An Analysis of Transcript Variation in Human Xp11.23

By

Tamsin Lee Eades

Thesis submitted for

the degree of Doctor of Philosophy

The Wellcome Trust Sanger Institute

and

Darwin College,

University of Cambridge

August 2005

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration, except where specifically indicated in the text.

This dissertation does not exceed the word limit set by the Biology Degree Committee.

Tamsin Eades

Abstract

Prior to its completion, it was conjectured that the human genome sequence contained in excess of 100,000 genes, whereas current estimates from the completed genome sequence have reduced this figure to between 20,000 and 25,000. This dramatic reduction could be partially attributed to the phenomenon of alternative splicing, where one gene has the ability to produce more than one functional mRNA transcript.

The aim of this thesis was to obtain a greater appreciation for the diversity of transcripts that can be generated from a single gene. Towards this end, the human genome sequence provided the definitive substrate for the study of transcript variation. A gene-rich region of the X chromosome was selected for analysis, where the frequency and the associated functional consequences of alternative splicing were assessed.

This thesis describes an analysis of the genome features and gene content in human Xp11.22-p11.3. Human Xp11.23 has the highest density of genes in the human X chromosome, and this analysis identified 77 known and 11 novel genes. The pseudogene content of this region was also high - it contained 59 processed and 7 non-processed pseudogenes. More detailed investigation of these revealed that the number and diversity of gene products generated from genes within Xp11.23 greatly exceeded the number of coding regions that it contained. In order to further study the impact of alternative splicing in Xp11.23 detailed analysis was completed on 18 genes using bioinformatic and comparative analysis together with a targeted RT-PCR sequencing strategy. This analysis identified more than 120 transcripts variants. Preliminary tissue profiling of these transcripts was completed using RT-PCR.

The functional consequences of alternative splicing were then investigated for one gene, polyglutamine binding protein 1, *PQBP1*. This ubiquitously expressed gene has been associated with various disease phenotypes including X-linked mental retardation, Renpenning syndrome and other neurodegenerative disorders. In concert with expression and evolutionary analysis, a cloned open reading frame

collection was generated, for 16 transcript variants. The relative abundance of minor transcript variants was determined in a panel of 20 human tissues where it was found that together the minor variants accounted for less than 10% of the all *PQBP1* transcripts.

Following *in silico* analysis of the predicted protein sequences, it was found that transcript variation was associated with variable inclusion of a nuclear localisation signal. Sub-cellular localisation analysis of the transcript variants in CHO-K1 and Cos7 mammalian cell lines showed that three isoforms were not redirected to the nucleus after translation. Further analysis revealed that 11 of the *PQBP1* transcripts harboured a premature termination codon. Preliminary mRNA kinetic assays, and nonsense mediated decay inhibition assays confirmed that at least five of these transcripts promoted rapid degradation via transcript surveillance pathways.

Acknowledgements

I would like to thank my supervisor Mark Ross, for his encouragement, support and advice throughout the entirety of this project. I feel incredibly fortunate to have had a supervisor who was has always made time for me and has taught me the value of meticulous experimental planning, execution and analysis. Special thanks are also extended to Alison Coffey and Colette Johnston who not only proof-read this thesis but also gave valuable advice on *everything*. I have benefited so much from your wisdom, and friendship – thank you.

I am also thankful for the computational support that I have received from numerous teams at the Sanger Institute throughout this project. Carol Scott and Sarah Hunt from human genetics informatics have provided perl scripts and advice on routine computational analysis of large amounts of sequence. Thanks to Darren Grafham for prioritising some human and mouse clones for finishing. James Gilbert and Stephen Keenan from the Institute's informatics team carried out the computational analysis of the genome sequence, while some of the genome annotation described in this thesis was completed by Jennifer Harrow and the human and vertebrate analysis (HAVANA) team at the Sanger Institute.

Thanks are also extended to people who helped me with my experiments. Firstly, to the current and past members of the X chromosome group and DNA collections team, Christine Burrows, Frances Lovell, Ruth Bennett, Eleanor Howard and Hazel Arbury thank you to you all for helping "find my feet" at the Sanger, for your practical advice and for lending solutions/equipment in times of need. I shared a panel of human total RNA samples with Jackie Bye and am extremely grateful and thankful for her advice and patience with all RNA work. Many resources used for the sub-cellular localisation experiments were kindly provided by team 62 (molecular genomics and proteomics). Special thanks are extended to its members especially Ian Dunham, John Collins, Charmain Wright and James Grinham. Philippe Couttet, gave me much needed advice on NMD and mRNA stability experiments and provided me with the pTRE-TIGHT (tet-off) plasmid. Thanks also to Ian Barrett and Gareth Howell for giving me such a friendly introduction to the Sanger Institute.

Finally, thank you to all of my friends and family who have been so patient and tolerant over the past 3 ½ years. It's now time to repay many favours and perhaps I will now get to spend a Christmas at home!!! I am especially indebted to my parents and to Graeme, I know that I wouldn't have made it to this stage without you, thank you.

This project was funded by the Wellcome Trust and the Sanger Institute.

Table of Contents

List of Figures

List of Tables

List of Appendices

Abbreviations

1st EF	First exon finder
ABI	Applied Biosystems
ACeDB	A *C. elegans* database
ACTB	beta actin
AEDB	alternative exon database
ANS	anisomycin
ASAP	the Alternative Splicing Annotation Project
ASD	alternative splicing database project
ASDB	alternative splicing database
ASG	alternative splicing gallery
AT	annealing temperature
ATP	adenosine 5'- triphosphate
BAC	bacterial artificial chromosome
bis-acrylamide	(N, N'-methylene-)bis-acrylamide
BLAST (-n -p)	basic local alignment search tool (-nucleotide -protein)
BLAT	basic local alignment tool
Blixem	BLast matches In an X-windows Embedded Multiple alignment
BMD	Becker muscular dystrophy
ß-ME	ß-mercaptoethanol
bp	base pair
BSA	bovine serum albumin
ºC	degrees Celsius
cDNA	complementary DNA
CDS (c-)	coding sequence (consensus)
CEN	centromere
CHX	cycloheximide
CNS	central nervous system
CpG	cytidyl phosphoguanosine dinucleotide
Cps	counts per second
C_T	cycle threshold
CT- antigen	cancer testis antigen
CTP (d-)	cytidine 5'-triphosphate (deoxy-)
DAPI	4',6-Diamidino-2-phenylindole
dbEST	database of expressed sequence tags

DDBJ	DNA DataBase of Japan
DEPC	diethyl pyrocarbonate
DKFZ	Deutsches Krebs Forschung Zentrum (German cancer research centre)
DMD	Duchenne muscular dystrophy
DMSO	dimethyl sulphoxide
DNA	deoxyribonucleic acid
Dnase I	deoxyribonuclease A
dNTP	deoxyribonucleotide 5' triphosphate
dsDNA	double-stranded deoxyribonucleic acid
dsRNA	double-stranded ribonucleic acid
EBI	European Bioinformatic Institute
ECL	enhanced chemiluminescense
ECR	evolutionary conserved region
EDTA	ethylenediamine tetra-acetic acid
EJC	exon junction complex
EMBL	European Molecular Biology Laboratory
EMBOSS	European Molecular Biology Open Software Suite
ePCR	electronic polymerase chain reaction
ES cell	embryonic stem cell
ESE	exon splicing enhancer
ESS	exon splicing silencer
EST	expressed sequence tag
FITC	fluorescein isothiocyanate
g	gram
g	force of gravity (relative centrifugal force)
GAPDH	Glyceraldehyde-3-phosphate dehydrogenase
G banding	Geisma banding
GTP (d-)	guanosine 5'-triphosphate (deoxy-)
HASDB	human alternative splicing database
HAVANA	human and vertebrate genome annotation and analysis
HEPES	N-[2-hydoxyethyl]piperazin-N'-[2-ethansulphonic acid]
HGMP	Human Genome Mapping Project
HGNC	HUGO Gene Nomenclature Committee
HGP	Human Genome Project
hnRNP	heterogenous nuclear ribonucleoprotein
HRP	horseradish peroxidase

HMM	hidden Markov model
HS	hierarchical shotgun
HUGO	Human genome organisation
ICI	Imperial Chemical Industries
INDELs	insertion or deletion of deoxyribonucleic acid
IHGSC	International Human Genome Sequencing Consortium
IPTG	isopropyl ß-D-thiogalactoside
kb	kilobase pairs
kDa	kilodalton
l	litre
LB	Luria-Bertani
LINE	long interspersed nuclear element
LTR	long terminal repeat
M	molar
Mb	megabase pairs
MER	medium reiterative repeat
MGC	Mammalian Gene Collection
µg	microgram
µl	microlitre
µM	micromolar
min(s)	minute(s)
mg	milligram
ml	millilitre
mM	millimolar
mm	millimetre
mya	million years ago
NCBI	National Centre for Biotechnology Information
NEDO	New Energy and Industrial Technology Development Organisation
ng	nanogram
nm	nanometre
NMD	nonsense mediated decay
O/N	overnight
oligo-dT	deoxyribothymidyl oligonucleotide
OMIM	On-line Mendelian Ineritance in Man
ORF	open reading frame
PAC	P1-derived artificial chromosome

PAGE	polyacrylamide gel electrophoresis
PBS (-T)	phosphate buffered saline (-Tween 20)
PCR	polymerase chain reaction
RASL	RNA annealing selection and ligation
pg	picogram
PALSdb	putative alternative splicing database
PE	Perkin Elmer
poly(A)	polyadenylation
PTC	premature termination codon
PUR	puromycin
RNA	ribonucleic acid
RNAi	ribonucleic acid interference
mRNA	messenger ribonucleic acid
rRNA	ribosomal ribonucleic acid
snRNA	small nucleolar ribonucleic acid
hnRNA	heteronuclear ribonucleic acid
hnRNP	heteronuclear ribonucleoprotein
snRNP	small nucleolar ribonucleoprotein
dsRNA	double-stranded ribonucleic acid
siRNA	short-interfering ribonucleic acid
tRNA	transfer ribonucleic acid
miRNA	micro-ribonucleic acid
RISC	ribonucleic acid induced silencing complex
RNase A	ribonuclease A
rpm	revolutions per minute
RT	room temperature
RT-PCR	reverse transcription polymerase chain reaction
SAGE (Long-)	serial analysis of gene expression
SDS	sodium dodecyl sulphate
s	seconds
siRNA	short interfering RNA
SINE	short interspersed nuclear element
SNP	single nucleotide polymorphism
SSAHA	sequence Search and Alignment by Hashing Algorithm
ssDNA	single-stranded ribonucleic acid
SSH	suppressive subtractive hybridisation

SSX	synovial sarcoma X-linked
STS	sequence tagged site
TAP	transcript assembly programme
tBLAST -n, -x	translated blast local alignment sequence tool -vs nucleotide database -vs protein database
TLC	thin layer chromatography
TEL	telomere
TEMED	N,N,N',N'-tetramethylethylenediamine
TFB (-I, -II)	Transformation buffer -I, -II
TIGR	the Institute for Genomic Research
Tris	tris(hydroxymethyl)aminoethane
TSS	transcription start site
TTP	thymidine 5'-triphosphate
TTS	transcription termination site
U	unit
UCSC	University of California, Santa Cruz
UTR	untranslated region
uv	ultraviolet
V	volt
VEGA	vertebrate genome annotation
v/v	volume/volume
W	watt
w/v	weight/volume
WAS	Wiskott Aldrich syndrome
Wash U.	Washington University
WGS	whole genome shotgun
WTSI	Wellcome Trust Sanger Institute
Xace	X chromosome version of ACeDB
X-gal	5-bromo-4-chloro-3-indolyl-ß-D-galactoside
XIC	X-inactivation centre
Xist	X inactive specific transcript
YAC	yeast artificial chromosome

Chapter 1

Introduction

A fundamental description of biological organisms can be derived from their DNA sequence, which contains all of the necessary information for development, growth and reproduction. It is only in the past two decades that scientists have acquired the necessary knowledge and technologies to undertake efficient and accurate sequencing of genomes (Smith and Cantor 1986; Hunkapiller 1991). The first genome of a free-living organism to be sequenced was that of the pathogenic bacterium, *Haemophilus influenzae* (Fleischmann *et al.,* 1995). This has been followed by eukaryotic genomes ranging from the yeast, *Saccharomyces cerevisiae* (Yeast sequencing consortium, 1997), and the nematode, *Caenorhabditis elegans* (The *C. elegans* sequencing consortium) to more complex organisms including *Drosophila melanogaster* (Adams *et al.,* 2000; Myers *et al.,* 2000), *Mus musculus* (Waterston *et al.,* 2002) and *Homo sapiens* (International Human Genome Sequencing Consortium, IHGSC, 2004).

1.1 The Human Genome Project.

Launched in 1987, the human genome project (HGP) aimed to define accurately the euchromatic sequence of the human genome through the creation of genetic physical and sequence maps. The idea to sequence the entire human genome stemmed from several key experiments. Firstly, the sequencing of the genomes of the bacterial viruses ΦX174 (Sanger *et al.,* 1977; Sanger *et al.,* 1978) and lambda (Sanger *et al.,* 1982) and the animal virus *SV40* (Fiers *et al.,* 1978) demonstrated the feasibility of genome sequencing and its inherent value in understanding each organism. Secondly, the initiation of programmes to construct human genetic maps facilitated the identification of genes involved human disease based solely on their inheritance patterns (Botstein *et al.,* 1980) while physical maps of the model organisms *S.cerevisiae* (Olson *et al.,* 1986) and *C. elegans* (Coulson *et al.,* 1986) provided the unique opportunity to isolate genes and other regions solely on their chromosomal location. Finally, concurrent advances in high-throughput DNA sequencing techniques facilitated the genome sequencing of organisms with smaller, simpler genomes (Hunkapiller 1991). Together this work provided the necessary framework for the HGP, and a co-ordinated international collaboration began the mammoth task of sequencing the 24 human chromosomes.

Two different strategies were employed to obtain draft sequences of the human genome. The HGP adopted a hierarchical shotgun (HS) methodology combining both mapping and sequencing. Here, the genome was fragmented into overlapping

segments and then cloned into intermediate sizes, ranging from 40 kilobases (kb) (cosmids), to ~200 kb P1-artificial chromosomes (PACs) and bacterial artificial chromosomes (BACs). These fragments were assembled into maps and individual clones were then sequenced using a shotgun approach. The HS approach ensured that all assembly problems were contained within a small segment of the genome. A continuous sequence for each chromosome was produced by overlapping and merging the sequence of each clone with that from neighbouring clones (Lander *et al.,* 2001).

An alternative approach, the whole genome shotgun (WGS) sequencing strategy was adopted by a private company, Celera, to generate a draft sequence of the human genome (Venter *et al.,* 2001). The genomic DNA was sheared into differently sized fragments and directly inserted into clones, which were sequenced without any prior mapping. The genome sequence was then assembled from the shotgun clones and the HGP mapping data. Although this approach hastened the sequencing procedure by eliminating the preliminary mapping work, there was a greater risk of long-range mis-assembly.

In February 2001, draft sequences of the human genome were published using both the HS of the HGP (Lander *et al.,* 2001) and shotgun sequencing strategy adopted by Celera (Venter *et al.,* 2001). The genome sequences presented in these publications were not complete and represented 90% of the euchromatic sequence. The concurrent publication of two draft sequences also sparked much debate about the merits of each approach. However, it is not possible to perform a strict comparison of the HS and WGS strategies because the Celera group incorporated perfectly decomposed shotgun reads and mapping data from the public effort. Moreover, this comparison is purely academic as both drafts have been completely superseded by the HGP's finished sequence, where most of the draft sequence has been finished and many of the gaps have been closed to produce a virtually complete sequence (IHGSC 2004). The primary approach to close the gaps was an iterative procedure that involved 'walking' from the ends of previously positioned clones where a stretch of sequence from a terminal clone was used to identify a new clone (either experimentally or computationally) with overlapping sequence. This technique was repeated until either the gap was closed or a dead end was reached and reduced the number of gaps 400 fold from approximately 150,000 (Lander *et al.,* 2001) to 341 (IHGSC 2004). This work was also complemented by

'finishing' of clone sequences from their fragmented draft composition to contiguous genomic sequence.

The accuracy of the sequence has also been rigorously scrutinised. For example, the quality of the genome sequence was assessed by an independent group who examined approximately 34 megabases (Mb) of finished sequence (IGHSC, 2004). New sequence data and new sequence assemblies were generated that found an error rate of 1.1 per 100 kb for small events (≤ 50 base pair (bp), average 1.3 bp) and 0.03 per 100 kb for large events (> 50 bp; IHGSC, 2004). As of October, 2004 the assembled human genome sequence was 2.85 billion nucleotides in length and covered approximately 99% of the euchromatic genome with an error rate of approximately 1 event per 100,000 bp (IHGSC, 2004).

1.1.1 The human X chromosome

The human X chromosome represents 5% of the total human genome and is approximately 155 Mb in length (Ross *et al.,* 2005). It was sequenced by a number of centres, led by the Wellcome Trust Sanger Institute and including the Baylor College of Medicine (Texas, USA), the Washington University Genome Sequencing Centre (St. Louis, USA), the Max Planck Institute for Molecular Genetics (Berlin, Germany), and the Institute of Molecular Biotechnology (Jena, Germany).

Analysis of the sequence found that the X chromosome has a low GC content (39%) compared with the human genome (41%) but it is enriched for interspersed repetitive sequences, which account for 56% of its sequence compared with a genome average of 45%. In particular, the X chromosome is enriched in L1 repeats. Gene annotation confirmed that the X chromosome is gene poor as only 1,098 genes were identified (Ross *et al.,* 2005). The region of the X chromosome with the greatest gene density is Xp11.23 which contains approximately 10% of the gene content of the chromosome in 5% of its sequence (Ross *et al.,* 2005). This has been attributed to an expansion of cancer-testis antigen gene families. Seven hundred pseudogenes were also identified on the X chromosome, 92% of which (644/700) were processed.

The human X chromosome has many unique properties which make it a fascinating substrate for biological investigations. The X chromosome is one of the two sex chromosomes found in mammals. Females have two X chromosomes, while males

have one X chromosome and one Y chromosome. While it is thought that the X and Y chromosomes evolved from a pair of autosomes (Ohno 1967), the human X chromosome is both physically and genetically distinct from the Y chromosome (Charlesworth 1991).

Unlike autosomes, the X chromosome does not undergo recombination along the entire length during male meiosis. Instead, recombination is confined to the tips of the X and Y chromosomes, which are referred to as the pseudoautosomal regions (PARs). The shared homology between the X and Y chromosomes outside the PARs has been divided into five regions or 'evolutionary strata' on the X chromosome (Lahn and Page 1999; Ross *et al.,* 2005). Each strata differs in the extent of divergence between the X and Y chromosomes (Ross *et al.,* 2005) and the regions that have not undergone recombination for the longest time are the most diverged (Graves 1998). Only 54 of the X chromosome genes have active counterparts on the Y chromosome (Ross *et al.,* 2005). Most of these lie on the p arm of the X chromosome, while they are singletons or in small clusters throughout the euchromatic region of the Y chromosome (Lahn and Page 1999; Ross *et al.,* 2005).

The distinctive features of the X chromosome make it one of the most intensively studied human chromosomes. Approximately 10% of all human diseases with a Mendelian pattern of inheritance have been mapped to the X chromosome including colour-blindness and haemophilia (Online Mendelian Inheritance in Man, OMIM, http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM, Ross *et al.,* 2005). The presence of a single X chromosome in males reveals the effects of recessive mutations in X-linked genes while X-linked dominant phenotypes are manifested in both males and females. For hemizygous males, there only needs to be one copy of an X-linked recessive gene in order for the trait or disorder to be expressed and therefore males inheriting a recessive X-linked disorder are normally severely affected. Females, on the other hand, are usually heterozygous (obligate carriers) for X-linked recessive genes and are often asymptomatic. Mutation analysis of males affected with an X-linked disorder is more straightforward than that for autosomal diseases, since the male DNA contains only the one affected chromosome which can therefore by analysed directly. These types of inheritance patterns were first described in the 19th century for both colour-blindness and haemophilia. Although unable to define the precise mode of inheritance, 'Horner's

law' (1876) says that colour-blind fathers have colour-normal daughters; and these colour-normal daughters are the mothers of colour-blind sons (Jaeger, 1992*)*.

