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.

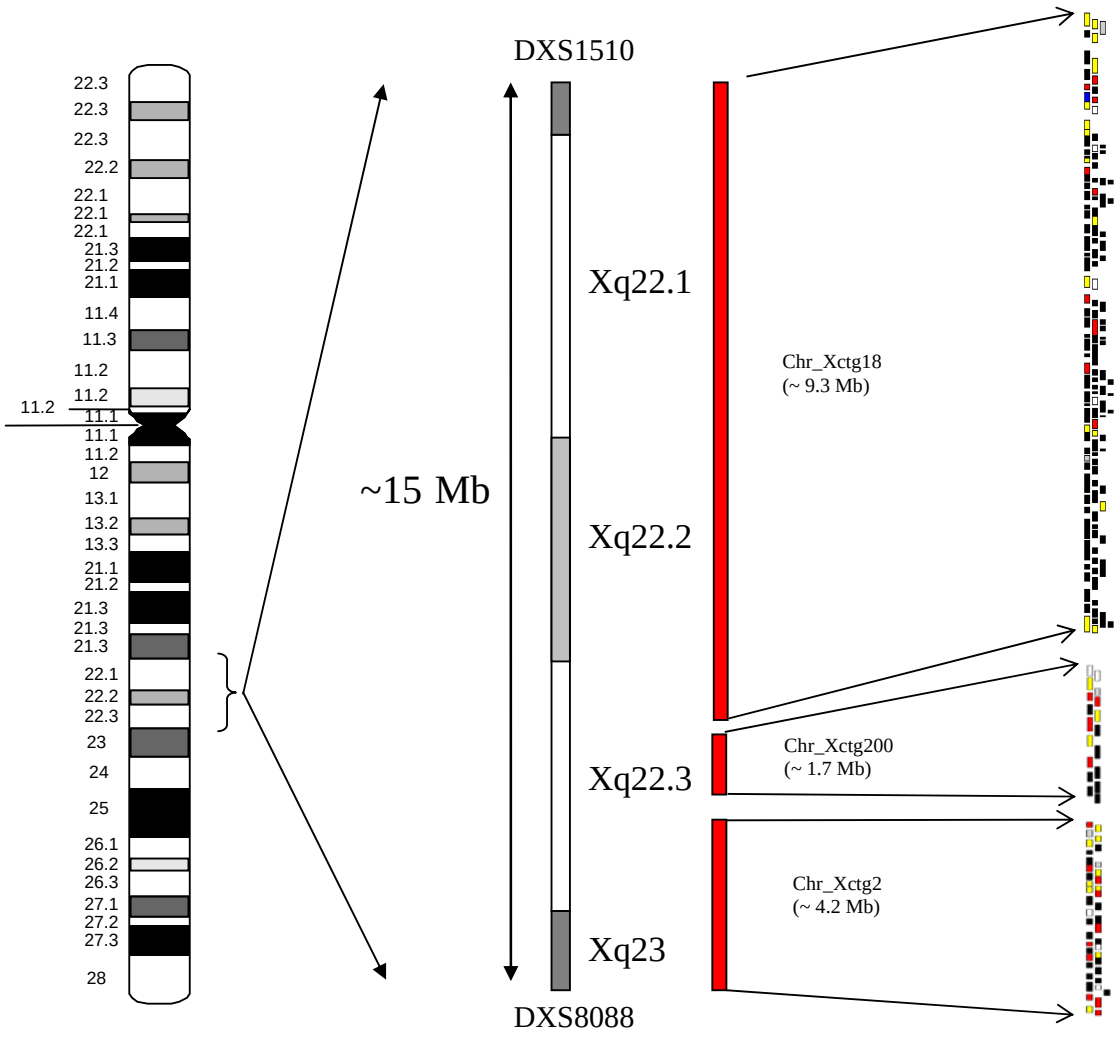


Figure 3-1 G-banded ideogram of the human X chromosome (Francke), illustrating the region Xq22-q23 chosen for this study. The ~15 Mb region bounded by markers *DXS*1510 (Cen) and *DXS*8088 (Tel) is shown. The three sequence-ready contigs spanning the region are shown (red bars) and a depiction of the 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.

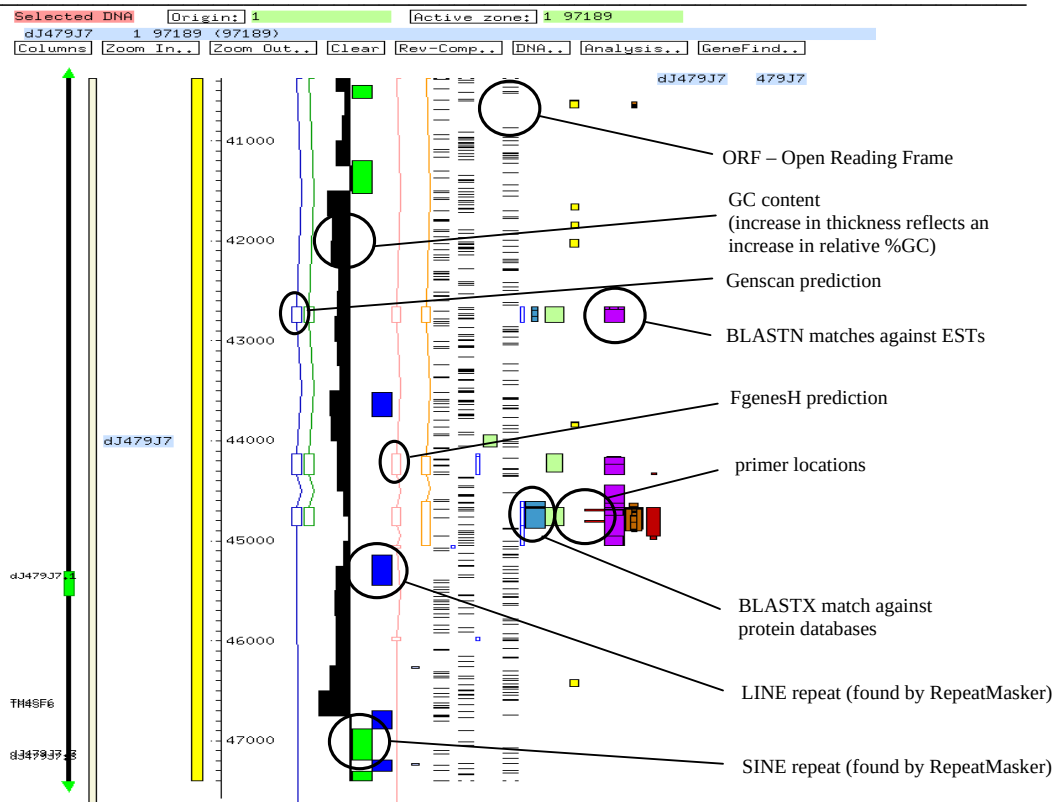


Figure 3-2 The image above illustrates how results of sequence analyses were collated and viewed within Xace. The yellow bar to the left of the image represents a section of the genomic clone's sequence.

