare full-length cDNA libraries for rapid discovery of new genes. *Genome Res* 10**:** 1617-30.

Chapman, M. A., Charchar, F. J., *et al.* (2003). Comparative and functional analyses of LYL1 loci establish marsupial sequences as a model for phylogenetic footprinting. *Genomics* **81:** 249-59.

Chowdhary, B. P., Raudsepp, T., *et al.* (1998). Emerging patterns of comparative genome organization in some mammalian species as revealed by Zoo-FISH. *Genome Res* **8:** 577-89.

Chudley, A. E., Tackels, D. C., *et al.* (1999). X-linked mental retardation syndrome with seizures, hypogammaglobulinemia, and progressive gait disturbance is regionally mapped between Xq21.33 and Xq23. *Am J Med Genet* **85:** 255-62.

Cohen, D., Chumakov, I., *et al.* (1993). A first-generation physical map of the human genome. *Nature* **366:** 698-701.

Collins, J. and Bruning, H. J. (1978). Plasmids useable as gene-cloning vectors in an *in vitro* packaging by coliphage lambda**:** cosmids. *Gene* **4:** 85-107.

Collins, J. E., Cole, C. G., *et al.* (1995). A high-density YAC contig map of human chromosome 22. *Nature* **377:** 367-79.

Crnogorac-Jurcevic, T. and Brown, J. R. (1997). Tetraodon fluviatilis, a new puffer fish model for genome studies. *Genomics* **41:** 177-84.

Davuluri, R. V., Grosse, I., *et al.* (2001). Computational identification of promoters and first exons in the human genome. *Nat Genet* **29:** 412-7.

De Leo, A. A., Guedelha, N., *et al.* (1999). Comparative chromosome painting between marsupial orders**:** relationships with a 2n = 14 ancestral marsupial karyotype. *Chromosome Res* **7:** 509-17.

Dear, P. H., Cook, P. R. (1989). Happy mapping: a proposal for linkage mapping the human genome. *Nucleic Acids Research* **17**: 6795-6807

Dehal, P., Satou, Y., *et al.* (2002). The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. *Science* **298:** 2157-67.

Deloukas, P., Schuler, G. D., *et al.* (1998). A physical map of 30,000 human genes. *Science* **282:** 744-6.

Dietrich, W. F., Miller, J., *et al.* (1996). A comprehensive genetic map of the mouse genome. *Nature* **380:** 149-52.

Donis-Keller, H., Green, P., *et al.* (1987). A genetic linkage map of the human genome. *Cell* **51:** 319-37.

Down, T. A. and Hubbard, T. J. (2002). Computational detection and location of transcription start sites in mammalian genomic DNA. *Genome Res* **12:** 458-61.

Dunham, I., Shimizu, N., *et al.* (1999). The DNA sequence of human chromosome 22. *Nature* **402:** 489-95.

Duyk, G. M., Kim, S. W., *et al.* (1990). Exon trapping: a genetic screen to identify candidate transcribed sequences in cloned mammalian genomic DNA. *Proc Natl Acad Sci U S A* **87:** 8995-9.

Emanuel, B. S. and Shaikh, T. H. (2001). Segmental duplications**:** an 'expanding' role in genomic instability and disease. *Nat Rev Genet* **2:** 791-800.

Faria, T. N., LaRosa, G. J., *et al.* (1998). Characterization of genes which exhibit reduced expression during the retinoic acid-induced differentiation of F9 teratocarcinoma cells: involvement of cyclin D3 in RA-mediated growth arrest. *Mol Cell Endocrinol* **143:** 155-66.

Fleischmann, R. D., Adams, M. D., *et al.* (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269:** 496-512.

Force, A., Lynch, M., *et al.* (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151:** 1531-45.

Friedman, R. and Hughes, A. L. (2001). Gene duplication and the structure of eukaryotic genomes. *Genome Res* **11:** 373-81.

Frints, S. G. *et al.* (2002). X-linked mental retardation: vanishing boundaries between non-specific (MRX) and syndromic (MRXS) forms. *Clin. Genet.* **62**: 423-432.

Frohman, M. A., Dush, M. K., *et al.* (1988). Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer. *Proc Natl Acad Sci U S A* **85:** 8998-9002.

Gibbs, R. A., Belmont, J. W., *et al.* (2003). The international HapMap project. *Nature* **426:** 789-96.

Glas, R., Marshall-Graves, J. A., *et al.* (1999). Cross-species chromosome painting between human and marsupial directly demonstrates the ancient region of the mammalian X. *Mamm Genome* **10:** 1115-6.

Gottgens, B., Barton, L. M., *et al.* (2002). Transcriptional regulation of the stem cell leukemia gene (SCL) - comparative analysis of five vertebrate SCL loci. *Genome Res* **12:** 749-59.

Gow, A. (1997). Redefining the lipophilin family of proteolipid proteins. *J Neurosci Res* **50:** 659-64.

Graves, J. and Westerman, M. (2002). Marsupial genetics and genomics. *Trends Genet* **18:** 517.

Gregory, S. G., Howell, G. R., *et al.* (1997). Genome mapping by fluorescent fingerprinting. *Genome Res* **7:** 1162-8.

Gregory, S. G., Sekhon, M., *et al.* (2002). A physical map of the mouse genome. *Nature* **418:** 743-50.

Grutzner, F., Crollius, H. R., *et al.* (2002). Four-hundred million years of conserved synteny of human Xp and Xq genes on three tetraodon chromosomes. *Genome Res* **12:** 1316-22.

Gu, X., Wang, Y., *et al.* (2002). Age distribution of human gene families shows significant roles of both large- and small-scale duplications in vertebrate evolution. *Nat Gen* **31:** 205-9.

Guigo, R., Agarwal, P., *et al.* (2000). An assessment of gene prediction accuracy in large DNA sequences. *Genome Res* **10:** 1631-42.

Herold, A., Suyama, M., *et al.* (2000). TAP (NXF1) belongs to a multigene family of putative RNA export factors with a conserved modular architecture. *Mol Cell Biol* **20:** 8996-9008.

Hescheler, J. and Schultz, G. (1993). G-proteins involved in the calcium channel signalling system. *Curr Opin Neurobiol* **3:** 360-7.

Holland, J., Coffey, A. J., *et al.* (1993). Vertical integration of cosmid and YAC resources for interval mapping on the X-chromosome. *Genomics* **15:** 297-304.

Howell, G. R., (2002). Physical, transcriptional and comparative mapping on the human X chromosome. PhD thesis, The Open University.

Huang, S.-H., Yang, A. Y, *et al.* (1993). Amplification of genes ends from gene libraries by polymerase chain reaction with single-sided specificity. In *Methods in Molecular Biology, PCR Protocols: Current Methods and Applications*. (B. A. White, ed.). Humana Press, Totowa, New Jersey. pp 357-363.

Hudson, T. J., Church, D. M., *et al.* (2001). A radiation hybrid map of mouse genes. *Nat Genet* **29:** 201-5.

Huff, T., Muller, C. S., *et al.* (2001). beta-Thymosins, small acidic peptides with multiple functions. *Int J Biochem Cell Biol* **33:** 205-20.

Inoue, K., Osaka, H., *et al.* (1996). A duplicated PLP gene causing Pelizaeus-Merzbacher disease detected by comparative multiplex PCR. *Am J Hum Genet* **59:** 32-9.

Jurka, J. (2000). Repbase update**:** a database and an electronic journal of repetitive elements. *Trends Genet* **16:** 418-20.

Kikuno, R., Nagase, T., *et al.* (2002). HUGE**:** a database for human large proteins identified in the Kazusa cDNA sequencing project. *Nucleic Acids Res* **30:** 166-8.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol* **16:** 111-20.

Kitagawa, K., Sinoway, M. P., *et al.* (1993). A proteolipid protein gene family: expression in sharks and rays and possible evolution from an ancestral gene encoding a pore-forming polypeptide. *Neuron* **11:** 433-48.

Kong, A., Gudbjartsson, D. F., *et al.* (2002). A high-resolution recombination map of the human genome. *Nat Genet* **31:** 241-7.

Konietzko, U. and Kuhl, D. (1998). A subtractive hybridisation method for the enrichment of moderately induced sequences. *Nucleic Acids Res* **26:** 1359-61.

Korf, I., Flicek, P., *et al.* (2001). Integrating genomic homology into gene structure prediction. *Bioinformatics* **17:** S140-8.

Kumar, S. and Hedges, S. B. (1998). A molecular timescale for vertebrate evolution. *Nature* **392:** 917-20.

Kumar, S., Tamura, K., *et al.* (2001). MEGA2**:** molecular evolutionary genetics analysis software. *Bioinformatics* **17:** 1244-5.

Kurochkin, I. V., Yonemitsu, N., *et al.* (2001). ALEX1, a novel human armadillo repeat protein that is expressed differentially in normal tissues and carcinomas. *Biochem Biophys Res Commun* **280:** 340-7.

Lander, E. S., Linton, L. M., *et al.* (2001). Initial sequencing and analysis of the human genome. *Nature* **409:** 860-921.

Li, X., Zimmerman, A., *et al.* (1996). The mouse thymosin beta 4 gene: structure, promoter identification, and chromosome localization. *Genomics* **32:** 388-94.

Lovett, M., Kere, J., *et al.* (1991). Direct selection: a method for the isolation of cDNAs encoded by large genomic regions. *Proc Natl Acad Sci U S A* **88:** 9628-32.

Lyon, M. F. (1998). X-chromosome inactivation: a repeat hypothesis. *Cytogenet Cell Genet* **80:** 133-7.

Lyon, M. F. (1999). X-chromosome inactivation. *Curr. Biol.* **9**: R235-R237

Ma, Y., Zhang, S., *et al.* (2002). Molecular characterization of the TCP11 gene which is the human homologue of the mouse gene encoding the receptor of fertilization promoting peptide. *Mol Hum Reprod* **8:** 24-31.

