Chapter 4

Identification of alternative transcripts in human Xp11.23

4.1 Introduction

In the process of annotating genes in human Xp11.22-p11.3 a high level of transcript diversity was observed. EST and mRNA data suggested that approximately 70% of the genes within this region were alternatively spliced, which was substantially higher than the then current predicted frequency for the entire human genome (25-59%) (Modrek and Lee 2002). This prompted further investigation to determine the full extent of transcript diversity for some of the genes annotated within human Xp11.22-p11.3.

The experimental strategy used to achieve this was a targeted PCR, cloning and sequencing protocol where gene fragments were amplified from a cDNA panel composed of 29 human tissues (Figure 4.1). The resulting sequences were aligned to genome sequence to identify additional transcript variants. The strategy was chosen because it does not require prior knowledge of the internal splicing patterns to generate an expression profile of transcript variants, but it still allows RNA samples from multiple tissues to be sampled in concert. By assaying fragments of genes, small amplicons were produced and this permitted the use of high-throughput sequencing strategies to generate the sequencing information. Novel transcripts could also be easily identified by comparing the size of amplified cDNA fragments to predicted amplicon sizes that were determined from existing cDNA and EST information. These could then be targeted for sub-cloning and sequence analysis.

A second part of the strategy was to find evidence for additional variants using transcript information from another organism. Transcript sequences and splicing patterns of mouse orthologues (*M. musculus*) were compared to their human counterparts. Any mouse-specific splicing patterns could be targeted for analysis using the PCR and sequencing strategy outlined above.

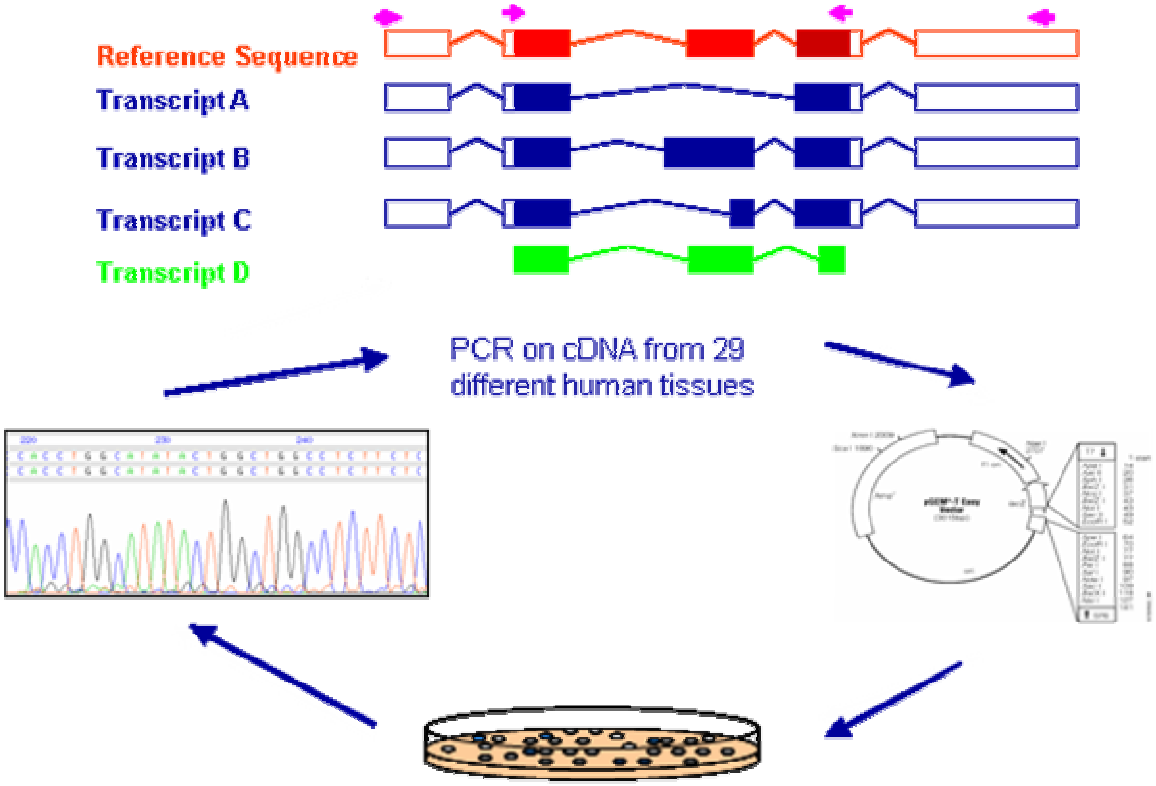


Figure 4.1 Experimental strategy to identify novel transcripts in human Xp11.23
Gene fragments are amplified from a panel of 29 human tissues (primers shown as pink arrows). The PCR products are then cloned into a holding vector and sequenced. The resulting sequences are then aligned against the genome sequence and additional alternative variants are identified.

In order to ensure that a detailed analysis of possible alternative transcripts was obtained, it was decided to focus on a subset of the protein-coding genes from the gene catalogue of human Xp11.22-p11.3. Xp11.23, a region of significant biological interest, was selected as it contains approximately 10% of all X chromosome genes in less than 4% of the X chromosome's DNA sequence (Ross *et al.*, 2005) and it is a clinically relevant region of X chromosome. The analysis focused on a genomic region flanked by the markers *DXS6491* and *DXS9784*.

A region of approximately 600 kb containing 18 protein-coding genes of contiguous genomic sequence was selected for further analysis. Four of these genes are associated with pathological phenotypes, Wiskott-Aldrich (*WAS)*, GATA binding protein 1 (*GATA1)*, proviral integration site 2 (*PIM2)* and polyglutamine binding protein 1, (*PQBP1)*. Two of the genes included have no known function (AC115617.1 and AF207550.5) while the function of the remaining genes has been

determined (10 genes) or inferred (six genes). The genes are compact - their average length is 12,417 bp, compared to the average for the entire X chromosome, 25,226 bp. They also exhibit a wide range of splicing complexity with the number of exons per gene ranging from 1, ES cell expressed Ras (*ERAS*), to 29, histone deacetylase 6 (*HDAC6*). Further information about the 18 genes included in this study is listed in Table 4.1.

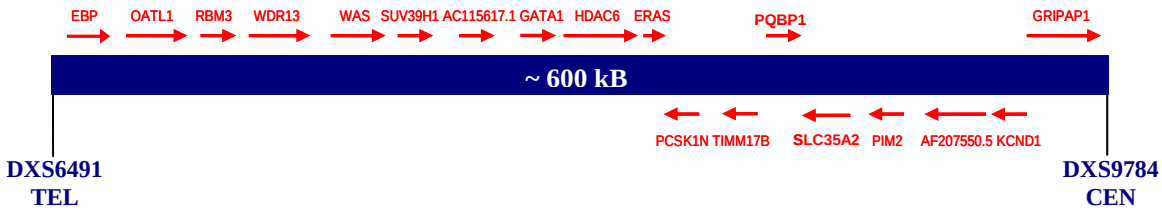


Figure 4.2 Order and orientation of genes involved in this study

Prior to commencing this study, there was a paucity of information available about transcript variation in human Xp11.23. Thirteen transcript variants had been identified for four genes in this region: *RBM3* (3 variants) (Derry *et al.,* 1995), *WAS* (1 variant) (Derry *et al.,* 1995), *PQBP1* (*4* variants) (Iwamoto *et al.,* 2000) and *SLC35A2* (4 variants) (Ishida *et al.,* 1996). It was therefore anticipated that the information created from this study could not only be a model dataset for more comprehensive alternative splicing analysis (e.g. understanding the underlying mechanisms of alternative splicing) but would also contribute valuable information towards understanding the function of each of the genes included in the study.

Table 4.1 Number of alternative transcripts annotated for each gene between the markers between markers *DXS6941* and *DXS9784*

Gene	Exons	Transcripts	Function GO classification	Published transcript variants	Disease (OMIM)
EBP	5	5	* Cholesterol delta-isomerase activity * Drug transporter activity * Isomerase activity		X-linked dominant chrondrodysplasia punctata (OMIM: 300205)
OATL1	6	2	* Transaminase activity * Transferase activity		
RBM3	7	10	* RNA binding * Nucleic acid binding * Nucleotide binding	Variants described by Derry *et al.*, 1995	
WDR13	10	10	* No known annotation		
WAS	12	2	* Small GTPase regulator activity	Alternative splice patterns associated with WAS	Wiskott-Aldrich syndrome (OMIM: 300392)
SUV39H1	6	3	* S-adenosylmethionine-dependent methyltransferase activity * Chromatin binding *Histone-lysine N-methyl transferase activity * Protein binding *Transferase activity *Zinc iron binding		
AC115617.1	4	1	* Novel		
GATA1	6	2	* Metal ion binding * Transcription factor activity		X-linked dyserythopoietic anaemia (OMIM:305371)
HDAC6	29	4	*Actin binding *Histone deacetylase binding		

			*Hydrolase activity *Specific transcriptional repressor activity *Zinc ion binding		
ERAS	1	1	* GTP binding		
PCSK1N	3	1	* Endopeptidase inhibitor activity * Receptor binding		
TIMM17B	5	1	* Protein translocase activity		
PQBP1	6	7	* DNA binding * Transcription co-activator activity	Four transcript variants described by Iwamoto *et al.*, 2000	X-linked mental retardation (OMIM:300463)
SLC35A2	5	4	* UDP-galactose transporter activity * Nucleotide-sugar transporter activity * Sugar porter activity	Described by Ishida *et al.*, 1996	
PIM2	5	1	* ATP binding * Protein serine/threonine kinase activity		
AF207550.5	7	2	* Novel		
KCND1	3	1	* Voltage-gated potassium channel activity		
GRIPAP1	29	2	* GTP binding * RNA binding		

Results

4.2 Using mouse transcript and genome information to identify additional alternative transcripts

Comparative analysis between the human and mouse was carried out with aim of using genomic and transcript information from the mouse to enhance the annotation of the 18 human genes targeted for detailed analysis. This was achieved by first identifying orthologous genes and the corresponding genomic BAC clones in the mouse using BLAST analysis. These clones were mapped and sequenced as a part of the mouse genome project (Gregory *et al.,* 2002; Waterston *et al.,* 2002). All genes (and transcript variants) were annotated in the mouse and novel exon junctions were identified by comparing the annotated structures of the mouse genes to their human counterparts.

4.2.1 Identification of mouse orthologues

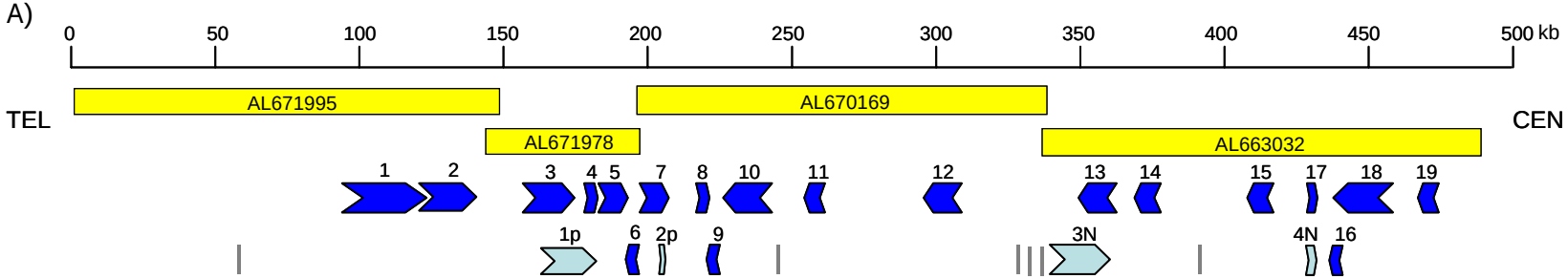
Orthologous genes in mouse for each of the 18 human genes included in this study were identified by BLASTp analysis performed at the National Center for Biotechnology Information (NCBI). Potential orthologues were identified for all 18 genes, and the nucleotide and amino acid sequences were aligned using ClustalW (Pearson 1990). The average amino acid identity between orthologous gene pairs was, 90.1%, while the average nucleotide identity was 85.6% (Table 4.2). In addition, the Ka/Ks ratio (an indicator of evolutionary selective pressure) was determined for each of 18 orthologous gene pairs (http://www.bioinfo.no/tools/kaks, Table 4.2). These ranged from 0.009 (*WDR13-Wdr13*) to 0.234 (*HDAC6-Hdac6*) and indicate that the gene pairs are under selective pressure to remain conserved throughout evolution. Subsequent analysis revealed that the gene order was identical to that recorded in the human (Figure 4.2).

Table 4.2 Sequence identity of human and mouse orthologues.

Human gene	Mouse orthologue	Protein identity (%)	Nucleotide identity (%)	Ka/Ks
EBP	*Ebp*	80.0	77.8	0.206
OATL1	*Oatl1*	93.9	87.1	0.047
RBM3	*Rbm3*	94.8	88.2	0.038
WDR13	*Wdr13*	99.2	90.3	0.009
WAS	*Was*	89.6	85.5	0.100
SUV39H1	*Suv39h1*	95.4	88.0	0.033
AC115617.1	NM_027227	83.0	84.2	0.160
GATA1	*Gata1*	86.0	83.3	0.114
HDAC6	*Hdac6*	81.1	81.1	0.234
ERAS	*Eras*	78.4	75.8	0.213
PCSK1N	*Pcsk1n*	83.7	81.9	0.150
TIMM17B	*Timm17b*	95.9	89.3	0.036
PQBP1	*Pqbp1*	88.2	84.7	0.086
SLC35A2	*Slc35a2*	96.6	90.4	0.054
PIM2	*Ppim2*	89.4	86.3	0.109
AF207550.5	DXImx46e	98.2	90.4	0.015
KCND1	*Kcnd1*	95.1	87.4	0.039
GRIPAP1	*Gripap1*	93.7	89.0	0.058

BLAST analysis of the mouse gene sequences to the mouse genome sequence suggested that all 18 genes were contained within four BAC clones. Their EMBL accession numbers (and clone names) are: AL671995 (RP23-109E24), AL671978 (RP23-443E19), AL670169 (RP23-198C2), and AL663032 (RP23-27I6). The WTSI Informatics team analysed the clone sequences in accordance with the approach described in chapter 3 and the clones were also annotated as described in chapter 3. In total, 22 genes were annotated, 19 of which were known genes, as well as one novel transcript and two putative genes (Figure 4.3). Five pseudogenes were also annotated. One hundred and three transcript structures were annotated and the average number of transcripts identified for each gene was 4.6. A transcript map of the four analysed clones is displayed in Figure 4.3.

A)



B)

Known Genes				Putative/Novel Genes		Pseudogenes	
No.	Gene	No.	Gene	No.	Gene		Cen → Tel
1	*Gripap1*	11	*Gata1*	1p	Rp23-109E24.8	1	Rp23-198C2.8
2	*Kcnd1*	12	NM_027227	2p	Rp23-198C2.2	2	Rp23-198C2.6
3	DXImx46e	13	*Suv39h1*	3N	Rp23-27I6.2	3	Rp23-198C2.9
4	*Pim2*	14	*Was*			4	Timm17p2
5	*Slc35a2*	15	*Wdr13*			5	Rp23-27I6.4
6	*Pqbp1*	16	*Rbm3*				
7	*Timm17b*	17	2900002K06				
8	*Pcsk1n*	18	*Oatl1*				
9	*Eras*	19	*Ebp*				
10	*Hdac6*						

C)

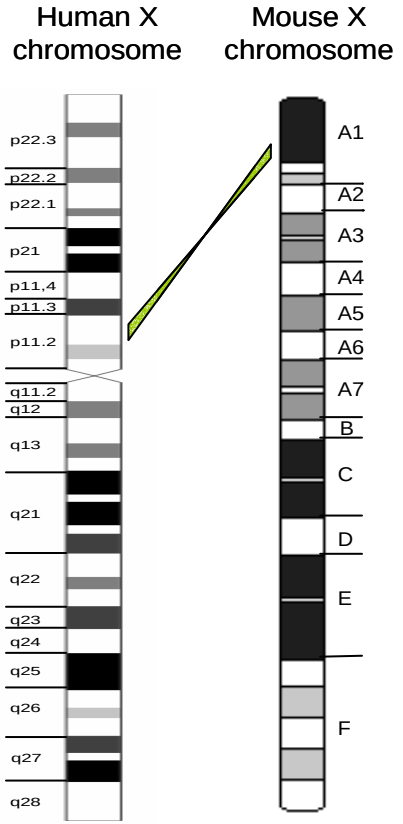


Figure 4.3 Annotation of mouse orthologues

A) The location and orientation of known genes (blue) and novel_transcripts and putative genes (light blue) and pseudogenes (grey) in the genome landscape is displayed. Also displayed are the sequenced clones (yellow). B) List of annotated genes. C) Putative location of orthologous regions on human and mouse X chromosomes.

4.2.2 Features of interest in the mouse genome

Annotation of the mouse genome sequence also identified two mouse specific loci that had not been identified or confirmed in the human. In the mouse, antisense transcripts were annotated for 3 genes, *Rbm3*, *Wdr13*, *Suv39h1*. Transcripts antisense to human *RBM3* and *SUV39H1* had already been identified, but the expression of a transcript antisense to *WDR13* remains to be confirmed. BLAST analysis of the sequence from the novel antisense exon failed to identify any region of shared homology between the two species (Figure 4.4).

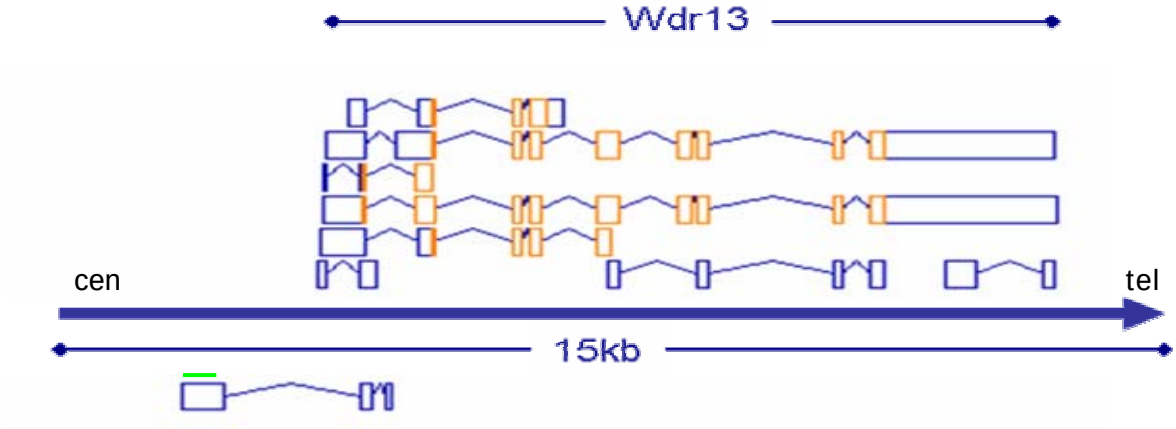


Figure 4.4 Annotation of an antisense transcript to *Wdr13* in the mouse.
The genome sequence is displayed from centromere to telomere. Annotated transcripts on the genome sequence are displayed – orange CDS, blue non-coding. Gene structures displayed above the genome sequence are transcribed from left to right while gene structures below the genome sequence are transcribed in the opposite orientation. The sequence used in the BLAST analysis is displayed by a green line.