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.

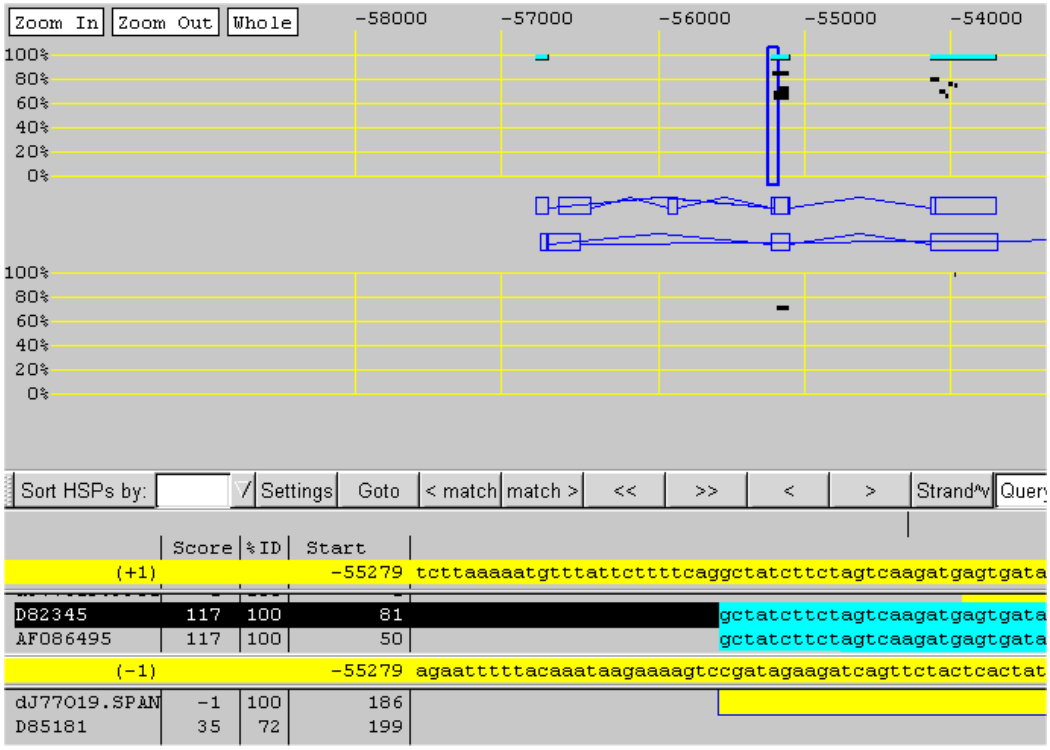


Figure 3-3 Example of an mRNA match viewed using BLIXEM. The diagram illustrates BLASTN matches to mRNA accession D82345 to genomic clone AL035609. The vertical blue box represents the position of the region of alignment highlighted in the lower section in context with other matches to the highlighted 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.

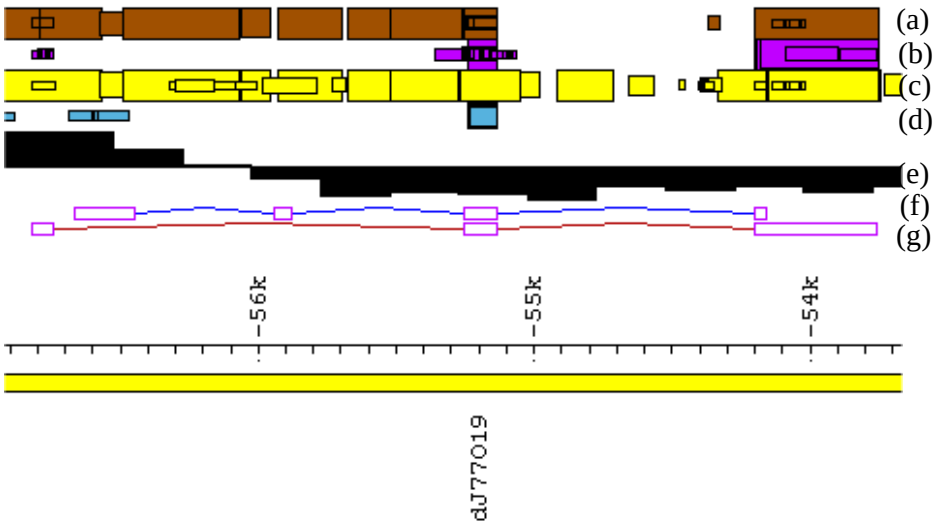


Figure 3-4 Example of a "gene" (GD_mRNA) structure, for locus dJ77O19.CX.1. In this case, the gene was annotated from mRNA accession D82345. The diagram shows an ACeDB representation of the gene structure. Key – (a) mRNA BLASTN matches, (b) EST BLASTN matches, (c) genomic sequence BLASTN matches (d) 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.

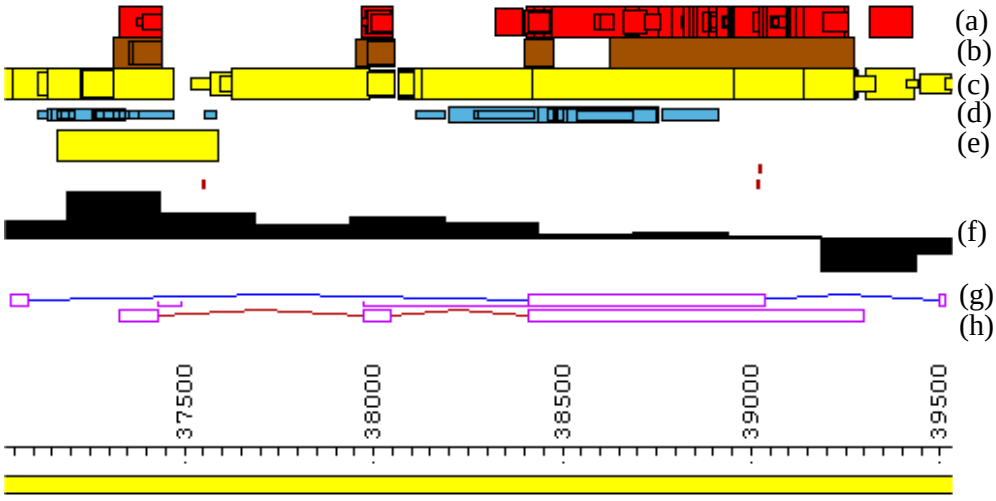


Figure 3-5 Example of a "predicted gene" (GD_composite) structure, for locus cV857G6.CX.2. This locus was annotated from overlapping EST sequences. The diagram shows an ACeDB representation of the gene structure. Key – (a) EST BLASTN matches, (b) mRNA BLASTN matches, (c) genomic sequence BLASTN matches (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.

