# Chapter 4

# Genome Landscape of Xq23-24

## 4.1 Introduction

The generation of large contiguous segments of human genome sequence allows for the systemic identification of the genes and other functional units contained within. As discussed in Section 1.3, one of the aims of studying the human genome is to identify all the genes present in the human genome which will provide the basis for furthering our understanding of the biological systems in which they are involved. There are also a large number of diseases for which the gene responsible remains uncloned. Identifying the genes within a region of interest because of the association with a particular disease enables screening for the causative mutations. For instance, a systematic search for the gene responsible for X-linked lymphoproliferative disease (XLP), previously localised to Xq25, resulted in the identification of SH2D1A, which was subsequently found to be mutated in patients with XLP (Coffey, A. J*., et al.*, 1998).

As part of the project to map and sequence the human X chromosome, the techniques developed during the construction of the sequence-ready bacterial clone contig in Xq22 (see previous chapter) were applied to extend this contig, and generate new contigs across all regions of the human X chromosome (Bentley, D. R*., et al.*, 2001). The sequencing of the X chromosome, based on these contigs, progressed rapidly in the Xq23-Xq25 region and was therefore chosen for an in depth study of the features encoded in the sequence.

# RESULTS

## 4.2 Identification of genes

As a result of the work carried out for the X chromosome mapping and sequencing project by the X chromosome group, four contigs covering 25 Mb of Xq22-Xq25 were generated and a minimum set of clones identified and sequenced (see Figure 4.1). This provided the raw data for the analysis and gene identification across an 8 Mb region between DXS7598 and DXS7333 that encompasses the distal portion of Xq23, Xq24, and the proximal portion of Xq25. There are currently twelve contiguous segments of finished sequence covering 7 Mb and a further 600 kb of draft sequence is available.

A combination of similarity searches (BLASTX and BLASTN) and *de novo* gene prediction were carried out on finished sequence (see Figure 4.2). Prototypical and consensus repeats were masked using RepeatMasker (Smit, AFA & Green, P. RepeatMasker at http://ftp.genome.washington.edu/RM/RepeatMasker.html), and the remainder of the sequence was aligned using BLAST (Altschul, S. F*., et al.*, 1990) to all known cDNAs, ESTs and other sequences to identify similarities with previously known genes from human and other species. A modification of this approach was to translate the genomic sequence into all possible reading frames and compare the sequence of the translation products using BLAST with databases of known protein sequences.

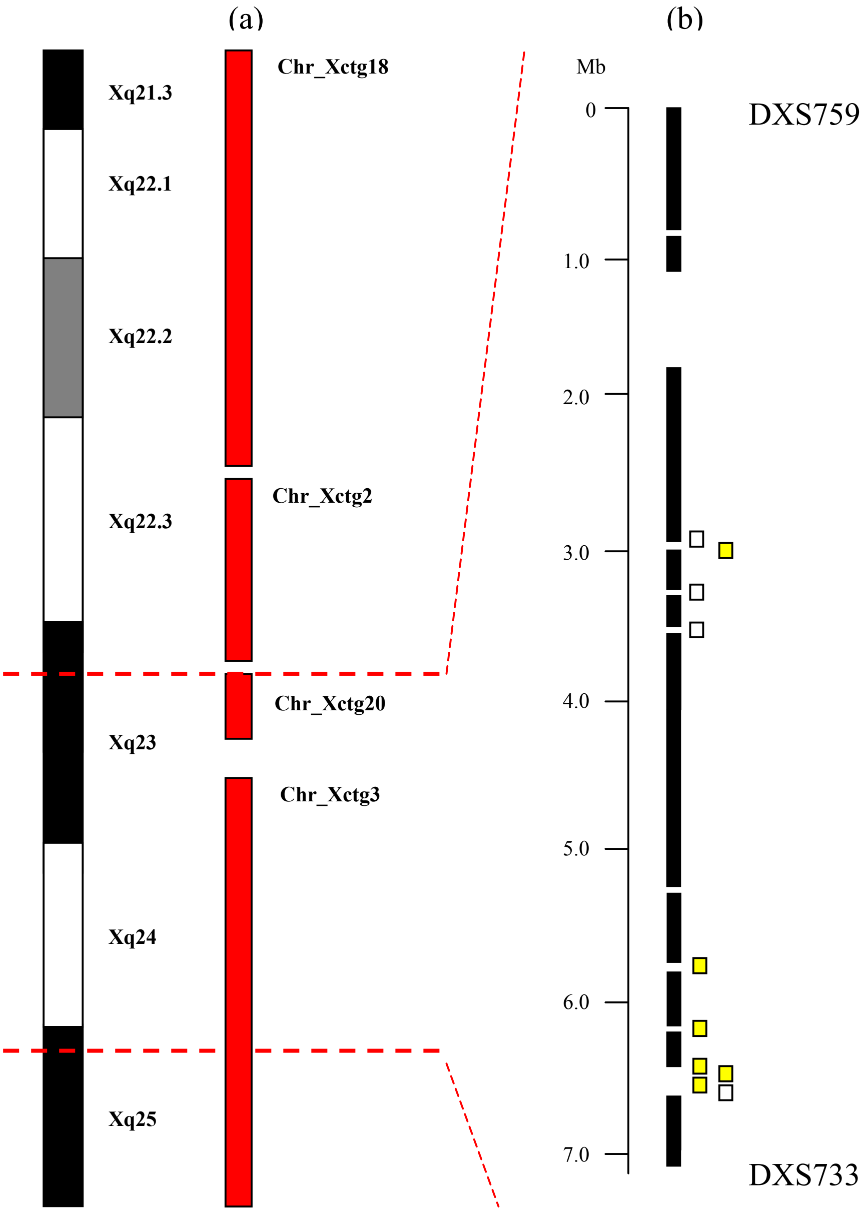


**Figure 4.1:** *Status of the region between Xq21.3 and Xq25 (a) Extent of bacterial clone mapping showing part of the X chromosome between Xq21.3 and Xq25, and the status of mapping and sequencing. Vertical red bars indicate contigs and their number from the Chromosome X mapping project are shown. The dotted red lines indicate the region identified for gene identification between DXS7598 and DXS7333. (b) The status of genomic sequence: black bars are continuous segments of finished sequence, yellow bars indicate clones with draft sequence available, white bars signify clones identified for sequencing, for which sequence is not yet available*

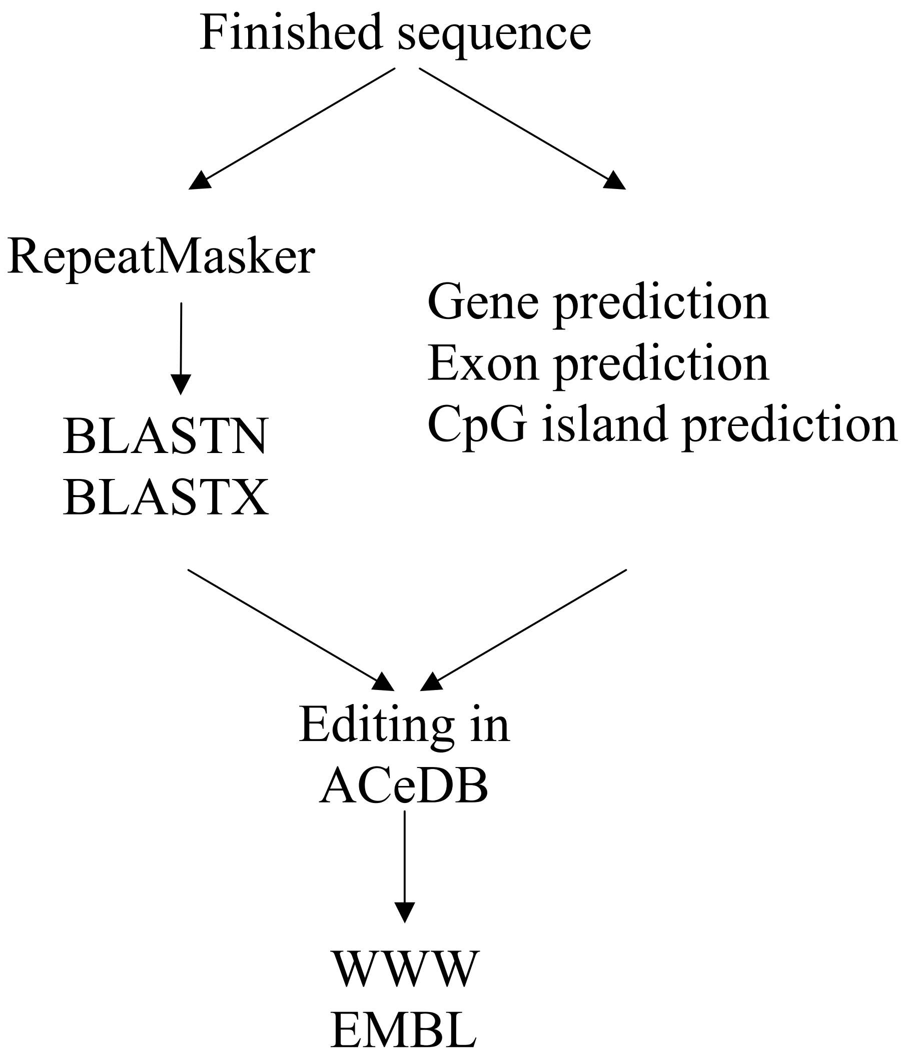


**Figure 4.2:**

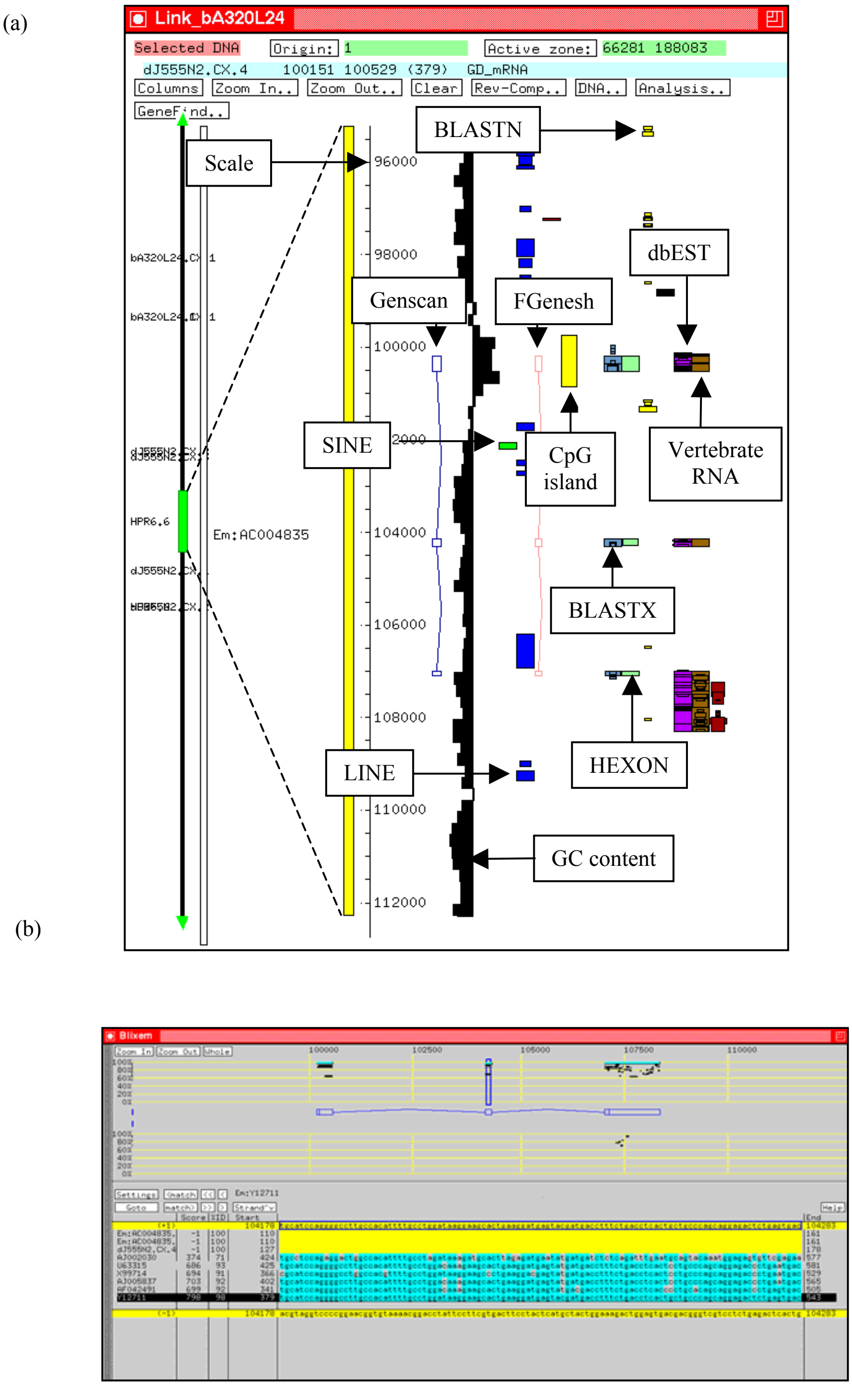
*Genomic sequence analysis. Finished sequence is analysed for repeats using RepeatMasker and compared to known protein and DNA sequences. De novo gene prediction is carried out on non-Repeatmasked sequence. The annotated sequence is viewed in an X chromosome ACeDB, Xace (see Section 2.23.3) and made available on the WWW (Sanger FTP site and EMBL).*

*De novo* gene prediction was carried out using a variety of prediction programmes to identify putative exons and genes. Also, given that CpG islands are associated with approximately 56% of genes (Hannenhalli, S.*, et al.*, 2001), a CpG island finder was used (courtesy of Gos Micklen). Members of the sequence annotation group at the Sanger Centre carried out the initial annotation of the genome sequence and the results were visualised in an X chromosome specific implementation of ACeDB, Xace (see Figure 4.3).

Initial analysis of the genomic sequence identified 19 genes that were previously known (see Table 4.1). In all cases the known genes were identified because there was a full length mRNA sequence aligning exactly to the genomic sequence.

**Table 4.1:** *Known Genes with full length mRNA sequence*

| Gene Name | Accession number | Reference |
|---|---|---|
| HOM-TES-85 | AF124430 | direct submission |
| hATB0+ | AF151978 | Sloan, J. L.*, et al.*, 1999 |
| ANT2 | J02683 | Battini, R.*, et al.*, 1987 |
| NDUFA1 | U54993 | Au, H. C.*, et al.*, 1999 |
| LAMP2 | J04183 | Kannan, K.*, et al.*, 1996 |
| GLUD2 | X66310 | Shashidharan, P.*, et al.*, 1994 |
| GRIA3 | X82068 | direct submission |
| T-plastin | M22299 | Lin, C. S.*, et al.*, 1993 |
| NRF | AJ011812.2 | Nourbakhsh, M.*, et al.*, 2000 |
| IL13R | X95302 | Caput, D.*, et al.*, 1996 |
| ZNF-kaiso | XM_010435 | direct submission |
| HPR6.6 | Y12711 | Gerdes, D.*, et al.*, 1998 |
| ZNF183 | X98253 | Frattini, A.*, et al.*, 1997 |
| UBE2A | M74524 | Koken, M. H.*, et al.*, 1996 |
| ATP1B4 | AF158383 | Pestov, N. B.*, et al.*, 1999 |
| SMT3B | X99585 | Lapenta, V.*, et al.*, 1997 |
| SEP2 | D50918 | direct submission |
| RPL39 | U57846 | Delbruck, S.*, et al.*, 1997 |
| U69a | Y11163 | direct submission |



Link_bA320L24

Selected DNA  Origin: 1  Active zone: 66281 188083

dJ555N2.CX.4  100151 100529 (379)  GD_mRNA

Columns  Zoom In..  Zoom Out..  Clear  Rev-Comp..  DNA..  Analysis..  GeneFind..

Blixem

Em:Y12711

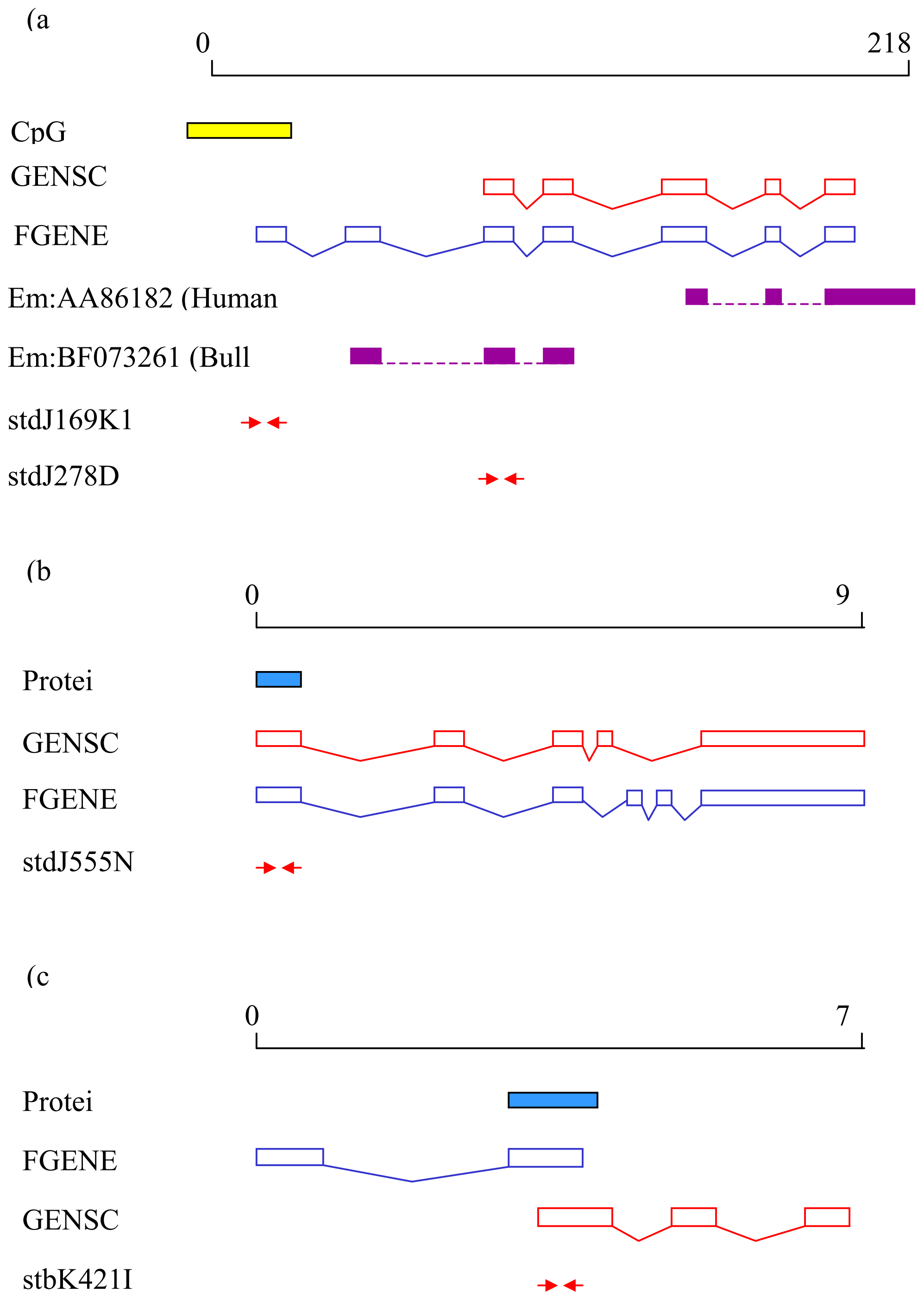
```
                  Score %ID   Start                                                                              End
(+1)                          104178 tgcatccaggggccttgccacattttgcctggataaggaagcactgaaggatgagtacgatgacctttctgacctcactgctgcccagcaggagactctgagtgac 104283
Em:AC004835.    -1    100     110                                                                                161
Em:AC004835.    -1    100     110                                                                                161
dJ555N2.CX.4    -1    100     127                                                                                178
AJ002030       374     71     424                                                                                577
U63315         686     93     425                                                                                581
X99714         694     91     366                                                                                529
AJ005837       703     92     402                                                                                565
AF042491       699     92     341                                                                                505
Y12711         798     98     379                                                                                543
(-1)                          104178 acgtaggtccccggaacggtgtaaaacggacctattccttcgtgacttcctactcatgctactggaaagactggagtgacgacgggtcgtcctctgagactcactg 104283
```

**Figure 4.3:** *ACeDB and BLIXEM (a) Example of annotated sequence in Xace. Features such as Genscan and Fgenesh predictions are labelled. The width of the each bar is an indication of the similarity between the genomic sequence and feature. (b) View of BLIXEM showing alignment of vertebrate mRNA sequences to the genomic sequence (see Section 2.23.4).*

Fourteen of the nineteen known genes have associated publications and the mRNAs of the remaining five genes were deposited directly into the sequence databases. The gene names are given as recommended by the Human Gene Nomenclature Committee (HGNC – http://www.gene.ucl.ac.uk/nomenclature). Although these genes needed no experimental verification, the precise exon/intron structure for each gene has been elucidated and their position and transcriptional direction on the genomic sequence in relation to neighbouring genes determined by alignment to the genomic sequence as part of this study.

In order to identify novel genes in the region, the genomic sequence was analysed for regions predicted to represent exons, based on sequence similarity searches and gene prediction programmes. Primers for the PCR were designed to regions with a variety of evidence suggesting the presence of a gene. For instance, eight pairs of primers were designed to regions predicted to be coding only by gene prediction programs, three pairs of primers were designed to regions predicted only by protein homology and two pairs of primers were designed to regions predicted only by EST homology. In some cases a protein or DNA sequence spliced across a series of exons in the genomic sequence and a predicted gene structure was identified and in other cases only a single exon was suggested. Examples of both predicted gene structures and a single exon region are shown in Figure 4.4.

**Figure 4.4:** *(see over) Examples of features for which STSs were designed for cDNA isolation. (a) A region of 218 kb was predicted to be coding by both GENSCAN (red boxes and lines) and FGENESH (blue boxes and lines), the 5' end suggested by the presence of a CpG island (yellow box). Two ESTs (purple boxes, splicing indicated by dotted lines) matched the genomic sequence exactly. Two STSs (indicated by red arrows) were designed, to generate novel cDNA sequence in regions not covered by the human cDNA sequence. (b) A region of 9 kb was predicted to be coding both by GENSCAN and FGENESH. A protein match (light blue box) was also observed in the first exon. (c) An example of a single exon feature where one exon from a GENSCAN prediction overlapped with one exon from a FGENESH prediction. A protein match was also observed.*


(a
0
218
CpG
GENSC
FGENE
Em:AA86182 (Human
Em:BF073261 (Bull
stdJ169K1
stdJ278D
(b
0
9
Protei
GENSC
FGENE
stdJ555N
(c
0
7
Protei
FGENE
GENSC
stbK421I

Analysis of the genomic sequence identified twenty-two predicted gene structures and twenty single exon regions. During the experimental verification process, mRNA sequences for five genes (T-plastin, ZNF-kaiso, UPF3B, NRF and HATB0+) were published and/or submitted to sequence databases by external groups and these previously predicted genes became "known genes" and are listed in Table 4.1. In order to analyse the 42 predicted gene structures and single exon predictions, a total of 58 primer pairs were designed either automatically using PRIMER (http://www.sanger.ac.uk/cgi-bin/primer3.cgi) or manually (see Section 2.15.1). Where possible each primer was between 18 and 20 nucleotides in length and had a GC content of approximately 50%. Primers were designed within a single predicted exon and pre-screened to determine the optimal annealing temperature for the PCR. PCR was carried out on pools of up to 19 different cDNA libraries (see Section 2.8.3) to identify the individual cDNA pool likely to contain the cDNA of interest. Each cDNA library in the panel comprises approximately 500,000 clones divided into twenty-five pools, each containing 25,000 cDNA clones. Five pools were combined to form a super pool containing 100,000 cDNA clones (cDNA library resources were kindly provided by Jackie Bye). For cDNA isolation using SSPCR, primers were initially screened against the super pools and then against pools representing up to five different positive super pools. For cDNA isolation using vectorette PCR, only the super pools were screened.  The results are summarised in Table 4. 2.

