Chapter 3

Gene annotation and analysis of human Xp11.22-p11.3

3.1 Introduction

Much of the genetic information contained with the genome sequence can be deciphered with the use of predictive and evidence based computational programmes. Detailed analysis and annotation of genome sequences empowers scientists to decipher much of their genetic information and maximise their utility. This chapter aims to demonstrate the utility of the human genome sequence in human gene discovery.

Several analytical processes can be employed in human gene discovery, the merits of which are discussed in section 1.4. Work presented in this chapter, used a variety of computational analyses to define the location of gene features including exons, transcription start sites and transcription termination sites. With the appropriate evidence these features can be annotated and built into gene structures.

However, computational analysis using either *ab initio* prediction programmes or sequence similarity searches frequently fails to identify complete gene structures and often requires additional evidence to complete them. In particular, the 5' and 3' ends of genes are frequently incomplete or are absent from gene structures derived from transcript sequences. Gene structures predicted by *ab initio* analysis require experimental evidence to confirm their transcription. Experimental techniques such as targeted cDNA screening and sequencing can be employed to obtain such evidence after which the resulting sequences can be overlaid onto previously annotated gene structures. This not only creates a more comprehensive gene set but it also increases its value for future investigations. In this chapter both experimental and computational techniques are employed to describe the genetic content of human Xp11.23-p11.3.

3.1.1 Xp11.22-p11.3

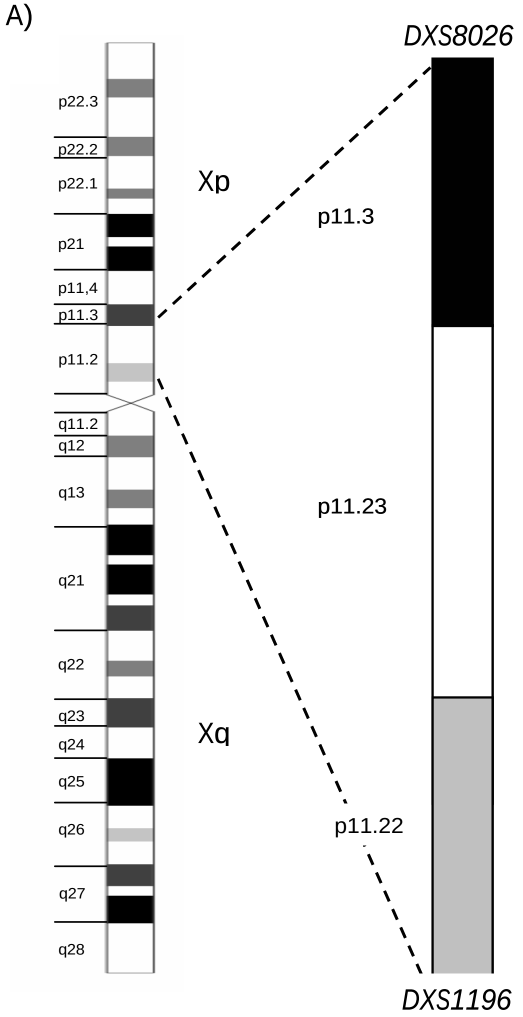
This chapter focuses on approximately 7.3 Mb of genomic sequence in human Xp11.22-p11.3. The region is encompassed by the markers *DXS8026* and *DXS1196* and was mapped and sequenced as part of the HGP by the Wellcome Trust Sanger Institute (http://www.sanger.ac.uk/HGP/ChrX) and the Institute for Molecular Biotechnology (Jena, Germany). The physical map of Xp11.23-p11.3 used in this study superseded several high-resolution previously published physical maps of the same region (Coleman *et al.,* 1994; Knight *et al.,* 1994). Problems were, however, encountered establishing a consensus marker order from these maps - a problem that was later attributed to clonal instability.

The mapping strategy adopted by the HGP to delineate the clone order of Xp11.22-p11.3 used landmark-based mapping and restriction fingerprinting techniques to generate bacterial clone contigs (Bentley *et al., 2001*). These were then positioned onto physical and genetic maps of the X chromosome. In May 2002, the physical map that was generated by the HGP had placed human Xp11.22-p11.3 in one contig. This contig was composed of a variety of bacterial clones including bacterial artificial chromosomes (BACs), P1 artificial chromosomes (PACs), fosmids and a yeast artificial chromosome (YAC). The finished clones representing the minimal tiling path for this region are displayed in Figure 3.1. The clones were sequenced (Ross *et al.,* 2005) which provided the information for the transcript analysis in this chapter.

The aim of this chapter was to analyse and annotate the genomic sequence which spans the region of Xp11.22-p11.3. Creation of a transcript map would provide a framework for more detailed research in this region. This region was of particular interest because was proposed to contain several clinically important genes. Thirty-two genes have entries in the OMIM database including Wiskott-Aldrich syndrome (*WAS*), GATA-binding protein 1 (*GATA1*) and retinitis pigmentosa (*RP2*). For example, the X-linked recessive Wiskott-Aldrich syndrome, is caused by mutations in the gene, *WAS*. Clinically, Wiskott-Aldrich syndrome is an immunodeficiency disease characterized by thrombocytopenia, eczema, and recurrent infections (Lemahieu *et al.,* 1999). Furthermore, the *WAS* protein may provide a link between the actin-cytoskeleton and Cdc42. It may function as a signal transduction adaptor downstream of Cdc42, and that, in affected males, the cytoskeletal abnormalities in males affected with Wiskott-Aldrich syndrome may result from a defect in Cdc42 signaling (Kolluri *et al.,* 1996).

In addition, several neurogenetic disorders appear to be localised to this region. Some of these include Graves disease (Zinn *et al.,* 1998; Imrie *et al.,* 2001), optic atrophy (Assink *et al.,* 1997) as well as several X linked mental retardation subtypes (Chiurazzi *et al.,* 2001). Moreover, human Xp11.22-p11.3 also harbours the hypothesised fusion point between the ancient X chromosome and an autosome (see Section 1.1). Accurate definition of this region may also give rise to further studies on the evolutionary origins of the individual genes.

A first-pass annotation of the *DXS8083-ELK1* interval on Xp11.23-Xp11.3 has been described (Thiselton *et al.,* 2002). This work identified 28 expressed, and 37 putative transcripts using NIX analysis (http://www.hgmp.mrc.ac.uk/NIX/), and described the mapping of transcript sequence clusters to the genomic sequence but the expression of these candidate genes was not experimentally verified. The genomic region analysed by Thiselton and colleagues (Thiselton *et al.,* 2002) partially covers the region analysed in this chapter.



B)

#	EMBL accession	Internal identifier		#	EMBL accession	Internal identifier		#	EMBL accession	Internal identifier
1	AL591491.5	bA435K1		26	AL590283.7	bB83I24		50	BX546463.3	yM787F5
2	AL590364.5	bA342D14		27	Z84466.2	dJ230G1			GAP	
3	AL020989.2	dJ192P9		28	AL009172.1	dJ212G6		51	AL591367.5	bA805H4
4	AL139810.16	bA576G22		29	AL031967.1	dJ473A14		52	AL357894.6	bA576P23
5	AL157701.2	dJ628F15		30	AL022578.1	dJ393P12		53	AL121865.7	bA119E20
6	AL391134.13	bA245M24		31	AL591394.11	bA1145B22		54	AL445491.18	bA88O18
7	AL034412.2	dA99M1		32	Z98304.1	dJ54B20		55	AL359272.9	bA554P16
8	AL031584.2	dJ1158E12		33	AL356464.15	bA38O23	H	56	AL359914.14	bA637B23
9	AL031585.2	dA227L5		34	AL683817.11	bA552E4	H	57	AL954833.5	bB346H10
10	AL139402.11	dJ808F24		35	AL606490.14	bA344N17		58	AL391379.12	bB171J5
11	AL034410.9	dJ774G10		36	AF196972	AF196972	H	59	AL158055.12	bA348F1
12	AL390917.12	bA255P16		37	LL0XNC01-30I4	AF196969	H	60	AL390060.14	bA56H2
13	AL034402.9	dJ30G7		38	RP11-1148L6	AC115618	H	61	AL662886.6	bB339K12
14	AL139811.30	bA75A9		39	AF196970	AF196970	H	62	AL929101.4	bA22B10
15	AL031393.1	dJ733D15		40	XX-A0424	AC115617	H	63	AL929410.7	bA234P3
16	AL022165.1	dJ71L16		41	AF196971	AF196971	H	64	BX322790.2	bA114H20
17	AL162583.9	dJ250J21		42	LL0XNC01-36H8	AF207550	H		GAP	
18	AL512633.6	bA198M15		43	BX530088.6	BX530088	H	65	AL591212.11	bA236P24
19	AL050307.13	dJ689N3		44	AF196779	AF196779	H	66	AL139396.18	bA258C19
20	AL627143.13	bA14O9	H	45	LL0XNC01-7P3	AF235097	H	67	AL590123.9	bA473J14
21	Z83822.1	dJ306D1	H		GAP			68	AL161779.32	dA29D12
22	AL513366.11	bK2522E6		46	LL0XNC01-30L17	AF235098	H	69	Z97054.1	dJ339A18
23	AL096791.13	dJ659F15		47	AF238380	AF238380	H	70	AL592046.7	bA155O24
24	AL591503.9	bA571E6		48	AF222686	AF222686	H	71	Z94044.2	dJ154P24
25	AL590223.12	bB479F17		49	AL663118.2	bB377G1	H			

Figure 3.1 The human X chromosome and clones mapped to human Xp11.22-p11.3

A) Ideogram of the human X chromosome. The region analysed in this chapter is also displayed.

B) EMBL accession numbers and Sanger Institute identifiers for clones analysed in this study. GAP = a gap in the physical map at the time of analysis. H – clones annotated by the Human and Vertebrate analysis team (HAVANA).

Results

3.2 Sequence analysis

The sequence composition of individual clones was analysed using a standard automated process described in section 2.26. This analysis was performed by the Informatics Team at the Sanger Institute and the programmes employed in this process are described in section 2.27. In total, 71 clones were analysed. The analysis results were visualised in Xace, a chromosome specific application of ACeDB (section 2.25.1, Figure 3.2).

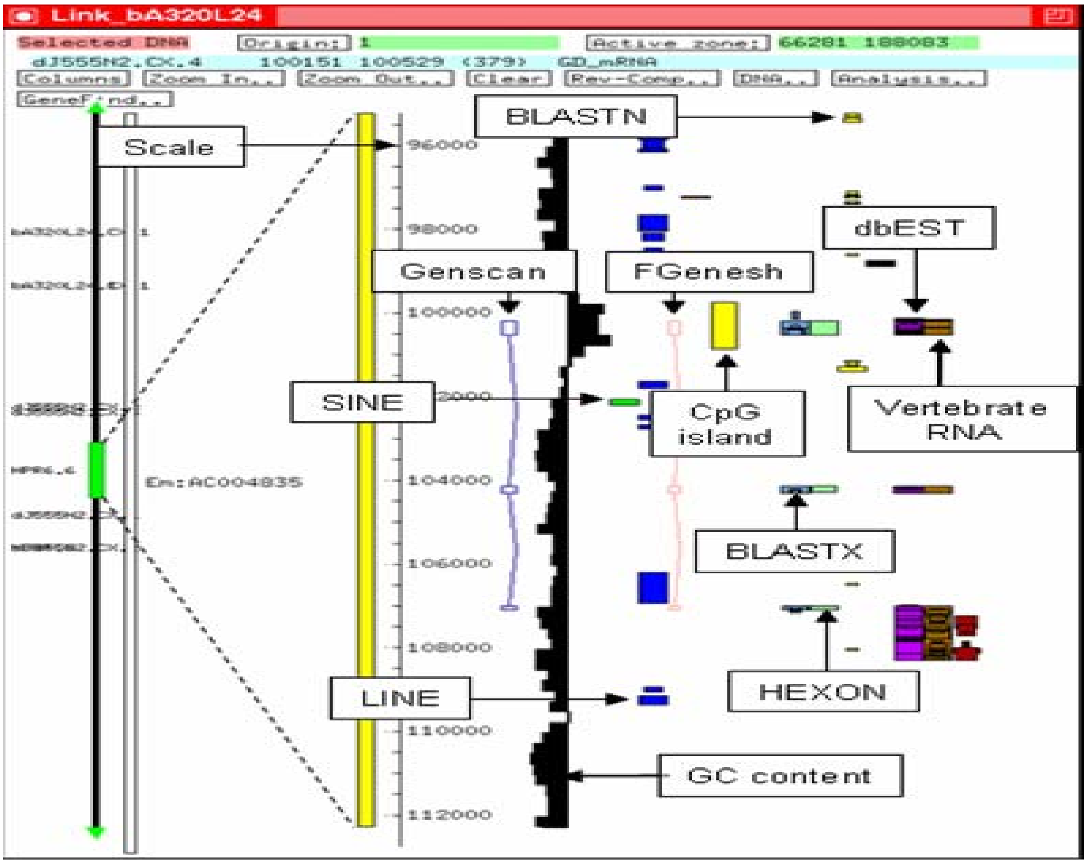


Figure 3.2 Visualisation of sequence data in ACeDB

Each analysis programme utilised in the analysis process is displayed in a separate track. The strength of identified sequence similarity matches are also displayed in the width of the corresponding band (a wider band equates to a stronger match). The columns labelled with vertebrate RNA, dbEST, BLASTN and BLASTX display the results from sequence similarity searches for cDNA, EST, genomic and protein sequences respectively. The columns labelled HEXON, FGenesh and Genscan display the results from *de novo* exon (HEXON) and gene (FGenesh and Genscan) prediction programmes. The genome landscape is described by displaying the GC content, a scale, predicted CpG islands and the retroposed LINE and SINE repeat elements.

3.2.1 Repeat analysis

The repeat content of human Xp11.22-p11.3 was assessed using RepeatMasker (Smit, Hubley and Green, 1990, http://repeatmasker.org), and was compared to the repeat content of the entire X chromosome and the genome average (Table 3.1). Compared to the genome average, the human X chromosome is enriched in LINE repeats, but has a slightly lower proportion of SINE repeats (Ross *et al.,* 2005). Human Xp11.22-p11.3 has a high repeat content. Compared to the genome average, this region is enriched in SINE, LINE and LTR repeats. Analysis of Xp11.22-p11.3 revealed this region to be very similar to the overall repeat content of the entire X chromosome. However, the distribution of repeat families in Xp11.22-p11.3 differs from that of the whole X chromosome. There is a decrease in the overall LINE repeat density (27.10% for Xp11.22-p11.3 versus to 32.18% for the entire X chromosome) and an increase in SINE repeat density (17.46% versus 10.3%). More specifically, the SINE increase is accounted for by a rise in the abundance of *Alu* repeats.

Table 3.1 Repeat content of human Xp11.23

Displayed in table are the percentages of LINE (L1, L2 and L3), SINE (*Alu*s and MIRs), LTR elements, and DNA elements.