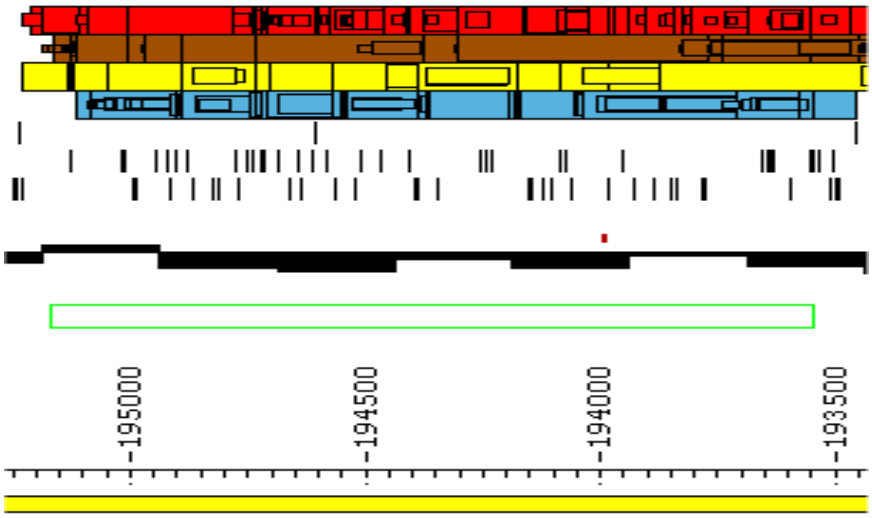


Figure 3-6 Diagram illustrating a "pseudogene" (pseudogene structure) structure, for the pseudogene locus dJ232L22.CX.2. The diagram shows an ACeDB representation of the gene structure. Key: mRNA/protein homologies as in Figure 3-5 above, vertical lines – boundaries of open reading frames (one row for 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.	

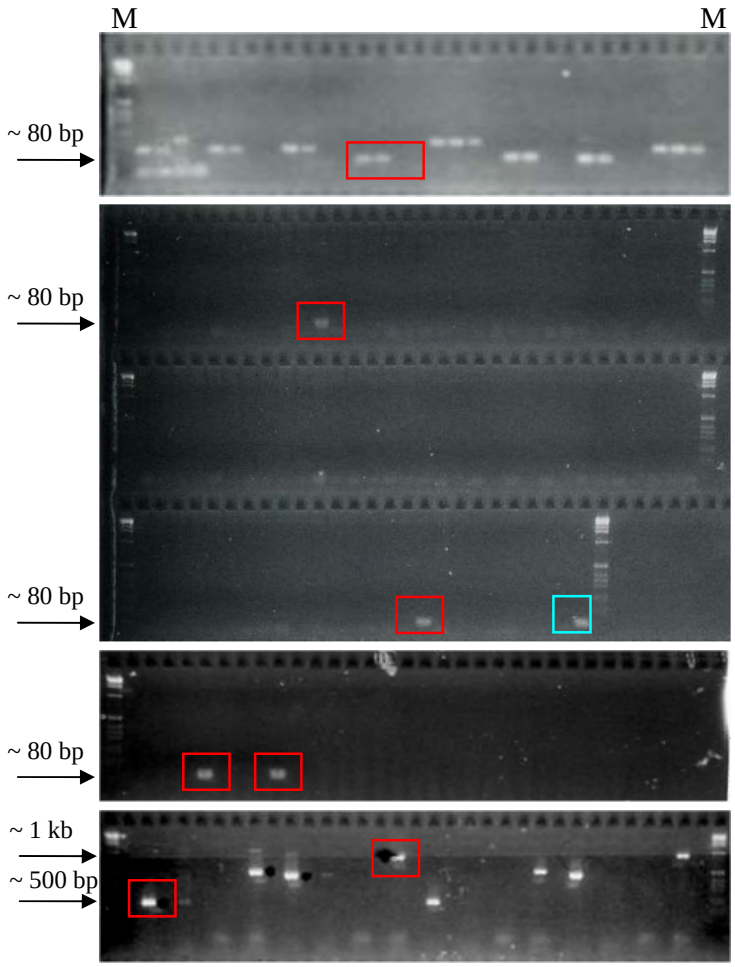
	gene					
		88384	60	NONE		
dJ769N13.CX.2	gene	88385	FAIL			
dJ839M11.1	predicted gene	88386	60	DAU, HPB, FB, SK, FLU	DAU 1. HPB 24	
dJ839M11.2	predicted gene	88387	FAIL			
dJ889N15.1	predicted gene	88388	60	T	T 9.	6397/6398
dJ889N15.1	predicted gene	88389	60	T	T 9.	6399/6400/6401/6402
dJ889N15.CX.1	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 55^0C are shown. M denotes 1kb ladder.