**Table 4.2:** *Experimental verification of predicted genes (see 2.8.3 for library pool codes); STSs are described in Table 2.6. Those superpools for which the equivalent pools were screened are shown in red. A gene name is given (last column) if the STS that was designed was used to generate novel cDNA sequence for confirmation of a predicted gene structure.*

| STS name | Evidence | Library Pools | Positive Superpool | Positive pool | Gene |
|---|---|---|---|---|---|
| stbA45J1.1.1 | EST, GENSCAN, FGENESH | V 1-11 | FL:B | - | bA45J1.CX.1 |
| stbA125M24.1.1 | Protein, mRNA (human) | V 1-11 | No pools +ve | - | bB125M24.1 |
| stbK421I3.1 | Protein, GENSCAN, FGENESH | S 1-17 | No pools +ve | - | - |
| stbK421I3.2 | mRNA, GENSCAN, FGENESH | S 1-17 | Uact:C | Uact:14 | bK421I3.CX.2 |
| stbK421I3.3 | EST, Protein, FGENESH | S 1-18 | WU:E, T:ABCDE | WU:22.23, T:1.2.4.6.7.9.10 | bK421I3.CX.1 |
| stdA155F9.1 | Protein | V 1-11 | WH:D, DAU:BCDE | - | dA155F9.CX.1 |
| stdA155F9.2 | Protein | V 1-11 | ALU:AE | - | dA155F9.CX.1 |
| stdJ29I24.1 | GENSCAN, FGENESH | S 1-18 | No pools +ve | - | - |
| stdJ57A13.1 | GENSCAN, FGENESH | S 1-17 | No pools +ve | - | - |
| stdJ57A13.2 | EST, GENSCAN | S 1-17 | WH:CE, NK:CE, DAU:B | WH:14, 24 | genomic contamination |
| stdJ93I3.1 | GENSCAN, FGENESH | S 1-18 | No pools +ve | - | - |
| stdJ93I3.2 | EST | S 1-18 | No pools +ve | - | Plastin 3 |
| stdJ169K13.1 | FGENESH, CpG island | S 1-19 | HPB:C | HPB:13.15 | dJ169K13.CX.1 |
| stdJ169K13.2 | Protein, EST | S 1-17 | No pools +ve | - | - |
| stdJ169K13.3 | Protein, EST | S 1-17 | No pools +ve | - | - |

| | | | | | |
|---|---|---|---|---|---|
| stdJ170D19.1 | GENSCAN, FGENESH | S 1-18 | T:E | T:21 | no cDNA product |
| stdJ170D19.2 | EST – no splice | S 1-18 | No pools +ve | - | - |
| stdJ222H5.1 | EST – no splice | S 1-17 | No pools +ve | - | - |
| stdJ222H5.2 | Protein | S 1-17 | No pools +ve | - | - |
| stdJ278D1.1 | Protein, EST, GENSCAN, FGENESH | S 1-17 | HPB:ABC, SK:AB, T:E | HPB:5, SK:5, T:21.22 | dJ169K13.CX.1 |

|  |  |  |  |  |  |
|---|---|---|---|---|---|
| stdJ278D1.1 | Protein, EST, GENSCAN, FGENESH | S 1-17 | HPB:ABC, SK:AB, T:E | HPB:5, SK:5, T:21.22 | dJ169K13.CX.1 |
| stdJ318C15.1 | Protein, EST, GENSCAN, FGENESH | S 1-17 | DAU:BCDE, HPB:BC, Uact:AD, FB:B | DAU:7.8.9, HPB:8, Uact:3.18, FB:8 | dJ318C15.CX.1 |
| stdJ321E8.1 | GENSCAN, FGENESH | S 1-19 | No pools +ve | - | - |
| stdJ321E8.2.1 | EST, GENSCAN, FGENESH | V 1-11 | T:BDE | - | dJ321E8.CX.2<br>dJ321E8.CX.3 |
| stdJ321E8.3.1 | EST, GENSCAN, FGENESH | V 1-11 | T:E | - | dJ321E8.CX.2<br>dJ321E8.CX.3 |
| stdJ327A19.1 | EST, FGENESH | S 1-17 | YT:ABCDE, HPB:AC, FB:ABCE, FL:BC, HL:ABCDE, SK:ABCDE, FLU:BCDE, DX3:ABCDE | YT:1.3.4.5, FB:7, SK:1.4.5, FLU:9, DXS:1.2.3.5 | UPF3B |
| stdJ327A19.2 | mRNA (mouse), EST, GENSCAN | S 1-17 | No pools +ve | - | dJ327A19.CX.4 |
| stdJ327A19.3 | BLASTX, GENSCAN, FGENESH | S 1-17 | YT:CD, HPB:B, FB:D, FL:C, HL:A, SK:ABC, T:E, AL:A, FLU:A | HPB:6.7, SK:2.3.4.5, FLU:5 | dJ327A19.CX.3 |
| stdJ327A19.4 | BLASTX, GENSCAN, FGENESH | S 1-17 | YT:ABCDE, HPB:BCE, FB:D, FL:C, HL:AE, SK:ABCD, T:CE, AL:AE, FLU:AD | YT:2.4, HPB:6.7, HL:5, SK1.5, FLU:5 | dJ327A19.CX.3 |
| stdJ327A19.5 | BLASTX, GENSCAN, FGENESH | S 1-17 | WP:A | WP:3 | dJ327A19.CX.3 |
| stdJ327A19.6 | BLASTX, GENSCAN, FGENESH | S 1-17 | WU:CE, YT:CD, DAU:D, HPB:ABCDE, FL:C, SK:ABCD, FLU:ACDE | WU:11.12, YT:13, \|HPB:7.8.10, SK:6.7.8.9.10, FLU:2.5 | dJ327A19.CX.3 |
| stdJ378P9.1 | EST – no splice | S 1-18 | DX3:ABDE, FB:ACE, FL:D, HL:E, FLU:BCDE | FB:4, FLU:4.6 | genomic contamination |
| stdJ394H4.1 | GENSCAN | S 1-18 | No pools +ve | - | - |
| stdJ404F18.1 | EST, FGENESH | S 1-18 | WU:AB, NK:ABDE, HPB:BE, | NK:3, HPB:10, BM:7 | dJ1139I1.CX.1 |

| | | | | | |
|---|---|---|---|---|---|
| | | | BM:B, HL:A, FLU:A, AL:E | | |
| stdJ404F18.2 | EST, GENSCAN, FGENESH | S 1-17 | WU:BC, YT:CD, NK:C, DAU:E, HPB:ABDE, BM:ACE, FB:CE, SK:BDE, FLU:E, DX3:C | YT:14, NK:15, SK:6.8 | dJ876A24.CX.1 |
| stdJ404F18.3 | EST, GENSCAN, FGENESH | S 1-18 | WU:BCE, YT:BCDE, HPB:ABDE, Uact:ABDE, DX3:AC, FB:CDE, HL:B, SK:BCDE, FLU:BE, ALU:B, AH:CE | ALU:6.8.10 | dJ876A24.CX.1 |
| stdJ452H17.1 | mRNA (mouse), EST | S 1-18 | No pools +ve | - | - |

| | | | | | |
|---|---|---|---|---|---|
| stdJ452H17.1.1 | mRNA (mouse), EST | S 1-19 | T:CD | T:13 | no cDNA product |
| stdJ525N14.1 | Protein, GENSCAN, FGENESH | S 1-17 | WU:E, T:ABCDE | WU:22, T:21.25 | dJ525N14.CX.1 |
| stdJ525N14.2 | Protein, EST | S 1-17 | NK:C | NK:15 | genomic contamination |
| stdJ525N14.3 | FGENESH | S 1-17 | No pools +ve | - | - |
| stdJ525N14.4 | mRNA (mouse) | S 1-17 | WU:ACE, DAU:ABCD, HPB:AB, BM:A, Uact:CE, SK:A, FLU:CD | WU:5, DAU:4.5, Uact:1.3, SK:4 | ZNF-kaiso |
| stdJ525N14.5 | mRNA (mouse) | S 1-17 | NK:ABCDE, DAU:ABCDE, BM:ACDE, Uact:D, FB:BE, HL:E, T:E, AL:C | NK:1.2.5, DAU:4.5, BM:1, FB:6, T:5 | ZNF-kaiso |
| stdJ525N14.6 | mRNA (mouse) | S 1-18 | All pools +ve | Stopped due to poor primer design | ZNF-kaiso |
| stdJ525N14.7 | mRNA (mouse) | S 1-18 | WU:CE, WH:DE, DAU:ABD, HPB:B, SK:AE, FLU:CD | DAU:4, HPB:9, FLU:14, SK:4, WH:16, WU:12.15 | ZNF-kaiso |
| stdJ525N14.10 | Protein | S 1-18 | No pools +ve | - | |
| stdJ555N2.1 | Protein, GENSCAN, FGENESH | S 1-18 | No pools +ve | - | dJ555N2.CX.1 |
| stdJ562J12.1 | Protein | S 1-19 | FB:ABC | FB:1 | dJ562J12.CX.1 |
| stdJ655L22.1.1 | mRNA (human) | V 1-11 | WU:ADE, FB:ABCD, FL:ABD, FLU:C, HL:C, ALU:AE, T:ABCDE, SK:ABCDE | - | dJ655L22.CX.1 |
| stdJ755D9.4 | Protein | S 1-18 | No pools +ve | - | - |
| stdJ808P6.1 | Protein | S 1-18 | FB:ADE, FL:ABCE, HL:ACE, SK:A | stopped | HATB0+ |
| stdJ808P6.2 | Protein | S 1-18 | WH:BC, YT:C, NK:A, DAU:DE, Uact:C, FB:AC | stopped | HATB0+ |
| stdJ808P6.3 | Protein | S 1-18 | HPB:CD, DX3:A | stopped | HATB0+ |

| | | | | | |
|---|---|---|---|---|---|
| stdJ876A24.1 | EST | S 1-18 | WU:CE, DAU:ABCE, HPB:BCDE, BM:AE, Uact:ABCDE, FB:ABCD | DAU:2.3, HPB:6, BM:5, Uact:5 | NRF |
| stdJ876A24.2 | Protein, mRNA (human) | S 1-18 | WU:ABCDE, YT:ABCDE, NK:ABCDE, DAU:ABCE, HPB:ABCDE, BM:ABCDE, Uact:ACE, DXS3:ABCDE, FB:ABCD, HL:CE, SK:ABCE, T:ABCE, FLU:ABCDE, AH:ABCDE | BM:1.2.3.5, T:1.3.5, AH:2.4.5, FB:3 | Sep2 |
| stdJ876A24.4 | EST | S 1-18 | YT:ABD, :DAU:BE, HPB:ABCDE, HL:B, SK:BCDE, FLU:C | T:3, HPB:1, SK:8.9, Dau:8, FLU:11 | NRF |
| stdJ878I13.1 | GENSCAN, FGENESH | S 1-18 | No pools +ve | - | - |
| stdJ1139I1.2 | FGENESH, CpG island | S 1-18 | No pools +ve | - | - |
| stdJ1152D16.1 | EST, GENSCAN, FGENESH | S 1-18 | WU:BD, YT:BCDE, DAU:AE, HPB:ABDE, BM:E, DX3:C, FB:E, SK:BCDE, FLU:BE | HPB:4.5, BM:21, FB:21, SK:8.10 | dJ876A24.CX.1 |

Thirty-six of the 58 primer pairs screened gave positive superpools in the libraries tested. Analysis of the twenty-two that failed to give positive superpools showed that eight were designed to regions predicted to be coding by gene prediction programs alone. The remaining fourteen were predicted by a combination of protein matches, EST matches and gene prediction program. Twenty-six of the thirty-six primer pairs were screened against the cDNA library pools for cDNA isolation by SSPCR and as expected all gave positive pools. Ten STSs gave positive superpools but were not subsequently screened against the pools, because six of the ten were to be used for cDNA isolation using vectorette PCR and a the remaining four were stopped because a mRNA was deposited into the sequence databases for the hATB0+ gene making cDNA isolation unnecessary.

cDNA isolation from individual positive pools was carried out using either SSPCR (Huang, see Figure 4.5) or vectorette PCR (adapted from Riley, J., *et al*. (1990), see Figure 4.6) (see also Section 2.22.3 (Figure 2.1) and 2.22.4 (Figure 2.2) for schemas). For each predicted gene, cDNA isolation was carried out on three pools or super pools from different cDNA libraries in order to increase the likelihood of generating a cDNA sequence covering the entire predicted gene. When different sized products were generated in different pools, cDNA products were chosen for sequencing based on length (where possible the largest band was sequenced), but also intensity (the strongest band took precedence over the largest band). All products generated for sequencing were assigned an Sanger Centre cDNA number (sccd) prior to sequencing.

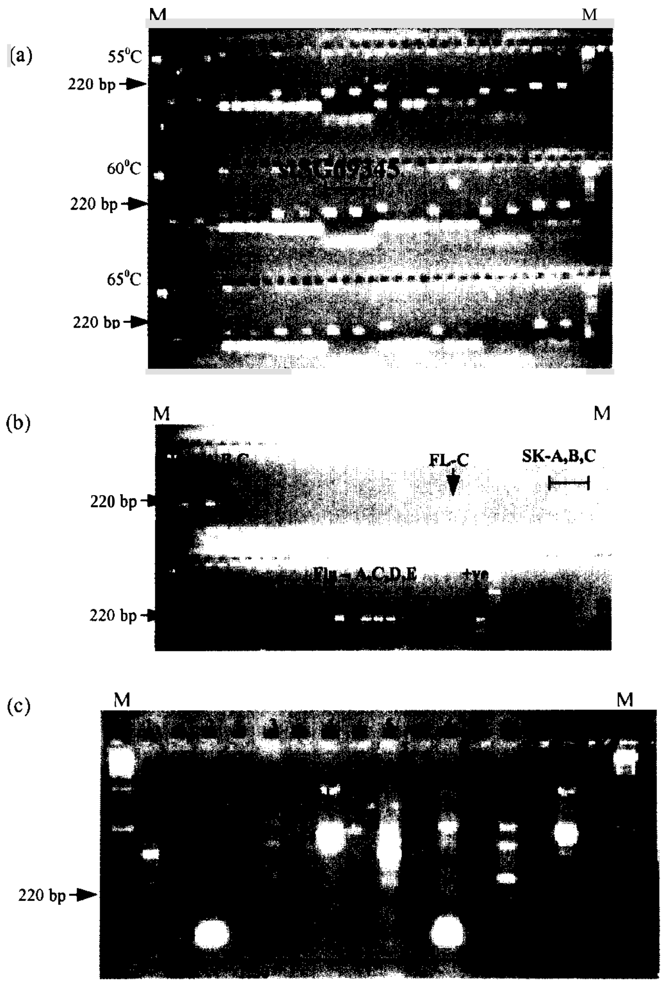


**Figure 4.5:** *cDNA isolation by SSPCR (a) Eight STSs designed to exons ofpredicted genes were testedfor the ability to amplify unique sequences in the human genome, an X chromosome specific hybrid and a hamster cell line, at three different annealing temperatures (M = marker). (b) One of the STSs, stSG69345, was used to amplify DNA ofpools of cDNA clonesfrom 14 different libraries (see Section 2.9). (c) The results of the second round of SSPCRprotocol (see 2.22.1 for schema). A combination of nested sequence-specificprimers (stSG77080S and stSG77080A) and vector-specificprimers (1RP and 2FP) were used to amplify theproductsfrom thefirst round of SSPCR. Two different dilutions of template were used (1:50, lanes 1-4, 1:500, lanes 5-8). Bandsfrom lanes two andfour were excisedfor sequencing.*

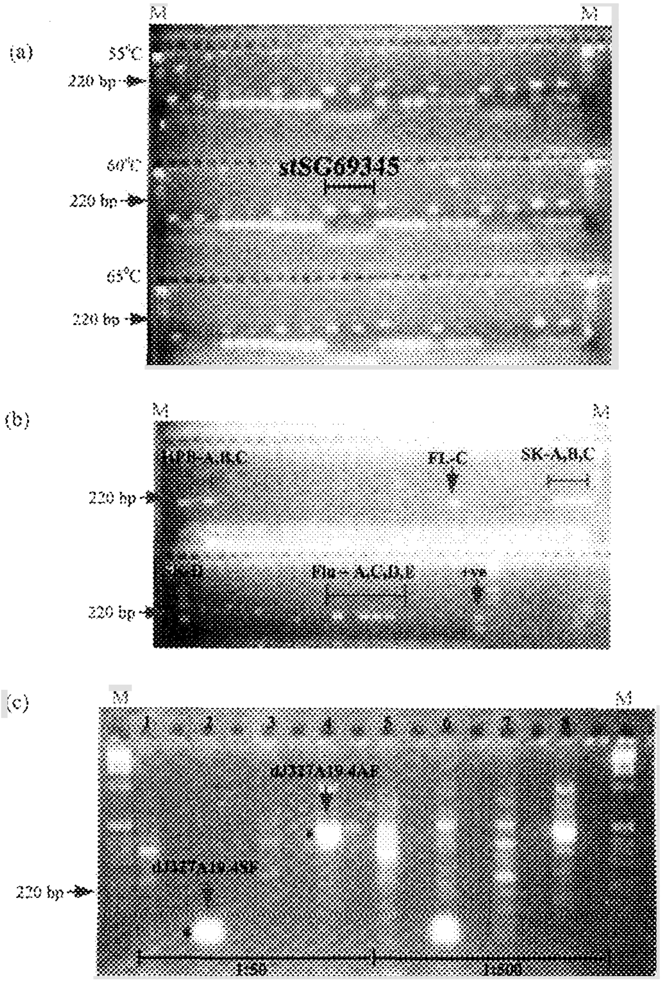


**Figure 4.5:** *cDNA isolation by SSPCR (a) Eight STSs designed to exons of predicted genes were tested for the ability to amplify unique sequences in the human genome, an X chromosome specific hybrid and a hamster cell line, at three different annealing temperatures (M = marker). (b) One of the STSs, stSG69345, was used to amplify DNA of pools of cDNA clones from 14 different libraries (see Section 2.9). (c) The results of the second round of SSPCR protocol (see 2.22.1 for schema). A combination of nested sequence-specific primers (stSG77080S and stSG77080A) and vector-specific primers (1RP and 2FP) were used to amplify the products from the first round of SSPCR. Two different dilutions of template were used (1:50, lanes 1-4, 1:500, lanes 5-8). Bands from lanes two and four were excised for sequencing.*

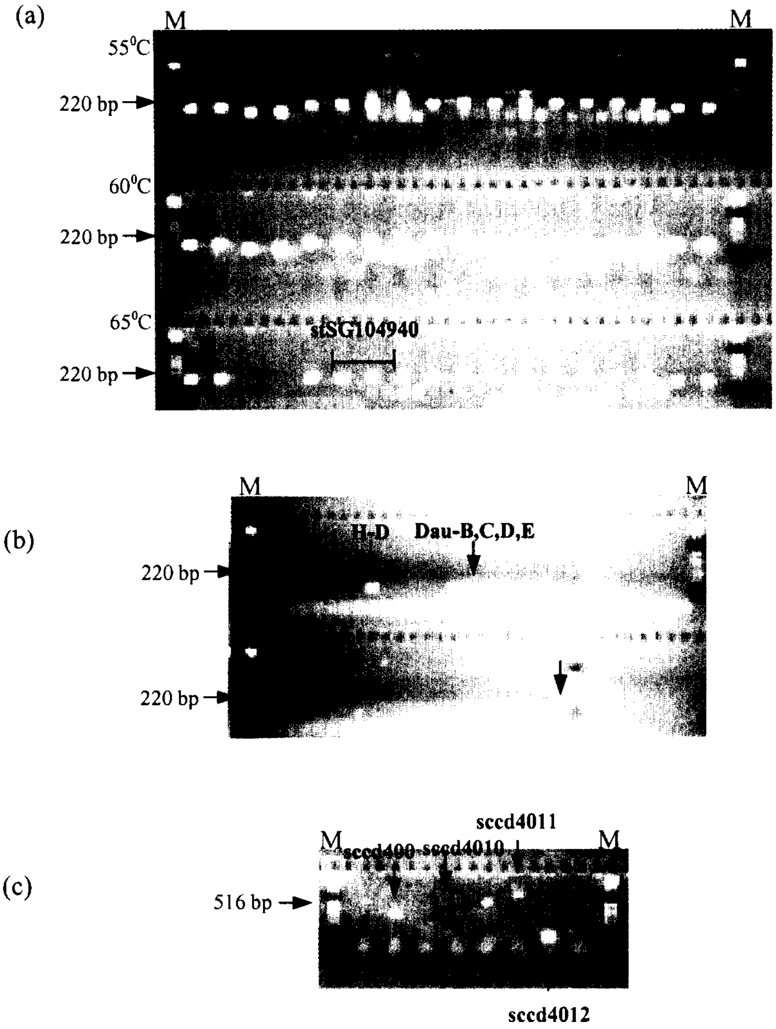


**Figure 4.6:** *cDNA isolation by vectorette PCR (a) Nine STSs designed to exons of predicted genes were testedfor the ability to amplify unique sequences in the human genome, an X chromosome specific hybrid and a hamster cell line, at three different annealing temperatures of the PCR (M = marker). (b) One of the STSs, stSG104940, was used to amplify DNA ofpools of 100,000 cDNA clonesfrom 11 different vectorette libraries. (c) Results of vectorette PCR. A combination of sequence-specific primers (stSG104940S and stSG104940A) and vectorette primer 224 was used to amplify DNA of two superpools (H-D and DauB – see Section 2.9for library informatation) at two different concentrations (1:100 and 1:1000). Bandsfrom lanes 2, 5, 6 and 7 were excised,purified, and sequenced.*

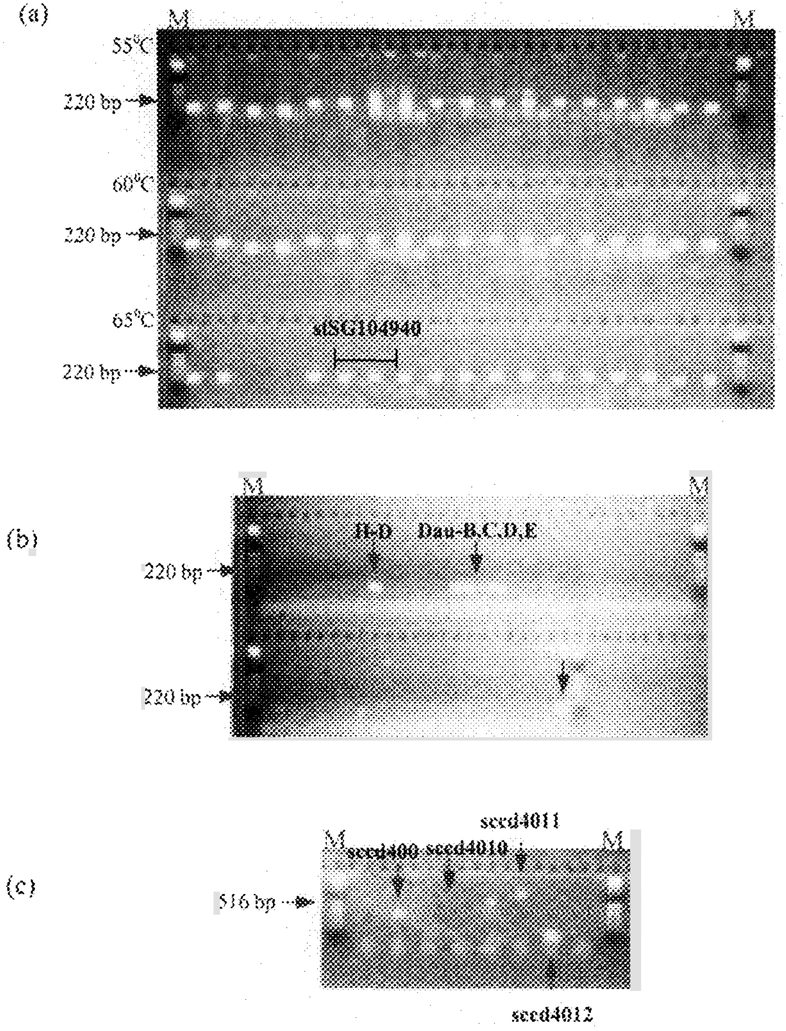
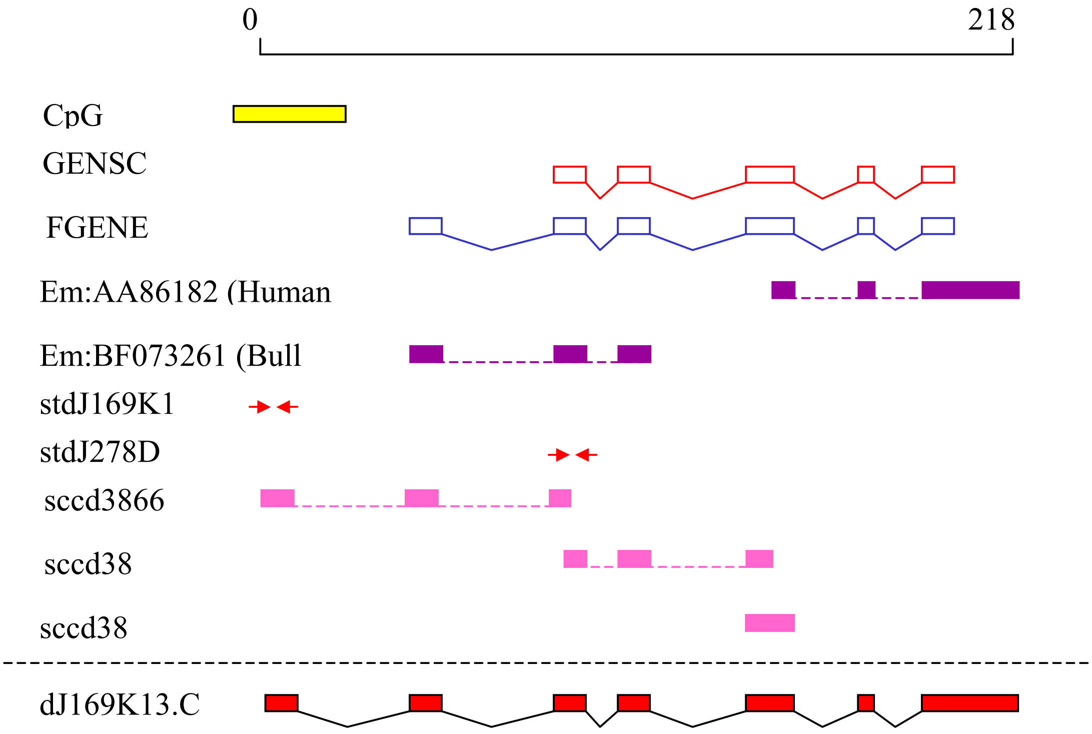


**Figure 4.6:** *cDNA isolation by vectorette PCR (a) Nine STSs designed to exons of predicted genes were tested for the ability to amplify unique sequences in the human genome, an X chromosome specific hybrid and a hamster cell line, at three different annealing temperatures of the PCR (M = marker). (b) One of the STSs, stSG104940, was used to amplify DNA of pools of 100,000 cDNA clones from 11 different vectorette libraries. (c) Results of vectorette PCR. A combination of sequence-specific primers (stSG104940S and stSG104940A) and vectorette primer 224 was used to amplify DNA of two superpools (H-D and DauB – see Section 2.9 for library informatation) at two different concentrations (1:100 and 1:1000). Bands from lanes 2, 5, 6 and 7 were excised, purified, and sequenced.*

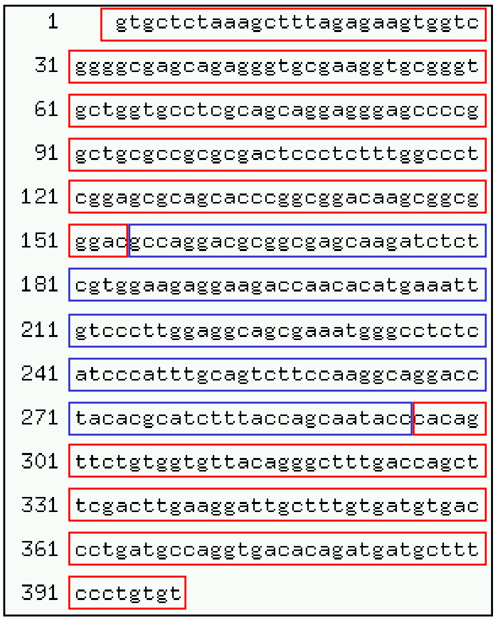
cDNA sequence was aligned to the genomic sequence and the gene structure evaluated for possible extension. Confirmation of a predicted gene was considered complete when there was human cDNA sequence covering at least the predicted protein-coding region, and as much untranslated region (UTR) as possible. A total of fourteen predicted genes were confirmed and eleven gene structures remain unconfirmed. An example of the confirmation of one gene is shown in Figure 4.7.

