Chapter 5

PQBP1 transcript diversity: comparative and expression analysis

5.1 Introduction

5.1.1 Polyglutamine binding protein 1, PQBP1

PQBP1 was first identified by two independent groups searching for novel proteins involved in protein-protein interactions (Komuro *et al.,* 1999a; Waragai *et al.,* 1999). The gene contains seven exons and spans approximately 4.6 kb of genomic DNA in human Xp11.23. It is closely flanked by its neighbours *SLC35A2* and *TIMM17B*, (Figure 5.1) and lies head-to-head with *TIMM17B* with the first exon of *PBQP1* overlapping the first exon of *TIMM17B* by 212 bp. The 3' UTR of *PQBP1* is 46 bp upstream from the 3' UTR of *SLC35A2*. The compact nature of *PQBP1* makes it suitable for further analysis as only limited number of potential splice sites can be contained within its sequence and increased intron length has been associated with increased rates of alternative splicing (Berget *et al.,* 1995). It is also predicted that full-length of *PQBP1* transcripts can be amplified using conventional PCR techniques as the reference cDNA is just over 1 kb long. Moreover, the gene is ubiquitously expressed (Iwamoto *et al.,* 2000; Kalscheuer *et al.,* 2003) and has highest expression levels observed in the brain (Komuro *et al.,* 1999; Waragai *et al.,* 1999; Waragai *et al.,* 2000; Zhang *et al.,* 2000; Okazawa *et al.,* 2001; Kalscheuer *et al.,* 2003).

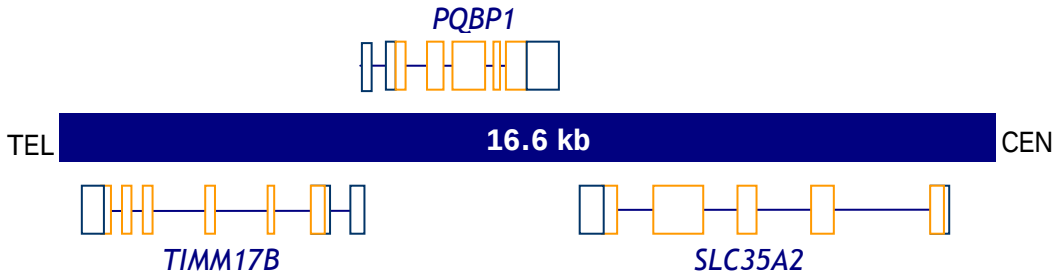


Figure 5.1 *PQBP1* **and its neighbours**
The genome sequence in the centre of the diagram is 16.6 kb long and runs from the telomere to the centromere. Transcript sequences above the genome sequence are transcribed in the forward direction while transcripts beneath the genome sequence are transcribed from the opposite stand. The transcript structures for the reference sequences of each *TIMM17B, PQBP1* and *SLC35A2* are displayed. UTR sequences are shown in blue, while coding sequence is shown in orange.

The *PQBP1* protein contains an N-terminal WW domain, a nuclear localisation signal and a unique carboxyl-terminal domain. The WW domain is a protein-protein interaction motif which has conserved tryptophan (W) residues found between an acidic region and an acidic-basic amino acid repetitive region. This domain has been shown to act as a transcriptional activator and it interacts with various proteins including POU domain, "class 3", transcription factor 2 (*PRU3F2*) and RNA polymerase II (Waragai *et al.,* 2000; Zhang *et al.,* 2000; Okazawa *et al.,* 2001). The carboxyl-terminal domain with the U5 component of the spliceosome (Waragai *et al.,* 2000). Based on these interactions and its sub-cellular localisation to the nucleus (Okazawa *et al.,* 2001; Kalscheuer *et al.,* 2003), *PQBP1* is thought to be a bridging molecule between mRNA transcription and splicing. *PQBP1* is also well conserved; potential orthologues have been identified in the mouse and in the more distantly related species, *C. elegans* and *Arabidopsis thaliana* (Okazawa *et al.,* 2001).

Work described in previous chapters has demonstrated that *PQBP1* has a high degree of transcript diversity. Fourteen alternatively spliced transcripts were identified in chapter 4 that displayed diversity in the 5' UTR and the CDS. These results supported and extended a previous study on *PQBP1* transcript diversity where PCR amplification and sequencing identified four transcript variants in the human brain (Iwamoto *et al.,* 2000). These transcripts all retained the WW domain, but two variants *PQBP1*-a and *PQBP1*-d did not contain the c-terminal domain nor the putative nuclear localisation signal. Only one of these transcripts was not identified in chapter 4 providing evidence that yet more diversity exists beyond those that have already been described.

A greater description of variation of transcripts from the *PQBP1* locus may contribute towards understanding its associated disease phenotypes. Mutations in the *PQBP1* gene are manifested as X-linked mental retardation phenotypes. Five out of 29 families studied were found to have a mutation in exon 4, that was shown to affect the sub-cellular localisation of the protein (Kalscheuer *et al.,* 2003) where isoforms were not targeted to the nucleus. Other mutations were also identified which caused frame-shifts that introduced PTCs (Kalscheuer *et al.,* 2003). In total, *PQBP1* has been associated with five syndromic X-linked mental retardation (XLMR) phenotypes and one non-syndromic condition (MRX55). These are listed in Table. 5.1.

Table 5.1 ***PQBP1*** **associated X-linked mental retardation phenotypes.**

Name	Mutation (location*)	Reference
Sutherland-Haan syndrome (MRXS3)	2 bp insertion, AG (3898)	Sutherland *et al.*, 1988 Kalscheuer *et al.*, 2003
Hamel's syndrome	2 bp deletion ,AG (3898)	Hamel *et al.*, 1994 Kalscheuer *et al.*, 2003
Golabi-Ito-Hall syndrome	A → G (194)	Golabi *et al.*, 1984
Porteous syndrome	2bp ins, AG (3898)	Porteous *et al.*, 1992
Renpenning syndrome	1 bp ins, C (64)	Lenski *et al.*, *2004* Stevenson et al., *2004* Renpenning *et al.*, 1962
MRX55	4 bp del AGAG (3896)	Deqaqi *et al.*, 1998 Kalscheuer *et al.*, 2003

*The location of the mutation is described in relation to the reference *PQBP1* transcript

It has been proposed that *PQBP1* is also involved in the pathology of neurodegenerative disorders such as spinocerebellar ataxia-1 (SCA1, Enokido *et al.*, 2002). The expansion of glutamine tracts in *ATXN1* leads to the death of cells in the cerebellar cortex. The interaction between *PQBP1* and *ATXN1* increases in proportion with the expanded glutamine sequences (Okazawa *et al.*, 2002) and the two proteins act co-operatively to repress transcription and induce cell death (Enokido *et al.*, 2002; Okazawa *et al.*, 2002). Over-expression of *PQBP1* in mice also results in the SCA1 phenotype (Okuda *et al.*, 2003).

5.1.2 Work described in this chapter

Analysis of transcript variation carried out in chapter 4 identified 14 transcript variants of *PQBP1*. Only four of the transcript variants spanned the entire open reading frame. Without full-length gene structures for the other transcript variants it is difficult to make any predictions about the affect of transcript variation on gene function. In order to overcome this limitation, it was decided to clone full-length open reading frames for one gene, polyglutamine binding protein 1, *PQBP1*. This analysis serves to test both the efficacy of the work undertaken in chapter 4 and to create a resource that could be used for downstream studies.

Comparative sequence analysis was performed to assess the evolutionary conservation of *PQBP1* transcript variants. Potential *PQBP1* orthologues were identified in eight vertebrate species. The sequence of each orthologous gene was

used to assess the evolutionary conservation of *PQBP1*, with particular emphasis on splice site conservation.

In order to further our understanding of the biological impact of transcript diversity in the *PQBP1* locus, the expression of *PQBP1* and its transcript variants in 20 different human tissues was determined by quantitative PCR. Here, it was hypothesised that alternative transcripts generated through controlled splicing events will be more abundant than those transcripts produced by aberrant splicing events. The expression profile of the reference transcript was first established to which the abundance of its variant transcripts were compared.

Results

5.2 Identifying additional *PQBP1* transcripts by random open reading frame cloning

The method employed to amplify *PQBP1* open reading frames was based upon the protocol developed by the Experimental Gene Annotation Group at the Sanger Institute (http://www.sanger.ac.uk/Teams/Team69/corf.shtml). The aim of this project is to create a cloned ORF for every protein-coding gene in the human and mouse genomes and differs from conventional full-length cDNA cloning strategies because it uses manually annotated gene structures as the initial template. To amplify the ORFs, nested primers are designed from the annotated gene structures. The transcript is amplified from either pools of MGC human, full-length cDNA clones or pooled cDNA samples (Universal cDNA® - Stratagene) using a proof reading *Taq* polymerase. Following successful amplification the size of the transcript is confirmed by agarose electrophoresis; the appropriately product is extracted from the agarose gel and purified. Before sub-cloning, the 3' ends of the purifed PCR products is adenylated to permit ligation with the holding vector, pGEM-T Easy. Following chemical transformation into *E.coli* strain JM109 individual colonies are selected for further analysis. Plasmids are purified and the fidelity of the insert confirmed by sequencing.

Two significant alterations were made to this procedure to facilitate the cloning of multiple transcript variants from a single gene. Firstly, the source of cDNA was altered from MGC cDNA clones or universal cDNA to unpooled cDNA samples prepared from total RNA as described in section 2.4.5. This was undertaken so that the source of *PQBP1* transcript variants could be traced to a single tissue type. Furthermore, to ensure that the maximum number of variant transcripts from each PCR were obtained, the PCR products were not gel purified prior to A-tailing and sub-cloning. Instead, the PCR product was purified directly after amplification (section 2.4.6).

5.2.1 Amplification, cloning and identification of alternative variants

Primer pairs *PQBP1*.corf1F & R (outer primers) and *PQBP1*.corf2F & R (nested primers) were used to perform PCR on 29 different cDNA samples (section 2.18) (Figure 5.2). The products from 5 tissues (brain, thymus, small intestine, kidney and adrenal gland) were selected for subcloning. These tissues were selected as

they displayed unique banding patterns when visualised by agarose gel electrophoresis (section 2.18, Figure 5.2).

Following shotgun cloning of PCR products in pGEMT-Easy and *E. coli* JM109 or DH10B, individual plasmids were prepared for sequencing. In total, 192 clones were sequenced using six primers, to ensure that the entire insert was sequenced. These primers were M13F and M13R which are located with the plasmid pGEM-T easy and stSG 483741S & 483741A, 483742A and 473744A which are located in the *PQBP1* insert. (N.B. These primers were not located in all *PQBP1* variants but adequate sequence coverage was obtained using 2 of the 3 primer pairs). The sequence reads were assembled and analysed using GAP4 (section 2.25.4). The resulting consensus sequence for each clone was aligned against the reference sequence for *PQBP1* (NM_003177) and the genomic sequence using Spidey (http://www.ncbi.nlm.nih.gov/spidey/spideyweb.cgi). Transcripts with novel splicing patterns were noted and the clones were retained for future analysis, while any sequences with aberrant splice sites, or less than 98% sequence identity to the reference human genome sequence were excluded from further analysis.

Fifteen variant transcripts of the *PQBP1* locus were identified. The number of clones generated for each transcript is displayed in Figure 5.3 and a multiple sequence alignment of the nucleotide sequences is displayed in Appendix IV. As expected the most abundant transcript represented in the *PQBP1* ORF collection was the reference transcript (38% of clones sequenced). Only one representative clone was sequenced for each variant transcript 5, 10, 11, 13 and 14. A description of each transcript variant is listed in Table 5.2. A diagram displaying the exon/intron structure of each transcript together with their splice site scores (section 5.2.3) is displayed in Figure 5.4. The sequences for these clones have been submitted to the GenBank nucleotide database. EMBL accession numbers for each clone can be found in Table 5.2.

Seven of the nine ORF altering transcript variants identified in chapter four were also identified in this study. The transcripts that were not confirmed here may be located in the 5' UTR (which was not assayed) or may be present in extremely low quantities.

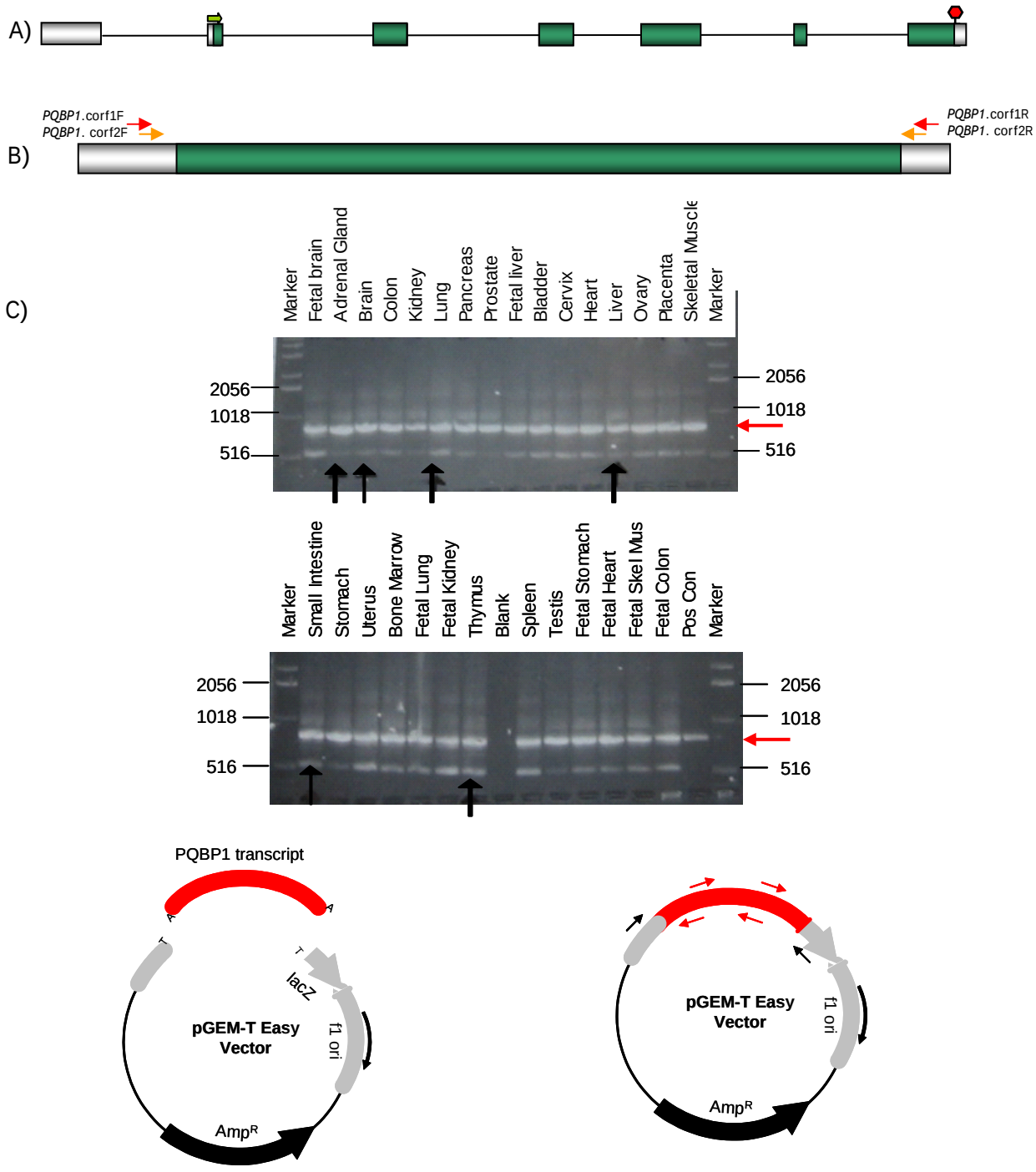


Figure 5.2 Overview of the ORF cloning protocol

A) The open reading frame representing the reference transcript was identified for *PQBP1*. Green boxes represent the location of the open reading frame, the start codon is denoted with a green arrow, while the stop codon is denoted with a red hexagon. B) Nested pairs of primers were designed to amplify the entire open reading frame. C) Agarose gel visualisation of the amplified ORF for 29 different tissues. Tissues selected for further analysis are denoted with a black arrow, while the expected reference transcript size PCR product is highlighted with a red arrow. Pos con = Universal cDNA (Stratgene); Blank = $T_{0.1}E$ negative control. D) A-tailing of the resulting purified PCR products and ligation with the holding vector pGEM T- Easy. Black arrows depict the orientation and location of sequencing primers located in the host plasmid, while red arrows denote the location of sequencing primers in the *PQBP1* transcripts.