Makalowski, W. (2001). Are we polyploids? A brief history of one hypothesis. *Genome Res.* **11:** 667-70.

Marra, M. A., Kucaba, T. A., *et al.* (1997). High throughput fingerprint analysis of large-insert clones. *Genome Res* **7:** 1072-84.

Mayor, C., Brudno, M., *et al.* (2000). VISTA : visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics* **16:** 1046-7.

Mazarakis, N. D., Nelki, D., *et al.* (1991). Isolation and characterisation of a testis-expressed developmentally regulated gene from the distal inversion of the mouse t-complex. *Development* **111:** 561-71.

Mazzarella, R. and Schlessinger, D. (1998). Pathological consequences of sequence duplications in the human genome. *Genome Res* **8:** 1007-21.

McLysaght, A., Hokamp, K., *et al.* (2002). Extensive genomic duplication during early chordate evolution. *Nat Genet* **31:** 200-4.

Mukai, J., Suvant, P., *et al.* (2003). Nerve growth factor-dependent regulation of NADE-induced apoptosis. *Vitam Horm* **66:** 385-402.

Mukai, J., Shoji, S., *et al.* (2002). Structure-function analysis of NADE: identification of regions that mediate nerve growth factor-induced apoptosis. *J Biol Chem* **277:** 13973-82.

Mukai, J., Hachiya, T., *et al.* (2000). NADE, a p75NTR-associated cell death executor, is involved in signal transduction mediated by the common neurotrophin receptor p75NTR. *J Biol Chem* **275:** 17566-70.

Mullikin, J. C. and Ning, Z. (2003). The phusion assembler. *Genome Res.* **13:** 81-90.

Murphy, W. J., Stanyon, R., *et al.* (2001). Evolution of mammalian genome organization inferred from comparative gene mapping. *Genome Biol* **2:** REVIEWS0005.

Neitz, M. and Neitz, J. (1995). Numbers and ratios of visual pigment genes for normal red-green color vision. *Science* **267:** 1013-6.

Neuman, S., A. Kaban, A., *et al.* (2001). The dystrophin / utrophin homologues in Drosophila and in sea urchin. *Gene* **263:** 17-29.

Nicholas, K.B., Nicholas, H.B. Jr., and Deerfield, D.W. II. (1997). GeneDoc**:** Analysis and Visualization of Genetic Variation, *EMBNEW.NEWS* **4:**14

Nowell, P., Hungerford, D. (1960). A minute chromosome in human chronic granulocytic leukaemia. *Science* **132**: 1497.

O'Brien, S. J., Menotti-Raymond, M., *et al.* (1999). The promise of comparative genomics in mammals. *Science* **286:** 458-62, 479-81.

Ohno, S. (1999). Gene duplication and the uniqueness of vertebrate genomes circa 1970-1999. *Semin Cell Dev Biol* **10:** 517-22.

Parra, G., Agarwal, P., *et al.* (2003). Comparative gene prediction in human and mouse. *Genome Res* **13:** 108-17.

Peifer, M., Berg, S., *et al.* (1994). A repeating amino acid motif shared by proteins with diverse cellular roles. *Cell* **76:** 789-91.

Pennisi, E. (2003). Drafting a tree. *Science* **300:** 1694.

Perry, J., Short, K. M., *et al.* (1999). FXY2/MID2, a gene related to the X-linked Opitz syndrome gene FXY/MID1, maps to Xq22 and encodes a FNIII domain-containing protein that associates with microtubules. *Genomics* **62:** 385-94.

Pillutla, R. C., Shimamoto, A., *et al.* (1999). Genomic structure and chromosomal localization of TCEAL1, a human gene encoding the nuclear phosphoprotein p21/SIIR. *Genomics* **56:** 217-20.

Rapp, G., Freudenstein, J., *et al.* (1990). Characterization of three abundant mRNAs from human ovarian granulosa cells. *DNA Cell Biol* **9:** 479-85.

Rappold, G. A., (1993). The pseudoautosomal regions of the human sex chromosomes. *Hum. Genet.* **92**: 315-324.

Reboul, J., Vaglio, P., *et al.* (2003). *C. elegans* ORFeome version 1.1: experimental verification of the genome annotation and resource for proteome-scale protein expression. *Nat Genet* **34:** 35-41.

Roberts, R. G. (2001). Dystrophins and dystrobrevins. *Genome Biol* 2**:** REVIEWS3006.

Roberts, R. G., Freeman, T. C., *et al.* (1996). Characterization of DRP2, a novel human dystrophin homologue. *Nat Genet* **13:** 223-6.

Rosenfeld, M. R., Eichen, J. G., *et al.* (2001). Molecular and clinical diversity in paraneoplastic immunity to Ma proteins. *Ann Neurol* **50:** 339-48.

Rozen, S. and Skaletsky, H. (2000). Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* **132:** 365-86.

Rozen, S., Skaletsky, H., *et al.* (2003). Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. *Nature* **423:** 873-6.

Saito-Ohara, F., Fukuda, Y., *et al.* (2002). The Xq22 inversion breakpoint interrupted a novel ras-like GTPase gene in a patient with Duchenne Muscular Dystrophy and profound mental retardation. *Am J Hum Genet* **71:** 637-45.

Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4:** 406-25.

Sanger, F., Coulson, A. R., *et al.* (1978). The nucleotide sequence of bacteriophage phiX174. *J Mol Biol* **125:** 225-46.

Sargent, C. A., Boucher, C. A., *et al.* (2001). Characterization of the human Xq21.3/Yp11 homology block and conservation of organization in primates. *Genomics* **73:** 77-85.

Scherf, M., Klingenhoff, A., *et al.* (2000). Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach. *J Mol Biol* **297:** 599-606.

Schmucker, D., Clemens, J. C., *et al.* (2000). Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell* **101:** 671-84.

Schwartz, S., Zhang, Z., *et al.* (2000). PipMaker - a web server for aligning two genomic DNA sequences. *Genome Res* 10**:** 577-86.

Segalat, L. (2002). Dystrophin and functionally related proteins in the nematode *Caenorhabditis elegans*. *Neuromuscul Disord* **12:** S105-9.

Serluca, F. C., Sidow, A., *et al.* (2001). Partitioning of tissue expression accompanies multiple duplications of the Na+/K+ ATPase alpha subunit gene. *Genome Res* **11:** 1625-31.

Shibata, K., Itoh, M., *et al.* (2000). RIKEN integrated sequence analysis (RISA) system--384-format sequencing pipeline with 384 multicapillary sequencer. *Genome Res* **10:** 1757-71.

Shizuya, H., Birren, B., *et al.* (1992). Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. *Proc Natl Acad Sci U S A* **89:** 8794-7.

Sidman, R. L., Dickie, M. M., *et al.* (1964). Mutant mice (Quaking and Jimpy) with deficient myelination in the central nervous system. *Science* **144:** 309-11.

Simmons, D. L. (1993). Cloning cell surface molecules by transient expression in mammalian cells. In *Cellular Interactions and Development*. (D. Hartley, ed.). IRL Press at Oxford Universiyu Pres, Oxford. pp 93-127.

Soderlund, C., Longden, I., *et al.* (1997). FPC: a system for building contigs from restriction fingerprinted clones. *Comput Appl Biosci* **13:** 523-35.

Solovyev, V. and Salamov, A. (1997). The Gene-Finder computer tools for analysis of human and model organisms genome sequences. *Proc Int Conf Intell Syst Mol Biol* **5:** 294-302.

Sonnhammer, E. L. and Durbin, R. (1995). A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene* **167:** GC1-10.

Spencer, J. A., Sinclair, A. H., *et al.* (1991). Genes on the short arm of the human X chromosome are not shared with the marsupial X. *Genomics* **11:** 339-45.

Srivastava, A. K., McMillan, S., *et al.* (1999). Integrated STS/YAC physical, genetic, and transcript map of human Xq21.3 to q23/q24 (DXS1203-DXS1059). *Genomics* **58:** 188-201.

Stankiewicz, P. and Lupski, J. R. (2002). Genome architecture, rearrangements and genomic disorders. *Trends Genet* **18:** 74-82.

Sternberg, N. L. (1992). Cloning high molecular weight DNA fragments by the bacteriophage P1 system. *Trends Genet* **8:** 11-6.

Strausberg, R. L., Feingold, E. A., *et al.* (1999). The mammalian gene collection. *Science* **286:** 455-7.

Sun, M., Zhou, G., *et al.* (2004). SAGE is far more sensitive than EST for detecting low-abundance transcripts. *BMC Genomics* **5:** 1.

Takai, Y., Sasaki, T., *et al.* (2001). Small GTP-binding proteins. *Physiol Rev* **81:** 153-208.

Tennyson, C. N., Klamut, H. J., *et al.* (1995). The human dystrophin gene requires 16 hours to be transcribed and is cotranscriptionally spliced. *Nat Genet* **9:** 184-90.

The *C. elegans* Sequencing Consortium (1998). Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282:** 2012-8.

Thomas, J. W., Prasad, A. B., *et al.* (2002). Parallel construction of orthologous sequence-ready clone contig maps in multiple species. *Genome Res* **12:** 1277-85.

Thomas, J. W., Touchman, J. W., *et al.* (2003). Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* **424:** 788-93.

Thompson, J. D., Higgins, D. G., *et al.* (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22:** 4673-80.

Thompson, J. D., Gibson, T. J., *et al.* (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* **25:** 4876-82.

Toder, R., Wakefield, M. J., *et al.* (2000). The minimal mammalian Y chromosome - the marsupial Y as a model system. *Cytogenet Cell Genet* **91:** 285-92.

Tourmen, Y., Baris, O., *et al.* (2002). Structure and chromosomal distribution of human mitochondrial pseudogenes. *Genomics* **80:** 71-7.

Ureta-Vidal, A., Ettwiller, L., *et al.* (2003). Comparative genomics: genome-wide analysis in metazoan eukaryotes. *Nat Rev Genet* **4:** 251-62.