Two of the best known X-linked diseases are Duchenne Muscular Dystrophy (DMD) and the less severe Becker Muscular Dystrophy (BMD) both of which result from mutations in the dystrophin gene, *DMD*. DMD is estimated to affect on in 3,000 live male births. Its most distinctive feature of muscular weakness usually presents between the ages of two and five years. The onset of Duchenne muscular dystrophy usually occurs before age 3 years, and the victim is confined to a wheelchair by the ages of seven to twelve years. Death usually occurs by the age of 20 years. The onset of Becker muscular dystrophy is often in the third and fourth decade and survival to a relatively advanced age is frequent.

The *DMD* gene is the largest known human gene (Koenig, 1988) and spans 2.2 Mb of the X chromosome. It is located in Xp21.1, contains 79 exons and encodes a long rod-like molecule that links actin fibres to the extracellular basal lamina. A large proportion of dystrophin mutations are partial gene deletion or duplications (approximately 60% and 6%, respectively). Deletions that maintain the reading frame in the deleted transcript generally cause BMD while those that cause a frameshift usually cause DMD (Roberts *et al.,* 1995).

In mammalian female cells, one of the two X chromosomes is silenced in order to maintain a transcriptional balance between females (XX) and males (XY). To achieve this, one of the two X chromosomes is converted from a transcriptionally active euchromatic to an inert heterochromatic state. The inactive X chromosome contains hypoacetylated histones and methylation at many of its CpG islands. X inactivation is initiated by a counting step so that only one X is functional per diploid adult cell. In humans, the chromosome that is inactivated is chosen randomly (reviewed by Avner and Heard, 2001) but the precise mechanism that determines how this choice is made remains unsolved. However, in the extra-embryonic tissues of the mouse and in marsupials it is always the paternal X that is inactivated. X chromosome inactivation occurs early in development and is maintained in a clonal fashion throughout subsequent cell divisions (reviewed by Avner and Heard, 2001).

However, not all genes are silenced on the inactivated X chromosome (X_i). Some genes on the inactive X escape silencing, remain active and are transcribed from both the active and inactive X chromosomes. A global analysis of human X-linked transcribed sequences showed that approximately 15% of genes on the human X chromosome are able to escape inactivation (Carrel *et al.,* 1999; Carrel and Willard 2005). The distribution of these genes along the X chromosome is non-random with the majority of the genes that escape inactivation being located on the short arm (Xp) of the human X chromosome. The frequency of genes that escape inactivation is similar to that observed for autosomal genes in X: autosome translocations (Sharp *et al.,* 2001). This may be consistent with the recent autosomal origin of part of Xp (Graves 1998). Furthermore, genes with differential levels of expression have also been identified (Anderson and Brown, 2002).

Comparative gene mapping studies of eutherian, marsupial and monotreme mammals show that only a part of the eutherian X chromosome is shared with its marsupial and monotreme counterparts (Spencer *et al.,* 1991; Wilcox *et al.,* 1996). This region is represented by the long arm (Xq) and possibly parts of the proximal short arm, which may correspond to a conserved region on the X. Conversely, several genes on the short arm of the human X chromosome that map distal to Xp11.23 are clustered on chromosomes 5p and 1p of the tammar wallaby (Toder and Graves 1998; Glas *et al.,* 1999a). Thus, the region of the human X chromosome distal to Xp11.23 represents a recently acquired region that has been added to the ancient X in the eutherian lineage. The fusion point between the conserved region and the recently added region of the X chromosome appears to be located in human Xp11.23 (Ross *et al.,* 2005).

1.1.2 Applications of the human genome sequence

The HGP has created a reference sequence that is a suitable starting point for a wide variety of studies that serve to accelerate biomedical research. Amongst many possible applications including gene identification the human genome sequence also provides a suitable framework to map genetic variations, to aid the completion of genome sequences for other organisms and can also be used for comparative genomic studies.

Sequence variation

Although human DNA sequences are approximately 99.9% identical between individuals, variations in DNA sequence form the basis of heritable phenotypes.

Differences between the genomes of two individuals occur on average every 0.3 to 1 kb, which equates to 5-10 million differences in a 3.2 billion base pair genome. Types of mutation events that give rise to genetic variation include single nucleotide polymorphisms, (SNPs) insertions or deletions of DNA, (INDELS), as well duplications, inversions and translocations. Together, SNPs and INDELs account for inherited phenotypes and their information provides markers for linkage and association studies. Analysis of sequence variation provides information regarding the differences between individuals; variants may act as surrogate markers for an adjacent functional variant or they can have direct functional consequences if they occur in coding or regulatory regions.

SNPs are the commonest source of variation which account for approximately 90% of mutations events (The International SNP map working group, 2001). When comparing two genomes, SNPs with a frequency of > 1% occur approximately every 1,000 bp. In concert with the HGP, high-throughput platforms have been developed to identify SNPs. At the time of writing, ~5 million SNPs had been mapped to the human genome assembly (dbSNP at NCBI, build 124). They provide the basis for further disease association studies and for the HapMap project which is designed to understand more about long-range haplotype structure in human populations (The International HapMap Consortium, 2003).

Comparative genomics

Now that the human genome is complete, research efforts have shifted to acquiring the genome sequences of other organisms. Many of techniques developed in the HGP can be applied to aid the completion of genomic sequence in other model organisms. The completed human genome can also provide a reference sequence for the mapping and sequencing of other organisms. For example, the construction of a BAC map of the mouse genome was facilitated and accelerated by the availability of the human genome sequence (Gregory *et al.,* 2002). Genomes of other metazoan organisms are now being sequenced using a combination of both the clone-based sequencing strategy and WGS method. The WGS approach provides a useful sequence resource at an early stage, which is then combined with physically-mapped, clone-based sequence to provide a "finished" genome. The approach has been successfully used in the mouse genome project.

Comparing sequences in different species is a powerful tool for increasing the confidence of a predicted functional unit, or identifying novel functional units

(Frazer *et al.,* 2003). The underlying belief of this principle is that when two species diverge from a common ancestor, those sequences that maintain their original function are likely to remain conserved in both species throughout their subsequent independent evolution.

The information acquired by comparing genome sequences is dependent upon the evolutionary distance between the species that are compared. In general, a greater evolutionarily distance between the species is reflected by more divergent sequences and fewer shared functional units. Comparing sequences that diverged from a common ancestor approximately 450 million years ago (mya) (for example, human and fish) aids the identification of coding sequences while functional non-coding regions are generally not identified. If the evolutionary distance between the two species is reduced to approximately 90 mya, eg human and mouse, both non-coding and coding units are commonly conserved. A large number of features are conserved between recently evolved species such as human and chimp. The inclusion of a closely related species in a comparative analysis makes it possible to identify coding and non-coding sequences, but also those genomic sequences that may be responsible for traits that are unique to the reference species.

A central principle of the HGP was that its information be made readily available to the entire research community. This has resulted in many advances being made in our understanding of genome architecture, sequence variation, human disease and evolution. The work described in this thesis has used the human genome sequence as the primary substrate for cataloguing the transcript diversity of a small region on the X chromosome.

1.2 Describing the genomic landscape

The digital nature of the DNA enables descriptive analysis to be performed on whole genomes at base pair resolution. Some of the commonly described sequence features of the human genome are discussed below. These include the distribution of G+C content, CpG islands, repeat content (including segmental duplications) and gene content.

1.2.1 G+C content

It has long been accepted that the human genome is a mosaic of regions with fluctuating G+C content. Traditionally, the genome sequence was partitioned into G+C-rich and G+C-poor regions using techniques with poor resolution such as

differential staining or density gradient separation. However, the finished genome sequence allows the G+C composition to be addressed directly and precisely. The sequence has been used to record long range deviations from the genome average of 41% (Lander *et al.,* 2001). For example, the most distal 48 Mb of chromosome 1p (telomere to the STS marker D1S3279) has an average content of 47.1% while a 40-Mb region on chromosome 13 has an average G+C content of 36% (approximately between STS markers *A005X38* and *stST30423*, (Lander *et al.,* 2001)).

Long range variations in G+C content across the human genome have been correlated with the staining of chromosomal banding patterns (Saccone and Bernardi, 2001), methylation patterns (Caccio *et al.,* 1997), repeat coverage (Smith and Higgs, 1999), gene distribution and tissue specific expression profiles of products (Vinogradov 2003).

1.2.2 CpG islands

Methylation in the human genome occurs at cytosine residues in CpG di-nlueotides. Like other vertebrate genomes, the human genome has a conspicuous shortage of CpG di-nucleotides. CpG depletion of genomic sequence has been attributed to spontaneous deamination to methylated cytosine residues to form thymine residues. The human genome, however, contains regions where CpG di-nucleotides occur at a frequency closer to that predicted by the local G+C content. These regions or "islands" contain non-methylated CpG di-nucleotides and constitute a distinctive part of the human genome (Bird 1986).

CpG islands are stretches of DNA greater than 200 bp in length with a G+C content greater than 50% and an observed CpG/expected CpG ratio in excess of 0.6 (Gardiner-Garden and Frommer 1987). These distinctive features have been used to develop programmes that predict CpG islands within the genomic sequence (e.g. as part of the GRAIL program, and Gos Micklem, unpublished). The human genome sequence is predicted to contain approximately 290,000 CpG islands (Lander *et al.,* 2001) and approximately 60% of protein coding genes have a CpG island at their 5' end (Ponger *et al.,* 2001). On account of this association CpG islands are now commonly used as gene markers.

CpG island methylation is a mechanism by which gene expression may be regulated. Methylation of CpG islands in the promoter regions of genes has been

associated with biological processes, such as gene silencing by genomic imprinting and X chromosome inactivation or carcinogenesis (reviewed by Strathdee *et al.*, 2004).

1.2.3 Repeat Elements

In contrast to lower eukaryotes, the human genome has a high proportion of repetitive sequence. There are many types of human repeats which constitute approximately 50% of the human genome sequence. This fraction is substantially higher than the volume of genomic DNA that encodes proteins (approximately 2% of human genome sequence). The two major classes of repeats, transposons derived repeats, and tandem repeats differ in both their abundance and the way in which they have amplified throughout the genome.

Constituting approximately 45% of the genome, transposable elements have directly contributed towards the expansion of the human genome. The predominant repeats are LINEs (Long Interspersed Elements) and SINEs (Short Interspersed Elements), with smaller contributions from LTR elements and DNA transposons. Their genomic distribution appears to correlate with factors such as G+C content and gene density.

The other type of repeat element found in the human genome is tandem array repeats, or satellites. These repeats are juxtaposed copies of sequence motifs that range in size between 1 bp or 100 kb while the arrays range in size from 5 bp to over a 1 Mb. Together they occupy about 4% of the human genome. Perhaps the best known examples are alpha satellites found at the centromere of human chromosomes.

1.2.4 Segmental duplications

Segmental duplications are duplicated regions of genomic DNA that are greater than 1 kb in length, and have a sequence identity in excess of 90%. They may create inversions and other types of chromosomal rearrangements, and have been implicated in human disease (Mazzarella and Schlessinger, 1998; Emanuel and Shaikh, 2001). The human genome has a high proportion of segmental duplications covering approximately 5.3% of the euchromatic genome (IHGSC, 2004).

1.2.5 Pseudogenes

Pseudogenes are segments of DNA derived from normal genes. While the vast majority of pseudogenes appear to serve no biological function, several examples of functional pseudogenes have been published (Dahl *et al.,* 1990; Bristow *et al.,* 1993; Moreau-Aubry *et al.,* 2000). There are two classes of pseudogenes; processed and non-processed. Non-processed pseudogenes are generated by genomic duplication events. Compared to their functional counterpart, non-processed pseudogenes retain an exon/intron structure but acquire modifications that result in the loss of function at the transcriptional and/or translational level. Processed pseudogenes are produced by reverse transcription and subsequent re-integration of an mRNA transcript into the genome. As they are not under the same selective pressures as their functional counterparts to maintain function, processed pseudogenes can accumulate random mutations throughout the course of evolution.

The human genome is thought to contain approximately 20,000 processed pseudogenes (Zhang *et al.,* 2003), which are found in both euchromatic and heterochromatic DNA. The number of processed pseudogenes on each chromosome is proportional to length of the chromosome, but they are not uniformly distributed across the genome sequence. They are clustered in regions of intermediate G+C content and tend to be close to telomeres (Torrents *et al.,* 2003).

Compared to the total gene count, the human genome contains a high proportion of processed pseudogenes with ratio of approximately one functional gene per processed pseudogene. By comparison, the mouse has approximately 5.26 functional genes per processed pseudogene (5.26:1), *C. elegans* 200:1 and *D. melanogaster,* 400:1 (D'Errico *et al.,* 2004). It is suggested that these variations may be partially attributed to increased expression levels of the parent genes and an increase in the amount of sequence available for re-integration (Friedman and Hughes, 2001).

1.2.6 Gene content

One of the major aims of the HGP was to describe the entire catalogue of human genes. Techniques commonly employed in human gene identification, such as cDNA sequencing and computational analysis are discussed in greater detail in section 1.4. The current computationally determined version of the human gene catalogue

(extracted from Ensembl version 31.35d) has 24,194 gene loci with a total of 35,845 transcripts (1.48 transcripts per locus). These genes have a total of 245,231 exons with approximately 10.1 exons per gene. The total length of the genome sequence covered by coding exons is approximately 51 Mb which represents approximately 1.8% of the euchromatic sequence. Non-coding RNA genes and untranslated regions (UTRs) together represent almost 33 Mb of sequence, or 1.1% of the genome. This, however, is likely to be a severe underestimate because of difficulty in finding these features computationally.

1.2.7 Viewing genome sequence information

Genome sequences provide a natural scaffold for visualisation and organisation of genome features such as repeats, G+C content or homologous sequences. Listed below are some commonly used genome browsers.

ACeDb

ACeDB was originally developed for the *C. elega*ns genome project (Durbin and Thierry-Mieg, 1991). It permits integrated visualisation of genomic sequence, genome features (such as repeat location and G+C content) and information generated by gene prediction programmes and similarity searches. These data can be used to identify human genes, but manual curation is required to annotate their structures. This is discussed in more detail in section 3.1. ACeDB documentation code and data are available from the World Wide Web (www) site, http://www.acedb.org/.

Ensembl

Ensembl (http://www.ensembl.org, Hubbard *et al.,* 2005) provides access to data for 18 genomes including 12 vertebrates (human, chimpanzee, dog, cow, rat, mouse, chicken, fugu, zebrafish, tetraodon, opossum and frog), three chordates (two nematodes, *Caenorhabditis briggsae* and *Caenorhabditis elegans* and the sea squirt *Ciona intestinalis)* and three insects (fruitfly, mosquito and honeybee). Each of these genomes is automatically analysed for genomic features and genes. Gene models are assembled from alignments of protein, cDNA and EST sequences to the genome sequence. The gene models are assembled prior to their release in the public domain.

University of California Santa Cruz (UCSC) genome browser

Like Ensembl, the UCSC genome browser (http://genome.ucsc.edu, (Kent *et al.,* 2002)) automatically generates gene sets for genome sequences. In contrast to Ensembl, the UCSC browser quickly incorporates information from new species or new genome assemblies and, in general, releases the data before Ensembl. The UCSC also releases sequence information prior to the completion of a gene build.

Vertebrate Genome Annotation (Vega) database

The Vega database (http://vega.sanger.ac.uk (Ashurst *et al.,* 2005)) houses manually annotated genome data from human, mouse and zebrafish. Human annotation is performed on a chromosome by chromosome basis, and the database currently contains information for chromosomes 6, 7, 8, 10, 13, 14, 20, 22, X and Y. The gene structures are manually annotated and are therefore more accurate and detailed. Vega contains information that is frequently missing from other gene builds such as splice variants, polyadenylation features and non-coding genes. The data contained within the Vega database has been integrated into other genome browsers such as Ensembl and UCSC.

1.3 The human transcriptome

One of the challenges facing the scientific community in the post-genomic era is understanding how the functional information stored in the sequence of a genome can be conveyed to the rest of the cell. DNA-dependent RNA transcription is one way this transfer occurs. Unlike the genomic sequence which remains mostly static throughout the life of a cell, the transcriptome varies greatly over time and between cells that have the same genome, and because of this, our current understanding of this process is limited.

The transcriptome may be defined as the complete collection of RNA molecules transcribed and processed from the DNA of a cell. In addition to protein-encoding mRNAs, the transcriptome also contains non-coding RNAs, which are used for structural and regulatory purposes.

1.3.1 Non-coding genes

Significant advances have recently been made in the identification and characterisation of non-coding RNA molecules (Johnson *et al.,* 2005). Much of this has stemmed from the availability of the completed human genome sequence, in addition to large-scale cDNA sequencing projects (section 1.4). Non-coding RNAs

(ncRNAs) can be classed into two categories that can perform either housekeeping or regulatory functions. Housekeeping RNAs are usually small, constitutively expressed and necessary for cell viability. They include ribosomal RNAs (rRNAs), transfer RNAs (tRNAs), small nuclear RNAs (snRNAs) and small nucleolear RNAs (snoRNAs). Housekeeping RNAs have been implicated in such functions as mRNA splicing, protein synthesis and ribosomal rRNA modifications (Eddy 2001). Regulatory RNAs include the microRNAs (miRNA) and antisense RNAs.

MicroRNAs encode 21-25 nucleotide RNAs that are derived from longer transcripts. They were first identified in *C. elegans* but have also been found in nematodes, plants, insects and mammals where they are thought to act as post-transcriptional modulators of gene expression. More than 200 microRNAs are encoded by the human genome (Lim *et al.,* 2003).

Antisense RNAs are non-coding RNAs that overlap protein coding or non-coding genes, but are transcribed from the opposite strand. The regulation of gene expression by antisense transcripts is well established in prokaryotic systems (Wagner and Simons, 1994); however, the precise role of antisense transcripts in eukaryotic organisms remains to be clarified. Antisense transcripts are frequently associated with imprinted loci and can show a mutually exclusive expression pattern with their sense counterpart. The association is more apparent in the mouse where antisense transcripts have been found for 38% (22/58) of imprinted genes (Kiyosawa *et al.,* 2003). A well-characterised antisense transcript with a regulatory role is *Tsix* (Lee *et al.,* 1999). In early stages of X-inactivation *Tsix* represses expression of the gene *Xist*, and regulates choice of which X chromosome will be silenced. It is predicted that this phenomenon of antisense control is widespread in humans, and at least 20% of human genes have antisense transcripts (Chen *et al.,* 2004).

1.3.2 Protein Coding Genes

The delineation of the human protein coding gene catalogue has been one of the major aims of the HGP. It is predicted that the human protein-coding gene catalogue contains between 20,000 and 25,000 genes (IHGSC 2004). The coding sequence (cds) of a human gene is flanked by UTRs which serve both to regulate translation and stabilize transcripts. Coding exons, account for approximately 1.8%

of the euchromatic genome and tend to be found in regions of high G+C content. A discussion on human gene identification follows.

1.4 Gene identification

Unlike gene identification in prokaryotes or simple eukaryotes, deciphering the human genome presents daunting challenge in part because of the size of the genome which makes it difficult to distinguish genetic signal from noise. Simpler organisms have much more compact genomes. In the case of yeast, *S. cerevisiae,* for example, genes that encode proteins account for about 70% of the genome (Feldmann *et al.,* 1994). In humans the process of gene identification is further complicated by how genes are organized in the human genome; introns and intergenic regions can be extremely long, whereas exons are usually extremely small. It takes many more of these short, scattered exons to make one gene. Methods used to identify the exonic regions of the human genome can be divided into those that use expression data, and those that use only these sequences of one or more genomes (*de novo* or *ab initio* methods).

Prior to the availability of the human genome sequence, gene identification methods focused on the isolation and characterisation of RNA transcripts from various tissues and cell-types. RNA is an ideal substrate for gene identification as it is the product of transcription from all genes and analysis of processed mRNA permits inferences to be made about a exon/intron of a gene structure and its coding potential. Genes have been identified using Northern blots that contain total RNA or mRNA from a variety of different tissues to which genomic clones were hybridised. Also, the hybridisation of gene fragments to Northern blots can be used to extend gene structures, confirm transcript size or identify alternative variants. Genes identified using these methodologies include those mutated in DMD (Monaco *et al.,* 1986) and Menkes disease (Chelly *et al.,* 1993).