In addition, two EST sequences (BM949497 and BQ714158) that spanned two genes, *Pim2* and DXImx4e were observed in mouse (Figure 4.5). In an attempt to identify these transcripts in human, cDNA samples from 29 different human tissues (Section 2.26) were screened using primer pairs designed to amplify the fusion transcript only with one primer located in each gene (487045A/487026S and 487045A/487054S). These screens failed to generate any evidence to support its expression of this composite transcript in any of the tissues sampled. Fusion transcripts have also been observed in other mouse genes (I. Barrett, personal communication) but it remains to be determined if they are functional variants. *Pim2* and DXImx46e are in close proximity to each other (intergenic region of 3,452 bp) and it is therefore possible that *RNA pol II* failed to dissociate from the

genomic sequence during transcription and managed to transcribe both of these genes in one round of transcription.

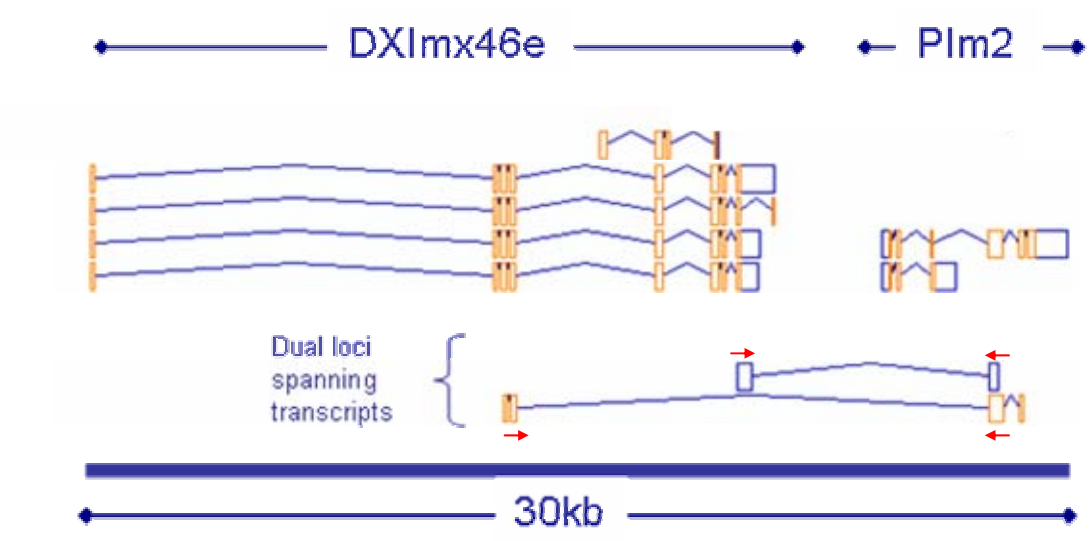


Figure 4.5 Annotation of EST sequences spanning *PIM2* and DXIm46e
The genome sequence is displayed from the centromere to the telomere. Annotated transcripts on the genome sequence are displayed – orange cds, blue non-coding. Gene structures displayed above the genome sequence are transcribed from left to right. Primer pairs (487045A/487026S and 487045A/487054S) used for the experimental confirmation of these transcripts are displayed by red arrows.

4.2.3 Comparison of human and mouse transcript variants

The 18 orthologous gene pairs were first assessed for their transcript complexity by comparing the number of transcript variants for each gene pair (Figure 4.6). Five gene pairs shared the same number of annotated transcripts; six genes had fewer transcript variants in the mouse, while seven genes had more transcripts in the mouse. The transcript complexity increased substantially in the mouse for the genes *RBM3* (14:10, mouse:human) and *HDAC6* (9:4). These results suggest that either the transcript coverage or splicing complexity is greater in mouse than in human. However, it is expected that the transcript coverage is greater in the mouse as genome sequence was annotated approximately 12 months after the human sequence had been annotated. Thus, the mouse annotation may have benefited from the addition of transcript information into the sequence databases.

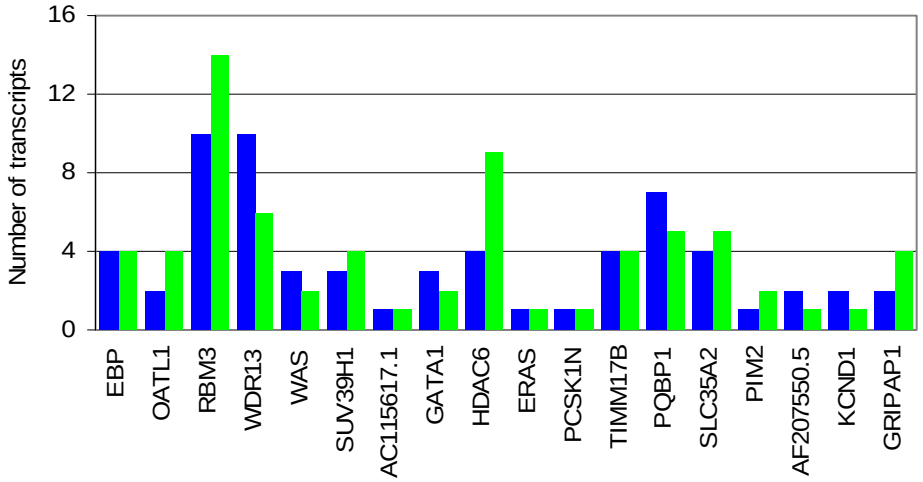


Figure 4.6 Number of alternative transcripts that were annotated for each of the gene pairs.
Human (blue), mouse (green).

Alternative splicing events are often conserved between human and mouse transcripts (Thanaraj and Stamm 2003). The 103 mouse transcript variants identified in section 4.2.1 reflected a detailed coverage of mouse genes in transcript databases, and the proposal here is that these would be a suitable resource for identifying additional human transcript variants. The splicing patterns of human and mouse orthologous gene pairs were compared using sequence tags derived from neighbouring exons. The tags were composed of 20 bp of exonic sequence from neighbouring exons (ten bp from each exon) and were used to search a list of tags from the orthologous gene. A sequence identity cut-off of 80% between the matching exon tags was imposed when searching for conserved exon junctions. This figure was chosen because it is close to the average nucleotide identify of the orthologous each gene pairs, 85.6% (Table 4.2). Analysis was completed using both human and mouse tags so that human and mouse specific exon junctions and shared exon junctions could be identified. These are listed for each gene pair in listed in Table 4.3. In addition, the numbers of species specific first and terminal exons were determined for each gene.

The reference transcript structures were the same for 15 of the 18 orthologous gene pairs. The exceptions were *ERAS*:*Eras*, *HDAC6*:*Hdac6* and AF207550.5:DXImx46e which had different gene structures. Figure 4.7 illustrates differences in the transcript structures of the *ERAS*:*Eras* gene pair. Here, the 5'

UTR of *Eras* (mouse) spanned two exons while the entire *ERAS* (human) transcript was contained within one exon.

Table 4.3 Analysis of the conservation of alternative exon junctions and first and last exons in human and mouse genes

Species specific and shared alternative splicing patterns, and novel or extended first and last exons were identified using the annotated transcripts for 18 orthologous gene pairs. Hs = human; Mm = mouse, Both = event shared by human and mouse.

Gene	Exon junction			First exon			Last exon		
	Hs	Mm	Both	Hs	Mm	Both	Hs	Mm	Both
EBP	3	1	0	0	1	0	0	0	0
OATL1	0	1	1	0	1	0	0	0	0
RBM3	1	1	5	0	1	2	0	1	3
WDR13	2	1	4	0	0	0	1	0	0
WAS	1	0	0	0	0	0	0	1	0
SUV39H1	2	2	0	1	1	0	0	1	0
AC115617.1	0	0	0	0	0	0	0	0	0
GATA1	2	0	0	0	0	0	0	1	0
HDAC6	0	1	4	0	2	1	0	0	1
ERAS	0	1	0	1	1	0	0	0	0
PCSK1N	1	0	0	1	0	0	0	0	0
TIMM17B	4	3	0	0	1	0	0	0	0
PQBP1	3	1	1	0	1	0	0	0	0
SLC35A2	0	0	3	1	0	0	0	0	0
PIM2	0	0	0	0	0	0	0	1	0
AF207550.5	3	2	0	0	0	0	0	1	0
KCND1	1	0	0	1	0	0	0	0	0
GRIPAP1	0	4	0	0	3	0	0	3	0
Total	**23**	**18**	**18**	**5**	**12**	**3**	**1**	**9**	**4**

A total of 59 alternative exon junctions were identified and 31% (18/59) of these were shared between the two species. This figure is lower than previous studies (Thanaraj and Stamm, 2003) and may reflect the need for additional transcript evidence in the two species. More alternative first and last exons were annotated in the mouse transcripts. Approximately 58% (11/19) of all novel or extended first exons were specific to mouse transcripts and 69% (9/13) of all novel or extended 3' exons were identified in the mouse. An example of an extended first exon that was only observed in mouse transcripts is shown for the *EBP*:*Ebp* gene pair (Figure 4.7).

The alternative splicing events are not conserved between the human and mouse and it is possible that mouse specific splicing events may suggest that these are

additional unidentified human transcript variants (and vice versa). Therefore, the decision was made to target mouse-specific splicing events for incorporation into future human transcript profiling assays. The dataset of mouse specific transcripts was filtered several times before novel transcripts were selected for further analysis. Internal mouse-specific exon junctions were removed from the dataset as it was assumed that they would be identified in the detailed overlapping cDNA screening process (section 4.3). Subsequently, homology searches were carried against the human genome sequence using the sequence that harboured a novel mouse specific exon (BLAST analysis –Altschul *et al.,* 1990). A positive identification was further pursued, where primers were designed to the orthologous region in the human. Twenty-one novel mouse exons were aligned to the human genome sequence and thirteen were found to be conserved in the human. These exons were included in all subsequent analyses (section 4.3) and are listed in Table 4.4.

Table 4.4 Mouse specific exons that are conserved in human

Gene	5' or 3' exon	Mouse sequence (EMBL accession)
EBP	5'	BC004703
RBM3	3'	AK049575
WDR13	3'	BU936044
WAS	5'	BF537073
SUV39H1	5'	CA885821
GATA1	5'	BY227941
HDAC6	5'	BF141098
PQBP1	5' 5'	AV097864 AK010658
PIM2	3' Gene fusion to DXmxi46e	 BM949497 BQ714158
AF207550.5	3'	BQ444253
GRIPAP1	5'	BY100293

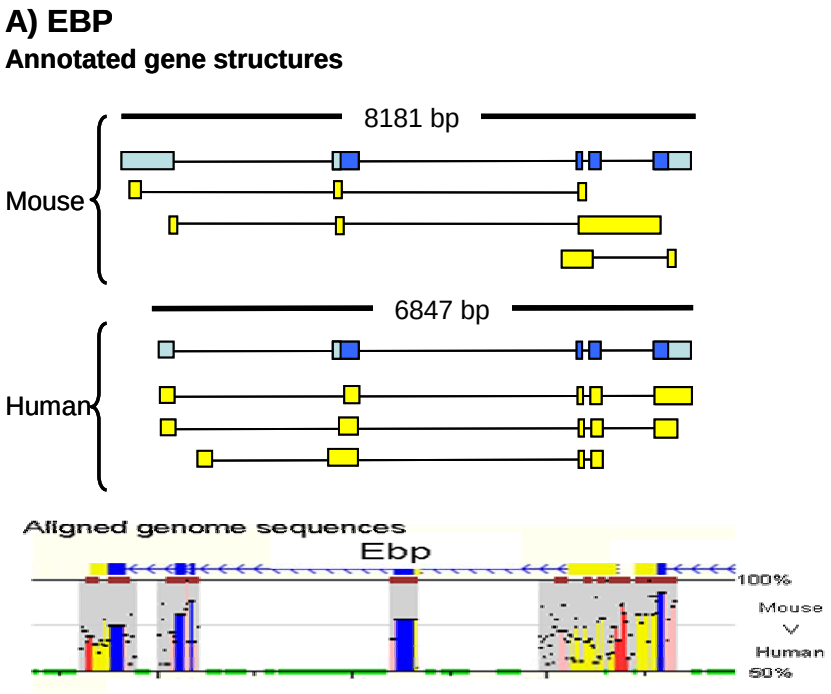


Figure 4.7 Identification of novel transcribed regions using comparative analysis

Annotated gene structures.
Transcripts are aligned to the genome sequence and are displayed 5' to 3' (left to right). Structures annotated in blue are reference sequences for each gene; dark blue exons are coding and light blue exons are UTR. Transcript variants are displayed in yellow. The size of each gene is displayed above the reference sequence.

Aligned genome sequences.
Genome sequences of the annotated genes (and 2 kb flanking sequence) were aligned using zPicture which identified evolutionarily conserved regions of at least 70% identity and 100 bp in length. The mouse genome sequence runs from left to right along the X axis. The conservation between the two species is displayed as a black bar where the height of the band represents the sequence homology and the length of the homologous region is also displayed. Mouse gene structures are annotated on top of the graphs. Exons (blue), UTR (yellow), intragenic (pink) and intergenic (red) regions are also highlighted. Green bars are repeats.
A) Comparative analysis of *EBP* confirms shared homology in the 5' UTR and upstream from exon 2.
B) Comparative analysis of *ERAS* highlights the lack of conservation of a mouse specific exon.

4.3 Identification of novel transcripts for human Xp11.23 by cDNA screening and sequencing

Annotation of existing expressed human and mouse transcripts has provided evidence of substantial transcript variation in human Xp11.23. These approaches enhanced the current description of transcription variation for 18 protein coding genes. However, it is unlikely that all cDNA libraries used in EST sequencing projects have been exhaustively sequenced and that all transcript information for these genes has been obtained. In an attempt to enhance further the description of transcript variants in human Xp11.23 an experimental approach of targeted PCR and sequencing was employed. This method is discussed below.

4.3.1 Primer Design

One hundred and forty-one primer pairs were designed to identify alternative transcripts from the 18 genes. To ensure amplification of as many alternatively spliced transcripts primers as possible, primers were designed to overlapping regions of each gene. Primers were designed to both reference and alternative transcript sequences to generate a PCR product of at least a 100 bp and to span at least two exon junctions. Regions of genes that displayed a high level of transcript complexity, such as *RBM3*, *PQBP1* and *HDAC6* were screened in greater detail. Here, primers were designed so that no more than three known transcript variants were amplified per PCR amplification. For example, 12 primer pairs were designed to *RBM3* to ensure that its highly variable region, which spanned exons 3-6, was adequately surveyed. All primers are listed in Appendix II, while the primer combinations and their predicted amplicon sizes are listed in Appendix III. The smallest predicted amplicon was 100 bp in length, while the largest was 3,197 bp in length.

4.3.2 Optimisation of PCR conditions

PCR conditions were optimised to ensure that a wide range of transcript sizes could be amplified in one reaction. Four different reaction conditions were tested on four different primer pairs that amplified a range of differently sized cDNA transcripts. Optimal PCR conditions used a combination of Amplitaq and Advantage *Taq* polymerases to amplify cDNA fragments. The PCR cycling conditions were also optimised for each primer pair using three different annealing temperatures (55°C, 60°C and 65°C) and cDNA synthesised from the brain, liver or lung total RNA. These tissues were chosen as they expressed between them the majority of genes

analysed in this study (14/18, determined using UniGene expression profiles, www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene). Negative pre-screen results could be attributed to failed primer design or they could also indicate that the cDNA fragment was not expressed in the three cDNA samples used. When a PCR product was not generated in the pre-screen process an annealing temperature of 60°C was used to screen the larger cDNA panel.

4.3.3 cDNA screening

cDNA was synthesised from total RNA from 29 different human tissues (using oligo-dT primers as outlined in section 2.13.3). All samples were confirmed to be free from genomic contamination (Ian Barrett, personal communication).

In total, 109 screens were performed on the panel of 29 cDNA samples, 93 of which were successful. The amplification of transcript variants was first monitored by comparing the experimentally generated PCR banding patterns to the banding pattern that was predicted from the gene's known transcript variants. Successful amplification of an appropriately sized amplicon during the cDNA screen suggested that a known transcript variant was expressed. The presence of additional PCR bands in a reaction may represent novel transcript variants, or non-specific PCR products. When an amplicon was expected but not observed, it was supposed that the corresponding transcript variant may not be expressed in the tissues studied or that reaction conditions were not appropriate to amplify the transcript.

The primers were redesigned for all failed reactions and the cDNA screening process was repeated. For successful reactions, a minimum of three samples deemed to represent all of the observed products in the cDNA screen were selected for further analysis. This selection step was introduced to reduce the redundancy encountered when processing multiple samples with the same expression profiles.

4.3.4 Cloning and sequencing

The selected PCR samples were purified (section 2.14.1) and their 3' end was adenylated (section 2.19.1) to facilitate ligation with the plasmid pGEM-T Easy (Promega) which has a 3' T overhang. Ligated plasmids and gene-fragments were introduced into JM109 cells by chemical transformation (section 2.17). In total, 344 ligations were carried out (the number of ligations performed for each gene listed in Table 4.5). For each ligation, thirty-two white colonies were tested by PCR

for the successful subcloning of fragments using the M13F and M13R primers that are located within the pGEM-T Easy vector. Sequences of primers M13F and M13R are listed in Appendix II. Clones of unique size were selected for sequencing, which was completed by the Research and Development team at the WTSI. Table 4.5 summarises the different phases of the project for each of the 18 genes.

The resulting sequences were processed prior to assessment for novel splicing patterns. The flanking vector sequence was first removed from the output files using an in house perl programme sccd2ace (C. Scott) or manually using Gap4 (section 2.26.4). The clipped sequences were then aligned to the genomic sequence together with the appropriate reference and variant cDNA and EST sequences using Spidey (http://www.ncbi.nlm.nih.gov/spidey). The fidelity of each sequence was assessed manually, and transcript sequences with greater than 95% sequence identity to the human genome reference sequence were used to identify variant transcripts. Variations in transcript structure were identified by comparing the exon co-ordinates of the sequenced transcripts to those of the transcripts that had already been annotated.