Venter, J. C., Adams, M. D., *et al.* (2001). The sequence of the human genome. *Science* **291:** 1304-51.

Veyrune, J. L., Campbell, G. P., *et al.* (1996). A localisation signal in the 3' untranslated region of c-myc mRNA targets c-myc mRNA and beta-globin reporter sequences to the perinuclear cytoplasm and cytoskeletal-bound polysomes. *J Cell Sci* **109:** 1185-94.

Wall, J. D. and Pritchard, J. K., (2003). Haplotype blocks and linkage disequilibrium in the human genome. *Nat Rev Genet* **4:** 587-97.

Wang, P. J., McCarrey, J. R., *et al.* (2001). An abundance of X-linked genes expressed in spermatogonia. *Nat Genet* **27:** 422-6.

Waterston, R. H., Lindblad-Toh, K., *et al.* (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520-62.

Weissenbach, J., Gyapay, G., *et al.* (1992). A second-generation linkage map of the human genome. *Nature* **359:** 794-801.

Whistler, J. L., Enquist, J., *et al.* (2002). Modulation of postendocytic sorting of G protein-coupled receptors. *Science* **297:** 615-20.

Williams, J. G. and Firtel, R. A., (2000). HAPPY days for the Dictyostelium genome project. *Genome Res* **10:** 1658-9.

Wingender, E., Chen, X., *et al.* (2001). The TRANSFAC system on gene expression regulation. *Nucleic Acids Res* **29:** 281-3.

Wolfe, K. H. and Shields, D. C., (1997). Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387:** 708-13.

Xu, Y., Einstein, J. R., *et al.* (1994). An improved system for exon recognition and gene modeling in human DNA sequences. *Proc Int Conf Intell Syst Mol Biol* **2:** 376-84.

Yoshida, M., Shan, W. S., *et al.* (1999). Conserved and divergent expression patterns of the proteolipid protein gene family in the amphibian central nervous system. *J Neurosci Res* **57:** 13-22.

Yudate, H. T., Suwa, M., *et al.* (2001). HUNT: launch of a full-length cDNA database from the Helix Research Institute. *Nucleic Acids Res* **29:** 185-8.

Zhang, M. Q. (1998). Statistical features of human exons and their flanking regions. *Hum Mol Genet* **7:** 919-32.

Appendix A

This appendix lists primers used in the studies described in Chapter 3. The first section lists primers used in screening of the cDNA libraries. Section two lists primers used in the alternative polyadenylation studies. Section three lists primers used in the NXF2/TCP11-like variant studies. stSG numbers and the primer sequences are given in each section. In section one, bold-type STSs denote primary poolscreen STSs used in the transcript map. Any SSPCR nested primers designed are given in the "SSPCR STS" column. Prescreen results are given in Chapter 3. For Sections two and three, expected product sizes, pre-screen optimal annealing temperature in ^{0}C and whether the primers span introns are given. Where an STS was not pre-screened due to it spanning an intron, n/a is stated in the pre-screen column.

Section 1					
stSG number	Primer 1	Primer 2	SSPCR STS		
stSG101247	CTGTGTCACGTTGCTGCAAC	AAGCGGACTTGGCAAGCATG			
stSG101248	ACCAATCCAACTGACCATAC	AGTGCGTACTGGATTAGTAC			
stSG101249	ATCCTTCTGCTCAAATCTCC	GAGATTTGAGCAGAAGGATC			
stSG101250	TGAAGTGGCCTGTGGCATCG	GCAACAGTAGGAAGAAGTTG			
stSG101252	GACTCTCAGAAGCTGTTACC	CCAGCTGGCACTCGGTGTAC			
stSG101443	ATGCATATAGAGATTATACG	ACTGACCCATTGAAACCATG			
stSG101533	GTGACGTGGGTGAAGGAGTT	AAACCACAGTCAGTTGATGGG			
stSG118040	AGACAGGTGTGAGGCACCTT	ACTGGATGAGTCCCAACAGC			
stSG118961	TAATGGTGAGTCCATGTAAGGC	GGAAAGGGAAAGAAAGAAGAGG			
stSG118962	CCACCTTCAGTGACCAGGTT	CGGCATATTGATTTGGTCCT			
stSG118963	GAAGCTGGGACTGAAAGTGC	TCGAGATTATTGTCTGCCAGG			
stSG118964	TCCGAAAGTCCCTTCCTTTT	CTCACTGAAATGGTAGAGCCG			
stSG118965	TGGAACAGATTGTGTGTTTGC	CATATCTCACTGGCCTGAAGC			
stSG118966	AGCCATTAGGCAAACACCAC	TCCCAGACATGTTTCACATCA	119560		
stSG118967	CAAGTCAGCAGTGAAAAATTGC	TCATTTCTCTGGTGTGGCAG			
stSG118968	TGAAGTCTGCACCTCCCTG	TTGTAATCGTGGTGATTTCAGG			
stSG118969	TGGTGAATAACAACCAGCTCC	AGTGCCTACCTTGCAGCAGT	119556		
stSG118970	CAGAGGGTTTAGCATGGGAA	CTATCCCATGGCCAAGATTG			

stSG118971	AGAATTAGGTCCCTGTAACCCC	TCTCATGTTGGTGAATCATGTG	119557		
stSG118972	ATACCAACCCACAGAAATGAGG	ACGGGGACTGTAGGTGGAG	119558		
stSG118973	GAAAAGAAAGCCAGGATCCC	AGAAGAATGGCATGGGGAG			
stSG118974	GCACCATCTTCAAATTTGTGG	AAACAAATGGAAGAGCCAACA			
stSG118975	ATAGCATGGCAGAAAAGAGAGG	GGCACAACTCCCCACATATT	119559		
stSG118976	CCCTGCAACTTCCTCTTCAG	GTGAAAGGAAAGATGACCTTGG			
stSG118977	ACGTGATTGCACATTTGGAA	CTGCACTCTCACTGGTCCAA	119555		
stSG118978	AGCTAGCCTTGGCATCAGAA	GTCAATGTTTTGTGTGGTCTGG			
stSG118979	CACTGTCCATCTCTCCCCAT	TTCAATGATGCAAACTCAAATG			
stSG118980	CCTCTCCCTTGATGATCCAC	AGATGGGCATTGTGTCCCAT			
stSG119002	TGTTTGCTTGCTCTATGTTTCA	GCTGCTTTATTTTTCCACGAA	119562		
stSG119008	GAAAGAAGAAACCATCAGCCC	GCATCTTTCTCCAGAGCCTG	119565		
stSG119009	TTGCTACATTGCAGACCCTG	CCTCAGACCAGAGGCCTAAT	119566		
stSG119010	TCATGATCAGCTTCAGTGGG	ACAAAAAAAATCCCATCAGCC	119567		
stSG119011	CAGGCAAGGTCCATGATCTT	TTGGGGGAATAGGTCATTCA			
stSG119013	TGAAGTTTCCAAGCTGAGCA	CTGTTTGGAGGCAAAATGGT			
stSG119014	GCACTGTTGCTAGGTCCTCT	TTTTAGTATGGGTCTGAGTATGC	119568		
stSG119015	ACCTTTTGGCATTCTTCTGC	TGCAATAAACATCTTTCCATGA	119563		
stSG119016	TTGAGCTGGATCATTTTGTG	AAACAACGAATAGCAGGGTAGC	119564		
stSG119017	CCAAATGTGCGCACTCTACA	TGTTGCGGAACGAGTATCTG	119771		
stSG119018	TGCTAGGACATTTCTGAGAGCA	TGGCTCTGAATGACAGGC	119569		
stSG119022	CCTCTGTCCTGTGTAAAAAGGG	TTGGAAGCAGACCTGCTTTT	119570		
stSG119023	ACCGAAGTGAATTTGAACAACA	CAGCTTCCCGTGGGTCAT	119571		
stSG119024	TGGAATGAAATGCCTCTTCC	GTACTCAAAAGCAGCAGGATCC			
stSG119025	GTCAGAGGCCCATGTGAAAT	AGGAGAAGCAAGGAAGTGAGG	119572		
stSG119027	GTGCTGAGTCCCCGACTG	ACCTCCAACAGGGGTTTTG	119573		
stSG119075	TCAGTGGTGGTGTCACTGCT	TCAGGTCTTTGGACATGGTG			
stSG119076	GGATCAGAGCCTATGGCAGT	TGCCTTTTCATCAGCCTCTT			
stSG119098	GAAACTGCTAACAGCCCAGC	TCCTTCCTCCACGCTAAGTG			
stSG119120	ATTCCTGAAGCCCCTTCCT	GTTCACTTCCTTGCATGTGG			
stSG119121	GATCACTTTACTGTAGAAGAAATGAGG	TTTCTTCTTCTTGATCTTCATTGG	119772		
stSG119555	GTGAATGTGTTCATGTGCAC	GGTCTACTGTCCCTAGGATG			
stSG119556	TGAGCAGGAATGCAGTCCTG	GTTCTGTCTCTGTTCAGGTA			
stSG119557	AGTGAGAAGCTACATTATCC	TAGAAATCATAAGGGAATAG			
stSG119558	TTCCTCCAGTGTATGCTGAG	CATAAGCAAATTTAACTCTG			
stSG119559	GCTCATCATTCTGAGACAAG	TGTACTACCTAATCTAATGC			