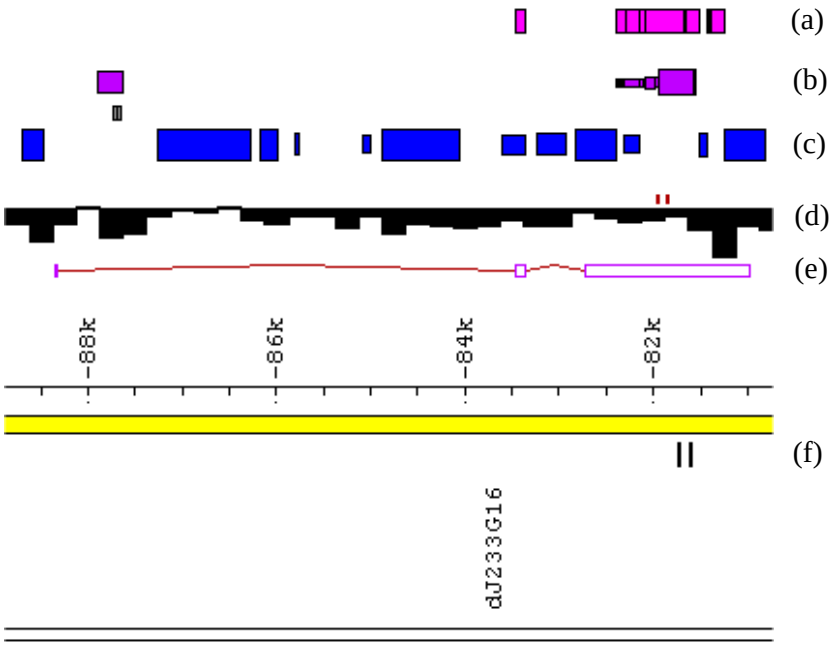


Figure 3-8 Diagram illustrating annotation of a gene structure using SCCD sequence. The diagram shows an ACeDB representation of locus dJ233G16.CX.1. Key – (a) SCCD sequence BLASTN matches, (b) EST BLASTN matches, (c) LINE repeats, (d) GC content (increasing upward thickness of bars represents increased %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.

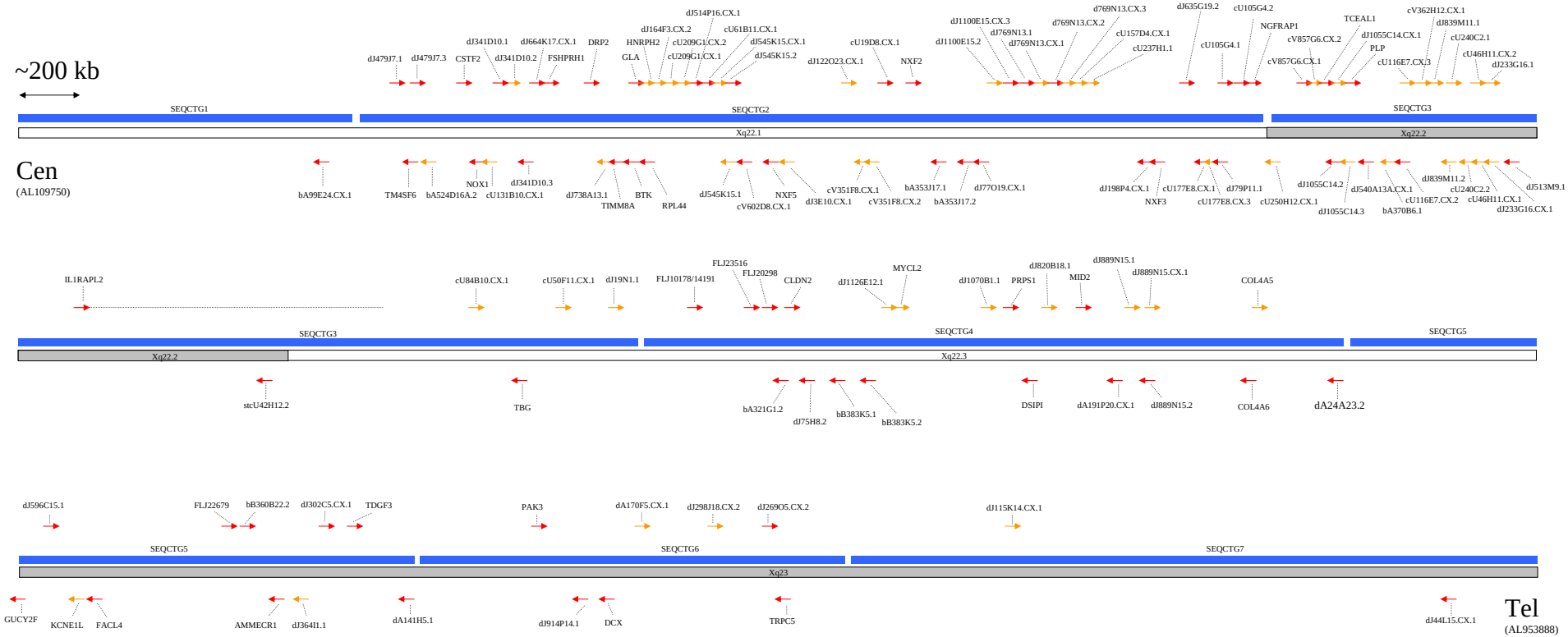


Figure 3-9 Genes annotated on finished sequence of the human Xq22-q23 region from clone dJ902O5 (AL109750) to dJ137P21 (AL953888), annotated as described in this Chapter. The region beginning is at top left, continuing onto the lower sections of the diagram. "Cen" denotes the centromeric end, "Tel" the 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.

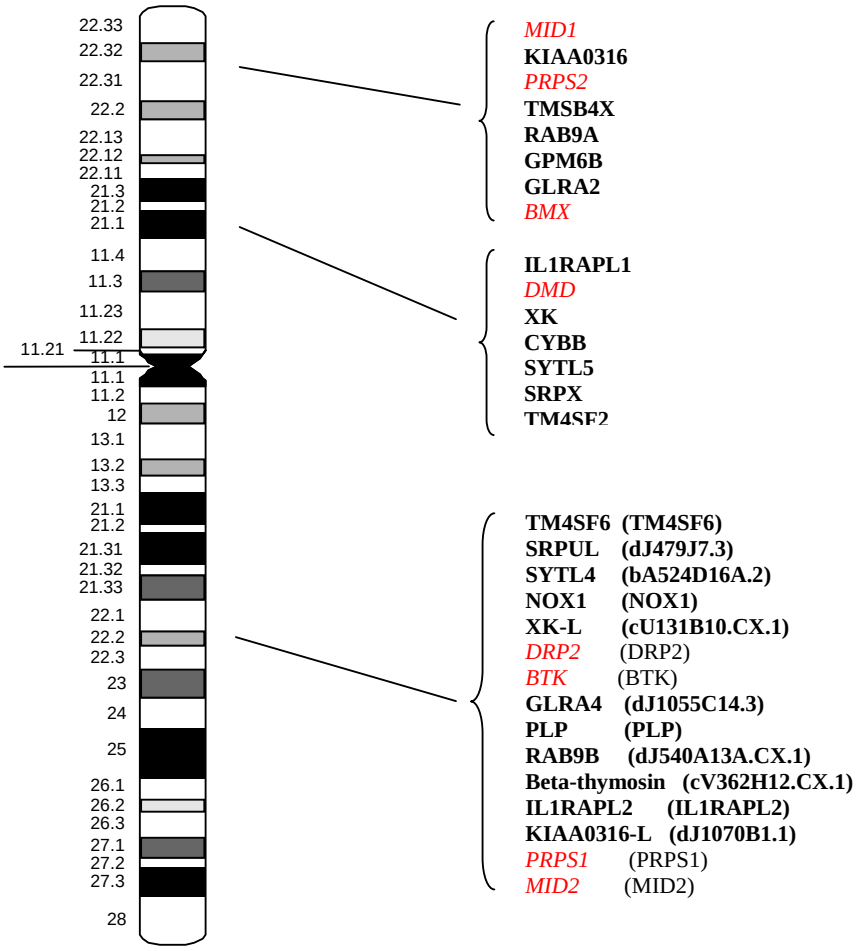


Figure 3-10 Observations of Xp/Xq paralogues. Xp/q paralogues noted by Perry *et al.*, (1999) are in red italic type, new observations are in bold type. Locus names assigned during annotation of Xq22 are given in parentheses.