Table 4.5 Summary of cDNA screening, ligations and sequencing reactions for each gene analysed in this study

Human Gene	Primers designed	cDNA screens	Ligations	Sequences
EBP	8	4	18	11
OATL1	20	5	18	27
RBM3	24	12	36	103
WDR13	18	6	12	76
WAS	16	6	21	50
SUV39H1	12	4	33	16
AC115617.1	6	6	14	136
GATA1	12	3	9	26
HDAC6	32	14	57	90
ERAS	6	2	None	None
PCSK1N	10	5	None	None
TIMM17B	14	5	3	57
PQBP1	24	14	57	74
SLC35A2	20	8	6	39
PIM2	10	2	9	15
NOVEL_2	16	3	15	44
KCND1	22	4	21	52
GRIPAP1	12	6	15	93
Total	**282**	**109**	**344**	**909**

A detailed description of the results for the two genes, *RBM3* and *PQBP1* is given below.

4.3.5 Identification of novel RBM3 transcripts

At the time of this analysis, ten transcript variants were annotated from existing cDNA and EST sequences for *RBM3*. The reference transcript is composed of seven exons and spans approximately 6,200 base pairs. 0 shows the various transcript structures for *RBM3* and highlights its high degree of transcript diversity. In addition to internal splice variations, two alternative first exons (*RBM3*.2 and *RBM3*.10) and two alternative final exons (*RBM3*.7 and *RBM3*.8) have been identified. An additional final exon was also identified in the mouse transcripts. To ensure that the expression of all known transcript variants was assessed and any additional novel transcripts were detected, twelve cDNA screens using different primer combinations PCR were completed. The primers and targeted transcripts are listed in Figure 4.8. Eleven of the twelve cDNA screens were successful in amplifying at least one transcript of the expected size, screen one was unsuccessful (Figure 4.8). Additional unexpected PCR bands were identified in eight out of 12 of the cDNA screens, which may represent novel transcripts.

Following the cDNA screens, 36 samples were selected for further analysis; three samples were selected for all screens except screen 1 (no samples selected) and screen 5 (six samples selected). These PCR products were cloned according to the protocol outlined in section 2.17. Here, 1,152 individual white colonies were tested for the presence of bands observed in the initial cDNA screening analysis (results not shown). These are displayed in Figure 4.9, and show that additional clones were obtained for four cDNA screens. These may have been missed from previous observations as their appearance may have been masked by more abundant PCR products in the cDNA screens. The predicted number of clones was generated for four cDNA screens while a lower than expected product number was observed for three cDNA screens. Approximately 10% (103) of the screened colonies were selected for sequencing from which eight novel splicing variants were identified by aligning the processed transcript sequences to the genome sequence of clone AC115618 in Spidey (www.ncbi.nlm.nih.gov/spidey/). All RBM3 transcript variants are listed in Table 4.6.

A)

Screen	Primer 1		Primer 2		Targeted transcript
1	A	486772S	C'	486774A	All except *RBM3*.2, *RBM3*.9, *RBM3*.10
2	B	486773S	C'	486774A	*RBM3*.9
3	D	486775S	E'	486776A	*RBM3*, *RBM3*.3, *RBM3*.5, *RBM3*.8
4	D	486775S	F'	486777A	*RBM3*, *RBM3*.3, *RBM3*.8, *RBM3*.9
5	C	486774S	F'	486777A	*RBM3*.8
6	C	486774S	E'	486776A	*RBM3*, *RBM3*.3, *RBM3*.8, *RBM3*.9
7	K'	457763A	H'	486779A	*RBM3*.2, *RBM3*.10
8	F	486777S	H'	486779A	*RBM3*, *RBM3*.2, *RBM3*.6, *RBM3*.9, *RBM3*.10
9	H	486779S	I'	486780A	*RBM3*, *RBM3*.2, *RBM3*.4, *RBM3*.6, *RBM3*.9
10	H	486779S	J'	486781A	*RBM3*, *RBM3*.2
11	G	486778S	J'	486781A	*RBM3*, *RBM3*.2, *RBM3*.4, *RBM3*.6, *RBM3*.9
12*	L	498796A	L'	498794S	Homologous mouse exons

B)

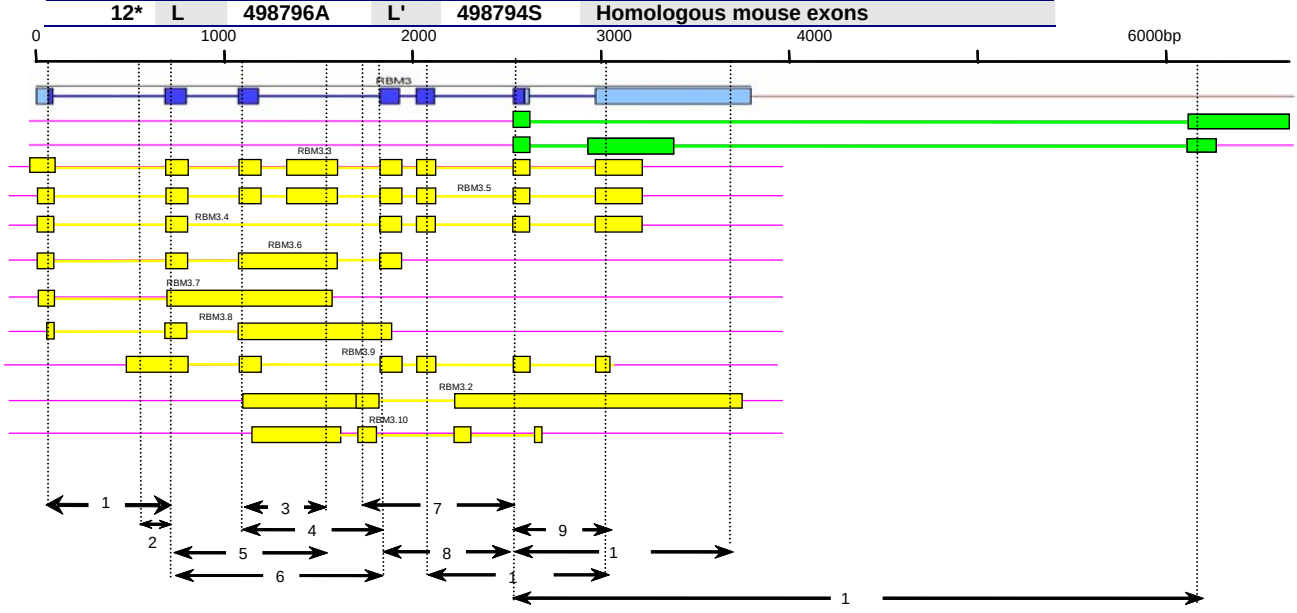


Figure 4.8 Transcript profiling of *RBM3* in 29 different tissues

A) List of primer combinations and *RBM3* transcript variants that should be amplified are listed. These primers were used to screen the expression of *RBM3* in 29 different tissues.

B) Exon/intron structures of *RBM3* transcript variants identified from existing cDNA and EST sequences. The transcripts run 5' to 3' (left to right). The reference transcript for *RBM3* is displayed with coding exons (dark blue) and UTR (light blue). The exon/intron structure of alternative variants is shown in yellow and mouse specific transcripts are displayed in green. Overlaid on this diagram is the location of primers. The regions of transcripts amplified in each cDNA screen outlined and numbered in accordance with (A).

Note transcripts *RBM3.3* and *RBM3.4*, and *RBM3.5* and *RBM3.6* are displayed on the same line.

A)

Primer combination	Product size						
1	566	237					
2	228						
3	480	329	219	800*	1000*	1800*	
4	646	425	318	156	900*		
5	655	508	401	239	132	900*	1100*
6	840	562	416	309			
7	323						
8	233	400*	550*				
9	654	299	550*				
10	1096	733	1150*				
11	1155	800	1600*	2000*			
12	1104	1000	1150*				

B)

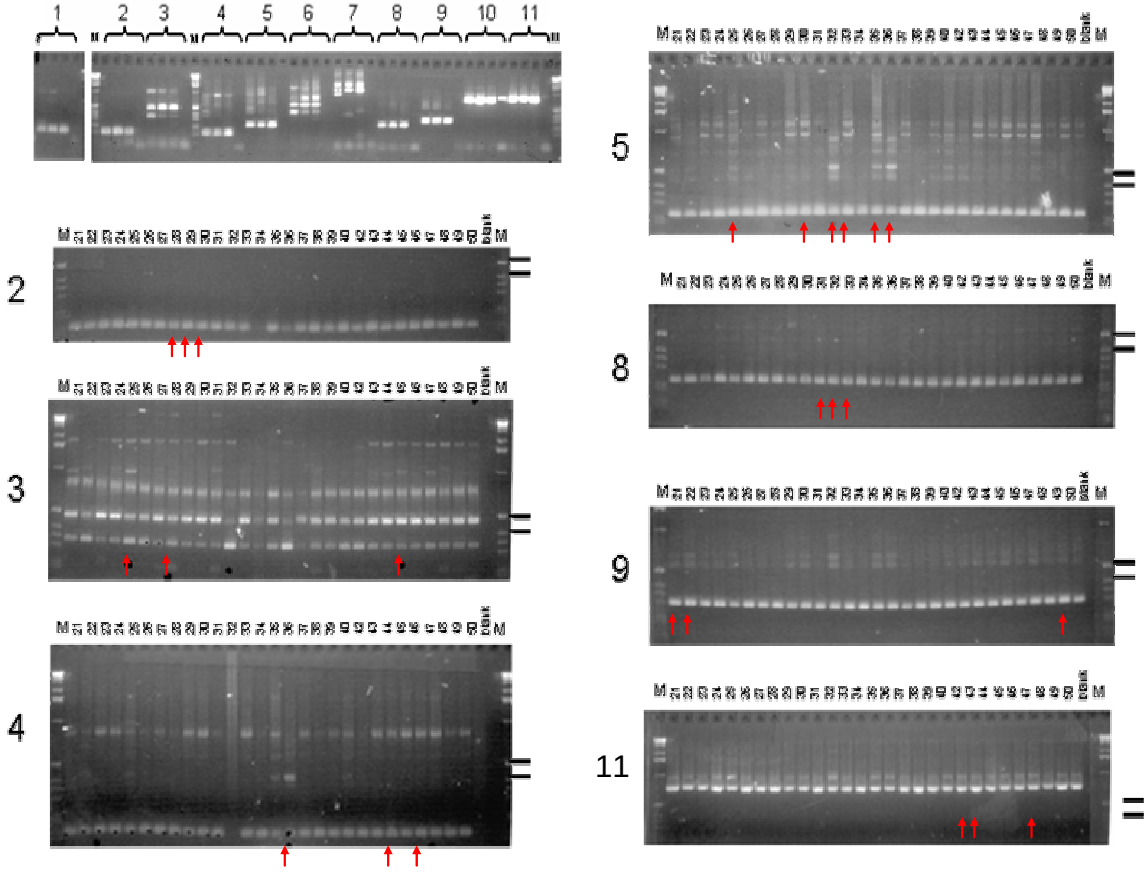
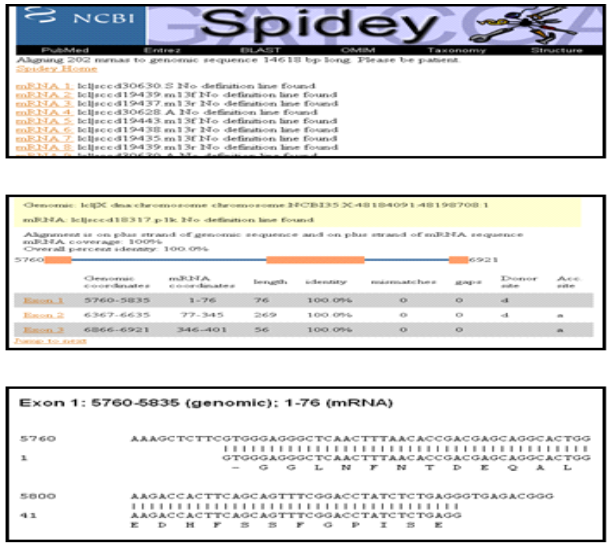


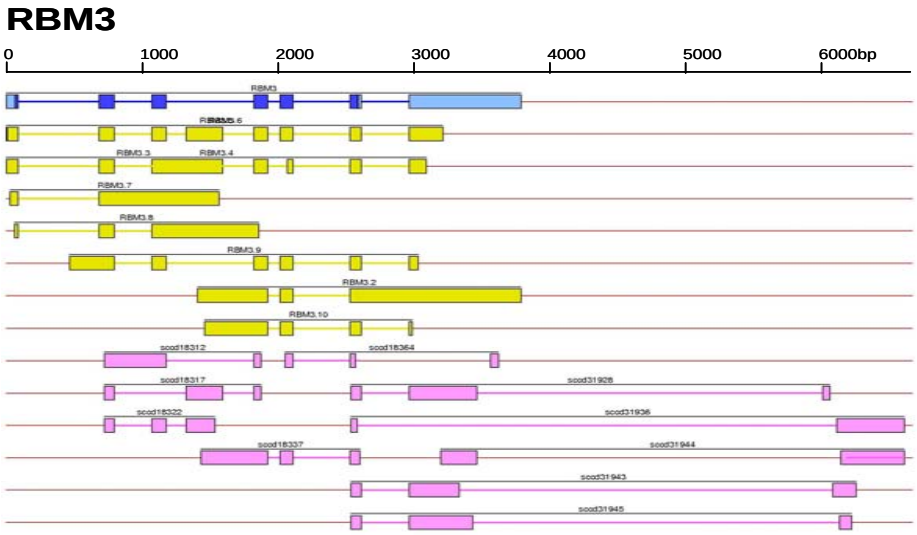
Figure 4.9 cDNA screens of *RBM3*

A) Expected PCR product sizes for *RBM3* gene fragments predicted from EST and cDNA sequences. Bands observed but not expected are denoted with an asterisk, *. B) Pre-screens of seven primer pairs designed to amplify fragments of *RBM3*. All pre-screen reactions were performed at 60°C and reactions for three tissues brain, lung and liver are shown in that order, followed by a negative control.
2-11 Selected cDNA screens for *RBM3* on 29 different tissues. Tissues are numbered 21-50 (excluding 41) and are listed in section 2.8. Molecular weight markers are denoted, M, and the locations of the 506 and 419 bp fragment are shown on the right hand side of each gel. Samples selected for further analysis are denoted with a red arrow.

A)



B)



C)

Screen	Ligations	Tissues	Colonies screened	Bands expected	Bands observed	Clones sequenced
1	Failed PCR					
2	3	28, 29, 30	96	1	1	3
3	3	25, 28, 45	96	6	5	9
4	3	36, 46, 44	96	5	4	6
5	6	29, 30, 32, 33, 35, 36	192	7	8	19
6	3	36, 39, 25	96	4	4	8
7	3	38, 39, 32	96	1	3	2
8	3	31, 32 33	96	3	4	11
9	3	49, 21, 22	96	3	3	12
10	3	23, 22, 26	96	3	2	5
11	3	47,43,42	96	4	4	7
12	3	25, 36, 35	96	3	5	12

Figure 4.10 Sequencing of *RBM3* fragments

A) Analysis of individual sequences using Spidey (Whelan *et al.*, 2001). This programme aligns EST and cDNA sequences to genomic sequence. It screens for splice site consensus sequences, and it was used to identify novel transcripts for *RBM3*. The screenshots displayed show the summary page, the exon/intron structure of the transcripts and an alignment between the transcript sequence (sccd18387 – AJ973555) and the genomic sequence (AC115618).

B) Exon/intron structures of *RBM3* transcript variants. The transcripts run 5' to 3' (left to right). The reference transcript for *RBM3* is displayed with coding exons (dark blue) and UTR (light blue). The exon/intron structure of alternative variants is shown in yellow. The novel transcripts identified in this study are displayed in pink.

C) Overview of the cloning and sequencing analysis of *RBM3*. Number of ligations completed to identify novel transcript fragments of *RBM3*, number of individual white colonies screened for each cDNA screen, the expected number of different clones and the observed number of different clones, and the number of samples sequences are displayed.

Table 4.6 Summary of *RBM3* transcript variants

(Blue – reference transcript, dark red-transcripts identified in this study)

Variant	Sequence	Identified in screen	Description	Size change (+/- bp)	Location	Region
RBM3.1	AK000859		n.a.	n.a.	n.a.	n.a.
RBM3.2	BM702340		novel first exon	415	exon 1	5' UTR
RBM3.3	AU137487		5' gain	416	exon1	5' UTR
RBM3.4	CB110977		whole exon deletion	107	exon 1	5' UTR
RBM3.5	BG708929		whole exon addition	269	exon 3a	CDS
RBM3.6	AL540984		intron retention	424	intron 2,3	CDS
RBM3.7	AL539019		intron retention	147	intron 3	CDS
RBM3.8	BM786866		first exon extension	322	exon 2	5' UTR
RBM3.9	AV703485		first exon extension	364	exon 4	5' UTR
RBM3.10	AJ973551 AJ973552	*RBM3*-7	internal deletion	1001	exon 7	CDS
RBM3.11	AJ973553	*RBM3*-5	intron retention	278	intron 2	CDS
RBM3.12	AJ973555	*RBM3*-5	whole exon deletion, whole exon addition	101	exon 3, 3a	CDS
RBM3.13	AJ973556	*RBM3*-5	intron retention	230	intron 3a	CDS
RBM3.14	AJ973585	*RBM3*-12	internal deletion 5' loss final exon extension	-619	exon 8	3' UTR
RBM3.15	AJ973558	*RBM3*-12	antisense	559	antisense	n.a.
RBM3.16	AJ973560	*RBM3*-12	5' loss, final exon extension	231	exon 8b	3' UTR
RBM3.17	AJ973562	*RBM3*-12	5' loss final exon extension	446	exon 8c	3' UTR
RBM3.18	AJ973564	*RBM3*-12	internal deletion 5' loss final exon extension	-616	exon 8d	3' UTR

In addition to aiding the identification of novel transcripts, the screening process also generated tissue expression profiles for the fragments of *RBM3*. Of particular interest were the tissue specific banding patterns in cDNA screens 3, 4 and 5, where the expression profiles in the ovary, prostate and skeletal muscle (tissues 32, 35 and 36 respectively) differed from all other tissues. Moreover, it was possible to predict the transcript structures of *RBM3* that displayed the tissue specific expression profiles using the predetermined, calculated banding patterns (Figure 4.11).