Table 5.2 Description of *PQBP1* transcript variants

The tissue from which the transcript was identified is also listed. (Acc = intron acceptor, Don = intron donor)

Transcript	EMBL ACCESSION	Isolated from (tissue)	Description
Reference	NM_003177	All (except adrenal gland)	6 exons
1	AJ973593	Brain	Loss of exon 4
2	AJ973594	Adrenal Gland	21 bp deletion exon 4
3	AJ973595	Adrenal Gland	21 bp deletion exon 4, intron 4 retained
4	AJ973596	Adrenal Gland	21 bp deletion exon 4, 132 bp deletion exon 4 (don)
5	AJ973597	Adrenal Gland	21 bp deletion exon 4, 3 bp deletion exon 5 (acc)
6	AJ973598	Thymus	132 bp deletion exon 4 (don)
7	AJ973599	Brain	Retains intron 4
8	AJ973600	Small Intestine	Retains introns 4 and 5
9	AJ973601	Brain	Novel exon 2a
10	AJ973602	Kidney	304 bp addition exon 2 (don)
11	AJ973603	Small Intestine	Novel exon 2a, retention of intron 4
12	AJ973604	Small Intestine	195 bp deletion exon 4 (don), loss of exon 5, 105 bp deletion exon 6
13	AJ973605	Kidney	90 bp deletion exon 3 (don), 108 bp deletion exon 4 (acc)
14	AJ973606	Small Intestine	3 bp deletion exon 5 (acc)
15	AJ973607	Brain	17 bp deletion exon 2 and loss of exon 4

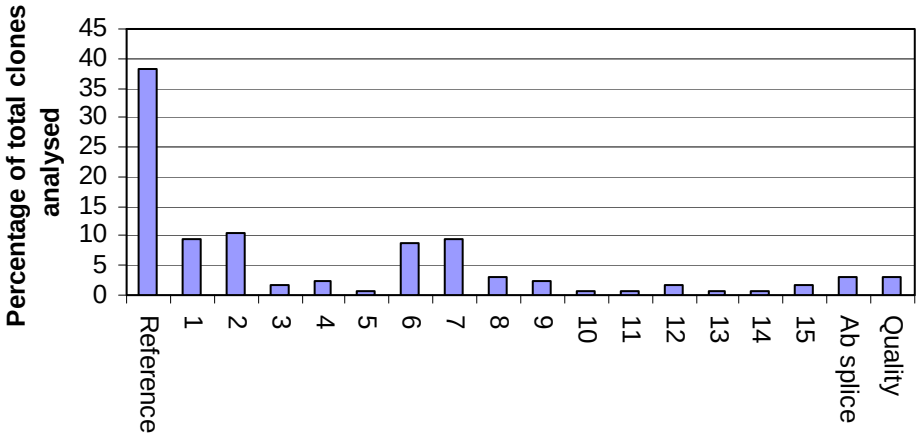


Figure 5.3 Frequency of clones representing different *PQBP1* variants

Ab splice - clones rejected as they displayed aberrant splicing patterns in more than 2 introns. Quality – clones rejected due to poor quality sequence reads.

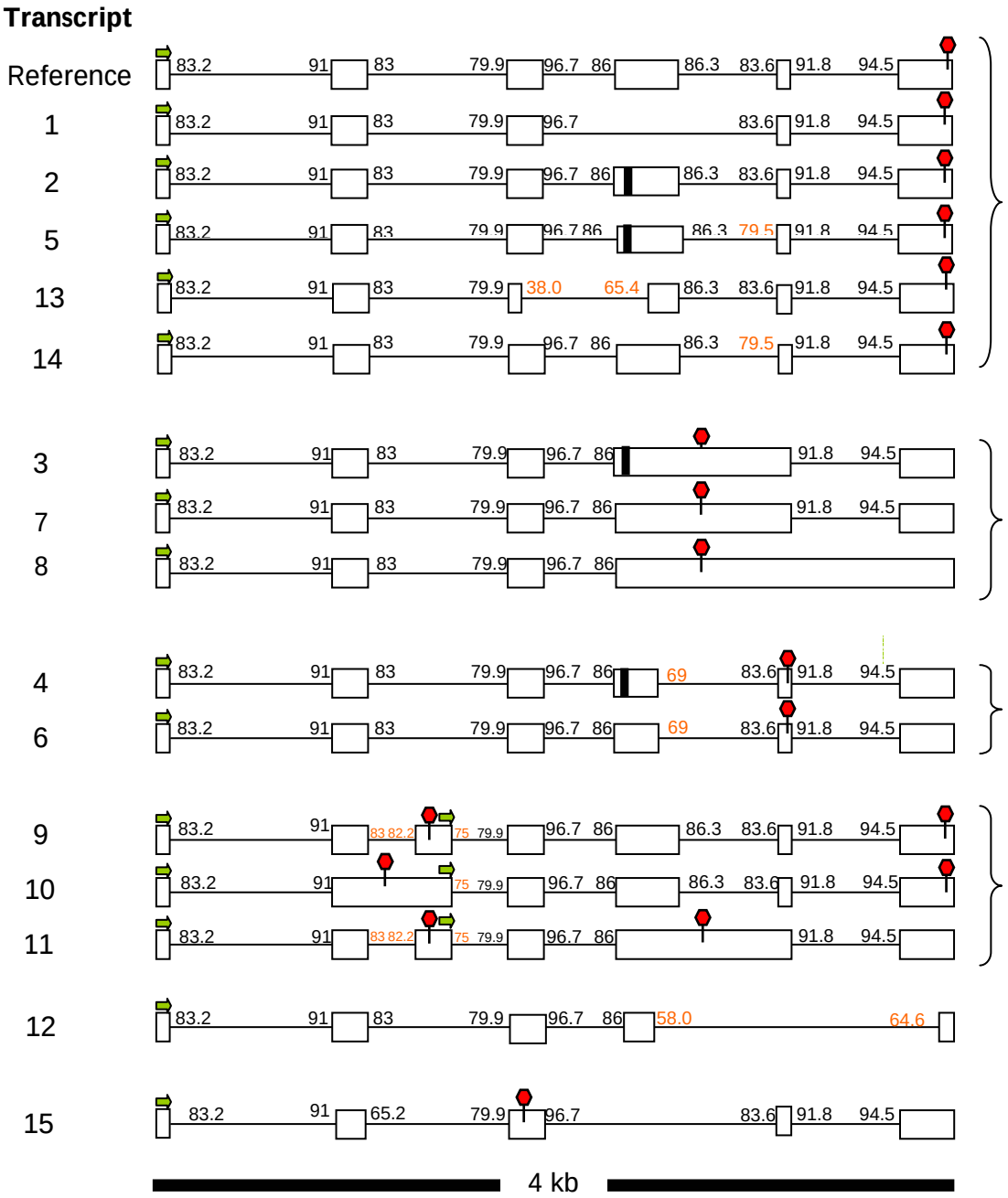


Figure 5.4 Exon/intron structures of *PQBP1* alternative transcripts

Boxes represent the approximate size and location of exons for each *PQBP1* transcript variant. Possible translation start sites (green arrows) and termination codons (red hexagons) are also displayed. Black bars represent a 21 bp deletion that is exclusive to transcripts identified from the adrenal gland. Acceptor and donor splice site scores that were determined using Shapiro and Senapathy's algorithm (Shapiro and Senapathy 1987) are displayed to the left and right of each exon. The transcripts are clustered according to the predicted location of its stop codon(s). Black splice site scores represent the scores of reference splice sites while orange splice site scores denote the scores of alternative splice sites.

5.2.2 Description of transcript variants

The *PQBP1* alternative transcripts were clustered into groups that share common exon junctions as discussed below.

Transcripts 2-5

A 21 bp deletion from exon 4 was shared between transcripts 2, 3, 4 and 5. This variation was discussed in greater detail in section 4.3.6. These samples were included in future functional studies in order to assess the impact of this sequence variation on *PQBP1* function.

Transcripts 9-11

These transcripts have a novel 89 bp exon which is located within an *AluSq* repeat. The location of this exon may have been missed from previous annotations as, like all other repetitive elements, it would have been masked out of the genome sequence by RepeatMasker. This particular alternative splicing event was been discussed in greater detail in section 4.3.6.

Transcripts 5 and 14

These clones both have 3 bp deleted from the 3' exon 5 acceptor site. The open reading frame is maintained but the encoded protein is one amino acid shorter than the reference protein.

Transcripts 3, 7, 8 and 11

These transcripts retain intron 4.

Transcripts 4 and 6

Share a common 132 bp deletion from the donor site of exon 4.

Transcripts 1 and 15

Exon 4 is skipped in both of these transcripts.

5.2.3 Analysis of splice site scores for PQBP1 alternative transcripts

Splice sites scores were calculated for each exon-intron boundary using the programme Splice Site Finder (http://www.genet.sickkids.on.ca/~ali/splicesitefinder.html). This programme is

based upon the algorithm developed by Shapiro and Senapathy (1987). Donor and acceptor splice site scores are displayed in Figure 5.4.

The reference *PQBP1* transcript has 10 splice sites, yet only 3 were used in all *PQBP1* transcript variants (1 donor and 2 acceptor splice sites). Greatest variability in splice site use was observed around the exon 4 donor and exon 5 acceptor sites. Sixty percent of the transcripts did not use the same exon 4 donor site as the reference transcript and 47% of the transcripts do not use the same exon 5 acceptor splice site. When alternative exon 4 donor and exon 5 acceptor splice sites were used, their splice site scores were not significantly different from the reference (donor 86.3 v 88, acceptor 83.6 v 87).

All other alternative donor or acceptor splice site scores were slightly weaker than those of the reference splice sites. Exon 4 is skipped in both alternative transcripts 1 and 15 but the splice site scores for this exon were not significantly different from the average splice site scores for the all other exons (acceptor 86 v 87, donor 86.3 v 88). Likewise, the splice site scores of the novel exon (exon 2a) do not differ from the average. Transcripts 4 and 6 use an alternative donor splice site whose strength is greatly reduced compared to the splice site score for the reference transcripts. The splice site score decreases from 83.6 for the reference splice site to 69 for the alternative splice site.

These results suggest that splice site strength does play a role in the selection of splice sites used when processing *PQBP1* transcripts. However, additional motifs such as regulatory elements may also influence the splicing patterns.

A total of 16 transcript variants of the gene *PQBP1* (including the reference sequence) were identified by screening cDNA from five different human tissues. Additional analysis was required to evaluate how the biological function of the alternative transcripts differs from that of the reference transcript. Therefore, these transcripts were cloned into a holding vector pGEM T-easy which is a suitable resource for future functional analysis.

The remainder of this chapter is focused on further characterisation of the *PQBP1* transcripts. Preliminary inferences concerning the biological significance of

transcript variation in *PQBP1* were generated by assessing the evolutionary conservation and tissue expression patterns of the alternative transcripts.

5.3 Comparative analysis of *PQBP1* locus

Comparative analysis can identify regions that are under selective pressure to remain conserved throughout evolution. This approach has been used to identify functional elements in the human genome including genes, alternative exons and regulatory elements (Dubchak and Frazer 2003; Frazer *et al.,* 2003) Comparative analysis of the *PQBP1* locus has been carried out using gene and transcript sequences from eight vertebrate species (Table 5.3) in order to assess the conservation of *PQBP1* splice sites throughout vertebrate evolution.

Table 5.3 Species and sequence assembly versions used in comparative analysis.

Species	Common Name	Genome Release
H. Sapiens	**Human**	NCBI35
M. musculus	**Mouse**	NCBI m33
R. norvegicus	**Rat**	RGSC 3.1
C. familiaris	**Dog**	CanFam1
P. troglodytes	**Chimpanzee**	CHIMP1
M. domestica	**Opossum**	Version 0.5
D. rerio	**Zebrafish**	WTSI Zv4
F. rubripes	**Fugu**	Fugu 2.0

5.3.1 Comparative Gene Analysis

Identification of orthologous genes

In order to identify potential orthologues of *PQBP1*, the RefSeq protein sequence of human *PQBP1*, (NP_005701), was used to interrogate the peptide and nucleotide databases at Ensembl (version 27) and the NCBI. BLASTp searches were performed to interrogate protein databases, while TBLASTX searches were performed to interrogate nucleotide databases with a peptide sequence. This search identified two homologues in all species except *P. troglodytes* and *H. sapiens*, where only one potential homologue was identified. The corresponding nucleotide sequences were analysed for evidence that the genes were related. This included analysis of the exon/intron structures of each transcript and the gene neighbourhoods surrounding the predicted *PQBP1* loci. Accession numbers, chromosomal location and percentage sequence identity to human *PQBP1* are listed in Table 5.4.

Table 5.4 Identification of homologues of the *PQBP1* locus in other vertebrates

Species	Chr	(Predicted) Gene sequence	% identity to Hs *PQBP1*(aa)
H. sapiens	X	*PQBP1*	100%
P. troglodytes	X	*PQBP1* - ENSPTRG00000021873	>99%*
C. familaris	X	ENSCAFG00000015672	93%
M. musculus	X	*PQBP1*	87%
R. norvegicus	X	ENSRNOG0000000776	86%
M. domestica	Unk	Built_from_Q9QYY2_And_Others_2	74%
D. rerio	8	*PQBP1*l - OTTDARG000000136	48%
D. rerio	8	Q90X39 - ENSDARG00000030032	48%
F. rubripes	Unk	SINFRUG00000152774 (fugu 1)	62%
F. rubripes	Unk	SINFRUG000000157836 (fugu 2)	57%

* if 1 bp insertion discussed below is ignored
Unk = unknown chromosomal location

The exon-intron sizes of the putative *PQBP1* orthologues are listed in Table 5.5 where it was noted that the exon sizes were similar between all species for at least one gene, except for exon 4 in the opossum which was 89 bp long (196 bp shorter than the human exon 4). Other changes of interest include a 1 bp insertion located 23 bp downstream from the predicted translation start site in the predicted *P. troglodytes* (chimpanzee) orthologue. The insertion would result in a frame shift and the subsequent loss of the open reading frame. It is possible that this insertion may be a species specific event, or the result of a sequence assembly error. However, searches of both the *Pan troglodytes* EST database and trace archive (http://www.ncbi.nlm.nih.gov/Traces/trace.cgi?) failed to identify any other evidence for the presence of the insertion. This insertion is likely to represent an error in the genomic sequence.