Repeat Family	Xp11.22-Xp11.3	X chromosome*	Genome average*
LINE	27.10	32.18	20.42
SINE	17.46	10.30	13.14
LTR	10.04	10.45	8.29
DNA elements	2.25	2.67	2.84
Total	56.86	55.61	44.69

* Figures taken from Ross et al, 2005

3.2.2 G+C content

The G+C content of Xp11.22-p11.3 was determined using the programme geecee at Emboss (section 2.25.6) and is 42%. This figure is closer to the genome average of 41%, than to the average G+C content of the whole X chromosome (39%).

Xp11.22-p11.3 has high G+C levels and high *Alu* levels which are in keeping with the high gene density in the region.

3.3 Annotation of transcripts mapped to human Xp11.22-p11.3.

The genomic sequence of 56 clones was analysed as a part of this thesis (79%, of clones included in the analysis) for the presence of both known genes and potential gene features requiring more in-depth investigation and experimental analysis. The remaining 15 clones were analysed by the human and vertebrate analysis (HAVANA), team at the Sanger Institute. These are appropriately noted in Figure 3.1.

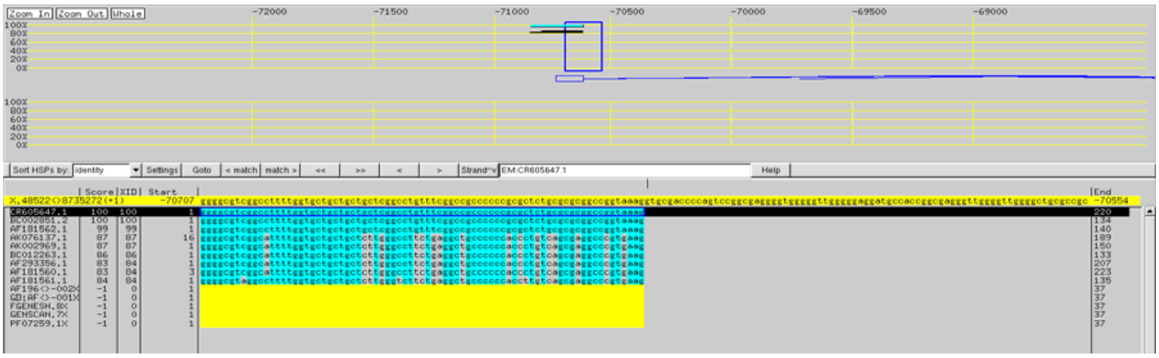
Genes identified in the annotation process were grouped according to the evidence that was used to identify them and their state of completion. These groupings are:

1. **Known** – curated genes that are identical to known human cDNAs or protein sequences. They have an entry in Entrez Gene or LocusLink at the NCBI.
2. **Novel coding sequence (CDS)** – has an open reading frame (ORF). Identical or similar to human cDNAs or proteins from other species.
3. **Novel transcript** – as novel CDS but with no clear ORF.
4. **Putative** – identical or homologous to human spliced ESTs but do not contain an ORF.
5. **Pseudogene** – sequence similar to a known spliced mRNA, EST or protein but contains a frameshift and/or stop codon(s) which disrupts the ORF. This class of gene was further divided into 2 classes; processed and non-processed pseudogenes. Processed pseudogenes typically have:

- A single exon structure
- A poly-adenylation tail in the genomic sequence
- Flanking repeat sequences (LINE repeats).

Non-processed pseudogenes have an exon/intron structure that is similar to their functional counterpart. These genes have a disrupted ORF.

Gene structures were manually annotated onto the genomic sequence using the Xace computational interface. Alignments to mRNA and protein sequences were visualised using Blixem (section 2.26.2), a BLAST result visualisation tool in Xace to confirm sequence identity and splice site fidelity (Figure 3.3). Annotated gene structures without an official gene name from the HUGO gene nomenclature committee (HGNC) were labelled with their clone name followed by sequential numbers (e.g. RP11-339A18.4 represents the fourth gene annotated in the clone RP11-339A18).



	Score	%ID	Start	End
X,48522<>8735272(+1)			-70707	-70554
CR605647.1	100	100	1	220
BC002051.2	100	100	1	134
AF181562.1	99	99	1	140
AK076137.1	87	87	16	189
AK002969.1	87	87	1	150
BC012263.1	86	86	1	133
AF293356.1	83	84	1	207
AF181560.1	83	84	3	223
AF181561.1	84	84	1	135
AF196<>-002X	-1	0	1	37
GD:AF<>-001X	-1	0	1	37
FGENESH.8X	-1	0	1	37
GENSCAN.7X	-1	0	1	37
PF07259.1X	-1	0	1	37

Figure 3.3 Visualisation of sequence alignments using BLIXEM

The diagram illustrates BLASTN matches between mRNA accession CR605647 and genomic sequence AF196971. The blue box in the top section represents the position of the alignment that is highlighted in the lower section and a predicted spliced sequence is displayed. The location of where the matches between the RNAs and genomic sequence end are displayed with yellow boxes. The genome sequence is also displayed in yellow. All homologous transcribed sequences identified by sequence similarity searches are also listed.

3.3.1 Annotation of known Genes

Seventy-seven known gene sequences were annotated onto the genomic sequence of human Xp11.22-p11.3 using this approach. Where appropriate, the HUGO gene names and the conserved CDS (CCDS) numbers for these genes are listed in Table 3.3. An example of an annotated known gene structure is displayed in Figure 3.4.

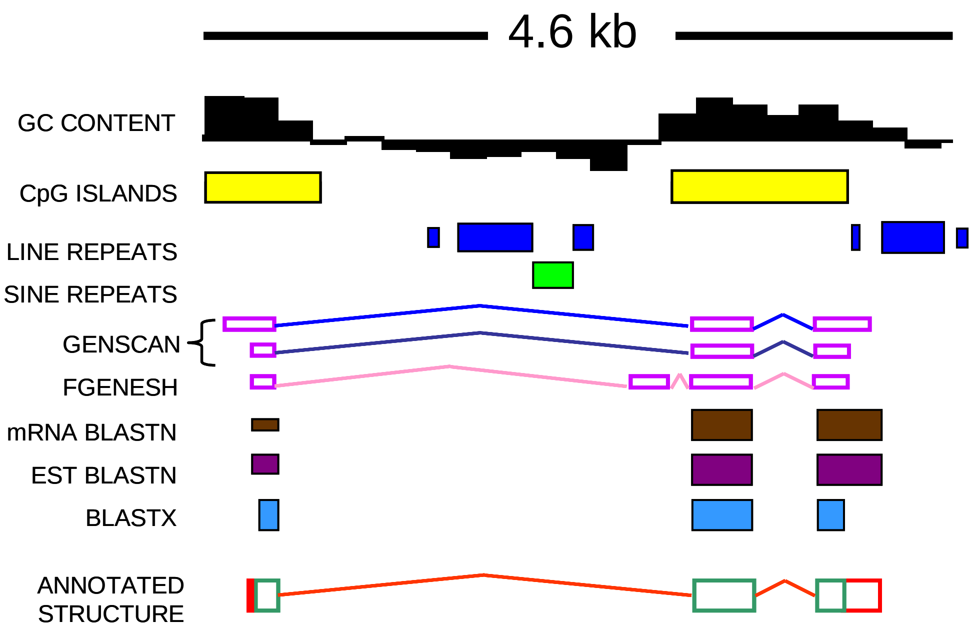


Figure 3.4 Example of a known gene structure, ***PCSK1N*****.**
In this case, the gene was annotated using mRNA accession AF181562 which aligned to the clone AF207550. The diagram shows an ACeDB representation of the gene structure. Displayed from top to bottom are the descriptive analysis of the genome landscape; G+C content (increasing upward thickness of bars represents increased %GC relative to a midrange value of 50% adjacent sequence), CpG Islands (yellow) LINE repeats (blue) and SINE repeats (green). GENSCAN predictions and FGENESH gene predictions are displayed. This is followed sequence homology matches: mRNA BLASTN matches (brown), EST BLASTN matches (purple) and protein BLASTX matches (pale blue). Finally the annotated gene structure of PCSK1N is displayed. Exons are depicted as outlined boxes, with introns represented as coloured lines connecting the exons.

In most cases, the entirety of the mRNA sequence (excluding the polyA tail) matched to the X chromosome sequence. However, the transcripts for the genes *GATA1, CCBN3* and *GRIPAP1* could not be mapped completely to the genome sequence. Subsequent BLAST analysis identified the remainder of *GATA1* in a sequenced clone that had not been analysed (AC115618). It was not possible to identify any genome sequence that contained either a 231 bp internal fragment of *CCBN3* or the first 196 bp of *GRIPAP1*. The gap containing *CCBN3* has not been closed and is now classed as a type 4 gap (a gap in the physical map of the human X chromosome that cannot be closed with existing technologies and resources). A gap was created within the clone AF207550 that was hypothesised to contain the 5' end of *GRIPAP1* and additional mapping was undertaken to close it. The genome sequence that flanked this gap was aligned to end sequences from the fosmid library WIBR2, in the UCSC genome browser (http://www.genome.ucsc.edu/). It was hypothesised that apparently short or orphan fosmids (i.e. only one end matching the genome sequence) would contain this missing sequence. Eight short/orphan fosmids were identified and were screened using markers designed against the missing mRNA sequence (primer pairs 487051) and the first annotated exon of *GRIPAP1* (primer pair 487049, Table 3.2, Figure 3.5). Three fosmids were identified that contained the 5' end of *GRIPAP1.* The fosmid G248P89409A6 was selected for sequencing as both ends of the clone matched to the genome sequence. Subsequent analysis of its sequence (ACC No: BX530088) confirmed that it contained the missing 196 bp section of *GRIPAP1*.

Table 3.2 "Orphan" or "short" fosmids that could harbour the 5' end of *GRIPAP1*.

"Orphan" fosmids have an end sequence that does not match the genome sequence. "Short" fosmids have both ends within the sequence but the separation of the ends is unfeasibly short.

Number	Clone	Orphan/size?	PCR screen	
			487049	487051
1	G248P89563G8	Orphan	+	+
2	G248P80210A10	Orphan	-	-
3	G248P82559C11	Orphan	-	-
4	G248P88846B10	24407	-	+
5	G248P89409A6	22928	+	+
6	G248P8389F4	Orphan	-	+
7	G248P87705G10*	24407	+	+
8	G248P88738H5*	Orphan	**+**	**+**

*results not shown

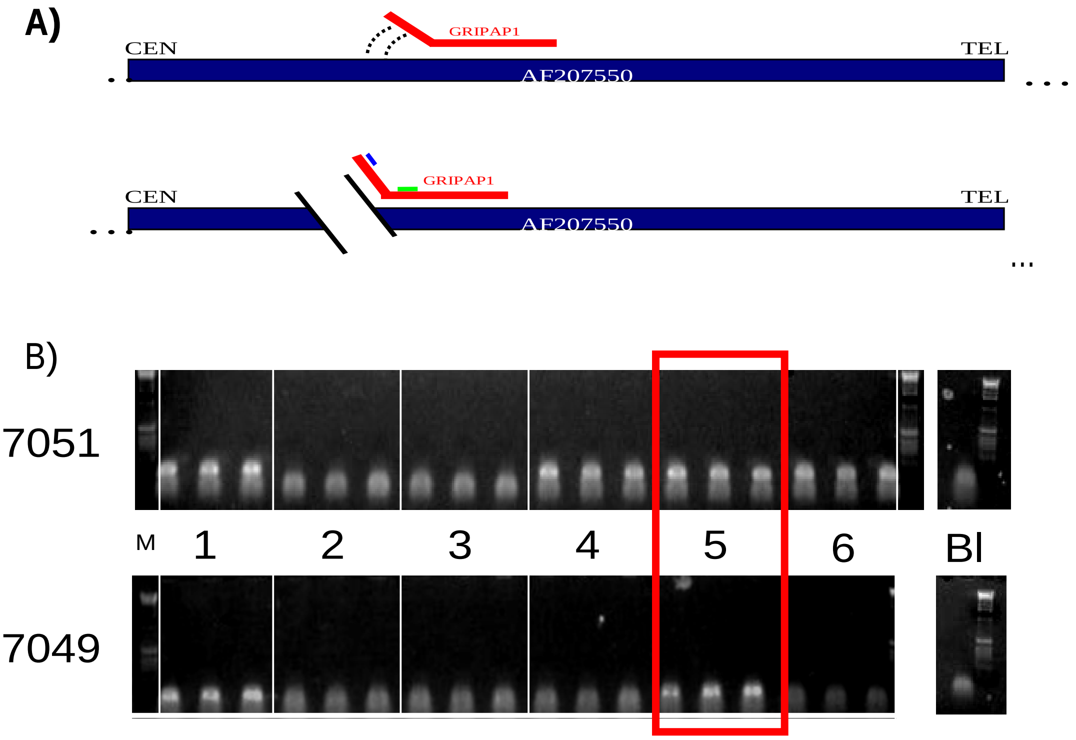


Figure 3.5 Screening of fosmids for 5' end of GRIPAP1.#

A) Cartoon displaying how the first 196 bp of *GRIPAP1* could not be identified in the clone AF207550. The location of primers that were designed to *GRIPAP1* to screen the short/orphan fosmids shown by blue and green bars.

B) Three individual colonies were screened for each fosmid numbered 1-6 (listed in Table 3.2) by PCR. Negative controls (Bl) and markers (M) are shown. PCRs were performed using primers 487051 (green) and 487049 (blue) that amplify the known and missing mRNA fragments of *GRIPAP1*. The 3 colonies from the fosmid G248P89409A6 are highlighted in a red box. This fosmid was subsequently submitted for sequencing.

Table 3.3 Known genes annotated in human Xp11.22- p11.3

CCDS –a database that contains of a core set of human protein coding regions that are consistently annotated and of high quality (http://www.ncbi.nlm.nih.gov/CCDS/).