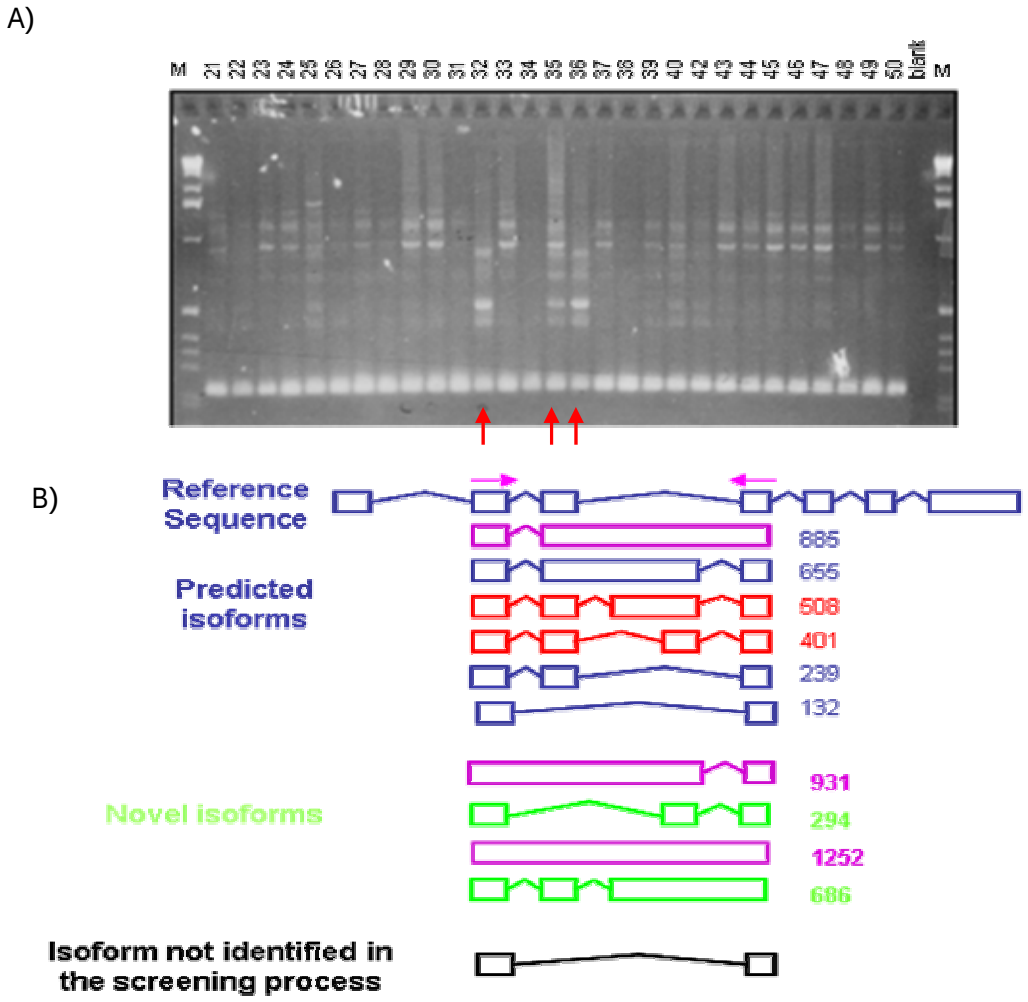


Figure 4.11 Tissue specific expression patterns of *RBM3*

A) A cDNA screen on 29 cDNA samples displays tissue specific expression patterns of *RBM3.* Tissues displaying a varied banding pattern are denoted with a red arrow.
B) Transcript fragments amplified in the PCR process. Blue – predicted from existing cDNA and EST evidence but predicted not to display expression specific to the ovary, placenta and skeletal muscle. Red - predicted from existing cDNA and EST evidence and predicted to display expression specific to the ovary, placenta and skeletal muscle. Green - identified as part of this study, but predicted not to display expression specific to the ovary, placenta and skeletal muscle. Magenta - identified as part of this study, but predicted not to display specific to the ovary, placenta and skeletal muscle. Black - predicted but not identified in this study.

4.3.6 Identification of novel *PQBP1* transcripts

At the time of analysis seven transcript variants of *PQBP1* had been identified using existing cDNA and EST sequences. The gene has seven exons and spans approximately 5.3 kb. The primer pairs used to identify alternative transcripts in *PQBP1* are listed and displayed in Figure 4.12. Fourteen cDNA screens were completed to cover the *PQBP1* gene comprehensively. All pre-screens and cDNA screens generated products of the predicted size. Additional bands were observed in 71% of the cDNA screens (10 out of 14 screens) which may represent novel transcripts or non-specific amplification. The cDNA screens are displayed in Figure 4.13.

Following the cloning and sequencing of specific PCR products for the cDNA sequences seven novel transcripts were identified for *PQBP1* (see Figure 4.18 and Table 4.7). The majority (57%) of these events were located within the 5' UTR of the gene. For example, the novel transcript fragment AJ973535 was generated in cDNA screen 3, where a novel PCR product of approximately 500 bp was observed. This product was amplified in all tissues except the heart, skeletal muscle and foetal skeletal muscle, and analysis of its sequence confirmed that the product was generated through the retention of intron 1. It is anticipated that this sequence was generated from reverse transcribed RNA rather than genomic DNA because the PCR product was consistently amplified in a variety of tissues. In addition, previous experimental analysis has confirmed that all samples were free from genomic contamination. All other variations that were observed in the 5' UTR were spliced.

All *PQBP1* transcript variants are listed in Table 4.7 and displayed in Figure 4.18.

A)

Screen	Primer 1		Primer 2		Product Size						
1	A	483749S	I'	476748A	144						
2	J	486749S	L'	483751A	203	134	400*	500*			
3	I	483748S	L'	483751A	99	500*					
4	K	483750S	L'	483751A	260	100*					
5	L	483751S	B	483741S	143						
6	A	483740S	D	483743A	239	320*					
7	A	483740S	C'	483742A	251	285	546	763	525*	850*	
8	A	483740S	E	483744S	501	480*					
9	C	483742S	G	483746S	498	328	476*				
10	C	483742S	F	483745S	216						
11	D	483743S	E	483744S	284	150*	100*				
12	D	483743S	C'	483742A	329	543	450*				
13	D	483743S	H'	483747A	329	621	753	1200 *			
14	B'	483741A	C'	483742A	170	204	465	682	320	444	700*

B)

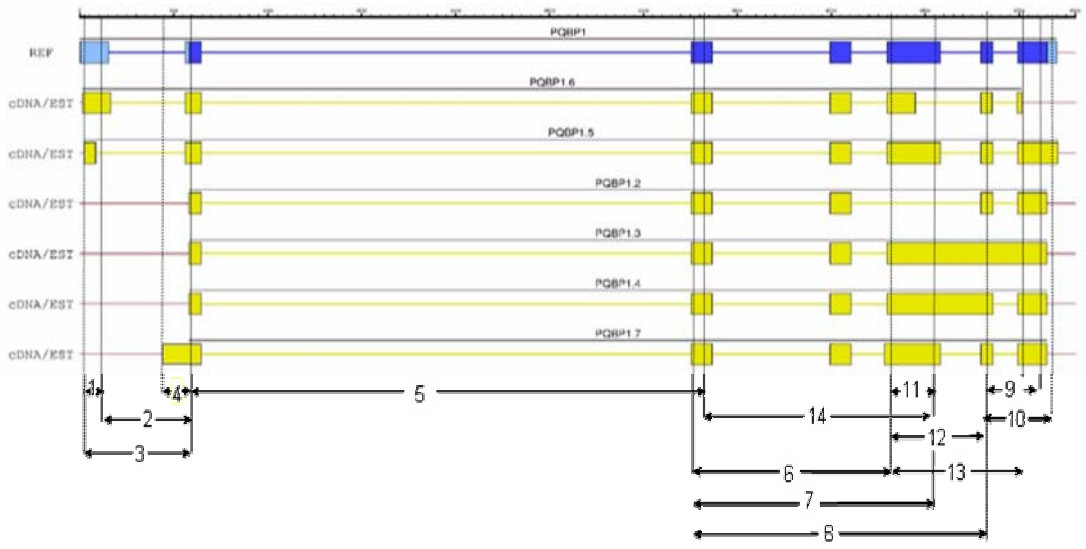


Figure 4.12 The location of primers used to screen for novel *PQBP1* transcripts.

A) List of primer combinations and *PQBP1* transcript variants that should be amplified are listed. These primers were used to screen the expression of *PQBP1* in 29 different tissues. Expected PCR product sizes for *PQBP1* gene fragments predicted from EST and cDNA sequences. Light grey, expected but not observed; dark grey not expected but observed.

B) Exon/intron structures of *PQBP1* transcript variants identified from existing cDNA and EST sequences. The transcripts run 5' to 3' (left to right). The reference transcript for *PQBP1* is displayed with coding exons (dark blue) and UTR (light blue). The exon/intron structure of alternative variants is shown in yellow. Overlaid on this diagram is the location of primers. The regions of transcripts amplified in each cDNA screen outlined and numbered in accordance with (A).

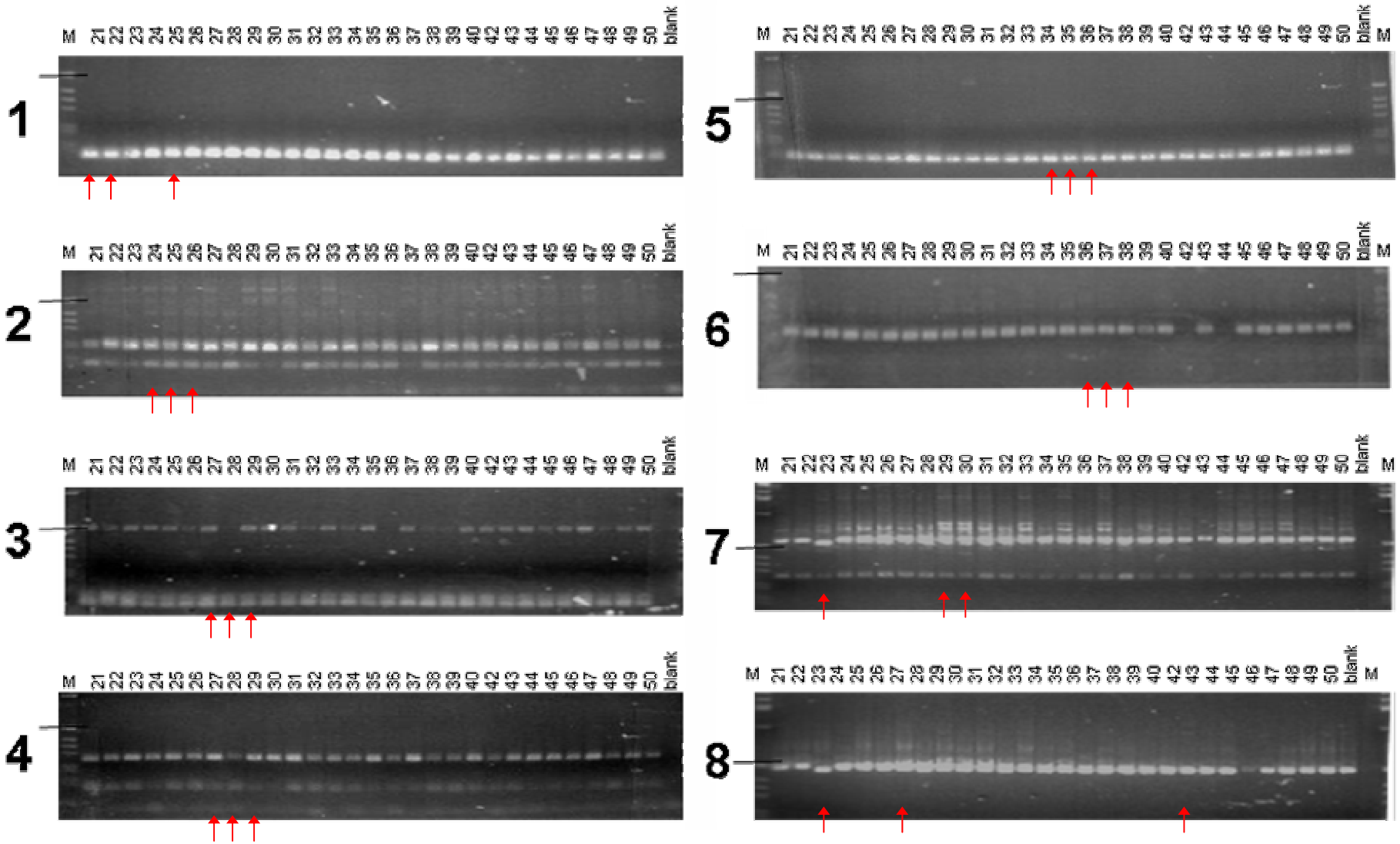


Figure 4.13 Identification of novel transcripts for *PQBP1*- cDNA screens

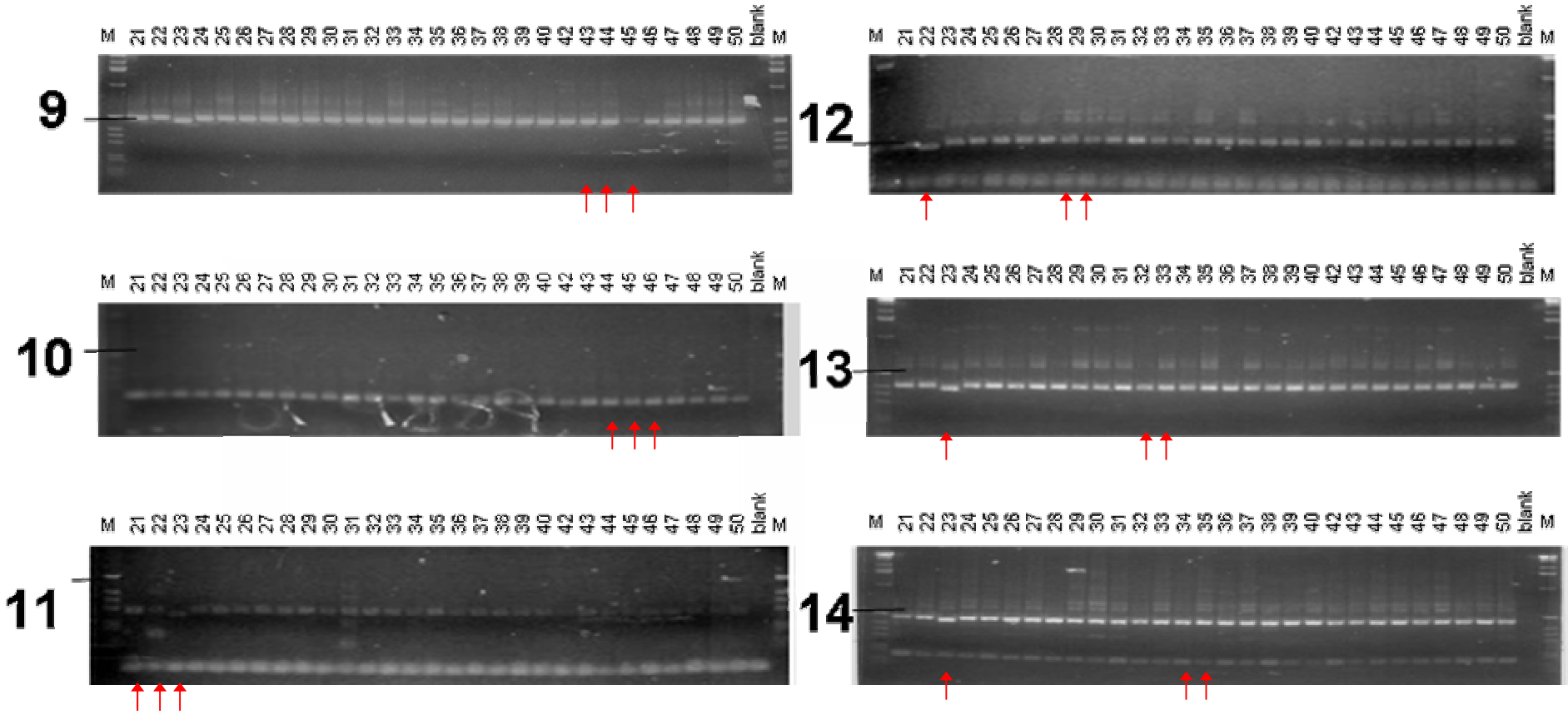


Figure 4.13 Identification of novel transcripts for *PQBP1*- cDNA screens

Shown are the 14 cDNA screens for *PQBP1*. The 29 tissues are numbered in accordance with section 2.8.1. The molecular weight marker used (kb ladder) is denoted (M) and for each screen the marker band 506 bp fragment is marked with a black line. Samples selected for further analysis are denoted with a red arrow.

Table 4.7 Summary of *PQBP1* transcript variants.