**Figure 4.7:** *(see over) Confirmation of a novel gene. (a) dJ169K13.CX.1 (exons shown as red boxes, introns as black lines) was predicted by GENSCAN (exons shown as open red boxes linked by red lines) and FGENESH (exons shown as open blue boxes linked by blue lines), and two ESTs (shown in purple), one human and one from bull. A CpG island upstream of the predicted genes suggested a possible location for the 5' end of the gene. Three cDNA sequences (shown as pink boxes) were generated to confirm the 5' end of this gene. (b) The cDNA sequence for sccd3866. Sequence corresponding to exons as they appear in the genomic sequence are coloured as alternating red and blue open boxes.*

(a



(b



```
  1      gtgctctaaagctttagagaagtggtc
 31 ggggcgagcagagggtgcgaaggtgcgggt
 61 gctggtgcctcgcagcaggagggagccccg
 91 gctgcgccgcgcgactccctctttggccct
121 cggagcgcagcacccggcggacaagcggcg
151 ggacgccaggacgcggcgagcaagatctct
181 cgtggaagaggaagaccaacacatgaaatt
211 gtcccttggaggcagcgaaatgggcctctc
241 atcccatttgcagtcttccaaggcaggacc
271 tacacgcatctttaccagcaataccccacag
301 ttctgtggtgttacagggctttgaccagct
331 tcgacttgaaggattgctttgtgatgtgac
361 cctgatgccaggtgacacagatgatgcttt
391 ccctgtgt
```

Twenty pseudogenes were identified within the region and are predicted to have arisen due to the reverse transcription of mRNAs into the genomic sequence. They all appear to have a functional counterpart elsewhere in the human genome. The pseudogenes were identified because they have no introns, a poly A tail within the genomic sequence and a disrupted ORF (see Figure 4.8) (see appendix, Table 4.6 for a full list of the pseudogenes identified in the region).

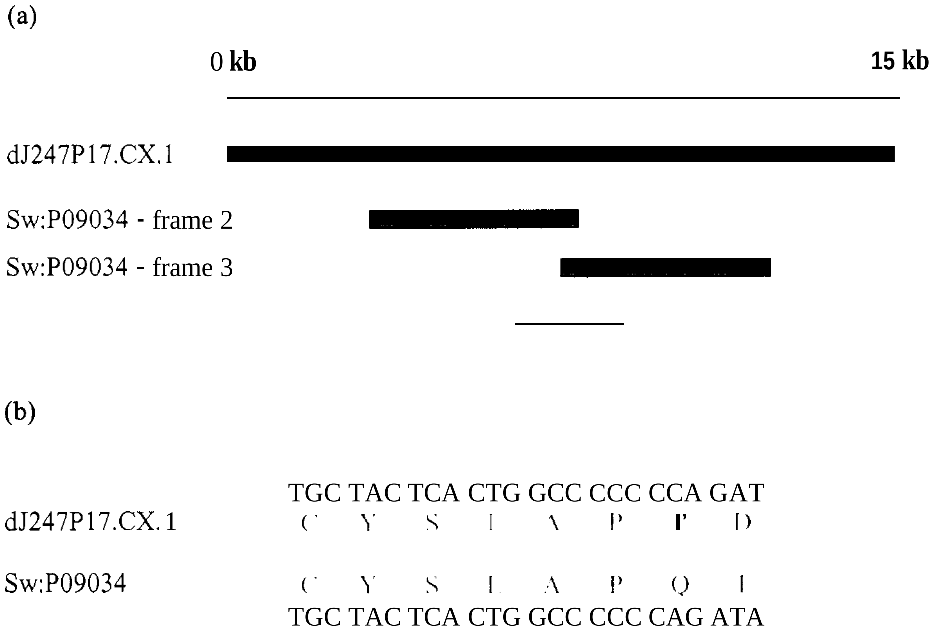


**Figure 4.8:** *Example of a pseudogene. (a) The extent* **of** *the dJ241P17.CX.1 (shown as a green box) is shown and is a pseudogene* **of** *arginosuccinate synthetase* **(ASS).** *Part of the protein sequence* **of ASS** *(Sw:P09034) aligns in two blocks (shown as blue boxes) (b) Disruption* **of** *the reading frame due to an insertion of a C within a run of seven C's (shown as red letters). The alignment of the nucleotide sequence (black letters) and the amino acid sequence (blue letters, using the one letter code) surrounding the insertion are shown.*

In summary, a gene map encompassing the distal portion of Xq23, Xq24 and the proximal portion of Xq25 between DXS7598 and DXS7333 covering 8 Mb has been constructed (see Figure 4.9). The region contains 33 confirmed genes (of which 14 were confirmed during this study), 11 predicted genes and 20 pseudogenes.

**Figure 4.9:** *(see over) A summary of the gene map between DXS7598 and DXS7333. The red bars indicate the contigs status and the black bars indicate the extent of finished sequence. Each link represents a series of individual clones (see appendix to this chapter). Yellow bars indicate clones for which draft sequence is available, and white bars indicate clones selected for sequencing, but not sequenced as of September 2001. A scale is given in megabase pairs (Mb). Approved names are given for known genes (see Table 4.1). Genes are indicated by arrows (black – complete, blue – predicted, green – pseudogene), the direction of each arrow reflects the direction of transcription. Genes on the plus strand are positioned above the dotted line, genes on the minus strand are positioned below the dotted line.*

# Chapter 4

# Genome Landscape of Xq23-24

## 4.1 Introduction

The generation of large contiguous segments of human genome sequence allows for the systemic identification of the genes and other functional units contained within. As discussed in Section 1.3, one of the aims of studying the human genome is to identify all the genes present in the human genome which will provide the basis for furthering our understanding of the biological systems in which they are involved. There are also a large number of diseases for which the gene responsible remains uncloned. Identifying the genes within a region of interest because of the association with a particular disease enables screening for the causative mutations. For instance, a systematic search for the gene responsible for X-linked lymphoproliferative disease (XLP), previously localised to Xq25, resulted in the identification of SH2D1A, which was subsequently found to be mutated in patients with XLP (Coffey, A. J.*, et al.*, 1998).

As part of the project to map and sequence the human X chromosome, the techniques developed during the construction of the sequence-ready bacterial clone contig in Xq22 (see previous chapter) were applied to extend this contig, and generate new contigs across all regions of the human X chromosome (Bentley, D. R.*, et al.*, 2001). The sequencing of the X chromosome, based on these contigs, progressed rapidly in the Xq23-Xq25 region and was therefore chosen for an in depth study of the features encoded in the sequence.

## RESULTS

### 4.2 Identification of genes

As a result of the work carried out for the X chromosome mapping and sequencing project by the X chromosome group, four contigs covering 25 Mb of Xq22-Xq25 were generated and a minimum set of clones identified and sequenced (see Figure 4.1). This provided the raw data for the analysis and gene identification across an 8 Mb region between DXS7598 and DXS7333 that encompasses the distal portion of Xq23, Xq24, and the proximal portion of Xq25. There are currently twelve contiguous segments of finished sequence covering 7 Mb and a further 600 kb of draft sequence is available.

A combination of similarity searches (BLASTX and BLASTN) and *de novo* gene prediction were carried out on finished sequence (see Figure 4.2). Prototypical and consensus repeats were masked using RepeatMasker (Smit, AFA & Green, P. RepeatMasker at http://ftp.genome.washington.edu/RM/RepeatMasker.html), and the remainder of the sequence was aligned using BLAST (Altschul, S. F*., et al.*, 1990) to all known cDNAs, ESTs and other sequences to identify similarities with previously known genes from human and other species. A modification of this approach was to translate the genomic sequence into all possible reading frames and compare the sequence of the translation products using BLAST with databases of known protein sequences.

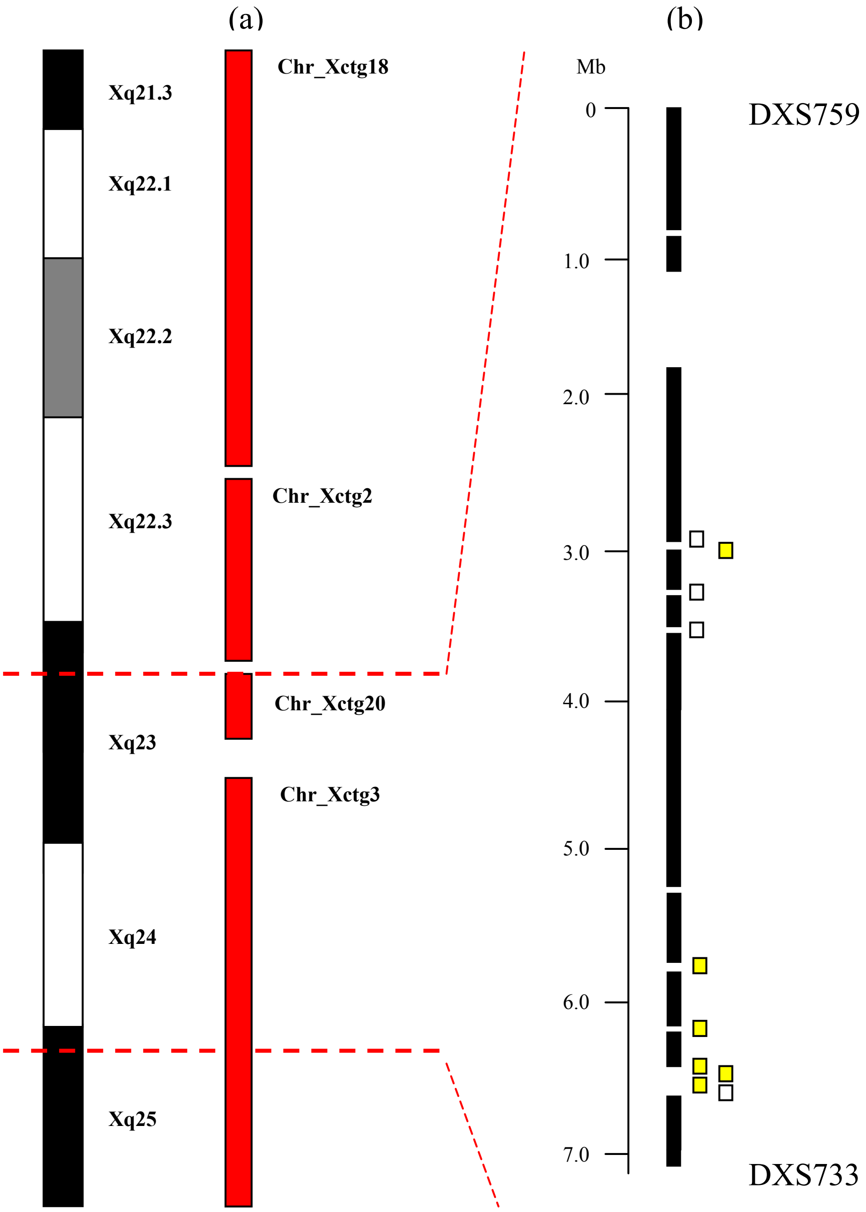


**Figure 4.1:** *Status of the region between Xq21.3 and Xq25 (a) Extent of bacterial clone mapping showing part of the X chromosome between Xq21.3 and Xq25, and the status of mapping and sequencing. Vertical red bars indicate contigs and their number from the Chromosome X mapping project are shown. The dotted red lines indicate the region identified for gene identification between DXS7598 and DXS7333. (b) The status of genomic sequence: black bars are continuous segments of finished sequence, yellow bars indicate clones with draft sequence available, white bars signify clones identified for sequencing, for which sequence is not yet available*

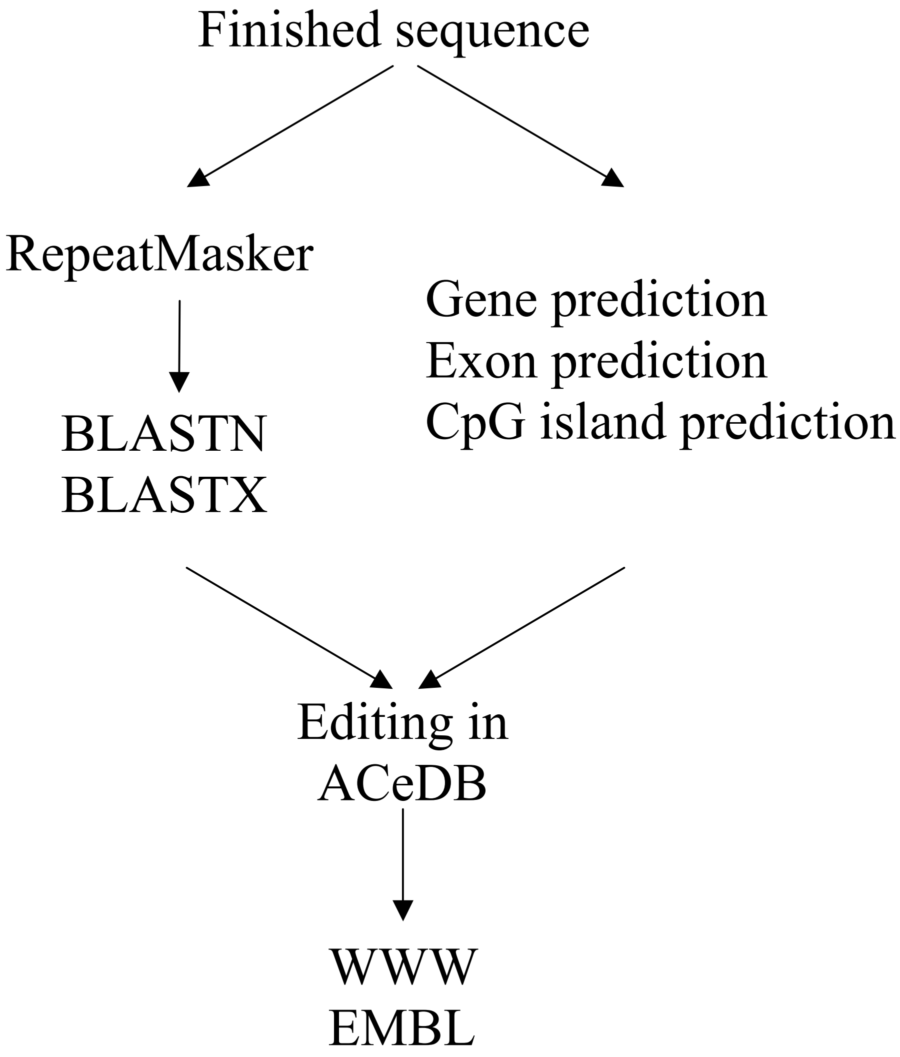


**Figure 4.2:**

*Genomic sequence analysis. Finished sequence is analysed for repeats using RepeatMasker and compared to known protein and DNA sequences. De novo gene prediction is carried out on non-Repeatmasked sequence. The annotated sequence is viewed in an X chromosome ACeDB, Xace (see Section 2.23.3) and made available on the WWW (Sanger FTP site and EMBL).*

*De novo* gene prediction was carried out using a variety of prediction programmes to identify putative exons and genes. Also, given that CpG islands are associated with approximately 56% of genes (Hannenhalli, S*., et al.*, 2001), a CpG island finder was used (courtesy of Gos Micklen). Members of the sequence annotation group at the Sanger Centre carried out the initial annotation of the genome sequence and the results were visualised in an X chromosome specific implementation of ACeDB, Xace (see Figure 4.3).

Initial analysis of the genomic sequence identified 19 genes that were previously known (see Table 4.1). In all cases the known genes were identified because there was a full length mRNA sequence aligning exactly to the genomic sequence.

**Table 4.1:** *Known Genes with full length mRNA sequence*

| Gene Name | Accession number | Reference |
|---|---|---|
| HOM-TES-85 | AF124430 | direct submission |
| hATB0+ | AF151978 | Sloan, J. L*., et al.*, 1999 |
| ANT2 | J02683 | Battini, R*., et al.*, 1987 |
| NDUFA1 | U54993 | Au, H. C*., et al.*, 1999 |
| LAMP2 | J04183 | Kannan, K*., et al.*, 1996 |
| GLUD2 | X66310 | Shashidharan, P*., et al.*, 1994 |
| GRIA3 | X82068 | direct submission |
| T-plastin | M22299 | Lin, C. S*., et al.*, 1993 |
| NRF | AJ011812.2 | Nourbakhsh, M*., et al.*, 2000 |
| IL13R | X95302 | Caput, D*., et al.*, 1996 |
| ZNF-kaiso | XM_010435 | direct submission |
| HPR6.6 | Y12711 | Gerdes, D*., et al.*, 1998 |
| ZNF183 | X98253 | Frattini, A*., et al.*, 1997 |
| UBE2A | M74524 | Koken, M. H*., et al.*, 1996 |
| ATP1B4 | AF158383 | Pestov, N. B*., et al.*, 1999 |
| SMT3B | X99585 | Lapenta, V*., et al.*, 1997 |
| SEP2 | D50918 | direct submission |
| RPL39 | U57846 | Delbruck, S*., et al.*, 1997 |
| U69a | Y11163 | direct submission |

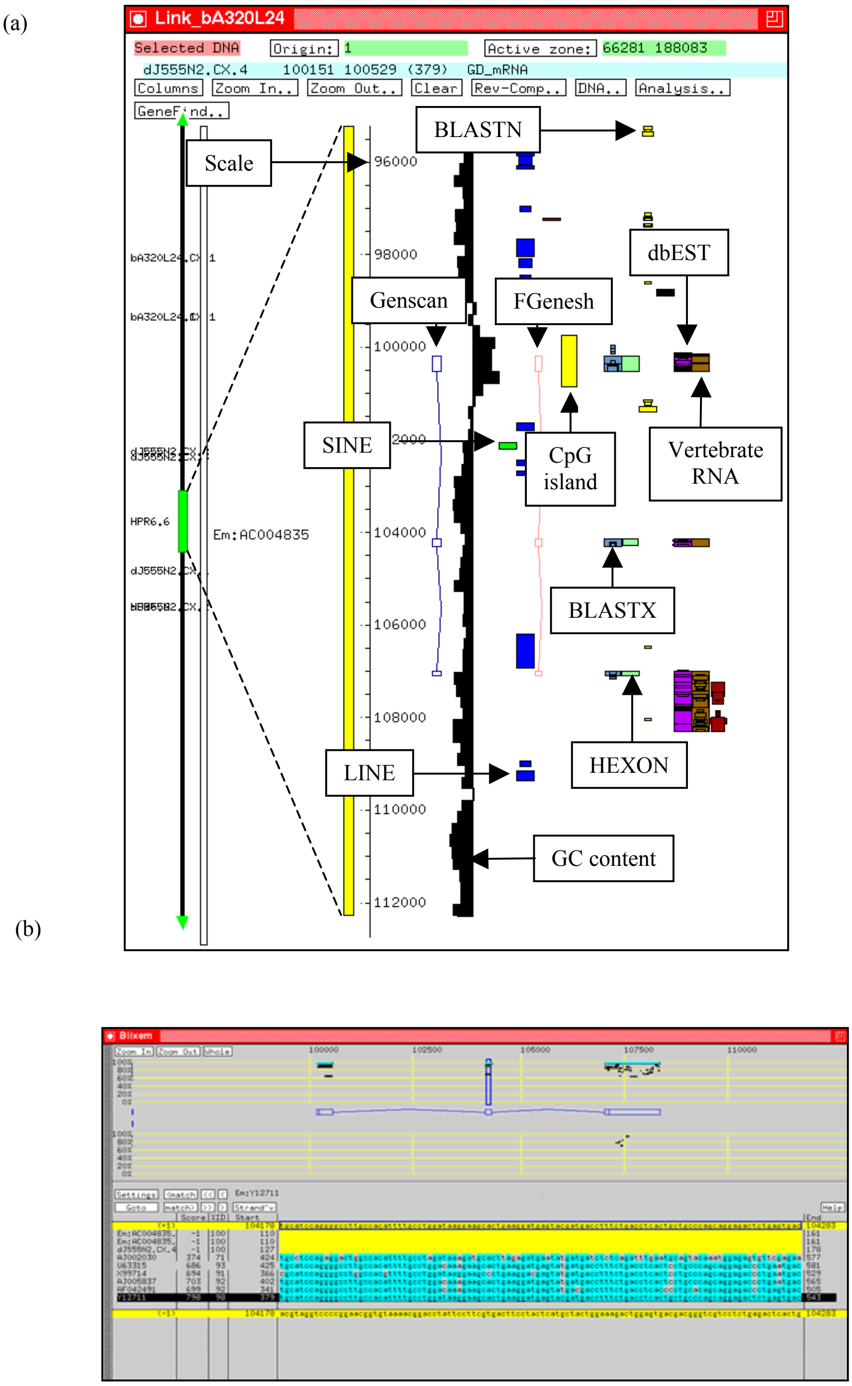
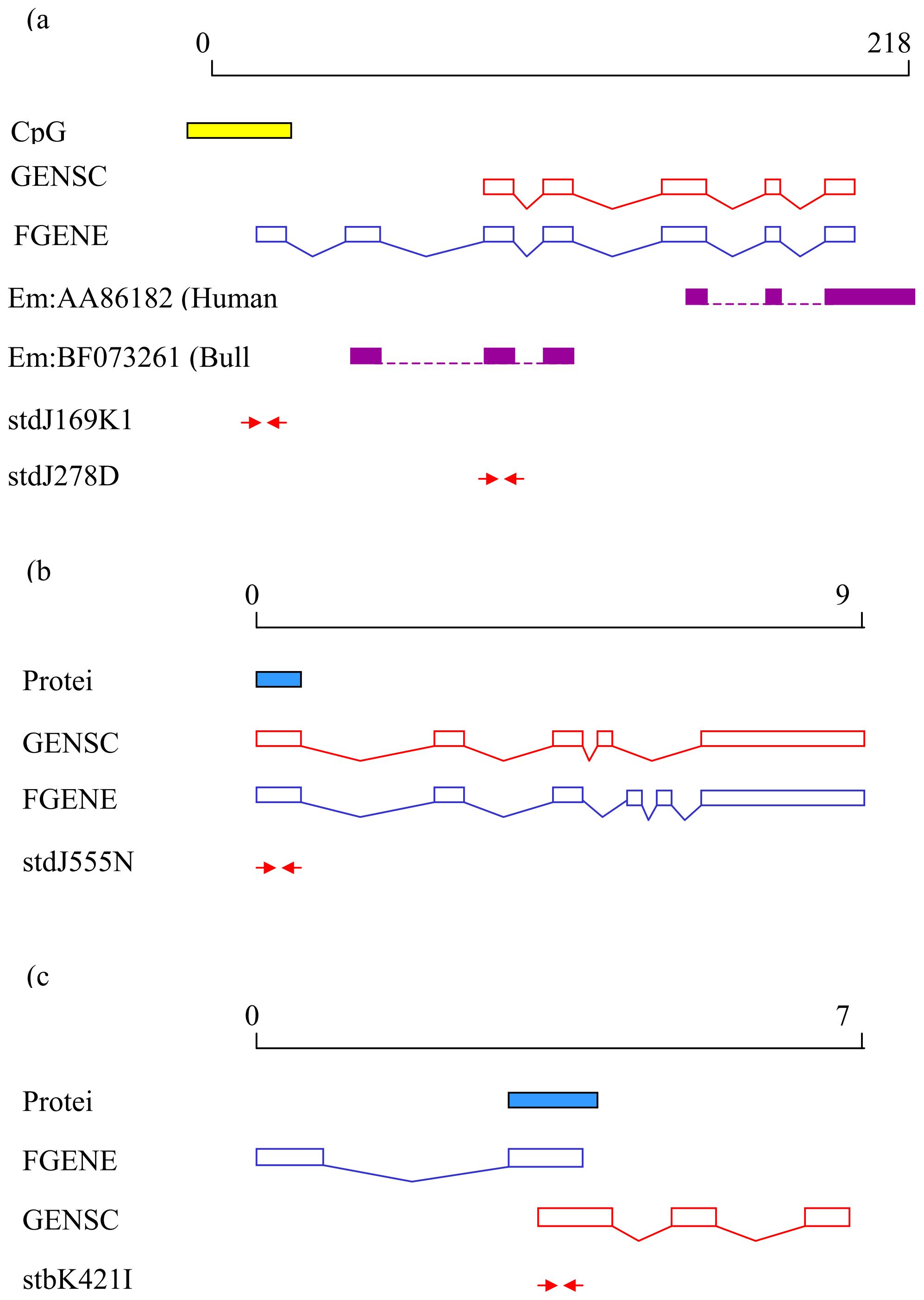


**Figure 4.3:** *ACeDB and BLIXEM (a) Example of annotated sequence in Xace. Features such as Genscan and Fgenesh predictions are labelled. The width of the each bar is an indication of the similarity between the genomic sequence and feature. (b) View of BLIXEM showing alignment of vertebrate mRNA sequences to the genomic sequence (see Section 2.23.4).*

Fourteen of the nineteen known genes have associated publications and the mRNAs of the remaining five genes were deposited directly into the sequence databases. The gene names are given as recommended by the Human Gene Nomenclature Committee (HGNC – http://www.gene.ucl.ac.uk/nomenclature). Although these genes needed no experimental verification, the precise exon/intron structure for each gene has been elucidated and their position and transcriptional direction on the genomic sequence in relation to neighbouring genes determined by alignment to the genomic sequence as part of this study.