stSG119560	CTTCATGTGACCTATCACTG	AATTCTGTTCTCCAGGTTGG			
stSG119561	ATTAATGCAAATATATTTCC	TGAAATAGATTCATATTG			
stSG119562	GTAGCAGATTTTCACACATG	TCATGACGTCTCTGCAATGC			
stSG119563	TTCTGATTGGAGATTGCA	AAGTTGATGAAAGTCCTC			
stSG119564	GATGCCAAATGAATATCA	ACTTAGGTGACACAGTTC			
stSG119565	ATAGTTGAAGGACTGCGCTG	TGTTATTCTTGGAGATTGAG			
stSG119566	GAGCATTACAGTGGAAAC	AGATTCTTCCTATTCACA			
stSG119567	CTGTAACAGCTATGCACATC	CGTACTGTCACGGGATAATG			
stSG119568	ATCATGTTCCTGGGTCTTCT	AGACCCAGCACTCATGTTAG			
stSG119569	ATAACATTGCCTAGAAACCG	CTACATTTAACCACATCAAC			
stSG119570	CCAATCACAGTCATCATGG	CAAGGCTACCTGCTGAGCAG			
stSG119571	CATCTCAGACCCATTCTACG	GGTCATGAGAGTCTGGTGTC			
stSG119572	TATAAAGATGTACACTGTGC	GGCTGGATTCTGCAGAAGA			
stSG119573	GCTTGCCATCACACAAAG	GAGGCTGGGATGTATTTC			
stSG119771	CAGGAGTTCCTATAGGCTAT	TGATACAGCTGGAACAGGAG			
stSG119772	GAGTGAGGGAGAACAACAAC	GTGCATTAGAGTTGTTGGGC			
stSG46776	GTCCAGGGATCAAAAATGACA	CAAAATGAAGAATTGTGTGGAA	119561		
stSG84327	GAAGAGGATCTCAGGACTG	AGTTGGTGCCGCTGACCAG	85560		
stSG84328	CTGAGGAATCACCTAAGC	CAGATCTCAGAAGTGAAC	85559		
stSG84329	TGAATGTACACAGCGAAC	GTGCACAGGAATCTCAGC	85558		
stSG84336	TCCAGGTACCATTGCTTTCC	TCCCAAAACCAATACTTGTCG	87976		
stSG84337	TGTGAGAGTGGCTTCCTCCT	AAACGGTTTTTCACCAATGC			
stSG84338	GGTTAGGGCTGAATGTTTTCC	CCCAAGGATTTTTCTTGTGTAG			
stSG84339	GGGGCTCCAATCCTAGAATT	GGATCCAGACTGGAATATCGG			
stSG84340	CATGGAGCACACTTTCTACAGC	TCGTGCACTTCCAATGTTTC	87977		
stSG84341	TAGAATTTGATCCCATGCTGG	CTCGTCCTCCTTGGTAGCAG			
stSG85558	CATTGAAGTGCGCTCTTTGC	TAATTGACCTGCGCTTCAGC			
stSG85559	CCTCTGAAGTCAATGTTAAC	CATACATGTGGATTAGTGG			
stSG85560	ACTCAACCTTCTGAGGAAG	TCCATAGAAGGACACAATG			
stSG87849	CTTGGATTTTCAAAGTCCTTGA	CCTCTGATGCAGCAGGATG			
stSG87850	CTTGTTCGTGCTTTTAGCAGG	GGGCTCTTCCCTTGGTAAAT	88201		
stSG87852	TTACATAGGCAGACTGGTCCG	GGTGAAAGGGCTGAAGCTC			
stSG87853	TTCTGCTGTCTTTGGAACTTTG	TCTTCAACATCGCCACCC	88203		
stSG87854	AGCTTTTTAGGCCAAGTGGC	TTGCTAGCAGCCATGACGT			
stSG87855	AGCTTTTTAGGCCAAGTGGC	CAGATGTAGTTTGGCACAGCA			
stSG87856	TCATGAGTGGGGCAAATACA	AGACCACTCTAAGGACTGGGG			

stSG87864	GAATCTGCTCCCACAGTC	GTAGAATAAGGTGGTGATG			
stSG87865	GCGCGCACCCAATTCAGTC	CGGCAGGTGGCCCAGATC			
stSG87976	AAGCCTGTTTCGACTTCTGC	GGATAAATACAGAAGCCATGCC			
stSG87977	TCAGAAGACAATGCAAATCTGG	GCTGAAATGCATCGTAACCA			
stSG88115	CGCTCCCGTTCGGCTCCTC	TGGCTTCGAGGAGCCGAAC			
stSG88151	TACTCTGAGGCAAGCACTGG	CATGGATGTCAACCAGCAAC	88152		
stSG88152	AGTCATCATGGAGACCTG	AAGGTGGTGATGACTGAG			
stSG88153	TGCACTCACGAGCTTCTGAG	ATATCTTCCCATCTGCGTTG	88154		
stSG88154	AGATTGCCAACCAGCAGC	TTGGCGCCTTTGCCTGCG			
stSG88155	TCTGTGTAACCTTCAACC	TGAGCCTGTAGTGGCATG	88156		
stSG88156	GAATCTGCTCCCACAGTC	AGAGGCTTAGGTGCACTG			
stSG88162	GCGCGCACCCAATTCAGTCG	AAGTTGGTGCCGCTGACCAG	88163		
stSG88163	TTCGGCTCCTCGAAGCCATG	AATGCTCCATAGAAGGACAC			
stSG88169	CGATTAACCTTCCAACGCAT	GAGGAGGCCAAAGCAAAAG			
stSG88170	TGACTCAACCGCCATAGTTG	TATCAGCAACAGCAGCAGCT			
stSG88171	CCAGATTGTGAAGGTAGAAGGG	GTGTGGCTTGTGTTGACAATG			
stSG88172	TTAAAACTACTGCCCCCCCT	GTAATTGTTCCTGCCTCCTCC			
stSG88173	TGTGTCAAGGGCTCAGGAG	TTTCTGGCCCCTACCTCC			
stSG88174	CACACCCGTTAGGCATTTG	TGCGTGAATGTCTTTGCTTC			
stSG88175	GCAAGCAGAGGAAGCGAG	AAGAATCCATGACACTCACGG			
stSG88176	TTACCAACAATGCTACGAATGC	CCATCAAGGCCTTAAGGACA	93105		
stSG88201	TGATTACATGGGCCGCATTC	AGCAGGTGCCAGCCCAGATC			
stSG88203	ACAGTTAGAAGATGCAGC	GAGCCCTTGGGCTGCATC			
stSG88327	GGAGCAGAGTCTTGGAGCC	TTCACGACATCTTGGAGCAC			
stSG88328	CGAGCAGTACCGCATGTG	AGCTTCGCTGTTTGTCCTGT	93106		
stSG88329	AAGCCAAAAATGTGGAGAAGC	TTTCCTCCATCTTCTTCTTGAA			
stSG88330	ACTGAAGATGGTCACACACAGG	GAGAGCCAAAGGGAGCAAG			
stSG88331	TCAAAGAAGAGCAAGATGTTGG	TGGAGAGTCCATGTGGGC			
stSG88332	AAAATTGAGCTTGCAGGGG	GTTTCATCAGCTCCAAAGTGC			
stSG88333	CTCGTGTGGCCCTACACC	CACAACCCCTCTGGGTAATG			
stSG88334	TGTCTTCTGCTGCTGCTTGT	CCACTGGCAGTTGCCTCT			
stSG88335	TCTTCGGGGACAGCTTTG	CGATGCGGTCCAGTATGTC			
stSG88336	TTGCTTTTGGTGTACCTGAGG	GATCCATGACATATTGGACCG			
stSG88337	CCTGAATGCCAGGATGATG	TTTTGCAGCTGTTTCTGGTG	93107		
stSG88338	TTGTCCGTGGTCACCTCAT	ACCGGGTGAATGTCTTCTTG			
stSG88339	TACCAACTATAGCACGGCCC	TCCACTGTCTTGTGCCTCAG			

stSG88340	GGGAGCTGGACAGGAAGAG	GGAGACCACAGAGAAGCTGG			
stSG88341	ACCTTAGGATGACTAGGGAGCC	TGGTTGCTAATATGCAGTCCC	93108		
stSG88342	AAGCGAAAATACACAACCCG	ATGTGTCAGCTCCTGCAGG	93109		
stSG88343	CTGGGAGTTGTTCAAGGGAA	ACTTGGAAGCAATGACTTGTCA	93110		
stSG88344	TCTGGTGTCATCTCAGGCTG	TGTCCGTTGAGTAGCGTCAG			
stSG88345	CTGGGAAAGAAGTTCGTGCT	TTCATGGTCACATGATGGCT			
stSG88346	AGCCTCTGAACATCCCCTG	CACCTTCTCTAATGTGACCTGC			
stSG88347	TCTGGGTCTGACATCAACCA	CTCAGTGGCTACCAACCCAT	93111		
stSG88348	CAAAAAGAAAAACAGACAGGGG	TCCAATGAAAACCACCCATT	93112		
stSG88349	TGGTTGCTAATATGCAGTCCC	ACCTTAGGATGACTAGGGAGCC			
stSG88350	TGCGAAAGAAGAATGAAATGG	CCAGTTCCTTGTGTATCCATCC			
stSG88351	TTCCTGAGGGAAGACCTCAA	CCATGCCTTCCACAGGTC	93113		
stSG88352	ACATAATCTTCAGCCGGGC	AGGTGATGAGGGACAACTGG			
stSG88353	TGCAGCCCTTGTTTCACTC	AAGGAAAGCCAGAGAGTGAGG	93114		
stSG88354	TGCTAATAAAGGGGAGCCCT	TCTCACCTCCTCCCCAATC			
stSG88355	CAGTGGCCCTTTTCAGATCT	TGGAAGATCTTCACCGAAGG			
stSG88356	ACTTGCAGGGGTATCGATTG	CTATTCTCATGGCGAGTCCTG			
stSG88357	TGCAGTTCACGCAGTAAACC	TTGCATGTTTCAAGACAAGAAG			
stSG88358	CCCTCAAACCAGATGCAAGT	CCAGTTTGGGCCAGAATAGA	93117		
stSG88359	TCCTTATCCTGGGAGGAGGT	ACCCTGCGGTTTTGTAACTG			
stSG88360	AAAAAGCGACTGAGAAAAGTGC	TATGGGAGGGACAAAGTTGC			
stSG88361	AGTTTCCGCGTGGTTTCTC	TCTATTTCCGTGGCTATGGC			
stSG88362	CAGTGATGGTAGCAGCTCTCC	CATCTACCTTCTGAAGATGGGG			
stSG88363	TGATTGAAACTCCTCTGGGG	CGTGATTTGCAGTGGCAG	93118		
stSG88364	AAGAAGAAAAACCCAGAGGTCC	CTTCTTGCCTCCCTCAACTG	93119		
stSG88365	CCCACAAGGGTTGAGAGTGT	ACGTGAGCCAGCTGTTCC			
stSG88366	CCACCAAAATGTAACACAATGC	TGATTTTTCATGTCACTGGTCA			
stSG88367	TCTCACCTCCTCCCCAATC	TGCTAATAAAGGGGAGCCCT			
stSG88368	GGAATTCCCCCAGGTGTC	TTGGCTCAGTGACCTTGATG			
stSG88369	AGATACTCCAGTCAGCAGAGGG	CGAAGAAAGTGGTTCAACAGC			
stSG88370	AGTTAATCAAACACTGGCCCC	TCCTAAGGTTTTCTGGGCAA			
stSG88371	GAGAGGGAGGCAACGAGAG	CTCAATGGTCAGGTGCCC			
stSG88372	GTGATGCAGGAAGAGAGCG	GATCGAACCACCACAGGC			
stSG88373	TGAAGAAGGAAGCTCTCCTCC	TAACAGTGGTCCGACCATCA	93121		
stSG88374	TTGAAATCCTGATGGAGATGC	TTGTGATCTCTGCAACATGATG			
stSG88375	AGCCAGCATGAAAGGACG	CAGTCCCTGCACACATATGG			