(Blue – reference transcript, pink transcripts identified in this study)

Variant	Sequence	Identified in screen	Description	Size change (+/- bp)	Location	Region
***PQBP1*.1**	**NM_005170**		**n.a.**	**n.a.**	**n.a.**	**n.a.**
***PQBP1*.2**	**AB041836**		**whole exon deletion**	**285**	**exon 5**	**CDS**
***PQBP1*.3**	**AB041834**		**Intron retention**	**2113**	**introns 5,6**	**CDS**
***PQBP1*.4**	**AB041835**		**Intron retention**	**211**	**intron 5**	**CDS**
***PQBP1*.5**	**BC012358**		**5' loss**	**-68**	**exon1**	**5' UTR**
***PQBP1*.6**	**BC255007**		**5' loss**	**-132**	**exon 5**	**CDS**
***PQBP1*.7**	**AB041837**		**3' gain**	**14**	**exon 5**	**CDS**
***PQBP1*.8**	**AJ973535**	***PQBP1*-2**	**3' gain**	**11**	**exon 1**	**5' UTR**
***PQBP1*.9**	**AJ973536**	***PQBP1*-2**	**Intron retention**	**410**	**intron 1**	**5' UTR**
***PQBP1*.10**	**AJ973538**	***PQBP1*-7**	**whole exon addition**	**87**	**exon 3a**	**CDS**
***PQBP1*.11**	**AJ973541**	***PQBP1*-4**	**Novel first exon**	**46**	**exon 1a**	**5' UTR**
***PQBP1*.12**	**AJ973540**	***PQBP1*-4**	**Novel first exon**	**102**	**exon 1b**	**5' UTR**
***PQBP1*.13**	**AJ973543**	***PQBP1*-6**	**5' loss**	**-48**	**exon 3**	**CDS**
***PQBP1*.14**	**AJ973545**	***PQBP1*-7**	**internal deletion**	**21**	**exon 5**	**CDS**

A novel exon was identified in transcript fragments sequenced from cDNA screen 7 (representative sequence: AJ973538). When aligned to genome sequence the 89 bp addition was found to be located within an *AluSq* repeat. *Alu* repeats have been implicated in shaping the human transcriptome as they harbour sequence motifs that resemble splice sites, which can result in their introduction of the *Alu* into a mature mRNA transcript. Most transcribed *Alu* repeats are observed in alternative transcripts (Jasinska and Krzyzosiak 2004). It has been demonstrated that the insertion of *Alu* elements into protein coding regions frequently disrupts the open reading frame and these types of events are often selected against. *In silico* analysis of the sequence AJ973538 suggested that the inclusion of this exon in *PQBP1* transcripts would introduce a premature termination codon that would reduce the length of the encoded protein by 198 amino acids (from 265 to 68 amino acids). However, the same analysis also identified a putative translation start site located within the novel exon that could be used as an alternative to the normal translation start site. Utilisation of this site could restore the correct reading frame and the resulting CDS would encode a 224 amino acid protein. This protein would lack the first 27 amino acids of the wild-type *PQBP1* protein which would be replaced by 18 amino acids from the *Alu* repeat (shown in Figure 4.14). The employment of the alternative translation start site *in vivo* is unlikely as the ATG is

not preceded by a Kozak consensus sequence (GCC(R)CC**ATG** consensus vs CGGCTT**ATG** actual (Kozak 1987)). There is no experimental evidence to support or refute the use of this alternative translation start site.

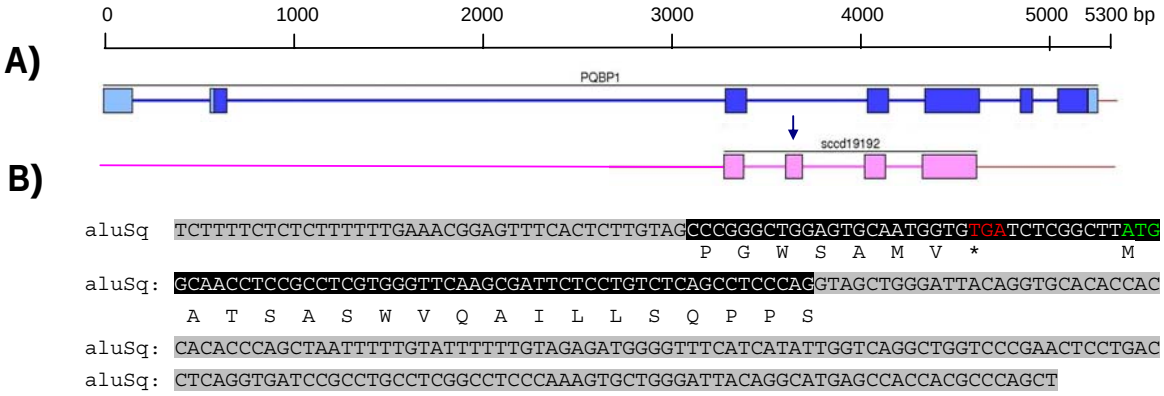


Figure 4.14 The novel exon 2a lies within an *Alu* repeat

A) Exon/intron structures of *PQBP1* reference sequences and a transcript variant, AJ973535 (sccd19192). The transcripts run 5' to 3' (left to right). The reference transcript for *PQBP1* is displayed with coding exons (dark blue) and UTR (light blue). The novel transcript, AJ973535 (sccd19192), is displayed in pink and the novel exon is denoted with an arrow.
B) The sequence of the *Alu* repeat is shown in grey with the novel exon shown in black. The predicted termination and start codons are displayed as well as the the predicted amino acid sequence.

A banding pattern exclusive to PCR products amplified from the adrenal gland was observed in cDNA screens, 7, 8, 9 and 11 to 14. All transcripts amplified from this tissue contained an internal 21 bp deletion in exon 4. The deletion is located within a repetitive region of the *PQBP1* gene that contains 7 copies of a 21 bp unit (Figure 4.15). However, the deletion does not lie on an exon/intron boundary, and neither its size nor location conforms to the spatial requirements of intron excision (Wieringa *et al.,* 1984). This novel transcript is therefore, not predicted to be the result of a tissue specific alternative splicing event, but rather a small genomic deletion. To confirm this observation genomic DNA from the same donor could also be screened using the same primer pairs. This, however, was not possible because the appropriate genomic DNA sample could not be obtained. Proposed mechanisms to explain changes in minisatellite size include unequal crossing over (Jeffreys *et al.,* 1998), gene conversion (Jeffreys *et al.,* 1994), and replication slippage (Levinson and Gutman 1987).

This deletion would remove seven amino acids from the encoded *PQBP1* protein (amino acids 83- 89 - HDKSDRG). It is anticipated that this deletion will have little effect on the cognate protein's structure or function (see section 6.2).

A) Exon 4

```
Ref:ATGCTGAAGAAAAGTTGGACCGGAGCCATGACAAGTCGGACAGGGGCCATGACAAGTCGGACCGCAGCCATGAGAAACTAGACAGGGGC
Adr:ATGCTGAAGAAAAGTTGGACCGGAG----------CCATGACAAGTCGGACCGCAGCCATGAGAAACTAGACAGGGGC

Ref:CACGACAAGTCAGACCGGGGCCACGACAAGTCTGACAGGGATCGAGAGCGTGGCTATGACAAGGTAGACAGAGAGAGAGAGCGAGACAG
Adr:CACGACAAGTCAGACCGGGGCCACGACAAGTCTGACAGGGATCGAGAGCGTGGCTATGACAAGGTAGACAGAGAGAGAGAGCGAGACAG

Ref:AGGGAACGGGATCGGGACCGCGGGTATGACAAGGCAGACCGGGAAGAGGGCAAAGAACGGCGCCACCATCGCCGGGAGGAGCTGGCTCC
Adr:AGGGAACGGGATCGGGACCGCGGGTATGACAAGGCAGACCGGGAAGAGGGCAAAGAACGGCGCCACCATCGCCGGGAGGAGCTGGCTCC

Ref:CTATCCCAAGAGCAAGAAGG
Adr:CTATCCCAAGAGCAAGAAGG
```

B)

Reference	Adrenal Gland
CTGAAGAAAAGTTGGACCGGA	CTGAAGAAAAGTTGGACCGGA
GCCATGACAAGTCGGACAGGG	---------------------
GCCATGACAAGTCGGACCGCA	GCCATGACAAGTCGGACCGCA
GCCATGAGAAACTAGACAGGG	GCCATGAGAAACTAGACAGGG
GCCACGACAAGTCAGACCGGG	GCCACGACAAGTCAGACCGGG
GCCACGACAAGTCTGACAGGG	GCCACGACAAGTCTGACAGGG

Figure 4.15 Identification of a 21 bp deletion that is exclusive to the adrenal gland sample

A) Sequence alignment of exon 4 for the reference *PQBP1* sequence and the sequence identified in the adrenal gland (AJ973535). The 21 bp deletion is displayed in red.

B) Sequence alignment of the 21 bp repeat contained in exon 4. Bases that are conserved in all repeat units are displayed in blue. Bases displaying moderate conservation are displayed in grey.

4.3.7 Tissue specific amplification profiles

Using the data generated in this chapter, inferences about tissue specific regulation of alternative splicing can be made. Here, it is possible to analyse the banding patterns created by PCR amplification of the cDNA panel for variations that were indicative of tissue specific splicing patterns. Examples of cDNA screens that displayed variable expression profiles are displayed in Figure 4.16.

Figure 4.16-a illustrates an example of one tissue having uniquely sized amplicons. This particular example was discussed in section 4.3.6 and is hypothesised to be a sequence polymorphism rather than a transcript variation. An example of a cDNA that generated a highly variable expression profile is displayed in Figure 4.16-b (cDNA screen *WAS-4)*. Two novel alternative transcripts were identified from this cDNA screen. One transcript was detected from a small amplicon that was generated in the brain (tissue 25), while the other, larger transcript was amplified from the lung (tissue 31). Figure 4.16-c depicts an example of amplification of alternative transcripts in some but not all tissues. This variation may be the result of an deletion such as one recorded in *GATA1*-4. Figure 4.16-d again shows an example of amplification of alternative transcripts in some but not all tissues. Larger PCR products were amplified in *HDAC6*-8 in most tissues except heart (tissue 28) and skeletal muscle (tissue 36).

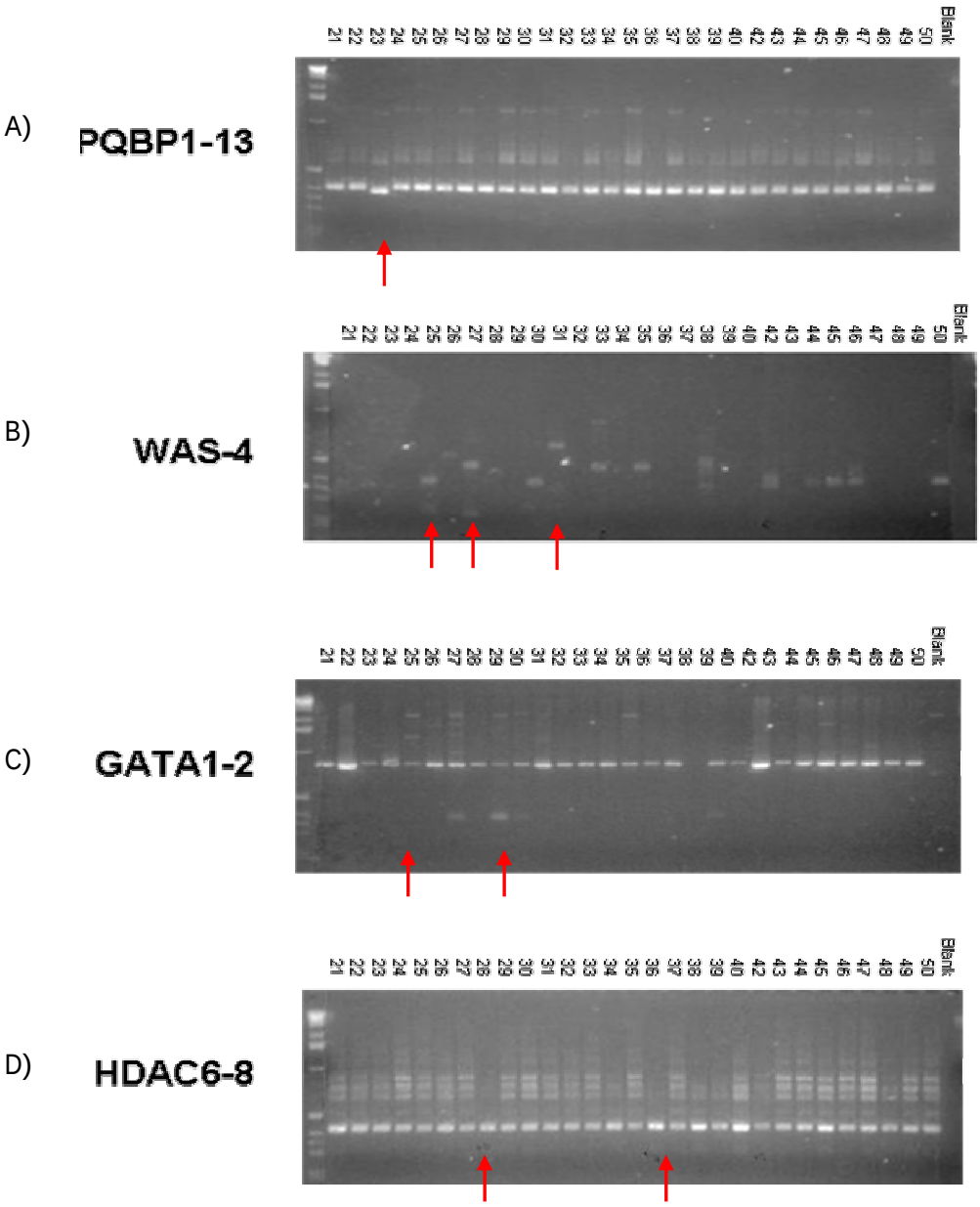


Figure 4.16 Example of cDNA screens that displayed tissue specific expression profiles

cDNA screens that displayed tissue specific amplification profiles are displayed. Tissues are numbered 21-50 and are listed in section 2.8. Molecular weight standards (1 kb ladder) are shown on the left hand side of each gel while negative controls are shown on the right hand side of each gel image. Tissues displaying unique profiles are indicated with a red arrow.

4.3.8 Summary of results

In total, 61 novel transcripts were identified using the targeted cDNA screening and sequencing strategy. This almost doubles the previous number of novel transcripts identified from EST and cDNA sequences (64 transcripts). The largest number of novel transcript fragments were identified for *HDAC6* (13 fragments), *RBM3* (8) and *PQBP1* (7) while no novel fragments were identified for *ERAS*, PCKSN1, *PIM2*, *KCND1* and *SLC35A2*. The number of novel splicing variants that have been identified for each gene is displayed in Figure 4.17. All but two genes had more than one transcript with an average of 7.7 transcripts identified per gene.

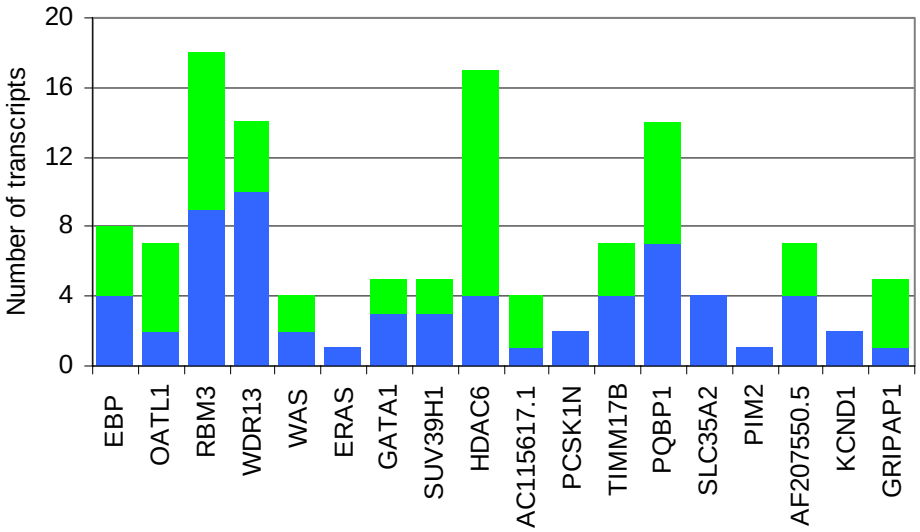


Figure 4.17 Summary of the number of novel transcripts identified by cDNA screening for 18 genes in Xp11.23
The number of transcripts identified from EST and cDNA sequences is displayed in blue while the number of transcripts identified in this study is displayed in green.

All novel sequences obtained in this study have been submitted to the EMBL nucleotide database (accession numbers – AJ973481 to AJ973591). A list of the novel transcripts that were identified for the 18 eighteen genes analysed in this study is displayed in Appendix IV. Appendix IV lists the sequences from which the novel transcript was identified, the type of variation observed, the size of the variation and its location. The exon/intron structures for all of the transcript variants are displayed in Figure 4.18.

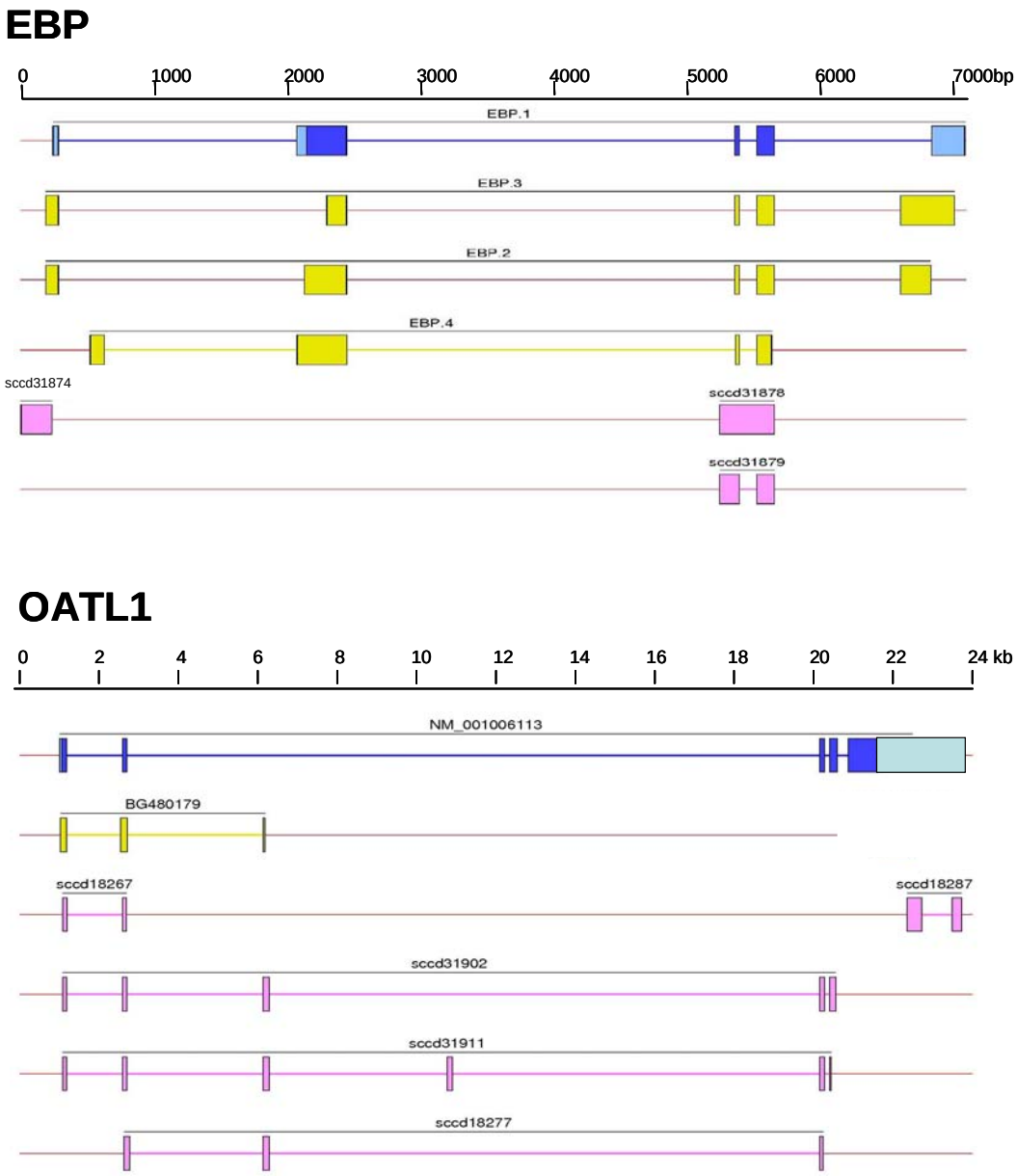


Figure 4.18 Transcript structures for genes in human Xp11.23

Continued overleaf.