Other variations were observed in the exon length of *PQBP1* homologues. These differences tended to occur in multiples of 3 which would have maintained the ORF. For example, exon 5 is 64 bp long in the human, chimp, mouse, rat and dog but is 3 bp shorter in the opossum and 6 bp shorter in the *Fugu*. This exon has expanded by 12 bp in the zebrafish.

Table 5.5 Exon/Intron Structure of *PQBP1* homologues

Species	E	I	E	I	E	I	E	I	E	I	E
Human	**67**	2607	**112**	628	**113**	190	**285**	214	**64**	132	**209**
Chimp	**68**	2612	**112**	626	**113**	190	**285**	214	**64**	132	**215**
Mouse	**67**	2138	**112**	195	**113**	110	**279**	375	**64**	297	**345**
Rat	**67**	2338	**112**	175	**113**	106	**279**	184	**64**	155	**157**
Dog	**67**	2339	**112**	267	**113**	210	**285**	228	**64**	197	**157**
Oposs	**67**	1493	**112**	189	**110**	106	**89**	452	**61**	140	**157**
Fugu1	**67**	174	**112**	208	**101**	58	**234**	86	**58**	492	**49**
Fugu2	**67**	88	**112**	191	**101**	210	**234**	228	**58**	191	**157**
Zeb 1	**67**	91	**112**	840	**101**	3670	**N.I**	N.I	**76**	417	**359**
Zeb 2	**67**	91	**112**	838	**101**	3670	**N.I**	N.I	**76**	417	**359**

N.I., not identified E = exon, I = intron Sizes are given in bp.

BLAST analysis also identified additional homologous, but unspliced matches within the mammals *M. musculus, R. norvegicus, M. domestica,* and *C. familiaris* (Table 5.6). Each of these had characteristics that are usually associated with retroposed pseudogenes:

- The sequence identity for these matches was lower than that between human *PQBP1* and (spliced) orthologue, suggesting that these intronless copies are not under the same selective pressures to remain conserved.
- These genes did not cover 100% of the human *PQBP1* locus. They were truncated, and harboured premature termination codons.
- BLAST searches against each species' EST database failed to confirm transcription from these loci. However, transcription was confirmed for the of *PQBP1* orthologues (100% identity matches to EST sequences).
- The second intronless copies of *PQBP1* were not located on the X chromosome. A chromosomal location could not be assigned to the second *PQBP1* match in the opossum.

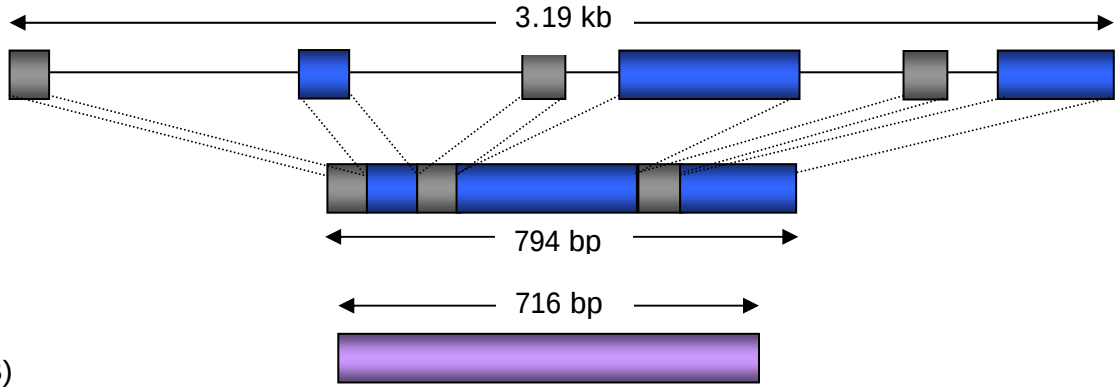
It is therefore suggested that the mouse, rat, dog and opossum have one functional copy and one retroposed non-functional copy of *PQBP1*. An example of the retroposed copy of *PQBP1* identified in the dog is displayed in Figure 5.5.

All subsequent analysis was completed using the orthologous copy of *PQBP1* from the mouse, rat dog and opossum which have introns and were mostly found on the X chromosome. Sequences from both *PQBP1* copies in *D. rerio* and *F. rubripes* have been used.

Table 5.6 Potential *PQBP1* retroposed pseudogenes

Species	Chr	Retroposed pseudogene	Chr	Amino acid id to human *PQBP1* (%)
Human	X	None	n.a.	n.a.
Chimp	X	None	n.a.	n.a.
Dog	X	ENSCAFG00000004230	19	90%
Mouse	X	ENSMUSG00000021001	12	83%
Rat	X	ENSRNOG00000015114	18	94%
Opossum	Unk	Built_from_Q80WW2_ And_others_1	Unk	82%

A)



B)

C)

```
PQBP1:ATGCCGCTGCCTGTTGCATTGCAGACCCGCTTGGCCAAGAGAGGCATCCTCAAACATCTGGAGCCCGAACCAGAGGAAGAGATTATTGCTGAGGACTACG
pseud:.................G.........................C.............G....C....G.....................C...

PQBP1:ATGATGACCCTGTGGACTATGAGGCCACCCGGTTGGAGGGCCTGCCACCAAGCTGGTACAAAGTGTTTGACCCATCCTGTGGGCTCCCTTACTACTGGAA
pseud:.............................C.................C....C...--.......TGA.........C........C.-......C..

PQBP1:TGTGGACACCGACCTCGTATCCTGGCTCTCCCCGCATGACCCCAACTCCGTCGTTACCAAATCTGCCAAGAAACTCAGGAGCAATAATGCAGATGCTGAG
pseud:.................CG..............................C.............................................

PQBP1:GAGAAGCTGGAGCGTGGCCACGAGAAATCAGACCGGAACCATGAGAAGCCAGACCGGAGCCACGAGAAGTCAGACCGGAGCCATGAGAAGTCGGACCGGA
pseud:.............................G...----------------------------------------------------------------.

PQBP1:GCCATGAGAAGTCTGACAGAGATCGAGAACGTGGTTATGACAAAGTGGACAGAGAGAGAGACCGGGACCGGGACCGGGACCGGGACCGCGGGTATGACAA
pseud:....C.G.G....C.....................G.....................................------.............G

PQBP1:GTCAGACAGGGAAGAGAGCAAGGAGCGGCGCCATCATCGCAGAGAGGAGCTGGCTCCCTACCCCAAGAGCAAGAAGGCGGCAAGCCGGAAGGATGAAGAG
pseud:.......C...G........G.............A.T............C......G.....-....T..G........G......-...C....--

PQBP1:TTAGACCCCATGGACCCCAGCTCCTACTCGGACGCGCCTCGGGGCACGTGGTCAACAGGACTCCCCAAGCGGAATGAAGCCAAGACGGGCGCAGACACCA
pseud:.........G........G....C.....................AC..C.................G..............A.............

PQBP1:CAGCAGCGGGGCCGCTCTTCCAGCAGCGTCCCTACCCGTCTCCGGGAGCTGTGCTCCGGGCCAATGCTGAGGCCTCCCGAACCAAGCAACAGGACTGA
pseud:.....................-.............G.G-...........C.......-.....--.C...........G..........T......
```

Figure 5.5 Identification of a *PQBP1* like pseudogene in *C. familiaris*.

(A) The exon structure of the predicted *PQBP1* orthologue spanning 3.19 kb is shown and its predicted transcript is 794 bp in length. Exons are coloured alternatively (blue and grey). (B) The intronless pseudogene is contained in one exon spanning 716 bp (purple). (C) Nucleotide alignment of the predicted *PQBP1* transcribed sequence (coloured according to alternating exons) and the pseudogene sequences. Sequence identity is indicated by a dot (.) and gaps are shown by dash (-) . A premature termination codon (PTC) in the pseudogene sequence is displayed in red.

Possible *PQBP1* gene duplication in the fishes.

Analysis of the genome sequence identified two copies of the *PQBP1* locus in each of the fish species included in the study. This observation is not entirely unexpected as it has been hypothesised that a fish-specific whole genome duplication event occurred before the origin of the teloeosts (approximately 450 mya) and that this addition of genomic material may have facilitated their radiation (Amores 1998; Meyer and SchartI, 1999; Taylor 2003; Christoffels, 2004).

The two predicted *PQBP1* like loci for *D. rerio* displayed extremely high levels of sequence similarity to each other. All exons and introns were identical in size, except for a 2 bp deletion in intron 3. The high level of sequence similarity between these loci, together with the similarity of gene neighbourhoods could reflect either a recent gene duplication event, a gene conversion event or a sequence assembly error. It is unlikely that this observation was the product of a sequence assembly error as both copies of the *PQBP1* gene are found within assembled BAC sequences and not WGS assemblies.

The sequence homology between the two putative *PQBP1* loci in F. *rubripes* was not as pronounced as that observed in *D. rerio*. The two predicted *Fugu* orthologues shared 57% and 62% identity with the human *PQBP1* reference protein (proteins encoded by the *F. rubripes* genes SINFRUG00000157836, SINFUG00000152774 respectively) and 95% identity with each other. The majority of the sequence differences were observed at the carboxyl end of the protein (Figure 5.6). The exon sizes were consistent between the two loci while all of the introns differed in size. A comparison of the gene neighbourhood surrounding the putative orthologues was not possible, as one of the predicted orthologues was located in a small sequence scaffold (scaffold_9097-Fugu 2.0) without any neighbouring genes. The other predicted orthologue was embedded in a much larger scaffold (scaffold_998–Fugu 2.0).

A) Alignment of predicted human and *F. rubripes PQBP1* protein sequences.

```
human : MPLPVALQTRLAKRGILKHLEPEPEEIIAEDYDDDPVDYEATRLFGLPPSWYKVFDPSCGLPYYWNADTDLVSWLSPHDFNSVVTKSAK
fugu1 : MPLPPALLARLAKRGIVKPSDQEVDEEIIAEDYDDNNVDYEATKHENLPPNWYKVFDPACGLPYYWNVETDLVAWLSPNDPSSVVTKAAK
fugu2 : MPLPPALLARLAKRGIVKPSDQEVDEEIIAEDYDDNNVDYEATKHENLPPNWYKVFDPACGLPYYWNVETDLVAWLSPNDPSSVVTKAAK

human : KLRSSNADAEEKLDRSHDKSDRGHDKSDRSHEKLDRGHDKSDRGHDKSDRDRERGYDKVDRERERDRERDRDRGYDKADREEGKERRHHR
fugu1 : KVR----------------AWLWFVSFSSPAFIAEGGEERTERHEKLEREREREREREREREREREREREREREREREREREREREREREPR
fugu2 : KVR----------------AWLWFVSFSSPAFIAEGGEERTERHEKLEREREREREREREREREREREREREREREREREREREREREPR

human : REELAPYPKSKKAVSRKDEELDPMDPSSYSDAPRGTWSTGLPKRNEAKTGADTTAAGPLFQQRPYPSPGAVLRANAEASRTKQQD
fugu1 : RNEIAPYSKSKRGK--KDDEMDPMDPSAYSDAPRGSWSSGLPKRNEAKTG----------------------------------
fugu2 : RNEIAPYSKSKRGK--KDDEMDPMDPSAYSDAPRGSWSSGLPKRNEAKTGADTTAAGPLFQQRPYPSPGAVLRANA---------
```

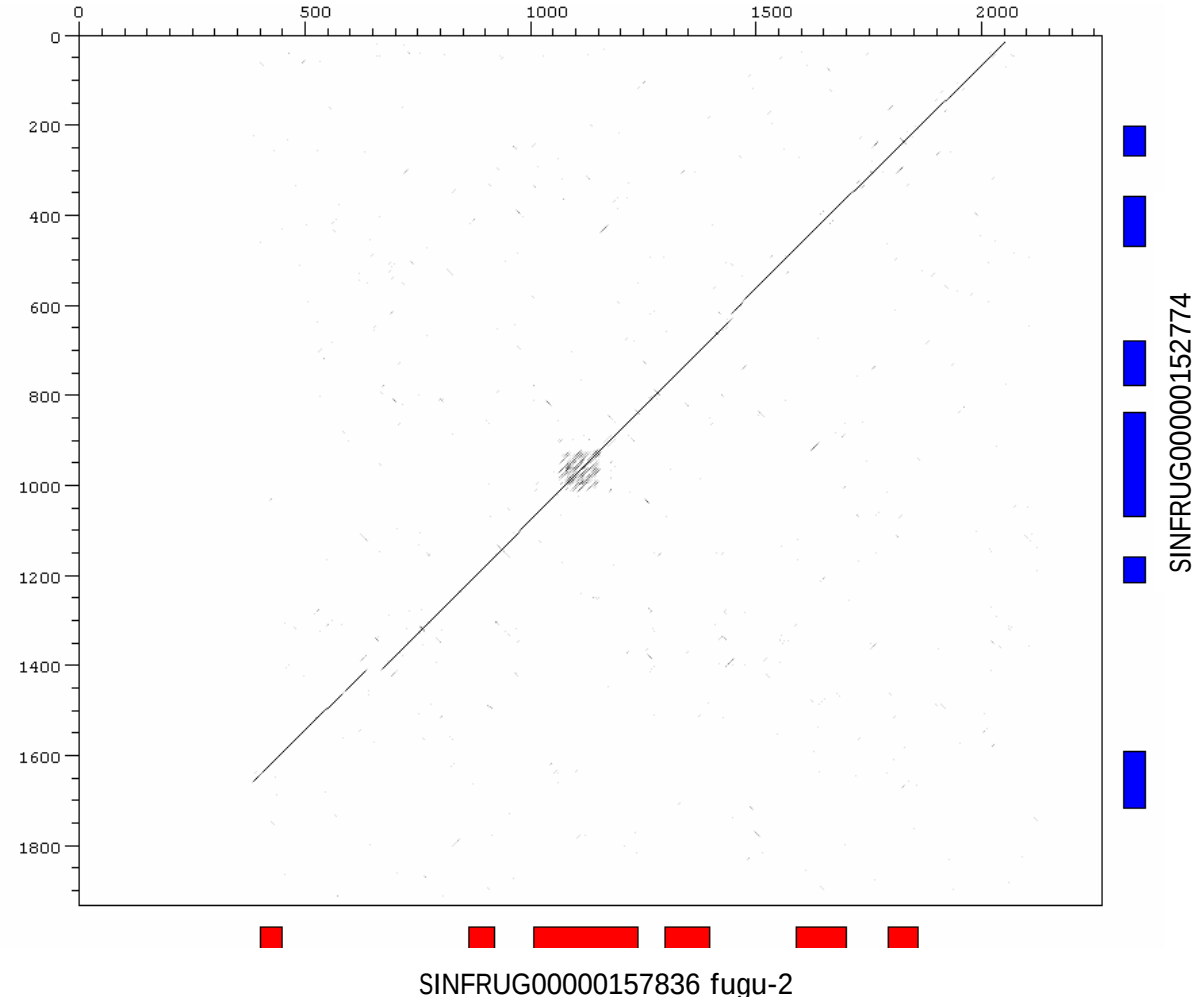


Figure 5.6 Potential duplication of *PQBP1* locus in *F. rubripes*

Fugu -1 SINFRUG00000157836 extracted from scaffold_9097 (assembly Fugu 2.0)
Fugu -2 SINFUG00000152774 extracted from scaffold _998 (assembly Fugu 2.0)

A) Amino acid alignment of predicted *PQBP1* proteins in *F. rubripes* and human *PQBP1*. Amino acids residues shaded in black are conserved in all 3 proteins while residues shaded in grey are conserved in the two *F. rubripes PQBP1* proteins.