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.

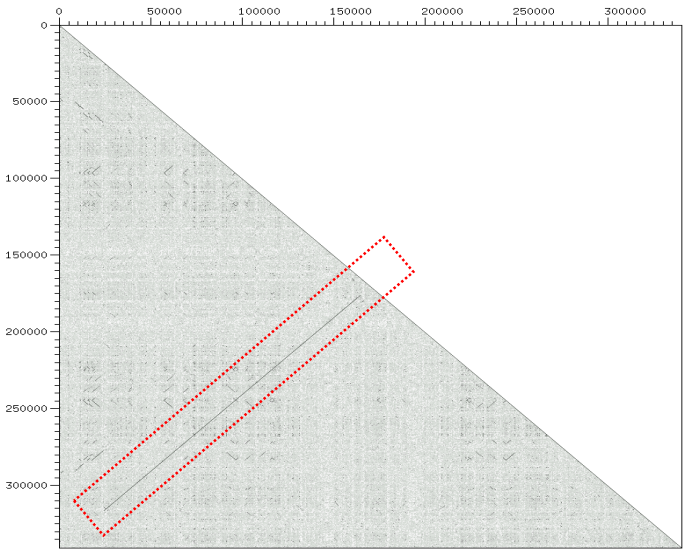


Figure 3-11 Diagram showing results of Dotter analysis of the genomic sequence flanking the two NXF2 loci (against itself). The red box highlights the inverted repeat.

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.

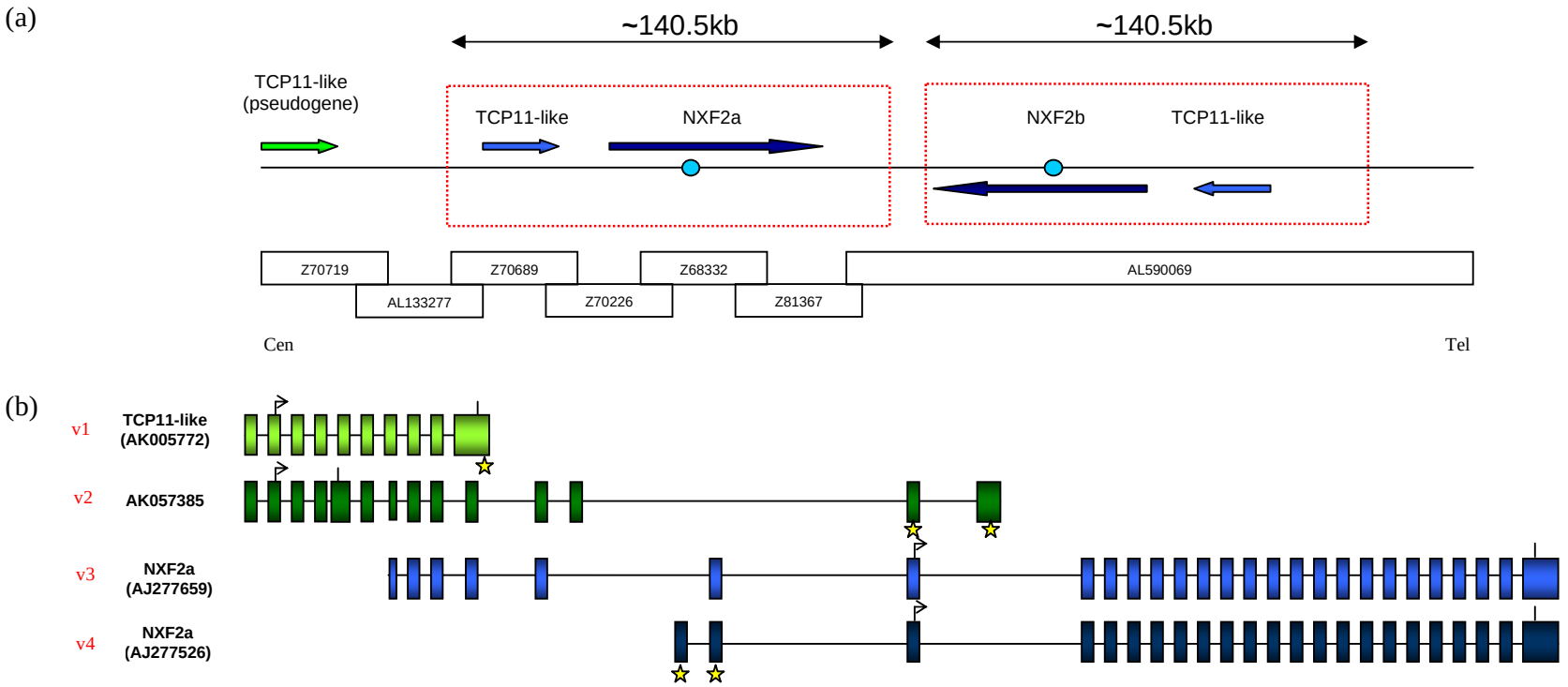
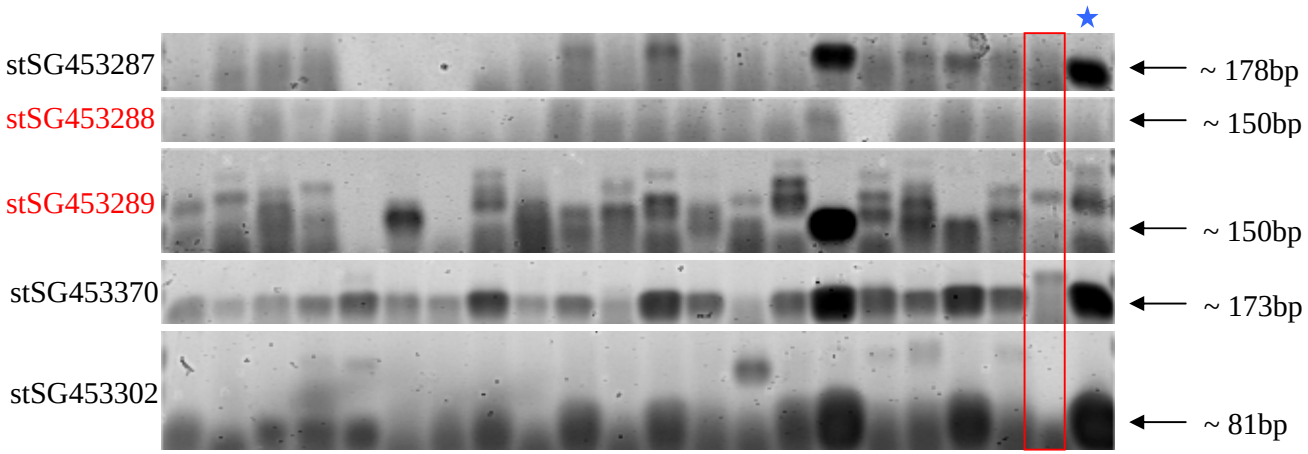