In all cases the direction of transcription runs from left (5') to right (3'). Reference sequences are displayed (blue), UTR (light blue). Novel transcripts identified from existing EST and cDNA sequences are displayed in yellow while sequences identified in this study are displayed in pink.

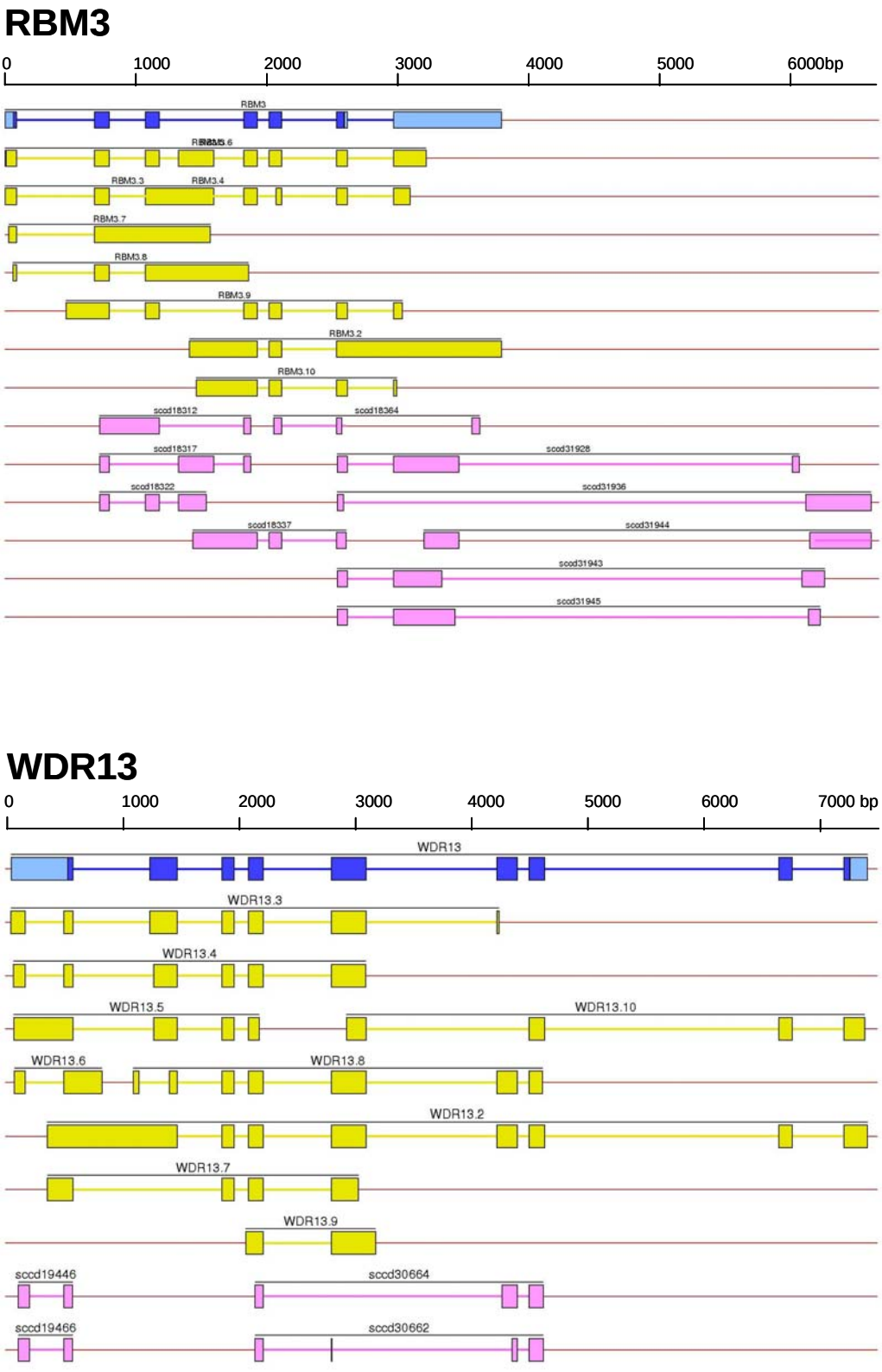


Figure 4.18 Transcript structures for genes in human Xp11.23 – continued overleaf

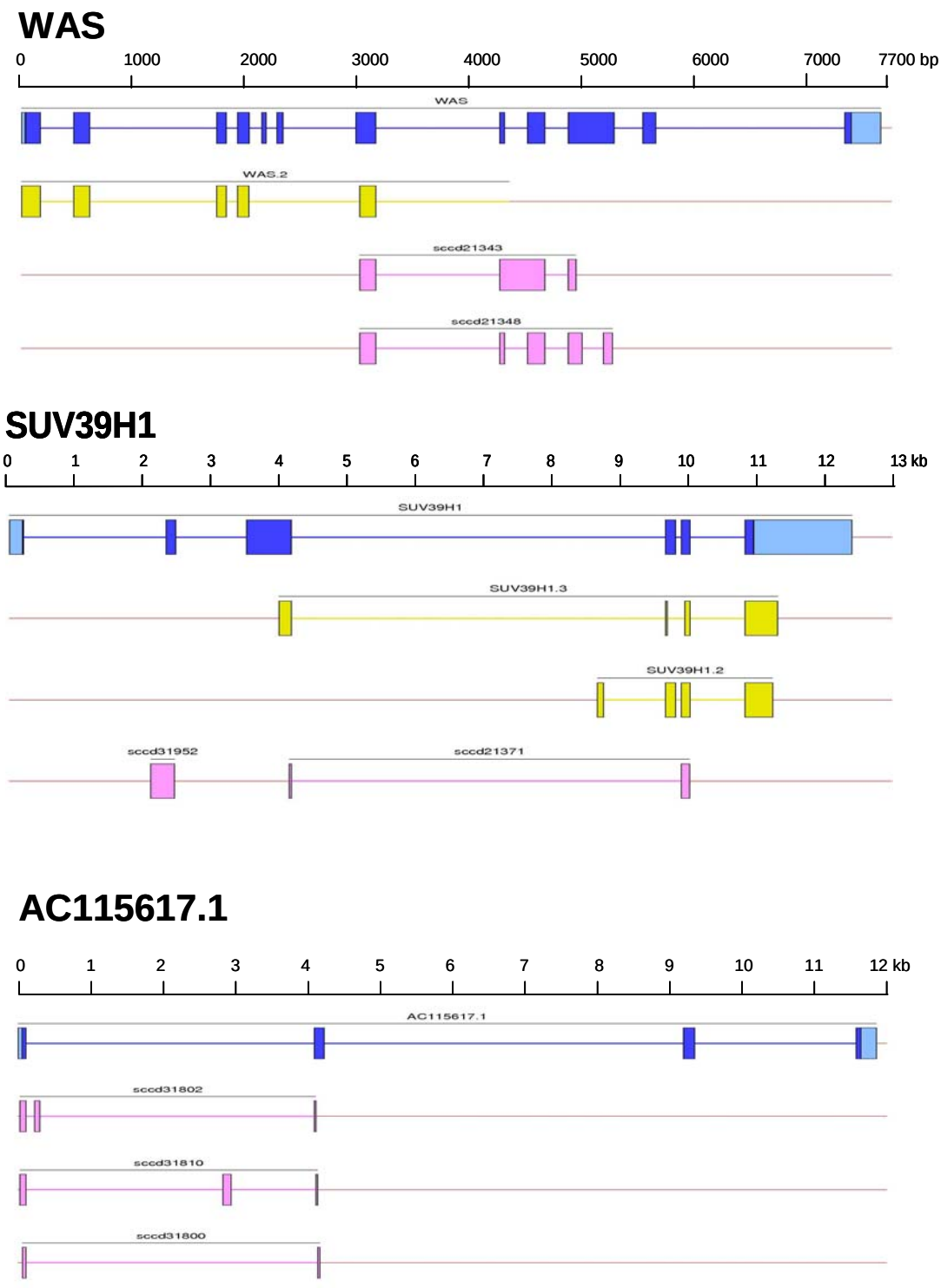


Figure 4.18 Transcript structures for genes in human Xp11.23 – continued overleaf

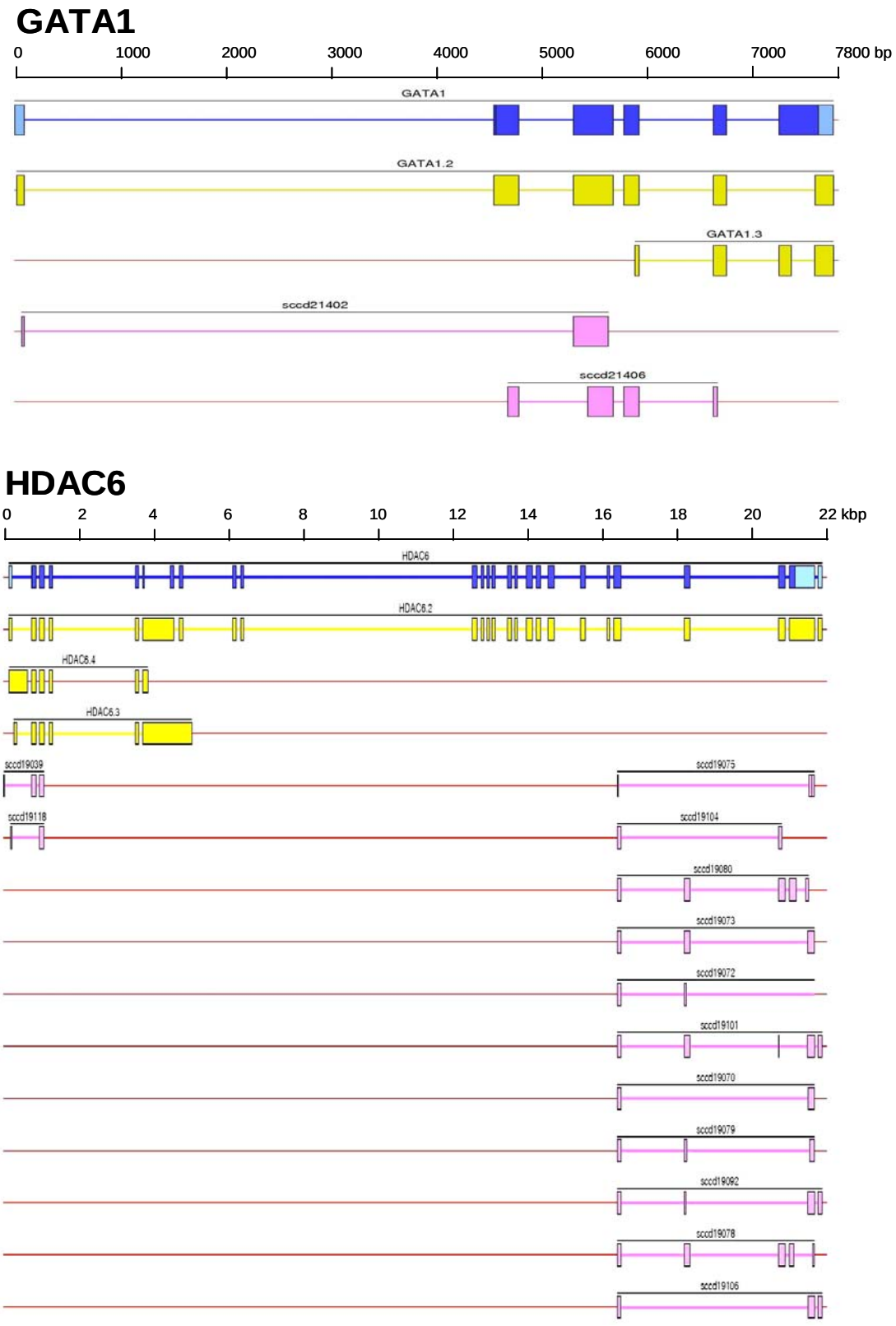


Figure 4.18 Transcript structures for genes in human Xp11.23 – continued overleaf

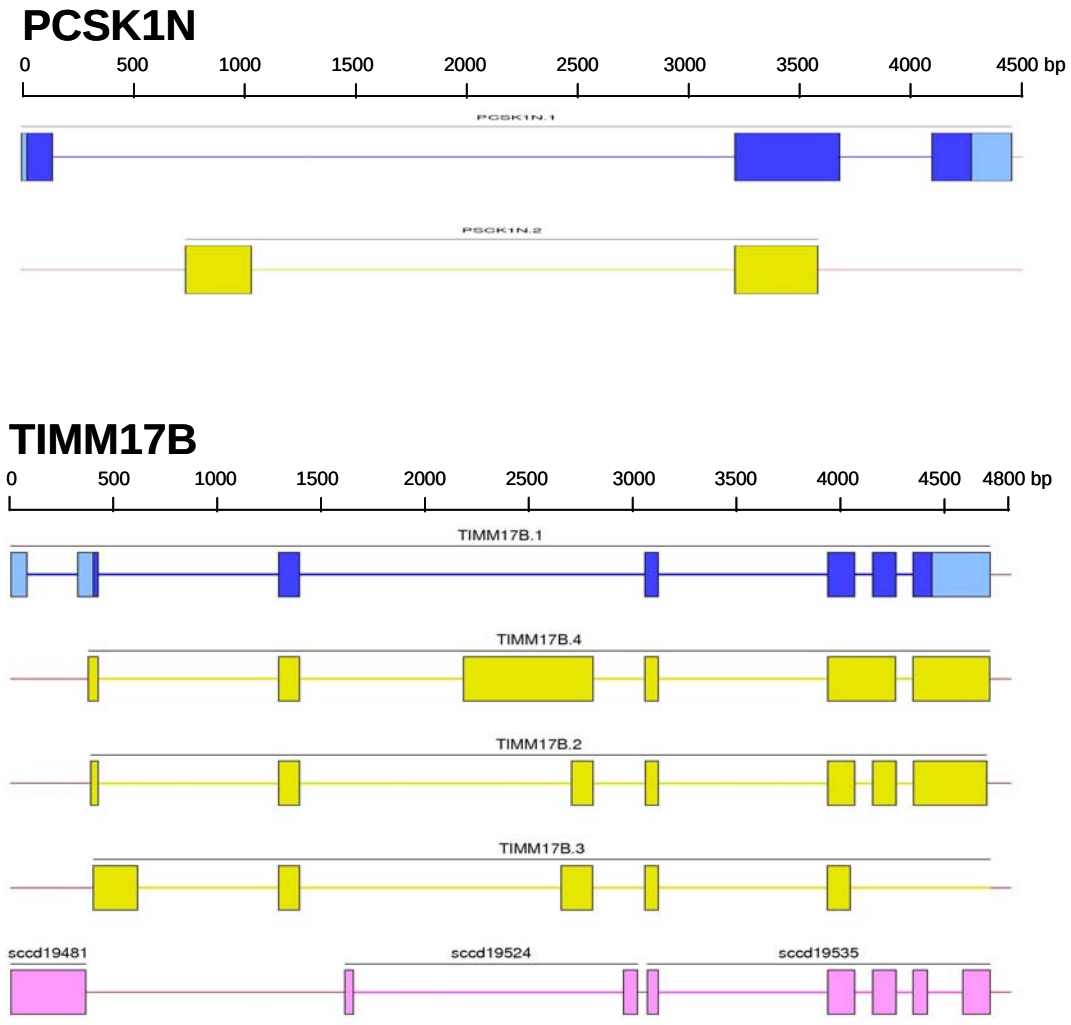


Figure 4.18 Transcript structures for genes in human Xp11.23 – continued overleaf