In order to identify novel genes in the region, the genomic sequence was analysed for regions predicted to represent exons, based on sequence similarity searches and gene prediction programmes. Primers for the PCR were designed to regions with a variety of evidence suggesting the presence of a gene. For instance, eight pairs of primers were designed to regions predicted to be coding only by gene prediction programs, three pairs of primers were designed to regions predicted only by protein homology and two pairs of primers were designed to regions predicted only by EST homology. In some cases a protein or DNA sequence spliced across a series of exons in the genomic sequence and a predicted gene structure was identified and in other cases only a single exon was suggested. Examples of both predicted gene structures and a single exon region are shown in Figure 4.4.

**Figure 4.4:** *(see over) Examples of features for which STSs were designed for cDNA isolation. (a) A region of 218 kb was predicted to be coding by both GENSCAN (red boxes and lines) and FGENESH (blue boxes and lines), the 5' end suggested by the presence of a CpG island (yellow box). Two ESTs (purple boxes, splicing indicated by dotted lines) matched the genomic sequence exactly. Two STSs (indicated by red arrows) were designed, to generate novel cDNA sequence in regions not covered by the human cDNA sequence. (b) A region of 9 kb was predicted to be coding both by GENSCAN and FGENESH. A protein match (light blue box) was also observed in the first exon. (c) An example of a single exon feature where one exon from a GENSCAN prediction overlapped with one exon from a FGENESH prediction. A protein match was also observed.*


(a
0
218
CpG
GENSC
FGENE
Em:AA86182 (Human
Em:BF073261 (Bull
stdJ169K1
stdJ278D
(b
0
9
Protei
GENSC
FGENE
stdJ555N
(c
0
7
Protei
FGENE
GENSC
stbK421I

Analysis of the genomic sequence identified twenty-two predicted gene structures and twenty single exon regions. During the experimental verification process, mRNA sequences for five genes (T-plastin, ZNF-kaiso, UPF3B, NRF and HATB0+) were published and/or submitted to sequence databases by external groups and these previously predicted genes became “known genes” and are listed in Table 4.1. In order to analyse the 42 predicted gene structures and single exon predictions, a total of 58 primer pairs were designed either automatically using PRIMER (http://www.sanger.ac.uk/cgi-bin/primer3.cgi) or manually (see Section 2.15.1). Where possible each primer was between 18 and 20 nucleotides in length and had a GC content of approximately 50%. Primers were designed within a single predicted exon and pre-screened to determine the optimal annealing temperature for the PCR. PCR was carried out on pools of up to 19 different cDNA libraries (see Section 2.8.3) to identify the individual cDNA pool likely to contain the cDNA of interest. Each cDNA library in the panel comprises approximately 500,000 clones divided into twenty-five pools, each containing 25,000 cDNA clones. Five pools were combined to form a super pool containing 100,000 cDNA clones (cDNA library resources were kindly provided by Jackie Bye). For cDNA isolation using SSPCR, primers were initially screened against the super pools and then against pools representing up to five different positive super pools. For cDNA isolation using vectorette PCR, only the super pools were screened.  The results are summarised in Table 4. 2.

**Table 4.2:** *Experimental verification of predicted genes (see 2.8.3 for library pool codes); STSs are described in Table 2.6. Those superpools for which the equivalent pools were screened are shown in red. A gene name is given (last column) if the STS that was designed was used to generate novel cDNA sequence for confirmation of a predicted gene structure.*

| **STS name** | **Evidence** | **Library Pools** | **Positive Superpool** | **Positive pool** | **Gene** |
|---|---|---|---|---|---|
| stbA45J1.1.1 | EST, GENSCAN, FGENESH | V 1-11 | FL:B | - | bA45J1.CX.1 |
| stbA125M24.1.1 | Protein, mRNA (human) | V 1-11 | No pools +ve | - | bB125M24.1 |
| stbK421I3.1 | Protein, GENSCAN, FGENESH | S 1-17 | No pools +ve | - | - |
| stbK421I3.2 | mRNA, GENSCAN, FGENESH | S 1-17 | Uact:C | Uact:14 | bK421I3.CX.2 |
| stbK421I3.3 | EST, Protein, FGENESH | S 1-18 | WU:E, T:ABCDE | WU:22.23, T:1.2.4.6.7.9.10 | bK421I3.CX.1 |
| stdA155F9.1 | Protein | V 1-11 | WH:D, DAU:BCDE | - | dA155F9.CX.1 |
| stdA155F9.2 | Protein | V 1-11 | ALU:AE | - | dA155F9.CX.1 |
| stdJ29I24.1 | GENSCAN, FGENESH | S 1-18 | No pools +ve | - | - |
| stdJ57A13.1 | GENSCAN, FGENESH | S 1-17 | No pools +ve | - | - |
| stdJ57A13.2 | EST, GENSCAN | S 1-17 | WH:CE, NK:CE, DAU:B | WH:14, 24 | genomic contamination |
| stdJ93I3.1 | GENSCAN, FGENESH | S 1-18 | No pools +ve | - | - |
| stdJ93I3.2 | EST | S 1-18 | No pools +ve | - | Plastin 3 |
| stdJ169K13.1 | FGENESH, CpG island | S 1-19 | HPB:C | HPB:13.15 | dJ169K13.CX.1 |
| stdJ169K13.2 | Protein, EST | S 1-17 | No pools +ve | - | - |
| stdJ169K13.3 | Protein, EST | S 1-17 | No pools +ve | - | - |

| | | | | | |
|---|---|---|---|---|---|
| stdJ170D19.1 | GENSCAN, FGENESH | S 1-18 | T:E | T:21 | no cDNA product |
| stdJ170D19.2 | EST – no splice | S 1-18 | No pools +ve | - | - |
| stdJ222H5.1 | EST – no splice | S 1-17 | No pools +ve | - | - |
| stdJ222H5.2 | Protein | S 1-17 | No pools +ve | - | - |
| stdJ278D1.1 | Protein, EST, GENSCAN, FGENESH | S 1-17 | HPB:ABC, SK:AB, T:E | HPB:5, SK:5, T:21.22 | dJ169K13.CX.1 |

| | | | | | |
|---|---|---|---|---|---|
| stdJ278D1.1 | Protein, EST, GENSCAN, FGENESH | S 1-17 | HPB:ABC, SK:AB, T:E | HPB:5, SK:5, T:21.22 | dJ169K13.CX.1 |
| stdJ318C15.1 | Protein, EST, GENSCAN, FGENESH | S 1-17 | DAU:BCDE, HPB:BC, Uact:AD, FB:B | DAU:7.8.9, HPB:8, Uact:3.18, FB:8 | dJ318C15.CX.1 |
| stdJ321E8.1 | GENSCAN, FGENESH | S 1-19 | No pools +ve | - | - |
| stdJ321E8.2.1 | EST, GENSCAN, FGENESH | V 1-11 | T:BDE | - | dJ321E8.CX.2<br>dJ321E8.CX.3 |
| stdJ321E8.3.1 | EST, GENSCAN, FGENESH | V 1-11 | T:E | - | dJ321E8.CX.2<br>dJ321E8.CX.3 |
| stdJ327A19.1 | EST, FGENESH | S 1-17 | YT:ABCDE, HPB:AC, FB:ABCE, FL:BC, HL:ABCDE, SK:ABCDE, FLU:BCDE, DX3:ABCDE | YT:1.3.4.5, FB:7, SK:1.4.5, FLU:9, DXS:1.2.3.5 | UPF3B |
| stdJ327A19.2 | mRNA (mouse), EST, GENSCAN | S 1-17 | No pools +ve | - | dJ327A19.CX.4 |
| stdJ327A19.3 | BLASTX, GENSCAN, FGENESH | S 1-17 | YT:CD, HPB:B, FB:D, FL:C, HL:A, SK:ABC, T:E, AL:A, FLU:A | HPB:6.7, SK:2.3.4.5, FLU:5 | dJ327A19.CX.3 |
| stdJ327A19.4 | BLASTX, GENSCAN, FGENESH | S 1-17 | YT:ABCDE, HPB:BCE, FB:D, FL:C, HL:AE, SK:ABCD, T:CE, AL:AE, FLU:AD | YT:2.4, HPB:6.7, HL:5, SK1.5, FLU:5 | dJ327A19.CX.3 |
| stdJ327A19.5 | BLASTX, GENSCAN, FGENESH | S 1-17 | WP:A | WP:3 | dJ327A19.CX.3 |
| stdJ327A19.6 | BLASTX, GENSCAN, FGENESH | S 1-17 | WU:CE, YT:CD, DAU:D, HPB:ABCDE, FL:C, SK:ABCD, FLU:ACDE | WU:11.12, YT:13, \|HPB:7.8.10, SK:6.7.8.9.10, FLU:2.5 | dJ327A19.CX.3 |
| stdJ378P9.1 | EST – no splice | S 1-18 | DX3:ABDE, FB:ACE, FL:D, HL:E, FLU:BCDE | FB:4, FLU:4.6 | genomic contamination |
| stdJ394H4.1 | GENSCAN | S 1-18 | No pools +ve | - | - |
| stdJ404F18.1 | EST, FGENESH | S 1-18 | WU:AB, NK:ABDE, HPB:BE, | NK:3, HPB:10, BM:7 | dJ1139I1.CX.1 |

| | | | | | |
|---|---|---|---|---|---|
| | | | BM:B, HL:A, FLU:A, AL:E | | |
| stdJ404F18.2 | EST, GENSCAN, FGENESH | S 1-17 | WU:BC, YT:CD, NK:C, DAU:E, HPB:ABDE, BM:ACE, FB:CE, SK:BDE, FLU:E, DX3:C | YT:14, NK:15, SK:6.8 | dJ876A24.CX.1 |
| stdJ404F18.3 | EST, GENSCAN, FGENESH | S 1-18 | WU:BCE, YT:BCDE, HPB:ABDE, Uact:ABDE, DX3:AC, FB:CDE, HL:B, SK:BCDE, FLU:BE, ALU:B, AH:CE | ALU:6.8.10 | dJ876A24.CX.1 |
| stdJ452H17.1 | mRNA (mouse), EST | S 1-18 | No pools +ve | - | - |

| | | | | | |
|---|---|---|---|---|---|
| stdJ452H17.1.1 | mRNA (mouse), EST | S 1-19 | T:CD | T:13 | no cDNA product |
| stdJ525N14.1 | Protein, GENSCAN, FGENESH | S 1-17 | WU:E, T:ABCDE | WU:22, T:21.25 | dJ525N14.CX.1 |
| stdJ525N14.2 | Protein, EST | S 1-17 | NK:C | NK:15 | genomic contamination |
| stdJ525N14.3 | FGENESH | S 1-17 | No pools +ve | - | - |
| stdJ525N14.4 | mRNA (mouse) | S 1-17 | WU:ACE, DAU:ABCD, HPB:AB, BM:A, Uact:CE, SK:A, FLU:CD | WU:5, DAU:4.5, Uact:1.3, SK:4 | ZNF-kaiso |
| stdJ525N14.5 | mRNA (mouse) | S 1-17 | NK:ABCDE, DAU:ABCDE, BM:ACDE, Uact:D, FB:BE, HL:E, T:E, AL:C | NK:1.2.5, DAU:4.5, BM:1, FB:6, T:5 | ZNF-kaiso |
| stdJ525N14.6 | mRNA (mouse) | S 1-18 | All pools +ve | Stopped due to poor primer design | ZNF-kaiso |
| stdJ525N14.7 | mRNA (mouse) | S 1-18 | WU:CE, WH:DE, DAU:ABD, HPB:B, SK:AE, FLU:CD | DAU:4, HPB:9, FLU:14, SK:4, WH:16, WU:12.15 | ZNF-kaiso |
| stdJ525N14.10 | Protein | S 1-18 | No pools +ve | - | |
| stdJ555N2.1 | Protein, GENSCAN, FGENESH | S 1-18 | No pools +ve | - | dJ555N2.CX.1 |
| stdJ562J12.1 | Protein | S 1-19 | FB:ABC | FB:1 | dJ562J12.CX.1 |
| stdJ655L22.1.1 | mRNA (human) | V 1-11 | WU:ADE, FB:ABCD, FL:ABD, FLU:C, HL:C, ALU:AE, T:ABCDE, SK:ABCDE | - | dJ655L22.CX.1 |
| stdJ755D9.4 | Protein | S 1-18 | No pools +ve | - | - |
| stdJ808P6.1 | Protein | S 1-18 | FB:ADE, FL:ABCE, HL:ACE, SK:A | stopped | HATB0+ |
| stdJ808P6.2 | Protein | S 1-18 | WH:BC, YT:C, NK:A, DAU:DE, Uact:C, FB:AC | stopped | HATB0+ |
| stdJ808P6.3 | Protein | S 1-18 | HPB:CD, DX3:A | stopped | HATB0+ |

| | | | | | |
|---|---|---|---|---|---|
| stdJ876A24.1 | EST | S 1-18 | WU:CE, DAU:ABCE, HPB:BCDE, BM:AE, Uact:ABCDE, FB:ABCD | DAU:2.3, HPB:6, BM:5, Uact:5 | NRF |
| stdJ876A24.2 | Protein, mRNA (human) | S 1-18 | WU:ABCDE, YT:ABCDE, NK:ABCDE, DAU:ABCE, HPB:ABCDE, BM:ABCDE, Uact:ACE, DXS3:ABCDE, FB:ABCD, HL:CE, SK:ABCE, T:ABCE, FLU:ABCDE, AH:ABCDE | BM:1.2.3.5, T:1.3.5, AH:2.4.5, FB:3 | Sep2 |
| stdJ876A24.4 | EST | S 1-18 | YT:ABD, :DAU:BE, HPB:ABCDE, HL:B, SK:BCDE, FLU:C | T:3, HPB:1, SK:8.9, Dau:8, FLU:11 | NRF |
| stdJ878I13.1 | GENSCAN, FGENESH | S 1-18 | No pools +ve | - | - |
| stdJ1139I1.2 | FGENESH, CpG island | S 1-18 | No pools +ve | - | - |
| stdJ1152D16.1 | EST, GENSCAN, FGENESH | S 1-18 | WU:BD, YT:BCDE, DAU:AE, HPB:ABDE, BM:E, DX3:C, FB:E, SK:BCDE, FLU:BE | HPB:4.5, BM:21, FB:21, SK:8.10 | dJ876A24.CX.1 |

Thirty-six of the 58 primer pairs screened gave positive superpools in the libraries tested. Analysis of the twenty-two that failed to give positive superpools showed that eight were designed to regions predicted to be coding by gene prediction programs alone. The remaining fourteen were predicted by a combination of protein matches, EST matches and gene prediction program. Twenty-six of the thirty-six primer pairs were screened against the cDNA library pools for cDNA isolation by SSPCR and as expected all gave positive pools. Ten STSs gave positive superpools but were not subsequently screened against the pools, because six of the ten were to be used for cDNA isolation using vectorette PCR and a the remaining four were stopped because a mRNA was deposited into the sequence databases for the hATB0+ gene making cDNA isolation unnecessary.

cDNA isolation from individual positive pools was carried out using either SSPCR (Huang, see Figure 4.5) or vectorette PCR (adapted from Riley, J., *et al.* (1990), see Figure 4.6) (see also Section 2.22.3 (Figure 2.1) and 2.22.4 (Figure 2.2) for schemas). For each predicted gene, cDNA isolation was carried out on three pools or super pools from different cDNA libraries in order to increase the likelihood of generating a cDNA sequence covering the entire predicted gene. When different sized products were generated in different pools, cDNA products were chosen for sequencing based on length (where possible the largest band was sequenced), but also intensity (the strongest band took precedence over the largest band). All products generated for sequencing were assigned an Sanger Centre cDNA number (sccd) prior to sequencing.

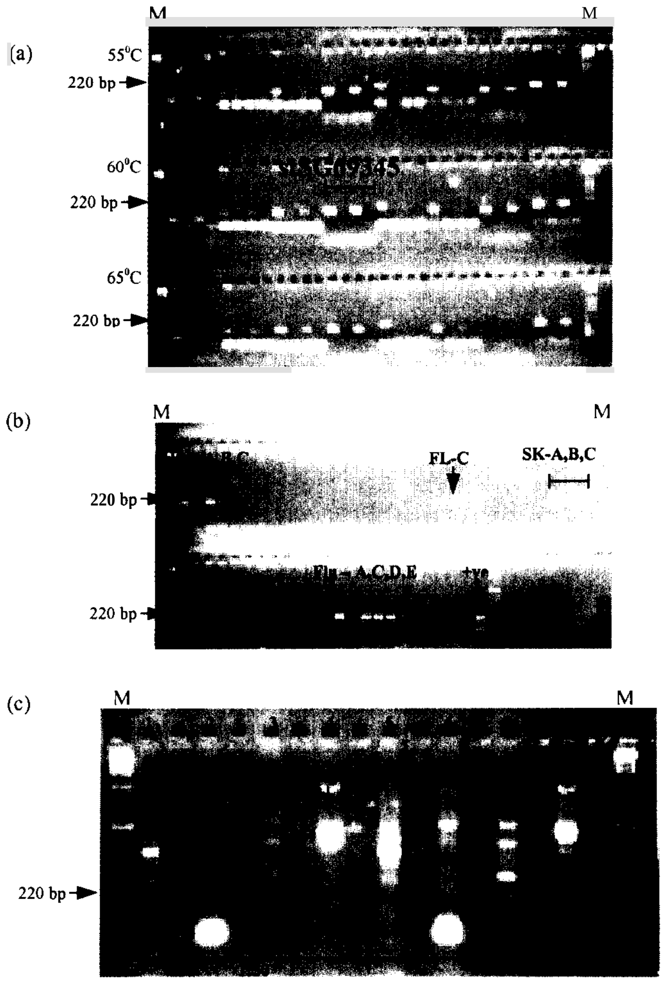


**Figure 4.5:** *cDNA isolation by SSPCR (a) Eight STSs designed to exons ofpredicted genes were testedfor the ability to amplify unique sequences in the human genome, an X chromosome specific hybrid and a hamster cell line, at three different annealing temperatures (M = marker). (b) One of the STSs, stSG69345, was used to amplify DNA ofpools of cDNA clonesfrom 14 different libraries (see Section 2.9). (c) The results of the second round of SSPCRprotocol (see 2.22.1 for schema). A combination of nested sequence-specificprimers (stSG77080S and stSG77080A) and vector-specificprimers (1RP and 2FP) were used to amplify theproductsfrom thefirst round of SSPCR. Two different dilutions of template were used (1:50, lanes 1-4, 1:500, lanes 5-8). Bandsfrom lanes two andfour were excisedfor sequencing.*

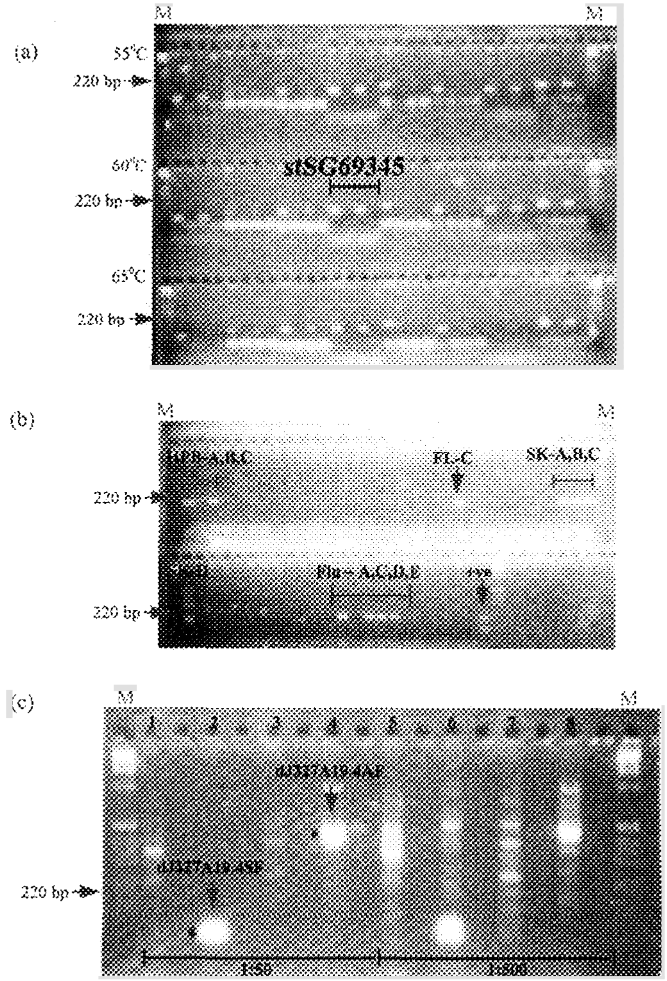


**Figure 4.5:** *cDNA isolation by SSPCR (a) Eight STSs designed to exons of predicted genes were tested for the ability to amplify unique sequences in the human genome, an X chromosome specific hybrid and a hamster cell line, at three different annealing temperatures (M = marker). (b) One of the STSs, stSG69345. was used to amplify DNA of pools of cDNA clones from 14 different libraries (see Section 2.9). (c) The results of the second round of SSPCR protocol (see 2.22.1 for schema). A combination of nested sequence-specific primers (stSG77080S and stSG77080A) and vector-specific primers (1RP and 2FP) were used to amplify the products from the first round of SSPCR. Two different dilutions of template were used (1:50, lanes 1-4, 1:500, lanes 5-8). Bands from lanes two and four were excised for sequencing.*

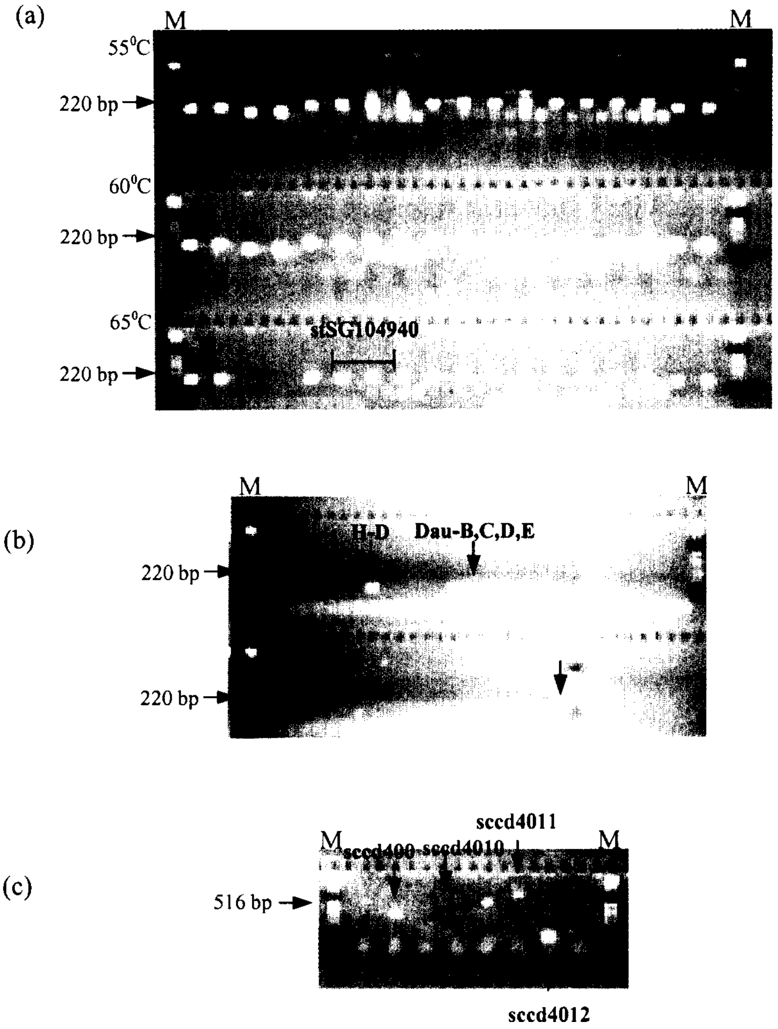


**Figure 4.6:** *cDNA isolation by vectorette PCR (a) Nine STSs designed to exons of predicted genes were testedfor the ability to amplify unique sequences in the human genome, an X chromosome specific hybrid and a hamster cell line, at three different annealing temperatures of the PCR (M = marker). (b) One of the STSs, stSG104940, was used to amplify DNA ofpools of 100,000 cDNA clonesfrom 11 different vectorette libraries. (c) Results of vectorette PCR. A combination of sequence-specific primers (stSG104940S and stSG104940A) and vectoretteprimer 224 was used to amplify DNA of two superpools (H-D and DauB – see Section 2.9 for library informatation) at two different concentrations (1:100 and 1:1000). Bandsfrom lanes 2, 5, 6 and 7 were excised,purified, and sequenced.*