	CCDS	HUGO Identifier (Alias)	Function	Reference
1	ccds14266.1	CxORF36	Unknown	Ota *et al.*, 2004
2	ccds14267.1	None (FLJ20344)	Unknown	Ota *et al.*, 2004
3	ccds14268.1	*CHST7*	Generates sulfated glycosaminoglycan (GAG) moities during chondroitin sulphate biosynthesis	Kitagawa *et al.*, 2000
4	ccds14269.1	*SLC9A7*	SLC9A7 plays a role in maintaining the cation homeostasis and function of the trans-golgi network	Numata and Orlowski 2001
5	ccds14270.1	*RP2*	The RP2 locus has been implicated as one cause of X-linked retinitis pigmentosa	Schwahn *et al.*, 1998
6	ccds14271.1	*PHF16*	Unknown	Nagase *et al.*, 1995
7	ccds14272.1	*RGN*	Plays an important role in calcium homeostasis	Fujita *et al.*, 1995
8	ccds14274.1	*RBM10*	Has significant homology with several RNA-binding proteins	Nagase *et al.*, 1995
9	ccds14275.1	*UBE1*	Catalyzes the first step in ubiquitin conjugation	Handley *et al.*, 1991
10		*INE1*	Contains an open reading frame encoding a 51-amino acid polypeptide with a single zinc finger domain	Esposito *et al.*, 1997
11		*NUDFB11* (Ba2522E6.CX.2)	Neuronal protein 17.3. Function unknown	Ota *et al.*, 2004
12	ccds14276.1	*PCTK1*	May play a role in signal transduction cascade in terminally differentiated cells	Meyerson *et al.*, 1992
13	ccds14277.1	*USP11*	May have a role in regulating the eukaryotic cell cycle	Swanson *et al.*, 1996

14	ccds14278.1	*ZNF157*	Also contains KRAB A and KRAB B boxes which may be involved in signal transduction	Derry *et al.*, 1995
15	ccds14279.1	*ZNF41*	Contains 18 contiguous zinc fingers and a KRAB/FPB domain	Franze *et al.*, 1991
16		*ARAF1*	Encodes a cytoplasmic protein serine/threonine kinase that plays a critical role in cell growth and development	Mark *et al.*, 1986
17	ccds14280.1	*SYN1*	Neuronal protein - regulation of axonogenesis and synaptogenesis	Sudhof and Rizo 1996
18	ccds14281.1	*TIMP1*	Multi-functional; is able to inhibit collagenase in addition to having erythroid-potentiating activity	Gossen and Bujard 1992
19	ccds14282.1	*PFC*	Has a role in complement-mediated clearance	Nolan *et al.*, 1992
20	ccds14283.1	*ELK1*	Involved in the ras signalling cascade	Rao *et al.*, 1989
21	ccds14284.1	*UXT*	Abundantly expressed in tumours, and is likely to be involved in tumorigenesis	Schroer *et al.*, 1999
22		*ZNF81*	Hypothetical zinc finger protein	Marino *et al.*, 1993
23	ccds14287.1	*SSX6*	Member of *SSX* family, CT antigen expressed in normal testis and in cancer cells	Gure *et al.*, 2002
24	ccds14288.1	*SSX5*	Member of *SSX* family	Gure *et al.*, 2002
25	ccds14289.1	*SSX1*	Member of *SSX* family	Gure *et al.*, 2002
26		*SSX9*	Member of *SSX* family	Gure *et al.*, 2002
27	ccds14291.1	*SSX3*	Member of *SSX* family	Gure *et al.*, 2002
28	ccds14292.1	*SSX4*	Member of *SSX* family	Gure *et al.*, 2002
29	ccds14293.1	*SLC38A5*	Transmembrane amino acid transporter protein	Nakanishi *et al.*, 2001
30	ccds14294.1	*FTSJ*	Member of FtsJ cell division family	Pintard *et al.*, 2000
31	ccds14296.1	*PORCN*	Member of MBOAT (Membrane o-acyl transferases) family	Caricasole *et al.*, 2002
32	ccds14300.1	*EBP*	ER membrane protein that is involved in the formation of cholesterol. High affinity binding protein for anti-ischemic phenylalkylamine	Hanner *et al.*, 1995

33		*OATL1*	Similar to ornithine-delta-aminotransferase	Geraghty *et al.*, 1993
34	ccds14301.1	*RBM3*	Has significant homology to several RNA-binding proteins	Ye *et al.*, 2001
35	ccds14302.1	*WDR13*	Member of WD repeat protein family. Function is unknown, but it may mediate protein interactions	Singh *et al.*, 2003
36	ccds14303.1	*WAS*	Wiskott-Aldrich syndrome family member. Involved in the transduction of signals from cell surface receptors to the actin cytoskeleton	Kwan *et al.*, 1988
37	ccds14304.1	*SUV39H1*	A heterochromatic protein that transiently accumulates at centromeric position during mitosis	Aagaard *et al.*, 1999
38	ccds14305.1	*GATA1*	Member of *GATA* family of transcription factors; involved in regulation of the switch from foetal to adult haemoglobin	Gumucio *et al.*, 1991
39	ccds14306.1	*HDAC6*	Histone deacetylase has activity and represses transcription	Nagase *et al.*, 1998
40	ccds14307.1	*PCSK1N*	Acts to process latent precursor proteins into their biologically active proteins. Endogenous inhibitor of the proprotein convertase subtilisin/kerin type 1	Fricker *et al.*, 2000
41	ccds14308.1	*TIMM17B*	Mitochondrial inner membrane translocase subunit; translocates nuclear encoded proteins into the mitochrondrion	Bauer *et al.*, 1999
42	ccds14309.1 ccds14310.1	*PQBP1*	Activates transcription and binds to polyglutamine tracts	Komuro *et al.*, 1999
43	ccds14311.1	*SLC35A2*	UDP-galactose translocator 2; transports nucleotide sugars	Ishida *et al.*, 1996
44	ccds14312.1	*PIM2*	Serine/threonine kinase, may have a role in proliferating cells as well as during mitosis	Baytel *et al.*, 1998
45	ccds14313.1	NONE (DKFZp761A052, AF207550.5)	Unknown	Strausberg *et al.*, 2002
46	ccds14314.1	*KCND1*	May act as an A-type voltage gated potassium channel	Isbrandt *et al.*, 2000
47		*GRIPAP1*	A neuron-specific guanine nucleotide exchange factor for the ras family of small G proteins (RasGEF) and is associated with the GRIP/AMPA receptor complex in brain	Ye *et al.*, 2001

48	ccds14315.1	*TFE3*	A member of the helix-loop-helix family of transcription factors and binds to the mu-E3 motif of the immunoglobulin heavy-chain enhancer	Macchi *et al.*, 1995
49	ccds14316.1	NONE (JM11)	Unknown	Strausberg *et al.*, 2002
50	ccds14317.1	*PRAF2*	Unknown	Strausberg *et al.*, 2002
51	ccds14318.1	*WDR45*	Unknown	Strausberg *et al.*, 2002
52		*GPKOW*	Unknown	Strausberg *et al.*, 2002
53		None (AF196779.6, FLJ21687)	Unknown	Strausberg *et al.*, 2002
54	ccds14319.1	*PLP2*	Proteolipid 2 protein which may multimerise to from an ion channel	Oliva *et al.*, 1993
55	ccds14320.1	*LMO6*	Contains three LIM domains which are cysteine rich motifs that bind zinc atoms to form a specific protein-binding interface for protein-protein interactions	Fisher *et al.*, 1997
56	ccds14321.1	*SYP*	Membrane protein of small synaptic vesicles in brain and endocrine cells	Sudhof *et al.*, 1987
57		*CACNA1F*	Role in X-linked congenital stationary night blindness	Fisher *et al.*, 1997
58	ccds14322.1	CXorf37 (AF235097.3)	Unknown	Strausberg *et al.*, 2002
59	ccds14323.1	*FOXP3*	A member of the forkhead/winged-helix family of transcriptional regulators	Brunkow *et al.*, 2001
60		*PPP1R3F*	Protein phosphatase 1, regulator (inhibitor) subunit 3F	Ceulemans *et al.*, 2002
61		None (AF235097.3)	As for GAGE1 (below)	Strausburg *et al.*, 2002
62	ccds14325.1	*GAGE1*	Member of GAGE family	Van den Eynde *et al.*, 1995
63	ccds14327.1	*PAGE1* (*GAGEB1*)	Member of GAGE family	Chen *et al.*, 1998
64		*PAGE4*	Member of GAGE family	Brinkmann *et al.*, 1998

		(*GAGEC1*)		
65	ccds14328.1	*CLCN5*	Chloride channel 5. Mutation results in renal tubular disorders complicated by nephrolithiasis	Fisher *et al.*, 1995
66	ccds14329.1 ccds14330.1	*AKAP4*	The encoded protein is localized to the sperm flagellum and may be involved in the regulation of sperm motility	Mohapatra *et al.*, 1998
67	ccds14331.1	*CCNB3*	Cyclin B3. Cyclins function as regulators of CDK kinase	Lozano *et al.*, 2002
68		NONE (KIAA1202, bA119E20.1)	Unknown	Nagase *et al.*, 1998
69	ccds14334.1	*BMP15*	Member of bone morphogenetic protein family. May be involved in oocyte maturation and follicular development	Dube *et al.*, 1998
70		*NUDT11*	Unknown	Hidaka *et al.*, 2002
71		*IQSEC2*	Unknown	Nagase *et al.*, 1998
72	ccds14336.1	*GSPT2*	A GTP binding protein that plays a role at the G1-S phase transcription of the cell cycle	Hoshino *et al.*, 1998
73	ccds14337.1	*MAGED1*	Member of the melanoma antigen (MAGE) family	Pold *et al.*, 1999
74	ccds14352.1	*SMC1L1*	Putative chromosome segregation protein has NTP binding site; coiled coil region	Rocques *et al.*, 1995
75	ccds14353.1	*RIBC1*	Unknown	Strausberg *et al.*, 2002
76	ccds14315.1	*HADH2*	Neurotoxic peptide that has been implicated in the pathogenesis of Alzheimer's disease	Yan *et al.*, 1997
77		NONE RP11-339A18.4 (dJ339A18.CX.6)	Contains HECT domain which is associated with ubiquitin protein-ligase activity	Gu *et al.*, 1995

3.3.2 Annotation of novel transcripts in Xp11.22-p11.3

Novel CDS and novel transcript genes were identified by aligning homologous protein and cDNA sequences to the genome sequence. Eight genes had a clear ORF and were classified as novel CDS genes. Eleven cDNA transcripts aligned to the genome sequence but did not contain a definitive open reading frame and were classified as novel transcript genes. The location of novel CDS and novel transcript sequences annotated in human Xp11.22-p11.3 are displayed in Figure 3.8. These genes are listed in Table 3.5. Further experimental analysis was completed on some of these loci to enhance their annotation (see section 3.4).

An additional five putative genes were identified from human spliced EST sequences. These genes were also targeted for further experimental verification. These genes are also listed in Table 3.5 and their locations displayed in Figure 3.8.

3.3.3 Annotation of pseudogenes

The genomic sequence was also scanned for the presence of both processed and non-processed pseudogenes. Loci that satisfied a combination of the following criteria were classed as processed pseudogenes: (i) high sequence similarity (> 80%) at the nucleotide level with the paralogous gene; (ii) very highly similar ESTs (< 95%); (iii) loss of introns when compared to the functional gene; (iv) loss of the ability to express mature proteins as detected by the presence of premature stop codons in the open reading frame, insertion or deletions resulting in a frameshift or loss of methionine start codon; or (vi) presence an imperfect poly A tract located in the genomic sequence at the 3' end of the gene. Loci that satisfied a combination of the following criteria were classed as nonprocessed pseudogenes: (i) high sequence similarity (> 80%) at the nucleotide level with the paralogous gene; (ii) very highly similar ESTs (< 95%); (iv) loss of the ability to express mature proteins as detected by the present of premature stop codons in the ORF, insertion or deletion reulsting in a frameshift or loss of methionine start codon. An example of an annotated processed pseudogene is displayed in Figure 3.6.

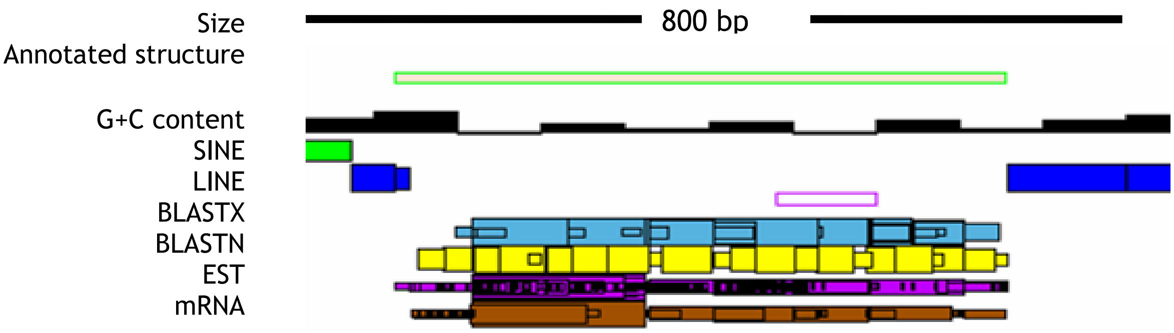


Figure 3.6 Diagram illustrating a "pseudogene" (pseudogene structure), for the locus bA198M15.1.

The diagram shows an ACeDB representation of the gene structure. Key: mRNA/protein homologies as in Figure 3.4 above. In this case, an intronless BLASTX match to the spliced gene chromatin modifying protein 5, *CHMP5* (EMBL accession: Q14410) and has an in-frame stop codon.

In total, sixty-four pseudogenes were annotated. They are listed in Table 3.4 and displayed in Figure 3.8. Ninety-one percent (58/64) of the pseudogenes identified were processed i.e, they were produced by the incorporation of processed mRNAs into the genome, and nine percent (6/64) were non-processed. The chromosomal origin of the functional copy of each pseudogene was also determined (Table 3.4). Eleven pseudogenes were retroposed copies of chromosome 10 genes which was largely attributed to multiple (9) copies of the gene ornithine aminotransferase (*OAT*). These pseudogenes are found amongst the multiple members of the *SSX* gene family (section 3.6.1) and are hypothesised to have been generated from a single retroposition event followed by multiple genome duplications. Four processed pseudogenes were retroposed copies of genes found on the X chromosome.

Five of the six non-processed pseudogenes identified in human Xp11.22-p11.3 have been generated by a series of regional intra-chromosomal duplication events. These five pseudogenes (Ψ*SSX2,* Ψ*SSX3,* Ψ*SSX7,* Ψ*SSX8 and* Ψ*SSX9*) are all non-functional members of the *SSX* gene family which is also located in human Xp11.23. These pseudogenes are discussed in more detail in section 3.6.1. The other non-processed pseudogene, Ψ*SAH*, has resulted from an inter-chromosomal duplication event. The functional copy of the gene SA hypertension-associated homolog, *SAH*, is located on chromosome 16.