B) Dotter alignment highlighting the sequence conservation between the two loci. Exon structures of these genes are displayed as coloured boxes.

5.3.2 Phylogenetic analysis of PQBP1 peptide sequences

To determine the relationships between *PQBP1* loci, peptide sequences for the full length *PQBP1* sequences of nine species. These were chosen as they represent a large group of highly divergent eukaryotic organisms whose genomes have been sequenced and are readily available. The (predicted) peptides were identified by BLASTp analysis and were extracted from Ensembl (v27). Following manual editing, the resulting alignment was entered into phylo_win (Galtier *et al.,* 1996) and phylogenetic analysis was performed using the neighbour joining method (section 2.28.3, Figure 5.7).

The phylogenetic analysis suggested that the two copies of *PQBP1* in each of the fishes, *D. rerio* and *F. rubripes* arose independently after their divergence from a common ancestor.

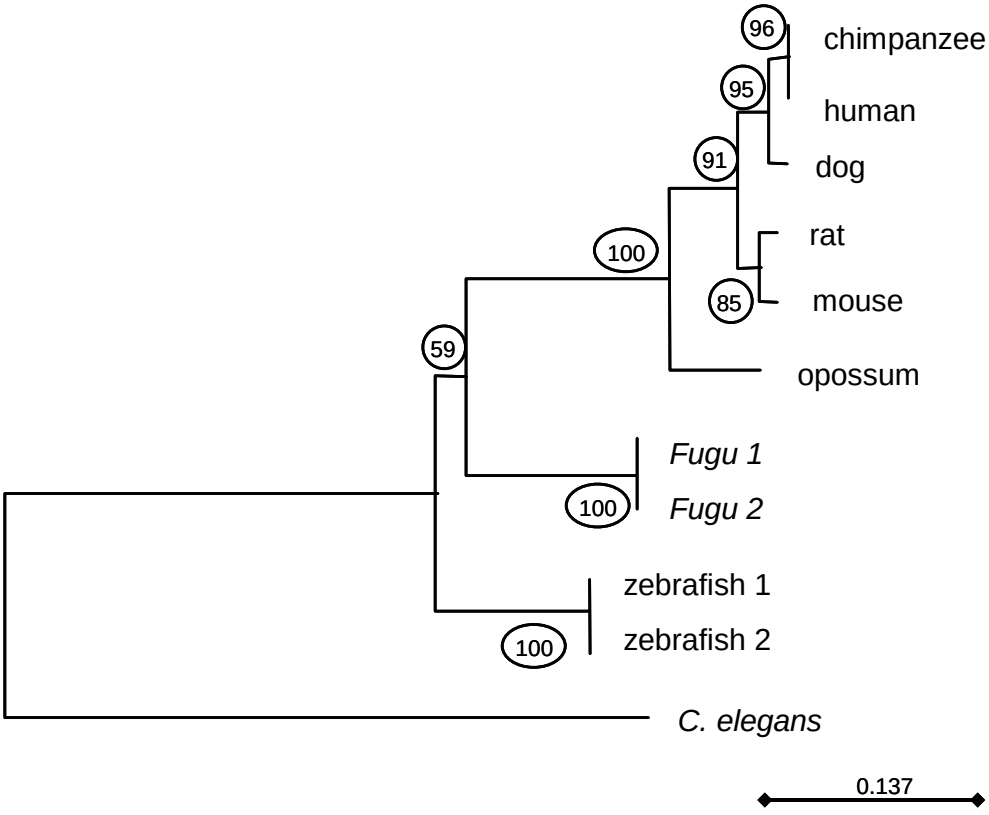


Figure 5.7 Phylogenetic tree of *PQBP1* peptides

PQBP1 transcripts were extracted from the protein database at the NCBI and aligned using clustalw. Phylogenetic analysis was completed using Phylo_win using the nearest neighbour methods with 500 bootstrap replicates. Bootstrapping values are indicated in circles. Duplicated genes accessions are: *F. rubripes* 1 (SINFRUG00000157836), *F. rubripes* 2 (SINFRUG00000152774), *D. rerio* 1 (*PQBP1*l), *D. rerio* 2 (ENDARG00000030032). *C. elegans* was included to provide an outgroup for this analysis. The scale represents the number of substitutions per site.

5.3.3 *Comparative genome analysis of the PQBP1 locus in eight vertebrate species*

The sequence conservation of *PQBP1* loci from eight different vertebrate species was analysed in order to define the evolutionary history the *PQBP1* gene further. Particular emphasis was placed upon the evolution of *PQBP1* alternatively spliced exons (exons 2a and exon 4). Genome sequences encompassing the *PQBP1* loci were extracted the UCSC genome browser or Ensembl (section 2.11). In the species where *PQBP1* has been duplicated, both loci were extracted for further analysis. The sequence assemblies and chromosome co-ordinates of the genomic DNA used in this analysis are listed in Table 5.7. Sequences were aligned using the programme zPicture (http://zpicture.dcode.org), which extracts the genome sequence directly from the UCSC genome browser (not available for all species) and automatically generates an annotation file for the species of interest. The programme aligns sequences using the local blast alignment programme blastz, and identifies and tags evolutionarily conserved regions that meet user defined thresholds. In this case, the imposed threshold was a sequence identity greater than 70% for at least 100 bp. The resulting alignments, including the location of ECRs are displayed in Figure 5.8.

Table 5.7 Sequence assemblies and chromosome co-ordinates of genomic sequences used for comparative analysis

Species	Common Name	Genome Release	Chromosome (or scaffold) co-ordinates
H. Sapiens	**Human**	NCBI35	chrX:48509000-48519000
M. musculus	**Mouse**	NCBI m33	chrX:6181000-6190000
R. norvegicus	**Rat**	RGSC 3.1	chrX:26636000-26645000
C. familiaris	**Dog**	CanFam1	chrX:41870000-41878000
P. troglodytes	**Chimpanzee**	CHIMP1	chrX: 49844000-49854000
M. domestica	**Opossum**	Version 0.5	Scaffold_11400:50000-57000
D. rerio	**Zebrafish**	WTSI Zv4 -1 - 2	chr8:29854000-29864000 chr8:30100000-30110000
F. rubripes	*Fugu*	Fugu 2.0 – 1 - 2	Scaffold_9097:1-1760 Scaffold_998:62000-68000

The zPicture analysis of the aligned genome sequences shows that the exons of the *PQBP1* reference transcript are conserved in all species except the fishes where exons 4 and 5 were not identified. Both of these exons were identified in the *Fugu* using BLAST analysis (section 5.3.1), which suggests that the inbuilt BLASTz parameters used by zPicture are not as sensitive as TBLASTX search parameters. Likewise, exon 5 in the zebrafish was detected TBLASTX but not blastz but exon 4

was not identified using either of the BLAST programmes. The lower sequence identity shared by exons 4 and 5 in the human and fish suggests that these exons may have been added later in vertebrate evolution or that they may have diverged more rapidly in the fishes. However, it is predicted that these exons have undergone more rapid divergence, because they were detected by TBLASTX analysis in the *fugu* and both the human and fish *PQBP1* homologues share a similar gene structure (section 5.3.1).

Exon one was identified in all species except either of the duplicate zebrafish genes.

Of additional interest was the identification of a non-coding intronic ECR that is shared between human, chimpanzee, dog, rat and mouse. The conservation of this ECR ranged between 99% (chimpanzee) and 72% (mouse) and spanned 213 bp (chimpanzee) to 147 bp (rat). The region has remained conserved between species that diverged from a common ancestor approximately 90-100 million years ago, but this cannot be dated in the marsupial (opossum), which shared a common ancestor with the eutherian mammals approximately 180 million years ago. This region may encode an alternative exon or may be a regulatory region. However, despite an in depth analysis of this locus, no evidence for transcription of this ECR was observed in the present study.

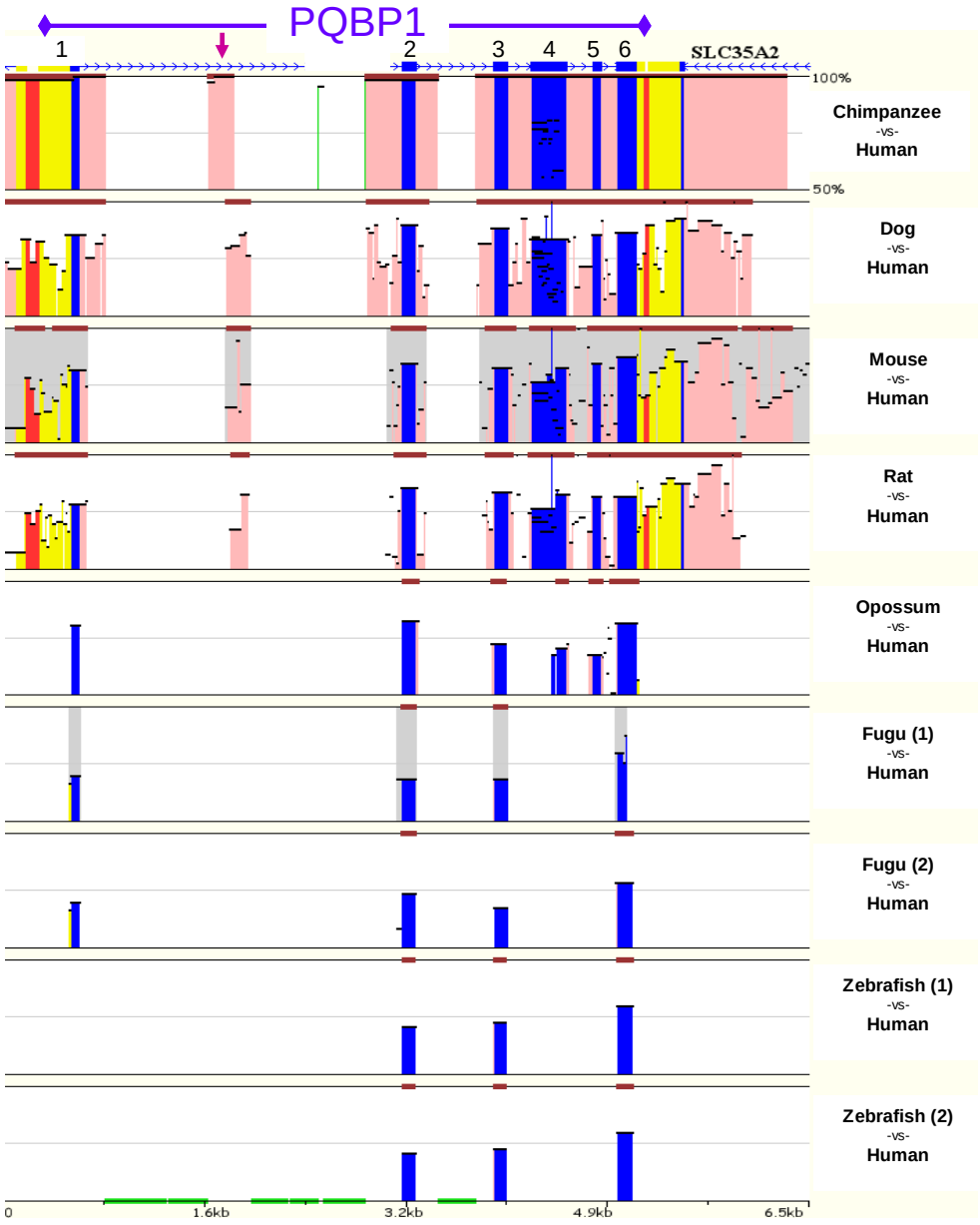


Figure 5.8 Comparative Genome Analysis of the *PQBP1* locus

Genome sequences from 8 different vertebrate species were aligned using zPicture (http://zpicture.dcode.org). All species were aligned to the reference species, human. The human sequence runs from left to right while the percentage similarity shared between the two sequences is displayed on the y-axis (lower limit display is 50% identity). Regions of conserved sequences between any two species are displayed as black horizontal bars, where the length of the bar represents the length of the conserved region (bp), and the height of the bar represents the percentage identity. Evolutionarily conserved regions (ECRs, defined as 100 bp of sequence identity over 70%) are displayed as brown boxes above the aligned sequences. Additional features displayed are: conserved coding regions (blue), conserved UTR (yellow), conserved introns (pink); conserved intergenic regions (red); masked repeats (green). All *PQBP1* exons are numbered above the diagram. A pink arrow denotes a ECR whose expression has not been confirmed. The additional matches at a lower percentage identity that are seen within exon 4 denote its repetitive nature.

5.3.4 Sequence variation around splice sites

Constitutive splice sites

The generation of aligned genome sequences allowed the opportunity to assess the sequence conservation around the splice site boundaries, for both constitutively, and alternatively expressed exons. Genomic sequences eight bp upstream and five (donor) or six bp downstream of each intron donor and acceptor site were extracted for analysis. Pictorial representations of the base usage around the exon boundaries of the reference transcripts are displayed in Figure 5.9. From this analysis it appears that the intron/exon junctions have remained well conserved, with little deviation from the AG-GT consensus splicing sequences. The lowest degree of variation was observed in exon 5 acceptor site where 10 out of the 12 sites analysed displayed variation. However, the AG acceptor dinucleotide was conserved in all species.

Alternative splice sites

Interspecies comparisons were also carried out to assess the conservation alternative splice sites. Here, sequences were extracted from the UCSC genome browser and manually aligned to view the conservation (Figure 5.10). Please note that splice sites could not be identified for each species analysed. In total, the conservation of six alternative splice sites were evaluated in seven species. Most of these sites displayed a high degree of conservation (71% - 100). Not all alternative splice sites were conserved in each of the species analysed. As expected, when compared to human splice site sequences, greatest variation was observed in the fishes, *Fugu* and zebrafish. These species displayed variation in 4/6 alternative splice sites that had the potential to affect the splice site selection. For example, the alternative exon 4 donor site recorded in transcripts 4 and 6 was not conserved in either *Fugu* (GA) or zebrafish (CG). The fish species also contained a 1 bp insertion at the exon/intron boundary which may affect the ability of the splicing machinery to recognise the alternative splice site. Additional analysis is required to determine the diversity of *PQBP1* transcript structures in orthologous genes. This could be achieved by cloning and sequencing cDNA samples from each species and by using existing EST data.