stSG88376	GCTCCTGCAGCTGAAACTG	AGCTGAAGCAAGAGCAGGAG	90808		
stSG88377	AGTGTTTGTGTCCATTGGCA	GTCCCTTCCCCTACCACAAT			
stSG88378	TTTTGCAATAGGCTGGGTTC	ACCTTTTCTAGCAGCTGCCA			
stSG88379	AGAAGAATCCCGCCATTTG	TCTTGATTTCCATCATCTGCC			
stSG88380	GATGTTCAAGAAAATTCGAGGG	TAACATTGCCAGCTGTCTGTG	93122		
stSG88381	TTTGACTCCTGCCTTCCG	CATAATGAGAAGCTGGGGGA	93123		
stSG88382	AACTATGGTGGTTCTGTGAGGG	AATGGCAAAGACGAGTCCC			
stSG88383	TGAGACCAAGTCTGTGCCTG	AACGAGCGCCACCTACTG			
stSG88384	GTGTCTGTGCCCGGAAAG	AATCAGGGCCCCGAGTAG			
stSG88385	CAAGGTTGTTGCTGAGACGA	TCATTTTTACATGTGGAGCTGG			
stSG88386	TCTTCGGGGACAGCTTTG	CGATGCGGTCCAGTATGTC			
stSG88387	AAGAATCCATGACACTCACGG	GCAAGCAGAGGAAGCGAG			
stSG88388	GTGTGGTGCAAGTGACCATC	GGCTCCATCTCCTTCTTATGG	93126		
stSG88389	AGAAGTTGGAATCATTGTTGGG	AGTTCCGCGATGGTCTTAGA	93127		
stSG88390	AAAAATGTGCCGTGTCCTTC	AGGGGTTTCCAGGCTGAG	93128		
stSG88391	CCTTCCCCGGAAAGAAATAG	ATTTCCTCATTGCAGCAGCT	93129		
stSG88392	TATCCCAGGAGCCCTGCT	AAAGCCAGAATCTTTGTTCAGC			
stSG90808	GCTGTGGTCCCTCAGCGGTT	GGAGGAGCCGCCCCAGACGA			
stSG93106	CAGTCTAAAGCCATGGATGA	GCCTGTTTATGAAGATGCGT			
stSG93107	CACCACCAGCTCCACTCACA	TGGGCAGACGATCTCCACTT			
stSG93108	CGGTAGATCAGGAGCAGCTT	GATGGCAGAACATGCCAAG			
stSG93109	GAAACCTGTGACAATGAGA	GTGATGATCAGATTCTCAA			
stSG93110	GGATTATTCACATCTTATGA	CACTACTCATAAGATGTGAA			
stSG93111	CTCTGTTCCTGTGTGTTTCT	CAAAGAGGACATATGATTCT			
stSG93112	CAAGGACTCTCAGGAGGACT	CCTTGACACATCTCCACATT			
stSG93114	CTCTCCTTCACTCTCTGGCT	GAGGGAGGGAGAGTCAGAGA			
stSG93117	GATATTGATGTGGATCTCAA	CACATGGTTGTTCCAGAGGA			
stSG93118	CCAAGATCCCACCTTATCAT	CTTCGGATTAGGCCCATACT			
stSG93119	CCAAGAAGTGCTCTGAGAAG	CATTATCCTTCATTTCCTGA			
stSG93121	CATCCCTAATGTCCTGAAG	CTTGATCTGCCCATTTTCT			
stSG93122	GTTTACATGACTATGGCTTA	GTAGTGCATTGGAGGATCGT			
stSG93123	GTGCCAGATTATTCCTGTGT	GATGGAGCAATACAGCCAAT			
stSG93126	GACGGTTTCGTGAACGTGA	GAAAGACCACTGGATGGAA			
stSG93127	GATTGGTAGCCTGGTAGGTG	GCCTTATTCCTTGCGAAGCA			
stSG93128	GTGCTGACACAGCTGGTGA	GCAGAGGTGCCCTTACTCT			
stSG93129	CATGGAGCCTAGGAGCAGAG	GACTTGCTGTAGCCATATAA			

stSG95448	AGTACGAGGGCATTGAGTTCA	TCTGGGCTTATCCTCCACAG			
stSG95449	CCCTTTGCAATATGTTGGCT	GAGCACTAGGAAGGGCACAG			
stSG95450	AAGGAAAGTCGAGAGGCTACG	CAGATAGGCAGGTACTGATCCC			
stSG95451	CAGATGTAGTTTGGCACAGCA	TGCTTTTTCAGACCTCTTTCTG			
stSG95452	GATGGTGTCCCATGCTCAG	CTGGACTGGATCTGCTACTGTG			
stSG95453	CTGCAGCTGTCAAAGGTGAA	ATTCTGCCTGTCACTGTCCC	101248		
stSG95454	TACTGAGATGGGTCTTCGGG	TGCTTCAGTTTTTCATCTGTGA	101249		
stSG95455	GAGACCGGTGGCCTGTTT	TCACCTGAGTTCTGGGCTG			
stSG95456	CAATCGGATGTCTGGGTCTT	TGCTGCTGCTAGTAGAAACTGC			
stSG95457	GGAGCTCTCTACCCTGGACC	TTCTTTGAGCCTGTGGCC			
stSG95458	GACGCTTAGAAGCAGAGCGT	TAGCATGCTACCCCCTGTG			
stSG95459	CCCGACATTTGTCTTGGTCT	CCTGGGGAATCCTTCTCTTC	101252		
stSG95460	AAGGACACCACCTGGTCAAG	TCATCCATGGACAGCCCT			
stSG95461	CACACTGGAATGGCATTGAC	GAGTTGTTATCCCAGTTGGAGG			
stSG95462	CAGCCTCACAGCCCTACTTC	GCCTGATTAGGAAGGAAATGC			
stSG95463	GATCAAAGTACTCCAGCAGCG	TCGTCTCGCTTTTCTTCCAT			
stSG95464	CATCAATGAAGTGTGTCGGG	GAACCTTAGCACATCCACTTCC	101247		
stSG95465	CAATTCCATTCTCCTCCGAA	GGAAGAAGAGCACAGCCAAC	101250		
stSG95466	AAGTTCAGATGACCCAAACAGA	CCAACCATTTGTCTGCTGC			
stSG95467	CTGAAGGAAAGCTCCACATTG	GGCATTCTACAAAGCAAAAACC			
stSG95468	GAACCTCGTGCTTTCCCC	GAAGCCTCCGAAGCCTGT			
stSG95469	AGAAGAAAGAAGGAGGAAATGG	CTTACCCACACGCTGCAGC			
stSG99719	CAGTTGTGAGCCTGAATACAGG	TCTGATAGTCACACTGTTGGCC	101254		
stSG99720	TAGGAAGTCAACCCCCATTG	TGGTTACTTTGGGACATCATTG	101255		
Section 2					
stSG number	Primer 1	Primer 2	Product size (bp)	Pre-screen	Intron spanning?
stSG158910	CTATTCATTTTCTCCACCTTGTTT	TGTCACTTATGCAGAAGAAATAGC	225	60	no
stSG158921	TGGGTAGAAGAATAATGAGTGATCTTT	CACTTGCTTCATTCCCAACA	130	60	no
stSG158922	CAGTTTGATGTACCTGCGTGA	TTTGCAACCCTTCACTCTGA	196	60	no
stSG158923	TAAGGCTTTGCCCTCTGAAA	TTGCTGACACACTCAAACCAG	238	60	no
stSG158924	TGTCAAGCAAAAGAATGCAAA	TGCATAGCCAACATCCAAGT	204	60	no
stSG158925	AGCCTGGTGTGATTGATGTG	TGAGGGGTATTCTGACAAAGAGA	199	60	no
stSG158926	TCCCCAGATCCTTCAACAAC	CCTCAGGGCTAGAATTTCAGA	157	60	no

Section 3					
stSG number	Primer 1	Primer 2	Product size (bp)	Pre-screen	Intron spanning?
stSG453287	GAGATCCTAAAACACATCATCCAT	TGGCTATAGCGAGGAAGAGG	178	60	no
stSG453288	TTTCCAAGTTCGTTCTGTGAGA	CTGGCGTTTATTGAGGGAGA	150	n/a	yes
stSG453289	GGGAAGTGTAGCTCCCAGGT	AAGTGTCAGGGCTGTGGAAG	150	n/a	yes
stSG453370	GAGATCCTAAAAACCCTCATTTCC	GCAGGGCCTGGGATAGAA	173	60	no
stSG453302	AGCCTAAGGCCTGGCTGAC	CAGAGCAACTGGCTCTTTGG	81	60	no

Appendix A.1

This appendix lists the clones/accession numbers from human Xq22 and represented in the sequence contigs of Figure 3-9. The column headed 'SEQCTG' indicates which clones/accessions are contained in the numbered sequence contigs of Figure 3-9.