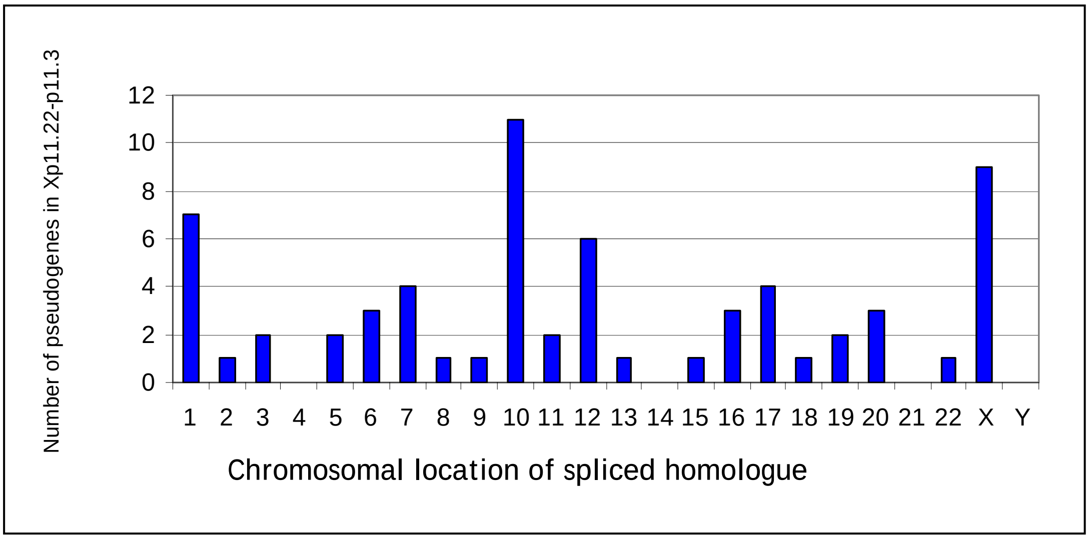


Figure 3.7 Chromosomal origin of pseudogenes identified in human Xp11.22-p11.3.

Table 3.4 Processed pseudogenes located in Xp11.22-Xp11.3

Locus	Origin	Functional gene	Locus	Origin	Functional gene
dJ192P9.1	chr16	TM4SF11	bA344N17.9	chr1	S100A11
bA254M24.1	chr12	KRT8	bA344N17.13	chr10	OAT
dA99M1.1	chr2	C2orf33	bA344N17.12	chr10	OAT
bA1158E12.1	chr20	SFRS6	AF196972.2	chr10	OAT
dA227L5.1	chr12	KRT18	AF196972.10	chr10	OAT
dJ774G10.1	chr20	PCNA	AF196972.4	chr10	OAT
dJ30G7.1	chr7	ACTB	bA1148L6.6	chr7	NOVEL
bA75A9.1	chr20	CTNNBL1	AC115617.2	chr18	ACAA2
bA75A9.2	chr12	GAPD	AF235097.12	chr7	HSPB1
bA75A9.3	chr12	VEZATIN	dJ8N8.10	chr5	VDAC1
dJ71L6.6	chr19	ZNF657	dJ8N8.11	chr16	SALL1
dJ71L6.1	chr10	PGAM1	AF238380.3	chr8	CPSF1
dJ71L6.3	chr1	NSEP1	bA637B23.1	chr1	H3F3A
bA198M15.1	chr9	CGI-38	bA637B23.3	chr13	HMGB1
bA571E6.1	chr3	NICN1	bA104D21.1	chrX	MAGE
bK2522E6.2	chr7	MAGED4	bA346H10.2	chr12	NUDT4
bA479F15.2	chr5	NPM1	bA56H2.1	chr17	PRR6
dJ230G1.4	chr6	C6ORF68	bA339K12.1	chr17	PRR6
dJ212G6.4	chrX	SMS	bA22B10.3	chr16	ZNF23
dJ393P12.2	chr1	WASF2	bA234P3.3	chr11	IPO7
dJ393P12.3	chr6	RPL7L1	bA234P3.4	chr22	HSPC051
bA38O23.3	chrX	ZNF81	bA234P3.5	chr6	TPMT
bA38O23.4	chr10	OAT	bA236P24.1	chrX	NOVEL
bA38O23.6	chr1	S100A11	bA258C19.3	chr17	RPL27
bA552E4.2	chr10	OAT	bA258C19.4	chr17	ACTG1
bA38O23.7	chr1	S100A11	dJ29D12.4	chr10	SUV39H2
bA344N17.2	chr1	S100A11	dJ29D12.3	chr3	RPSA
bA344N17.11	chr10	OAT	dJ29D12.2	chr19	FKSG24
bA344N17.6	chr10	OAT	bA339A18.3	chr15	FLJ20516

In total, 165 genes and pseudogenes were annotated in Xp11.22-p11.3. Forty-seven percent of the annotated transcripts were fully characterised known genes, and the region was also heavily populated with pseudogenes (38% of all structures annotated onto the genome sequence). Novel sequences that required additional experimental verification represented only fifteen percent of all structures, demonstrating the contribution of cDNA sequencing projects to describing the gene content of the genome.

3.3.4 Distribution of genes

Figure 3.8 displays the orientation and distribution of the genes and pseudogenes on the genome sequence. The tiling path of sequenced and analysed clones is also displayed. This figure highlights the non-uniform distribution of genes in human Xp11.22-p11.3. Genes are found in clusters on the genome sequence. Other clusters of human genes have been associated with a high G+C content, gene function and repeat content (Arhondakis, *et al.,* 2004).

Like functional genes, the pseudogenes annotated onto Xp11.22-p11.3 were not uniformly distributed across the genome sequence (Figure 3.). Analysis of the pseudogene content in chromosomes 21 and 22 found that pseudogenes tend to be found in ‘hot-spots’, with most being located near the centromeres (Harrison *et al.,* 2002a). Chromosomes 21 and 22 contain approximately 3,000 pseudogenes, and by extrapolating the figure to the entire genome it has been estimated that the human genome will contain approximately 20,000 pseudogenes (Harrison *et al.,* 2002b). From this estimate it could be predicted that the 5.6 Mb region studied would have approximately 37 pseudogenes, which is well below the observed figure. This suggests that Xp11.22-p11.3 is enriched not only for genes but also pseudogenes.

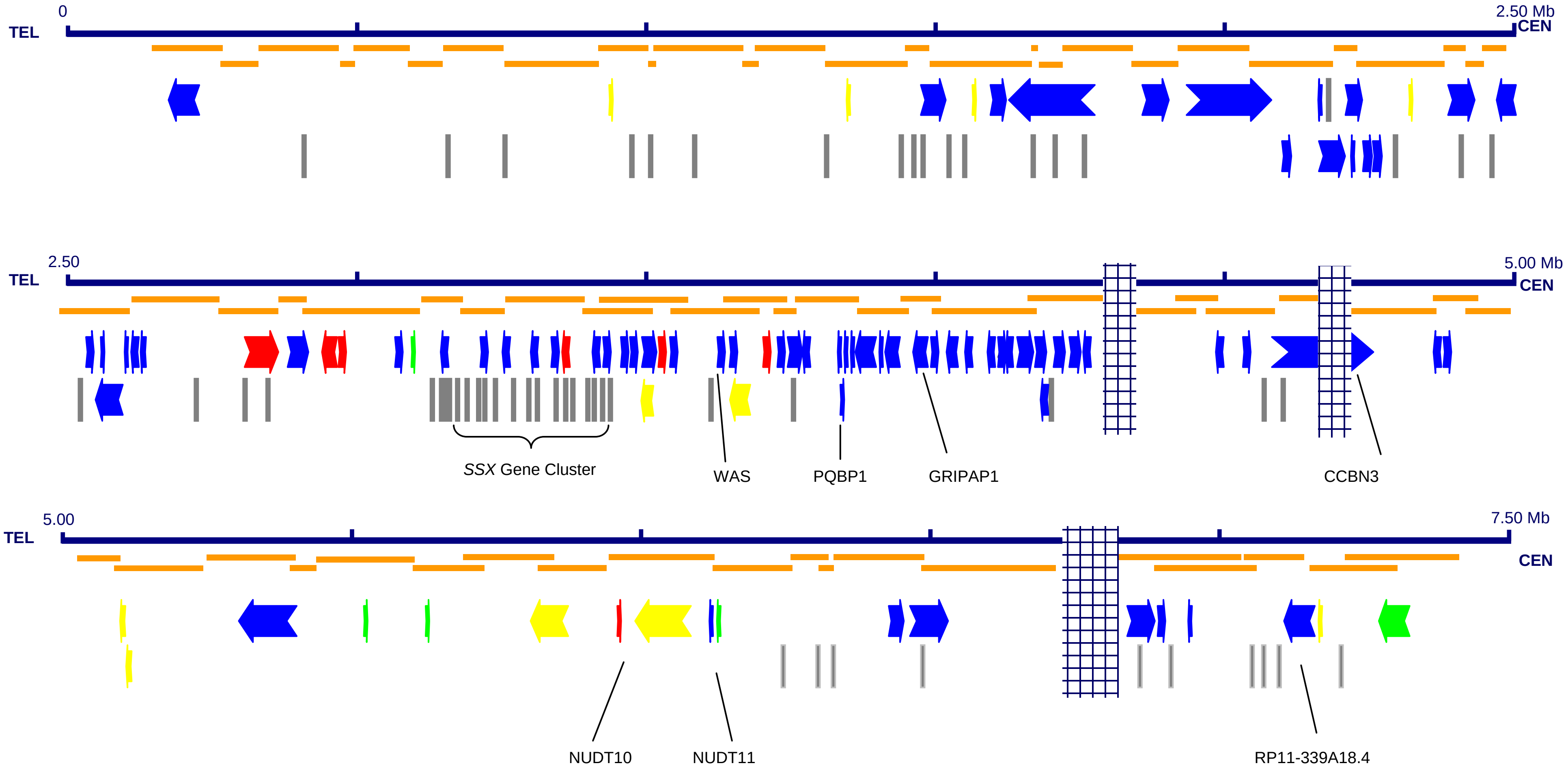


Figure 3.8 Gene annotation of human Xp11.22-p11.3.

Displayed are: the tiling path (orange), known genes (blue), novel_cds (red), novel_transcript (yellow), putative transcripts (green) and pseudogenes (grey). Gaps in the genome sequence are denoted with a hashed bar. Direction of transcription is indicated by the direction of the arrow.

3.4 Experimental verification of novel and putative genes

In order to verify experimentally transcription of novel and putative genes, a panel of vectorette cDNA libraries constructed from 19 human tissues was screened by vectorette PCR (section 2.15.3). Primer pairs were designed within a potential exon sequence for each novel CDS, novel transcript and putative genes. In total, 36 primer pairs were designed to 29 genes, including positive and negative controls. Primers were also designed to three known genes *FOXP3*, *PHF16* and *UBE1* which served as positive controls. Negative control primer pairs were designed to a retroposed pseudogene, bA637B23.1, and to a non-transcribed region of genomic sequence. All of these primers are listed in Table 3.5, while their sequences are listed in Appendix I.

In order to ensure specificity, each primer pair was pre-screened on the following templates: human genomic DNA, clone 2D (a human-hamster cell hybrid containing the human X chromosome as its only human component) and hamster genomic DNA. A negative control was also included. Reactions were performed using three different primer annealing temperatures (55°C, 60°C and 65°C) to determine the optimum cycling parameters (section 2.15.1).

Vectorette libraries consisting of pools of 20,000 clones were screened using the primer pairs at the optimal reaction conditions, and positive pools were selected for further analysis (described in section 2.15.2). This procedure identified a positive clone pool for 15 genes, including the three positive control genes. Positive identification was heavily dependent on the type of evidence from which the gene structure was predicted. Five of the novel CDS sequences (62.5%), six novel transcript sequences (54%) and one putative transcript sequence (20%) were identified in at least cDNA pool. The positive pools identified for each primer pair are listed in Table 3.5.

Failure to identify positive clones for the cDNA pools could be attributed to the following:

- The predicted gene not being real.
- Inadequate coverage of different cDNA samples types, including different tissue types, cell types, developmental stages or cell-cycle stages.
- Inefficiencies in the construction of the vectorette libraries.

Further analysis was completed on positive pools for up to three different tissues for each primer pair. Positive cDNA screening results were pursued to obtain novel sequence data flanking the original exon prediction. The vectorette products were amplified using each of the gene specific primers together with a universal primer (primer 224) that was located within the vectorette “bubble”. The resulting PCR products were purified and sequenced in-house by the Research and Development Group. Sequences were aligned to the genome sequence in Xace and were used to extend and confirm gene structures.

Sequence data was not generated for all genes positive in the pool screens as difficulties were experienced amplifying specific vectorette products. This has been attributed to the use of the general vectorette primer, 224 in the PCR reactions. Attempts were made to optimise the amplification conditions by varying the PCR cycling conditions, primer concentrations and using nested primer, but these were not successful.

In total, twenty-four annotated gene structures were tested using vectorette cDNA libraries and expression was confirmed for 12 genes. The annotated gene structures were extended with additional sequence for seven genes. Figure 3.9 summaries the utility of this approach in producing informative and novel sequence data. Sequence data were generated for eight genes and the relevant in-house accession numbers (sccd numbers) for these sequences are listed in Table 3.5.

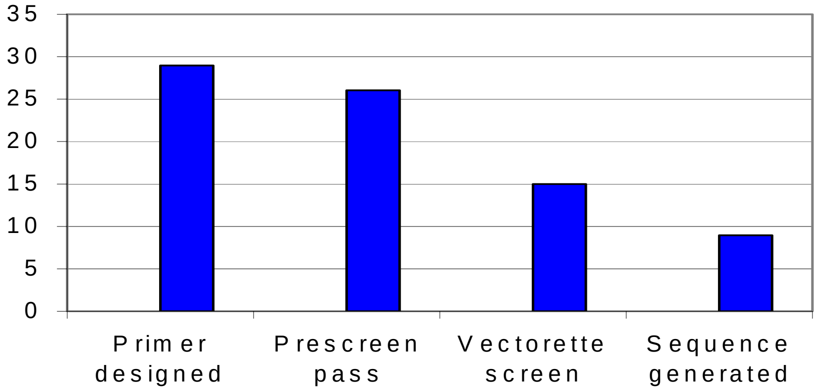


Figure 3.9 Success rates at various stages of analysis to attempt to confirm and extend novel and putative gene structures

During the course of this study, full-length cDNA sequences with homology to two predicted genes in Xp11.22-p11.3 were deposited in the public databases. The two novel cds genes AF196971.4 and RP11-348F1.1 became *ERAS* and *NUDT10* respectively and all future reference to these genes will use the HGNC approved gene names. The expression of *ERAS* is confined to embryonic stem cells (ES cells) (Takahashi *et al.,* 2003). This tissue was not included in this study.

An example of a novel cds whose structure was confirmed and extended using the aforementioned protocol is RP11-54B20.4 (Figure 3.10). This gene was predicted by GenScan, FgenesH and Span, and by EST (100% similarity) and protein (55% identity) sequences. Protein homology suggested that this gene encodes a zinc finger protein. Screening of vectorette cDNA libraries confirmed expression in adult heart, uterus, bone marrow, lymphocytic leukaemia T cells, and neuroblastoma cells (see Figure 3.10). Further amplification of the flanking cDNA from adult heart revealed two bands with specific primer A. Only one of these bands to has sequence similarity with RP11-54B20 (upper bands with specific primer A), while the sequence of lower band had little similarity to the gene, RP11-54B20.4.