However, an RNA precursor is not always required to identify transcribed sequences. For example, the exon-trapping assay identifies candidate exons from randomly cloned genomic DNA fragments. To achieve this a genomic DNA fragment is inserted into a "mini-gene" vector that contains two exons separated by an intron (Duyk *et al.,* 1990). If the genomic DNA fragment contains an exon, it would be spliced and fused between the two flanking exons contained in the vector. Following transient transfection into a mammalian cell, splicing to include the

putative exon is confirmed by PCR analysis of harvested mRNA. This approach is limited to small-insert clones and can be limited by slow or cryptic splicing. Exon trapping has, though, been used to identify genes such as neurofibromatosis type 2, *NF2* (Trofatter *et al.,* 1993).

In order to facilitate experimental analysis, mRNA is frequently assayed in its more stable, reverse transcribed form, cDNA. Gene identification strategies can use cDNA clone libraries to identify clone(s), and hence gene(s) of interest, which can be then be sequenced. Selection techniques can include hybridisation with a gene fragment representing the coding region of a genes, oligonucleotide mixtures or genomic clones. This technique has been used successfully to identify many genes including factor IX (Choo *et al.,* 1982).

Random amplification of genes from cDNA synthesised by oligo dT priming (Verma *et al.,* 1972; Wickens *et al.,* 1978) and advances in sequencing technologies have been combined to facilitate high-throughput sequencing of human cDNAs. Expressed sequence tag (EST) sequencing projects rely on a cDNA library constructed from a tissue of interest under particular set of conditions. From this, randomly isolated clones are sequenced from either end until further sequencing no longer yields an acceptable frequency of novel cDNAs (Boguski and Schuler 1995). All EST sequences generated in high throughput studies are deposited in dbEST (Boguski *et al.,* 1993). Since its conception, the amount of data deposited in dbEST has increased exponentially. In April 2005, the database contained 26,630,649 sequences from 862 different organisms. A drawback of this method is the repeated sequencing of abundant transcripts, as a result of which there is a large amount of sequence redundancy in dbEST. For example, the human beta-actin (*ACTB*), gene is covered by 15,712 EST sequences in dbEST. In order to manage this redundancy, and increase the efficiency of gene identification and characterisation, EST sequences are clustered into non-redundant sets of sequences that represent distinct transcripts. These clusters are housed in various databases including UniGene (Boguski and Schuler 1995), the Merck Gene Index (Boguski *et al.,* 1995) and the TIGR Human cDNA Collection (Lee *et al.,* 2005).

ESTs have also been used to identify alternative transcripts, extend known gene structures and generate expression profiles for genes of interest. Indeed, the success of this strategy in defining the gene complement even raised suggestions

that EST sequencing could obviate the need for whole genome sequencing (Brenner *et al.,* 1990). While it is now known that expressed sequences alone cannot be used to describe the molecular composition of an organism, EST sequencing projects have still made a substantial contribution towards describing the human gene catalogue. There are, however, many situations where EST sequences alone do not provide adequate information to allow further analysis of gene function. This is partly because ESTs often fail to span entire transcripts. Full-length cDNA sequencing projects partially overcome this limitation by providing high quality sequence information from cDNA clones. Unlike EST sequences, full-length cDNA sequences are sequenced in both directions and cover the entire length of the cloned cDNA. Table 1.1 lists the human full-length cDNA sequencing projects and the number of sequences that they have generated to date.

Table 1.1 Full-length cDNA sequencing projects.

Sequencing centre (or consortium)	Focus	Number of sequenced clones	Reference (or URL)
NEDO (New Energy and Industrial Technology DevelopmentOrgansiation)	Full-length enriched cDNA clones obtained from oligo-capping derived cDNA libraries	**21243**	http://www.nedo.go.jp/bio-e/
DKFZ (The German cDNA sequencingconsortium)	Human specific - full length cDNA sequencing	**15069**	http://www.dkfz-heidelberg.de/mga/groups.asp?siteID=48
Kazusa (Kazusa cDNA sequencing project)	Sequence and analyse long (> 4 kb) human cDNAs	**2037**	http://www.kazusa.or.jp/cDNA
MGC (Mammalian Gene Collection)	To sequence cDNA clones containing the full-length open reading frame in human, mouse and rat	**Hs 18234** **Mm 14901** **Rn 4023**	http://mgc.nci.nih.gov/
WTSI cORF project (The Wellcome Trust Sanger Institute	To clone full length open reading frames using manually annotated DNA as the starting substrate	**341**	Collins *et al.,* 2004

To date, most cDNA cloning strategies have been biased towards genes that are abundantly expressed in readily accessible tissues and over time the discovery rate of new genes via cDNA sequencing has decreased. Based on a random selection scheme the identification of rare mRNAs from a cDNA library can be difficult because of their low representation. To overcome this, various normalising and subtracting techniques, such as suppressive subtractive hybridisation (SSH), have been developed to select and enrich samples for rare mRNAs (Diatchenko *et al.,* 1996). In SSH, mRNAs of the test and control samples are prepared and reverse-

transcribed into cDNA. Each cDNA is digested with the enzyme *Rsa*I to obtain shorter, blunt-ended fragments. The test cDNA is annealed with one of two adaptor sequences and is hybridised with an excess of control cDNA. A mixture of hybridisation products is formed, but a tiny fraction of cDNA remains unhybridised and single-stranded. This represents transcripts specific to the test sample. Another round of selection follows after which the specific fragment is amplified by PCR to make sure that sufficient amounts are available for further processing. Cloning, sequencing and comparison with a gene database establishes the identity of the gene(s). This method was used in a number of studies, such as the identification of human renal cell carcinoma associated genes (Stassar *et al.,* 2001).

Another limitation of full-length cDNA sequencing is as many as one-third of all cDNA sequences are truncated and do not extend to the CAP structure or poly(A) tail (Gerhard *et al.,* 2004). New techniques are being to overcome this shortfall (Carninci and Hayashizaki, 1999; Shibata *et al.,* 2001; Sugahara *et al.,* 2001; Carninci *et al.,* 2002; Gerhard *et al.,* 2004). One technique that is being routinely used to extend to the 5' ends of genes is CAP-TRAPPER where the 5' cap structure of mRNAs is biotinylated to permit the selection of capped transcripts. Combined with treatment to increase reverse transcriptase efficiency, this has extended the 5' ends for up to 63% of transcripts studied (Sugahara *et al.,* 2001).

1.4.1 Using computational analysis to identify genes

The completion of the human and other metazoan genomes has resulted in a wealth of raw genomic sequence data being deposited in the public domain. Scientists are now faced with the task of deciphering much of the information that the sequences contain. In theory, this could be completed using traditional low-throughput methods of gene identification, such as those discussed previously, but these methods simply cannot keep pace with the amount of genomic sequence that requires analysis. Computational analysis, on the other hand, can analyse vast amounts of data extremely quickly.

<u>De novo gene identification</u>

D*e novo* gene prediction programmes use probabilistic models to recognise sequence patterns that are characteristic of splice sites, translation initiation and termination sites, protein-coding regions, poly-adenylation (Brent and Guigo, 2004). For example, given a DNA sequence, a splice site donor model assigns

likelihood to the proposition that the sequence does indeed function as a splice site donor. Higher probabilistic values are assigned to true splice donor sites and lower likelihoods to other sequences. *De novo* gene prediction programmes differ in both the sequence characteristics that they identify and the number of genomes that are used to perform the process.

De novo programmes commonly use hidden Markov models (HMMs) and related models to identify exons and whole gene structures. For example, the programmes Genscan (Burge and Karlin 1997), GENIE (Kulp *et al.,* 1996) and HMMGENE (Krogh 1997) all use HMMs to predict the location of human genes. Whilst all programs suffer to varying degrees from lack of specificity and sensitivity (Guigo, *et al.,* 2000), they have nevertheless proved invaluable in the annotation of genomic sequence and can attain high levels of accuracy in some instances (>90% for Genscan (Guigo *et al.,* 2000)). Multiple programs can be used to increase sensitivity and confidence in prediction. In order to further aid gene identification, dual genome predictive algorithms exploit the higher levels of conservation that are found in functional sequences. Such programmes include SLAM (Parra *et al.,* 2003), SGP-2 (Korf *et al.,* 2001) and TWINSCAN (Flicek *et al.,* 2003). While the underlying algorithms of these programmes are beyond the scope of this thesis, it is worthy of note that when used on the human and mouse genomes these programmes have greater sensitivity and specificity than single-genome predictors (Guigo *et al.,* 2003). To optimise dual-genome gene prediction, genomes of suitable phylogenetic distance must be used. For example, the optimal reference for comparative analysis of the human genome would be a species more distant than the mouse (Zhang *et al.,* 2003). All *de novo* gene predictions, regardless of how their accuracy, still require experimental verification to confirm the existence of the predicted gene.

To date, d*e novo* prediction programmes have focused on the accurate identification of single spliced, non-overlapping protein coding regions with canonical splice sites. They often fail to predict the location of UTRs, alternative splice variants, overlapping or embedded genes, short intronless genes, or non-coding genes. Efforts are currently being made to address these shortfalls with the aim of producing a more comprehensive automatic gene catalogue of eukaryotic genomes (John *et al.,* 2004; Nam *et al.,* 2005; Sorek *et al.,* 2004a)

Sequence similarity searches using expressed sequence information

Automated gene prediction is commonly complemented by the use of sequence similiarity searches. The location of genes is confirmed using existing sequence information from expressed sequence tags (ESTs), cDNA and protein sequences. These sequences are housed in sequence databases such as EMBL, DDBJ, Genbank or SwissProt (proteins only) and can be overlaid on the human genome sequence using the Basic Local Alignment and Search Tool, BLAST (Altschul *et al.,* 1990). This programme or its derivatives recognise regions of shared sequence identity between two sequences and can indicate the exon structure of a gene. The type of BLAST analysis that is used to locate gene structures is listed in Table 1.2. cDNA and EST sequences are mapped onto the genome sequence using either BLASTn or the less specific alignment tools SSAHA (Sequence Search and Alignment by Hashing Algorithm, (Ning *et al.,* 2001)) and BLAT (Kent 2002). Protein sequences must be converted into a nucleotide sequence before they can be aligned onto the genomic sequence. This is achieved using tBLASTn or tBLASTx.

Table 1.2 Type of BLAST (or BLAST like analysis) used in gene annotation

BLAST type	Description
BLASTn	Nucleotide aligned against nucleotide.
tBLASTn	tBLASTn compares a protein sequence to the six-frame translations of a nucleotide database. It can be a very productive way of finding homologous protein coding regions in unannotated nucleotide sequences.
tBLASTx	Blastx compares translational products of the nucleotide query sequence to a protein database. It translates the query sequence in all six reading frames and provides combined significance statistics for hits to different frames.
SSAHA	SSAHA is a very fast tool for matching and aligning DNA sequences. It is most useful when looking for exact or 'almost exact' matches between two sequences.
BLAT	DNA Blat is designed to find sequence of 95% and greater similarity of length 40 bases or more. To achieve this, an index (11-mers) of the entire genome is created in memory.

In general, annotation of the human genome uses existing sequence information in concert with *de novo* analysis to produce high quality gene models. Analysis of unusual features such as non-canonical splice sites and conflicting sources of data can be assessed on individual genes. This procedure is discussed in greater detail in chapters 3 and 4. This type of analysis sometimes identified partial genes that require additional sequence information to complete their structures.

1.4.2 Gene expression analysis aids gene identification

The process of human gene identification is also aided by adapting experimental techniques that were traditionally used to generate gene structures. The use of serial analysis of gene expression (SAGE) and genomic tiling microarrays in human gene identification are discussed below.

SAGE analysis can be used to determine expression patterns and identify transcribed sequences. Here, cDNA "tag" libraries are constructed by isolating a defined region of a cDNA transcript rather than the entire transcript. The tags were originally between 9 and 14 bp long and were located at the 3' ends of genes (Velculescu *et al.,* 1995; Zhang and Frohman, 1997). The transcript identification procedure takes advantage of high-throughput sequencing technology to obtain a digital expression profile of cellular gene expression. It is possible to sequence many thousands of such tags from a tissue or cell specimen in order to obtain an accurate quantitative analysis of the relative levels of the genes expressed in that specimen. The ability to count many thousands of genes allows the detection of those that are expressed at very low levels in a high-throughput manner. However, they often lacked specificity with one tag frequently mapping to two related transcripts. To overcome this problem, a modified version of SAGE (LongSAGE) was developed (Saha *et al.,* 2002). Here, the tag length was increased to 21 bp to achieve greater specificity, to permit direct assignment of the tag to genomic sequence, and facilitate gene identification by RT-PCR amplification. This method has been used to identify novel transcript fragments throughout the entire length of an RNA transcript (Saha *et al.,* 2002; Wahl *et al.,* 2005). Further modifications to the LongSAGE methodology have been used to extend the length of known transcripts to initiation sites and polyadenylation sites (Wei *et al.,* 2004).

Genomic tiling arrays assay transcription at intervals of the genome using regularly spaced probes that can be either overlapping or separated. These probes may be selected to be complementary to one strand or both strands and may be synthesised oligonucleotides or spotted PCR products. Recent experiments using this technology have been performed on human chromosomes 20, 21 and 22 (Kapranov *et al.,* 2002; Rinn *et al.,* 2003; Kampa *et al,.* 2004; Schadt *et al.,* 2004). For example, Kapranov and colleagues prepared a tiling array for chromosomes 21 and 22 using 25mer oligonucleotides spaced at 35 bp resolution. When probed with double-stranded cDNA samples for 11 different cell lines, it was found that 94% of

the positive probes lay outside known exons (Kapranov *et al.,* 2002). Subsequent grouping of these positive probes into contiguous blocks representing parts of the same transcript found that approximately half of the groups lay outside known ESTs or mRNAs (Kampa *et al.,* 2004). These regions may be novel protein-coding or novel non-coding genes, alternative or antisense transcripts, or extensions to the 5' or 3' ends of known genes.

One limitation of these arrays is the accurate definition of exon structure. The highest resolution used to date is 35 bp which does not permit accurate identification of splice sites. As a result subtle changes in transcript structure may be missed. This problem may be overcome with use of higher resolution tiling arrays. At the time of writing, details of a tiling array for chromosome 10 using probes spaced at 5 bp resolution were emerging in the scientific literature (Cheng *et al.,* 2005).

1.4.3 Completion of gene structures

In addition to determining gene structure, complete gene annotation also requires identification of elements that regulate expression. This process is not trivial because regulatory regions are short sequences and have high levels of heterogeneity. Regulatory regions may be identified using experimental techniques such as DNase I hypersensitivity assays. This approach was used for example to identify novel regions with regulatory function in intron 21 of the *CFTR* gene (Phylactides *et al.,* 2002.)

One particular class of regulatory regions, the promoter, has been the focus of considerable attention. Promoters are modular DNA structures that contain a complex array of *cis*-acting regulatory elements that initiate transcription and control gene expression. Promoter sites typically have a complex structure consisting of multi-functional binding sites for proteins involved in the transcription initiation process. Because of this, they are difficult to identify accurately. To aid promoter identification computational algorithms can either predict transcription start sites or promoter elements themselves. However, these programmes often lack the specificity and sensitivity to locate promoters accurately. A recent investigation of promoter prediction programmes found that some programmes do not perform better than random guessing (Bajic *et al.,* 2004). The most accurate programmes include Eponine, which predicts transcription start sites (Down and

Hubbard 2002), First Exon Finder, which predicts both promoter sequences and transcription start sites (FirstEF, Davuluri *et al.,* 2001), and Promoter Inspector which localises promoter sequence within large regions of genomic sequence (Scherf *et al.,* 2000). It has been found that the accuracy of these programmes is further improved when they are used in concert with CpG island finders (Bajic *et al.,* 2004).

Despite advances in technologies and the availability of the completed genome sequence, promoter prediction and human gene identification are still non-trivial tasks. The process of gene identification is complicated by high levels of variation in both gene size and organisation. The largest human gene, *DMD*, spans 2.2 Mb (discussed in section 1.1.1.) (Koenig *et al.,* 1987; Tennyson *et al.,* 1995) while the genome also contains approximately 1,600 single exons genes, many of which less than 1 kb in length (http://www.ensembl.org). The longest open reading frame is in the titin gene (*TTN)* located on chromosome 2. This gene spans 280,000 base pairs, has 309 exons and produces a protein that is more than 33,000 amino acids long (Hillier *et al.,* 2005). The ability to describe the structure of a gene completely and precisely by a complete knowledge of all its transcripts is complicated by the use of multiple promoters, multiple polyadenylation sites and alternative splicing. These variations, with particular reference to alterations to exonic organisation, are discussed in more detail below.

1.5 Messenger RNA splicing

RNA transcription does not simply involve the direct copying of a DNA template into an RNA transcript. Before a transcript is transported from the nucleus it must undergo three major processing events to produce a fully mature mRNA transcript. A cap structure is first acquired at the 5' end, the splicing of introns from the body of the pre-mRNA and the addition of a poly(A) tail. Only, when all of these processes have taken place does the mature mRNA transcript contain all of the necessary information to perform its predetermined function.

1.5.1 Mechanisms of mRNA splicing

Using the technique of R-looping (DNA:mRNA hybridisation) and electron microscopy, Berget and Sharp published the first evidence for the presence of intron sequences in the adenovirus (Berget and Sharp, 1977). Since their identification there has been much debate about their origin, evolution and significance (Stoltzfus *et al.,* 1994). Intronic sequences are common in eukaryotic

genes and have been hypothesised to play an important role in the evolution of higher eukaryotes. Introns also contain motifs that regulate gene expression and organisation (Berget, 1995).

The majority of mammalian genes contain introns which must be removed from the premature mRNA template for a functional mRNA to be produced. This process known as splicing, is conducted by the spliceosome, a very large dynamic complex consisting of protein and RNA molecules. The precise composition of the spliceosome remains to be determined, but it is known to contain in excess of 150 protein molecules and numerous mRNA molecules (reviewed by Yong *et al.,* 2004). The best characterised of component of this complex are the small nuclear RNAs (snRNAs) which play a pivotal role in spliceosomal assembly and the two catalytic steps of the splicing reaction (Bentley 2002; Proudfoot *et al.,* 2002).

The precise recognition of intron-exon junctions (splice sites) and the correct pairing of the 5' splice site with its cognate 3' splice site is critical for splice site selection. To ensure that the appropriate splice sites are utilised spliceosomal assembly occurs in a step-wise fashion (reviewed Kalnina *et al.,* 2005). snRNPs are recruited to sequences located near the 5' and 3' ends of an intron. The arrangement, spacing and sequence context of adjacent splice sites, also contributes to accurate splice site selection. These regulatory elements are discussed in section 1.5.3 while the different sequences that are utilised in the splicing reaction are discussed below.

The most common dinucleotide sequences used in U2 dependent splicing are GT at the 5' end of the intron, and AG at the 3' end of the intron (frequently termed GT-AG introns). The 5' exon-intron junction utilises the consensus sequence AG|GURAGU, while the 3' exon-intron junction is marked by the sequence YAG|RNNN (R, purine; Y, pyridimine). Located approximately 100 bp upstream from the 3' splice site, the branch point is defined by the sequence CURAY, and contains a highly conserved adenosine followed by a pyridimine-rich tract. GC-AG introns are also utilised by the spliceosome in U2 dependent splicing.

Excision of intron sequences from a premature mRNA can be achieved by at least two functionally distinct pathways; the U2 dependent pathway and the U12 dependent pathway. These pathways are used at different frequencies, with the

U2 dependent pathway being used in more than 99% of splicing reactions. The two pathways use different small nuclear ribonucleoproteins (snRNPs) in the splicing reaction; the U2 dependent pathway employs the 5 snRNPs U1, U2, U4, U5, and U6 while the snRNPs U11, U12, U4atac and U6atac are used in the U12 dependent pathway. These molecules differ in the sequence composition and are therefore complementary to different mRNA sequences at the 5' and 3' exon-intron junctions and branchpoint.