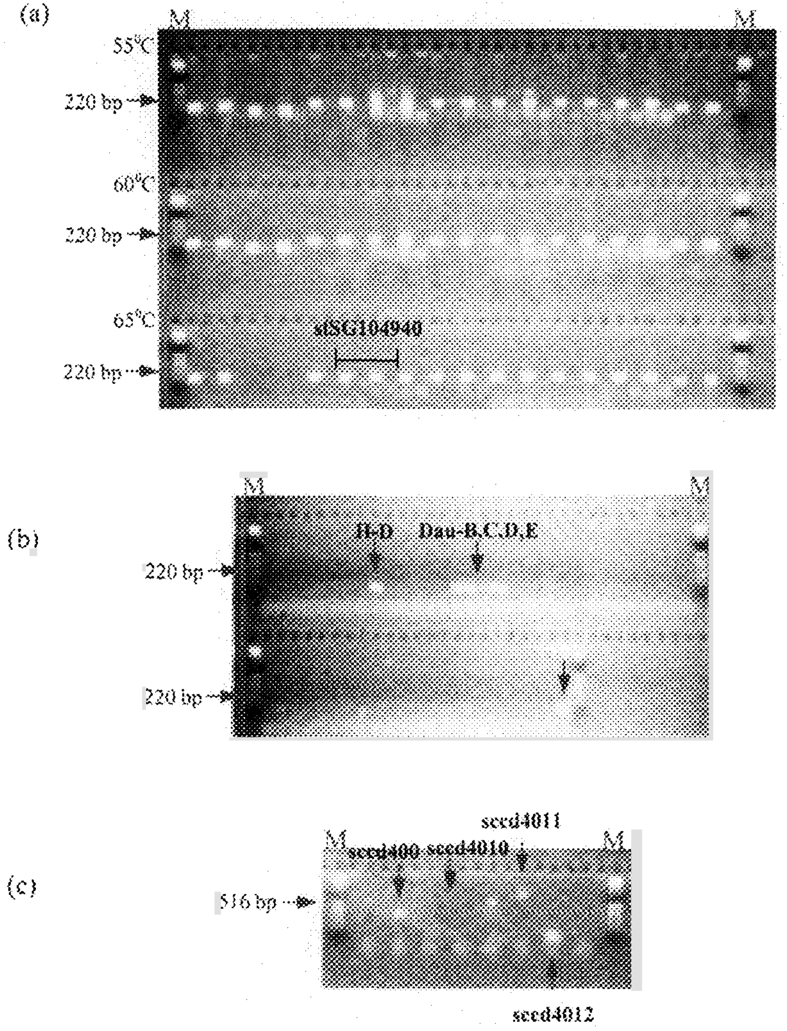
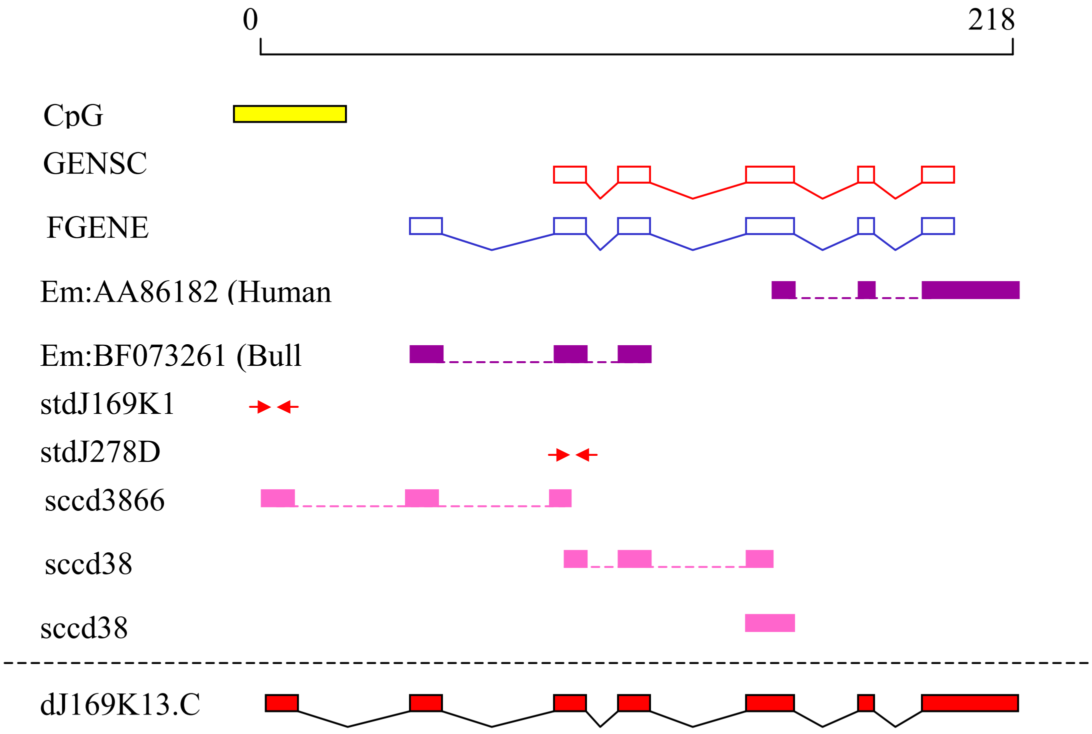


**Figure 4.6:** *cDNA isolation by vectorette PCR (a) Nine STSs designed to exons of predicted genes were tested for the ability to amplify unique sequences in the human genome, an X chromosome specific hybrid and a hamster cell line, at three different annealing temperatures of the PCR (M = marker). (b) One of the STSs, stSG104940, was used to amplify DNA of pools of 100,000 cDNA clones from 11 different vectorette libraries. (c) Results of vectorette PCR. A combination of sequence-specific primers (stSG104940S and stSG104940A) and vectorette primer 224 was used to amplify DNA of two superpools (H-D and DauB – see Section 2.9 for library informatation) at two different concentrations (1:100 and 1:1000). Bands from lanes 2, 5, 6 and 7 were excised, purified, and sequenced.*

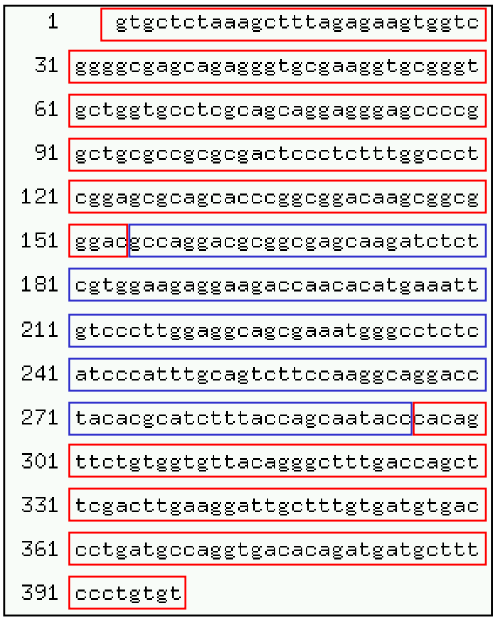
cDNA sequence was aligned to the genomic sequence and the gene structure evaluated for possible extension. Confirmation of a predicted gene was considered complete when there was human cDNA sequence covering at least the predicted protein-coding region, and as much untranslated region (UTR) as possible. A total of fourteen predicted genes were confirmed and eleven gene structures remain unconfirmed. An example of the confirmation of one gene is shown in Figure 4.7.

**Figure 4.7:** *(see over) Confirmation of a novel gene. (a) dJ169K13.CX.1 (exons shown as red boxes, introns as black lines) was predicted by GENSCAN (exons shown as open red boxes linked by red lines) and FGENESH (exons shown as open blue boxes linked by blue lines), and two ESTs (shown in purple), one human and one from bull. A CpG island upstream of the predicted genes suggested a possible location for the 5' end of the gene. Three cDNA sequences (shown as pink boxes) were generated to confirm the 5' end of this gene. (b) The cDNA sequence for sccd3866. Sequence corresponding to exons as they appear in the genomic sequence are coloured as alternating red and blue open boxes.*

(a

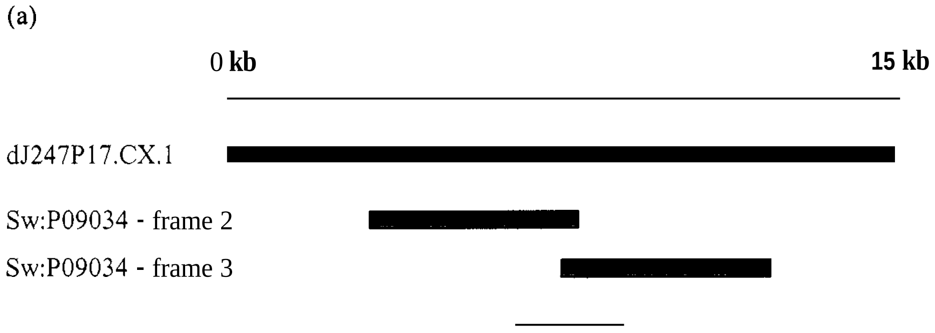


(b



```
  1      gtgctctaaagctttagagaagtggtc
 31   ggggcgagcagagggtgcgaaggtgcgggt
 61   gctggtgcctcgcagcaggagggagccccg
 91   gctgcgccgcgcgactccctctttggccct
121   cggagcgcagcacccggcggacaagcggcg
151   ggacgccaggacgcggcgagcaagatctct
181   cgtggaagaggaagaccaacacatgaaatt
211   gtcccttggaggcagcgaaatgggcctctc
241   atcccatttgcagtcttccaaggcaggacc
271   tacacgcatctttaccagcaataccacag
301   ttctgtggtgttacagggctttgaccagct
331   tcgacttgaaggattgctttgtgatgtgac
361   cctgatgccaggtgacacagatgatgcttt
391   ccctgtgt
```

Twenty pseudogenes were identified within the region and are predicted to have arisen due to the reverse transcription of mRNAs into the genomic sequence. They all appear to have a functional counterpart elsewhere in the human genome. The pseudogenes were identified because they have no introns, a poly A tail within the genomic sequence and a disrupted ORF (see Figure 4.8) (see appendix, Table 4.6 for a full list of the pseudogenes identified in the region).



(b)

```
                TGC TAC TCA CTG GCC CCC CCA GAT
dJ247P17.CX.1    C   Y   S   L   A   P   P   D

Sw:P09034        C   Y   S   L   A   P   Q   I
                TGC TAC TCA CTG GCC CCC CAG ATA
```

**Figure 4.8:** *Example of a pseudogene. (a) The extent of the dJ241P17.CX.1 (shown as a green box) is shown and is a pseudogene of arginosuccinate synthetase* ***(ASS).*** *Part of the protein sequence of* ***ASS*** *(Sw:P09034) aligns in two blocks (shown as blue boxes) (b) Disruption of the reading frame due to an insertion of a C within a run of seven C's (shown as red letters). The alignment of the nucleotide sequence (black letters) and the amino acid sequence (blue letters, using the one letter code) surrounding the insertion are shown.*

In summary, a gene map encompassing the distal portion of Xq23, Xq24 and the proximal portion of Xq25 between DXS7598 and DXS7333 covering 8 Mb has been constructed (see Figure 4.9). The region contains 33 confirmed genes (of which 14 were confirmed during this study), 11 predicted genes and 20 pseudogenes.

**Figure 4.9:** *(see over) A summary of the gene map between DXS7598 and DXS7333. The red bars indicate the contigs status and the black bars indicate the extent of finished sequence. Each link represents a series of individual clones (see appendix to this chapter). Yellow bars indicate clones for which draft sequence is available, and white bars indicate clones selected for sequencing, but not sequenced as of September 2001. A scale is given in megabase pairs (Mb). Approved names are given for known genes (see Table 4.1). Genes are indicated by arrows (black – complete, blue – predicted, green – pseudogene), the direction of each arrow reflects the direction of transcription. Genes on the plus strand are positioned above the dotted line, genes on the minus strand are positioned below the dotted line.*

## 4.3 Evaluation of genes in region

Genes can be evaluated for “completeness” by analysing the genomic sequence for the functional signals (indicated by *italics*). In order for a gene to be transcribed, an RNA polymerase binds within the *core promoter sequence* and initiates transcription at a specific position in the genomic sequence, known as the *transcription start site*. As discussed in Section 1.3, the 5’ end of approximately 56% of genes lies within a *CpG island* (Antequera, F*., et al.*, 1993). Transcription proceeds along the DNA sequence until the RNA polymerase encounters a *polyadenylation signal* (consensus: AATAAA) and transcription terminates soon after. The pre-mRNA then undergoes a series of processing steps including the addition of a 3’ polyA tail and a 5’ cap (an additional guanine added at the 5’ end), and the splicing out of the introns. Three important splice signals (*5’ splice site*, *3’ splice site* and the *branch point*) are involved in the removal of introns and these are recognised by the spliceosome. Approximatley 99.9% of splice sites studied conform to consensus sequences, GT at the 5’ end of the intron (spliced donor), and AG at the 3’ end of the intron (splice acceptor) (Levine, A*., et al.*, 2001).

Ribosomes scan the processed mRNA for a *translation start site* (usually ATG, coding for methionine). Sequence in the mRNA preceding the translation start site is termed *5’ untranslated region* (5’ UTR). Translation continues until the ribosome encounters a *stop codon*, a run of three bases in frame that do not code for an amino acid (UGA, UGG and UAA). Sequence following the stop codon is termed *3’ untranslated region* (3’UTR).

However, because of the difficulty in identifying some of these signals, the presence or absence of such sequences within any one gene can only be used as a guide of the completeness. Furthermore, some of the signals are not found in all genes. At the 5' end of a gene, the transcription is initiated downstream of core promoter sequences such as tata boxes. Programmes such as PromoterInspector (Scherf, M.*, et al.*, 2000) and Eponine (courtesy of Thomas Down) attempt to predict regions likely to contain such sequences. The optimal context for initiation of translation is GCCACCatgG (Kozak, M., 1991), but within this motif, two bases exert the strongest effect, a G at the first base after the ATG, and a purine (preferably A) three nucleotides upstream.

Confidence that the true 3' ends of genes have been identified, can be increased by identifying the polyadenylation signal in either the genomic sequence or the cDNA sequence, of which the most common is AATAAA, but there are other less common sequences used such as TATAAA (Beaudoing, E.*, et al.*, 2000). In some cases aberrant cDNA clones can arise due to the priming of the poly dT primer from a polyA tract in contaminating genomic sequence, but these can be distinguished from real expressed clones as these contain a polyA tail in the cDNA sequence that is not present in the genomic sequence.

Genes within the 8 Mb region have been analysed for the presence of these features. The results are summarised in Table 4.2. All splice sites in the genes identified in this study conformed to the consensus splice site sequence - GT at the 5' end and AG at the 3' end of each intron.

**Table 4.3:** *Evaluation of the gene structures. PI = PromoterInspector, At least one = 5' end predicted by at least one method. A.S = Alternative Splice*

| | 5' end | | | | 3' end | |
|---|---|---|---|---|---|---|
| **Gene Name** | **CpG island** | **PI** | **Eponine** | **At least one** | **Poly A Signal** | **A.S** |
| IL13R | | | | | AATAAA | 2 |
| bA161I19.CX.1 | YES | | YES | YES | | |
| bA161I19.CX.2 | YES | | | YES | | |
| HOM-TES-85 | | | | | AATAAA | |
| T-plastin | | | | | AATAAA | |
| SMT3B | | -9 | | | | |
| hATB0+ | | | | | | |
| dJ169K13.CX.1 | YES | YES | YES | YES | AATAAA | |
| bB115H14.CX.1 | YES | YES | YES | YES | AATAAA | |
| dA155F9.CX.1 | YES | YES | YES | YES | | |
| bB125M24.1 | | | | | | |
| dJ562J12.CX.2 | | | | | GATAAA | |
| dJ170D19.CX.2 | | | | | AATAAA | |
| dJ555N2.CX.1 | | | YES | YES | | |
| HPR6.6 | YES | YES | YES | YES | AATAAA | |
| dJ1139I1.CX.1 | YES | | YES | YES | GATAAA | |
| ANT2 | YES | YES | YES | YES | AATAAA | |
| dJ876A24.CX.1 | | | | | ATTAAA | |
| SEP2 | YES | YES | YES | YES | ATTAAA | |
| NRF | YES | | YES | YES | AATAAA | 2 |
| UBE2A | YES | YES | YES | YES | | |
| RPL39 | YES | | | | AATAAA | |
| U69a | | | | | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| UPF3b | YES | | YES | YES | AATAAA | |
| ZNF183 | YES | YES | | YES | AATAAA | |
| dJ327A19.CX.3 | YES | | | YES | AATAAA | |
| dJ327A19.CX.4 | | -8 | | | | |
| NDUFA1 | YES | YES | | YES | AATAAA | |
| bG421I3.CX.1 | | | -1 | | AATAAA | |
| bG421I3.CX.2 | YES | | | YES | | |
| bG421I3.CX.3 | | | YES | YES | AATAAA | |
| dJ525N14.CX.1 | | -4 | YES | YES | AATAAA | |
| dJ525N14.CX.3 | | -6 | -6 | | | |
| ZNF-kaiso | YES | YES | | YES | ATTAAA | |
| bA45J1.CX.1 | YES | | | YES | AATAAA | 2 |
| ATP1B4 | | | | | | |
| dJ318C15.CX.1 | | -2 | -2 | | AATAAA | |
| LAMP2 | YES | YES | | YES | TATAAA | 2 |
| dJ655L22.CX.1 | YES | YES | YES | YES | ATTAAA | 3 |
| MCT-1 | | | | | AATAAA | |
| dJ321E8.CX.2 | YES | | | YES | | |
| dJ321E8.CX.3 | | | | | TATAAA | |
| GLUD2 | YES | | | YES | | |
| GRIA3 | YES | | | YES | | 2 |
| **Total (%)** | **55** | **30** | **34** | **60** | **64** | **14** |

### *4.3.1 Evaluation of the 5' ends*

CpG rich sequences (identified by CpGfinder) are present at the 5' end of 25 of the 44 genes (56%). PromoterInspector predicts regions likely to contain promoter sequences overlapping the 5' ends of 13 genes (30 %). Eponine predicts transcription start sites at the 5' end of 15 genes (34%). The 5' end of 26 genes (60%) lies within a region predicted by at least one of the programmes. An example of the analysis for the ANT2 gene is shown in Figure 4.10a. In general, PromoterInspector and Eponine predict sequences for those genes associated with a CpG island. *Promoter* sequences or transcription start sites are predicted within 10 kb of a further 5 genes (11%). In general, both PromoterInspector and Eponine predict the presence of promoter sequences or transcription start sites at the 5' end of genes for which a CpG island has been detected. This is consistent with the analysis of the genes on chromosome 22 (John Collins, personal communication).

### *4.3.2 Evaluation of the 3' ends*

As discussed in the previous section some of the cDNA libraries used both in this study and for other cDNA sequencing projects, have been generated by priming from the poly-A tail of mRNAs. In general, the true 3' end of an mRNA is represented in cDNA or EST sequence. However, the presence of a poly-A sequence within the mRNA, can lead to artefacts that represent sequences truncated at the 3' end in cDNAs that are generated using the poly(dT ) primer. An example is shown in Figure 4.10b, where a cluster of apparently 3' EST sequences are located upstream of a second cluster of 3' EST sequences within the 3'UTR of Septin2. The presence of a polyadenylation signal within the sequences of the second cluster increases the likelyhood that the complete 3'UTR has been identified. Analysis of all 44 genes shows that the most common polyadenylation signal (AATAAA), is present within 30 bp of the end of the transcribed sequence in 18 genes (42%). A further 10 genes (23%) contain one of three less common sequences (ATTAAA, GATAAA and AATATA) within 30 bp from the end of the transcribed sequence. These figures are slightly different to those reported by Beaudoing, E., *et al* (2000) who compared 8700 mRNAs and showed that 58% contained AATAA and 20% contained the three less common variants. It is not clear whether the genes with no detectable polyA signal represent incomplete structures, although it is possible that the genes contain other variations of the polyA signal (Beaudoing, E., *et al.*, 2000).

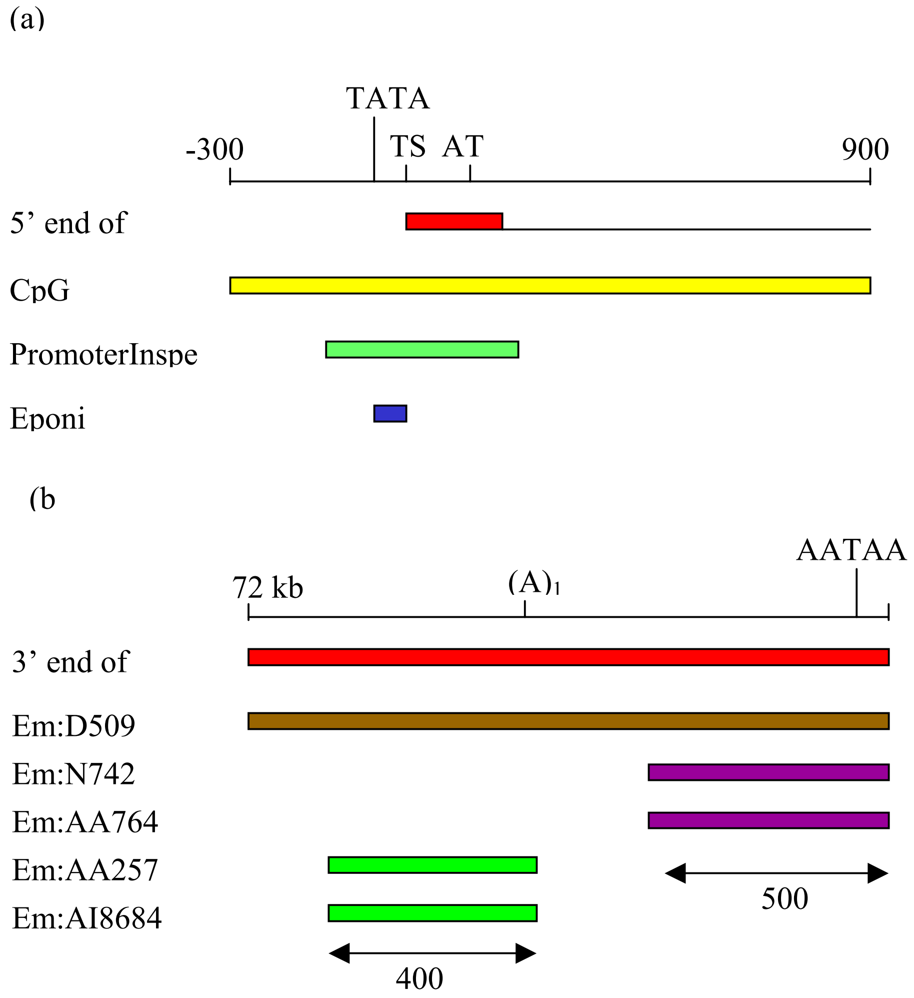


**Figure 4.10:** *Evaluation of gene structures (a) The 5' end of ANT2 (shown as a red box) coincides with the results of analysis using three prediction packages, CpGfinder (yellow box), PromoterInspector (green box) and Eponine (blue box) TATA = TATA box, TS = transcription start, ATG = translation start. (b) The 3' end of Septin2 gene (shown in red) aligns with the 3' end of the cDNA KIA01228 (Em:D50918, shown as a brown box), along with two EST sequences (shown as purple boxes) is the evidence for the correct 3' end of the gene. Two other EST sequences (shown as green boxes) indicate the position of a less likely 3' end on the basis that no polyA signal is observed. The correct 3' end is confirmed by the presence of poly-adenylation signal 15 bases upstream (AATAAA).*

### *4.3.2 Alternative splicing*

As was recently stated in the publication of the draft sequence of the human genome and associated publications (IHGSC, 2001), the number of genes in the human genome is expected to be between 30,000 and 40,000. However, the generation of alternatively spliced transcripts for individual genes increases the complexity of the transcriptome in higher eukaryotes without the need to increase the number of gene loci. Analysis of the 727 genes identified on Chromosome 20 showed there was evidence for alternative splicing for 29% of the genes, with an average of 1.65 transcripts per gene (exluding putatively predicted genes) (Deloukas, P*., et al.*, 1998). In this project described in this chapter, the primary aim was to identify as many genes as possible and at least one full-length transcript. However, where evidence of alternative splicing was observed, the partial transcripts were annotated but not confirmed. In the region studied, there is evidence for alternative splicing in six genes and the average number of transcripts per gene is predicted to be 1.16, which is lower than that observed on Chromosome 20.

An example of a gene with evidence of alternative transcripts is shown in Figure 4.11a. In the example shown, three different transcripts have been identified based on three separate cDNA sequences. The first transcript, dJ655L22.CX.1 aligns with a human cDNA sequence and contains three exons. cDNA sequence generated during this project, showed splicing of exon one to exon three of dJ655L22.CX.1, and this transcript has been termed dJ655L22.CX.1b. A third transcript, dJ655L22.CX.1c shows exon one of dJ655L22.CX.1 splicing onto a novel exon, which in turn splices onto exon three of dJ655L22.CX.1. The presence of this novel exon in dJ655L22.CX.1c is based on an EST sequence from *Bos taurus* and has not so far been seen in human cDNA sequence. A fifth exon may be present in the region, as predicted by GENSCAN, but has not been verified to date.