Figure 3-12 (a) A schematic representation of the inverted repeat containing the TCP11-like and NXF2 loci, and an adjacent TCP11-like pseudogene. The repeat boundaries are denoted by red boxes, and the location and transcriptional directions of relevant genes denoted by arrows. The blue circles represent 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).

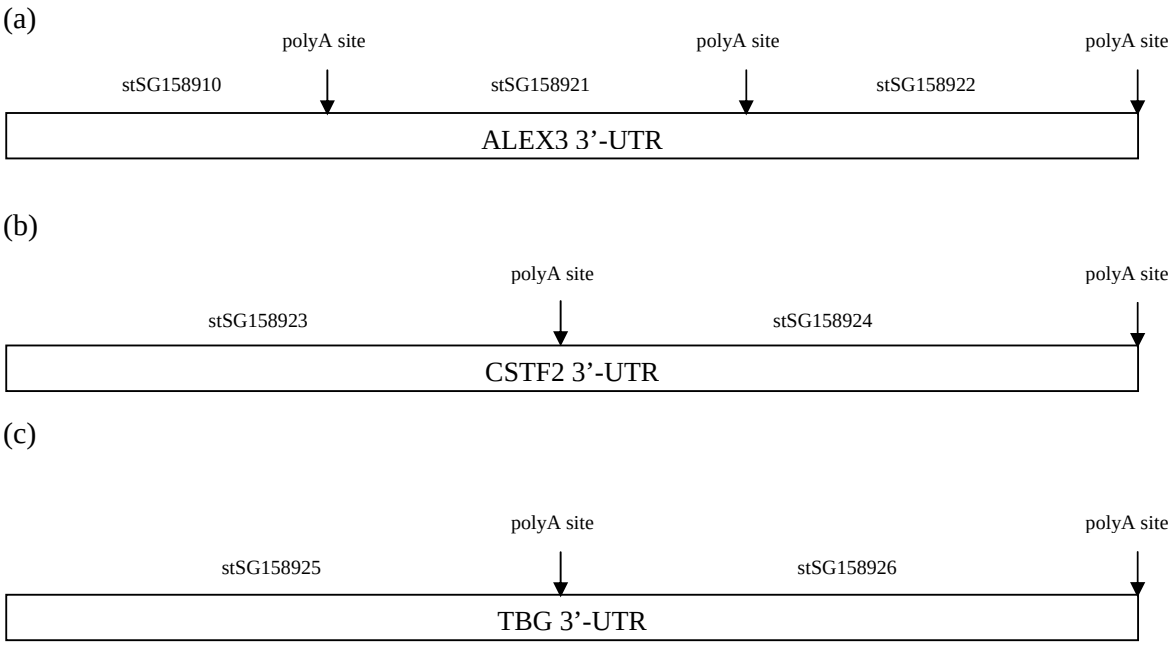


Figure 3-14 a schematic representation of the primer positions within the 3' UTRs of (a) ALEX3, (b) CSTF2 and (c) TBG.