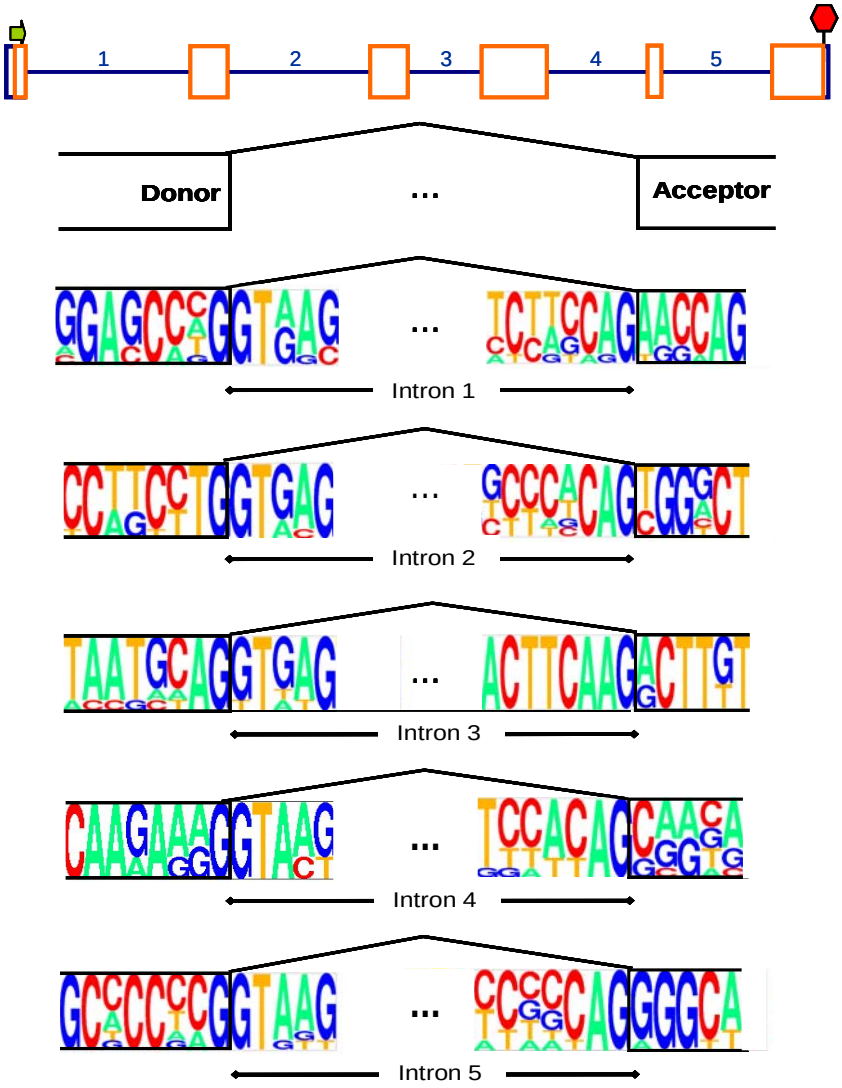


Figure 5.9 Sequence variations around exon/intron junctions in different species

Sequences were extracted from the sources outlined in Table 5.3 and aligned by blastz multiple sequence alignment. The height of each base in the pictogram is proportional to its frequency at that location.

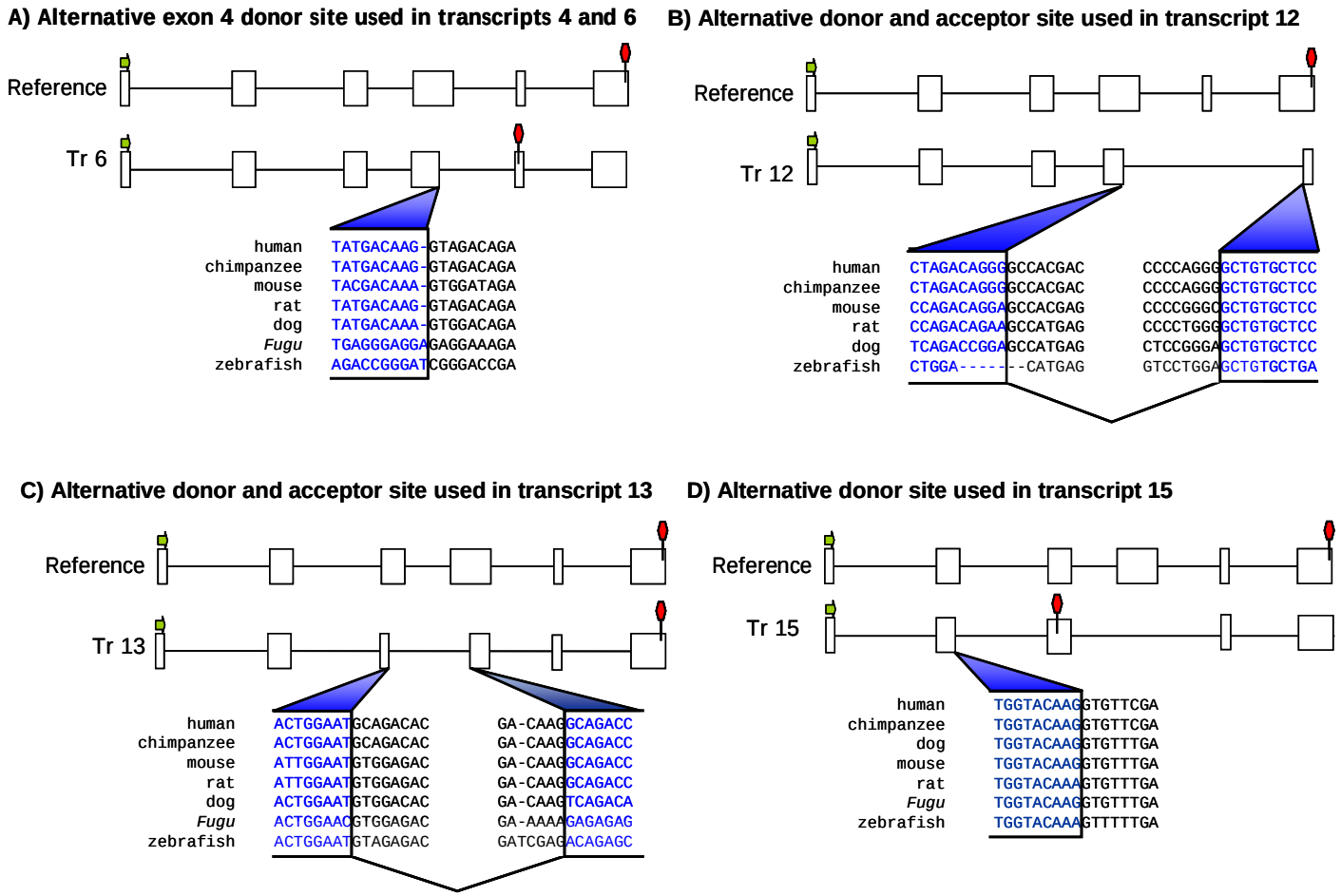


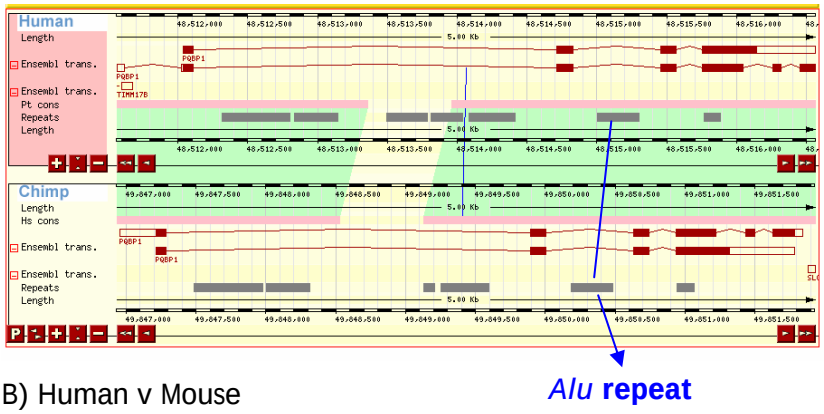
Figure 5.10 Multiple sequence alignments of splice sites used in alternative transcripts

Sequences were extracted from the UCSC genome browser using the Comparative Track for seven different species and were manually aligned. Coding sequences (in accordance with the alternative transcript) are shown in blue. Non-coding sequences are shown in black.

5.3.5 Conservation of exon 2a

The conservation of the novel exon (exon 2a) identified in this study was also analysed. This exon, which lies in *Alu* repeat, could only be identified in *H. sapiens* and *P. troglodytes,*where the two species share 98% sequence identity for this exon. *Alu* repeats are primate specific, and it was not anticipated that exon 2a would be observed in any of the other species analysed. This prediction was confirmed by comparing the size of intron 2 (which contains exon 2a). This was approximately 230 bp longer in the human and chimpanzee than in the mouse, rat, dog, opossum and *Fugu* which can be partially attributed to the incorporation of the *Alu* repeat into the genome (Table 5.5). Figure 5.11 displays the presence of the *Alu* repeat in *H. sapiens* and *P. troglodytes* but not *M. musculus*. tBLASTn analysis against the *P. troglodytes* EST and cDNA database failed to obtain any evidence for the expression of this exon in the chimpanzee.

A) Human v Chimp



B) Human v Mouse

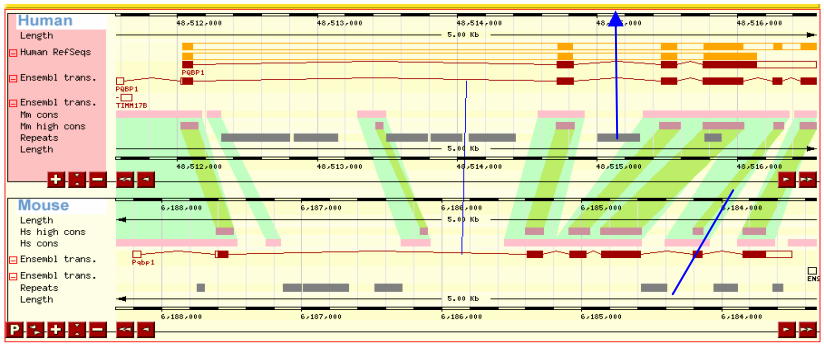


Figure 5.11 Global alignment of genome sequences containing the gene *PQBP1* using *MultiContigView* at Ensembl

Conserved regions between the two genomes (light green) were identified by global alignments on the untranslated genome sequence using BLASTz (Schwartz *et al.*, 2003). Further processing of these data scored highly conserved regions (dark green). Grey blocks represent the location of repetitive sequences and the conservation of an *Alu* repeat is depicted by blue arrows.

5.4 Expression profiling of *PQBP1*

Establishing the pattern of tissue expression for a gene is fundamental to understanding its function. Genes can develop restricted expression patterns in response to various stimuli, development processes or disease states. For example, the expression of some X-linked CT-antigen genes has been refined to the testis (Chen *et al.,* 2005). This could confer a fitness benefit to males, without being deleterious to females. At the other extreme, genes with ubiquitous expression profiles tend to be involved in basic cellular processes that are common to all cells, examples of which are the housekeeping genes, such as beta-actin (*ACTB*) or glyceraldehydephosphate dehydrogenase, (*GAPDH*). The expression pattern of human *PQBP1* was first determined using known expression data from EST sequences or Affymetrix expression microarrays. This analysis was followed by quantitative PCR, where global and transcript specific expression patterns for *PQBP1* transcripts were determined.

5.4.1 Known data on tissue-expression of PQBP1

Prior to experimental determination of *PQBP1* expression in human tissues two published sources of expression data were consulted. These databases were gene expression profiling *in silico,* (GEPIS) (Zhang *et al.,* 2004; http:www.cgl.ucsf.edu/Research/genentech/gepis/) and the Gene Expression Atlas (http://expression.gnf.org/cgi-bin/index.cgi#Q) which contain expression information derived from EST data or affymetrix expressions microarrays, repsectively.

Gene Expression Profiling *in silico* (Gepis)

ESTs and their associated tissue source information provide valuable expression information. In theory, EST clone frequency is proportional to expression levels in that tissue. However, the accuracy of this method is limited by several factors, such as insufficient sampling of all cell types, the use of normalised and subtracted libraries, and the need for further experimental validation of some EST derived results. Here, the web based programme GEPIS was used to extract associated tissue information from EST entries in dbEST at the NCBI (http://www.ncbi.nlm.nih.gov/dbEST). Samples from pooled tissues and normalised or subtracted libraries were removed from the database prior to analysis. The results are displayed in Figure 5.12.

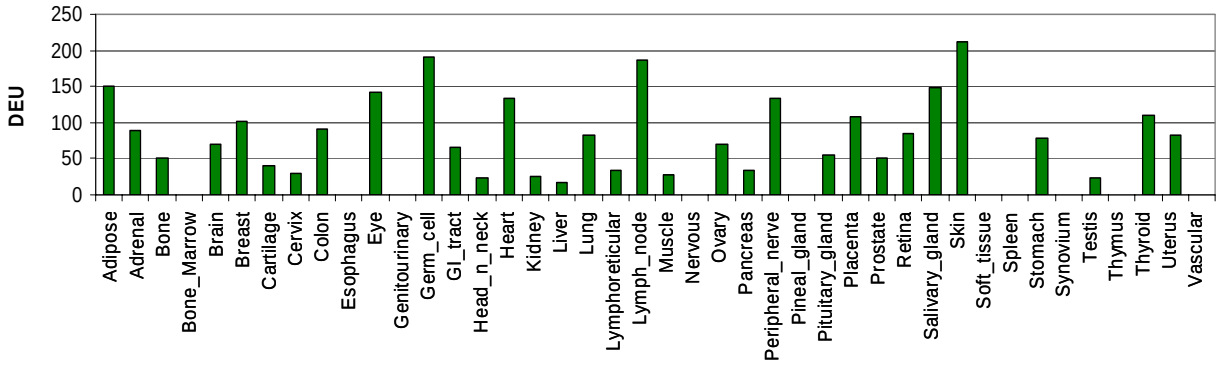


Figure 5.12 EST expression profile for *PQBP1*

Gene expression values were determined using the web programme GEPIS, and values displayed are digital expression units, DEU (total number of matching clones divided by the sum of the library sizes for both normal and tumor tissues and multiplied by 1,000,000).

The expression profile generated for *PQBP1* using this programme indicated no real trend. Highest values were obtained for adipose tissue, skin, lymph nodes and germ cells, while no ESTs have been sequenced from the bone marrow, esophagus, genitourinary tract, nervous system, pineal gland, soft tissue, spleen synovium, thymus and vascular tissue. Moderate *PQBP1* expression levels were recorded in the brain. It must be noted that these results are heavily biased by the EST library sizes; ESTs sequenced from a smaller library will have a disproportionately high DEU value.

Gene Atlas of Expression

The expression pattern of *PQBP1* was also extracted from the Gene Expression Atlas (http://expression.gnf.org; Su *et al.*, 2002). This information was derived from hybridisation of RNA from numerous tissues to human affymetrix chips (chip type – Human U95A) (Figure 5.13). These results demonstrate that relative expression of *PQBP1* is highest in the brain, ovary and uterus. This expression profile differs to that derived from existing EST data and confirms that current amount of EST coverage is inadequate in most tissues to give a good measure of relative expression levels.