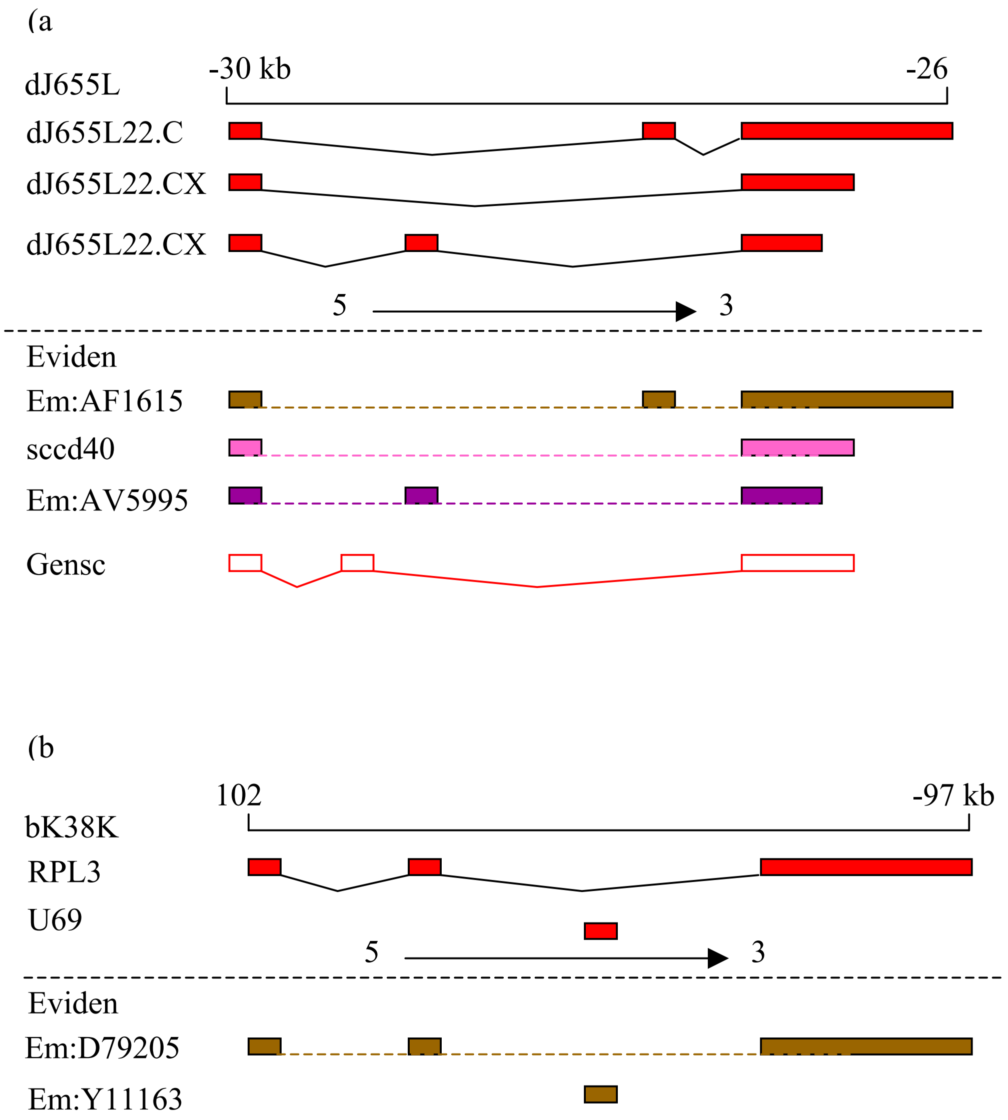


**Figure 4.11:** *Genes in their genomic context (1) (a) Three alternatively spliced transcripts of dJ655L22.CX.1 (shown as red boxes linked by black lines). The evidence for each transcript is shown below the dotted line. Aligned sequence is indicated by boxes linked by dotted lines. (b) The RPL39 gene (shown as red boxes linked by black lines) contains the snoRNA U69a (shown as a single red box) within intron 2. The alignment of each mRNA to the genomic sequence is shown as brown boxes below the dotted line.*

### *4.3.4 Genes in their genomic context*

One of the advantages of systematic gene identification across large regions of sequence is the placement of genes in their genomic context. One phenomenon is the presence of a gene within an intron of a second gene. This has previously been shown in humans and other species, for instance on the human X chromosome, the F8A gene is located within an intron of the Factor VIII gene (Naylor, J. A.*, et al.*, 1995). Figure 4.11b shows a similar example, where U69a, a gene encoding a small nucleolar (sno) RNA is located within intron 3 of RPL39, a ribosomal protein subunit. The positioning of these genes may play a significant role in their function as both are involved in the transcriptional machinery of the cell (Delbruck, S.*, et al.*, 1997; Eliceiri, G. L., 1999).

A second example of gene placement is shown in Figure 4.12a. ZNF183 and NDUFA1, two well-characterised genes are placed on opposing strands, and their transcription start sites are predicted to be only 12 bp apart. Further investigation is needed to identify their respective regulatory elements to determine how these genes are transcribed and controlled, and whether their close proximity plays an important role in their correct functioning.

During the identification of one gene, dJ876A24.CX.1, an EST was aligned to the genomic sequence of dJ876A24 but was assigned to chromosome 8 by the Gene map '98 project (Deloukas, P.*,* *et al.*, 1998, updated electronically 1999, see http://www.ncbi.nih.nlm.gov/genemap99) (see Figure 4.12b and c). Further analysis of the two regions, revealed a pseudogene of dJ876A24.CX.1 on chromosome 8. Primers designed within the EST for RH mapping, flanked an intron within dJ876A24.CX.1 and gave an expected genomic PCR product of 14.8 kb. This would be an X-specific product; none was observed during the RH mapping experiment. By contrast, the absence of the intron between the two primer sequences in the pseudogene on chromosome 8 resulted in the generation of a 287 bp chromosome 8-specific product. This product was detected during the original RH mapping experiments, resulting in the assignment of the EST AA085642 to chromosome 8.

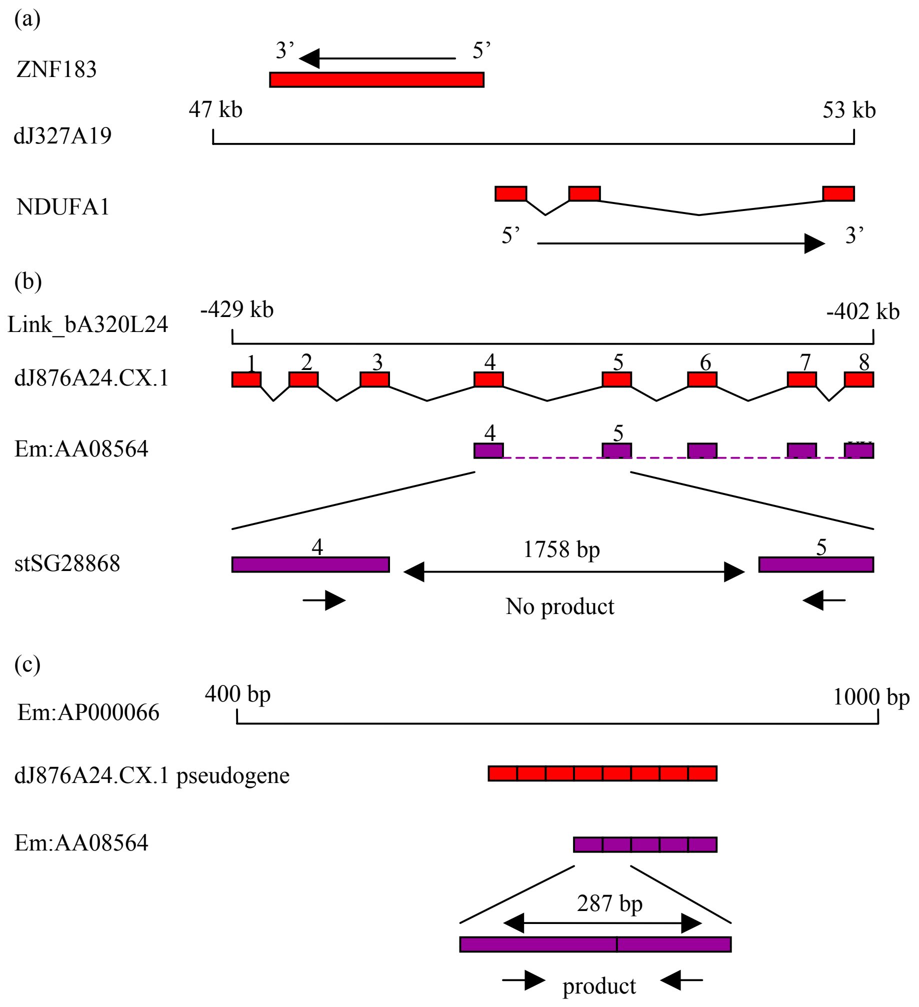
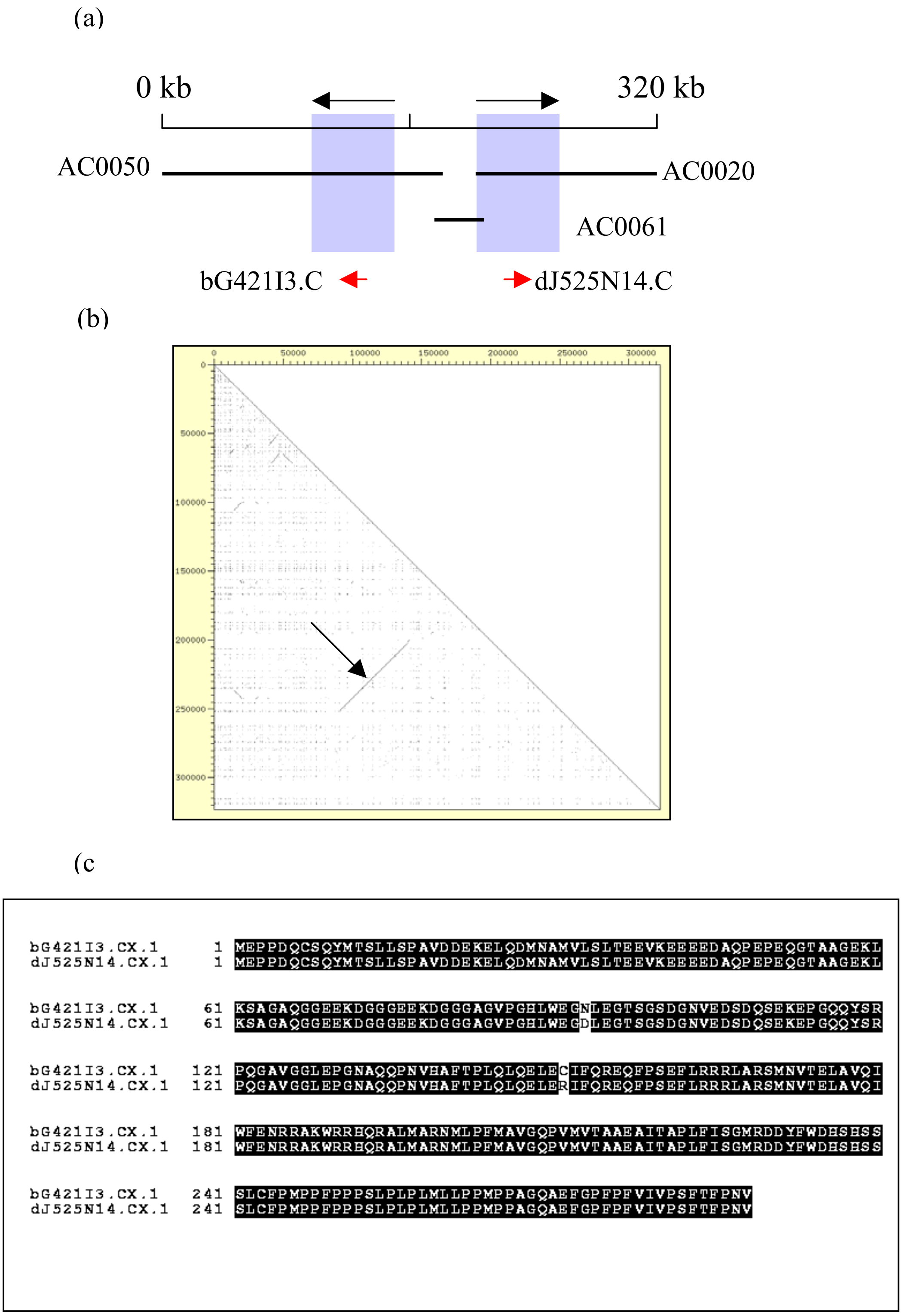


**Figure 4.12:** *Genes in their genomic context (2) (a) ZNF183 (shown as a single red box above the solid line) is transcribed in the opposite orientation to NDUFA1 (shown as red boxes linked by black lines below the line). The position of the regulatory elements for each gene is unknown. (b) dJ876A24.CX.1 (shown as red boxes linked by black lines) was identified in part with an EST (Em:AA085642 shown as purple boxes). The expansion of the alignment of the EST to exons four and five and the position of the primers for stSG28868 is shown. Intron five is 1758 bp and RH mapping using stSG28868 did not reveal a location on the X chromosome. (c) A pseudogene of dJ876A24.CX.1 (shown as red bars) on Chromosome 8 contains no introns and the alignment of the EST (shown as purple bars) generates a product of 287 bp, and the RH mapping did reveal a location on Chromosome 8.*

As was discussed in the previous chapter (Section 3.4), the availability of the genomic sequence allows for the identification of previously uncharacterised low copy repeats. There is an inverted repeat of approximately 50 kb within the region and two very similar genes, dJ525N14.CX.1 and bG421I3.CX.1, are present within the repeat (see Figure 4.13). The predicted amino acid sequence of both genes shows that there are only two amino acid differences between them (Asn93 to Asp and Cys151 to Arg). The cDNA sequence generated to confirm the gene structures matched exactly to dJ525N14.CX.1. The lack of supporting evidence for bG421I3.CX.1 and the non-conservative nature of the amino acid differences does not mean that the gene is not functional. It may be expressed at much lower levels or in a limited number of tissues, and confirmation may be obtained by screening a wider variety of cDNA resources.

**Figure 4.13:** *(see over) Analysis of 50 kb duplication (a) Inverted duplication of a 50 kb region (shown as blue boxes, black arrows indicate orientation). The sequence contribution of each clone is shown as black bars, and EMBL accession numbers of the sequences that make up the region are given. The position of the each gene and the direction of transcription are shown as red arrows. (b) A DOTTER (Sonnhammer, E. L., et al., 1995) of the region matched against itself, the continuous black line along the diagonal is the match against itself, and the smaller black line perpendicular to the central diagonal indicates the position of the duplication. (c) The alignment of the predicted protein sequences of the two genes, dJ525N14.CX.1 and bG421I3.CX.1. Identifical matches are shown as black boxes, the two amino acids that differ (* N to D and ^ C to R) are shown as white boxes.*



```
bG421I3.CX.1      1  MEPPDQCSQYMTSLLSPAVDDEKELQDMNAMVLSLTEEVKEEEEDAQPEPEQGTAAGEKL
dJ525N14.CX.1     1  MEPPDQCSQYMTSLLSPAVDDEKELQDMNAMVLSLTEEVKEEEEDAQPEPEQGTAAGEKL

bG421I3.CX.1     61  KSAGAQGGEEKDGGGEEKDGGGAGVPGHLWEGNLEGTSGSDGNVEDSDQSEKEPGQQYSR
dJ525N14.CX.1    61  KSAGAQGGEEKDGGGEEKDGGGAGVPGHLWEGDLEGTSGSDGNVEDSDQSEKEPGQQYSR

bG421I3.CX.1    121  PQGAVGGLEPGNAQQPNVHAFTPLQLQELECIFQREQFPSEFLRRRLARSMNVTELAVQI
dJ525N14.CX.1   121  PQGAVGGLEPGNAQQPNVHAFTPLQLQELERIFQREQFPSEFLRRRLARSMNVTELAVQI

bG421I3.CX.1    181  WFENRRAKWRRHQRALMARNMLPFMAVGQPVMVTAAEAITAPLFISGMRDDYFWDHSHSS
dJ525N14.CX.1   181  WFENRRAKWRRHQRALMARNMLPFMAVGQPVMVTAAEAITAPLFISGMRDDYFWDHSHSS

bG421I3.CX.1    241  SLCFPMPPFPPPSLPLPLMLLPPMPPAGQAEFGPFPFVIVPSFTFPNV
dJ525N14.CX.1   241  SLCFPMPPFPPPSLPLPLMLLPPMPPAGQAEFGPFPFVIVPSFTFPNV
```

### 4.4. Predicting the function of novel gene products

One of the major challenges once genes have been identified is to predict the role they play within the cell. In order to ascertain the function of the genes experimentally, some initial hypothesis of potential function would greatly facilitate the experimental investigation. As discussed in Section 1.5, there are a variety of methods available to gain an insight into the particular function of novel genes. As part of the systematic identification of the genes, genomic sequence is aligned to previously generated nucleotide and protein sequence. Genes that show similarity to genes of previously determined function may have a related function. In some cases, novel genes are not similar to any genes for which function has been previously determined, and for these, *ab initio* prediction of function can be carried out. For instance, proteins are made up of functional units or domains, and identification of homologues of well-known functional domains with novel proteins may predict specific biochemical functions which may be ascribed to the novel protein.

The 44 genes identified in the 8 Mb region between DXS7598 and DXS7333 have been functionally characterised as far as the available information allows (see Table 4.4). The function of eleven genes within the region has been previously deduced experimentally by others (see reference column in Table 4.4). For instance, the Glutamate Receptor Subunit 3 gene (GRIA3) is a member of the glutamate receptor protein family that mediates most of the excitatory neurotransmission in the mammalian brain (Gecz, J*., et al.*, 1999). The function of a further ten genes is inferred based on their similarity at the protein level to genes whose function has been determined experimentally. For example, Septin2 is a member of the septin family of genes and is an orthologue of the KIAA00128 gene in rat. It has been shown that this rat homologue of the Septin2 is one of four septin proteins that form a filament around which actin bundling can occur (Kinoshita, M*., et al.*, 1997). The function of two of the fourteen novel genes identified and confirmed in this chapter (see Section 4.2) has been inferred based on their similarity to genes of known function. The first novel gene, dJ327A19.CX.4 is 70 % identical at the protein level to a rat gene, testis-specific A-kinase anchoring protein (TAKAP-80), which is involved in sperm motility by binding to a type II cAMP-dependent protein kinase which tightly associated with the fibrous sheath (Mei, X*., et al.*, 1997). The second novel gene with inferred function is dJ318C15.CX.1, which is 90% identical at the protein level to the human gene, cullin 4A (Osaka, F*., et al.*, 1998). Cullins are hydrophilic proteins found in yeast, worm and human that are involved in cell cycle regulation through protein degradation (Mathias, N*., et al.*, 1996).

**Table 4.4:** *Functional characterisation of Genes*

| Gene Name | Characterisation | Function or function prediction | Reference |
|---|---|---|---|
| HOM-TES-85 | Known | HOM-TES-85 tumour antigen | direct submission |
| hATB0+ | Known | Amino acid transporter | Sloan, J. L*.*, *et al.*, 1999 |
| ANT2 | Known | Adenine nuceotide translocator | Giraud, S*.*, *et al.*, 1998 |
| NDUFA1 | Known | Accessory protein in mitochondria complex I | Au, H. C*.*, *et al.*, 1999 |
| LAMP2 | Known | Lysosomal associated membrane protein | Kannan, K*.*, *et al.*, 1996 |
| GLUD2 | Known | Glutamate dehydrogenase | Shashidharan, P*.*, *et al.*, 1994 |
| GRIA3 | Known | Neurotransmitter receptor | Gecz, J*.*, *et al.*, 1999 |
| T-plastin | Known | Actin bundling protein | Lin, C. S*.*, *et al.*, 1999 |
| NRF | Known | Transcription regulation | Nourbakhsh, M*.*, *et al.*, 2000 |
| IL13R | Known | Cell surface receptor, binds IL13 | Aman, M. J*.*, *et al.*, 1996 |
| ZNF-kaiso | Known | Zinc finger transcription factor | Daniel, J. M*.*, *et al.*, 2001 |
| HPR6.6 | Inferred | Steroid binding membrane protein | Gerdes, D*.*, *et al.*, 1998 |
| ZNF183 | Inferred | C3HC4 Ring finger | Frattini, A*.*, *et al.*, 1997 |
| UBE2A | Inferred | Orthologue of yeast RAD6, a DNA repair enzyme | Koken, M. H*.*, *et al.*, 1996 |
| ATP1B4 | Inferred | K-ATPase beta sub-unit | Pestov, N. B*.*, *et al.*, 1999 |
| SMT3B | Inferred | Kinetochore associated protein | Lapenta, V*.*, *et al.*, 1997 |
| Septin2 | Inferred | Septin filament protein | Kinoshita, M*.*, *et al.*, 1997 |
| RPL39 | Inferred | Ribosomal protein | Delbruck, S*.*, *et al.*, 1997 |
| U69a | Inferred | Small nucleolar RNA | direct submission |
| dJ327A19.CX.4 | Inferred | similar to  rat testis-specific A-kinase anchoring protein | Mei, X*.*, *et al.*, 1997 |
| dJ318C15.CX.1 | Inferred | Homologue of cullin-4A | Osaka, F*.*, *et al.*, 1998 |
| bA161I19.CX.1 | Interpro prediction | Leucine rich repeat (IPR001611) | - |
| bA161I19.CX.2 | Interpro | EF-hand (IPR002048), 2 Calporin homology | - |

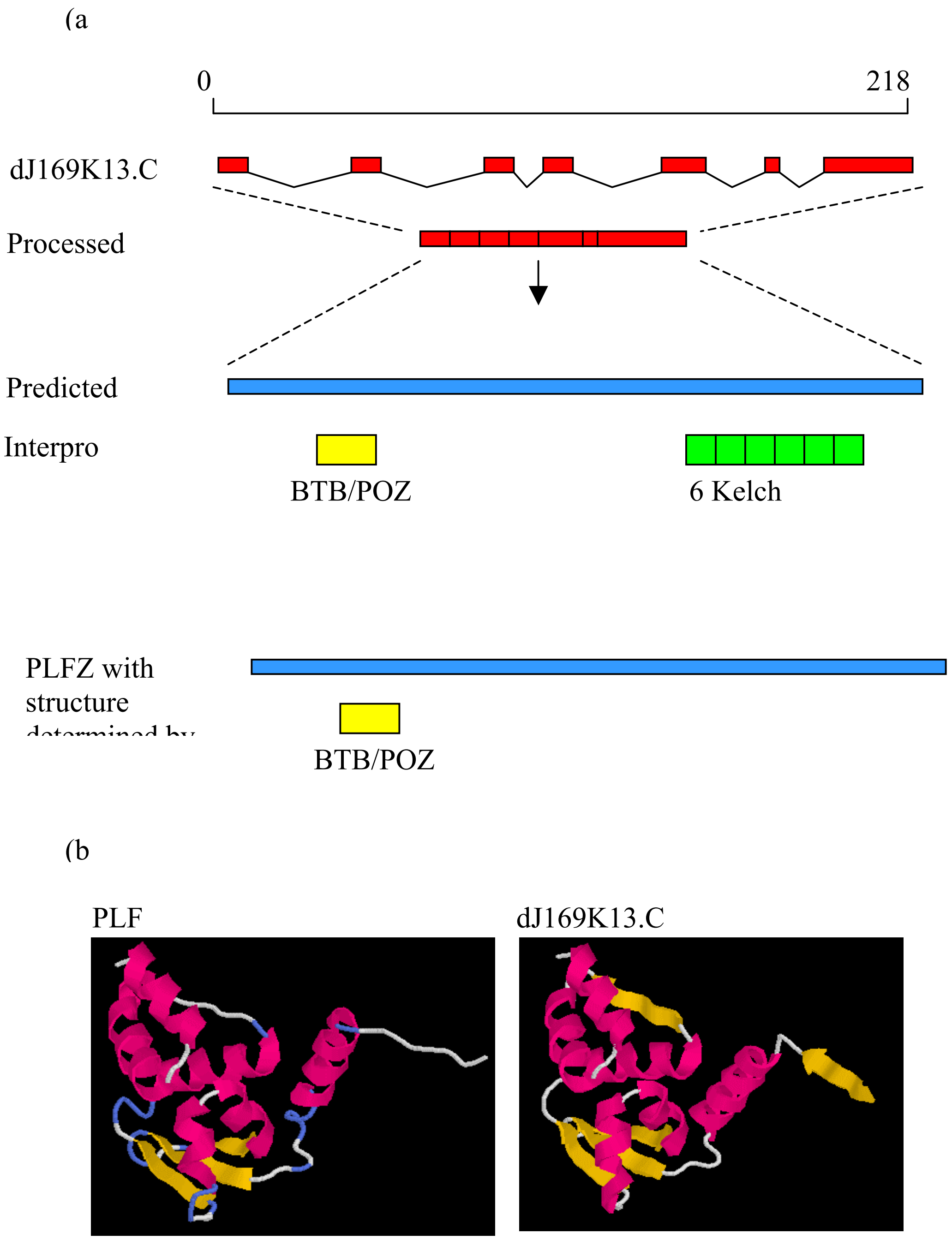
|  |  |  |  |
|---|---|---|---|
|  | prediction | (IPR001715) |  |
| dJ169K13.CX.1 | Interpro prediction | 6 kelch repeats (IPR001798), 1 BTB/POZ (IPR000210 | - |
| dA155F9.CX.1 | Interpro prediction | Plekstrin homology (IPR001849) | - |
| dJ562J12.CX.2 | Interpro prediction | Zinc finger CCHC type (IPR001878) | - |
| dJ555N2.CX.1 | Interpro prediction | 6 bipartite nuclear localisation signals (IPR001472) | - |
| dJ1139I1.CX.1 | Interpro prediction | 3 mitochondrial energy transfer (IPR001993 | - |
| UPF3b | Interpro prediction | 5 bipartite nuclear localisation signals (IPR001472) | - |
| bG421I3.CX.1 | Interpro prediction | Homeobox (IPR001356), proline rich (IPR000694) | - |
| bG421I3.CX.3 | Interpro prediction | Homeobox (IPR001356) |  |
| dJ525N14.CX.1 | Interpro prediction | Homeobox (IPR001356), proline rich (IPR000694) | - |
| dJ876A24.CX.1 | Tmpred | Transmembrane protein | - |
| bB115H14.CX.1 | Unknown | - | - |
| bB125M24.1 | Unknown | - | - |
| dJ170D19.CX.2 | Unknown | - | - |
| dJ327A19.CX.3 | Unknown | - | - |
| bG421I3.CX.2 | Unknown | - | - |
| dJ525N14.CX.3 | Unknown | - | - |
| bA45J1.CX.1 | Unknown | - | - |
| dJ655L22.CX.1 | Unknown | - | - |
| MCT-1 | Unknown | - | - |

|  |  |  |  |
|---|---|---|---|
| dJ321E8.CX.2 | Unknown | - | - |
| dJ321E8.CX.3 | Unknown | - | - |

The remaining 23 genes do not appear to show any similarity to genes of known function, and these have been analysed for the presence of protein domains. This analysis was carried out using INTERPRO, a web-based tool that uses a series of protein domain prediction programmes (see http://www.ebi.ac.uk/interpro/scan.html) which compare novel protein sequence to sequences of known protein domains.