The most common intron used in U12 dependent splicing is AT-AC. This was first reported in the human cartilage matrix protein (*CMP*) (Jackson 1991). The AT-AC intron is also complementary to different branch sites. These are ATATCCTY and TCCTTRAY. Other recognised intron boundaries include AT-AA and AT-AG (Wu and Krainer 1999).

The splicing reaction is a two step process that occurs in concert with mRNA transcription. Briefly, the 2' hydroxyl group of the branchpoint adenosine residues acts a nucleophile to attack the 5' exon-intron border. This exposes the 5' hydroxyl end of the exon and also creates a lariat structure that contains the intron sequence and the 3' end of the adjacent exon. The second, trans-esterification reaction fuses the exon where the free 5' hydroxyl end replaces the intron at the 5' end of the second exon. An overview of this process is displayed in Figure 1.1 and is reviewed by Kornblihtt and colleagues (Kornblihtt *et al.,* 2004).

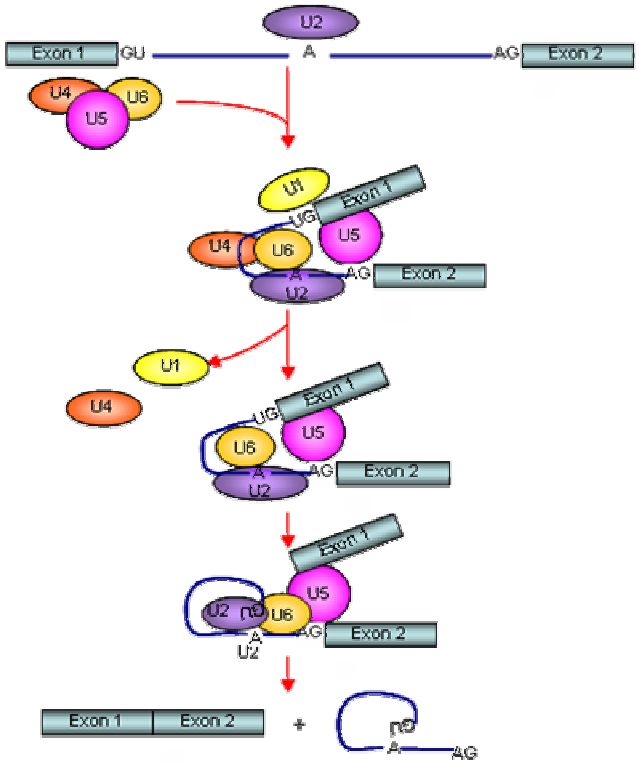


Figure 1.1 mRNA splicing

Splicing involves several RNA-protein complexes, called small nuclear ribonucleoproteins (snRNPs), which together make up the spliceosome. It occurs in several stages. U1 snRNP binds to the boundary between exon 1 and the intron by recognizing a specific sequence. U2 snRNP subsequently binds to the branch site (A) and then U4/U5/U6 triple snRNPs join in. After a dynamic rearrangement, U1 and U4 are destabilized, and the remaining snRNP complex is activated for the two steps that remove the intron and stitch together exons 1 and 2. Adapted from Gu, J. & Reddy, R. Cellular RNAs: varied roles in *Encyclopedia of Life Sciences* (Nature Publishing Group, London, 2001)

1.5.2 Alternative mRNA splicing

During mRNA splicing exons are selected to be included in the processed mRNA. Some exons, displayed variable expression patterns and are not always included in the mature mRNA. This phenomenon is known as alternative splicing, and allows the optional inclusion or substitution of some exons within the constant framework provided by constitutive exons (reviewed by Modrek and Lee, 2002). Variations in a transcript may result in altered expression patterns, or changes to a gene's cognate protein. The impact of alternative splicing on the size of an organism's transcriptome is most spectacularly illustrated by the highly characterised fruit fly gene, Down syndrome cell adhesion molecule (DSCAM) which has the ability to produce 38,016 transcripts through the variable selection of exons (Schmucker and

Flanagan 2004). This is substantially larger than the total number of genes contained within the organism's genome.

Many spliced genes have the potential to incorporate one or more changes into its exon structure. The combination of splice sites employed in the splicing process can vary within each gene and can be the result of either highly regulated or aberrant splicing events. At least five different forms of alternative splicing have been identified and are displayed in Figure 1.2: 1) Entire exons may be added or deleted. 2) Additional mRNA may be retained at either the exon acceptor (3') or 3) donor (5') splice site. Conversely, mRNA may be deleted at either of these locations. 4) Transcripts may also use mutually exclusive exons, where one of two possible exons is integrated into the transcript's structure. 5) Transcript variants have also been identified where an intronic sequence is not removed during the splicing process.

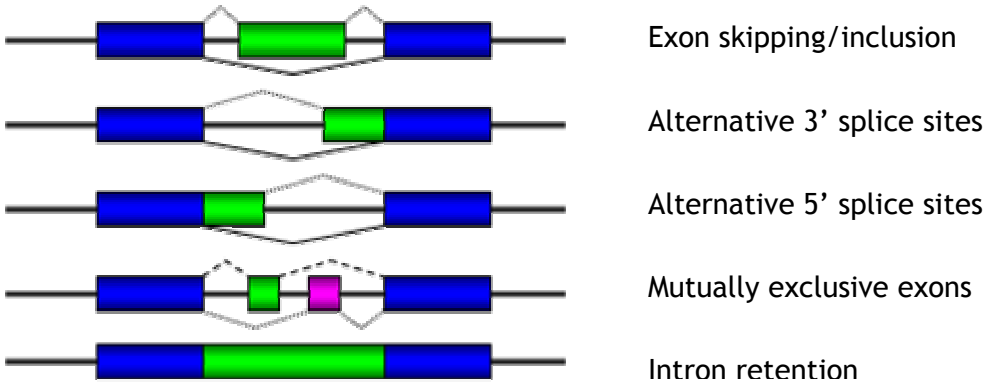


Figure 1.2 Types of splicing variation
Shown are the possible splicing patterns that can occur through alternative splicing. The genome sequence is displayed as a thick black line. Constitutive exons are displayed in blue, while alternative exons are displayed in green (or pink). The exon junctions between constitutive exons are displayed as solid black lines. Dashed lines represent alternative exon junctions.

1.5.3 Regulation of splicing

Although the plasticity of the human transcriptome is well documented, little is known about the molecular mechanisms that govern its variation. It is thought the alternative splicing is controlled by a combination of *cis* and *trans* acting factors. These include splice site strength, regulatory elements, protein regulatory factors and RNA secondary structure. Splicing patterns are inherently flexible, with variations observed in different cells and tissues and at different stages of

development. Inducible changes to splicing can also occur as a function of cell excitation in neuronal cells, T-cell activation, heat shock or cell-cycle progression.

Splice site strength

It is possible to measure the strength of a splice site by comparing its sequence to the consensus of the sequences that surround exon/intron junctions. Splice site scores are generated by comparing the sequence at each position to the frequency with which that base appears in splice sites. Higher scores are associated with stronger splice sites, and in general constitutively spliced exons will have a higher score than alternative exons.

Cis regulatory signals

In addition to the sequence composition of splice sites, other sequence motifs influence the patterns of mRNA splicing. The two major classes of *cis* elements regulate exon inclusion in either a positive or negative fashion. Splicing enhancer and splicing silencer elements have been identified in both exonic and intronic sequences. These have an important role in regulating mRNA splicing by interacting with components of the spliceosome. Exon splicing enhancers (ESEs) are required to promote the inclusion of exon sequences in a mature mRNA transcript. These sequence motifs can also act as barriers to prevent exon skipping (Ibrahim el *et al.,* 2005). Most ESE elements interact with members of the arginine/serine rich (RS) domain-containing protein family, which bind to the intronic branchpoint and to other components of the spliceosome to enhance the recognition of adjacent splice sites (Shen *et al.,* 2004). An example of this class of regulatory motif is the purine-rich exon splicing enhancer that is commonly found in both alternative and constitutive exons. Purine-rich ESEs are bound by members of the RS family and promote the use of weak 3' splice sites (Fairbrother *et al.,* 2002).

The second class of *cis* regulatory elements are exon splicing silencers (ESSs). These are prevalent in the human genome and approximately one third of randomly cloned genomic DNAs of 100 bp were shown to exhibit ESS activity (Fairbrother and Chasin, 2000). In contrast with ESE elements, ESS elements inhibit the use of adjacent splice sites. Until recently the sequence composition of these motifs was largely unknown. However, bioinformatic investigations and cell-based splicing

reporter assays have identified over 1,000 short sequence motifs that silence exon splicing (Sironi *et al.,* 2004; Wang *et al.,* 2004; Zhang and Chasin, 2004).

The inhibition of exon identification or splice-site usage involves ESS elements interacting with negative regulators. Members of the heterogeneous nuclear ribonucleoprotein (hnRNP) family are frequently involved in such interactions (Zheng *et al.,* 1998; Zhu *et al.,* 2001). The best characterised hnRNP family member is hnRNP1 commonly known as polypyridimine-tract-binding protein (*PTB*). A role for *PTB* in alternative splicing was first proposed by Mulligan and co-workers who studied the alternative splicing patterns of ß-tropomyosin transcripts (Mulligan *et al.,* 1992). They demonstrated that mutations within *cis* elements upstream of the skeletal muscle-specific exon 7 of the ß-tropomyosin pre-mRNA, resulted in its inclusion in *HeLa* cells *in vivo* and these mutations were demonstrated to disrupt the binding of *PTB in vitro*. The propensity of *PTB* to bind to stretches of pyrimidines has led to the hypothesis that it may compete with the splicing factor U2 for the polypyrimidine tract upstream of a regulated exon, thus causing skipping of that exon by blocking the recognition of the branch point by the splicing machinery (Singh *et al.,* 1995).

The selection of correct splicing variants in a given cell or tissue type is believed to be co-ordinated by multiple and sometimes overlapping ESE and ESS elements. Tissue-specific regulatory elements can be identified by comparing genes that are subject to alternative splicing in different tissues. The inclusion of tissue-specific exons can also be modulated by multiple sequence motifs acting in either a co-operative or an antagonistic manner. Molecular approaches have been used to identify a ternary combination of two exonic (UAGG) and one 5' proximal (GGGG) motifs that function co-operatively to generate brain specific transcripts of the glutamate NMDA R1 receptor, *GRIN1* (Han *et al.,* 2005).

RNA secondary structure

The ability of RNA molecules to form highly stable secondary structures has been established using *in vitro* and *in vivo* analysis (reviewed by Holbrook, 2005). Secondary RNA structures may influence splicing patterns in either a direct or indirect manner. They have been found in pre-mRNA exonic or intronic sequences and may affect accessibility of various sequence motifs to hnRNPs and regulatory proteins that are involved in the splicing cascade (Buratti and Baralle, 2004). For

example, secondary structure may affect the recognition of splice sites and branchpoints by hindering the accessibility of basic splicing factors. This may hinder intron processivity and promote exon skipping. Such structures may also affect the exposure of enhancer or silencer elements. The presence of secondary structures in mRNA transcripts has been proposed to influence the generation of human growth factor isoforms (Estes *et al.,* 1992).

RNA secondary structures may also alter the distance between different splice sites. Changes in the spatial distribution of splice site sequences resulting from formation of RNA secondary structures may serve to provide a greater level of flexibility in the control of splicing. The gene product of heterogeneous nuclear ribonucleoprotein A1, hnRNPA1, auto-regulates its own tissue specific splicing patterns by promoting the formation of mRNA secondary structure in non-neuronal cells. A loop structure is induced by polypyrimidine binding proteins and results in the removal of exon 7b from the transcript (Blanchette and Chabot, 1999).

1.5.4 cDNA and EST sequences facilitate the identification and characterisation of alternative splicing events

The identification of alternatively spliced genes has been facilitated by the large number of transcripts sequences generated in high throughput sequencing projects. As discussed in section 1.4, the number of cDNA and EST sequences available in public databases has increased by over 250% in the past four years and in general the deeper the sequence coverage of an EST or cDNA library, the more likely it is that alternative transcripts will be identified. Early estimates of the frequency of alternative splicing predicted that 5% of genes have more than one transcript (Sharp *et al.,* 1994). However, an increase in the number of transcript sequences in the public domain has seen the figure increase to between 25% and 70% (Brett *et al.,* 2000; Kan *et al.,* 2001; Modrek *et al.,* 2001; Mironov and Gel'fand, 2004).

Information about alternative transcripts may also be obtained from experimentally determined and characterised alternative splicing events. This can be extracted from databases such as PubMed (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=,SwissProt) (Bairoch *et al.,* 2005) and AEDB (Thanaraj *et al.,* 2004). AEDB contains high quality annotation, functional analysis and tissue expression profiles. In April 2005, it contained entries for 213 human genes.

Transcript variants are commonly identified by mapping transcript sequence clusters onto the genomic sequence. With this approach, alternative splicing forms of EST and cDNA sequences can be detected using computational alignment programmes such as BLAST (Altschul *et al.,* 1990), SIM4 (Florea *et al.,* 1998), FASTA (Pearson, 1990) and Spidey (Wheelan *et al.,* 2001).

As the sequence quality of transcript sequences, especially EST sequences, is highly variable it is important to exclude biologically non-relevant transcripts from the datasets. In order to do this, the fidelity of each alignment is assessed with the objective of removing false positives from the dataset. Alignments with low sequence homology and small, repetitive or unspliced alignments are frequently removed from datasets. Additional sequences whose splice site dinucleotides do not match consensus sequences may also be removed. After filtering, blocks shared between the expressed and genome sequences are considered to be exons and are linked to create an exon-intron structure. Transcript variations are identified by comparing the start and end position of each exon. With these computational methods it is possible to interrogate large datasets of sequence information and conduct analysis of alternative splicing on a genome wide scale (Modrek and Lee, 2002). Such analysis has been completed using different alignment and filtering criteria and has resulted in the production of several alternative splicing databases, which are listed in Table 1.3.

The problems associated with the use of EST and cDNA sequences to identify alternative transcripts are similar to those discussed in section 1.4. Since they are generated from single-pass sequence reads and so are of lower quality, EST sequences are often unreliable sources of transcript information. These sequences are also biased towards the 5' and 3' end of genes and rarely span their entire gene length. For these reasons it is best not to infer any functional changes from a novel alternative transcript that has been identified from EST sequences alone. However, an advantage of using ESTs in the identification of alternative transcripts is that additional information about the regulation and expression profiles may be obtained. Several studies have used normalised EST information to identify tissue specific expression patterns of alternatively spliced genes (Xie *et al.,* 2002; Gupta *et al.,* 2004; Yeo *et al.,* 2004).

Table 1.3 A comparison of alternative splicing databases.

Database	Type of sequence supported	Filtering process	Estimated frequency of alternative splicing	Reference
PALS db	EST, mRNA	Sequences must have greater than 95% sequence homology over 50 bp	21%	Huang *et al.*, 2002
HASDB	EST, mRNA	Sequences must have standard splice site consensus	42%	Modrek *et al.*, 2001
ASD		Manually curated from published transcripts		Thanaraj *et al.*, 2004
AEDB	mRNA	Sequences must conform to either the GT-AG or	1100 entries	
AltExtron	EST, mRNA	GC-AG splice sites. Duplicate genes (> 99% un-gapped sequence identity) removed. Hypervariable genes are removed	37%	
ASAP	mRNA, EST	Sequences are clustered and splice validated	44-57%	Lee *et al.*, 2003
SpliceNest	EST mRNA	Identifies structural variation in transcript clusters	45%	Krause *et al.*, 2002
ASDB	Protein, EST, mRNA	Catalogue of tagged alternative splicing events	Not given	Dralyuk *et al.*, 2000
ECGene	EST, mRNA	Part A – High quality cDNA sequences only Part B- More than one transcript	26% 44%	Kim *et al.*, 2005
ASG	EST, mRNA	Greater than 95% sequence identity over 100 bp. Sequences identified on the complementary strand to the gene are eliminated	65%	Leipzig *et al.*, 2004
TAP	EST, mRNA	Greater than 92% sequence identity, confirmed by more than 1 transcript. Sequences must have standard splice site consensus	33%	Kan *et al.*, 2001
ProSplicer	EST, mRNA, protein	Sequences must be greater than 15 bp long. Sequences identified on the complementary strand to the gene are eliminated	Not given	Huang *et al.*, 2003

1.6 Functional consequences of alternative splicing

Many questions remain about the functional influence of alternative splicing in human biology. These questions require evidence to define not only the spatial and temporal expression patterns of alternative transcripts but the molecular mechanisms that control them. The frequency of regulated and aberrant splicing events must be determined.

It has been estimated that up to 28% of alternative splicing variants result in a disruption to a functional domain (Kriventseva *et al.,* 2003). Some examples of the functional alterations induced by alternative splicing include induction or repression of apoptosis (Schwerk and Schulze-Osthoff, 2005), up or down regulating transcription (Kawai *et al.,* 2005) or changing ligand affinity (Pan *et al.,* 2005).

1.6.1 Expression profiling of alternative splice forms

Monitoring the abundance of alternative transcripts could help further define the physiological and pathological consequences of alternative splicing. Most experimental methods used to analyse transcript expression are based on hybridisation procedures using probes that generally exploit the unique exon-exon junctions of alternative transcripts. In order to quantify the abundance of alternative transcripts accurately, and avoid any cross-hybridisation, particular care must be taken to design oligonucleotides that are unique to each variant.

Traditional methods used to detect specific mRNAs such as Northern blot analysis, and RNase protection assays required microgram quantities of RNA. The recent introduction of fluorescence techniques has overcome this limitation by increasing the efficiency with which information about transcript variants can be acquired. The expression patterns of transcript variants can also be quantified accurately using real-time (RT) PCR while fluorescently labelled RNA can be used to probe thousands of different transcripts in parallel using techniques such as microarray analysis, bead based fibre optic arrays and polymerase colony (polony) technology. A brief discussion of these techniques with particular reference to the characterisation of transcript variants follows.

Microarray technology

Microarray technologies can be used to estimate the frequency of alternative splicing. Exon-junction microarrays contain oligonucleotides that span both constitutive and alternative exon junctions. In these arrays, a positive hybridisation with cDNA should only occur when both exons are adjacent to each other in the transcript (Clark and Thanaraj, 2002; Castle *et al.,* 2003; Modrek and Lee, 2003). Johnson and colleagues (2003) completed such an analysis using a microarray that contained approximately 125,000 exon junctions. They analysed the expression of 10,000 multi-exon human genes in 52 tissues and estimated the frequency of alternative splicing to be 74%. Limitations of this method include that it cannot be used to identify novel exon junctions. Moreover, as each exon-junction is assayed independently, alternative splicing events cannot be put into context with the full length sequence thereby prohibiting any functional inferences about the impact of alternative splicing events.

Variations in transcript structures can also be detected using overlapping oligonucleotide arrays designed to non-repetitive regions of the genome. The locations of exons can be defined by hybridising a transcript directly to small region of genome sequence. Unlike exon-junction arrays, this method does not require any prior knowledge of exon size or location, thereby allowing the identification of novel exons and exon boundaries. However, this method only identifies single exons and does not link them into a transcript structure. Using overlapping oligonucleotide arrays at a resolution of 35 bp, it has been estimated that 77-89% of genes on human chromosomes 21 and 22 have more than one transcript (Kampa *et al.,* 2004).

Real-time PCR analysis

Because of its higher sensitivity and greater accuracy in determining concentration, real-time PCR of mRNA is now used in preference to classical methods such as dot-blot analysis, limiting dilution PCR and competitive PCR. Unlike endpoint RT-PCR, real-time quantification is defined by C_T (cycle threshold number), a fixed threshold where PCR amplification is still in the exponential phase and reaction components are not limited. At this point, the amplification efficiency is constant and an increase in fluorescent signal during this phase corresponds directly to increase in the PCR product. Transcript specific real-time PCR has been employed

to determine the abundance of alternative variants for genes such as brain-derived neurotrophic factor (*BDNF*; Altieri *et al.,* 2004), Interleukin- receptor, (*HIL-5R alpha;* Perez *et al.,* 2003).