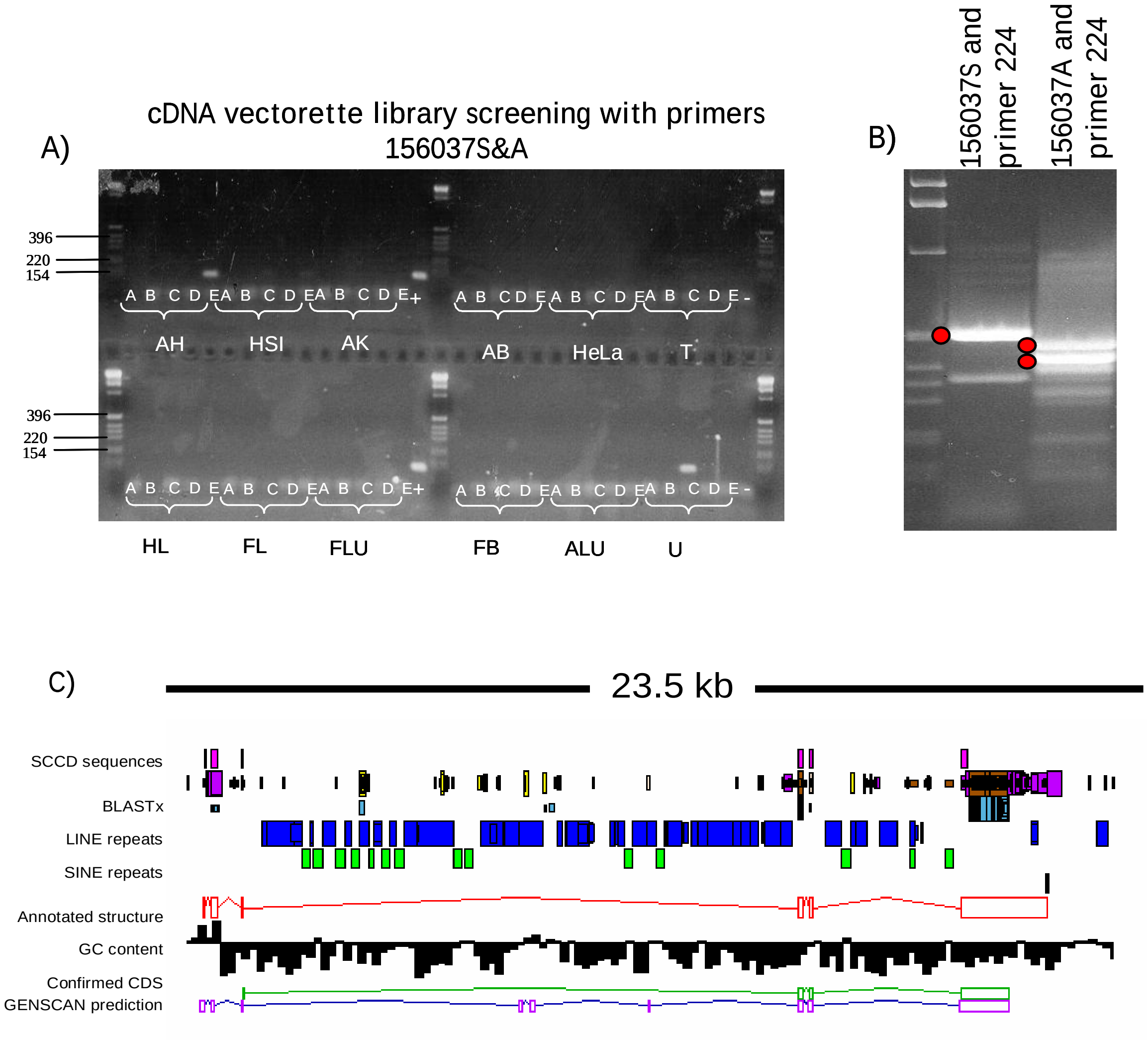


Figure 3.10 Confirmation and extension of novel cds RP11-54B20.4

A) Example of cDNA vectorette library screening for RP11-54B20.4. Products are observed in the vectorette cDNA libraries adult heart (AH) and monocyte (U).

B) Vectorette PCR for the positive superpool AH (E). Bands markers with a dot were excised for sequencing.

C) Gene structures as annotated in Xace. The diagram shows an ACeDB representation of the gene structure, RP11-54B20.4. The confirmed cds structure is then shown. The GENSCAN prediction for this RP11-54B20.4 is also displayed

Table 3.5 Experimental verification of novel genes and transcripts and putative genes.

The vectorette libraries used in this study are described in Table 2.4.

	Gene	Category	Primer Pair	Annealing temperature (°C)	Positve vectorette libraries	Specific cDNA end amplification	Sequenced	Comments
	FOXP3	Known	156841	60	H, SK, YT	n.a.	n.a.	Positive Control
	UBE1	Known	156840	55	All	n.a	n.a.	Postive Control
	PHF16	Known	156758	65	BM, FB	n.a	n.a.	Positive Control
	N.A.	n.a.	156806	60	None	n.a	n.a.	Negative Control
	bA637B23.1	Pseudogene	156802	60	None	n.a	n.a.	Negative Control
1	RP11-1145B22.1	Novel_cds	388756	60	HL	yes	sccd_10894	Encodes protein with a zinc finger domain
2	RP11-54B20.4	Novel_cds	156037	60	AH, U	yes	sccd_10895	Encodes protein with a zinc finger motif
3	RP11-38O23.2	Novel_cds	187910	60	T, U, HPB	yes	sccd_10898	Encodes protein similar to lysozyme C
4	AC115617.1	Novel_cds	486758 486759 486760 529890	60 65 60 60	HL	yes	sccd_31771 sccd_31772 sccd_31774	Encodes protein with a lactoylglutathione lyase domain – may be involved in amino acid metabolism.
5	RP11-348F1.1	Novel_cds	156807 172512	Failed 60	AK, AB, T, FB, HPB, SK	yes		Now known gene – NUDT10 (Hidaka *et al.*, 2002)
6	AF196971.4	Novel_cds	156785 498780	65 60	None	n.a.	n.a.	Known gene – ERAS (Takahashi *et al.*, 2003)
7	RP11-54B20.3	Novel_cds	156035	60	None	n.a.	n.a.	Encodes protein with a zinc finger motif
8	RP11-344N17.4	Novel_cds	Not analysed. Member of *SSX* gene family **-** difficulties experienced designing specific primers					

Table 3.5 continued….

	Gene	Category	Primer Pair	Annealing temperature (°C)	Positve vectorette libraries	Specific cDNA end amplification	Sequence	Comments
1	AF238380.5	Novel transcript	156846	60	H, HPB, FL, AB	Yes	sccd_10880 to sccd_10884	
2	RP11-1148L6.5	Novel transcript	486773	60	B, FB	Yes	sccd_31936	Antisense transcript to *RBM3*
3	RP11-805H4.2	Novel transcript	156812 380014 380015	60 60 65	FL FL FL	Yes	sccd_11421 sccd_11422 sccd_11435 sccd_11434	
4	AF196970.3	Novel transcript	172510	60	AB	Yes	No novel sequence generated – antisense transcript to *SUV39H1*	
5	AF235097.6	Novel transcript	380006	65	B, FB	No		
6	RP11-258C19.5	Novel transcript	156811	60	T, HeLa	No		
7	RP11-576P23.4	Novel transcript	156810	65	None			
8	RP1-30G7.2	Novel transcript	156783	60	None			
9	RP5-1158E12.2	Novel transcript	156757	65	None			
10	bA104D21.3	Novel transcript	156804	65	None			
11	RP1-71L6.2	Novel transcript	400506	Failed				

Table 3.5 continued….

Putative genes

	Gene	Category	Primer Pair	Annealing temperature (°C)	Positve vectorette libraries	Specific cDNA end amplification	Sequence
1	AF196972.9	Putative	387162	60	AH, HeLa, HL	5’ No 3’ Yes	No
2	bA56H2.2	Putative	156798 156871	Failed 60	n.a. None		
3	RP11-1158E12.1	Putative	156756	60	None		
4	RP11-348F1.3	Putative	none				
5	bA258C19.6	Putative	156809	60	None		

3.4.1 Antisense Transcripts

In the process of annotating genes to human Xp11.22-p11.3, two antisense genes were identified. These novel transcript genes were AF196970.3 (antisense to *SUV39H1)* and RP11-1146L6.5 (antisense to *RBM3,* Figure 3.11). To confirm their expression, primer pairs were designed to novel exons that did not overlap the exons from the sense strand (these are listed Table 3.5). The expression of only one gene, RP11-1146L6.5, was confirmed in human and foetal brain samples. RP11-1146L6.5 may have a role in regulating the expression of *RBM3*. *RBM3* is a glycine-rich RNA-binding protein, whose expression is enhanced under mildly hypothermic conditions and it has been postulated to facilitate protein synthesis at colder temperatures (Chappell *et al.,* 2001). As the expression of *RBM3* is tightly regulated would be interesting to determine if RP11-1146L5.6 has any influence in regulating its expression. This could be determined using techniques such as *in vitro* transcription – translation studies.

Screening of both vectorette-ligated and non-cloned cDNA samples failed to confirm the expression of AF196970.3.

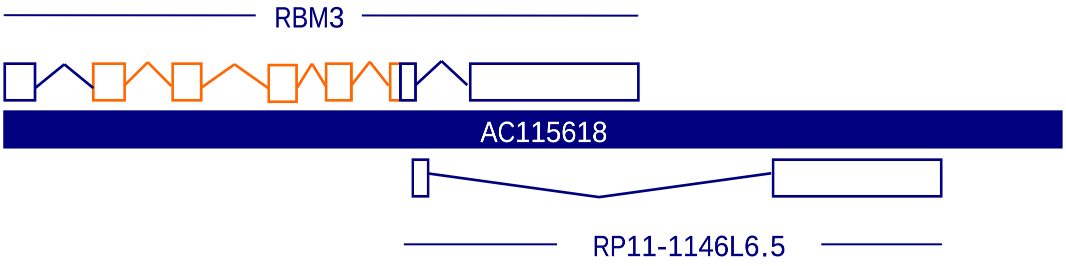


Figure 3.11 Annotation of RBM3 and its antisense gene AC1145L6.5 on the genome sequence, AC115618

The genome sequence is represented by a solid blue block in the middle of the figure. The exon/intron structure of the known gene RBM3 is displayed above the genome sequence and the direction of transcription runs from left to right (forward orientation). Blue boxes represent UTRs and orange boxes (exons) represent CDS. Exon structure of RP11-1145L6.5 is displayed below the genome sequence. It is transcribed in the opposite orientation to RBM3.

3.4.2 Transcript features

General features of the genes identified in section 3.3 were extracted for further analysis. In total, 101 sequences were analysed for a number of features which are shown in Table 3.6. Genes spanned 50.2% of the 7.3 Mb of genome sequence studied while 5.1% of the genome sequence is contained within exons. These percentages are considerably higher than the entire X chromosome and are in keeping with it being the most gene rich area of the chromosome (Ross et al, 2005).

Table 3.6 Overview of transcript features annotated in human Xp11.22-Xp11.3

Feature	Xp11.22-p11.3					All X chr genes**
	Known genes	Novel CDS	Novel transcripts	Putative transcripts	All genes	
Number in category	77	8	11	5	101	1098
Total gene coverage (bp)	1946288	119754	306087	18143	2390272	51724866
Mean gene length (bp)	25276	14969	27826	3629	23666	48883
Median gene length (bp)	12127	7609.5	12663	2315	11711	11542
Total number of exons*	822[@]	34	27	12	895	7668
Mean exons/gene*	10.82[@]	4.25	3.38	2.4	10.23	6.98
Median exons/gene*	7[@]	4	2	2	6	4
Total exon coverage (bp)*	218418[@]	11628	12964	3086	246096	2672795
Mean exon length (bp)*	266[@]	342	480	257	281	351
Total intron coverage (bp)*	1727870[@]	108126	293123	15057	2144176	44295727
Mean intron length (bp)*	2316[@]	4159	11274	579	3621	6965

* longest annotated transcript used

** figures from Ross *et al.*, 2005

[@] CCBN excluded from analysis because it spans a gap in the genomic sequence

3.5 Assessing the completeness of annotated genes

The completeness of all gene structures was assessed to determine if the annotation extended to transcription start (TSS) and termination sites (TTS). Three computer programmes CpGIsland (G. Micklem, unpublished), Eponine (Down and Hubbard, 2002) and First EF (Davuluri *et al.*, 2001) were used to determine if the 5' end of genes had been reached. These were used in concert to increase the likelihood of identifying potential TSSs. The 3' ends of genes were identified by searching transcribed sequences for polyadenylation signals and polyadenylation sites.

3.5.1 Polyadenylation sites

The 3′ ends of fully processed eukaryotic mRNAs (except most histone genes) often have a polyadenosine (poly(A)) tail. Poly(A) tails have been shown to influence mRNA stability, translation and transport and are thought to be involve in other transcriptional and post-transcriptional processes, such as splicing and transcriptional termination. Polyadenylation of pre-mRNA transcripts is a two-step process; transcripts are cleaved prior to the addition of the poly(A) tail. The cleavage and polyadenylation specificity factor (*CPSF*) binds to a 6 bp nucleotide sequence that is located approximately 15-30 bp upstream from the site of mRNA cleavage. The presence, frequency and sequence composition of polyadenylation signals for all genes that were annotated in Xp11.3-p11.22 was determined by visual inspection of cDNA and EST sequences that harboured a polyA tail. The most common signal used in the polyadenylation process is AATAAA, while 10 other signals have also been reported (Beaudoing *et al.* 2000).

All annotated gene structures were manually inspected for the presence of a polyA signal using Xace. Transcripts that contained a polyA tail were inspected for any of the 11 known polyA signals located approximately 15-30 bp upstream from the start of the polyA tail. Seventy-six percent of all genes identified in human Xp11.22-p11.3 were found to have a polyA signal (77/101) and the likelihood of finding a signal varied with the gene type (Figure 3.12A). The majority of known genes (68/77) had a polyA signal, while half of the novel CDS genes (4/8), three of novel transcripts (3/11) and only one putative gene had this feature (1/5). As expected, the completeness of annotated gene structures is closely associated with the presence of a polyA signal and tail.

It is also possible for a gene to have more than one polyadenylation site. Recent analysis has suggested that 54% of human mRNA species have more than one polyadenylation site which can be used to modulate gene expression in a tissue specific or developmental manner (Tian *et al.,* 2005). The presence of multiple polyadenylation signals was not observed in any of the novel CDS, novel transcript or putative genes. Nine known genes have more than one polyadenylation signal, further highlighting the complete nature of these transcripts.

The sequence composition of these signals was assessed in known genes. The highly utilised AATAAA hexanucleotide was present in 64 genes, with its most common variant being ATTAAA which was identified in five genes. Five other variants were also identified in known genes (Figure 3.12C).