The analysis shows that eleven genes are predicted to contain one or more previously defined protein domains and an example of this is shown in Figure 4.14. One gene, dJ169K13.CX.1 is predicted to contain a BTB/POZ domain and 6 kelch repeats. The kelch repeat was first identified in the kelch gene from *Drosophila*, and repeating kelch units form anti-parallel beta-sheets that come together to form a propeller like structure. Proteins containing six-kelch repeats have been shown to have a role in actin bundle formation (e.g. scruin, Way, M*., et al.*, 1995), and this may suggest a possible role for dJ169K13.CX.1.

**Figure 4.14:** *(see over) Functional analysis of genes (a) Identifying a possible function for dJ169K13.CX.1 (shown as red boxes linked by black lines). The processed transcript (shown as red boxes) is translated to generate the predicted protein sequence (shown as a blue box), which is compared to known protein and their domains. dJ169K13.CX.1 is predicted to contain a BTB/POZ domain (shown as a yellow box) and 6 kelch repeats (shown as green boxes). PLFZ, a protein whose structure has been determined by X-ray crystallography also contains a BTB/POZ domain and the two sequences corresponding the BTB/POZ domain were aligned and the structure of the domain within dJ169K13.CX.1 was predicted using MODELLER. (b) The predicted structure of the BTB/POZ domain of dJ169K13.CX.1 compared to the known structure of PLFZ (viewed in RASMOL). The domain is made up of a series of alpha helices (shown in red) flanked by beta sheets (shown in yellow). Looping regions are shown as either white or blue lines.*


(a
0
218
dJ169K13.C
Processed
Predicted
Interpro
BTB/POZ
6 Kelch
PLFZ with
structure
determined by
BTB/POZ
(b
PLF
dJ169K13.C

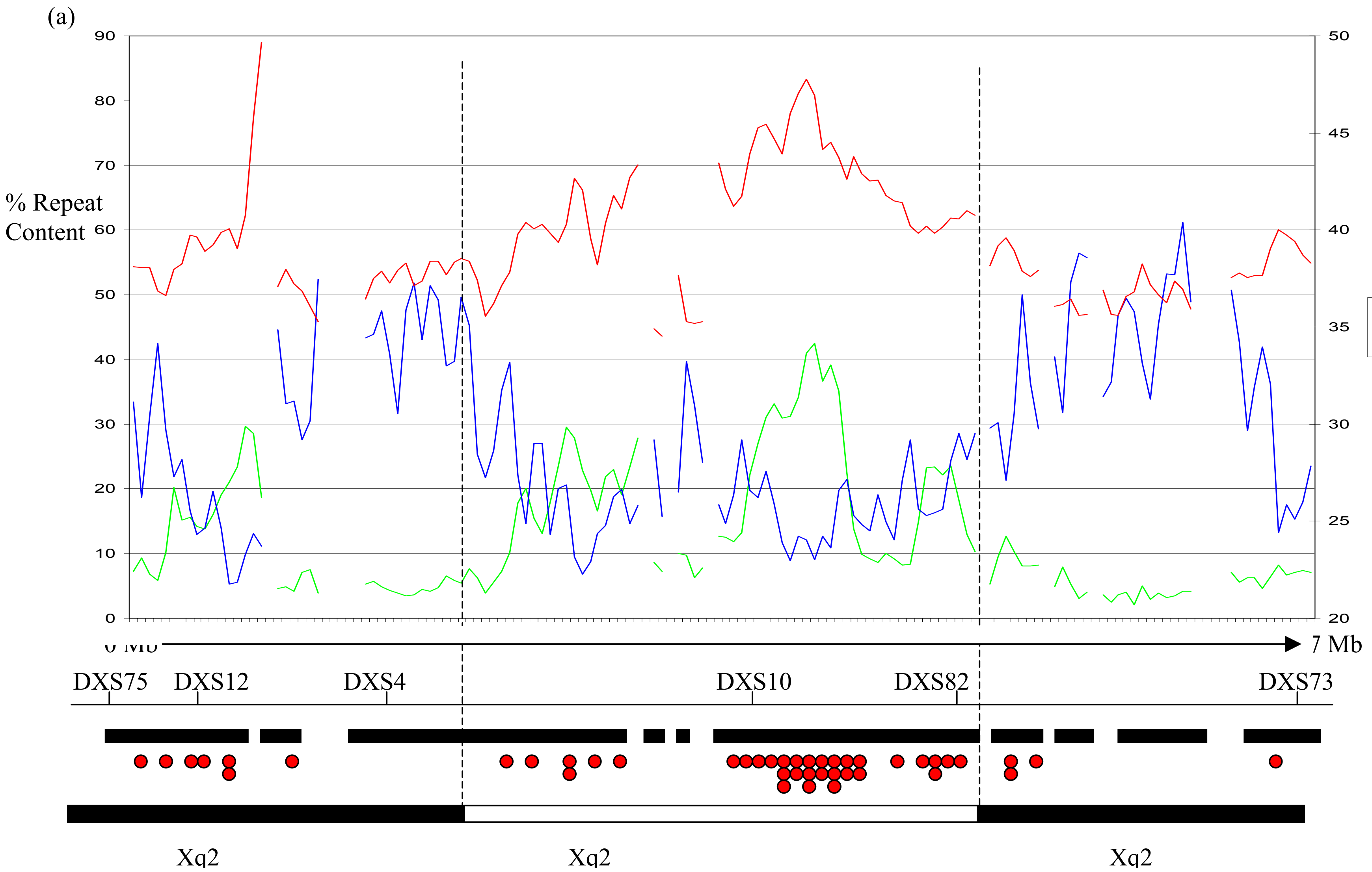
The BTB/POZ domain (BTB = BR-C, ttk and bab (Zollman, S.*, et al.*, 1994), and POZ = Pox virus and Zinc finger (Bardwell, V. J.*, et al.*, 1994) is a much simpler domain, comprising of a cluster of alpha-helices flanked by short beta-sheets, determined using X-ray chrystallography on the promyelocytic leukemia zinc finger oncoprotein (plzf) (Ahmad, K. F.*, et al.*, 1998). The BTB/POZ domain mediates homomeric and in some cases heteromeric dimerisation. An alignment of the BTB/POZ domains from plfz and dJ169K13.CX.1 was carried out using CLUSTALW (data not shown). The structure of the BTB/POZ domain in dJ169K13.CX.1 was then predicted using MODELLER which uses experimentally determined protein sequences to predict conformation of the other proteins with similar amino acid sequences (Sanchez, R.*, et al.*, 1997). This is possible because a small change in the sequence usually results in a small change in the 3D structure. The predicted structure for the BTB/POZ domain of dJ169K13.CX.1 is shown in Figure 4.14b.

### 4.5 Analysis of the sequence composition of the region in Xq23-Xq24

The availability of the genomic sequence and the identification of the genes contained within allow for a detailed analysis of the correlation between the overall sequence composition of the region, the distribution of the repetitive elements and the gene density. It also allows for a comparison with other regions of the human genome. Previous mapping and cytogenetic analysis predicted the region of interest contains the distal portion of Xq23, part of a G light band, and the whole of Xq24, a complete R dark band. As discussed in section 3.4, R-bands are thought to be generally gene rich and SINE rich, while G-bands are thought be generally gene poor and SINE poor. The availability of the genomic sequence and the identification of genes contained within allows for a detailed analysis to be carried out, to determine the relationship between gene content and sequence composition.

The twelve sequence contigs in the region were analysed by dividing each sequence contig into 100 kb segments that overlapped by 50 kb. Each segment was then analysed for the repeat content and GC content using RepeatMasker (Smit, AFA & Green, P. RepeatMasker at http://ftp.genome.washington.edu/RM/RepeatMasker.html) and the results plotted as single points on a graph (see Figure 4.15a). The positions of the genes relative to the sequence segments were identified. The lines indicate marked alterations in GC content, LINE and SINE content. These would be consistent with possible location of boundaries defined cytogenetically assuming one or more of these sequence features affects intensity of Giemsa staining. The GC content is highest in the region predicted to be the light band Xq24, and coincides with the highest density of genes. The areas of high SINE content and low LINE content correspond with a relatively high gene density. This observed correlation between SINE content and gene density agrees with observations carried out over the whole genome (IHGSC, 2001).

***Figure 4.15:*** *(a) (see over) Genome landscape of the region of interest. The X-axis represents the extent of the region divided into 100 kb segments, overlapping by 50 kb. The Y-axis on the left is the % repeat content, and the one on the right is the % GC content. Gaps in the plot indicate gaps in the finished sequence. The lines indicate marked alterations in GC content (red), LINE (blue) and SINE content (green). The genomic interval is represented as single thin black line below the chart, with key markers positoned. The extent of sequencing is shown as black boxes and the position of each gene is represented by a red circle. The hypothetical positions of the cytogenetic bands are also given (black dotted lines). In general, there is good correlation between SINE content and gene density.*


(a)
% Repeat Content
90
80
70
60
50
40
30
20
10
0
% GC Content
50
45
40
35
30
25
20
SINE
LINE
GC
0 Mb
7 Mb
DXS75
DXS12
DXS4
DXS10
DXS82
DXS73
Xq2
Xq2
Xq2

Comparisons were also carried out with two other light bands on the X chromosome (Xp11.23 and Xq26.1) as well as with the whole of the X chromosome, chromosome 22 and the whole genome (see Figure 4.15b). For each X chromosome band, a plot of the GC and repeat content was generated as was described for Xq23-Xq24 and average figures for GC content, SINE, LINE and particularly L1, given the proposed role LINE elements are thought to play in X inactivation (Lyon, M. F., 1998 and see Section 1.5). The extent of each band was identified using sequence content (high GC and SINE content) and the expected position of the band based on previous mapping information (FISH data and gene and marker placement). The average figures for the whole of the X chromosome and the whole genome were obtained from analysis of sequence in ENSEMBL (courtesy of Ewan Birney and Mark Ross) and figures for chromosome 22 were obtained from published sources (Dunham, I., *et al*, 1999).

Xq24 has an average GC content of 40% and gene density of 10 per megabase (taking 33 genes in the 3.3 Mb). In comparison, chromosome 22 has a much higher GC content (48%) and a higher gene density (22 genes per megabase). Xq24 has a relatively high SINE content of 21% when compared to the rest of the genome (13%), as expected, as the figures for the whole genome includes regions within R-band which are thought to be gene poor and SINE poor. However, Xq24 has a much lower SINE content when compared to Xp11.23 (33%), which is considered to be one of the most gene rich regions on the X chromosome (IHGSC, 2001). Although chromosome 22 is considered to be a gene-rich chromosome, and SINE content is thought to be linked to gene density, the average SINE content on chromosome 22 (17%) is less than that of both Xq24 (21%) and Xp11.23 (33%). However, chromosome 22 is gene rich when compared to other whole chromosomes, not individual light bands, and therefore as expected has a higher SINE content when compared to the average figures for the whole of the X chromosome (11%).

The Lyon hypothesis suggests an involvement of L1 elements in X-inactivation and hypothesises a higher LINE1 content on the X chromosome than for the other regions of the human genome (as discussed in Section 1.5). It has been shown that the X chromosome in general has a much higher L1 content (30%) when compared to the average figure for the whole genome (17%) (IHGSC, 2001). Xq24 has a higher L1 content (13%) than that reported previously for Chromosome 22 (9%) supporting Lyon hypothesis that there will be more LINE1 elements on the X chromosome.

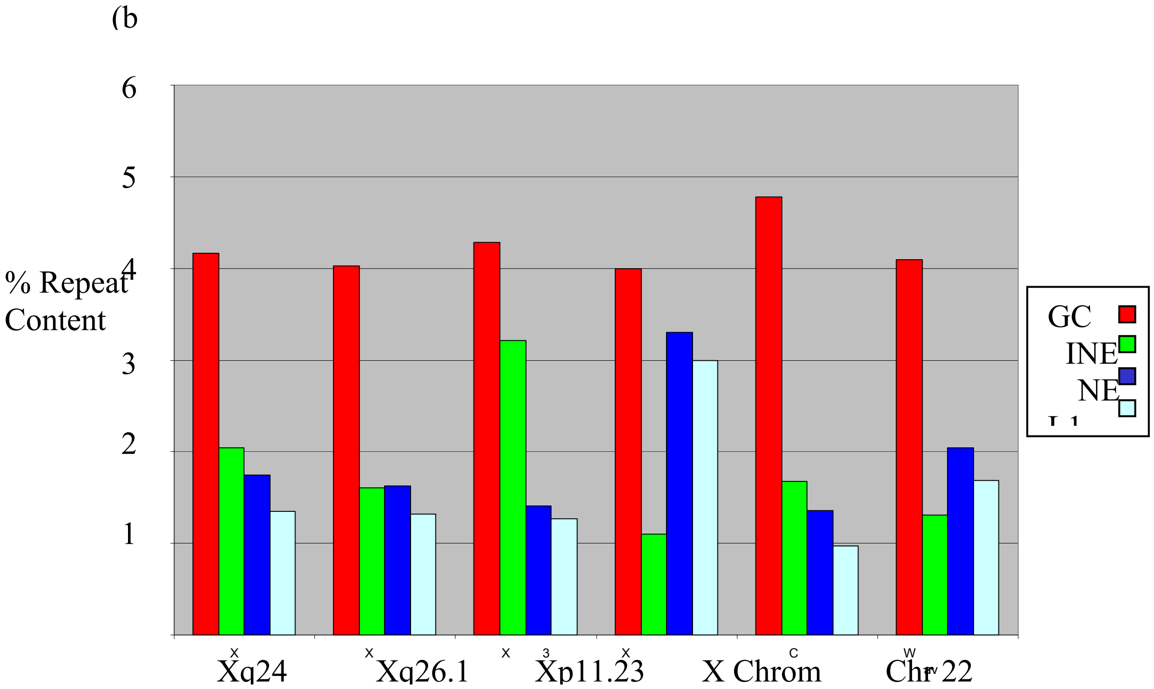


**Figure 4.15:** *cont. (b) Comparisons of genome landscapes of other regions of the genome. The average GC content (red bars), SINE content (green bars), LINE content (dark blue bars), and LINE1 content (light blue bars) across the region has been calculated. Assuming there is a correlation between SINE content and gene density, Xq24 appears slightly more gene dense than Xq26.1, but much less gene dense than Xp11.23, one of the gene dense regions of the X chromosome (IHGSC, 2001) . There appears to be a higher LINE1 content in the X chromosome bands than on Chromosome 22, which agrees with the hypothesis that LINE1 elements maybe involved in X-inactivation.*

### 4.6 Mutation screening for MRX23

The identification of genes within a given region provides a valuable resource in the search for disease causing mutations. The critical regions for a number of diseases for which no causative mutation has been identified, are contained within or overlap with the 8 Mb region between DXS7598 and DXS7333 (see Figure 4.16). The nature of X-linked non-specific mental retardation (MRX) syndromes (the absence of any distinguishing phenotype other than mental retardation) results in the assignment of a different MRX number to each family diagnosed with the condition. MRX23 (discussed further in Section 1.4.3) has previously been localised to a region of the X chromosome between DXS1220 and DXS424, a region of approximately 2.4 cM (Gregg, R. G*., et al.*, 1996). Analysis of the available sequence shows that these markers have now been accurately positioned. DXS1220 is located within Chr_Xctg20, 500 kb from the distal end, and DXS424 is located within Chr_Xctg3, 150 kb from the proximal end. The gap between the two contigs has been sized by fibre fish to approximately 300-500 kb (see Figure 4.9).

Work is currently underway to generate sequence across the gap between the two contigs as part of the X chromosome mapping and sequencing project. As a result the critical region for MRX23 can now be more precisely defined, with boundaries accurately placed on the genomic sequence. Allowing for upper and lower estimates of the size of the remaining gap, the critical region is thought to be between 950 kb and 1150 kb. To date, the critical region between DXS1220 and DXS424 has been shown to contain 3 genes, hATB$^{0+}$, T-plastin and SMT3B. hATB$^{0+}$ has been shown to be an amino acid transporter, with the particular affinity for hydrophobic amino acids (Sloan, J. L*., et al.*, 1999). T-plastin is a member of the plastin family of actin binding proteins and is expressed in most tissues in higher eukaryotes (Arpin, M*., et al.*, 1994, Lin, C. S*., et al.*, 1993). Although little is known about the function of SMT3B, it shares 87% amino acid identity with SMT3A, and both are thought to be human homologues of the yeast SMT3, which are essential in chromosome segregation (Lapenta, V*., et al.*, 1997). Genes with a wide variety of function are associated with mental retardation-related disorders therefore all genes were considered to be positional candidates and screened for disease-causing mutations. Mutation screening of all three candidate genes was carried out by amplification of the exons using DNA from an affected individual and DNA from an unaffected, unrelated individual and sequencing the products.

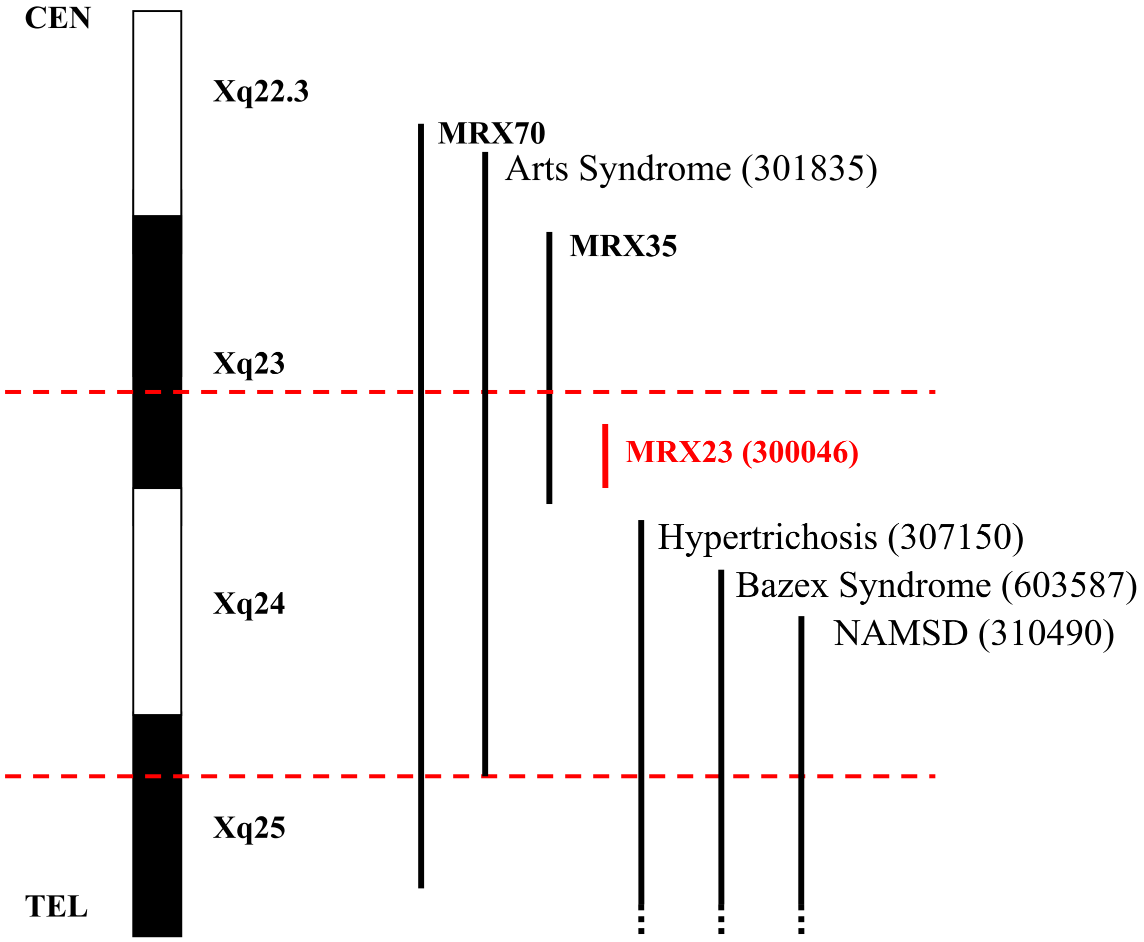


**Figure 4.16:** *Uncloned diseases mapping to region of interest. A section of the X chromosome showing the extents (vertical black lines) of the critical regions of uncloned diseases that include the region of interest between DXS7598 and DXS7333 (between the two dotted red lines). The critical region for MRX23 is shown in red.*

*OMIM numbers, where available are given in brackets. For information on MRX families see Toniolo, D., et al., 2000.*

A total of 31 primer pairs were designed to amplify each of the exons of the three genes including at least 50 bp of flanking intron sequence (see Table 4.7 in appendix to this chapter). The PCR was used to amplify DNA from one affected individual (kindly provided by Ron Gregg, see Section 2.7) and one unaffected, related individual. The products were excised, purified and sequenced (for example see Figure 4.17). In each experiment, a control STS lying close to the candidate gene was amplified as a positive control in the case of a deletion of an exon or exons where no product would be observed. To date, sequences from all 31 exons have been compared and only one difference has been observed between affected and unaffected individuals.

A single base difference, T to A, at nucleotide position 36 of exon 2 of the hATB$^{0+}$ gene was identified in the affected individual. This is a synonymous change as it alters GGT to GGA, both of which code for glycine (see Figure 4.18). This alteration in the sequence does not appear to cause the formation of a cryptic splice site, the single change does not appear to introduce a consensus sequence for a novel 5' or 3' splice site. Analysis of other sequence aligned to the genomic sequence in the region (cDNA and EST sequence) shows that only DNA sequence generated from the affected individual contains GGA, and all other sequence contained GGT.

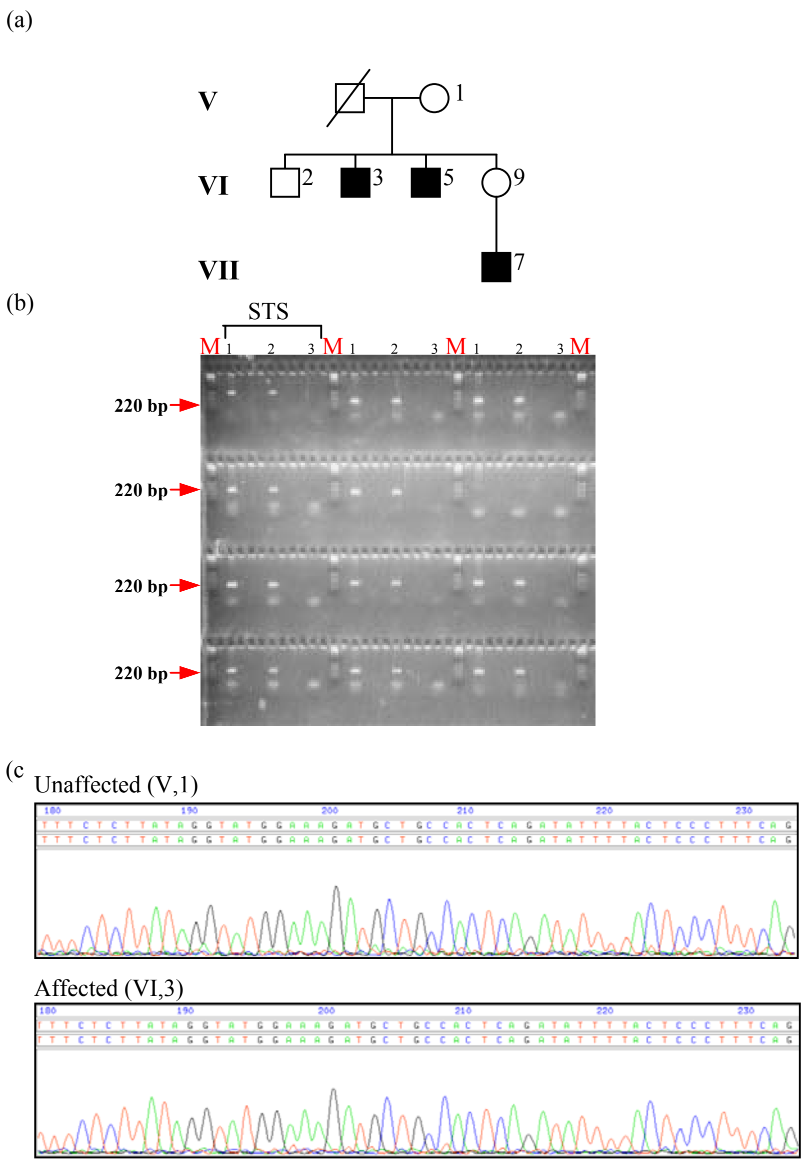


**Figure 4.17:** *Mutation screening for MRX23 (a) Part of the pedigree for the MRX23 family indicating the relationships between unaffected and affected family members (Males = squares, females = circles, affected = shaded black, deceased = line through). DNA from VI,3 was used for mutation screening. (b) Twelve STSs designed to amplify exons 1-12 of the T-plastin gene, screened against DNA from affected and unaffected individuals (lane 1 = unaffected, lane 2 = affected, lane 3 = $T_{0.1}$E, M = marker). (c) The sequence of part of exon 4 from both an affected and an unaffected family member. Alignment of the two sequences reveals no differences.*

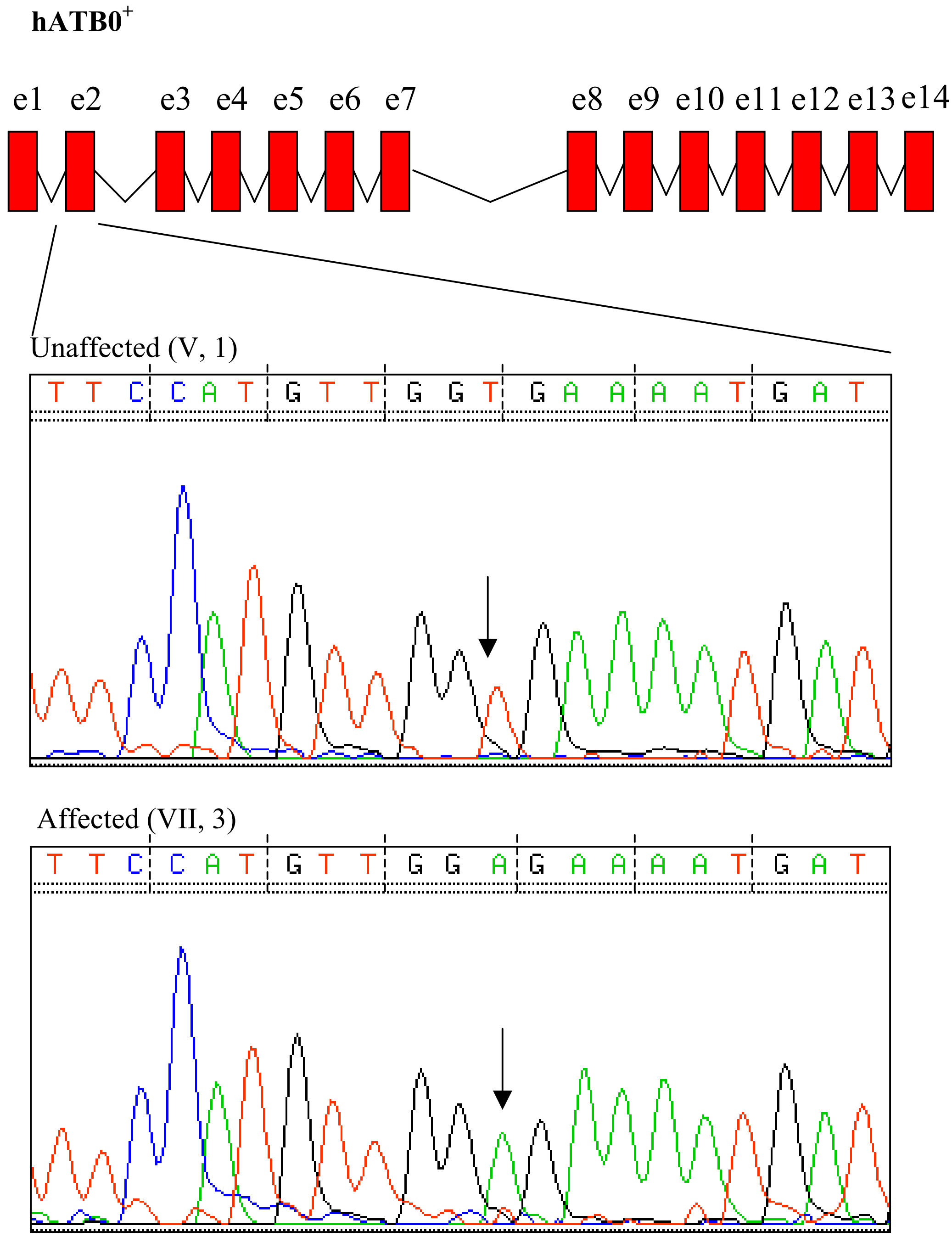


**Figure 4.18:** *Identification of a potential silent mutation. Top of the figure shows a schematic of the the hATB0+ gene (exons shown as red boxes, introns shown as 'v'-shaped lines). The bottom of the figure shows a potential silent mutation (GGT to GGA, indicated by the black arrow on the sequence (viewed in TREV) from both the unaffected and the affected) was observed within exon two of the hATB0+ gene (shown as red boxes linked together with black lines).*

A more detailed analysis has not been carried out while the sequence of the region remains incomplete and therefore while one or more other candidate genes may be discovered and included in the first phase of mutation detection. Further analysis of the GGT to GGA variant would involve testing other members of the family, as well as other unaffected individuals to exclude the possibility it is a polymorphism in the population and that the change is specific to the MRX23 family. If the polymorphism did appear to segregate with the affected individuals a functional assay could be carried out to to identify any functional implications of the variant.