Quantifying alternative transcripts by real-time PCR can be either absolute or relative. Absolute quantitation determines the quantity of mRNA using an absolute standard curve for each individual amplicon. This is then used to calculate the precise copy number of mRNA transcript per cell or unit of tissue mass. The relative quantitation method compares the expression of the "target" to a "reference" gene. Reference genes should have ubiquitous expression, characteristics commonly associated with housekeeping genes, such as beta-actin (*ACTB)*. Relative expression is detected using the formula $2^{-\Delta\Delta CT}$ (Livak and Schmittgen, 2001), which is based on the assumption that the amplification efficiencies of the "reference" and "target" genes are approximately equal.

Two different detection methods are commonly used to quantify mRNA levels, SYBR Green 1 and probe-specific detection e.g., TaqMan. SYBR Green1 binds in the minor groove of double stranded DNA, where its fluorescence increases over a hundred fold. Detection of dsDNA using SYBR Green is non-specific as the dye also binds to all dsDNA, including non-specific products and primer-dimers. If using SYBR green to detect mRNA transcripts, extreme care must be taken to ensure that all primer pair combinations are highly specific. The TaqMan method uses a transcript specific probe that emits fluorescence upon the successful amplification of the desired PCR product. Unlike SYBR Green, TaqMan analysis directly quantifies the amplification of one specific PCR product. Although they are highly specific, Taqman probes are expensive and therefore not suited to high-throughput analysis.

Bead based fibre-optic arrays

A novel approach to large scale analysis of known alternative splicing events based on a fibre-optic microarray platforms has also been described (Yeakley *et al.,* 2002). The bead arrays are assembled by loading a mixture of micro-spheres (beads) onto the tip of an etched fibre-optic bundle. Each bead contains a specific oligonucleotide (address) sequence which acts as a probe for hybridisation experiments. 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).

Polymerase colony (polony) technology

Polony technology allows alternative splicing events in the same molecule to be detected and quantified (Zhu *et al.,* 2003). Here, the cDNA template, is mixed into an acrylamide matrix together with all of the necessary components for PCR. The PCR is completed in the solid phase, each template giving rise to an individual "colony" of amplification products. These products are then targets for probe hybridisation with fluorescently labelled exon specific probes. The fluorescence is monitored using a standard microarray scanner. With the use of spectrally distinct fluorophores, this method can also be used to characterise several alternative splicing events in parallel.

1.6.2 Tissue specific regulation of alternative splicing

The expression of alternative mRNA transcripts and proteins encoded by a single gene can be regulated in a tissue, temporal or stimulus-dependent manner and may serve to control cellular function. Many examples of tissue specific alternative splicing are available in the literature. Sixty-five percent of human tissues display tissue specific alternative splicing events, and it has been suggested that the brain has the highest degree of alternative splicing (Xu *et al.,* 2002). High levels of transcript variation have been observed in systems that require a high level of variation in their proteome such as the immune and nervous systems (Grabowski and Black 2001; Lynch 2004). Defining the molecular mechanisms that govern the highly variable but regulated patterns of expression has proved to be extremely difficult. To date, no single critical factor has been shown to modulate tissue specific splicing patterns in any system.

Tissue specific splicing may be regulated by proteins whose expression is restricted to certain cell types. For example, the neuronal proteins, *Nova-1* and *Nova-2* regulate splicing events in the central nervous system (CNS; Buckanovich *et al.,* 1993; Yang *et al.,* 1998). These proteins are expressed almost exclusively in neurons of the CNS, and contain three RNA binding domains. *Nova-1* influences mRNA splicing by binding to an intronic sequence in transcript of the gene glycine receptor $\alpha 2$ (*GlyRα2*), in the brain stem and spinal cord but not the forebrain

(Jensen *et al.,* 2000). *Nova-2* has a broader CNS distribution than *Nova-1*, and is likely to function in RNA metabolism (Yang *et al.,* 1998).

1.6.3 Evolutionary conservation of alternative splicing

Proteomic diversity is a hallmark of complex eukaryotic organisms. One post-transcriptional mechanism that seems to increase protein diversity of such organisms is alternative splicing. This process may explain the disparity between the low number of protein-coding genes (20,000-25,000) and the total number of human proteins (>90,000). The frequency of alternative splicing is highest in humans and other mammals (Modrek and Lee 2002) which would also explain the diversity of mammalian proteomes when compared to the proteomes of *C. elegans* or *Drosophila*, which have a gene catalogue only 25% smaller than humans (The *C. elegans* sequencing consortium 1998; Adams *et al.,* 2000).

The appearance of multi-exon genes and constitutive splicing probably predated the appearance of alternative splicing (Ast, 2004). Although introns are present in most eukaryotes, evidence for alternative splicing has only been documented in multi-cellular eukaryotes. In yeast *S. cerevisiae,* introns are only found in only ~3% of its genes (~253 introns), and only six genes have two introns, none of which have been reported to be alternatively spliced (Barrass and Beggs, 2003). Alternative splicing may have originated from multi-intron genes through DNA mutations and/or the evolution of splicing regulatory proteins.

A greater understanding of the evolution of alternative splicing has been achieved by comparing sequence from the human genome and transcriptome to other species. Humans and rodents are considered to be separated by an ideal evolutionary distance to study the conservation of alternative variants (Zhang and Gerstein, 2003). Greater than 90% of the human and mouse genomes have been partitioned into regions of conserved synteny. The gene content of the human and mouse are surprisingly similar: 99% of all human genes have a functional orthologue in the mouse and *vice versa* (Dehal *et al.,* 2001; Mural *et al.,* 2002; Waterston *et al.,* 2002). The similarity extends to gene structures, where 90% of constitutive exons share the same boundaries. However, alternatively spliced exons do not share the same level of conservation. An analysis of the human, mouse and rat has found that only 72% of alternatively spliced exons were conserved between the

three species (Modrek and Lee, 2003). Using a smaller, higher quality dataset, Thanaraj and colleagues examined the conservation of splice junctions between the human and mouse (Thanaraj and Stamm, 2003). Here, 74% of constitutive exons junctions were conserved while this value decreased to 61% for alternatively spliced junctions. From these data, it cannot be discerned if the apparent lower of conservation of alternatively spliced exons is attributed to loss or gain of sequence in either species. It is also difficult to tell if these changes have arisen as the result of positive, negative or neutral selection. Additional species must be included in these analyses for greater insights to be made into the selective pressures that define alternative transcripts.

More recently, comparative genome analysis has been used to identify conserved features of alternative exons in the human and mouse (Sogayar *et al.,* 2004; Yeo *et al.,* 2005). Conserved alternative exons are more likely to be flanked by conserved intronic sequences. They are also shorter than constitutive exons, their size tends to be divisible by three and share a higher level of sequence identity to their mouse counterparts (Sorek *et al.,* 2004). These features, and the underlying datasets, have been used to generate computational algorithims to predict the expression of alternative exons *a priori.*

The lack of conservation of alternative exons in the human, mouse and rat genomes, combined with the relatively low representation of alternative exons in transcript databases suggests that alternative splicing may play a unique role in evolution, serving to reduce negative selection pressure against mutations such as exon creation and loss (Modrek and Lee, 2003; Boue *et al.,* 2003). Using alternative splice information obtained from EST and cDNA sequences, Xing and Lee (2004) found that alternative transcripts have a higher frequency of premature termination codons (PTCs) compared with the major transcript of each gene. Moreover, the frequency of PTC harbouring transcripts was lower on the X chromosome when compared to autosomes (Xing and Lee, 2004). This may be because the potentially deleterious consequences of alternative splicing are masked in the heterozygous state where the wild-type copy of the gene would ensure that the original transcript would still be produced at 50% of its original level. As discussed in section 1.1, X chromosome genes are hemizygous in males and in general, X gene expression is limited to one copy in females. The increased frequency of potentially toxic PTC harbouring transcripts in diploid chromosomes

may reflect a decrease in the selective pressure preventing the transcription of aberrant transcripts that may produce a dominant negative effect.

One possible source of transcript variation in primate species is achieved by the exonisation of *Alu* repeats. *Alu* repeats are primate specific and account for more than 10% of the human genome. More than 5% of alternatively spliced exons in the human genome are *Alu* derived (Sorek *et al.,* 2002), which is not unexpected as both strands of *Alu* repeats harbour motifs that resemble consensus splice sites (Makalowski *et al.,* 1994). Efficient splicing of *Alu* repeats may be induced by point mutations. By aligning transcribed *Alu* exons to their ancestral sequence it was possible to identify sequence changes that are most responsible for the exonisation process (Lev-Maor *et al.,* 2003). Here point mutations in one of two AG dinucleotides can produce a 3' splice site that is responsible for alternative splicing (Lev-Moar *et al.,* 2003).

Not all *Alu* derived exons are alternatively spliced. Newly created constitutively spliced *Alu* exons have been shown to generate new products at the expense of the original (Lev-Maor *et al.,* 2003). The biological impact of these changes remains to be determined.

The recent influx of comparative analyses describing selective pressures that regulate alternative splicing have raised at least two possible evolutionary models to describe its appearance in eukaryotic organisms. The first model suggests that alternative splicing may have resulted from mutations in the DNA sequence that produces weak splice sites. This would provide an opportunity for the splicing machinery to skip internal exons during mRNA processing. This gives the cell the potential to produce a new transcript with, perhaps, new function(s), without compromising the original repertoire of transcripts produced by the gene. Alternative exons have been shown to have weaker splice sites than constitutively spliced exons, which allows for sub-optimal recognition of exons by the splicing machinery and leads to alternative splicing (Carmel *et al.,* 2004, Sorek *et al.,* 2004).

The second model of alternative splicing suggests that *trans* acting mechanims may also promote alternative splicing. Here, splicing regulatory factors may apply selective pressures on constitutive exons to become alternative exons (Ast, 2004).

For example, the binding of Serine Arginine (SR) proteins in proximity to a constitutively spliced exon weakens the selection of that exon leading to alternative splicing. This releases the selective pressure from the splice sites, resulting in mutations that weaken those splice sites.

1.6.4 Aberrant mRNA splicing and the role of transcript variation in disease

To date, little attention has focused on the error rate of the splicing process and its pathogenic potential. Spliceosomal errors have been proposed to produce transcript variants with little biological relevance, and given the intricate, highly complex process of mRNA splicing it would be anticipated that mRNA splicing will not always proceed with absolute accuracy (Venables, 2004). Incorrect mRNA splicing may result from the spliceosomal machinery skipping constitutive splice sites, or pseudo-splice sites being used in preference to the correct sites.

It is often difficult to differentiate between functional and non-functional transcripts variants. Kan and colleagues hypothesised that transcripts generated by spurious mRNA splicing events will be present at a lower frequency than *bona fide* transcripts (Kan *et al.,* 2002). They applied stringent filters on EST sequences to perform statistical analysis of their frequencies of occurrence and predicted that only 17-28% of genes generated functional transcript variants. This analysis has been performed somewhat prematurely, and would be more informative if it were completed when the sequencing of EST libraries representing many tissues and disease states was exhausted. More recent analysis has found that between 73% and 78% of alternatively spliced exons neither changed the open reading frame nor introduced a premature termination codon (Thanaraj and Stamm, 2003; Sorek *et al.,* 2004).

Erroneous mRNA splicing is not only caused by the unregulated actions of the spliceosome. Mutations within splice sites, in splicing regulatory elements or in proteins that participate in mRNA splicing have been implicated in the production of aberrant mRNAs with deleterious functional alterations. Disease causing splice variants have been implicated in a variety of human conditions, such as cancer, Alzheimer's disease (Scheper *et al.,* 2004; Farris *et al.,* 2005), Parkinson's disease (Ferrier-Cana *et al.,* 2005), ataxia telangiectasia (Pagani *et al.,* 2002), and cystic fibrosis (Cuppens and Cassiman, 2004). Over 15% of human genetic diseases are

current thought to be caused by errors in mRNA splicing (Krawczak *et al.,* 1992). This figure is likely to be an underestimate, as the survey was completed in excess of 13 years ago, and was only completed on genes containing AG-GT introns.

Perhaps the most widely studied disease caused by alternative splicing is spinal muscular atrophy. Spinal muscular atrophy (SMA) is one of the most common autosomal recessive disorders and is caused by the absence of, or mutations in the gene, survival motor neuron 1 (*SMN1)*. This gene has a closely related homologue, survival motor neuron 2 (*SMN2*) which acts as a modifying gene and that can compensate for the loss of *SMN1*. The two genes undergo alternative splicing, with *SMN1* producing an abundance of full-length mRNA transcripts, whereas *SMN2* predominantly produces exon 7-deleted transcripts. The exclusion of exon 7 from *SMN2* is caused by a critical C-to-T substitution at position 6 of exon 7 in *SMN2* (C6U transition in mRNA) which introduces a PTC and protein is therefore unable to compensate for the loss of *SMN1*. It has been proposed that this substitution promotes that gain of a silencing element associated with *hnRNP A1* (Kashima and Manley 2003). The incorporation of exon 7 in *SMN2* can be restored using oligoribonucleotides that are complementary to exon 7 and contain exonic splicing enhancer motifs to provide trans-acting enhancers (Skrodis *et al.,* 2003).

1.6.5 Tools to study isoform function

The function of an individual protein can be determined by knocking out or inactivating individual genes and then subsequently assessing the phenotype of the mutated organism. In general, these methodologies do not consider transcript variation and therefore do not specifically down-regulate the expression of individual transcripts. Several approaches can be employed to analyse the phenotypic effects of individual transcripts. Individual variants can be introduced into a "clean" genetic background that does not contain the gene of interest. This could be either the same species from which both copies of the entire gene have been deleted or in a distantly related species that does not contain the gene of interest. Alternatively transcript specific knockouts or knock-downs can also be made. These have been generated in the mouse for several genes including Arginase AI, *Arg1* (Cederbaum *et al.,* 2004) and myosin light chain kinase, *Mlck* (Tinsley *et al.,* 2004).

RNA interference (RNAi) is a post-transcriptional gene-silencing process induced in diverse organisms by double-stranded RNAs (dsRNAs) homologous in sequence to the silenced genes (Fire *et al.,* 1998). In mammalian cells, long dsRNAs (>30 bp) have been used to activate a global, sequence-nonspecific response resulting in the blockage of protein synthesis and mRNA degradation (Bass, 2001). Small dsRNAs, between 21–23 nucleotides (nt) in length, can bypass the sequence-independent response of mammalian cells and induce transcript-specific degradation of target mRNA (Caplen *et al,.* 2001; Elbashir *et al.,* 2001). These small dsRNAs (or small interfering RNAs, siRNAs) may act as 'guides' within a nuclease complex, the RNA-induced silencing complex (RISC), to direct cleavage and degradation of target mRNA (Hutvagner and Zamore, 2002). Target recognition is a highly sequence-specific process mediated by the siRNA complementary to the target mRNA (Bass, 2000). Gene silencing by siRNA is commonly carried out by transient transfection of cells with synthetic siRNAs or by using expression vectors to produce cells that transiently or stably express siRNAs or short hairpin RNAs. This technique has been used successfully in *HeLa* cells to confirm the isoform specific functions of protein phosphastase, *PP1* (Okada *et al.,* 2004).

Inferences about isoform function can also be made using computational algorithms. As protein function cannot be predicted from sequence alone, predictive programmes rely upon the presence of patterns and motifs in a protein sequence to infer function. Here, sequence characteristics shared in functional domains are identified using multiple sequence alignments from which patterns and profiles for domains can be deduced. Patterns are solely reliant upon sequence identity to a defined motif while profiles are composed of position specific amino-acid weights and gap costs. Databases such as PROSITE use these patterns and profiles to infer potential domain structures within an amino-acid sequence (Falquet *et al.,* 2002).

Gene specific assays can also be used to assess the functional implications of alternative splicing. These include apoptotic, phosphorylation or intracellular localisation assays. For example, alternative splicing of the gene Uracil DNA glycosylase, *UNG* produces two distinct isoforms, *UNG1* and *UNG2,* which are targeted to different cellular compartments (Otterlei *et al.,* 1998). The isoforms differ in their N-terminal domain sequences; *UNG1* has a mitochondrial localisation signal while *UNG2* has a nuclear localisation signal. This observation was

confirmed by tagging both *UNG1* and *UNG2* with the green fluorescent protein and monitoring their subcellular locality in *HeLa* cells (Otterlei *et al.,* 1998).

1.6.6 Nonsense mediated decay

Approximately one-third of genetic disorders result from nonsense or frameshift mutations that truncate the full-length protein structure (Xing and Lee, 2004). These proteins may not be able to fulfil their intended biological function and because of this their transcripts are often targeted for rapid degradation by post-transcriptional surveillance pathways. Recognition of a PTC is essential for triggering the rapid removal of such mRNAs and mRNA surveillance pathways represent a nexus between the cell's machinery for mRNA turnover and translational fidelity (Ruiz-Echevarria *et al.,* 1998). One of the best characterised quality control pathways is nonsense mediated decay (NMD).

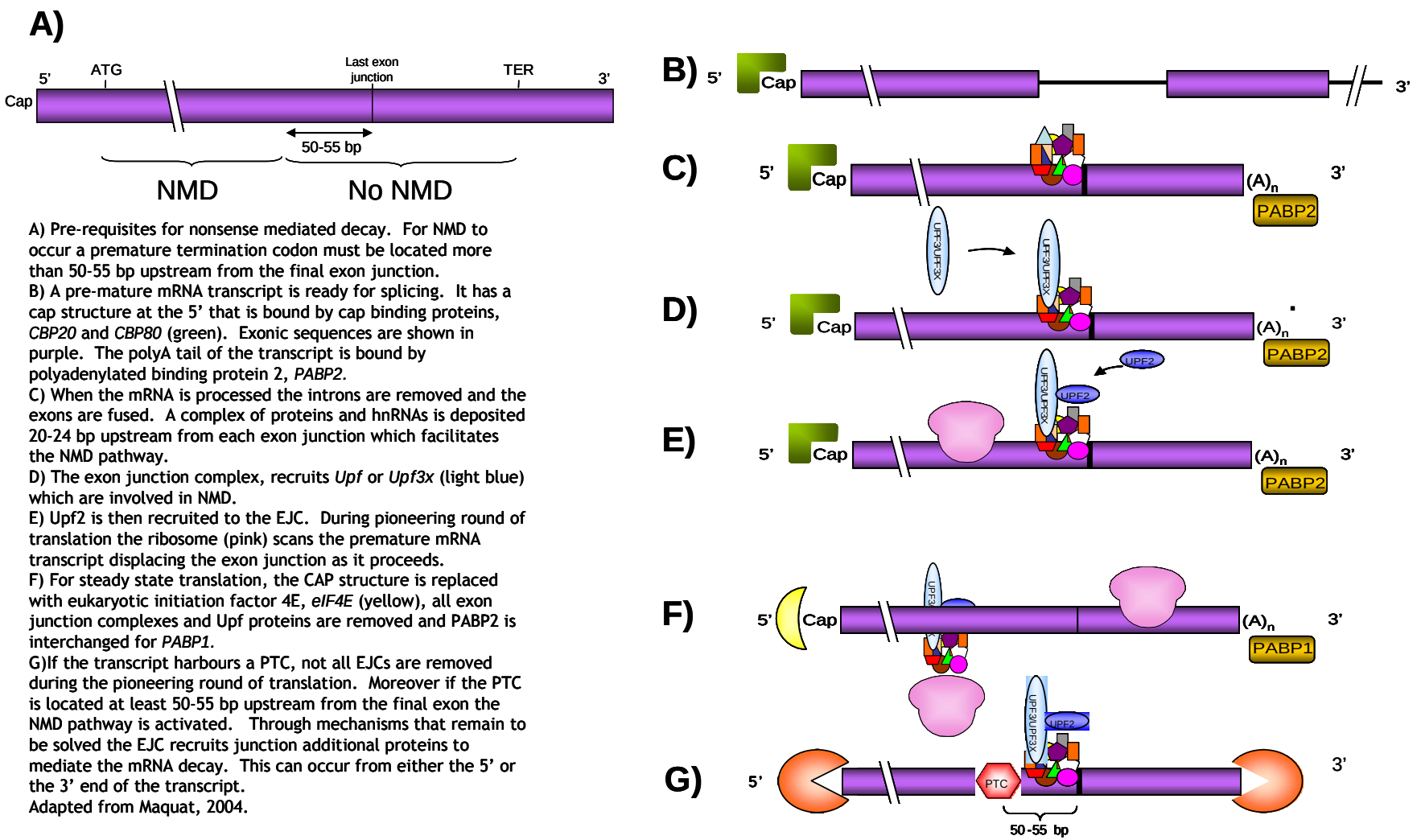
On account of NMD, PTC harbouring transcripts do not generally participate in the synthesis of truncated proteins, which could have a dominant negative effect on the organism. In this manner, NMD is a surveillance mechanism that recognises and degrades aberrant transcripts resulting from erroneous mRNA processing or rare mutations. NMD also contributes to the fine tuning of normal gene expression by degrading specific mRNAs that have naturally occurring premature stop codons. (Jacobson and Peltz, 1999; Wagner and Lykke-Andersen, 2002). For example, upstream open reading frames can control the translation of the downstream main open reading frames using the NMD pathway (Ruiz-Echevarria and Peltz, 2000). The gene splicing factor, arginine/serine-rich 2, *SFRS2* has been shown to auto-regulate its expression through regulated, unproductive alternative splicing which produces transcripts that are targeted for NMD (Sureau *et al.,* 2001).