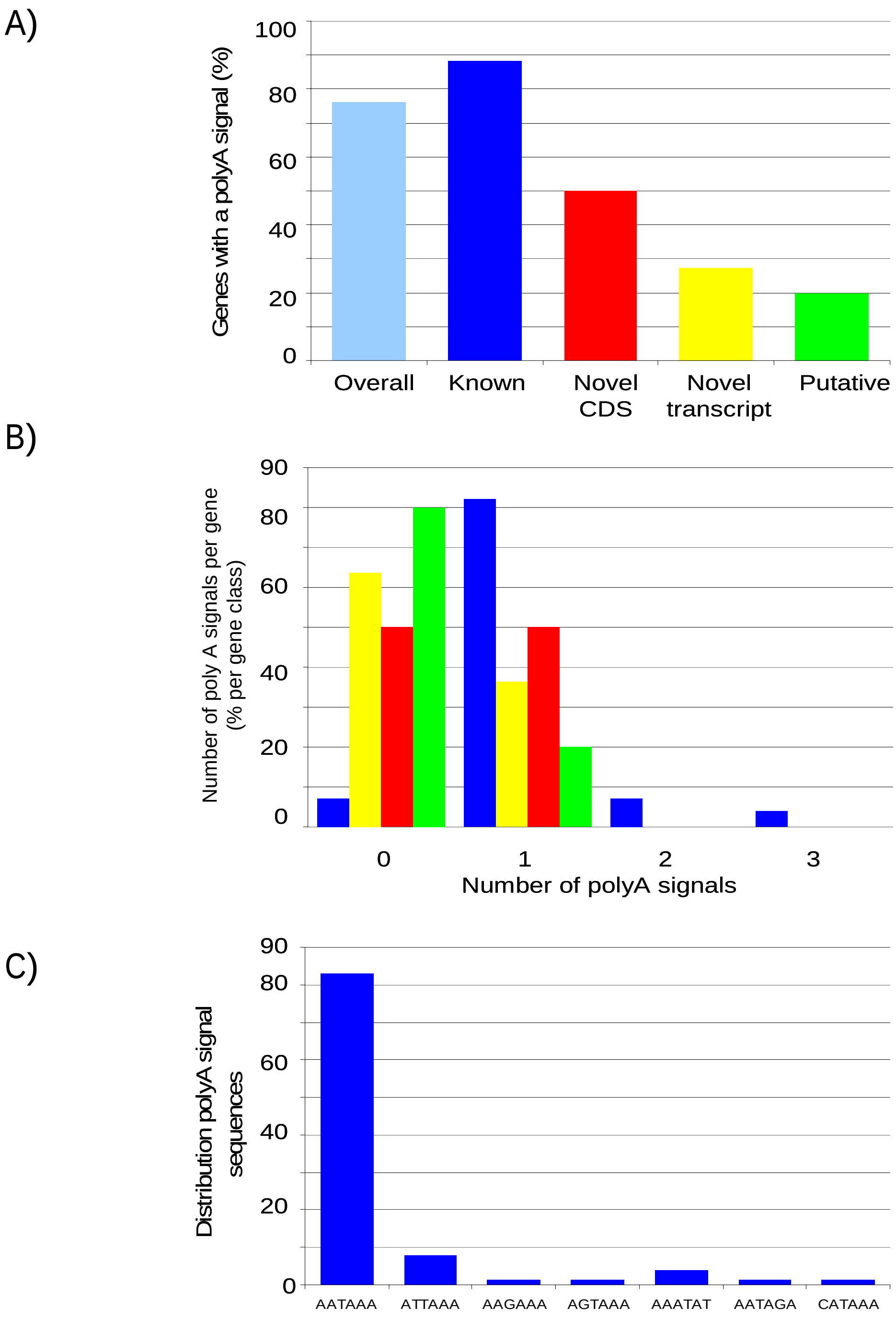


Figure 3.12 Analysis of polyadenylation signals in human Xp11.22-p11.3

A) Proportion of genes with a polyA signal. Known genes (blue), novel CDS (red), novel transcripts (yellow) and putative genes (green).

B) Percentage of polyA signals identified for each gene type: known genes (dark blue, n=77), novel CDSs (red, n=8), novel transcripts(yellow, n=11), and putative genes (green, n=5).

C) Distribution and sequence composition of polyA signals in all genes that have a polyA signal.

3.5.2 Transcription start sites

Computational identification of transcriptional activation regions, or promoters, is complicated by their diversity. One set of rules cannot be employed to identify all promoters. Rather programmes scan genome sequences for a series of different sequence motifs commonly associated with transcription initiation such a CpG islands, TATA boxes, and G+C rich regions, or homology with orthologous promoters.

Three different programmes, CpGIsland (G. Micklem unpublished), FirstEF (Davuluri *et al.,* 2001) and Eponine (Down and Hubbard, 2002) were used to predict if the TSS of annotated gene structures had been reached. CpGIsland identifies genomic regions enriched for CpG nucleotides that are longer than 300 bp in length. FirstEF identifies potential first exon donor sites, in both CpG associated and CpG non-associated transcription start sites. Eponine recognises multiple sequence specific motifs associated with transcription start sites. The following criteria were used to identify potential TSSs:

- A CpG Island located within 1 kb of the first annotated exon (CpG islands are greater than 300 bp in length and have a minimal GC content of 50%)
- A eponine prediction located within the first annotated exon (probability cut off 0.995)
- A First Exon Finder (1st EF) prediction located within the first exon (cut off values for the promoter of 0.4, first exon 0.5, first splice site of 0.4).

Annotated pseudogenes were also included in the analysis. Here, it was predicted that most pseudogenes would not be transcribed, would not have a functional TSS and therefore would not have any predicted TSSs.

When a TSS was predicted at the 5’ end of gene, it was most commonly identified by CpGIsland and 1st EF but not Eponine. Thirty-one out of 70 TSSs (43.2%) were predicted by all three programmes, while 33 TSSs (47.2%) were predicted by CpG island and 1st EF but not eponine. Only five TSSs were identified by only one programme, four by CpGIsland alone and one by 1st EF alone. Of the programmes used in this analysis, Eponine was the least sensitive and identified the smallest number of TSSs. The percentage of TSSs identified by each computer programme is displayed in Figure 3.13- A.

The proportion of genes with a predicted TSS at their 5' end is displayed in Figure 3.13-B. Potential TSSs were identified for 78% of the annotated known genes (60/77), 50% of novel CDS genes (4/8), 45.5% of novel transcript genes (6/11) and 20% of putative genes (1/5). These results suggest that the majority of novel transcript and novel CDS and putative genes may require additional evidence to extend their annotation to a TSS. It is, however, possible that the annotation of some genes may have extended to the TSS but these sites were not recognised because they did not meet the threshold for identification. Re-analysis with lower stringencies may identify additional TSSs that were not recognised in this analysis.

Eleven percent of annotated pseudogenes (7/64) had a predicted TTS (Figure 3.13-B). All predictions were associated with processed pseudogenes and it is likely that the TSSs contained within pseudogenes are not functional. They may be remnants from their functional counterparts.

It is predicted that approximately 75% of all gene structures are annotated to the TSS and TTS sites. The state of completion is heavily dependent on the gene type (Figure 3.14). The gene class with the most complete structures is known genes (TSS are predicted for 78% of genes and TTS for 88% of genes, while only one TTS and one TSS were identified on annotated putative gene structures (these sites were identified on two independent transcripts). Additional forms of evidence may be required to complete the gene structures that do not have an associated TSS or a polyA tail.

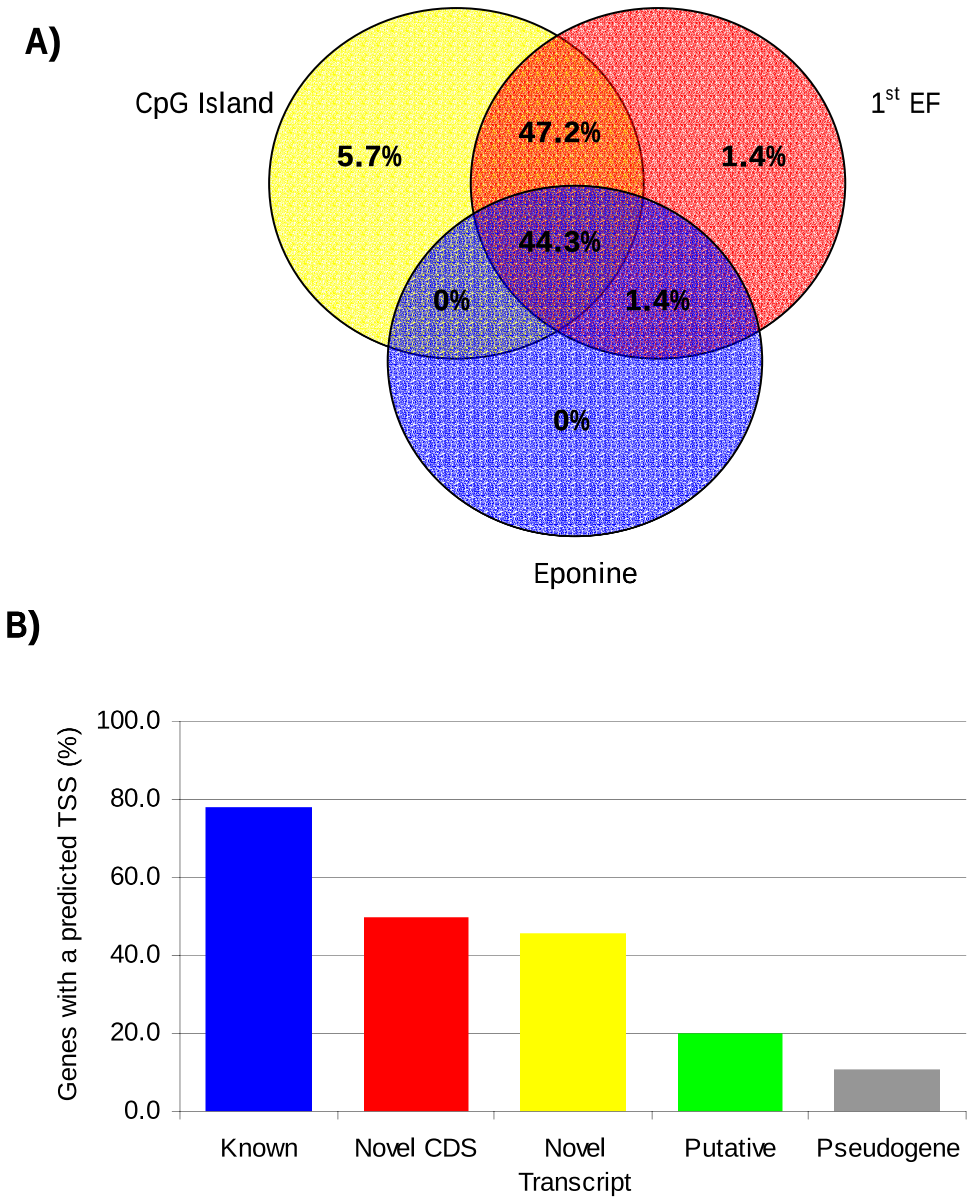


Figure 3.13 Distribution of predicted transcript start sites (TSSs) for genes and pseudogenes in human Xp11.22-p11.3

Three programmes, Eponine, 1st EF and CpGIsland were used to identify TSSs in the genome sequence from human Xp11.22-p11.3.

A) Venn diagram displaying the proportion of all predicted TSSs identified by each programme.

B) Percentage of annotated genes with a predicted TSS in their first exon.

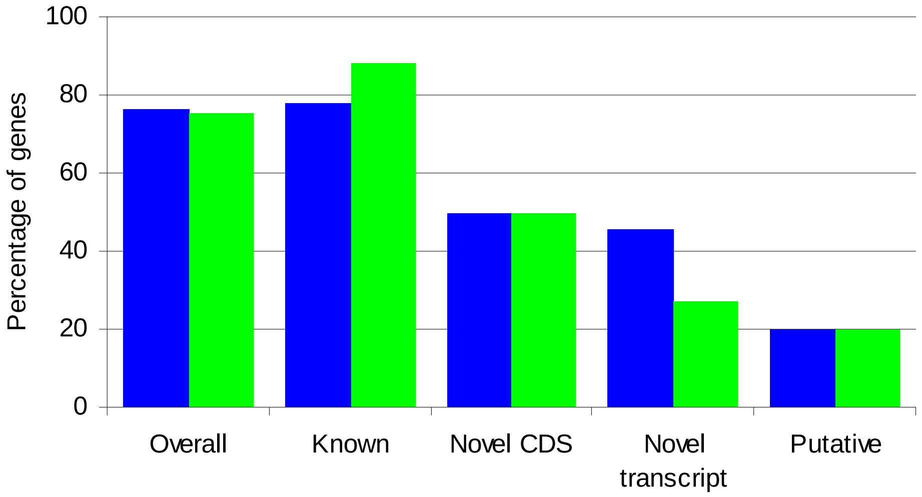


Figure 3.14 State of completion of annotated gene structures

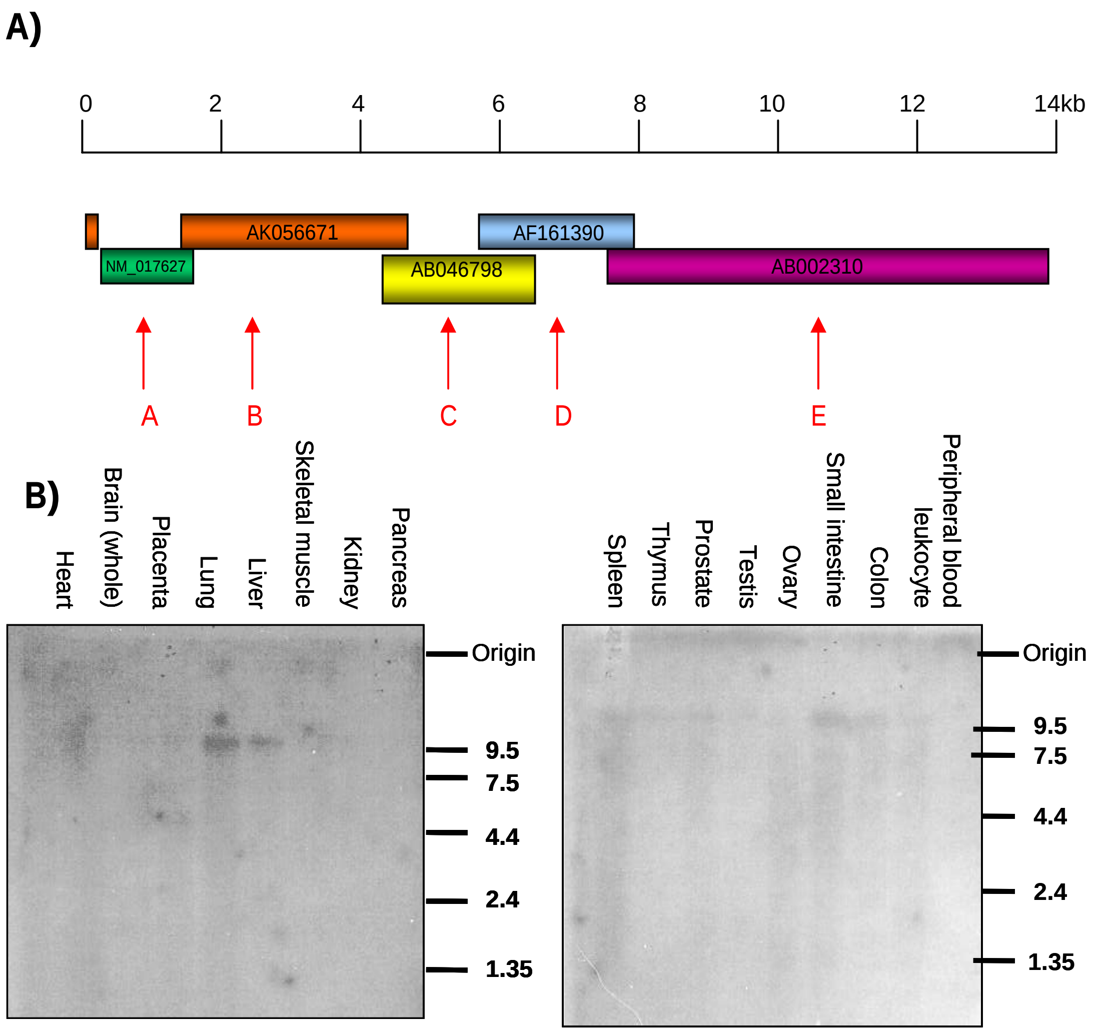
Completion at the 5' end was assumed in the presence of a predicted TSS. Completed 3' ends had a polyA signal and polyA in their transcript sequence (green). The results are displayed for each gene class.