## 4.7 Discussion

A gene map covering approximately 8 Mb of Xq23-24 between DXS7598 and DXS7333 has been constructed and contains 33 confirmed genes of which 14 have been described in this chapter, 11 genes predicted by a series of similarity searches and gene prediction packages and 20 pseudogenes. Contiguation of the bacterial clone contigs covering the region and subsequent completion of the genomic sequence, will provide the basis for the identification of further novel genes in this region. The largest sized gap in the region lies within the critical region for MRX23 and is thought to be approximately 500 kb.

Where possible, the entire predicted ORF for each gene has been confirmed in cDNA sequence. The methods described here for gene identification require the presence of cDNA sequence to confirm a predicted gene. A number of different cDNA sequencing projects around the world (for a full list see Table 4.8 in appendix to this chapter) are depositing cDNA sequence into public databases and Figure 4.19 shows the relative contribution of each collection to the confirmation of genes in the region studied. This shows that random cDNA sequencing can greatly facilitate the identification of genes.

The gene structures predicted by GENSCAN and FGENESH have been compared to the genes that were confirmed by cDNA or EST sequence using BLAST (Altschul, S. F*., et al.*, 1990), and GENSCAN correctly predicted 42.4 kb of the 80.9 kb of expressed bases (52%), whereas FGENESH correctly predicted 40.5 kb (50%) (see Figure 4.20a – red bars). As gene prediction programs only predict coding sequences, the figure is artificially lower due to the inclusion of the untranslated regions in the total amount of expressed sequence. Figure 4.19b shows two examples of genes predicted using GENSCAN and FGENESH. In general, they are better at predicting internal exons but are unable to predict the 3’ and 5’ most exons. This is consistent with analysis carried out on larger gene sets (e.g Chr22 gene set, Dave Beare personal communications, The Sanger Institute).

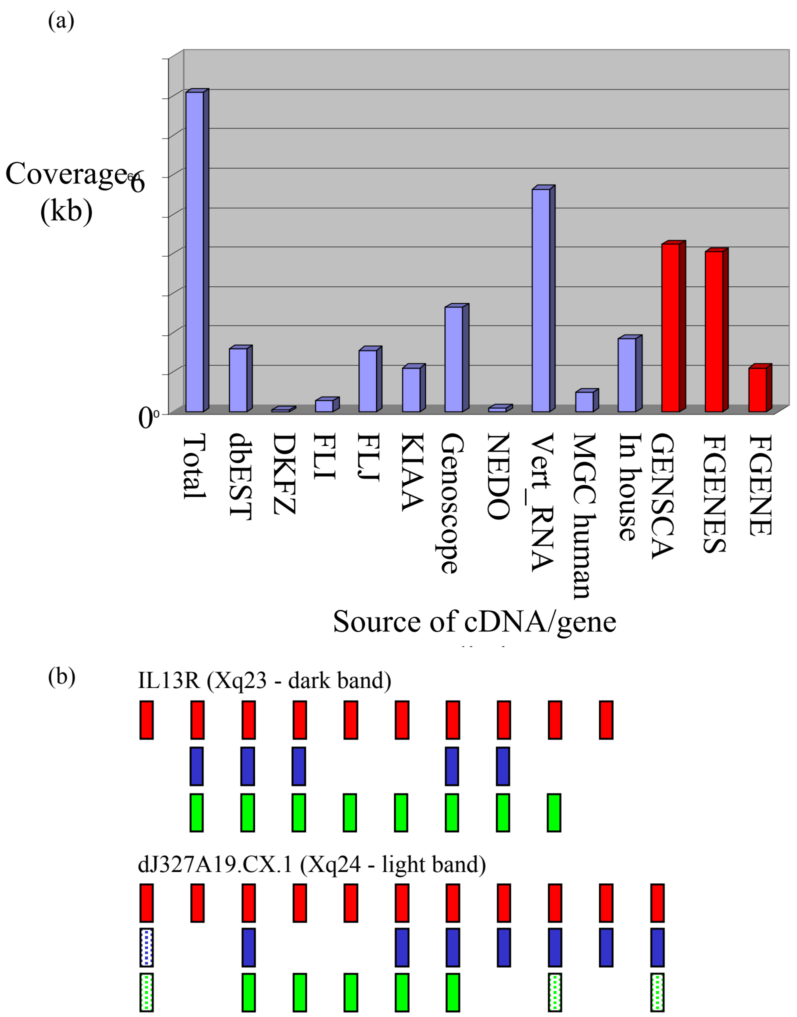


**Figure 4.19:** *Contributions of cDNA sequencing projects and prediction programs (a) Breakdown of the relative contribution of each cDNA resource (blue bars) and each gene prediction package (red bars). The total amount of expressed sequence (first bar) is approximately 80 kb. The vertebrate mRNA database (vert_mRNA) provides over 50 kb and 18 kb of cDNA sequence was generated in house. GENSCAN and FGENESH predict around 50% of the expressed sequence. This is slightly lower than expected as the total figure for expressed sequence includes UTR, regions the gene prediction programs do not predict. (b) Two examples of genes (exons shown as red boxes) predicted by GENSCAN (blue boxes) and FGENESH (green boxes). Partially predicted exons are represented by hashed boxes.*

The region of interest was contained in twelve sequence segments covering 7 Mb. A further 600 kb of sequence was available as unfinished sequence. Analysis of the unfinished sequence revealed no new genes and one pseudogene. This finding is not suprising as 90% of the draft sequence was predicted to be within gene-poor regions at the proximal portion of Xq25. In two cases though, the draft sequence contained exons from genes placed on the finished sequence (GRIA3 and Serotonin-5HTR2), and these genes both cover large regions of genomic sequence. Therefore one of the disadvantages of draft sequence is that the larger gene structures, the genes with large introns are more likely to be present on multiple draft sequence contigs.

As discussed in Section 4.3, it is difficult to identify the complete ORF and flanking UTR sequences for each gene. In this study a total of 22 primer pairs, designed to regions thought to contain genes, failed to identify any positive pools in the cDNA libraries. In one case, a PCR assay was designed to generate cDNA sequence at the 5' end of a gene (see Figure 4.20a). In a second example, the primer pair was designed to an exon of a gene predicted by both GENSCAN and FGENESH, and with a BLASTX homology (see Figure 4.20b). Both PCR assays failed to identify any positive pools in the libraries available. The absence of such confirmation for predicted genes does not mean the gene is not real. It may be expressed in a tissue that is not represented in the cDNA collection tested or at low levels, or for a short period of time, for instance during development. Screening of cDNA libraries from a wider variety of human tissues will increase the likelihood of confirming a particular gene but one complementary strategy (discussed in chapters five and six) is to use comparative sequence analysis.

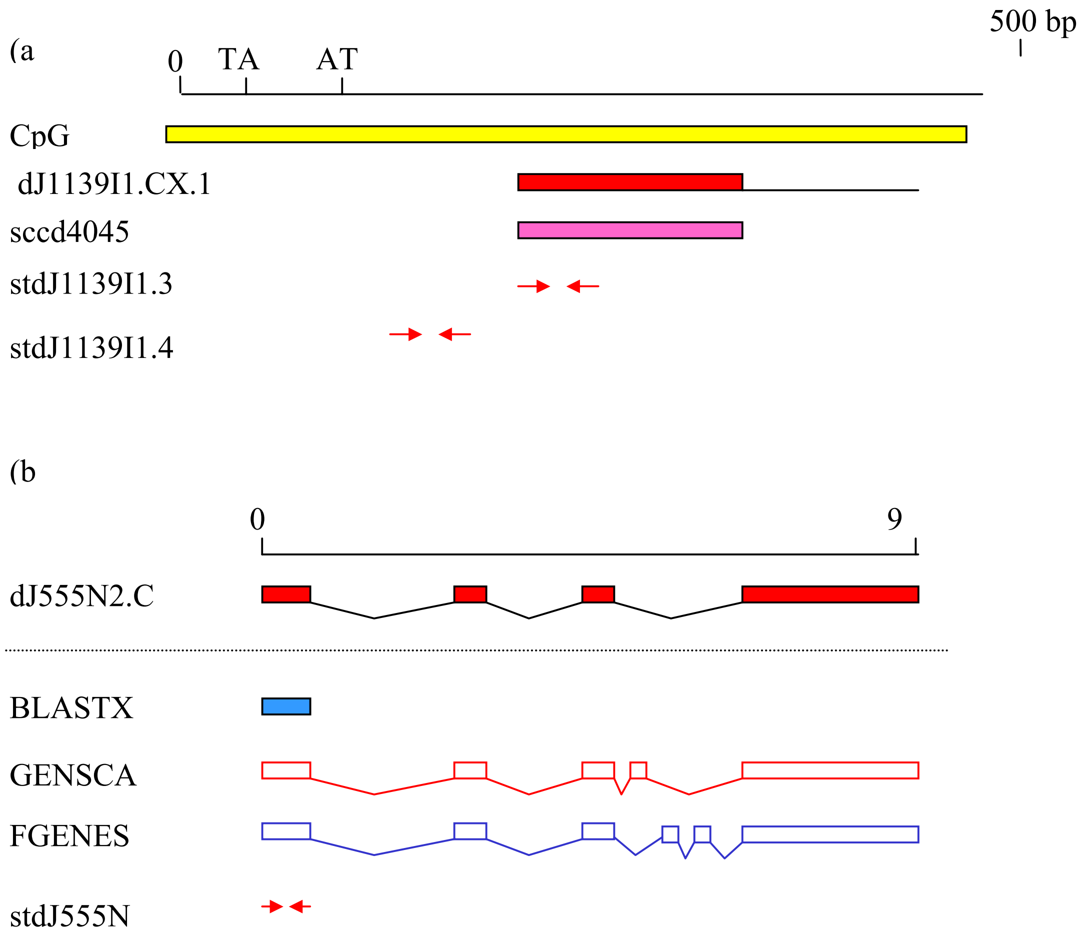


**Figure 4.20:** *Examples of unconfirmed genes (a) The 5' end of dJ1139I1.CX.1 (shown as a red box), the extent of the cDNA sequence confirming the 5' end of the gene is shown as pink boxes. The position of the most likely translation start site (ATG) is shown, a well as the first in-frame stop codon. Primers were designed (shown as red arrows), both within the cDNA sequence and just upstream but no 5' extension of the sequence was observed. (b) dJ555N2.CX.1 (shown as overlapping red boxes linked by black lines) was predicted by both GENSCAN (open red boxes linked by red lines) and FGENESH (open blue boxes linked by blue lines). A BLASTX homology (shown as blue box) was also observed. Primers designed to the first exon (shown as red arrows) brought up no positives when tested against all available cDNA libraries by PCR (data not shown).*

## 4.8 Appendix

**Table 4.5:** *Link information as described in Figure 4.9*

| Link | Accession | Clone Name | Link | Accession | Clone Name |
|---|---|---|---|---|---|
| Link_bA810O3 | AL355812 | RP11-810O3 | Link_dJ170D19 | AC004822 | RP1-170D19 |
| | AL121878 | RP6-204F4 | | AC007025 | RP4-673N16 |
| | AL445164 | RP4-736G20 | Link_dJ29I24 | AC007022 | RP1-29I24 |
| | AL589786 | RP11-161I19 | | AL589677 | RP11-12P4 |
| | AL109751 | RP1-237H22 | Link_bA320L24 | AL441887 | RP11-320L24 |
| | AL135921 | RP4-682C13 | | AC004835 | RP4-555N2 |
| | AL049591 | RP5-878I13 | | AC004973 | RP5-1139I1 |
| | AC003983 | RP1-93I3 | | AC004000 | RP3-404F18 |
| | AL589842 | RP11-268A15 | | AC005190 | RP5-1152D16 |
| | AC005000 | RP1-241P17 | | AC004913 | RP5-876A24 |
| Link_bB377L5 | AL513265 | RP13-377L5 | | AL355348 | RP13-163A20 |
| | AL121879 | RP5-961O8 | | AC005052 | 38K21 |
| | AL356314 | RP5-858F2 | | AC002477 | RP3-327A19 |
| | AL034411 | RP4-808P6 | | AC005023 | GS1-421I3 |
| | Z96810 | RP3-452H17 | | AC006147 | RP4-755D9 |
| Link_dJ2C6 | AL590156 | RP13-420K18 | | AC002086 | RP3-525N14 |
| | AC002071 | RP1-2C6 | | AL512286 | RP11-45J1 |
| | AL590157 | RP13-420K18 | | AC006962 | RP6-52J4 |
| | AC073306 | RP4-564D24 | | AC002476 | RP1-318C15 |
| | AC004823 | RP6-172A13 | | AL451005 | RP11-92B10 |
| | AC005002 | RP3-378P9 | | AC011890 | RP4-655L22 |
| | AC004959 | RP5-1098A23 | | AC008162 | RP1-321E8 |
| | AC006395 | RP3-394H4 | Link_dJ296G17 | AC006144 | RP1-296G17 |
| | AL591505 | RP11-67I12 | | AC002377 | RP1-222H5 |
| | AL031074 | RP1-143G3 | | AL450488 | RP11-161O12 |
| | AC006975 | RP5-1026B21 | | AC007486 | RP5-1015P16 |
| | AC008059 | RP3-409F10 | Link_dJ368G6 | AC007074 | RP3-368G6 |
| | AL445246 | RP11-247H9 | | AC006143 | RP1-74M20 |
| | AC006963 | RP1-278D1 | | AC006314 | RP1-314H24 |
| | AC006968 | RP4-649M7 | Link_bB16D10 | AL357562 | RP13-16D10 |
| | AL590114 | RP11-318H18 | | AL589847 | RP13-477D10 |
| | AC003012 | RP1-169K13 | | AL513487 | RP13-63I15 |
| | AC007088 | RP6-39H21 | Link_bA438H17 | AL359956 | RP11-438H17 |
| | AL391358 | RP11-197B12 | | AL109800 | RP6-64P14 |
| | AL391474 | RP13-115H14 | | Z83848 | RP1-57A13 |
| | AL391830 | RP13-318F20 | | AL035426 | RP3-370N13 |

|  |  |  |  |  |  |
|---|---|---|---|---|---|
|  | AL391803 | RP13-25C19 |  | Z82899 | RP1-181N1 |
|  | AC007021 | RP6-155F9 |  | AL356213 | RP5-1171F9 |
|  | AL391237 | RP13-125M24 |  |  |  |
|  | AL391280 | RP13-128O4 |  |  |  |
|  | AL589824 | RP11-76G11 |  |  |  |
|  | AL606485 | RP11-370H3 |  |  |  |
|  | AC006965 | RP4-562J12 |  |  |  |

**Table 4.6:** *Information of pseudogenes*

| Pseudogene | Description |
|---|---|
| bA320L24.CX.1 | similar to ADP-ribosylation factor-like protein 5 |
| bB192B19.CX.1 | similar to translationally controlled tumor protein |
| dJ29I24.CX.1 | similar to 60S Ribosomal protein |
| dA204F4.CX.2 | similar to zinc finger protein |
| dA204F4.CX.3 | similar to YES-associated protein |
| dJ1189B24.1 | similar to NADH-Ubiquinone Oxidoreductase MLRQ subunit |
| dJ1189B24.2 | similar to Tubulin Beta |
| dJ1189B24.3 | similar to Proto-oncogene Tyrosine-protein Kinase |
| dJ169K13.CX.2 | similar to Tubulin Beta |
| dJ169K13.CX.4 | similar to Ribosomal Protein L12 |
| dJ170D19.CX.3 | similar to Heat shock cognate 70 |
| dJ222H5.CX.4 | similar to mitochondrial heat shock protein 70 (hsp70) |
| dJ237H22.CX.2 | similar to activator of apoptosis Hrk (HRK) |
| dJ241P17.CX.1 | similar to arginosuccinate synthetase |
| dJ241P17.CX.2 | similar to elongation factor-1-gamma |
| dJ321E8.CX.1 | similar to cell cycle protein p38-2G4 homolog (hG4-1) |
| dJ378P9.CX.1 | similar to transcription factor, CA150 |
| dJ525N14.CX.2 | similar to elongation factor Tu family |
| dJ555N2.CX.2 | similar to heterogeneous nuclear ribonucleoprotein |
| dJ562J12.CX.1 | similar to aflatoxin aldehyde reductase AFAR |

**Table 4.7:** *STSs used for mutation screening of MRX23 patients*

| STS name | Primer 1 | Primer 2 | Size (bp) | AT ($^{o}$C) | Gene | Exon |
|---|---|---|---|---|---|---|
| stdJ808P6.4 | GAGCTTCTCTTCATAAATG | GTGGAGCACAAGGAACAG | 298 | 55 | hATBo+ | 1 |
| stdJ808P6.8 | TGTTGCTCTATGGATTTG | ATGCCTTCCTTCAACTCG | 354 | 55 | hATBo+ | 2 |
| stdJ808P6.5 | TAGAAAGCAGTGAACTTTAG | CAGTGAAGGTAGCTATATG | 361 | 55 | hATBo+ | 3 |
| stdJ808P6.6 | TGAGATACAGCTTTTTTATG | TGCTTTCACCAGTGACCTTTG | 366 | 55 | hATBo+ | 4 |
| stdJ808P6.13 | GATGCTGAATGTACATAGC | CACATTGCTGACTATGAGC | 517 | 53 | hATBo+ | 5 |
| stdJ808P6.14 | CTATCTGTGCCCTTCGTATTG | GCTTGATATTGAACTACCATG | 376 | 55 | hATBo+ | 6 |
| stdJ808P6.10 | TAACTTTGGTATATCATCAG | TTGCAAGCTATTACATTATG | 388 | 55 | hATBo+ | 7 |
| std452H17.12 | GTAGAAGGGTGACAATGATG | CTATTGGAGTTTCATAAGTG | 494 | 55 | hATBo+ | 8 |
| stdJ452H17.3 | AGATTTTTCTAATATCTTATG | CTTTCAGAAAGATCATTCTG | 343 | 55 | hATBo+ | 9 |
| stdJ452H17.4 | TGAATTCTGTGATTAACAG | ACCTGGACTTGTCACTAAG | 293 | 55 | hATBo+ | 10 |
| std452H17.13 | GAAACTAAGGAGCATATG | GTTGTGCAGTATATTGTAC | 343 | 53 | hATBo+ | 11 |
| stdJ452H17.6 | ACAGAAAGATAATTGATG | TTGCCTTTTGTCTTCAATG | 276 | 55 | hATBo+ | 12 |
| stdJ452H17.7 | TGATAAATCACATCTGAG | AGGTATAGAAGTAGCCAAG | 401 | 55 | hATBo+ | 13 |
| stdJ452H17.8 | CAGTTCAATATTTGCTTG | ACATGGCTGAGAATTAAGA | 405 | 55 | hATBo+ | 14 |
| stdJ93I3.4 | GTTGATGTGACAGGCTCG | TGAGCTTAACCGAGATGC | 315 | 55 | T-plastin | 1 |
| stbA268A15.1 | AGATGAGAACTTAGCAAG | AGAGAAATAACTTTGAGAC | 221 | 55 | T-plastin | 2 |
| stbA268A15.2 | GTGAGCTTATGAACTGAAC | GATATTCCAGCAGCTAAAAG | 433 | 60 | T-plastin | 3 |
| stbA268A15.3 | GAGTTCAATGCATGTAGC | GCCCGTCCTTGACATTAC | 448 | 60 | T-plastin | 4 |
| stbA268A15.4 | ACCACTGTGTTGCATCCAAG | GATTCATGGACAGACCTAG | 460 | 60 | T-plastin | 5 |
| stdJ241P17.1 | GAGTACATGAAAGAGATG | GAACAGGTCCTCAAACAG | 257 | 55 | T-plastin | 6 |
| stdJ241P17.2 | GAAAGGTCAAGAAGCAAGTG | GCACGAAAGTCTGCATGAC | 276 | 55 | T-plastin | 7 |
| stdJ241P17.3 | GACTGAATGAACTTGGCATG | CTTGGTGATACAGTGTTAGG | 313 | 55 | T-plastin | 8 |
| stdJ241P17.4 | CATGGGACAATAGGATAC | GCCAACTCTACTTCATACG | 262 | 55 | T-plastin | 9 |
| stdJ241P17.5 | GTAGAACTGTATACCCAG | GCCATCTACTTCTTGTAG | 403 | 53 | T-plastin | 10 |
| stdJ241P17.6 | GTGTATTGGCACTATATGC | CATCCATTCATGACATTCG | 201 | 55 | T-plastin | 11 |
| stdJ241P17.7 | CTTGTTTGACAATGTAGTG | CTCTAACAAATATATACAGC | 239 | 55 | T-plastin | 12 |

| | | | | | | |
|---|---|---|---|---|---|---|
| stdJ241P17.8 | CATTTACTCTTGTGCCTTTG | GTGTAGTTATCGACATATC | 255 | 55 | T-plastin | 13 |
| stdJ241P17.9 | CTCATAAAGTAGATGGTGAC | GTAAAGAATTGTGCCATTAG | 291 | 55 | T-plastin | 14 |
| stdJ241P17.10 | GGCTTCTTTGTGAGTGAG | CTATTAGCAGTCTCCCTTAC | 292 | 55 | T-plastin | 15 |
| stdJ241P17.11 | GTGTCCTTAACTGACAAG | GCCAAGAGTTCCTTAAGC | 282 | 55 | T-plastin | 16 |
| stdJ241P17.12 | CTTCTGCAGCTCCTGGTG | CGGTAGTAGTCAGGATGTG | 488 | 55 | SMT3B | 1 |

**Table 4.8:** *Information on cDNA sequencing projects*

| cDNA collection | Web Address |
|---|---|
| Mammalian Gene Collection (NIH) (MGC) | http://mgc.ncbi.nih.gov/ |
| NEDO database (Kazusa DNA Research Institute) (FLJ) | http://www.kazusa.or.jp/NEDO/ |
| HUGE database (Kazusa DNA Research Institute) (KIAA) | http://www.kazusa.or.jp/huge/ |
| The German Human cDNA project (DKFZ) | http://mips2.gsf.de/proj/cDNA/ |
| The Riken Mouse collection | http://genome.rtc.riken.go.jp/ |
| Genoscope | http://www.genoscope.cns.fr/ |
| EMBL (dBEST, vert_RNA) | http://www.ebi.ac.uk/embl/ |