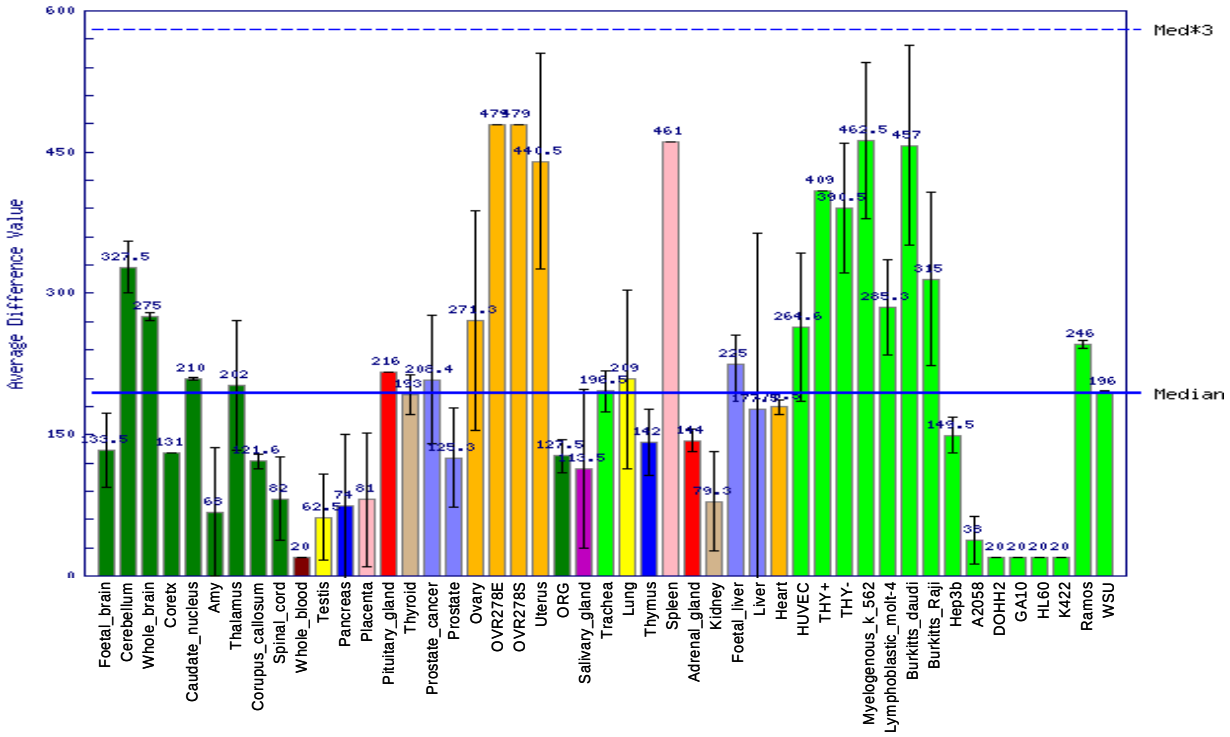


Figure 5.13 Expression profile of *PQBP1* as extracted from the Gene Expression Atlas

Relative levels of *PQBP1* expression were determined by RNA hybridisation to the affymetrix expression Chip (Human U95A). Samples are colour coded according to tissue type some of which are brain (dark green), female specific tissues (orange), and cultured cell lines (green). The average difference values (y-axis) are computed by Affymetrix software. These values are proportional to mRNA content in the sample.

5.4.2 Analysis of PQBP1 gene expression by quantitative PCR

A crude evaluation of the tissue expression patterns for *PQBP1* transcripts was described earlier, by visualising the PCR products produced by nested PCR using agarose gel electrophoresis (Figure 5.2). Here, it was apparent that the reference transcript of *PQBP1* was the most abundant, while the size of the minor band suggested that may represent the complete deletion of exon 4 could also be distinguished in most tissues. Additional faint bands corresponding to additional *PQBP1* variants were also observed in most tissues but no meaningful information could be derived on their identity or expression levels.

In order to achieve a more accurate description of *PQBP1* expression patterns the relative abundance of *PQBP1* transcript variants was determined by quantitative PCR (qPCR). This methodology has been used successfully to quantify the abundance of variant transcripts for several genes including neurotrophic factor,

BDNF, (Altieri *et al.,* 2004), Interleukin- receptor, *HIL-5Ra* (Perez *et al.,* 2003) and was chosen in preference to other hybridisation based techniques such as Northern blotting or RNase protection assays because it is more sensitive and can detect *PQBP1* transcript variants with a low abundance.

Quantitative analysis was completed on a panel of RNAs from 20 different human tissues (section 2.8.1). Reactions in were performed in an ABI7000 light cycler, and in all cases quantitation was determined by SYBR green fluorescence.

5.4.3 Primer design

Primer pairs were designed using the programme Primer Express (ABI Biosystems) to amplify either all *PQBP1* transcripts in concert or to distinguish between different transcript variants. Where possible discriminative primer pairs spanned exon-junctions of *PQBP1* transcript variants; one primer of each pair was designed to span an exon junction, while the second primer was designed to ensure that the amplicon remained between 50 and 150 bp in length. It was possible to discriminate between various transcripts by designing primers to novel exon junctions of transcript variants. All primers are listed in Table 5.8 and their sequences are listed Appendix VI. The location of all primers used in this study is displayed in Figure 5.14.

It was not possible to design primer pairs to assay all *PQBP1* transcript variants. In these cases, primer pairs were designed to amplify multiple transcripts that shared an alternative exon junction. For example, exon 4 was deleted in both transcripts 1 and 15. One of the primers designed specifically to amplify these transcripts spanned the unique exon junction created by the union of exons 3 and 5. Further analysis was required to differentiate between transcripts 1 and 15. This could be achieved by assaying another variant exon junction, spanning exons 2 and 3, that was exclusive to transcript 15. However, this analysis was not completed as specific primers could not be designed successfully to amplify this junction. Specific primer pairs could not be designed to transcript 2 (21 bp deletion in exon 4) as the deletion was flanked by repetitive sequence. Table 5.8 shows all such cases where a primer set gave data on multiple rather than individual transcripts. Additionally, specific primer pairs could not be designed to the 21 bp deletion

observed in transcripts 2-5. This is because deletion was flanked by repetitive sequence.

Quantitation of the alternative transcripts proceeded by generating standard curves by quantitative PCR using cloned transcripts to which sample cDNAs were compared. To enable a comparison between different transcript variants and different tissues, samples were normalised against the constitutively expressed region of *PQBP1*.

In order to ensure that each primer pair only amplified the transcripts of interest, PCR was carried out on the panel of *PQBP1* transcript variant clones generated in section 5.2.1. Amplification conditions were optimised for each primer pair by performing the reactions over a range of annealing temperatures, as outlined in section 2.15.1. The specificity of each PCR was first assessed by agarose electrophoresis (Figure 5.15). In addition, to ensure that the PCRs produced only one amplicon, the melting temperature of the PCR products was monitored using the melting curve option following quantitative analysis (results not shown). Here, analysis was performed at 0.1°C increments between 60°C and 90°C. Together these results demonstrated that most primer pairs were able to amplify the desired transcripts and only generated one PCR product. Three primer pairs (*PQBP1*.Q2, *PQBP1*.Q7 and *PQBP1*.Q8) failed this screening process.

This analysis confirmed that the primers only amplified products from the expected *PQBP1* clone(s).

Table 5.8 Primer sequences and additional details of primers used in quantitative analysis of alternative *PQBP1* transcripts.

Name	stSG #	Transcript	Comment
PQBP1.Q10	660506	All	Amplified desired transcripts
PQBP1.Q2	559357	1, 15	Not transcript specific – failed and redesigned
PQBP1.Q2b	810729	1, 15	Amplified desired transcripts
PQBP1.Q3	559358	4,6	Amplified desired transcripts
PQBP1.Q4	559359	3,7,8, 11	Amplified desired transcripts
PQBP1.Q5	559360	8	Low abundance transcript – detection not above background levels
PQBP1.Q6	559361	9, 10, 11	Amplified desired transcripts
PQBP1.Q7	559362	13	Not specific – failed, unable to redsign new primers
PQBP1.Q8	559363	12	Not specific –failed and redesigned
PQBP1.Q8b	810730	12	Redesigned – annealing temp 63 °C, low abundance transcripts
PQBP1.Q9	559364	5, 14	Amplified desired transcripts

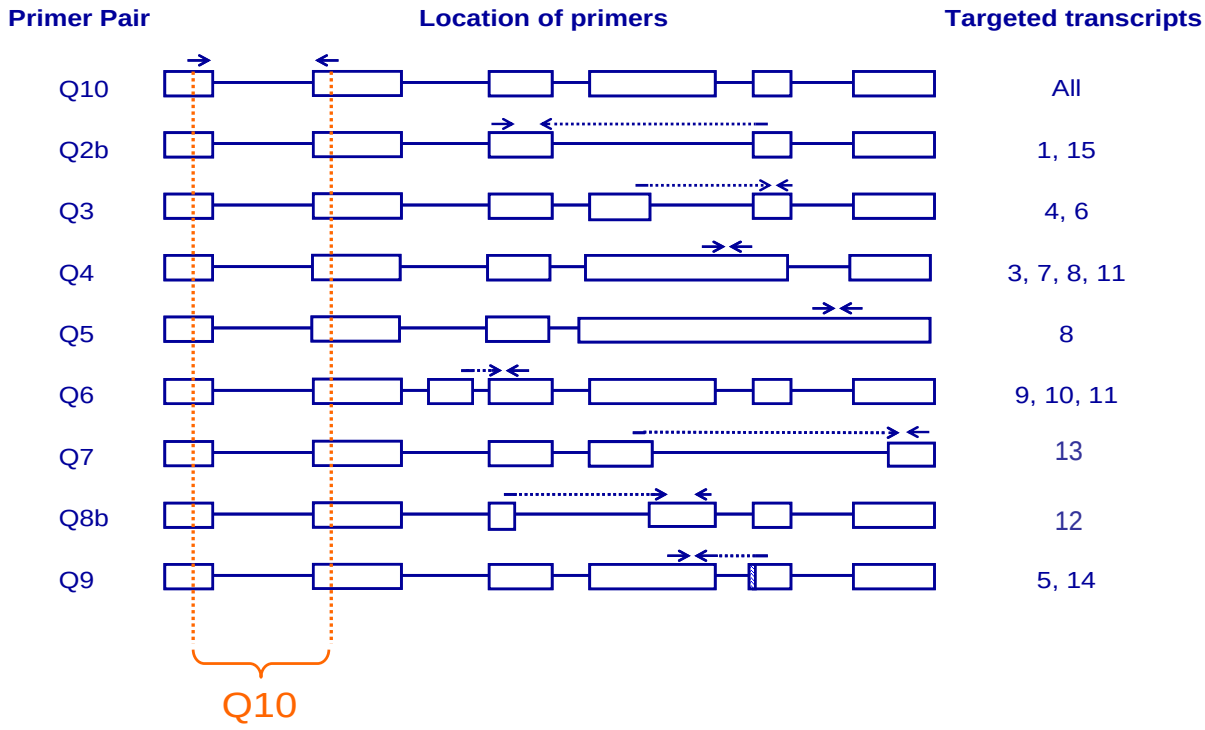
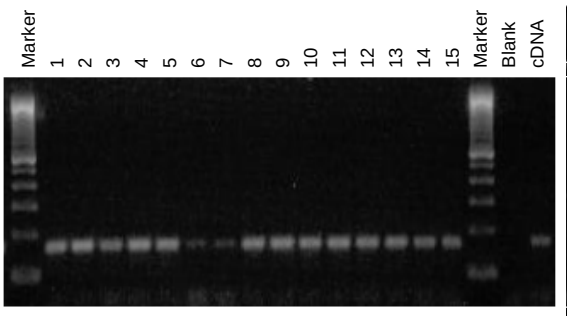


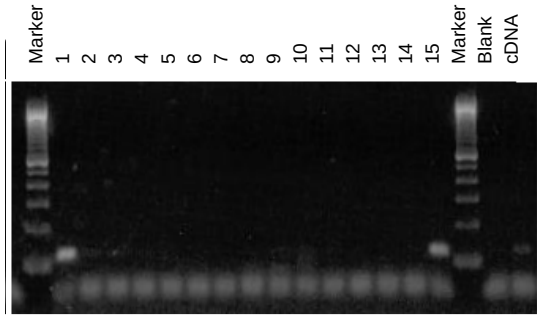
Figure 5.14 Location of primers used to determine the abundance of *PQBP1* alternative transcripts.

Primers used to amplify *PQBP1* transcripts are listed next to the exon-intron structure of the targeted transcripts. Arrows denote the location of primers while the transcripts to which they were designed are also listed (RHS).

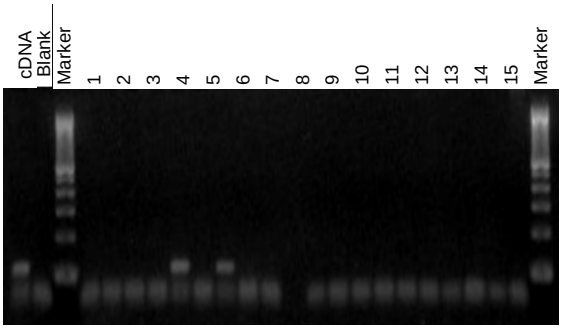
A) Primers *PQBP1*-Q10 (AII)



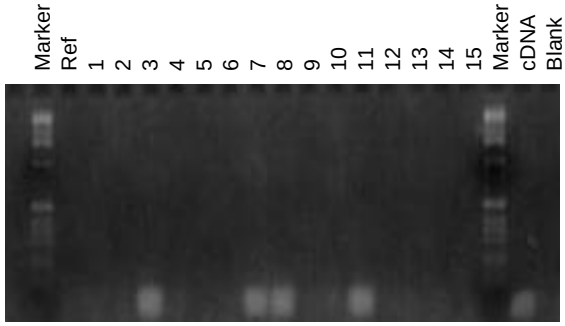
B) Primers *PQBP1*-Q2b (1, 15)



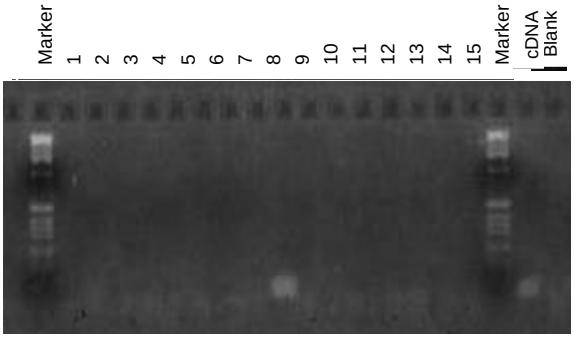
C) Primers *PQBP1*-Q3 (4, 6)



D) Primers *PQBP1*-Q4 (3, 7, 8, 11)



E) Primers *PQBP1*-Q5 (8)



F) Primers *PQBP1*-Q6 (9-11)

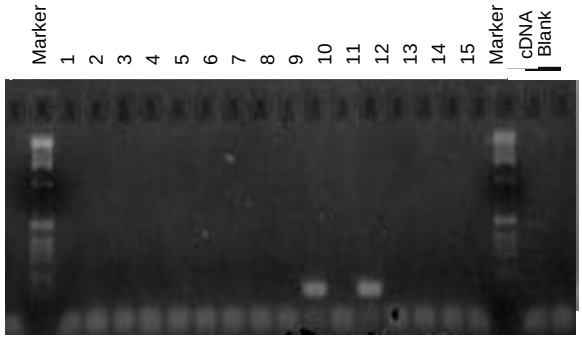
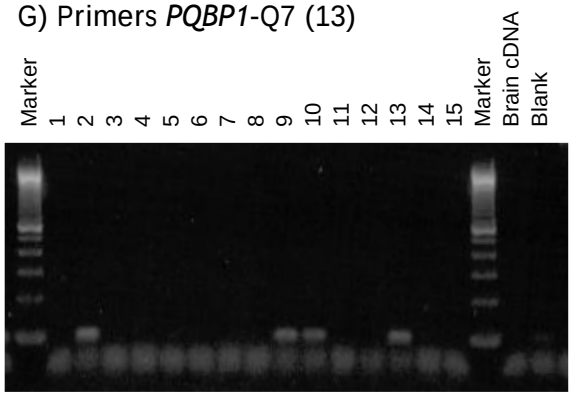


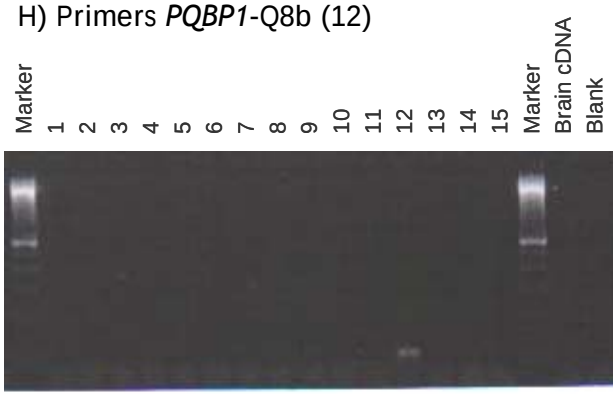
Figure 5.15 Specificity of *PQBP1* alternative transcript primers.