3.5.3 Experimental evidence confirms the transcript size of RP11-339A18.4

The completeness of annotated gene structures can also be assessed using experimental procedures. Northern blotting is suited to such analysis because the primary substrate is RNA (total or mRNA). Therefore, reverse transcription of mRNA to cDNA is not required to complete the experiment. In addition, Northern blotting is unbiased, as the size of the full-length transcript can be determined without prior knowledge of its TSS or TTS.

The gene RP11-339A18.4, was analysed because it was annotated from five separate overlapping mRNA transcripts. Northern blot analysis was carried out to confirm that the five RNAs did link up to produce the very large transcript. The gene spans 180 kb, consists of 83 exons and its processed transcript is in excess of 14 kb in length.

The expression of RP11-339A18.4 was confirmed by Northern blot analysis, using a probe designed to each of the five overlapping mRNA sequences shown in Figure 3.15. Hybridisation of these probes to total RNA confirmed the expression of a large transcript (in excess of 9.5 kb) in various tissues. Strongest signals were observed in the liver, lung and small intestine.



C)

Probe	stSG #	Designed to	Transcript size	Tissue
A	367537	NM_017627	> 9.5 kb	Liver, lung
B	367539	AK056671	> 9.5 kb	Liver, lung, small intestine
C	367541	AB046798	Failed	Failed
D	367543	AF161390	>9.5 kb	Liver, lung
E	367545	AB002310	>9.5 kb	Liver only

Figure 3.15 Confirmation of gene expression, RP11-339A18.4.

A) The composite structure of RP11-339A18.4. Arrows indicate the location where specific primers were designed (listed A-E).
B) Northern blots using probe B(stSG367539). Size standards (kb) are displayed on the right hand side of each blot. Data from other probes are not shown.
C) Summary of Northern blot results.

3.5.4 Alternative splicing

The gene annotation protocols employed in this chapter focused on identifying and annotating one full-length mRNA transcript per gene. However, in many cases more than one transcript variant was identified per gene. For example, 9 transcript variants were identified for the gene *UBE1* (Figure 3.16).

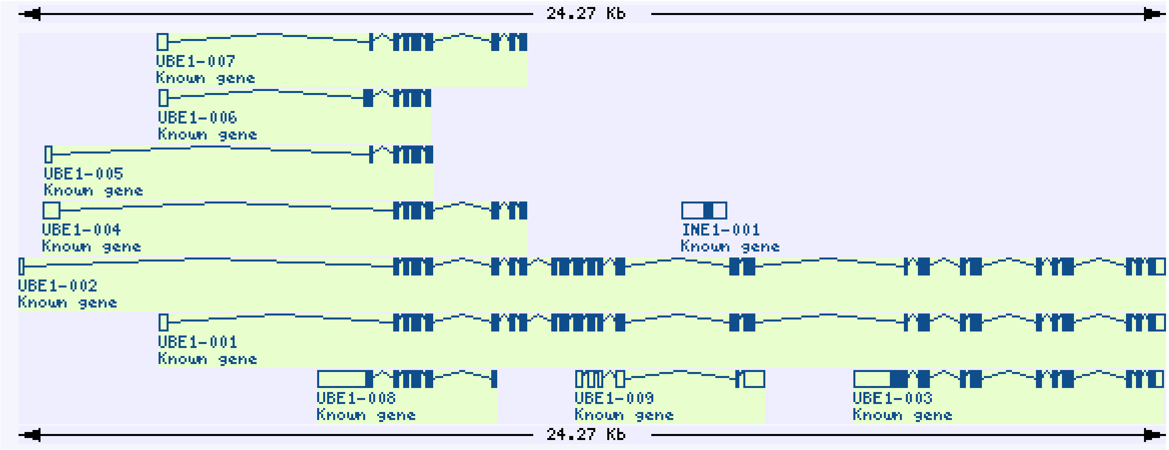


Figure 3.16 Transcript variants of the gene *UBE1*
Open boxes represents UTR sequence, while filled boxes depict coding sequence. This figure was extracted from Vega http://vega.sanger.ac.uk.

In order to detect the approximate level of transcript variation in human Xp11.22-p11.3 an analysis of the frequency of alternative splicing events was carried out using the database PALSdb (Huang *et al.*, 2002, http://palsdb.ym.edu.tw/). PALSdb predicts alternative splicing events using UniGene clusters of EST and mRNA sequences are aligned to the corresponding REFSEQ sequences. After filtering for sequence similarity (minimum 95% and a minimum 50 bp overlap), an insertion or deletion in any EST or mRNA entry is recorded as a splice variant. A gene name search was completed for all known genes identified in this region. Only forty-nine of these genes had a PalsDB entry and 75% of which were predicted to have more than one transcript. The average number of transcripts per gene is 3.6. It is acknowledged that the information contained in this database only gives a crude estimate of the frequency of alternative splicing. It does not monitor splice site consensus sequences and contaminating sequences are also recorded as alternative transcripts. Nevertheless, this simple analysis gave an indication of a considerable degree of alternative splicing of gene in the region. A more detailed discussion of alternative splicing is completed in chapter 4.

3.6 Duplication events

Throughout the annotation process multiple members of the *SSX* gene family, zinc finger and NUDT family were mapped to human Xp11.22-p11.3. The human genome contains a myriad of duplications varying in both ancestry and size. Approximately 5% of the human genome sequence is covered by duplications greater than 1 kb in length with more than 90% sequence identity. Duplication frequency has been weakly associated with regional gene density, repeat density, recombination rates and GC content (Zhang *et al.,* 2005). Locating and characterising such duplications has enabled scientists to define accurately the genome changes that may have contributed to species divergence, as well as identifying the potential role of genome duplication in human disease.

3.6.1 Duplication of the SSX family

The *SSX* gene family consists of 9 functional and 9 non-functional members that were first identified by sequencing the breakpoint of the t(X:18) translocation in synovial sarcomas (Clark *et al.,* 1994). The recorded translocation events generated a fusion product between *SYT* and one of the *SSX* genes, with the 5' end of the *SYT* linked to the 3' end of the *SSX1*, *SSX2* or *SSX4* genes. Transcript features of these family members such as exon/intron structure, sequence homology and expression profiles have recently been characterised, although this analysis failed to map the location of four family members (Ψ*SSX*2, *SSX*3, *SSX*4 and *SSX*6) on the genome sequence (Gure *et al.,* 2002). Detailed analysis of the finished genome sequence accurately localised four functional and four non-functional members of the *SSX* gene family to Xp11.23. A comparison of the gene annotation provided in this study to that of Gure and colleagues (2002) identified one anomaly in the published data where the novel cds gene, RP11-344N17.4 was incorrectly classified as a pseudogene, Ψ*SSX*6. The location and orientation of the *SSX* family members in Xp11.23 is displayed in Figure 3.17. cDNA and translated protein sequences of the *SSX* family member were extracted for further analysis. Their nucleotide and amino acid sequences were aligned using ClustalW to compare their similarity. These are also displayed in Figure 3.17.

Similarity Dot plot analysis of the genome sequence harbouring the *SSX* gene family members confirmed the high level of duplications within this region where a complex pattern of inverted and tandem duplications were observed.

An additional 15 processed pseudogenes were scattered amongst the *SSX* family members; nine pseudogenes were retroposed fragments of the gene, ornithine aminotransferase, *OAT* (see section 3.3.3) and three pseudogenes were retroposed fragments of the gene S100 calcium binding protein A11, *S100A11*. As discussed in section 3.3.3, it is predicted that the *OAT* pseudogenes arose from a single retrotransposition event followed by subsequent genomic duplication.

The annotation of both functional and non-functional gene families to this region may provide a useful basis for future evolutionary studies on the duplication events in Xp11.23. Detailed analysis of the *SSX* gene family members has confirmed that these genes are under selective pressure to remain conserved throughout evolution (M. Ross personal communication). Comparative sequence analysis has demonstrated that the *SSX* gene family has undergone independent amplification in both the human and mouse genomes (M.Ross personal communication).

A)

GENE	RefSeq
SSX1	NM_005635
SSX3	NM_021014
SSX4	NM_005636
SSX4L	BA344N17.4*
SSX5	NM_021015
SSX6	NM_017357
SSX9	NM_174962

* can be accessed at Vega (http://vega.sanger.ac.uk)

B)

% nucleotide identity

SSX1	SSX3	SSX4	SSX4L	SSX5	SSX6	SSX9	
100	82	83	84	77	89	85	SSX1
77	100	89	82	76	84	88	SSX3
79	80	100	92	75	87	89	SSX4
79	80	100	100	77	84	83	SSX4L
63	69	64	64	100	76	76	SSX5
81	78	76	76	62	100	86	SSX6
79	86	82	82	68	77	100	SSX9

% amino acid identity

C)

D)

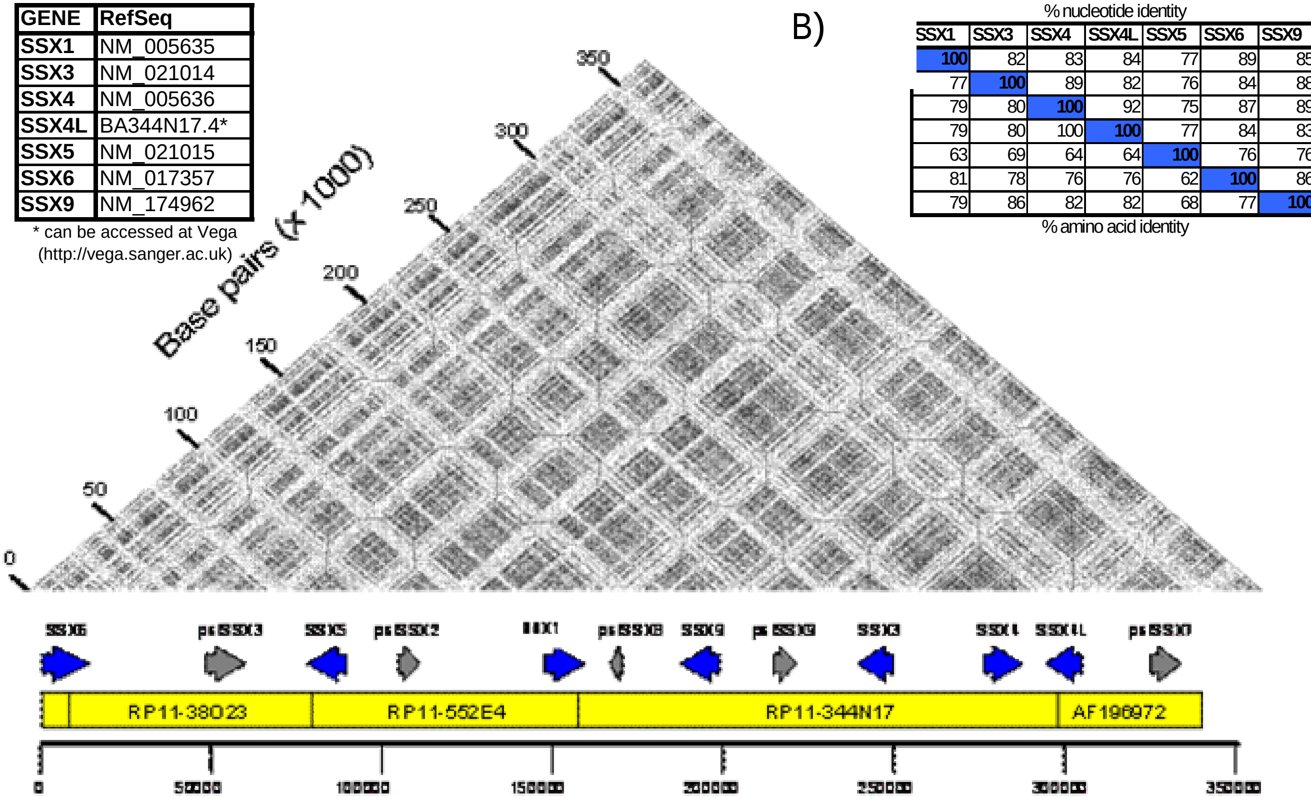


Figure 3.17 Annotation of *SSX* gene family members

A) RefSeq identifiers numbers for functional copies of the *SSX* families that were mapped to Xp11.23

B) Sequence identity at both nucleotide and amino acid level for functional *SSX* genes in Xp11.23.

C) Similarity dot plot analysis of the genome sequence. Horizontal lines represent duplicated sequences while vertical lines represent inverted repeats.

D) Transcript map of *SSX* region. Blue arrows represent functional genes while grey arrows represent *SSX* pseudogenes. Tilepath clones are shown in yellow.

3.6.2 Genes located in an inverted repeat

Recent analysis of the complete X chromosome sequence demonstrated that it contained a disproportionately high number of large, highly homologous inverted repeats that contained testes specific genes (Ross *et al.,* 2005). An inverted repeat was also observed in a single clone in Xp11.22 where the gene *NUDT11* was duplicated. Expression of *NUDT11* was supported by both mRNA and EST sequences and the gene encodes a 164 amino acid protein that is predicted to contain a NUDEX domain (involved in DNA repair). A duplicate of this gene, *NUDT10*, was identified 150 kb downstream from *NUDT11*. At the time of analysis this locus was not supported by a full-length cDNA sequence, however five ESTs with 100% homology to the genomic sequence were identified. Expression of the duplicate locus was also confirmed by screening vectorette cDNA libraries, using primers that were designed to a region with lower similarity to *NUDT11*. Sequencing of the resulting PCR products confirmed its expression of *NUDT10*. *NUDT10* shares 88% DNA identity with *NUDT11,* with the 3' UTR displaying the least sequence identity. The two encoded proteins share 99% identity, see Figure 3.18.