The biological importance of the NMD pathway is confirmed by its evolutionary conservation. Sequence homology for a defined set of proteins involved in the NMD pathway has been confirmed in a diverse range of eukaryotes such as yeast, *C. elegans,* and humans. Seven different proteins involved in NMD have been identified in *C. elegans.* Orthologues for three of these genes have been confirmed in *S. cerevisiae,* and homology searches have also identified potential orthologues in the mouse, rat and human.

The degradation of transcripts containing PTCs is less well understood in mammalian cells than in yeast. One of the main questions is how cells discriminate between normal termination codons and PTCs. In general, transcripts that harbour a premature termination codon at least 50 bp upstream from the ultimate exon junction will be targeted for rapid degradation by the NMD pathway (Maquat, 2004). This process is discussed below and is illustrated in Figure 1.3.

One of the essential components of the NMD pathway is a ribonuclear protein complex, the exon junction complex (EJC), which is a remnant from the splicing process. In mature mRNAs, the EJC is located approximately 20-24 bp upstream from all exon junctions (Ishigaki *et al.,* 2001; Le Hir *et al.,* 2001). In NMD additional proteins are also recruited to the EJC: first are the upstream processing factors 3/3x, *Upf3/3x*; second the perinuclear protein upstream processing factor 2 (*Upf2)* to the EJC.

Steady state translation requires the substitution of the CAP structure with eukaryotic initiation factor 4E (*EIF4E*), and polyadenylation binding protein 2, (*PABP2*) with polyadenylation binding protein 1 (*PABP1*) and the removal of all EJCs from the mRNA transcript. EJCs are removed during the pioneering round of translation by the translating ribosome. The ribosome scans the mRNA until a stop codon is reached. Once reached, translation is terminated and the ribosome dissociates from the mRNA. In wild-type mRNAs all EJCs will be removed during the pioneering round of translation. However, if the transcript contains a PTC, some EJCs remain bound to be mRNA. These trigger the NMD pathway through as yet unknown mechanisms. Degradation of the mRNA transcript can then occur from either the 5' or the 3' end of the transcript. Single-exon genes will not contain EJCs, at any point and so are not targets for NMD. Presumably, an alternative mechanism exists to target PTC containing transcripts in such cases.



A) Pre-requisites for nonsense mediated decay. For NMD to occur a premature termination codon must be located more than 50-55 bp upstream from the final exon junction.
B) A pre-mature mRNA transcript is ready for splicing. It has a cap structure at the 5' that is bound by cap binding proteins, *CBP20* and *CBP80* (green). Exonic sequences are shown in purple. The polyA tail of the transcript is bound by polyadenylated binding protein 2, *PABP2*.
C) When the mRNA is processed the introns are removed and the exons are fused. A complex of proteins and hnRNAs is deposited 20-24 bp upstream from each exon junction which facilitates the NMD pathway.
D) The exon junction complex, recruits *Upf* or *Upf3x* (light blue) which are involved in NMD.
E) Upf2 is then recruited to the EJC. During pioneering round of translation the ribosome (pink) scans the premature mRNA transcript displacing the exon junction as it proceeds.
F) For steady state translation, the CAP structure is replaced with eukaryotic initiation factor 4E, *eIF4E* (yellow), all exon junction complexes and Upf proteins are removed and PABP2 is interchanged for *PABP1*.
G)If the transcript harbours a PTC, not all EJCs are removed during the pioneering round of translation. Moreover if the PTC is located at least 50-55 bp upstream from the final exon the NMD pathway is activated. Through mechanisms that remain to be solved the EJC recruits junction additional proteins to mediate the mRNA decay. This can occur from either the 5' or the 3' end of the transcript.
Adapted from Maquat, 2004.

Figure 1.3 Overview of nonsense mediated decay

1.7 Aims of this thesis

When this study began, Xp11.22-p11.3 was comprehensively covered with finished genomic sequence and presented an ideal region for sequence-based gene identification and annotation approaches. Comprising approximately 7.3 Mb of DNA, the region spans light, dark and intermediate Giemsa-staining bands. As such, it was expected to display a heterogeneous gene density and repeat content. The first aim of this thesis was to annotate the genome sequence of Xp11.22-p11.3 in order to obtain a greater understanding its gene content.

The annotation of Xp11.22-p.11.3 described 101 gene structures, 77 of which were known genes. While annotating Xp11.22-p11.3 it became apparent that a substantial number of the genes had more than one transcript structure. EST and mRNA data suggested that approximately 70% of the genes within this region were alternatively spliced, which was higher than the frequency of alternative splicing in the genome predicted at that time (between 25-59% Modrek and Lee, 2002). The motivation for the rest of the work described in this thesis stemmed from the high frequency of transcript variation observed within Xp11.22-p11.3.

Existing cDNA and EST sequences are commonly used to estimate the frequency and type of alternative splicing events in the human genome. Consequently, these analyses are limited by the amount and type of sequence information available in public databases. cDNA and EST libraries have not been sequenced exhaustively from all tissues, and therefore the inferred frequency of alternative splicing is likely to be extremely conservative. In the second phase of work, a discovery project was undertaken to enhance our current understanding of the prevalence and variety of transcript variation. Here, a targeted PCR, cloning and sequencing approach was carried out on a panel of cDNAs from 29 different human tissues. To ensure that a detailed analysis of possible alternative transcripts was obtained, it was decided to focus on a subset of 18 protein-coding genes from the original transcript map.

This approach identified 61 gene fragments with novel splicing patterns. It confirmed that the use of random EST and cDNA data alone severely underestimates transcript variation levels. However, it was not confirmed if these variants were functional or if they were the result of imprecise mRNA splicing. If

the splicing events were functionally relevant, how did they alter the function of the cognate protein? Or, if the variants were products of mis-splicing events were they degraded rapidly in order to prevent any dominant negative effects? Were the alternative transcripts present in a biologically relevant concentration?

In order to begin to address some of these questions, a detailed study of one gene, polyglutamine binding protein, *PQBP1*, was carried out. By determining splice site sequence characteristics, relative mRNA expression levels, sub-cellular localisation of isoforms and mRNA transcript stability, it was hoped that any potential functional differences among the *PQBP1* transcript variants would be identified.

Chapter 2

Materials and Methods

Materials

2.1 Chemical reagents

All common chemicals were purchased from Sigma Chemical Co., BDH Chemical Ltd., and Difco Laboratories unless specified below or in the text.

Bio-Rad Laboratories	β-mercaptoethanol
Gibco BRL Life Technologies	Foetal Bovine Serum
	Phosphate buffered saline, PBS (pH7.2)
	ultraPURE™ agarose
	Ham's F12 media
Novagen	GeneJuice
	KOD polymerase
Novagiochem	X-gal
	(5-bromo-4-chloro-3-indolyl-β-D-galactoside)
Fluka	Formamide
Sigma	Anisomycin
	Cycloheximide
	Doxycycline
	Puromycin
	Streptomycin
	L-glutamine
	Dulbecco's Modified Eagle Medium (DMEM)
	4', 6-diamidino-2-phenylindole (DAPI) stain
Stratagene	Perfect Match™
	Taq Extender™

2.2 Enzymes and commercially prepared kits

All restriction endonucleases were purchased from New England Biolabs.

Ambion	DNA-free DNase treatment kit
Amersham Biosciences	Megaprime DNA labelling systems
	Sephadex G-50 Nick Columns
	Redivue™[α-^{32}P]-dCTP (AA 005) aqueous solution (370 MB/ml, 10 mCi.ml)
	2'-deoxynucleoside 5' triphosphates (dATP, dTTP, dGTP, dCTP)
BD Biosciences	Advantage-2 PCR Enzyme Mix

Bio-Rad	Bradford Protein Quantification
	Acrylamide Gels
Clontech	Luciferase Reporter Assay
	Human MTN Blot I
	Human MTN Blot II
Fuji	RX 1100 medical X-ray film
Invitrogen	Superscript II cDNA synthesis kit
	DNase I
New England Biolabs	T4 DNA ligase (1 U/μl)
PE Applied Biosystems	Amplitaq™
	AmplitaqFS
	SYBR Green Master Mix
Qiagen	Genomic DNA and DNA gel purification
	Midi- and Maxi- Prep Kits
	RNeasy RNA purification
Sigma Chemical Company	Ribonuclease A
	Deoxyribonuclease I
	DNA polymerase I (10 U/μl)

2.3 Hybridisation membranes, and X-ray and photographic film

Amersham Biosciences	Hybond-C™ Nylon (20 cm x 1 m)
	(used for western blotting)
Fuji	RX 100 Medical X-ray film
Polaroid	Polaroid 667 Professional film

2.4 Solutions and buffers

Solutions used in this thesis are listed below, alphabetically within each section. Final concentrations of reagents are given for most solutions. Amounts and/or volumes used in preparing solutions are given in some cases. Unless otherwise specified, solutions were made in nanopure water.

2.4.1 *Buffers*

10x DNase I Buffer
200 mM Tris-HCl (pH 8.4)
20 mM $MgCl_2$
500 mM KCl

10x Ligase Buffer
500 mM Tris-HCl (pH 7.5)
100 mM Dithiothreitol
100 mM $MgCl_2$

10x PCR buffer (Advantage)
400 mM Tricine – KOH (pH 8.7)
150 mM KOAc
35 mM $Mg(OAc)_2$
37.5 μg/ml BSA
0.05% Tween-20
0.05% Nonidet-P40

10x PCR buffer I
100 mM Tris-Hcl (pH 8.3)
500 mM KCl
15 mM $MgCl_2$

10x PBS pH 7.4
10.6 mM KH_2PO_4
1.5 M NaCl
30 mM Na_2PO_4-$7H_20$
5% v/v β-mercaptoethanol

PBS-T
0.1% v/v Tween 2
1 x PBS

10x TBE
890 mM Tris Base
890 mM Borate
20 mM EDTA (pH 8.0)

1x TE
10 mM Tris-HCl (pH 7.4)
1 mM EDTA

1x $T_{0.1}E$
10 mM Tris-HCl (pH 8.0)
0.1 mM EDTA

TFB I
30 mM KOAc
100 mM $RbCl_2$
10 mM $CaCl_2$
50 mM MnCl
15% v/v Glycerol
pH 5.8

TFB II
10 mM MOPS
75 mM $CaCl_2$
100 mM $RbCl_2$
15% v/v Glycerol
pH 6.5

2.4.2 *Northern blotting solutions*

20 x SSC
3 M NaCl
300 mM Trisodium Citrate

Hybridisation buffer
6 x SSC
1% w/v N-lauroyl-sarcosine
10 x Denhardt's
50 mM Tris-HCl (pH 7.4)
10% w/v Dextran sulphate

100 x Denhardt's Solution
20 mg/ml Ficoll 400-DL
20 mg/ml polyvinylpyrolidine 40
20 mg/μl BSA (pentax fraction V)

2.4.3 *Electrophoresis Solutions and Western Blotting Solution*

Blocking Solution
10% w/v Milk Powder
0.1% v/v Tween 20
PBS

6x Glycerol Dye
30% v/v Glycerol
0.1% w/v Bromophenol Blue
0.1% w/v Xylene Cyanol
5 mM EDTA (pH7.5)

1x Protein Sample Buffer
2% w/v SDS
10% v/v Glycerol
60 mM Tris pH6.8
0.01% w/v Bromophenol Blue

10x Running Buffer
0.25 M Tris
1.92 M Glycine
1% w/v SDS

1x Transfer Buffer
0.025 M Tris
0.192 M Glycine
0.1% w/v SDS
25% v/v Ethanol

2.4.4 *Immunofluorescence Solutions*

Blocking Solution
0.2% w/v Gelatine
0.05% w/v Saponin
PBS

Washing Solution
0.05% w/v Saponin
PBS

Quenching Solution
50 mM NH_4Cl

2.4.5 *Media*

All media were prepared in nanopure water and either autoclaved or filter-sterilised prior to use. When used for bacterial growth, 15 mg/ml bacto-agar was added to the appropriate media. Where appropriate Ampicillin (dissolved in 1 M sodium bicarbonate, stored at -20°C) was added to media at a final concentration of 75 μg/ml.

LB
10 mg/ml Bacto-tryptone
5 mg/ml Yeast extract
10 mg/ml NaCl
pH 7.4

2 X TY
15 mg/ml Bacto-tryptone
20 mg/ml Bacto-peptone
2% w/v dextrose
pH 5.8

LB plates
LB media
15 g/l agar
75 ug/ml Amplicillin

X-gal plates
As for LB plates plus
100 μg/ml Xgal
200 μg/ml IPTG

2.4.6 *General DNA preparation solutions*

GTE
50 mM Glucose
1 mM EDTA
25 mM Tris-HCl (pH 8.0)

3 M K^+/5 M Ac^-
60 ml 5M potassium acetate (pH 4.8)
11.5 ml glacial acetic acid
28.5 ml H_20

2.5 Size Markers

1 kb ladder (1 mg/ml) (Gibco BRL Life Technologies)

Contains 1 to 12 repeats of a 1,018 bp fragment and vector fragments from 75 bp to 1,636 bp to produce the following sized fragments in bp: 75, 142, 154, 200, 220, 298, 344, 394, 516/506, 1,018, 2,036, 3,054, 4,072, 5,090, 6,108, 7,125, 8,144, 9,162, 10,180, 11,198, 12,216.

100 bp ladder (Invitrogen Life Technologies)

The 100 bp ladder consists of 15 blunt ended fragments between 100 and 1500 bp in multiples of 100 with an additional fragment of 2072 bp.

SeeBlue Protein Standard (Invitrogen Life Technologies)

Consists of 10 pre-stained protein bands in the range of 4 – 250 kDa. The proteins and their approximate molecular weights (kDa) are: Myosin – 250, Phosphorylase – 148, BSA – 98, Glutamic Dehydrogenase – 64, Alcohol dehydrogenase – 50, carbonic Anhydrase – 36, Myoglobin Red – 22, Lysozyme – 16, Aprotinin – 6, Insulin, B chain – 4.

2.6 *E.coli* strains

The bacterial strains used in this study are listed in Table 2.1.

Table 2.1 Strains of *E. coli* used in this study

Strain	Source	Genotype
JM109	Clontech	*e14-(McrA-) recA1 endA1 gyrA96 thi-1 hsdR17*(rK- mK+) *supE44 relA1 ⊿(lac-proAB)* [F' *traD36 proAB lacI^{q}Z⊿M15]*
DH5α	Invitrogen	FΦ80*lacZ*ΔM15 Δ(*lac*ZYA-*arg*F) U169 *rec*A1 *end*A1 *hsd*R17(r_k^-, m_k^+) *pho*A *sup*E44 *thi*-1 *gyr*A96 *rel*A1 *ton*A
DH10b	Invitrogen	F- *mcr*A Δ(*mrr-hsd*RMS-*mcr*BC) φ80*lacZ*ΔM15 Δ*lac*X74 *deo*R *rec*A1 *end*A1 *ara*D139 Δ(*ara, leu*)7697 *gal*U *gal*K λ - *rps*L *nup*G *ton*A

2.7 Mammalian Cell Lines

The mammalian cell lines used in this study, together with their associated experiments and growth conditions are listed in Table 2.2.

Table 2.2 Mammalian cell lines used in this study

Cell Line	Cell Line Description	Use	Media *	Source
Cos-7	African Rhesus Monkey (kidney)	Intracellular Localisation	DMEM (Sigma)	Dr J Collins
CHO-K1	Chinese Hamster Ovary	Intracellular Localisation	Ham's F12 (Gibco)	Dr P Couttet
CHO-AA8-Luc-Off	Chinese Hamster Ovary transfected with tetracycline regulatory element	Transcript Stability	DMEM (Gibco)**	Dr P Couttet
Hek293FT	Human embryo kidney	Translation Inhibition	DMEM (Sigma)	Dr J Collins

* Media was supplemented with 10% v/v foetal bovine serum (Gibco, BRL), 100 U/ml Penicillin (Sigma), 100 μg/ml Streptomycin (Sigma), and 2 mM L-glutamine (Sigma).
** Media was supplemented with 10% v/v tetracycline approved foetal bovine serum (Clontech), 100 U/ml Penicillin (Sigma), 100 μg/ml Streptomycin (Sigma), and 2 mM L-glutamine (Sigma).

2.8 RNA samples

2.8.1 *Sources of human total RNA*

Total RNA was obtained from Ambion, Clontech and Stratagene. Tissue origins are given overleaf in Table 2.3. All commercial samples were extracted from single tissues and are not pools of samples.

Table 2.3 Total RNA samples used in this study

Supplier	Tissue panel number	Tissue	Supplier	Tissue panel number	Tissue
Clontech	1	Adrenal gland	Ambion	26	Cervix
Clontech	2	Bone marrow	Ambion	27	Colon
Clontech	3	Brain (cerebellum)	Ambion	28	Heart
Clontech	4	Brain (whole)	Ambion	29	Kidney
Clontech	5	Foetal brain	Ambion	30	Liver
Clontech	6	Foetal liver	Ambion	31	Lung
Clontech	7	Heart	Ambion	32	Ovary
Clontech	8	Kidney	Ambion	33	Pancreas
Clontech	9	Liver	Ambion	34	Placenta
Clontech	10	Lung	Ambion	35	Prostate
Clontech	11	Placenta	Ambion	36	Skeletal muscle
Clontech	12	Prostate	Ambion	37	Small intestine
Clontech	13	Salivary gland	Ambion	38	Spleen
Clontech	14	Skeletal muscle	Ambion	39	Stomach
Clontech	15	Spleen	Ambion	40	Testis
Clontech	16	Testis	Clontech	41	Thymus
Clontech	17	Thymus	Clontech	42	Bone marrow
Clontech	18	Thyroid gland	Stratagene	43	Foetal stomach
Clontech	19	Trachea	Stratagene	44	Foetal lung
Clontech	20	Uterus	Stratagene	45	Foetal heart
Clontech	21	Foetal brain	Stratagene	46	Foetal kidney
Clontech	22	Foetal liver	Stratagene	47	Foetal skeletal muscle
Ambion	23	Adrenal gland	Stratagene	48	Foetal colon
Ambion	24	Bladder	Clontech	49	Uterus
Ambion	25	Brain	Clontech	50	Thymus

2.8.2 *Additional sources of total RNA*

Total RNA from *S. pombe* was a gift from Dr. J. Bähler (WTSI).

2.9 cDNA libraries

Nineteen different cDNA libraries were used in the study (see Table 2.4). cDNA libraries were imported and maintained by Jacqueline Bye. Each library contains 500,000 cDNA clones, divided into 25 pools of 20,000 clones. Five pools were combined to form a superpool that contained 100,000 clones. Prior to their use in PCR, each superpool was diluted 1:100 in $T_{0.1}E$.