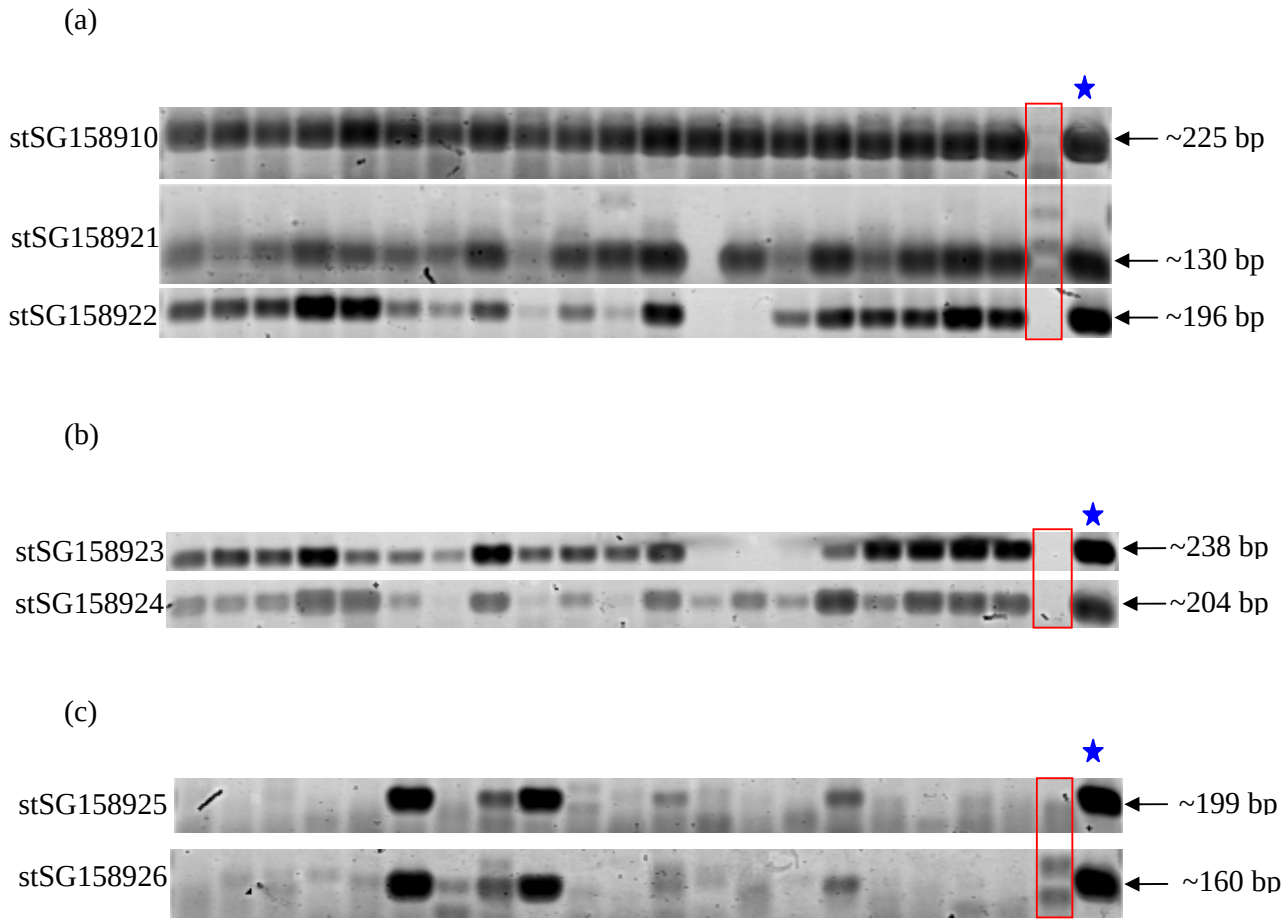


Figure 3-15 Images of 2.5% agarose gels containing RT-PCR products for primers designed to 3'-UTR variants, for (a) ALEX3, (b) CSTF2 and (c) TBG. Expected product band sizes are shown with an arrow. The red box in the gel images is the negative control lane, which whilst showing a faint product in some 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.

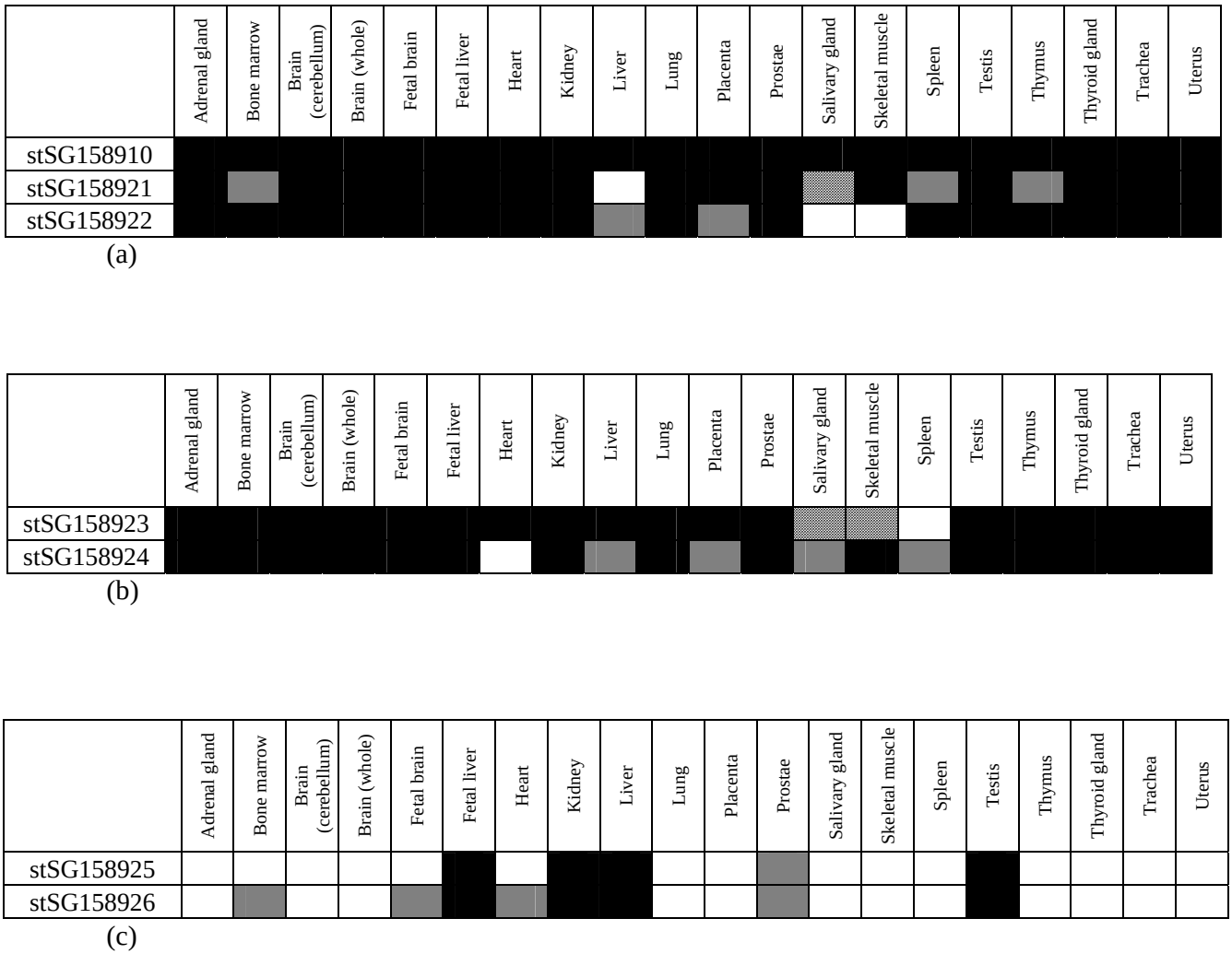


Figure 3-16 A tabulated summary of the RT-PCR results, with tissues tabulated according to the images shown in Figure 3-15, for (a) ALEX3, (b) CSTF2 and (c) TBG. Black filled cells denote medium to strong PCR product bands detected, grey cells denote weaker product bands detected and white cells denote 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.