The expression patterns of these genes were determined using a panel of twenty different human tissues by RT-PCR. Both genes were found to be ubiquitously expressed, no expression was observed in the liver for *NUDT11*. This could be attributed to tissue specific expression (or repression) or inconsistent cDNA preparations.

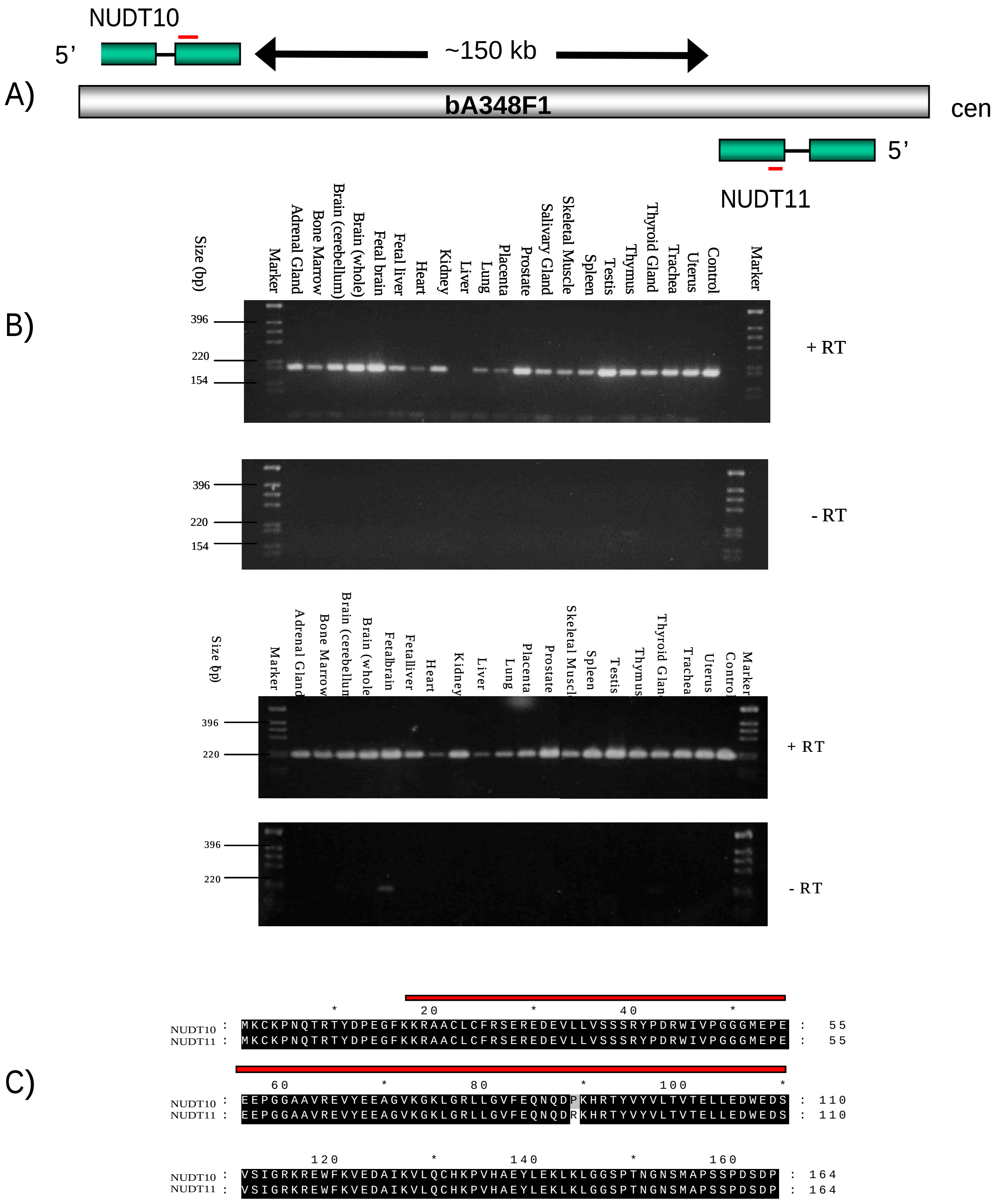

A)
NUDT10
5'
~150 kb
bA348F1
cen
5'
NUDT11
B)
Size (bp)
Marker
Adrenal Gland
Bone Marrow
Brain (cerebellum)
Brain (whole)
Fetal brain
Fetal liver
Heart
Kidney
Liver
Lung
Placenta
Prostate
Salivary Gland
Skeletal Muscle
Spleen
Testis
Thymus
Thyroid Gland
Trachea
Uterus
Control
Marker
396
220
154
+ RT
- RT
C)
NUDT10 : MKCKPNQTRTYDPEGFKKRAACLCFRSEREDEVLLVSSSRYPDRWIVPGGGMEPE : 55
NUDT11 : MKCKPNQTRTYDPEGFKKRAACLCFRSEREDEVLLVSSSRYPDRWIVPGGGMEPE : 55
NUDT10 : EEPGGAAVREVYEEAGVKGKLGRLLGVFEQNQDPKHRTYVYVLTVTELLEDWEDS : 110
NUDT11 : EEPGGAAVREVYEEAGVKGKLGRLLGVFEQNQDRKHRTYVYVLTVTELLEDWEDS : 110
NUDT10 : VSIGRKREWFKVEDAIKVLQCHKPVHAEYLEKLKLGGSPTNGNSMAPSSPDSDP : 164
NUDT11 : VSIGRKREWFKVEDAIKVLQCHKPVHAEYLEKLKLGGSPTNGNSMAPSSPDSDP : 164
20
40
60
80
100
120
140
160

Figure 3.18 Gene duplication in Xp11.23

A) Genomic landscape displaying the orientation and gene structures of *NUDT10* and *NUDT11*. *NUDT10* and *NUDT11* are shown in green, and the approximate location of STSs used in subsequent analysis is shown in red.

B)Gene expression profiling of *NUDT10* and *NUDT11* in 20 different human tissues by RT-PCR. No-RT controls for all 20 tissues are also included. The positive control is genomic DNA while the negative control is $T_{0.1}E$. In the expression studies of *NUDT10* a band that is indicative of genomic contamination can be seen in the foetal brain.

C) Alignment of amino acid sequences for *NUDT10* and *NUDT11*. The NUDEX domain is shown as a red box.

3.7 Discussion

It has been proposed that a complete understanding of the human gene complement could be obtained by sequencing cDNA samples from multiple tissues at multiple stages of development as well as disease states. While this approach produced a wealth of sequence information that has been an invaluable resource for gene identification, it cannot define the entire gene catalogue. cDNA libraries often do not contain full-length transcripts and it is difficult to determine if the entire transcript has been sequenced. cDNA sequencing also only offers a partial view of the human gene complement as these sequences often cannot be put into context with other sequence motifs that govern transcription such as regulatory elements. Nor can they be used alone to understand to identify splicing patterns or other transcript variants. On a larger scale, the cDNA sequences cannot be put into context with the genome landscape to understand gene duplication events or other aspects that may influence their expression and function, such as repeat content.

The ultimate substrate for defining the entire human gene complement is the human genome sequence. The genome sequence was used to delineate the gene complement of Xp11.22-p11.3 as it contains all of the sequence information required for transcription. However, exons sequences only constitute a small proportion (approximately 5%), of the genome sequence and are embedded amongst regions such as introns and repetitive elements. One of the challenges associated with defining the human gene complement has been the identification of functional transcript units and distinguishing them from their surrounding genomic neighbourhood. cDNA sequencing has facilitated this process, as transcripts can be overlaid directly on the genome sequence. Full-length cDNA sequences also provide useful training sets for exon and gene prediction programmes, such as GENSCAN (Burge and Karlin, 1997) and FGenesH (Solovyev *et al.,* 1995) and can be used in concert with protein sequence information, predictive programmes and the human genome sequence, to define the gene content. Work completed in this chapter employed this approach to define the gene complement of human Xp11.22-p11.3. This study annotated 101 genes to varying levels of completion.

Twenty-four partial gene structures were identified that required additional evidence to either confirm or extend their annotation. The method chosen for this was screening vecorette-ligated cDNA samples. This technique was chosen because a wide range of tissues could be studied simultaneously; but more importantly, a primer designed to the vectorette bubble could be used to amplify appropriate cDNA ends. This bypassed the need for 5' and 3' RACE analysis to extent the structure towards 5' and 3' ends. In total, novel sequences were obtained for five novel CDS genes, and three novel transcript genes (see Table 3.5), but expression was confirmed for five more genes.

During this analysis, difficulties were experienced in generating specific PCR products that were suitable for sequencing. For example, screening of the vectorette samples amplified products for three genes but then failed to generate a specific PCR product that could be sequenced. This problem was not solved by changing the primers, nor amplification conditions. The expression of ten genes could not be confirmed using the vectorette-ligated cDNA samples. These transcripts may have gone undetected because cDNA cloning fixes a gene product to a particular time and to particular conditions. The genes also may be unstable or transiently expressed and may not be present in these libraries, or they may simply be expressed in different tissues. Transcription of other novel or putative genes may have been missed due to cloning biases in library construction. It is unlikely that all transcribed genes are be cloned – in particular rare transcripts may not be represented.

Some of these problems may be overcome by using uncloned and/or subtracted cDNA samples or by increasing the number of tissue/cell-types that are sampled. Moreover, experimental evidence and computer gene/exon prediction programmes partially address the challenge of identifying human coding regions. But neither method is sufficient to ensure that all coding regions in the human genome are identified. Alternative methods such as SAGE (Saha *et al.,* 2002) or DNA tiling microarray analysis can also provide additional evidence for gene expression of a predicted coding region (Kampa *et al., 2004*).

Another method that can be used to identify novel human genes is comparative genome analysis. Comparing DNA sequences from different organisms provides a means of identifying common signatures that may have functional significance.

Throughout the course of the analysis described in this chapter mRNA transcripts generated in large scale sequencing projects were submitted to the nucleotide sequence databases and further enhanced the gene catalogue of human Xp11.22-p11.3. The contribution of large scale cDNA sequencing to the human gene catalogue was illustrated by the fact that 85% of the annotated structures were known genes having a complete mRNA sequence. The majority of highly expressed transcripts have now been sequenced, and cDNA sequencing projects have since been refined so that less abundant transcript sequences are acquired (Carninci *et al.,* 2001).

Genome annotation is an iterative process that benefits from re-analysis on new genomic DNA and transcript sequences. For example, in February 2001 the draft version of human genome sequence was interrupted by approximately 150 000 gaps (Lander *et al.,* 2001). By October 2004, this figure was reduced to 341 gaps in finished sequence (IHGSC, 2004). During this time significant changes were also made to the physical map of Xp11.22-p11.3. While none of the 4 gaps in this region were closed, their size was reduced by the introduction of 17 clones into the tilepath. In addition, the orientation of 4 further clones was changed, and a gap was created as the sequence generated from the YAC, yM787F5, was found not to be contiguous with the neighbouring clones.

To coincide with the completion and analysis of the entire human X chromosome the sequence of Xp11.22-p11.3 has recently been re-analysed (Ross *et al.,* 2005). This confirmed Xp11.23 to be the most gene rich region of the X chromosome, which was partly attributed to the expansion of two gene families; the G-antigen (*GAGE*) and the synovial sarcoma (*SSX*) families. The *GAGE* transcripts were annotated in a clone that was not analysed as a part of this project. This is also true for the remaining members of the *SSX* family. A comparison of the transcript annotation presented within this chapter to that completed in the whole chromosome analysis identified an additional 22 transcripts within the same genomic region. These additional transcripts were:

- supported by new cDNA and EST sequences (4 transcripts)
- found in recently finished sequenced clones (4 transcripts)
- only one EST sequence was used as supporting evidence (the analysis required at least two overlapping and splicing EST sequences

- they were not identified in this study (8 transcripts)

Together, these additional transcripts represent 12% of the genes annotated on Xp11.22-p11.3.

Re-analysis of Xp11.22-p11.3 to include new transcript sequences would ensure the resultant gene catalogue is increasingly comprehensive and accurate, and that it adequately represents the current state of transcript sequencing. For example, a second round of annotation of chromosome 22 increased the total length of annotated exons by 74% (Dunham et al., 1999; Collins et al., 2003*)*.

Targeted analysis will also be required to reach the 5' and 3' ends of incomplete genes. There is increasing evidence to suggest that 5′ and 3′ UTR regions of transcripts are important for translational regulation, transcript stability, and subcellular targeting (reviewed by Mignone *et al.,* 2002; Kuersten and Goodwin 2003). They are difficult to identify because they do not show the characteristic sequence bias of coding regions and they are less conserved across species. The 3' end of genes are often well represented in cDNA and EST libraries and are often more well represented in annotated gene structures. However, cDNA sequences frequently do not extend to a TSS. To address this shortfall technologies such as LongSAGE together with a 5' TSS library have been developed. Using these technologies, accumulation of the TSS data in a high throughput manner will be possible without degrading the data quality (Shiroki *et al.*, 2003; Hashimoto *et al.*, 2004). A direct outcome of high-throughput TSS and TTS mapping would be the identification of many alternative TSSs and TTSs, and therefore, the identification of novel 5′ and 3′ UTR regions of known transcripts. Moreover, dense mapping of TSSs on chromosomes would also help to provide quantitative measurements of differential TSS use and aid in the identification of putative promoter regions.

Complete genome annotation relies on comprehensive transcriptome characterisation. Apart from the genes that might be expressed in a cell, the additional complexity of a transcriptome is mostly created by three major mechanisms, namely alternative transcription initiation, alternative splicing, and alternative polyadenylation. In particular, one of these features was highlighted by the analysis completed in this chapter - the complexity of the transcriptome generated by alternative splicing. A high degree of variation was observed in the splicing patterns of the annotated genes. Transcript variants can arise as a result of highly regulated splicing events, and they may serve to enhance the diversity of the proteome that is create more functional products from a limited number of loci. Alternatively, transcript variants may be the result of mis-splicing events, and may not serve any function in the cell. As a preliminary measure to further investigate the diversity of the transcriptome it was decided to assess the presence of alternative transcripts in Xp11.23 in greater detail. The data generated within this chapter provided the basis for the experimental work described in the following chapters.

Annotated genome sequences provide a useful resource for a variety of genetic studies. In addition to providing the framework for the remainder of the work completed in this thesis, it is hoped that detailed annotation of human Xp11.22-p11.3 could provide a useful resource for future evolutionary and disease association studies. For example, association studies have implicated Xp11.23 with several X-linked mental retardation phenotypes. Now that this region of the genome has been annotated it would be possible to screen the exons of genes found this region for SNPs and other sequence variations that may have pathological consequences.