Table 2.4 cDNA libraries used in this study

cDNA library code	cDNA library description	Vector	Source/ Reference
1. U	Monocyte NOT activated-from a patient with promonocytic leukaemia (U937+)	pCDM8	Simmons (1993)
2. H*	Placental, full term normal pregnancy (H9)	pH3M	Simmons (1993)
3. P	Adult brain	pCDNA1	Pfizer
4. DAU	B lymphoma (Daudi)	pH3M	Simmons (1993)
5. FB	Fetal brain	pCDNA1	Invitrogen
6. FL	Fetal liver	pcDNA1	Invitrogen
7. HL	Peripheral blood (HL60)	pCDNA1	Invitrogen
8. SK	Neuroblastoma cells	pCDNA1	Invitrogen
9. T	Testis	pCDM8	Clontech
10. FLU	Fetal lung	pCDNA1	Invitrogen
11. AL	Adult lung	pCDNA1	Clontech
12. UACT*	(Monocyte PMA activated – from a patient with promonocytic leukaemia) (U937act)	pCDM8	Simmons (1993)
13. YT*	HTLV-1+ve adult leukaemia T cell	pH3M	Simmons (1993)
14. NK*	Natural killer cell	pH3M	Simmons (1993)
15. HPB*	T cell from a patient with acute lymphocytic leukaemia (HPBALL)	pH3M	Simmons (1993)
16. BM*	Bone Marrow	pH3M	Simmons (1993)
17. DX3*	Melanoma	pH3M	Simmons (1993)
18. AH	Adult Heart	pcDNA3-Uni	Invitrogen
19. SI **	Small Intestine	pcDNA3	Stammers

* Generously provided by Dr Simmons, Oxford (Simmons *et al*., 1993)
** Generously provided by Dr Stammers (Sanger Institute)

2.10 Primer sequences

Appendices I to III list the STSs (sequence tag sites) used in this thesis and give the sequence of each primer. Where appropriate, the clones, or genes from which the STSs were derived are also listed.

Primers were synthesised in house by Dave Fraser or externally by Sigma.

2.11 Key World Wide Web addresses

Table 2.5 Key world wide web addresses used in this study

Website	Address
Baylor College of Medicine Search Launcher	http://searchlauncher.bcm.tmc.edu/
CCDS	http://www.ncbi.nlm.nih.gov/CCDS/
Dotter	http://www.cgr.ki.se/cgr/groups/sonhammer/Dotter.html
EBI	http://www.ebi.ac.uk/
EBI- ClustalW	http://www.ebi.ac.uk/clustalw/index.html
EMBOSS	http://www.hgmp.mrc.ac.uk/Software/EMBOSS/
Ensembl	http://www.ensembl.org
Eponine	http://servlet.sanger.ac.uk:8080/eponine/
First Exon Finder	http://genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=42528583&c=chrX&g=firstEF
Gap4	http://staden.sourceforge.ent/overview.html
Gene Expression Atlas	http://expression.gnf.org/cgi-bin/index.cgi#Q
NCBI	http://www.ncbi.nlm.nih.gov/
NCBI – BLAST server	http://www.ncbi.nlm.nih.gov/BLAST/
NCBI – Entrez	http://www.ncbi.nih.gov/Entrez/
NCBI – Locus Link	http://www.ncbi.nlm.nih.gov/projects/LocusLink/
NCBI – Spidey	http://www.ncbi.nlm.nih.gov/spidey/
NCBI – UniGene	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene
OMIM	http://www3.ncbi.nlm.nih.gov/Omim/
Primer3	http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi
Prosite	http://au.expasy.org/prosite/
PSORT	http://psort.nibb.ac.jp/
RepeatMasker	http://ftp.genome.washington.edu/RM/webrepeatmaskerhelp.html
SpliceSiteFinder	http://www.genet.sickkids.on.ca/~ali/splicesitefinder.html
The HGMP Resource Centre	http://www.hgmp.mrc.ac.uk/
The Wellcome Trust Sanger Institute	http://www.sanger.ac.uk/
UCSC Genome Browser	http://genome.ucsc.edu/
Vega	http://vega.sanger.ac.uk/
zPicture	http://zpicture.dcode.org/

Methods

2.12 Mammalian cell culture

The mammalian cells used in this study are listed in Table 2.2.

2.12.1 *Growing and harvesting cells*

All cell lines were adherent and were grown at 37°C, 5% CO_2 in either 75 cm^2 or 150 cm^2 flasks, 8-well cell culture slides, or 6-well dishes.

When a cell density of greater than 80% confluency was reached, cells were sub-cultured. Briefly, the media was aspirated and the cells were washed twice in PBS. To detach the cells from the flask/wells the cells were then incubated with approximately 0.07 volumes of trypsin (Sigma) at 37°C for 2 to 5 minutes. Approximately 0.7 volumes of pre-warmed media were then added to inactivate the typsin. Cell clumps were disrupted with gentle pipetting. From here, the cells were distributed into fresh flasks for sub-culturing or were counted and aliquoted for future experimental analysis.

Cells were counted with a haemocytometer (Sigma) with a 0.1 mm sample depth and light microscope (Olympus).

For frozen storage, cells were harvested at 200 x g following trypsinisation. Cell pellets were resuspended at a density of approximately 3 x 10^6 cells per ml in 10% DMSO in foetal bovine serum in polypropylene vials. Vials were frozen at 1°C/per minute for 12 to 24 hours before permanent storage in the gas phase of a liquid nitrogen vessel (at -180°C). Cells were recovered by rapid thawing at 37°C. They were washed and then mixed directly with 10 to 15 ml of complete medium.

2.12.2 *Transfection*

On the day prior to transfection, cells were trypsinized and plated to the required density with the media, FCS and antibiotics. The next day, the appropriate media was mixed with GeneJuice and incubated at room temperature for between 5 and 15 minutes (volumes and amount of DNA used are listed in Table 2.6). DNA was added to the GeneJuice/media mix and incubated for a further 5 to 15 minutes at room temperature. This was then added to the cells in a dropwise manner, and was gently swirled to mix.

Cells were then grown at 37°C, 5% $C0_2$ with humidity for a further 24 to 48 hours.

Table 2.6 Amounts of DNA and reagents used in mammalian cell transfections.

Dish	Media (µl)	GeneJuice (µl)	DNA (µg)
8-well culture slide	20	0.75	0.25
24-well cell culture dish	20	0.75	0.25
6-well cell culture dish	100	4	2

2.13 RNA Manipulation

All reagents for RNA work were prepared with Diethylene Pyrocarbonate (DEPC) treated water. Bench surfaces, tubes and labware were cleaned before use with RNaseZap (Ambion).

2.13.1 *Preparation of RNA from cellular extracts*

Total RNA was prepared from mammalian cell pellets using the RNeasy RNA extraction kit (Qiagen) in accordance with the manufacturer's protocol. RNA was eluted with 50 µl of DEPC-treated water.

Following elution, the integrity of the RNA was confirmed by visualisation on 1% agarose gels using ethidium bromide staining. The concentration of RNA was determined by spectroscopy (GENE QUANT) where an absorbance of 1 at 260 nm equates to a concentration of 40 ng/µl. A_{260}:A_{280} ratios were also calculated for each sample. Samples with ratios smaller than 1.7 or greater than 1.9 were discarded.

2.13.2 *DNase treatment of total RNA*

Residual plasmid or genomic DNA was removed from RNA samples using DNase I (Invitrogen). Briefly, samples (up to 2 µg) of total RNA were equilibrated with 10 x DNase buffer to which 1 U of DNase was added. Samples were then incubated for 15 minutes at room temperature and the enzyme was denatured by incubation at 65ºC for 10 minutes.

2.13.3 *cDNA synthesis*

cDNA was synthesized from (DNase treated) total RNA (0.5 to 5 µg) using Superscript II (Invitrogen). Where appropriate, and as outlined in the text, either random hexamers (50 ng) or Oligo$(dT)_{12-18}$ primers (500 ng) were used to prime the cDNA synthesis. The synthesis was completed in accordance with the

manufacturer's instructions, with an incubation temperature of 25ºC (random hexamers) or 42ºC (Oligo(dT)$_{12-18}$).
The resulting cDNA was diluted to 400 μg/μl and was stored at -20ºC.

2.13.4 *Northern blotting*

2.13.4.1 Probe Preparation

Approximately 25 ng of purified, nested PCR product were randomly labelled with [α-^{32}P] dATP using the Megaprime DNA labelling Kit (Amersham Biosciences) in accordance with manufacturers instructions.

Incorporation of the label was confirmed by spotting 2 μl of radiolabelled probe onto a Polygram 300 PEI/UV Thin Layer Chromatography (TLC) plate in 1M KH_2PO_4 (pH3). TLC plates were exposed to autoradiography film for approximately 1 hour.

Unincorporated label was then removed by elution on a Sephadex G50 column (Pharmacia Biotech). The probe was diluted to 400 μl in $T_{0.1}E$ and applied to the column. Five 400 μl fractions were collected. The amount of label in each fraction was monitored in a scintillation counter, with fraction 2 containing the labelled probe. All probes were denatured at 95 °C for 5 minutes prior to hybridisation.

2.13.4.2 Hybridisation

Northern blots were pre-hybridised in 50 ml of hybridisation buffer for at least 2 hours at 65 °C in Hybaid tubes with gentle rotation. Twenty-five ml of hybridisation buffer were removed and the radiolabelled probe was added. The blots were hybridised at 65 °C overnight with gentle rotation (approximately 16 hours).

Following overnight hybridisation, the hybridisation solution was then discarded and the blots were washed under the following conditions:

2 x SSC	room temperature	2 x 5 minutes
2 x SSC, 1 % sarcosyl	65 °C	2 x 30 minutes
1 x SSC, 1 % sarcosyl	65 °C	1 x 30 minutes
0.5 X SSC, 1 % sarcosyl	65 °C	1 x 20 minutes
0.2 x SSC	room temperature	2 x 5 minutes.

All washes were carried out with gentle shaking.

Blots were washed until the signal from a Geiger counter dropped below ~ 5cps.
Blots were wrapped in Saran Wrap (Dow Chemical Co.) and were then exposed to pre-flashed autoradiography film at –70 °C.

2.13.4.3 Removal of radiolabelled probe from Northern blots.

Blots were stripped of any remaining hybridised probe by incubation in 5 mM Tris-HCl pH7.4 at 75 °C for 1 hour. Removal of the probe was confirmed by exposing the blots to pre-flashed autoradiography film for 1 week.

2.14 DNA manipulation

2.14.1 *Purification of DNA*

Ethanol precipitation

DNA was precipitated from solution by the addition of 0.1 volumes of 3M sodium acetate and either 2 volumes of 100% ethanol, or 0.7 volumes of isopropanol. Samples were centrifuged for 20 minutes at 13,000 x g. The pellet was washed with 1 volume of 70% ethanol, followed by centrifugation for 5 minutes at 13,000 x *g*. DNA pellets were air-dried and resuspended in $T_{0.1}E$.

Direct PCR product purification

PCR products (10 to 50 µl) were purified for subsequent analysis using the PCR purification kit (Qiagen) in accordance with the manufacturer's protocol.

ExoSAP purification of PCR products

PCR reactions were equilibrated with 0.1 volumes of 10 x EXOSAP buffer, 1 U/µl Shrimp Alkaline Phosphatase and 1 U/µl exonuclease I.
Reactions were incubated at 37°C for 30 minutes, followed by 70°C for 15 minutes.

Agarose gel purification

Following visualisation by agarose gel electrophoresis, the appropriately sized DNA fragment was excised from the gel using a clean scalpel. Purification proceeded using the gel purification kit (Qiagen) in accordance with the manufacturer's instructions.

2.14.2 *Alkaline phosphatase treatment of DNA*

The removal of the 5' phosphate group from up to 10 µg of digested DNA was completed using 50 units of Shrimp Alkaline Phosphatase (SAP) in SAP buffer. Samples were incubated at 37°C for 1 hour.

2.14.3 *Restriction digests*

Restriction digests of plasmid DNA (up to 10 µg) were completed using the appropriate buffer, and 20-50 units of enzyme. Samples were incubated at 37°C for 2 hours, and the resulting digest was confirmed by agarose electrophoresis.

2.14.4 *Mini-preps of plasmid DNA*

A single colony was inoculated in up to 10 ml of LB broth containing the appropriate antibiotic and grown overnight at 37°C, shaking at 250-300 rpm. On the following day the cells were pelleted at 1500 x *g*, and resuspended in 200 µl GTE on ice. To this, 400 µl of fresh 0.2M NaOH and 1% SDS were added with gentle inversion of the tube. The tube was left on ice for 5 minutes. Three hundred microlitres of 5M Acetate, 3M K^+ were added. The tube was inverted to mix, and the samples were incubated on ice for at least 10 minutes. The precipitate was pelleted by centrifugation at 13,000 x *g*. This procedure was repeated if the resulting supernatant was still cloudy. The DNA was pelleted by the addition of 0.7 volumes of isopropanol and subsequent centrifugation at 13,000 x *g*.

The DNA was then ethanol precipitated (see section 2.14.1), and washed with 70% ethanol before being resuspended in 30 µl $T_{0.1}E$.

2.14.5 *Midi and maxi preps of plasmid DNA*

A single colony was inoculated into 5 ml LB broth containing the appropriate antibiotic and grown for 8 hours at 37°C, shaking at 20-300 rpm. This inoculation was diluted 1:500 into 100-500 ml LB broth containing the same antibiotic and grown overnight at 37°C, shaking at 200 to 300 rpm.

Purification of plasmid DNA was completed using the appropriate kit (Qiagen Midi- and Maxi Prep Kits) in accordance with the manufacturer's protocol.

2.14.6 *Quantification*

Absorbance Spectroscopy

DNA was quantified by applying the Beer-Lambert equation relating absorbance, and extinction co-efficient to DNA concentration. Absorbance readings were measured at 260 nm and the extinction coefficients used were 50 for dsDNA and 33 for ssDNA. Absorbance readings were taken on either Gene Quant (Biochrom Ltd.) or BioQuant (Eppendorf).

Pico Green Quantification

Double stranded DNA was quantified using PicoGreen® dsDNA Quantitation Reagents (Molecular Probes). The procedure was completed in accordance with the manufacturer's protocol. Lambda DNA (Molecular Probes) was used to construct the standard curve in the range of 0-200 ng/µl from which the concentration of unknown samples was extrapolated. Absorbance readings were performed using a luminescence spectrofluorimeter (BioQuant) measuring at a wavelength of 480 nm for excitation and 520 nm for emission.

2.15 Polymerase Chain Reaction

Primers were designed using Primer3 (section 2.11) or Primer Express software packages (Applied Biosystems). Where possible primers were selected using the following parameters:

- Melting temperature between 57°C and 63°C.
- G/C content between 30-80%
- Length between 17 – 22 bp
- Less than 2°C difference in melting temperature between the two primers

ePCR or BLAST searches were performed to test the specificity of the primer pair.

PCR was performed in either a 96-well micro-titer plate (Costar Thermowell™ M-type plate or 0.2 µl tubes (Falcon) in a PTC-225 (MJ Research) thermocycler. Unless stated otherwise template DNA was amplified in a reaction volume of 15 µl. Reactions contained approximately 1.3 µM of each oligonucleotide primer, 67 mM Tris-HCL (pH 8,8), 16.6 mM $(NH_2)SO_4$, 6.7 mM $MgCl_2$, 0.5 mM of each deoxyribonucleoside triphosphate (dATP, dCTP, dGTP, dTTP), 1.5 U of Amplitaq™ (Cetus Inc.). 10 mM β-mercaptoethanol and 170 µg/ml of BSA (Sigma Chemical Co., A-4628) were also added to the reactions.

Unless specified otherwise, cycling conditions were as follows: an initial denaturing step of 5 minutes at 94 °C, followed by thermal cycling at 94 °C for 30 seconds, [primer-specific annealing temperature] for 30 seconds, and 72 °C for 30 seconds. A final extension step of 5 minutes at 72 °C completed the amplification reaction. PCR products were then separated on 2.5% agarose gels and were visualised by ethidium bromide staining.

2.15.1 *STS pre-screen*

STS pre-screens were performed using 5 ng of the following templates: human genomic DNA, Clone 2D (a hybrid containing only X chromosome DNA), hamster DNA and a negative control. Pre-screens were performed using three different annealing temperatures (55 °C, 60 °C, 65 °C) to determine the cycling parameters that will give a visible and specific DNA product. Buffer and PCR conditions are described above (section 2.15) and section 2.4.1.

2.15.2 *cDNA library screening*

Nineteen different cDNA libraries were subdivided into 25 subpools of 20,000 clones which were then combined to produce 5 superpools of 100,000 clones (J.Bye). The cDNA libraries are listed in Table 2.4.

To screen for expression, aliquots of the superpools of each library were arranged in a micro-titre plate to facilitate subsequent manipulation and gel-loading post PCR with a multichannel pipetting device. Five microlitres of each superpool were used as template in a 15 μl final reaction volume in the primary screens. Buffer and PCR conditions are described in section 2.15 and section 2.4.1.

2.15.3 *Vectorette PCR*

Vectorette PCR on cDNA libraries.

Vectorette PCR was performed on the cDNA library superools (listed in Table 2.4). PCR was performed using 5 μl of the diluted superpools (1:100 dilution in $T_{0.1}E$) as the template in a 15 μl final reactions volume using buffer conditions as described in section 2.15.1) Primer combinations were as follows: universal primer 224 and specific primer A, universal primer 224 and specific primer B for each STS.

The PCR was performed in a DNA thermocycler (Omnigene) using a hot start. Cycling conditions were as follows: an initial denaturation step of 5 minutes at 95 °C, followed by 17 cycles of: 94 °C for 5 seconds, 65 °C for 30 seconds and 72 °C for 3 minutes; followed by cycles of 94 °C for 5 seconds, 60 °C for 30 seconds and 72 °C for 3 minutes. An incubation of 72 °C for 5 minutes completed the reaction. The PCR was paused after 4 minutes of the initial denaturation and 2 μl of Taq premix (containing 0.12 units Amplitaq, 0.12 units Taq Extender and 0.l2 units of Perfect Match, 10 % sucrose + cresol red, and $T_{0.1}E$) were added to each reaction.

Re-amplification of vectorette PCR products.

In reactions where multiple bands or weaker bands were observed, bands were excised and placed in 100 μl of $T_{0.1}E$ for at least 10 hours. Re-amplification of each band was carried out by PCR using 5 μl of $T_{0.1}E$ taken from the 100 μl containing the excised band followed by the addition of PCR reagents as described in section 2.15.1). Cycling conditions were as follows: an initial denaturation step of 5 minutes at 95 °C followed by 25 cycles of 94 °C for 5 seconds, 60 °C for 30 seconds and 72 °C for 3 minutes, followed by 72 °C for 5 minutes.

The vectorette PCR products were separated by electrophoresis in 2.5% agarose gels and were visualised by ethidium bromide staining. Products were gel purified using the Qiagen gel extraction kit prior to sequencing.

2.15.4 *Colony PCR*

Following bacterial transformation (section 2.17.3) individual white colonies were picked using a sterile toothpick and resuspended in 100 µl of sterile water. PCR was performed using 5 µl of the resuspended colony as the template DNA. Different primer combinations were used to either confirm the presence and size of an insert, or to confirm the presence and orientation of the insert. These primer combinations are listed in Appendices II and III.

Reaction products were visualised by agarose gel electrophoresis and staining with ethidium bromide (section 2.16.1).

2.15.5 *RT-PCR*

PCR was performed using up to 5 µl of single stranded cDNA synthesised from total RNA (section 2.13.3) with the final reaction volume ranging between 15-25 µl. Included in the reaction were 0.01 volumes of 50 x Advantage 2 PCR Enzyme System (BD Biosciences) which was used with 0.1 volumes of 10 x Advantage 2 buffer, 1.5 U of Amplitaq™ (Cetus Inc.), 0.4 mM of each deoxyribonucleoside triphosphate (dATP, dCTP, dGTP, dTTP), and 0.5 µM of each oligonucleotide primer.

Unless otherwise noted in the text, 35 cycles of PCR were performed. Cycling conditions were as follows: an initial denaturation step of 5 minutes at 95 °C, followed by 35 cycles of: 94 °C for 30 seconds, 65 °C for 30 seconds and 72 °C for

30 seconds to 3 minutes depending on the expected size of the PCR product. The extension time was increased by 1 min per kb when longer PCR products were expected.