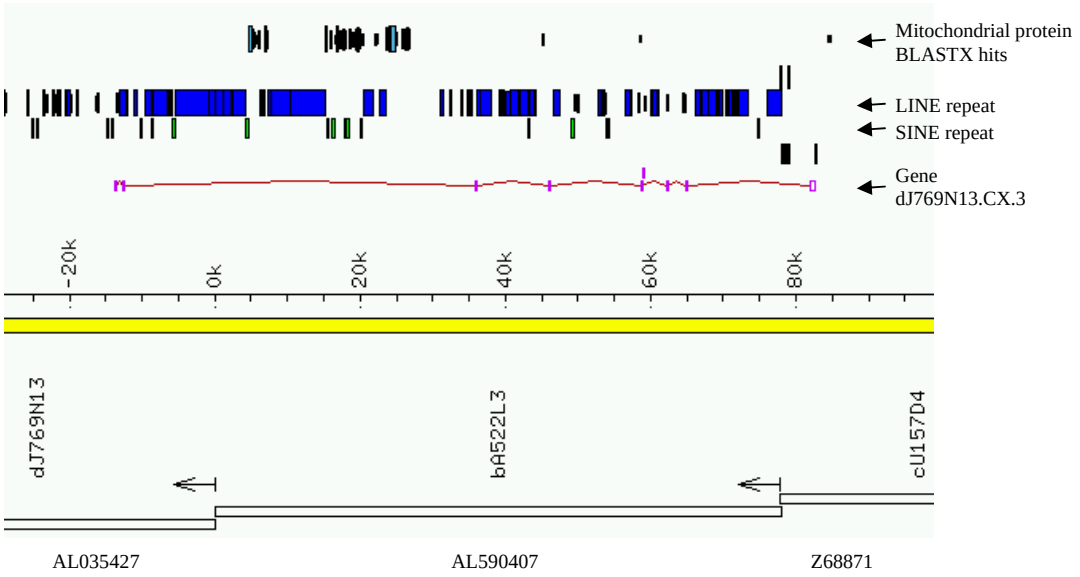


Figure 3-17 Xace representation of mitochondrial genome-encoded protein matches within clone bA522L3. Vertical magenta boxes represent exons of dJ769N13.CX.3, connected by horizontal lines representing introns. The positions of the mitochondrial protein homologies are shown. Genomic clones are depicted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

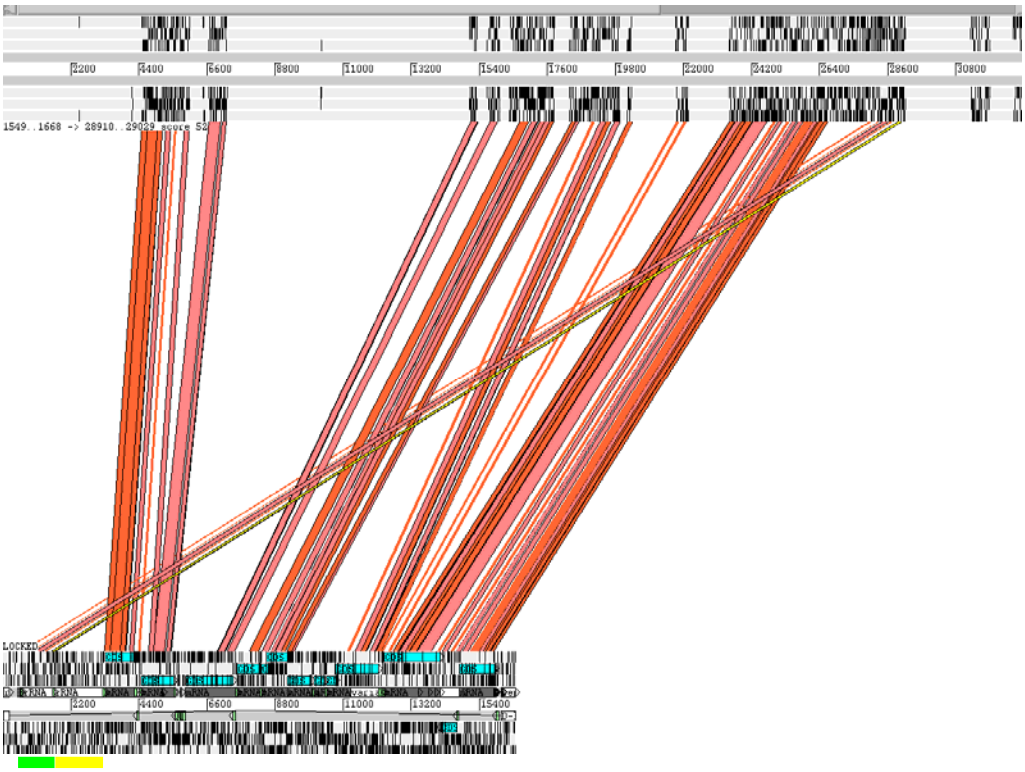


Figure 3-18 Diagram illustrating BLASTN matches between human Xq22 sequence (from clone bA522L3) and the mitochondrial genome. The matches were visualised using ACT (thanks to Dr. Julian Parkhill, Microbial Sequencing Unit, Wellcome Trust Sanger Institute). The red lines represent BLASTN matches. The 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.

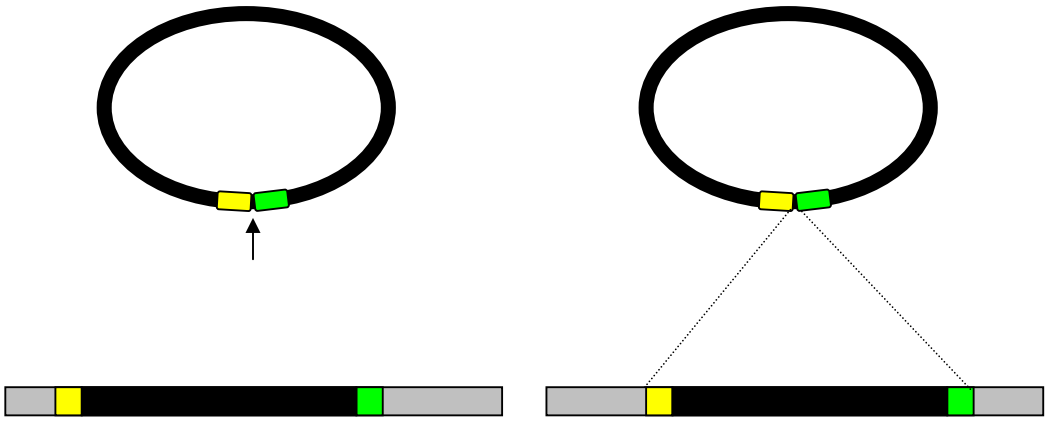


Figure 3-19 Schematic diagram of a model for integration of the mitochondrial genome in sequence accession AL590407 (bA522L3). Grey bars represent the nuclear genome, black bars the mitochondrial genome. The yellow boxes represents the 16S rRNA gene, the green boxes the 12S rRNA gene. The left section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