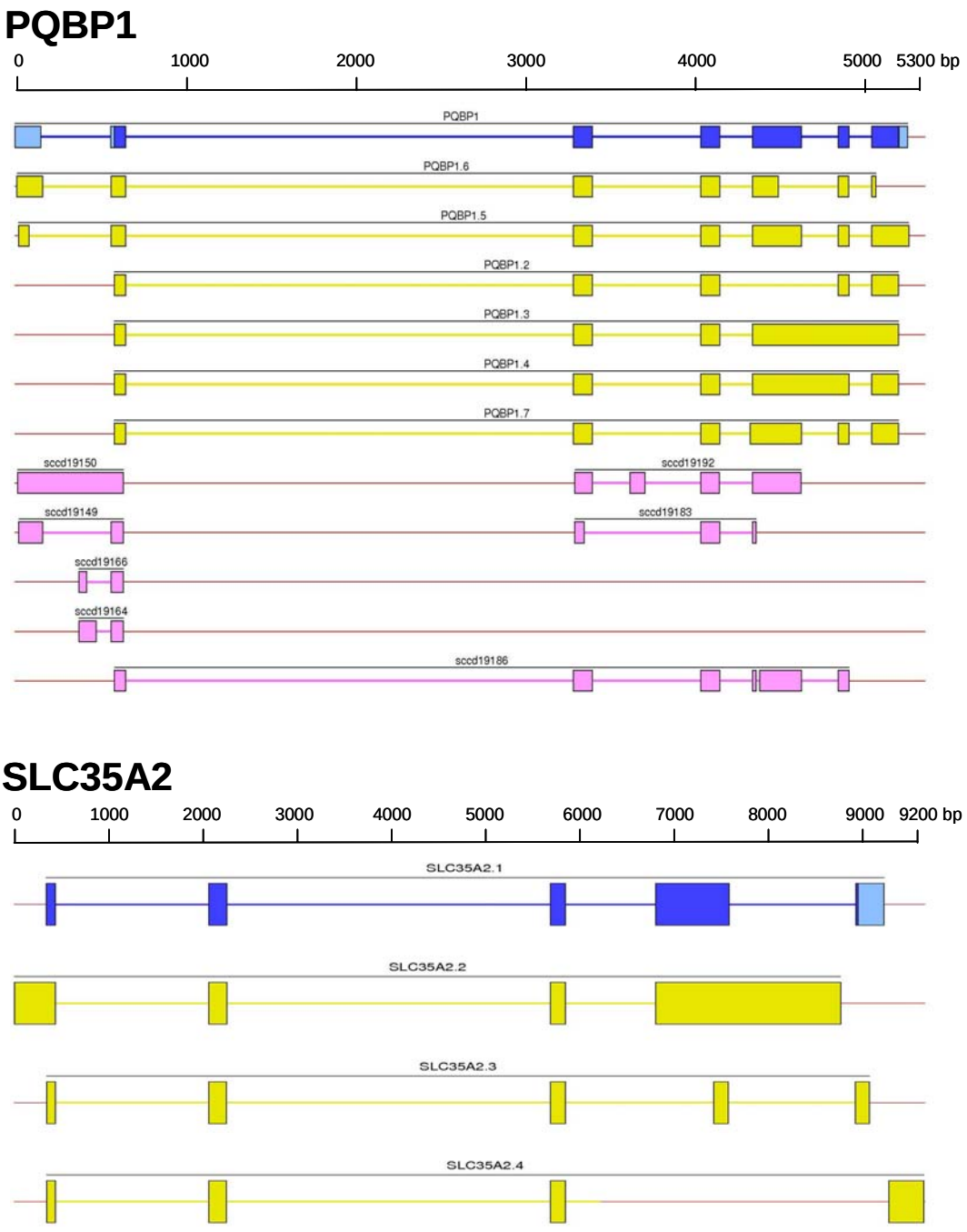


Figure 4.18 Transcript structures for gene in human Xp11.23 - continued overleaf

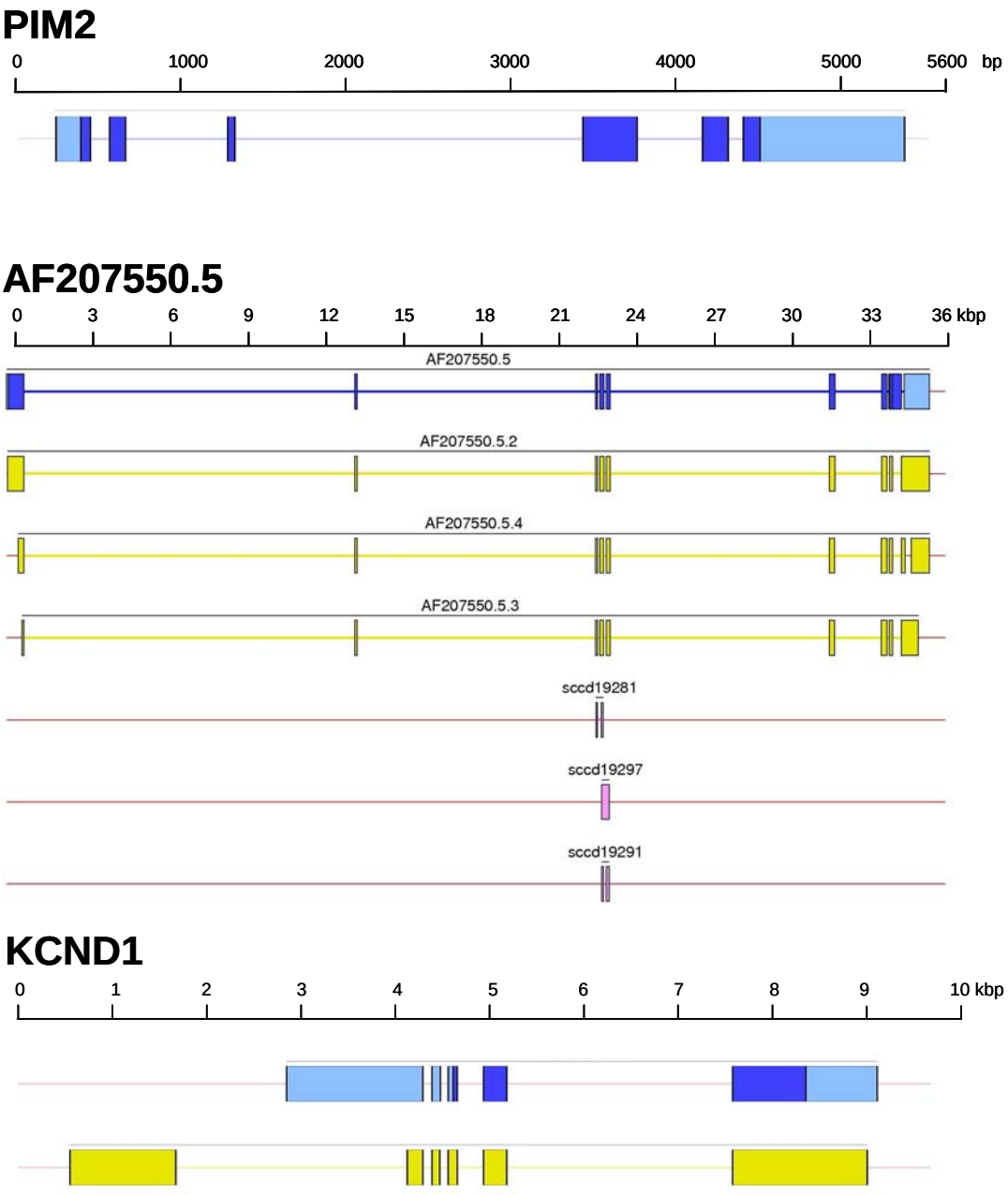


Figure 4.18 Transcript structures for genes in human Xp11.23 - continued overleaf

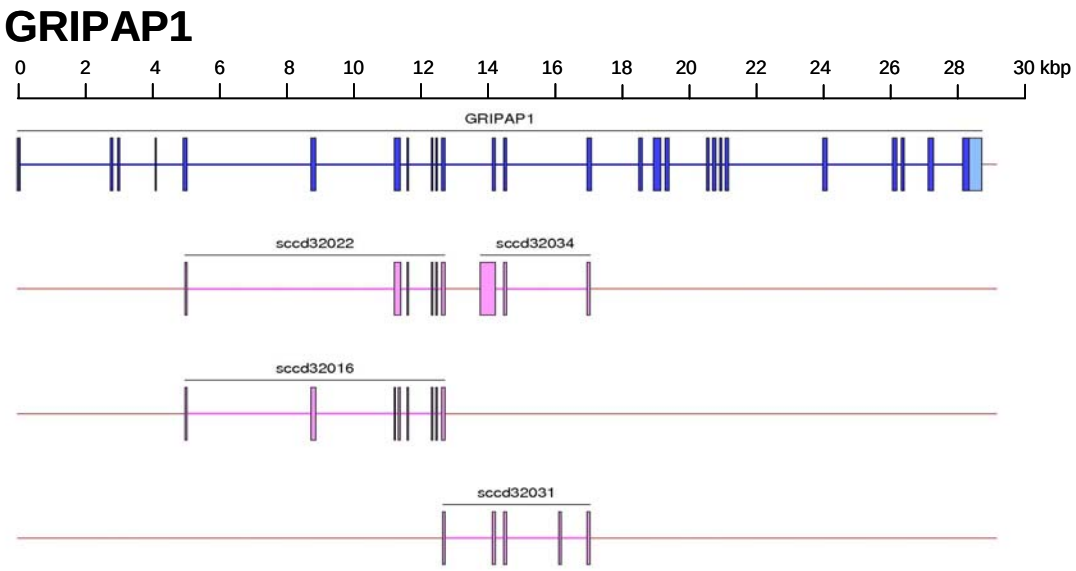


Figure 4.18 Transcript structures for genes in human Xp11.23

4.4 Analysis of alternative splicing events

Following the identification of novel transcripts, analysis proceeded to characterise their sequence properties. This analysis was performed on reference sequences, existing cDNA and EST sequences and the novel gene fragments identified in this study (125 sequences in total).

4.4.1 Describing the variation in transcript structures

The transcript variants were first classified for the type of splicing event that rendered them novel. The number of observations of each type of splicing event is displayed in Figure 4.19. The most common event was the removal of an entire exon (22 cases). An extension of the 3' end of the final exon was the least frequently observed event (novel final exon). This may be because the cDNA material from which many EST and cDNA sequences are sourced is synthesised using the oligodT primer method so that the 3' ends of genes are well represented in the existing transcript libraries.

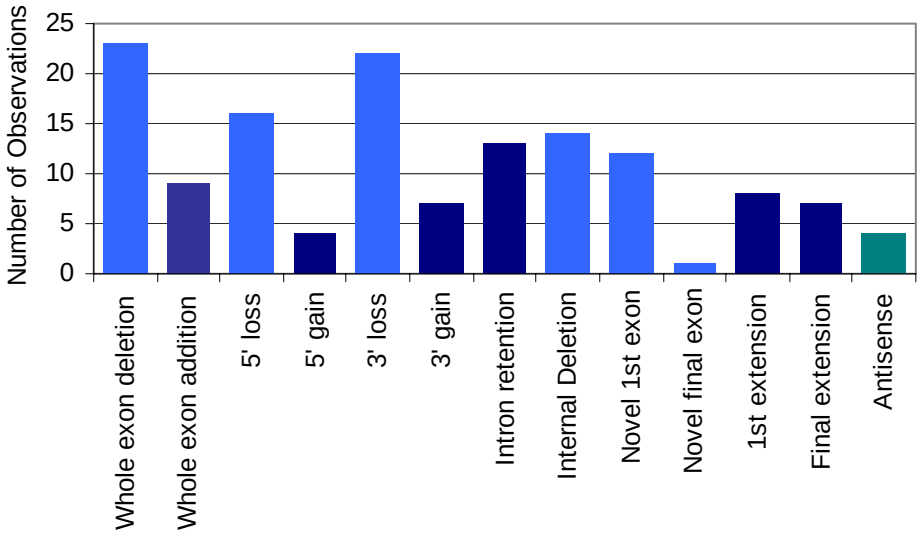


Figure 4.19 Types of alternative splicing events observed in 18 genes from human Xp11.23

4.4.2 Location of transcript variation

One of the many benefits of alternative splicing is the ability to increase the diversity of an organism's transcriptome and proteome without altering the gene

number. Alternative splicing may result in changes to either the UTR or CDS. Changes in the 5' and 3' UTR have been associated with changes in expression patterns (Beaudoing *et al.,* 2000). Change in the CDS may alter the domain structures of the cognate protein and hence affect its function (Kriventseva *et al.,* 2003). Interestingly, the majority of alterations were observed within the 5' UTR. This is consistent with finding of a positional bias of alternative splicing events to the 5' and 3' ends when using cDNA and EST sequences as the primary substrate (Jackson *et al.,* 2003) and may serve to alter the expression patterns of the genes included in this study.

4.4.3 Analysis of exon junctions

The sequence composition of splice sites was assessed by extracting transcribed sequence that flanked exon-intron junctions for each of the 17 spliced genes (the single exon gene, *ERAS*, was not included in this analysis). In total, 129 exon donor and acceptor sequences were extracted from the reference sequence for the genes, while 56 alternative donor and 60 alternative acceptor sequences were also extracted. These sequences were used to compare the sequence composition of di-nucleotide donor and acceptor sequences as well as the splice site scores for both reference and alternatively spliced exons.

The number and type of di-nucleotide exon donor and acceptor sequences is displayed in Figure 4.20. Greater than 99% of the reference junctions had the exon donor sequence GT (128/129), while one GC donor sequence was observed in the gene GRIPAP1. The acceptor dinucleotide sequence AG was observed in all reference exons. However, the dinucleotide sequences that border alternative exon junctions exhibited greater sequence variation (Figure 4.20, Table 4.8). In these cases, only just over 60% of exon junctions harboured the dinucleotide sequences used in U2 snRNA mediated mRNA splicing; GT (donor) and AG (acceptor).

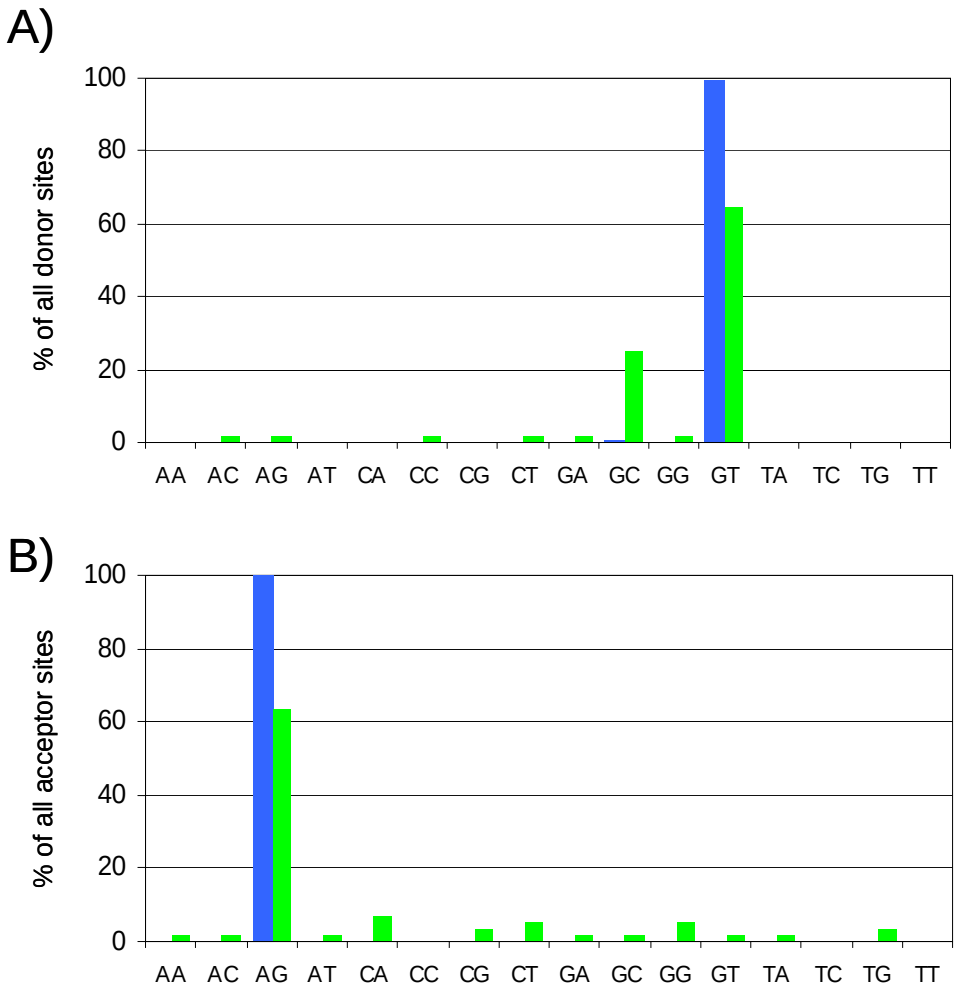


Figure 4.20 Classification of splice site sequences

A) 5' (donor) sequences for alternative (green) and the reference (blue) transcripts

B) 3' (acceptor) sequences for alternative (green) and reference (blue) transcripts

Table 4.8 Variation in splice site sequences for alternatively spliced exons

The number of instances of dinucleotides at the donor or acceptor site is shown for each gene.

A) 5' -donor

GENE	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT	Sum
EBP											1	1					**2**
OATL1												4					**4**
RBM3			1							1		7					**9**
WDR13										3		3					**6**
WAS										1							**1**
ERAS																	**0**
GATA1										1							**1**
SUV39H1										1		1					**2**
HDAC6		1						1	1	7		3					**13**
AC115617.1												2					**2**
PCSK1N												1					**1**
TIMM17B										1		2					**3**
PQBP1												7					**7**
SLC35A2																	**0**
PIM2																	**0**
AF207550.5												2					**2**
KCND1												1					**1**
GRIPAP1												2					**2**
Total	**0**	**1**	**1**	**0**	**0**	**0**	**0**	**1**	**1**	**15**	**1**	**36**	**0**	**0**	**0**	**0**	**56**

B) 3' -acceptor

	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT	Sum
EBP			2					1									**3**
OATL1			5														**5**
RBM3			4				1	1		1		1					**8**
WDR13			6					1			1				1		**9**
WAS			1														**1**
ERAS																	**0**
GATA1			1		1										1		**3**
SUV39H1							1										**1**
HDAC6			8	1	1			1				2					**13**
AC115617.1			3														**3**
PCSK1N																	**0**
TIMM17B			3		1						1						**5**
PQBP1			3														**3**
SLC35A2			1														**1**
PIM2																	**0**
AF207550.5			2														**2**
KCND1			1														**1**
GRIPAP1			2														**2**
Total	**0**	**0**	**42**	**1**	**3**	**0**	**2**	**4**	**0**	**1**	**2**	**3**	**0**	**0**	**2**	**0**	**60**

The strength of exon boundaries (splice sites) was then calculated for each of the donor and acceptor sites in both the reference and alternative transcripts. This was done using the equations derived by Shapiro and Senapathy (1987) in which the sequence composition of 8 bp around the 5' intron, (donor) and 15 bp around the 3' intron (acceptor) are analysed (Shapiro and Senapathy 1987). The programme weights each base according to the frequency of a particular base at a particular position and the sum of the weighted base scores reflects the strength of the splice site. Optimal acceptor and donor splice sites score 100.

The splice site scores for all genes are displayed in Figure 4.21. The average 5' donor score for reference transcripts was 86.7, which decreased to 70.1 for alternative exons. The average 3' donor score for reference transcripts was 84.8, decreasing to 61.4 for alternative transcripts. In almost all cases (15/17 donor, 17/17 acceptor), the splice site scores were higher in reference transcripts rather than alternative transcripts. Marginally higher scores for alternative donor sites were recorded for the genes *GATA1* (86.1 alt v 85.7 ref) and *TIMM17B* (88.1 v 86.7).

These results suggest that, as might be expected, the sequence composition of reference splice sites more closely resembles the consensus motif for efficient intron excision. These splice sites may therefore be recognised in preference to cryptic splice sites by the U2 splicing machinery.