Primers were designed to amplify *PQBP1* alternative transcripts. Primers were screened against each of the cloned transcripts (1-15) by PCR and the products were resolved by agarose electrophoresis on a 2.5% gel stained with ethidium bromide. Transcripts to which the primers were designed are denoted in parenthesis. Blank = $T_{0.1}E$ negative control, cDNA = brain cDNA positive control (50 ng). Continued overleaf.

G) Primers *PQBP1*-Q7 (13)



H) Primers *PQBP1*-Q8b (12)



I) Primers *PQBP1*-Q9 (5, 14)

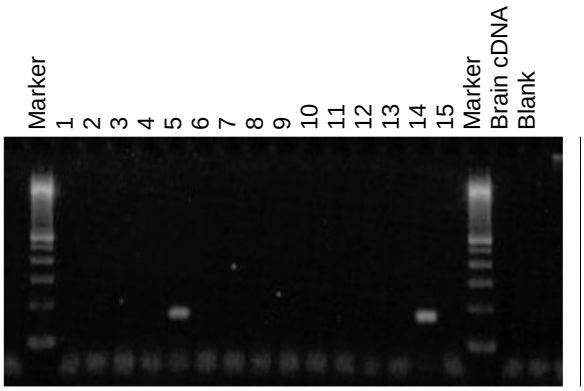


Figure 5.15 continued

5.4.4 Sensitivity, linearity and amplification efficiencies

The sensitivity of each primer pair in the quantitative PCR was evaluated using different starting amounts of the clone cDNA (section 2.18). SYBR Green fluorescence was tested over five orders of magnitude ranging between 1 x 10^3 to 1 x 10^8 molecules per reaction. The cycle threshold (Ct) value decreased in proportion with the amount of cDNA used in the reaction and all resulting standard curves had correlation coefficients greater than or equal to 0.99. Subsequent analyses were performed using standard curves between the range of 1 x 10^3 and 1 x 10^7 molecules per reaction.

These calibration curves were established using individual cloned cDNA rather than the complex mixture of transcripts commonly found in reverse transcribed RNA samples. The presence of additional cDNA transcripts could influence the reaction efficiency. In order to test the possibility that amplification was affected by the complexity of the sample, *S. pombe* cDNA was included in the reactions for one primer pair (*PQBP1*.Q10) at varying concentrations (0-100 ng per reaction). *S. pombe* cDNA was considered to be a suitable substrate to include in this experiment as the sample did not contain the *PQBP1* cDNA target (as assessed by ePCR). The influence of *S. pombe* cDNA levels on the PCR efficiency are shown in Figure 5.16. Negligible differences were observed in the amplification efficiencies of the *PQBP1* target.

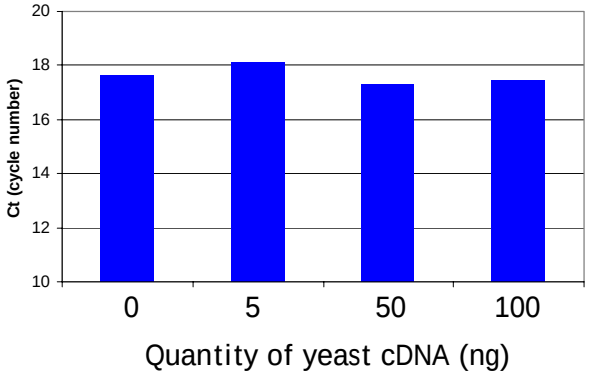


Figure 5.16 Effect of yeast cDNA on real-time PCR amplification efficiency of *PQBP1* transcripts

Reactions were established with 1 x 10^6 molecules of cloned reference *PQBP1* cDNA, (for the reference transcript) to which varying concentrations of yeast cDNA were added (0-100 pg).

The sensitivity of each primer pair in the PCR was also assessed using different starting amounts of adult brain cDNA sample. Concentrations tested ranged between 50 pg and 200 ng of total RNA. In most cases linearity of the Ct values was observed between 100 ng and 200 ng of starting material.

All subsequent quantification reactions were completed using cDNA synthesized from 100 ng of total RNA. The number of molecules present in each sample was extrapolated from a standard curve that was generated using the appropriate cDNA clone. For example, primer pair Q2b was designed specifically to amplify transcripts 1 and 15. The standard curve generated using cloned cDNA (*PQBP1* transcript 1) is displayed in Figure 5.17 from which the concentration in 20 different cDNA samples was extrapolated.

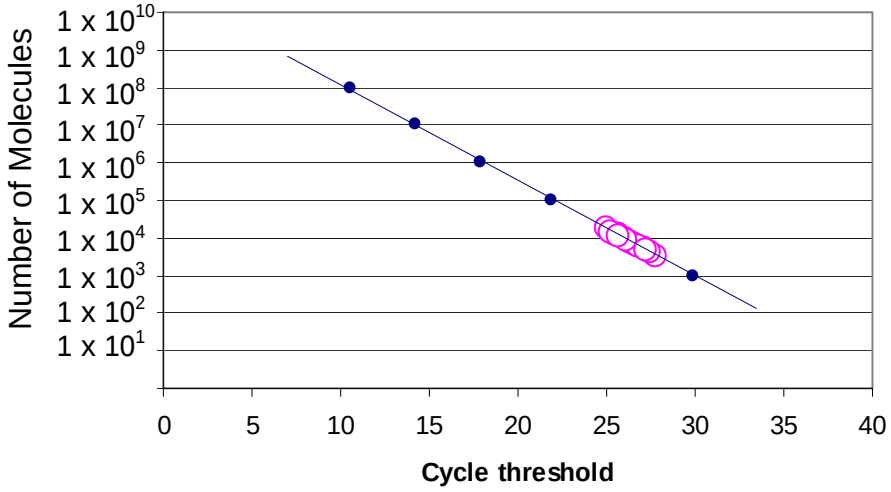


Figure 5.17 Quantification of *PQBP1* alternative transcripts by real-time PCR.
Standard curve for primer pairs *PQBP1*.2b is displayed in blue, while the C_T value of cDNA samples are shown by pink open circles.

5.4.5 Quantitation of reference PQBP1

Intron spanning primers *PQBP1*.Q10F and *PQBP1*.Q10R were designed to exons 1 and 2 of the *PQBP1* gene and were used to profile the overall transcript abundances from *PQBP1* in 19 human tissues (previously described in Section 2.8.2). These primers amplified all known *PQBP1* transcript variants (Figure 5.14). The panel of cDNAs was previously been shown to be free from genomic DNA contamination (Ian Barrett, personal communication) and each reaction was performed in triplicate.

To normalise the transcript levels primers, GAPDHF and GAPDHR, were designed in neighbouring exons of the housekeeping gene, *GAPDH*. The efficiencies of the *PQBP1* and GAPDH reactions were compared over four orders of magnitude confirmed that primer combinations were suitable for quantitative analysis (as outlined in Section 2.15.6, Figure 5.18). Each experiment was completed in duplicate. The purpose of this experiment is to shown that both genes have the same amplification efficiencies over the test range.

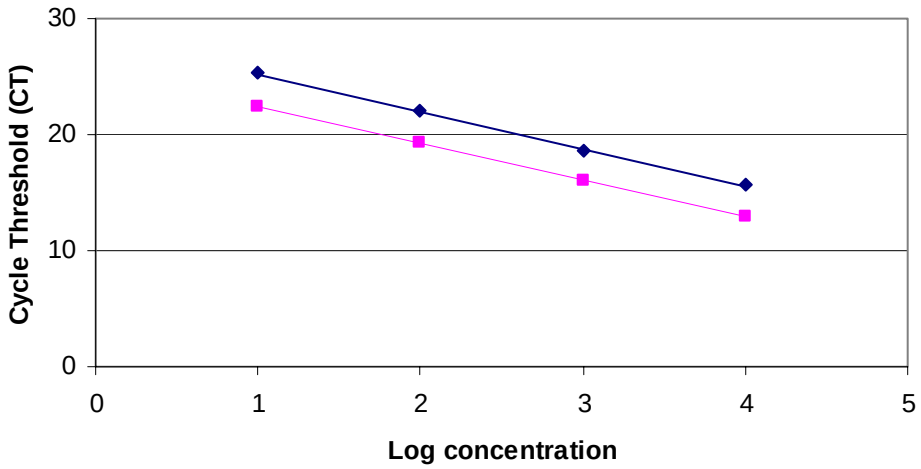


Figure 5.18 Comparison of amplification efficiencies for the primer pairs GAPDH and *PQBP1*.Q10.
Human brain cDNA was serially diluted between the range of 1 and 0.001 (100 ng to 100 pg of cDNA per reaction). The cycle value at which amplification was measured (C_T) and is displayed in pink for the GAPDH primer pair and blue for the primer pair, *PQBP1*.Q10. Linear regression lines are displayed being y= -3.211x + 25.71 (R^2= 0.9996 – GAPDH) and y = -3.219x + 28.475 (R^2 = 0.9994; *PQBP1*.Q10).

In Figure 5.19, the normalised relative abundance of the *PQBP1* amplicon is given in relation to the expression levels in the brain. Expression of *PQBP1* was observed in all tissues, and was relatively uniform. Highest expression levels were recorded in the heart (15% greater than brain), while lowest expression levels were observed in the ovary and stomach (each 78% lower than the brain).

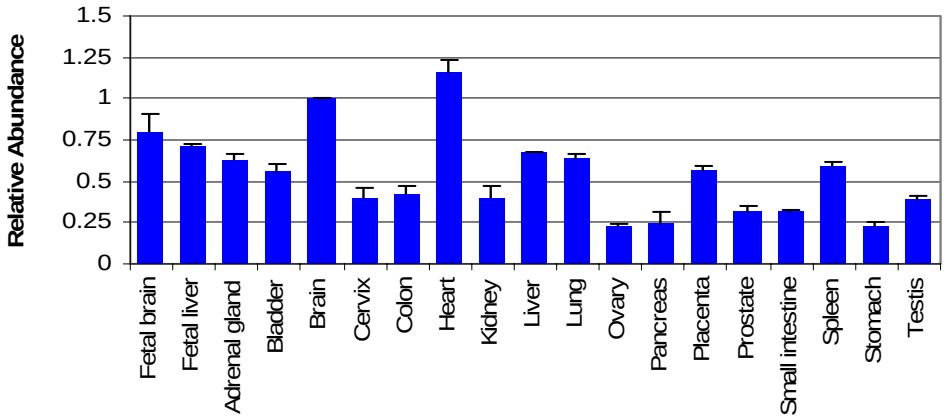


Figure 5.19 Relative abundance of *PQBP1* expression.
The relative abundance of *PQBP1* transcripts was determined for 19 different human tissues. Values are expressed relative to those in the brain and are normalised to the housekeeping gene, GAPDH. Values displayed are the results from duplicate experiments.

5.4.6 Expression profiling and quantitation of alternative variants

The quantitation of *PQBP1* expression described in the previous section produced a global view of the expression patterns of all *PQBP1* transcripts. Subsequently, the relative contribution of *PQBP1* transcript variants to the total was assessed. Motivation for this study stemmed from a publication that suggested that the abundance of transcript variants may be related to their capability to perform a biological function (Kan *et al.*, 2002). This publication suggested that functional transcripts produced by regulated splicing events may be (relatively) more abundant than variants produced by imprecise mRNA splicing events (Kan *et al.*, 2002).

Quantitative PCR was carried out on cDNA synthesised from a panel of 20 different human tissues as described in section 2.8.1. Results showing the abundance of each alternative transcript class expressed in relation to the total *PQBP1* abundance are shown Figure 5.20. All experiments were completed in duplicate, using independently synthesized batches of cDNA.

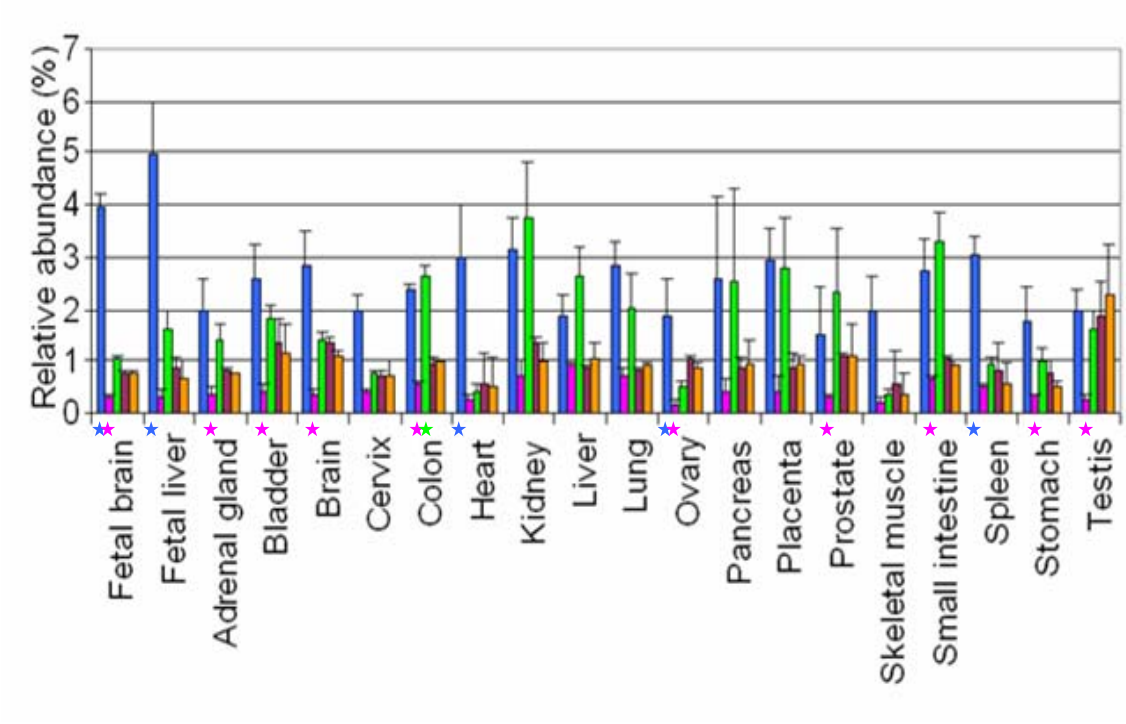


Figure 5.20 Relative abundances of *PQBP1* transcript variants
The abundance of each transcript variant group is expressed as a percentage of the constitutive region of *PQBP1*. Transcripts 1 and 15 (blue), 4 and 6 (pink), 3, 7, 8 and 11 (green), 9 and 11 (brown) and 5 and 14 (orange) are shown. Statistically significant changes in transcript abundance are denoted with a star (★) which are coloured in appropriately.