SEQCTG	Accession	Clone name		SEQCTG	Accession	Clone name
seqctg1	AL109750	dJ902O5		seqctg2	Z70228	cV411C1
seqctg1	AL022148	dJ435A7		seqctg2	Z70758	cV434E11
seqctg1	AL606759	bK2319N17		seqctg2	Z69304	cV311G7
seqctg1	AL590293	bA227B10		seqctg2	Z68868	dJ3E10
seqctg1	AL137843	dJ377O6		seqctg2	Z68326	cU163D10
seqctg1	AL358953	bA552B20		seqctg2	Z80107	dJ197J16
seqctg1	AL390027	bA368G3		seqctg2	Z97355	cU230G7
seqctg1	AL590012	bA40I8		seqctg2	AL035214	dJ122O23
seqctg1	AL359641	bA402K9		seqctg2	Z70719	cV351F8
seqctg1	AL355593	bA99E24		seqctg2	AL133277	dJ158I15
seqctg1	AL590412	bA557A17		seqctg2	Z70689	cU19D8
seqctg2	Z93928	dJ127B14		seqctg2	Z70226	cV1077H7
seqctg2	AL035608	dJ479J7		seqctg2	Z68332	cV775G11
seqctg2	AL391688	bA524D16A		seqctg2	Z81367	cV618H1
seqctg2	Z73900	cU85H7		seqctg2	AL590069	bA353J17
seqctg2	AL391689	bA524D16B		seqctg2	AL035609	dJ77O19
seqctg2	Z95327	dJ347M6		seqctg2	AL035551	dJ1100E15
seqctg2	Z83819	dJ146H21		seqctg2	AL035427	dJ769N13
seqctg2	Z73417	cU131B10		seqctg2	AL590407	bA522L3
seqctg2	Z97985	dJ341D10		seqctg2	Z68871	cU157D4
seqctg2	AL133275	dJ1053B6		seqctg2	AL669904	dJ1054G24A
seqctg2	AL109952	dJ664K17		seqctg2	Z95624	cU237H1
seqctg2	AL109963	dJ1188J21		seqctg2	AL645812	dJ1054G24B
seqctg2	Z70280	cV210E9		seqctg2	Z93943	cU235H3
seqctg2	AL022155	dJ302L24		seqctg2	Z75895	cU61F10
seqctg2	Z70281	cV526F1B		seqctg2	Z73361	cU73E8
seqctg2	Z68331	cV521F8		seqctg2	AL008708	dJ198P4
seqctg2	AL109801	dJ738A13		seqctg2	Z75746	cU221F2
seqctg2	AL035422	dJ164F3		seqctg2	Z85997	cU101D3
seqctg2	Z69838	cV1164A6		seqctg2	AL035494	dJ635G19
seqctg2	Z68873	cU209G1		seqctg2	Z81014	cU65A4
seqctg2	AL133280	dJ514P16		seqctg2	Z68694	cU177E8
seqctg2	Z73913	cU61B11		seqctg2	AL133348	dJ79P11
seqctg2	AL121883	dJ545K15		seqctg2	Z92846	cU105G4
seqctg2	Z83131	cV602D8		seqctg2	AL606763	bB349O20
seqctg2	AL392164	bA269L6		seqctg3	AL079333	dJ823F3
seqctg2	AL672206	dJ232L22		seqctg3	AL117327	dJ421I20

seqctg3	Z69733	cU250H12		seqctg3	Z75747	cU96H1
seqctg3	AL035444	dJ43H13		seqctg3	AL391070	bA539A6A
seqctg3	AL021308	cU246D9		seqctg3	Z70225	cU165H7
seqctg3	Z73965	cV857G6		seqctg3	AL391071	bA539A6B
seqctg3	Z68327	cU25D11		seqctg3	Z74619	cU232G2
seqctg3	AL049610	dJ1055C14		seqctg3	AL022168	cU247E12
seqctg3	Z93848	cU35G3		seqctg3	Z70051	cU92G6
seqctg3	AL034409	dJ764D10		seqctg3	AL133272	bA229F9
seqctg3	Z73964	cV698D2		seqctg3	Z69367	cU159B9
seqctg3	AL139228	dJ540A13A		seqctg3	Z99706	cU226D1
seqctg3	Z75896	cV461C10		seqctg3	Z73967	dJ315B17
seqctg3	AL139229	dJ540A13B		seqctg3	Z68330	cU9D4
seqctg3	AL390022	bA370B6*		seqctg3	Z68908	cU227D1
seqctg3	Z70273	cU116E7		seqctg3	Z68328	cU72E5
seqctg3	Z70227	cV362H12		seqctg3	AL139813	dJ312P4
seqctg3	AL034485	dJ839M11		seqctg3	Z70274	cU84B10
seqctg3	Z73497	cU240C2		seqctg3	Z97356	cU25E4
seqctg3	Z74620	cV467E10		seqctg3	Z69734	cU71B4
seqctg3	Z82254	cU46H11		seqctg3	Z68339	cU230B10
seqctg3	AL135959	dJ233G16		seqctg3	Z83850	dJ82J11
seqctg3	AL049631	dJ513M9		seqctg3	AL512661	bA560L11
seqctg3	AL121868	bA541I12		seqctg3	Z69722	cU212C1
seqctg3	AL136977	bA230E14		seqctg3	AL135922	dJ875J13
seqctg3	AL133385	dJ81E11		seqctg3	Z68289	cU50F11
seqctg3	AL133381	dJ406H21		seqctg3	AL133271	bA155N17
seqctg3	AL050401	dJ519P24		seqctg3	AL139812	dJ19N1
seqctg3	AL121866	bA191C22		seqctg4	AL590306	bA565G2
seqctg3	AL021812	cU86H4		seqctg4	AL606515	bA575K24
seqctg3	AL008642	cU18H8		seqctg4	AL590808	bB483F6
seqctg3	Z70224	cU144A10		seqctg4	AL606833	bA647M7
seqctg3	Z80774	cU173H7		seqctg4	AL391315	bA150F24
seqctg3	AL442225	bA40K1A		seqctg4	AL591849	bA321G1
seqctg3	Z69724	cU85B5		seqctg4	AL158821	dJ75H8
seqctg3	Z95333	cU116E9		seqctg4	AL390039	bB383K5
seqctg3	AL512445	bA258F8		seqctg4	AL136112	dJ1126E12
seqctg3	Z68872	cU161B10		seqctg4	AC004081	dJ320J15
seqctg3	Z70232	cU139A4		seqctg4	AL035088	dJ3D11
seqctg3	Z75745	cU107D4		seqctg4	AL137787	dJ1070B1
seqctg3	AL133276	dJ114N20		seqctg4	AL772400	bA540N4
seqctg3	Z71182	dJ248J6		seqctg4	AL590423	bB364K23
seqctg3	Z70040	cU174H1		seqctg4	AL109946	dJ820B18
seqctg3	Z75957	cU203H4		seqctg4	AL034399	dA191P20
seqctg3	Z74477	cU42H12		seqctg4	AL953860	dJ1107B12
seqctg3	Z81144	cU129B7		seqctg4	AL031177	dJ889N15
seqctg3	Z69721	cU201H11		seqctg4	AL136080	bA448E12

seqctg4	AL109943	dJ657D12		seqctg7	AL392112	bB266I11
seqctg4	AL034369	dA149D17		seqctg7	BX088563	bA1066D24
seqctg4	AL031622	dJ740A11		seqctg7	AC000114	dJ527C21
seqctg4	AL136364	dJ734E5		seqctg7	AL034450	dJ115K14
seqctg4	AL035425	dA24A23		seqctg7	AL137124	bA485M23
seqctg4	AL928697	bK2004P2		seqctg7	AC004827	dJ44L15
seqctg4	BX322556	yR4AC12		seqctg7	AL078580	dJ874H6
segqctg5	AL732308	bA199F23		seqctg7	AC002072	dJ218B13
segqctg5	AL670405	bB148E24		seqctg7	AL023876	dA111F4
segqctg5	AL928646	bK2328D6		seqctg7	AL023877	dJ1142C11
segqctg5	AL731796	bB344M15		seqctg7	AL772262	bA444C24
segqctg5	AL390836	bB179C6		seqctg7	AL031183	dJ1168A5
segqctg5	AL034403	dJ31B8		seqctg7	AL030995	dJ1170D6
segqctg5	AL031387	dJ596C15		seqctg7	AL442070	bA468C24
segqctg5	AL118496	dJ136J15		seqctg7	AL050311	dJ964N17
segqctg5	AL138968	dA13I1		seqctg7	AL049859	dJ826A18
segqctg5	AL590647	bA130P9		seqctg7	AC002449	dJ402K21
segqctg5	AC003013	dJ205E24		seqctg7	AC003014	dJ290B4
segqctg5	AL590384	bA349A16		seqctg7	AL513476	bK2335J1
segqctg5	AL360224	bA471A8		seqctg7	AC004998	dJ164D5
segqctg5	AL359079	bB360B22		seqctg7	AC000055	dJ73F11
segqctg5	AL079334	dJ300O13		seqctg7	AL953888	dJ137P21
segqctg5	AL031319	dJ364I1				
segqctg5	AL137844	dJ557A17				
segqctg5	AC000113	dJ302C5				
segqctg5	AL360174	dJ465D4				
segqctg5	AL512882	bA441A11				
segqctg5	AL049176	dA141H5				
seqctg6	AL591489	bA814C6				
seqctg6	AL590809	bA733H21				
seqctg6	AL356578	bA14G9				
seqctg6	AL357774	bA473N15				
seqctg6	AL117326	dJ944N9				
seqctg6	AL031117	dJ914P14				
seqctg6	AL450490	bA124N4				
seqctg6	AL035067	dA170F5				
seqctg6	AL589880	bA111F16				
seqctg6	AL096764	dJ298J18				
seqctg6	AL049563	dJ68D15				
seqctg6	AC005191	dJ269O5				
seqctg6	AL031388	dJ737M10				
seqctg7	AL929583	bB530G12				
seqctg7	BX119929	dJ1150P3				
seqctg7	AL031223	dJ1041B16				
seqctg7	AL031176	dJ124K22				