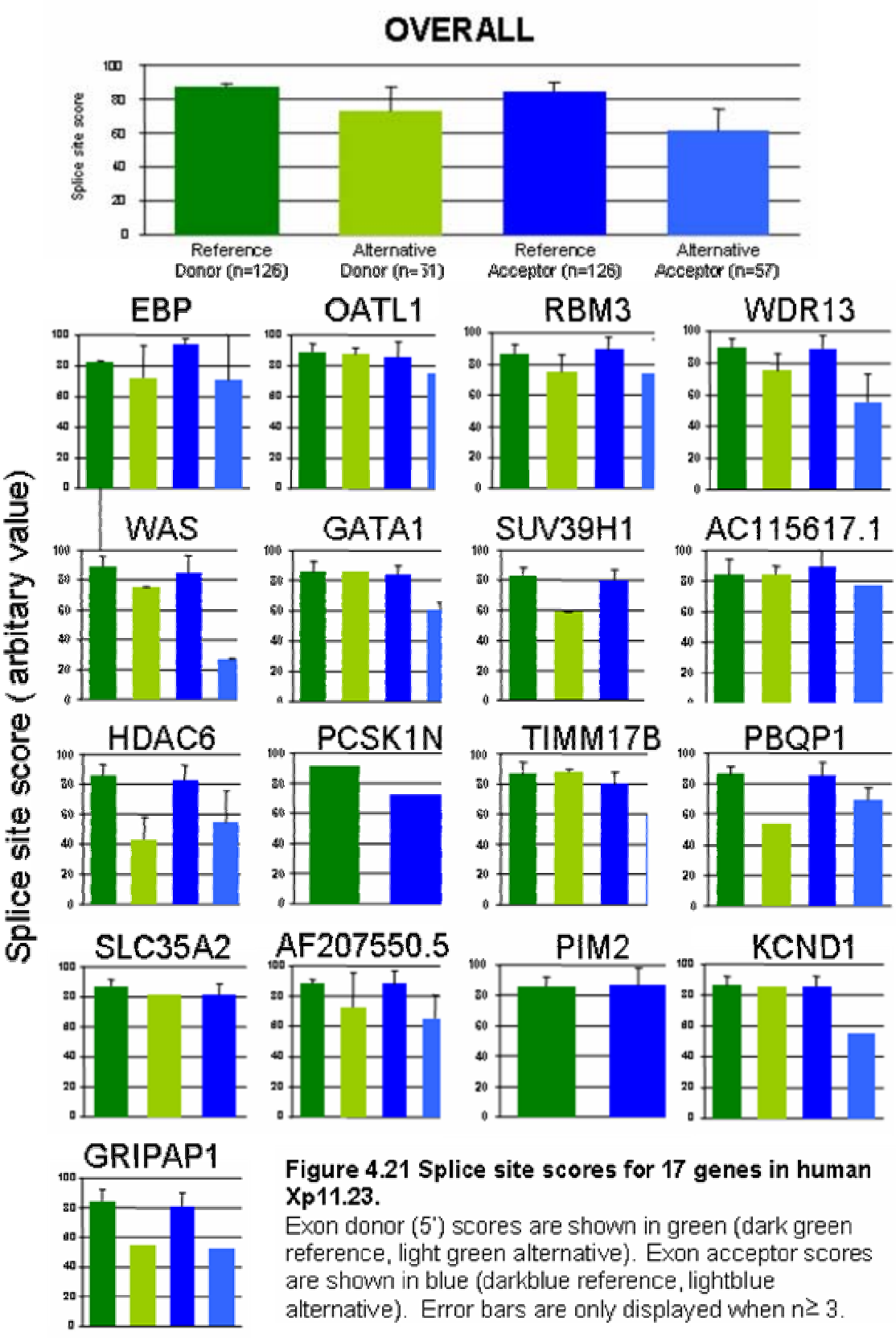


Figure 4.21 Splice site scores for 17 genes in human Xp11.23.
Exon donor (5') scores are shown in green (dark green reference, light green alternative). Exon acceptor scores are shown in blue (darkblue reference, lightblue alternative). Error bars are only displayed when n≥ 3.

4.4.4 The association of transcript diversity with other features

Attempts were made to correlate the number of alternative transcripts identified for each gene with a number of other genic parameters.

Transcript variation versus exon number

The fidelity of mRNA splicing is dependent on an intricate network of interactions using both RNA and proteins as substrates. These highly specific interactions serve to ensure that correct splice sites are utilised at exon-intron junctions in preference to numerous cryptic splice sites that resemble consensus sequences. Alternative splicing events may utilise alternative splice sites, which are often cryptic and are recognised less efficiently by the splicing machinery. Messenger RNA splicing is also modulated by branchpoint strength and regulatory elements in response to a variety of stimuli. If these events are very tightly controlled it could be assumed that all transcripts variants arise as the result of regulated splicing events. Conversely, it might the case that many transcript variants arise as the result of aberrant splicing events. In this case, it might be predicted that the greater number of exons in a gene, the greater the number of aberrant splicing events and the greater the level of transcript diversity (Figure 4.22 – A).

No correlation was observed between the exon number and the number of alternative transcripts for the 18 genes analysed in this study, (r^2= 0.016). This result is inconclusive but it may suggest that both exon number and imprecise splicing may affect the number of transcript variants. Additional analysis on a larger gene set is required to test either of these assumptions more rigorously.

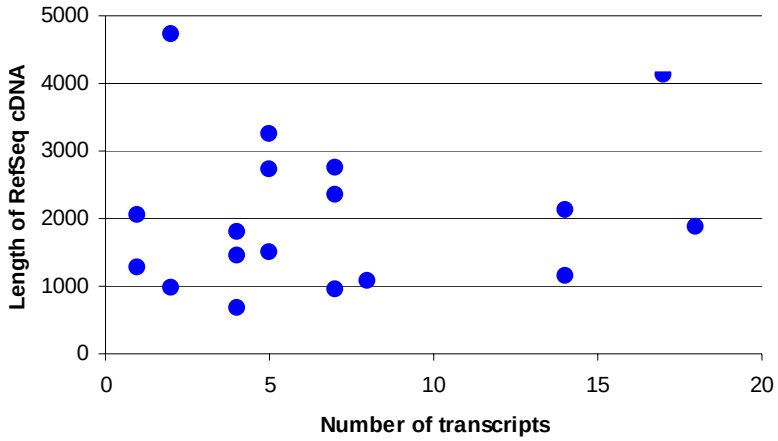
Transcript variation versus gene length

If mRNA processing were not tightly regulated it may also be hypothesised that the number of variant transcripts identified for each gene would increase in proportion with the length of the premature mRNA sequences. Genic sequences are scattered with potential cryptic splice sites that may act as decoys from genuine splice sites during mRNA processing. However, no correlation (r^2 = 0.019) was observed between the number of transcript variants and the length of the reference cDNA locus (Figure 4.22 – B).

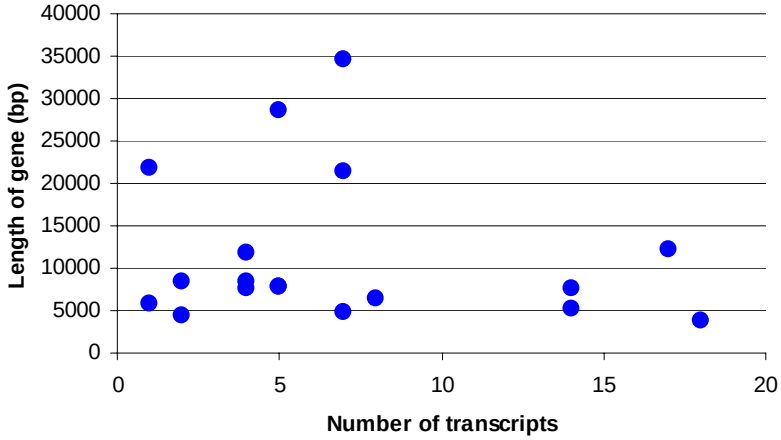
Transcript variation versus size of homologous UniGene cluster

It is also widely conjectured that the frequency of transcript variation increases with the depth of sampling (Zavolan and Kepler 2001; Kan *et al.*, 2002). To test this suggestion, the number of transcript variants for each gene was correlated with the total number of transcripts contained within the gene's UniGene cluster (http://www.ncbi.nlm.nih.gov/UniGene). Here, a moderate correlation (r^2=0.5) was observed between transcript number and UniGene cluster size, suggesting that the depth of sequence coverage is a reasonable indicator of transcript diversity (Figure 4.22 -C).

A) vs number of exons



B) vs gene length



C) vs size of UniGene cluster (number of sequences)

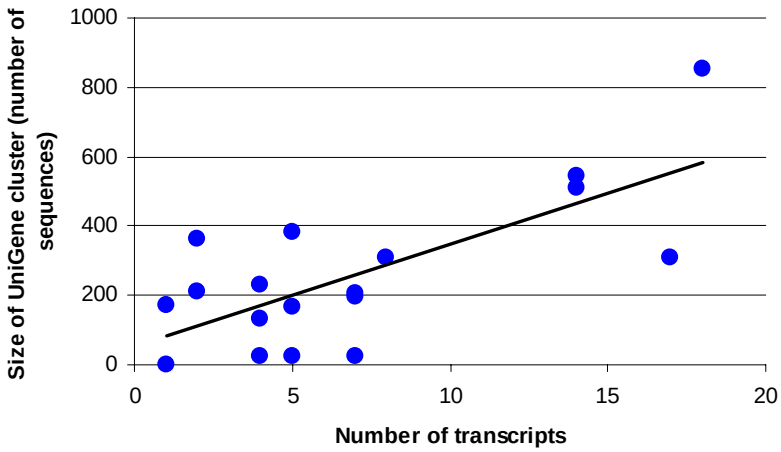


Figure 4.22 Correlation of gene features with transcript numbers

4.5 Discussion

Work presented in this chapter, extends upon previous knowledge of alternative splicing for 18 genes in human Xp11.23. Sixty-four transcripts were identified using existing human cDNA and EST sequence information while an additional 61 novel transcripts were identified in this study. The ease with which additional transcript variants were identified suggests that the cDNA samples used in this study have not been exhaustively sequenced and further sampling may identify even more transcript variants.

This frequency of alternative splicing is higher than current estimates of transcript variation in the human genome which range between 29%-59% (Modrek and Lee 2002). It is possible that this gene rich region in human Xp11.23 is also enriched for transcript variation; however, it is more likely that the increased frequency of transcript variations could be attributed to the increased depth of cDNA sequencing for the genes included in this study.

The method chosen to identify novel transcript variants has several advantages over alternative methods. The expression of each gene was comprehensively screened in a large number of tissues using a variety of different primer combinations. Standard PCR conditions were used for all profiling experiments which enabled easy detection of novel variants represented by unexpected PCR products, and permitted additional tissue profiling of defined regions of transcribed loci. Moreover, this method allowed multiple exon junctions to be analysed in concert, thereby enabling transcript variants to be put into context with the full-length gene.

This approach was also chosen as it produces exact sequence information to facilitate the precise identification of splice sites and novel splice forms. From this information possible functional effects may be inferred. In contrast, indirect identification techniques, such as hybridisation based methods, require assumptions to be made about possible splicing variants. This can be done by predicting what splice forms might be produced from the genomic sequence and then correlating these predictions to observed transcript sizes (Northern blot) or to hybridisation signal (on a microarray). Sequence data, on the other hand can accurately detect a new form by providing a direct readout of its sequence.

The success of the cDNA screening, cloning and sequencing strategy employed in this chapter is illustrated by the large number novel transcripts identified. One caveat of this strategy is its sensitivity. As cDNA transcript variants are identified following amplification it is possible that the method may detect transcripts that are present at low levels and that do not have biological relevance. That is, it may detect non-functional transcripts produced by imprecise splicing events. Additional analysis is required on all transcript variants to differentiate between functional and non-functional transcript variants.

It is acknowledged that several limitations are associated with this experimental approach used here. cDNA synthesis, PCR and cloning inefficiencies may have decreased the number of alternative transcripts identified. To overcome these limitations more cDNA screens were completed and more tissues were sampled in regions that displayed a high level of heterogeneity in their transcript structures. The use of overlapping primer pairs in the cDNA screening process also ensured that most splice sites were screened more than once. Moreover, efforts were made to ensure that the amplification conditions were optimised for each primer pair used and that, PCR products from several different tissue samples were analysed for each screening reaction.

With this approach it is difficult to identify small changes in the size of cloned PCR products. Hence, changes in splicing patterns that resulted in small length differences between two transcripts may not have been detected. This problem could be overcome by randomly sequencing more cloned transcripts or randomly sequencing the clone collection to considerable depth.

This study may also benefit from more detailed cDNA sampling, both in the depth of sequence coverage and type of tissue analysed. The primary resource used in this study was total RNA from 29 different tissues from healthy individuals. This represented all commercial samples available when the project was initiated. Previous analysis confirmed that two of the tissues included in this study had exhibited high level of transcript variation, the brain and testis (Dredge *et al.,* 2001). Detailed analysis of different cell types and developmental stages in these tissues may identify even more novel transcript variants. Transcript variation is also prevalent in the central nervous (Lee and Irizarry 2003) and the immune systems (Lynch 2004), which were not represented in the cDNA panel used in this

study. Using additional tissue sources and developmental stages in these systems may also identify more novel transcript variants.

Pooled of cDNA samples could facilitate the high-throughput identification of transcript variants. For example, a commercial source of mixed cDNA samples, (e.g., Universal cDNA Clontech, Stratagene) combines cDNA samples from over 30 different human tissues and would permit many gene fragments to be analysed in concert. This approach obviates the need for cDNA profiling as direct purification and cloning of the PCR products may identify novel transcripts without having to screen individual tissues. However, increasing the complexity of the sample may exacerbate the problem of identifying variants that are expressed at low levels. This approach would also require further analysis to determine the expression patterns of any novel transcript variants.

The approach used here also limited the size of alternative splicing events that can be identified, as deletions spanning several exons may be missed. As this technique focused on gene fragments rather than full length cDNA sequences it cannot be applied to identify all mRNA transcript variants definitively when multiple regions in a given transcript are subject to alternative splicing. Consequentially, this study may underestimate the true frequency of transcript variation in human Xp11.23. In these cases it is possible to use a targeted cDNA screening strategy to assess the use of such splice sites in different tissues, and developmental stages, but full-length cDNA sequencing is still required to appreciate the extent of transcript variation in the human transcriptome fully.

It is possible that transcript variations may be generated through mutations in the genomic sequence rather than alternative splicing events. These mutations may perturb the use of splice sites, alter the exonic sequence or regulatory regions of genes. This could, in turn, could alter their splicing patterns, sequence composition or expression profiles. The cause of the sequence variation can be determined by sequencing both the cDNA and its genomic DNA, after which both sequences could be aligned to a reference genome sequence. (N.B., alterations in expression patterns may also require sequencing of intronic and promoter sequences to identify mutations). Unfortunately, matching genomic DNA and RNA were not used in this analysis as the genomic DNA could not be sourced. Genomic

DNA or additional non-related RNA samples are needed to confirm the source of sequence variation recorded in the adrenal gland of *PQBP1*.

Over half of all functional alternative splicing events are conserved between human and mouse (Sorek *et al.,* 2004b). The contribution of mouse sequence information to human novel transcript identification was illustrated in the chapter where both transcript and genome comparisons identified 22 mouse specific exon junctions,12 mouse specific first exons and nine mouse final exons. Seven of these were confirmed in human by cDNA screening and sequencing. To further enhance this analysis genome and transcript sequences from other model organisms such as the rat (*R. norvegius*), dog *(C. familaris*) and opossum (*M. domesticas*) could also be used.

Much of the analysis in this chapter has focused on the identification of novel exon junctions. The transcripts of most genes included were further complicated by the use of alternative transcription start sites. Variable transcription starts sites may influence post-transcriptional gene regulation such as mRNA processing, export and translation and can reside in genomic locations that are far apart from each other. The combination of differential promoter usage and variable splicing in the 5' UTR may provide highly specific regulation of gene expression in response to a wide variety of intrinsic and extrinsic signals. In this study, three alternative first exons were identified for the gene, *RBM3*. Two of these were expressed in all tissues while the third failed the cDNA screening stage (*RBM3* cDNA screen 1). I*n vitro* expression studies have found of *RBM3* is up-regulated upon exposure to mild hypothermic conditions (Dresios *et al.,* 2005) and additional experimental analysis is required to determine if alternative promoters increase the expression of *RBM3* under such conditions.

Almost all of the reference transcript splice sites had AG-GT dinucleotides at their exon junctions suggesting that they utilise the U2 snRNP splicing machinery. These splice sites had higher splice site scores than alternative exons suggesting that they use the U2 snRNP splicing machinery during mRNA processing. The splice sites of alternative exons had lower splice site scores suggesting that they are less likely to be recognised by the U2 mediated splicing machinery. Analysis of alternative exon boundaries found that approximately 40% of the exon boundaries did not have the AG-GT dinucleotides necessary for U2 snRNP mediated splicing. This suggests that

intron excision may be mediated by the alternative U12-dependent splicing machinery. This figure is higher than the recorded use of the U12-dependent mRNA splicing pathway (Kalnina, 2004) and suggests that alternative splicing events studies identified in this study may have a preferentially use the alternative U12-dependent splicing machinery during mRNA processing or may be they generated by aberrant mRNA splicing events.

It is not clear at this stage whether the transcript variants identified in this study were produced by regulated or aberrant mRNA splicing events. The frequency of alternative splicing events observed in this study was weakly associated with gene length and exon number (Lee and Irizarry 2003). This suggests that the mRNA processing of the 18 genes included in this study was reasonably well controlled, or at the very least, chaotic mRNA processing was not taking place. This result was not entirely unexpected as regulated mRNA processing is crucial to maintain cellular viability and all samples used in these analyses were obtained from healthy individuals. It is however anticipated that mRNA splicing will not always proceed with 100% accuracy and perhaps aberrant splicing has an adaptive value by facilitating the evolution of genes.

Other genic sequence motifs can also influence mRNA splicing patterns, such a branchpoint strength, and regulatory elements such as exonic splicing enhancers (ESEs) and silencers (ESSs). For example, ESEs may compensate for weak splice sites due to greater selection pressure for weak splice sites to retain their ESE sequence (Fairbrother *et al.,* 2002). Intronic splicing branchpoints can also influence splice site selection of the beta tropomyosin gene (Libri *et al.,* 1992). The presence of these motifs in genic sequences can be predicted using a variety of computational programmes such as ESE-RESCUE which identifies ESE sequence motifs (Fairbrother *et al.,* 2002).

For functional genomic studies research programmes are underway to generate a clone resource of transcript sequences that span the full length ORF of protein coding genes. At the WTSI a cDNA cloning initiative has been implemented that uses manually annotated genome sequence to identify all protein-coding genes (Collins *et al.,* 2004). Primer sequences are designed to amplify a full-length ORF which are then amplified from pooled cDNA samples. The amplified ORFs are cloned into a holding vector and sequence verified. With greater sampling of the

cloned ORFs it may also be possible to identify more novel transcript variants and these clones may provide an additional resource for future functional studies. The power of this approach is shown in the following chapter, where primers flanking the open reading frame of one gene, *PQBP1,* were used to identify seven additional transcript variants.