The results were normalised to ensure that comparisons could be made between different tissues. All samples were normalised against a region *PQBP1* that was found in all transcripts using primer pair, *PQBP1*.Q10F &R. Results were given as a relative percentage of the reference amplicon. All results were assessed for statistical significance using the one-way ANOVA test in Microsoft Excel. Tests were performed to measure variations in the abundance of alternative transcripts both between different tissues within the same tissue.

All tissue types had comparable levels of *PQBP1* transcript variants and no statistically significant changes were recorded. It is anticipated that a greater sample size, with lower sample variation could yield results of statistical significance.

Intra-tissue variation was assessed by comparing the abundance all transcript variants within the one tissue. This analysis did produce results of significance and the results are listed in Table 5.9. Fifty percent (10/20) of the tissue samples analysed using primer pair Q3 (transcripts 4 and 6) had a statistically significant smaller abundance when compared to all of the other *PQBP1* transcript variants. All other statistically significant variations were the result of an increase rather than decrease in the relative transcript abundance. These were observed for transcripts 1 and 15 (amplified using primer pair Q2b) in the foetal brain, heart ovary and spleen and transcripts 3, 7, 8 and 11 (amplified using primer pair Q4) colon.

Table 5.9 Tissue samples with statistically significant differences variations in transcript abundance

Primer pair (transcripts amplified)	Tissue	Increase or decrease in abundance
Q2b (1 and 15)	Foetal brain, heart, ovary and spleen	Higher
Q3 (4 and 6)	Foetal brain, adrenal gland, bladder, brain, colon ovary, prostate, small intestine, stomach and testis	Lower
Q4 (3, 7, 8 and 11)	Colon	Higher
Q6 (9, 11)	None	n.a.
Q9 (5, 14)	None	n.a.

Variation in the abundance of *PQBP1* variant transcripts was assessed for twenty different tissues. Overall analysis suggested that the abundance of *PQBP1* alternative transcripts was low and represented less than 10% of all *PQBP1* transcripts. The impact of this degree of variation on *PQBP1* function remains to be solved.

5.5 Discussion

Work described in this chapter demonstrated how alternative transcripts can be easily identified by screening a large number of cloned PCR products amplified from cDNA. *PQBP1* was targeted for more detailed analysis and by screening 192 clones an additional six novel transcripts were identified. These variants had changes within the ORF section defined by the reference transcript. These novel transcripts supplemented information that was obtained in chapter 4 where fourteen novel transcripts were identified for the gene *PQBP1* (this analysis also identified variations in the 5' UTR) and highlight the amount of additional transcript variation information that can be obtained with a detailed screening of cDNA samples for variant transcripts. An additional advantage of this method is that it creates a resource that can be used for functional studies, as will be described in the following chapter.

5.5.1 Comparative sequence analysis highlights potential causes for PQBP1 transcript variation

Comparative sequence analysis was used to provide information on the evolution of the *PQBP1* gene structure. Particular emphasis was placed on using genomic information in *PQBP1* orthologues, such as splice site sequences, in order to obtain a greater understanding of the conservation and sequence requirements of splice sites. This sequence information would provide additional evidence for the intended use of a splice site during processing of the *PQBP1* transcripts. In order to complete this analysis genome sequences from eight vertebrate species were used. The sources of information employed in the analyses presented reflect the increase of sequence submissions (both EST and genomic) to the public repositories within a short period of time. This includes availability of zebrafish BAC resources, human genomic sequence information and also the generation of WGS assemblies, for *Fugu,* rat, opossum, chimpanzee and dog. The availability of even draft quality genomic sequence allows important contextual information to be considered in the generation and testing of hypotheses regarding the evolution of mRNA splicing as well as individual genes.

Two different types of BLAST analysis, tBLASTn and blastz, were used to perform the comparative analysis. tBLASTn was more sensitive than blastz analysis and detected more divergent exons. This could be attributed the sensitivity of search

parameters used in the analysis, where smaller "word sizes" yield more sensitive results. One limitation of zPicture is that the parameters are fixed and cannot be altered to increase the sensitivity of the blastz alignment. Hence, some conserved sequences may be missed in the comparative analysis carried out by zPicture.

In the process of identifying orthologues for this analysis the dynamic evolutionary history of *PQBP1* became apparent. Phylogenetic analysis of the *PQBP1* homologues suggests that only the fishes, the zebrafish and *Fugu,* have two fully processed copies of *PQBP1* and that these copies were acquired via independent duplication events. One functional copy and one non-functional copy of *PQBP1* were identified in four mammals – the dog, mouse, rat and opossum, while the human and chimpanzee only have one functional copy of the gene. Additional work needs to be performed to define the evolutionary ancestry of *PQBP1.* For example, the presence of two functional *PQBP1* genes in the zebrafish and *Fugu* must be confirmed. This could be achieved using more complete genome sequence assemblies or experimentally using *in situ* hybridisation techniques. If the two copies are confirmed, additional analysis could be performed to date the duplication event(s) that created two copies of this gene in the two fishes.

The architecture of the *PQBP1* homologues has also varied throughout evolution. Of particular interest was the variable presence of exon 4. The entire exon was detected in all eutherians analysed (human, chimp, dog, rat and mouse), however, only the 3' end of the exon was detected in the opossum whereas the entire exon was not detected in the either zebrafish or the *Fugu* (although the exon was detected in the *Fugu* using TBLASTN analysis). From this analysis it is not clear if exon 4 was present in ancient copies of *PQBP1* and has been lost in the zebrafish lineage or if it appeared after the divergence of the fishes from the tetrapods. This hypothesis would, however, require an independent appearance of exon 4 in the *Fugu.*

Interestingly, exon 4 also displayed the most heterogeneous splicing patterns, and was either truncated or deleted in 12 of the 16 *PQBP1* transcript variants. For example, the entire deletion of exon 4 observed in transcripts 1 and 15 was the most frequent *PQBP1* alternative splicing event. The splice site sequences of exon 4 were examined to see if they influenced its exclusion during mRNA processing. Several studies have demonstrated that acceptor splice site strength is an

important regulator of exon inclusion (Graveley *et al.,* 1998; Thanaraj and Clark 2001; Sorek *et al.,* 2004b) but in this case, the splice site score of its exon acceptor sequence was similar to all other exons. Other causes of exon skipping are promoted decreased exon length (Dominski, 1991) and increased intron length (Berget *et al.,* 1995) but neither of these observations support the exclusion of exon 4 from the processed *PQBP1* transcript. Other sequence elements such as exonic or intronic splicing silencers may promote the exclusion of exon 4 from *PQBP1* transcripts. It also remains to be solved if this alternative splicing event has any functional impact on its cognate protein. This notion is addressed in the following chapter.

Comparative sequence analysis was also used to analyse the conservation of exon boundaries. A high degree of similarity for the sequence motifs surrounding 5' and 3' splice sites has been observed in the genome sequences of the human, mouse and *Fugu* genomes (Yeo *et al.,* 2004) suggesting that functional splice sites may be conserved in the *PQBP1* orthologues. It was found that the sequence variation around exon junctions of alternative transcripts was higher than that observed for reference exon junctions. This suggests that reference splice sites may also be used in the *PQBP1* orthologues and that they represent *bona fide* human splice sites. The lack of conservation in the alternative splice sites suggests that these sites may not be under the same selective pressures to remain conserved. Therefore, it is possible to speculate that the poorly conserved alternative splice sites may not represent functional splice sites in the human. Experimental verification is required to confirm the preferential use of the reference splice sites in the *PQBP1* homologues which could be obtained by sequencing cDNA samples from the other species.

One interesting observation was the sequence composition of the alternative donor site used in exon 3 in transcript 13. The dinucleotide sequence of this splice site in primates, is GC while it is a GT in the other eutherian species analysed, as the dog, rat and mouse. This splice site was the weakest site used in all of the human *PQBP1* transcript variants (splice site score 38 versus average donor score 86.3) and may represent an aberrant splice site. However, the variant dinucleotide donor sequence observed in the non-primate species would effectively produce a stronger splice site because its sequence closely matches the consensus sequence donor sequence recognised by the U2 snRNP splicing machinery. The strength and

utilisation of this splice site in other vertebrate species needs to be determined. If it is used, would it compete with the reference splice site during mRNA splicing? What are the functional implications of this alternative splicing event?

5.5.2 Expression studies of PQBP1

The expression patterns of the *PQBP1* was assessed in human tissues by comparing known data from EST sequences and affymetrix expression microarrays to quantitative PCR expression patterns. Previous analysis of *PQBP1* expression had found to have highest expression levels was observed in the brain (Iwamoto *et al.,* 2000). Perhaps because of this observation, most functional analysis of *PQBP1* has focused on its biological role in the this tissue where it has been linked to Rennington disease (Stevenson *et al.,* 2005), XLMR (Kalscheuer *et al.,* 2003; Kleefstra *et al.,* 2004; Lenski *et al.,* 2004; Fichera *et al.,* 2005) and neurodegenerative disorders (Busch *et al.,* 2003). Although the expression profiles presented in this chapter confirmed expression of *PQBP1* in the brain, higher expression levels were recorded in heart by RT-PCR.

PQBP1 expression patterns using three types of data differed both in the tissues analysed and relative abundance of *PQBP1* in each tissue. Some variation between the experimental protocols was expected since dramatically different techniques were used to collect each dataset. For example, the EST data was a concatenation of random sequence reads from different cDNA libraries which have been sampled to various depths. Despite the obvious experimental differences, a common thread to the three datasets is the widespread distribution of *PQBP1* transcripts where strong expression levels of *PQBP1* observed in other tissues including the skin and spleen. Together these results suggest a widespread role for *PQBP1*.

Insights into the functional relevance of *PQBP1* transcript variants were gained by quantifying their abundance in 20 different human tissues. In all tissues the reference transcripts was the most abundant with the variants representing less than 10% of all *PQBP1* mRNAs. This study has several limitations. High levels of sequence homology between some exon junctions meant that probes could not be designed to detect the expression of three *PQBP1* transcript variants (transcripts 10, 12 and 13). Ideally, all transcripts should have been assessed in separate assays, however data from the end-point RT-PCR indicates that these transcripts are unlikely to be present at very high levels. A related limitation was the need to

group transcripts. However, despite this, the expression of the variants does not even come close to the expression levels of the reference transcripts.

The dataset was further reduced because the abundance of transcript 8 fell below informative detection levels, which suggests that transcript 8 is not functional. However, the sensitivity of real-time PCR reactions could be further improved through the introduction of transcript selection procedures prior to amplification of cDNA transcripts. For example, a series of RNA-mediated annealing, selection and ligation (RASL) reactions have been used to detect alternative splicing events from sub-nanogram quantities of RNA (Yeakley *et al.,* 2002). Alternative transcripts are selected prior to amplification, by annealing mRNA molecules to oligos that flank a presumed splice junction. If the predicted transcript is present, the two corresponding oligos are ligated and amplified using universal primers. The selected transcripts are then hybridised to the microarray spotted with unique exon-junction sequence in order to permit accurate quantification. This technique has been used to profile regulated alternative splicing events of the tyrosine phosphatase receptor (*PTPRC*) in human cancer cell lines (Yeakley *et al.,* 2002). Alternative methods that could be used to quantify the abundance of alternative transcripts include exon-junction arrays and "polony" technology (section 1.6.1). These methods are suited to large scale analysis of multiple alternative splicing events in concert but great care must be taken when designing probes to ensure that false positive results caused by cross hybridisation are minimised.

It is important to note that the expression profiles generated using this method reflect the relative abundance of the *PQBP1* transcripts at one discrete stage in both developmental and cell cycle progression. The abundance of alternative transcripts may be regulated by such processes and would not be detected using this approach. Human tissues are also a mosaic of different cell types. The expression of *PQBP1* transcript variants may be up- or down-regulated in certain cells. The analysis described used homogenised tissue samples and not individual cell types to quantify the abundance of *PQBP1* transcripts. *In situ* analysis of *PQBP1* alternative transcripts could shed further light on the cellular distribution of *PQBP1* variants.

5.5.3 Conclusions

The data presented in this chapter indicate that splicing within the *PQBP1* locus is complex. Although several alternative transcripts have been found previously for the PQBP*1* gene (Iwamoto *et al.,* 2000), multiple novel *PQBP1* splice variants were identified and here a survey of some of their functions was performed. A catalogue of transcript variants was generated for a single gene that permitted comparative analysis of the splicing site strength, and expression patterns to be obtained. Together with other functional data, these two lines of information could help to come to a conclusion about how variants of *PQBP1* arise and whether they have a function.

Competition between the alternative and constitutive splice sites in mRNA processing depends on the relative quality of the splice signals. In this chapter it was found that splice site strength may have an important role in splice site selection of *PQBP1* transcripts. All alternative *PQBP1* splice signals were weaker than constitutive splice signals. Furthermore, comparative sequence analysis confirmed that *PQBP1* reference splice sites have been conserved throughout evolution whereas alternative splice sites do not appear to be as highly conserved. Further definition of the mechanisms that control *PQBP1* splicing is required. This could be achieved by assessing the branch point strength, predicting the influence of mRNA secondary structure on the accessability of both donor and acceptor sites to the components of the spliceosome and scanning for potential regulatory elements such as exon splicing enhancers.

The functional significance of the transcripts identified in this study still remains unclear. What fraction of the splicing represents 'noise', caused by the relaxation of the RNA splicing, is currently unknown. Perhaps the biggest clue to functional capacity of the *PQBP1* transcript variants was obtained by quantifying their expression levels where an overwhelming excess of the reference *PQBP1* transcript over the alternatives was recorded. Low expression levels of some transcripts (e.g., transcript 8) highlighted the possibility that at least some of the variants were generated by aberrant mRNA splicing. Other transcript variants (e.g. transcript 4 and 6) displayed tissue specific expression patterns. While it is it not possible to make any conclusions from these expression patterns, it is tempting to speculate that the *PQBP1* alternative transcripts identified in this chapter may not be generated by regulated splicing events and do not serve any function. It is also

possible that some of the novel transcripts could potentially encode proteins of different sizes that have distinct roles. Further analysis of the *PQBP1* transcript variants and their encoded products is carried out in the following chapter.