Appendix B

This appendix contains primers used in the studies presented in Chapter 4. stSG numbers, primer sequences and optimal pre-screen annealing temperatures in ^{0}C where determined are given. ND denotes pre-screen not determined. A fail denotes an instance where the pre-screen failed to give a clear result.

stSG number	Primer 1	Primer 2	Prescreen
stSG136026	TTCTGGTGTCACTTGTTTCCC	TATACTGAGCATCTTCCCATGC	60
stSG136027	TTCTCTGAAGATGACATGGGG	TACGGATCTTCCCATGCAAT	FAIL
stSG136028	AACCAGCATGTGTTTAGCCC	GACCTCTCTTTGGATTCCTGG	ND
stSG136029	GTCACCAGCTTTAAGCTGAACC	AGCTGAGTAGGCCATTCACG	ND
stSG136030	TGGAATCTATTTCTTGGTTGGG	TGTTATTTCACTTTCCAACCCA	FAIL
stSG136031	ATGCTGGTGGCAATTCTACC	CGAGAACAACATTTAGAAGGGC	60
stSG136032	ATGGACTTTCCACCTGAACG	CCCTGTTGGTCTAAGGCTCA	60
stSG136033	AACAAAATGAGCTTCTGATGGG	TGGCAAATACAAATAAGCAGAA	60
stSG136034	GTTGATGCGTTAGTTGGTATGC	GCTAATGTTTTCGCAAAGGC	60
stSG136291	TGAACAATGAAGCTGCCACT	TTTCTTTTTGACACCATCTTCAA	60
stSG136970	GCCAAGCCCTAGCCTCTC	ACAGTGGCCAGCCAAAAG	60
stSG136971	CGAACTGGAAAGCAGACTCC	ATTTGCTGCTTTTGGGTCC	60
stSG155403	TGCTGCCACTTACAACTTCG	ATCAGTGGCAAAGGCAGAGT	60
stSG155404	ATCCATACTGTGACTCCCGC	AAGCTGGCAACACAAGCC	60
stSG155405	GGGTTTTGCTCATGAAGCAT	TGCGCATTGTAAATTGCATT	60
stSG155406	CGAGCAAGGTACTGAGTTTGC	CATCCTGGTGTTCAGGCC	60
stSG155407	GCCAATTACGTGTCTGGGTT	GGCACTGAATAGTTTTTTTGCC	60
stSG155408	CCTGTGGCAGGTTCTAGCTC	TGAGAGTTGGATCACAGTTTCC	60
stSG155409	CCCCGAAGAAGTGATAAGAGG	AGGCCTGAAGCACACAGG	60
stSG155410	AATGCCTTCAGCCTTTCCTT	GCACATGGGTGAAAGTCCTT	60
stSG155411	CCCCCTAAAAGCCCCATATA	CATTTGGGGGACAAAATTTG	60
stSG155412	TCCCCTGTCCAGGAACTTC	AAAAATGGAATGCTACAGAGCA	60
stSG155413	CACTTCCAAGTCCATGCCTT	GGAAAAGGCTTGCAGAAGAA	60
stSG155414	CCCTTGATACCCAAATCCCT	CTGTCATTTTCAGGGGCAAT	60
stSG155415	ACCCCAGAAGAAGCAGAAAT	TTCCATGACAAAGAAACAGAAA	ND
stSG155416	TAAGAATAAGCCCTCTCTTGGG	TGCATTGGACAGAAAGAGACC	60
stSG155417	TTCTTTGGCCTTCTTCTCCA	GCCCACCTCTCAATTATGGA	60
stSG155418	AACCAAGAGAGTAAAAGGAGCC	TTCTTCAAATCCATGAGATGG	FAIL
stSG155419	ATGAAAACAAACCACTGCAGG	GCTTCTGATCAAGATTTCCCC	55
stSG155420	CCATCCAGATAGCCAAATGG	TTGCAATTAATGACAGTGATGC	60
stSG155421	TGTCAAAACCTGAGTCATCCC	TTATCCAGTGAGGAATGTGGC	60
stSG155422	CGTCAGAAATTGTGGGAAGG	CTTCCTGCTATGCAGCCTCT	60
stSG155423	CATGCATAATGGACACCATACC	GACTGCATTGCAACTGAGTCA	60
stSG155424	CAGAGAGCCACATGAATGTCA	TGGGGCAGTGACTTTACACA	ND
stSG155425	CTCCAGGAAATCCTGATGGA	TCCACTAAATAGAATGGGGGG	60
stSG155426	AAAAAAAAAAAACAGAAGGGGC	GAGACTTGTGATGATGCTCAGC	60
stSG155427	ATGTCACGTCACACTGAGCTG	GGAATTTTGCCATTAGATGAGG	60
stSG155428	TGAGCTTTGTGTCAGGGATG	GGTCTCTTAGACTGGGATGCC	60

stSG155429	GCTGCGGTGTTTATACCTATG	CACATAAACGAAATCATGCTCA	60
stSG155430	ATGACATGGAGAAAGCAGGG	GGCACTAAGCAATTGGTGGT	60
stSG155431	TTGTCCCTGAAACAAAAGCA	TTTGTTATGTACAATGTTGGCC	60
stSG155432	TGTATCCATTTCCCTCATTGC	CAAGTACCCCACCTGATGCT	60
stSG155433	GATGGATCATTTCATGGATGG	ACAACAGCCAAATACAGCCC	60
stSG155434	ACCTTTCCTGGAACCCTCAT	CTGGAAAATAAGTTTCATGGCA	60
stSG155435	ATCATTCTCTAGGCCTGCCA	CCCAGCAAACATTCCATTG	60
stSG155436	TCTCCAGGCTGCTAGGATGT	GAAAAAACCCAAGCAGAAAGG	60
stSG155437	AGGGCAGCGATCTGTTTG	GGATTATCCCAGGGACACCT	60
stSG155438	CTGCTCATAACCCAAATCTTCC	ATTTCCACGGAGTGAGATGC	60
stSG155439	CCATTGTTCCACACAGCAAG	TTTACCCAAAGGGAGTGTGC	60
stSG155440	TCTAGCACGGCAACAGTCAC	ACTAACAGAAAGGGCCTTTGG	60
stSG155441	TTTCCTCTGAATGGACAGTGG	GGGGGGGAGGTGAGATATAA	60
stSG155442	TAGGCAGCGTTGATAAGTTCTG	CAACATGCCTATCCACAAAGG	60
stSG155443	TGGTTACGTGTTTTCCTTTGC	GTGGCAATATGCTACAGACTGC	60
stSG155444	AGTCGAAAGGGCTGTGAGAA	GGGTTACAGCTTGCTCTTCG	60
stSG155445	TCCTTTTTGTTGTTGTTGTTGC	AACACCTATCCACCCAAAACC	60
stSG155446	CCAGCCAGTTTTTAGTGGCA	AGGAGTCAATGGGATTTATCCT	ND
stSG155447	TTATGTACTGCGTGGAAGCG	GGGAATCTTCTGTTGACACCA	60
stSG156835	TGGGATTTATCCTAGGGCTTA	CATTGAAAATAACACTTCCATGACA	ND
stSG156883	CCATCCAGATAGCCAAATGG	TTGCAATTAATGACAGTGATGC	ND
stSG156884	TGTCAAAACCTGAGTCATCCC	TTATCCAGTGAGGAATGTGGC	ND

Appendix C

This appendix contains information for primers used in the studies presented in Chapter 5. stSG numbers, the locus the STS was designed to, the primer sequences, expected product sizes, optimal pre-screening annealing temperatures in ^{0}C and indication if the primers span introns are given. N/A denotes instances where a pre-screen was not performed as the primers span an intron.

stSG number	Locus	Primer 1	Primer 2	Product size (bp)	Pre-screen	Intron spanning?
stSG158852	dJ77O19.CX.1	AAACTTACCATTGGTGCATATG	CAATGGAAGCACCCATCA	137	60	no
stSG158853	cV362H12.CX.1	AAACTTCCCATTGGTATGTAAA	AACTATTGAGGCACCCATTG	139	60	no
stSG158855	dJ198P4.CX.1	CCCTAAAGTTATTACGGAAACAGA	CCACGTAAACAAGTGACAGGT	96	60	no
stSG158856	dJ79P11.1	CCCTGAAGTTAATAGGGAGACC	CCCACAAGAAATAGGTAACATCA	95	60	no
stSG158857	dJ635G19.2	AGGCACTATATGCGCTTCC	CATTAACCTCAGCGAAAACTTT	94	60	no
stSG158858	cV351F8.CX.2	ATGTGAACCTTTTGGCATTCT	ATGAAAGTTGATGAAAGTCCTCTT	103	60	no
stSG158860	NGFRAP1	GAACCCTATGTTATTTCCATGTGTC	ACTACTGCTGACAGAAACTTACACTG	99	60	no
stSG158862	dJ769N13.1	GTCCTGGTTCTGGGATGGA	CCACAGTTGCTTCCATTGT	246	60	no
stSG158864	dJ769N13.CX.2	CCATTTTCAGGGAAGTTAAAGAG	CTTACTTAAAACGAACTGGTTTGC	291	60	no
stSG158865	dJ1100E15.CX.3	TAAGACAAAACCCCTGGCAGA	AATGCTGGTCTCTTCCCTTC	199	60	no
stSG158866	dJ769N13.CX.1	GGTCCAAGCTCAGGACAAATAG	TTTCTGGGCCCTGAGTTTCA	161	60	no
stSG158869	cU250H12.CX.1	GCTCCAGGAAAAACTCTGGTTA	CCCGTGAGAAACTGAAAACTA	82	55	no
stSG158870	cU237H1.1	GCTCCAGGAAAAACTCTGAATG	CCCATGAGAAACTGAAAACGT	87	55	no
stSG158871	cU237H1.1	ACAAGATGTTCTACAAGATATGAAGC	TCTTACTGACCCATTGAAACCA	100	55	no
stSG158905	cU116E7.CX.3	CCATTTCCCTTTCATCTTTTC	CTGGAGCCTCTGCTTGTGTT	112	n/a	yes
stSG158906	cU46H11.CX.1	CATTTCCCTTTCATCCGTGA	TTTCCTGGAGCCTCTGCTC	119	n/a	yes
stSG158907	cU209G1.CX.1	TCTGGAAGTAATGCACATTGTAAA	CCACACAAAGGACCAAAGATTAC	226	60	no
stSG158908	cU61B11.CX.1	TATTGAGATATTTGCAGTTGGTACG	AACCAATTCATTGTCAGTTTAGGT	285	60	no
stSG158909	dJ545K15.1	GTGGACAAAGAACATCAAATTAC	GCGTGTAGAAGAAAGAGCAAA	174	60	no
stSG158910	dJ545K15.2	CTATTCATTTTCTCCACCTTGTTT	TGTCACTTATGCAGAAGAAATAGC	225	60	no
stSG158911	cV602D8.CX.1	CCCTCTATAATTCTTACGTGGAATC	GAAATTTCTTCAAGTCTTTTGACG	269	60	no
stSG158900	dJ122O23.CX.1	GGCTAGGGTGGAGGATAAAAG	CTGAAACATAGAGCAAGCAAACAC	289	60	no
stSG158901	cV351F8.CX.1	CTGCTGTGCACATCCCTATG	TGAGTATGCAGACCCAGCAC	115	60	no

stSG158902	cU177E8.CX.1	TCGAGGTGAGGGAAGAGAGA	CTTCTTCGGAAGGTTGAGGA	203	n/a	yes
stSG158903	cU177E8.CX.3	AAGGCAACTGCCCTACAGC	TTCTAGGTTTCTTTCATTCTCTGG	178	n/a	yes
stSG158904	cV857G6.CX.1	GGCTAAGGTGCAGGATGAGAA	TCCTACTGAGAACACTGCATTGC	284	60	no
stSG158913	cV857G6.CX.2	GAGATGTGTCAAGGGCTCAA	CCAGCAATATTCTCATCAGAGAAA	219	60	no
stSG158914	TCEAL1	CTCGCTTAAAGTTGAGGTTTCC	CAACCACTGTTATGCCTTTGAA	225	60	no
stSG158930	pp21 homolog	CCGTGGAAGGAGTCAAACT	GGATTTCATCTTGTATCTGTCTACC	193	n/a	yes
stSG453302	NXF2	AGCCTAAGGCCTGGCTGAC	CAGAGCAACTGGCTCTTTGG	81	60	no
stSG453303	NXF5	CACCAGCCTAAGACCACTCTAAG	CACAGAACAACTGGCTCTTCAG	112	60	no
stSG453304	NXF4	TACCAGCCTAAGGCCGTG	CACAGAACAACTGGTTCTTCAA	112	60	no
stSG453305	NXF3	GGGCTGACTCTCATCCTTCC	CCATACTTTATTGTGAAAAGGAACA	98	not tested	no
stSG453306	NXF1	CTCGCCTGCCTTCTGGAA	TAAGGAGGTCCTGGGGTTAAGT	135	60	no
stSG482247	cV857G6.CX.2	TGGAGGTAGGGGCCAGAGG	ACCAGGGAAAGCAGGGCCA	108	not tested	no
stSG482248	cV857G6.CX.2	GCTTCCGCGAGCCGGAGA	TCCTCCTTTTTAGGCCTCGC	93	n/a	yes
stSG482249	TMSB4X	TGCAAAGAGGTTGGATCAAG	CTTCCTTCACCAACATGCAA	152	not tested	

Appendix D

This appendix contains information for primers used in the studies presented in Chapter 6. stSG numbers, the gene the STS was designed to, the primer sequences, expected product size and optimal pre-screen annealing temperature in ^{0}C are given.

stSG number	Gene	Primer 1	Primer 2	Product size (bp)	Prescreen
stSG427961	ALEX1	AGATGGCTGGGCTAAGACTG	GATGCGAGCCCTTCATTTT	307	60
stSG187942	ALEX2	AAAATCAGGGCCGGCTTC	GATGCTGGCACTTGGGTACT	280	60
stSG407290	BMX	CCCACCTGCTGGTCAAGTA	TGCTTCCCAGAGGGTACATC	104	60
stSG407309	BTK	CCGGTACAACAGTGATCTGG	TTCCATTCCTGTTCTCCAAA	125	60
stSG427959	cU46H11.CX.1	TCAGCATTCTCAAGGAATTCAA	GCATGTTCTGCCATCCAATA	282	60
stSG427962	cU209G1.CX.1	TGTGCAGTTGGCTGGACTAA	TCAGTGATGTTGGTGCATTG	212	60
stSG187911	CYBB	TCACCCTTTCAAAACCATCG	CCACGCATCTTGAAACTCCT	235	60
stSG427960	dJ545K15.1	GCGGCAGGACTGATGATT	GAATGGCCGCTGTTTTATTG	285	60
stSG407294	DMD	TCCAGCATTACTGCCAAAGTT	GCTCCCCTCTTTCCTCACTC	102	60
stSG407311	DRP2	CTTTAACCACAGCCCTGGAA	CATTGAGGAGCCAATTGAGG	180	60
stSG407317	GLRA4	TCGTTTCTATTTCCGTGGCTA	CTGGTGGATATCTTCTGACCA	259	60
stSG187917	GLRA2	CCTGCATTGAGGTCAAGTTTC	CCTGAACTCTGGGTGGTCAT	184	60
stSG187903	GPM6B	ATGCTGCATCAAGTGTCTGG	CTCAGCAAGGCATGGTCAC	171	60
stSG187941	GRPR	CCAGACCTTCATTAGCTGTGC	CCTTCCACGGGAAGATTGTA	174	60
stSG187907	IL1RAPL1	TTGAAGATGTGGCAAGATGTG	CATTTTGGTCCATGCCATTT	253	60
stSG407310	IL1RAPL2	TATTCTCAGACGGGGATGGA	TTTTGGATCCCTTCCACTTG	188	60
stSG407318	KIAA0316-like	TGGCATCAGCCATGTTATTG	CCATGATGCTGGTCTCCAC	128	60
stSG407292	KIAA0316	TGACACAGGCAATCCCTTTT	TACTGAGCGAACGACCACTG	149	60
stSG187894	MID1	GAGGTCTAGACGGGCTCAAG	ATGTGAGAGTCCGGAATTGG	307	60
stSG407305	MID2	GCATCACCTGTGAGGTCTCC	AATCGATCATTCAGGGATGC	247	60
stSG187946	NADE	TGGCAAATATTCACCAGGAA	ATTTCCTGCAGGCTGGTG	104	60
stSG407312	NOX1	TTGAAGTGGATGGTCCCTTT	TTTGAGGTTGTGGTCTGCAC	152	60
stSG187944	NXF2	GTCTTTGAACTTGTGCAACAACA	ATTTTTGGAGAGATTCAGGGTCT	100	60
stSG407308	PLP	CCATGCCTTCCAGTATGTCA	CCTAGCCATTTTCCCAAACA	246	60

stSG187936	POLA	CGGAGTGGATTTGGTGAAAG	TTTTCCATTGGGATTACAACC	155	60
stSG407306	PRPS1	GTGACCTCCATTGCAGACAG	AGTGTCAGCCATGTCATCCA	144	60
stSG187896	PRPS2	CAGGTTGAATGTGGAATTTGC	AGTGTCAGCCATGTCATCCA	127	60
stSG187898	RAB9A	TACCATGCAGATTTGGGACA	TGGCATCTTTTGCACTTGTT	308	60
stSG407307	RAB9B	GGAGGTAGATGGACGCTTTG	GGGTAATCCCCATTCTCCAT	300	60
stSG187938	RAI2	TTCTGTGGCAAGATCAAAGG	TTTTTGATTGGGAGCATGTG	209	60
stSG187923	SAT	GCAGCATGCACTTCTTGGTA	TCACTCCTCTGTTGCCATTTT	149	60
stSG407313	SRPUL	AACGTCAACGTCAACTCAGC	TGCATTTTATAATATCGGTTGGAA	107	60
stSG187914	SRPX	ACGTCAATGTGGGTGTCAGA	ATTCCTAGCTGGAGCCGGTA	121	60
stSG407315	SYTL4	TAAAAAGGAAAGGGGCCAAG	CACCTCCAGGTACCATTGCT	169	60
stSG187916	TM4SF2	ATCACTGGGGTGATCCTGCT	GTTTCAGCATCCATGGGCTA	187	60
stSG407314	TM4SF6	ACTGGCGTTATCCTTCTTGC	TAGCATCCATGCAGAAGCTC	180	60
stSG427958	TMSNB	GATAAGCCAGACTTGTCGGAAG	TTTCCTTTGAGGGAAGAGTATTTT	94	60
stSG407316	XKL	ATCCTCTTTGAGCCCTGGAT	TCTCTGTCTGCCAACCTCAA	176	60
stSG187919	XK	CTTCCTGTGGAGGAGCTTTG	ATGGGACTTGCTGATGAGGT	304	60
stSG187921	SYTL5	ACCTGCTCCCTGATGATAGC	CAAACGAACTCCTCCCAGAA	194	60