2.15.6 *Quantitative PCR*

Quantitative PCR reactions were performed on an ABI7000, using SYBR-Green Master Mix (Applied Biosystems), ABI PRISM® 96 well optical reaction plates and ABI PRISM™ optical adhesive plate sealers. All reactions were completed in triplicate, with minus RT controls. Each 25 µl PCR reaction contained 0.02 to 0.1 µg cDNA (section 2.13.3), 2 x SYBR-Green Master Mix, and primers diluted to a final concentration of 0.5 µM.

The following cycling parameters were used: 50ºC for 10 minutes, then 95ºC for 10 minutes. This was followed by 40 cycles of 95ºC for 10 seconds and a combined annealing/extension temperature of 60ºC for 2 minutes. During each cycle of the PCR the fluorescence emitted by the binding of SYBR-Green to the dsDNA produced in the reaction was measured. To confirm the specificity of the reactions dissociation curves were constructed for each primer pair at 0.1ºC intervals between the temperatures of 60 ºC and 95ºC.

Analysis of Quantitative PCR

The SYBR-Green fluorescent spectra collected during the PCR were analysed using the Sequence Detection System Software (ABI). Firstly, background threshold levels were set at the number of cycles before any SYBR-Green fluorescence was detected. The detection threshold was set at the point where the increase in SYBR-Green fluorescence became exponential. Assuming specific amplification, the cycle number at which the sample's fluorescence intersected with the detection threshold, was directly proportional to the amount of DNA in the sample, and was expressed at C_T values. Two different methods were employed to quantify PCR products, absolute and relative concentration.

For absolution concentration analysis, standard curves were generated using a known amount of template (purified plasmid DNA). All unknown samples were then plotted on the standard curve and from which the concentration of DNA in the samples was extrapolated. This method was employed to determine the relative abundance of *PQBP1* alternative transcripts.

Determination of the relative abundance was achieved using a ubiquitously expressed gene as a calibrator. Calibrators used in this thesis, and their primer sequences are listed in Table 2.7. This approach requires the calibrator/sample reactions to have the same amplification efficiency which was determined by titrating the calibrator and sample 1,000 fold, where the gradient of the titration series equates to the amplification efficiency of the reaction. Calibrator/sample primer pairs with similar amplification efficiencies (< 0.01) were used for further analysis.

Table 2.7 Real time PCR control primers

Calibrator	Species	Forward Primer (5' → 3')	Reverse Primer (5' → 3')
GAPDH	*H. Sapiens*	GAAGGTGAAGGTCGGAGTC	GAAGATGGTGATGGGATTTC
β-actin	*C. griseus*	ACCAACTGGGACGACATGGAGAAGA	TACGACCAGAGGCATACAGGGACAA

The calculation for quantitation first determined the difference (ΔC_T) between the C_T values of the target and the calibrator:

$\Delta C_T = C_T$ (target) - C_T (calibrator)

This value was calculated for each sample after which one sample (either time = 0 for time course experiments or brain cDNA for expression profiles) was designated as the reference sample. The comparative ($\Delta\Delta C_T$) calculation was then used to determine the difference between each sample's ΔC_T and the reference's ΔC_T.

Comparative expression level = ΔC_T target - ΔC_T reference

Finally, these values were transformed to absolute values using the formula:

Absolute comparative expression level = $2^{-\Delta\Delta CT}$

2.16 Electrophoretic analysis of DNA, RNA and proteins

2.16.1 *Agarose electrophoresis*

Electrophoresis was carried out using gels containing agarose (BioRad, UK) melted in 1x TBE (section 2.4.3). The concentration of agarose ranged between 0.8% and 2.5% w/v depending on the resolution of separation required. Electrophoresis was

performed in 1x TBE with a voltage ranging between 25-150 V, for between 15-120 minutes depending on the resolution required.

2.16.2 *SDS-PAGE*

Denaturing polyacrylamide gels (15% w/v) were purchased from BioRad, UK. Gels were run in a mini-gel apparatus (BioRad, UK) in 1x Running Buffer (section 2.4.3) at 150 V for 45 minutes.

2.17 Bacterial cloning

2.17.1 *Preparation of chemically competent E.coli*

On the day before preparation an inoculation of a single colony representing the appropriate bacterial strain was established in LB broth and grown at 37°C for at least 12 hours. Cells were diluted 1:500 in LB broth and were grown until the A_{550} was between 0.4 and 0.7. Cells were then pelleted by centrifugation at 4000 x g for 20 minutes.

All subsequent procedures following collection of the bacterial cells were completed at 4°C with minimal agitation to the cells to preserve their viability.

Cells were then gently resuspended in 0.2 volumes TFB I and were again pelleted by centrifugation at 4,000 x g for 20 minutes. The cell pellets were then resuspended in 0.02 volumes TFB II. Aliquots of the prepared cells were stored at –70°C for use in bacterial transformations. To ensure that the cells were chemically competent, test transformations were completed used 1 µg of pUC control plasmid (Clontech).

2.17.2 *Subcloning*

The vector and insert to be used were digested with appropriate restriction enzymes (section 2.14.3). The products were gel purified (section 2.14.1) and the concentration of each product was estimated by visualisation on agarose gels.

A 10 µl ligation reaction was prepared using an approximate 3:1 molar ratio of the insert and vector (roughly 150 ng insert and 50 ng vector) together with 0.1 volumes of 10 x ligation buffer, and 1 unit of T4 DNA ligase. The reaction was incubated at 4°C for at least 12 hours.

2.17.3 *Transformations*

Between 20-50 µl of competent cells were thawed on ice, prior to incubation with 2 to 6 µl of ligation mix (section 2.17.2) for 20 minutes (also on ice).The cells were heat-shocked at 42°C for 45 seconds followed by a brief incubation on ice for 5 minutes. Following this, 950 µl of LB were added to each transformation, which was then incubated at 37°C degrees for 1 to 1.5 hours. The transformation mix was then plated out in varying quantities onto agar plates containing amplicillin and where necessary IPTG and Xgal. The plates were incubated overnight at 37ºC. Individual white colonies were resuspended in either 100 µl of sterile water for colony PCR (section 2.15.4), or 1 to 5 ml of media to harvest the plasmids (section 2.14.4 and 2.14.5).

2.18 Identification of transcript variants

PCR products were purified by direct PCR purification (section 2.14.1), and were ligated to pGEM®T- Easy vector (50 ng) (section 2.17.2) before being chemically transformed into JM109 cells (section 2.17.3). Colony PCR (section 2.15.4) followed by agarose electrophoresis with ethdium bromide stainin was completed to confirm the presence and the size of the inserts.

2.19 *PQBP1* open reading frame cloning

Nested primers were designed to amplify the open reading frame of *PQBP1*, and 16 bp upstream from the translation start site and 20 bp downstream from the stop codon. KOD, a proof reading DNA polymerase (Novagen), was used for all amplification procedures.

The first round of PCR was completed on the panel of cDNA from 20 different human tissues (section 2.13.3) using the following cycling profile: 94°C for 2 minutes, followed by 35 cycles at 94°C for 30 seconds, 60°C for 30 seconds and 72°C for 2 minutes, finally followed by 1 cycle at 72°C for 5 minutes. The reaction mix was diluted 1:200 with $T_{0.1}E$. Five microlitres of the diluted reaction mix were used as a template for the second round of PCR.

An additional round of PCR was completed using the internal set of nested primers and the number of amplification cycles was decreased from 35 to 25. Reaction products were purified directly as described in section 2.14.1 and visualised by agarose electrophoresis and staining with ethidium bromide (section 2.16.1).

2.19.1 *A-tailing of purified PCR products.*

The 3' end of PCR products generated using KOD polymerase were adenylated prior to ligation with pGEM®T-Easy. Between 2 and 6 µl of purified PCR product were incubated with 0.1 mM dATP and 0.12 units of *Taq* polymerase for 30 minutes at 37°C. The reaction was terminated by incubation at 70°C for 15 minutes.

2.19.2 *Ligation into pGEM®T-Easy.*

A-tailed PCR products were cloned into pGEM®T-Easy as outlined in section 2.17.2 and transformed into DH10B cells (section 2.17.3). Colony PCR (section 2.15.4) was performed to confirm the presence and the size of the inserts.

2.20 Preparation of constructs for immunofluorescence

The vectors pCDNA3.NT7 and pCDNA3.CT7 containing a T7 epitope sequence were a kind gift from Dr J. Collins. These vectors express a fusion protein consisting of the desired open reading frame and the T7 epitope (at either the amino- or carboxy- terminal). Here, the insert was ligated after digestion with restriction enzymes restriction enzyme sites located either side of the T7 tag.

2.20.1 *N-terminal T7 tag*

PCR using a proofreading enzyme (KOD) was used to introduce *Not*I and *Xba*I restriction sites flanking the open reading frame of the various *PQBP1* clones made in section 2.19. Following amplification, the PCR products were restriction digested with *Xba*I and *Not*I (2.14.3) and agarose gel purified (section 2.14.1). The translation termination codon was retained.

In preparation for ligation with pCDNA.3-Ntag, the vector was also digested with *NotI* and *XbaI* (section 2.14.3) and purified (section 2.14.1). Ligation was performed using 3:1 molar ration of insert and vector (300 ng of insert and 100 ng of vector, section 2.17.2). The ligation was transformed into DH10B and colony PCR (section 2.15.4) confirmed the presence of the inserts.

Individual white colonies were screened for the pCDNA-3-Ntag: PQBP1 plasmids by PCR. Colonies that contained the insert in the correct orientation were harvested and the plasmid purified by midi-prep purification (section 2.14.5). All plasmids were sequenced in order to verify the fidelity of the insert.

2.20.2 *C-terminal T7 tag*

PCR using a proofreading enzyme (KOD) was used to introduce *Hind* III and *Nhe*I restriction sites flanking the open reading frame of the various PQBP1 clones made in section 2.19. Inserts were prepared as outlined previously (section 2.20.1). The stop codon was removed from the PQBP1 ORF to ensure translation of the C-terminal T7 tag.

In preparation for ligation the vector was digested with *Hind*III and *Nhe*I (section 2.14.3) and purified (section 2.14.1). Ligation, and transformation and subsequent analysis of the construct were performed in the same way as described in section 2.20.1.

2.20.3 *Non-directional cloning*

Transcripts containing either a *Not*I, *Xba*I, *Hind*III or *Nhe*I restriction enzyme sites could not use the protocol outlined in sections 2.20.1 and 2.20.2. In such cases, vectors and inserts where digested with XmnI (section 2.14.3). Digested vector was prepared for ligation with the addition of a thymidine residues to generate a 5' overhang. The prepared vector was a kind gift from Dr J. Collins. To complement the thymidine overhang, the 3' end of the insert was adenylated as outlined in section 2.19.1. Ligation, transformation, propagation and sequence verification were completed as before (section 2.20.1).

2.21 Transcript stability assays

2.21.1 *Preparation of inserts*

The *PQBP1* transcripts were excised from the pGEM®T-Easy by restriction enzyme digest with *EcoR*I (section 2.14.3). The excised DNA containing the PQBP1 open reading frame with 24 bp upstream and 17 bp downstream was then gel purified (section 2.14.1).

2.21.2 *Preparation of constructs*

The vector pTRE-TIGHT (Clontech) was prepared for ligation by digestion with *EcoR*I (section 2.14.3) to generate ends complementary to the DNA insert. The digested vector was then treated with alkaline phosphatase (section 2.14.2) to remove the 5'-phosphate group from the exposed ends.

2.21.3 *Ligation of Vector and DNA*

PQBP1 transcripts were ligated with the pTRE-TIGHT (tet-off) vector as outlined in section 2.17.2 and transformed into DH10B cells (section 2.17.3). Colony PCR (section 2.15.4) was used to confirm the presence and orientation of the inserts.

Individual colonies were screened for the incorporation of the pTRE-TIGHT:*PQBP1* DNA construct by restriction digestion with *EcoR*I (section 2.14.3). Colonies that contained that appropriate plasmid were purified by bacterial mini-preps (2.14.4) prior to DNA sequencing to confirm the integrity of the plasmids.

2.21.4 *Transfection of mammalian cells*

CHO-AA8 Tet-off (Luc) cells were plated at a density of 30,000 cells per well on an 6-well culture dish (Falcon). Individual wells were used for each time point as well as each *PQBP1* transcript variant. Cells were incubated for 30 minutes at room temperature, before growing overnight at 37°C, 5% $C0_2$ with humidity. Transfection of the cells proceeded as described in section 2.12.2 using 1 µg of each the pTRE-TIGHT-*PQBP1* construct, and pCMVβ plasmid. Transfected cells were grown overnight prior to the addition of doxycycline.

2.21.5 *Time course experiment*

Doxycycline was added to the media of the transfected CHO-AA8 tet-off cells to final concentration of 50 ng/µl and was mixed by gentle swirling. Cells were then incubated at 37°C, 5% $C0_2$ with humidity for the appropriate length of time. Cells were harvested (section 2.12.1) after 0, 1, 2, 3, 4, 6 and 8 hours. Total RNA was harvested from the cells as described in section 2.13.1 and was treated with DNase (section 2.13.2). cDNA was synthesised from 2 µg of total RNA (section 2.13.3).

2.21.6 *Real-time PCR analysis*

Real-time PCR was established using the reagents and conditions described in section 2.15.6. The amount of cDNA used in each reaction varied with the primer combination that was used. The primer combination and amount of cDNA used (given in the amount of total RNA from which the cDNA synthesised) per reaction were: *ActB* (10 ng), PQBP1-Q10 (25 ng), *LacZ* (25 ng).

All real-time PCRs were performed in triplicate using cDNAs prepared in the presence and absence of RT. Some of the experiments were duplicated.

2.21.7 *Analysis of results*

The relative abundance of *PQBP1* transcript variants and *LacZ* cDNA were normalised to the abundance of *ActB* as outlined in section 2.15.6. To correct for varying transfection efficiencies the relative abundance of the *PQBP1* transcripts were normalised to the amount of LacZ.

Changes in the abundance of the *PQBP1* transcripts were then expressed in relation to t=0. mRNA decay plots were determined by plotting the logarithmic value of the transcript abundance against the incubation period with doxycycline. A linear relationship between the log(relative abundance) and the incubation time was observed between 1 and 4 hours and these time points were used to determine the mRNA half lives of the *PQBP1* transcript variants. The first-order rate constants for *PQBP1* degradation (k= mRNA half-life = gradient), correlation coefficient (r^2) and standard error were calculated from linear regression analysis of the mRNA decay plots.

2.22 Translation Inhibition Time Course Analysis

On the day prior to treatment with translation inhibitors, 2 x 10^5 HEK293FT cells were plated into 6 well culture dishes. The cells were grown to 50-80% confluency. Translation inhibitors cycloheximide (CHX, Sigma), anisomycin (ANS, Sigma) and puromycin (PUR, Sigma) were added to the cells at a final concentrations of 100, 10 and 20 µg/ml respectively. Cells were harvested after 0, 1, 2, 4 and 6 hours.
The samples were prepared for analysis by extracting the RNA (section 2.13.1), DNase I treatment (section 2.13.2) and cDNA synthesis (section 2.13.3).

2.22.1 *Real-time PCR analysis*

Real-time PCR was established using the reagents and conditions described in section 2.15.6. The amount of cDNA used in each reaction varied with the primer combination that was used. The primer combination and amount of cDNA used (given in the amount of total RNA from which the cDNA synthesised) per reaction were: *GAPDH* (10 ng), PQBP1-Q10 (25 ng), *PQBP1*- Q2b, Q3, Q4 and Q6 (50 ng).

All real-time PCRs were performed in triplicate using cDNAs prepared in the presence and absence of RT. Each experiment was duplicated.

2.23 Detection of luciferase activity

2.23.1 *Preparation of cell lysates*

On the day prior to treatment with doxycycline, 2 x 10^5 CHO-AA8 Luc Off cells were plated into 6-well culture dishes. The cells were grown to 50 to 80% confluency. Doxycycline was added to the cells in concentrations ranging between 0 to 1,000 ng/µl which were incubated at 37°C, 5% humidity for 4 hours. Following incubation, the media was removed from the cells which were washed twice in PBS prior to lysing. Cells were lysed in 1x lysis buffer (Clontech) at room temperature for 15 to 20 minutes and were centrifuged at 13 000 x *g* to remove the cellular debris. All samples were assayed immediately after lysis.

2.23.2 *Luciferase activity assay*

The luciferase assay was performed in a white opaque 96-well flat-bottomed plate (Costar) in accordance with the manufacturer's protocol (Clontech). Luciferase activity was measured on a luminometer (BioRad, UK). All samples were assayed in duplicate.

Luciferase activity was normalised against the sample's total protein concentration of the sample which was determined using the Bradford assay.

2.23.3 *Bradford Assay*

Total protein concentrations were measured from the cell lysates prepared in section 2.23.1). The Bradford protein quantification assay was performed in accordance with the manufacturer's instructions (BioRad, UK). Here, a standard curve was constructed using bovine serum albumin (BSA, Sigma) between the concentrations 0 to 1 mg/ml. Absorbance readings were taken at 595 nm, and all samples were measured in duplicate.

2.24 Western Blotting

2.24.1 *Preparation of samples for Western blotting.*

On the day prior to transfection, cells were trypsinized and diluted to a concentration of 8 x 10^4 cells/ml. Transfections were carried out as described in section 2.12.2 in a 24-well culture plate.

Cells were grown for an additional 48 hours and were harvested using 1 x protein sample buffer. All samples were denatured by boiling for 5 minutes, prior to loading on a 15% denaturing polyacrylamide gel (Biorad).

2.24.2 *Electrophoresis of proteins using SDS-PAGE*

SDS-PAGE was carried out using a Mini-PROTEAN® Electrophoresis cell (Biorad) using 1 x running buffer. Proteins were resolved using a 15% separating gel, with a 4% stacking gel (Biorad), and SeeBlue Protein standards (section 2.5). Gels were run at 150 V for approximately 70 minutes.

Electrophoretic transfer

Proteins were transferred to a nitrocellulose membrane using the Mini Trans-Blot® Electrophoresis Transfer cell (Biorad) in 1 x transfer buffer. The electrophoretic transfer was performed at 100 V for 1½ hours at 4ºC.

Detection of proteins

Once proteins were transferred onto the nitrocellulose membrane, the membrane was blocked in blocking buffer for 45 minutes. The blot was then incubated with a mouse anti-T7 monoclonal antibody (stock at 1 mg/ml) (Novagen #69522-4) at a dilution of 1/7,500 in blocking buffer for between 1-12 hours. The blot was washed 3 x10 minutes in PBS-T. The secondary antibody, a sheep-anti-mouse-IgG HRP-conjugate (stock at 0.32 mg/ml) (Sigma #67782) was used at a dilution of 1/7,500 in blocking buffer. Again, the blot was washed for 3 x 10 minutes in PBS-T and then 1 x 5 minutes in PBS. Visualisation of immunoreactivity was completed using ECL (NEN) according to the manufacturers instructions. Membranes were wrapped in Saran Wrap and were then exposed to autoradiography film for between 1 and 15 minutes.

2.25 Intracellular localisation

2.25.1 *Transfection of mammalian cells for immunofluorescence*

Cells were plated at a density of 15,000 cells per well on an 8-well culture slide (Falcon). Cells were incubated for 30 minutes at room temperature, before growing overnight at 37°C, 5% $C0_2$ with humidity. Transfection of the cells proceeded as described in section 2.12.2.

2.25.2 *Fixation of cells*

Twenty-four hours after transfection the cells were washed 3 times in 0.5 ml of 250 mM Hepes, and were then fixed in HISTOCHOICE™ MB® (Amresco), for 20 minutes. The fixation reagent was removed by washing the cells four times with PBS. Cells were then quenched in quenching buffer for at least 15 minutes.

2.25.3 *Antibody staining and visualisation*

The cells were rinsed 3 times with PBS before incubating with blocking buffer for 15 minutes. The well dividers were then removed from the slides.

Primary antibody mouse anti-T7 monoclonal antibody (Novagen) was diluted 1:500 in blocking buffer and cells were incubated with the antibody for 45-60 minutes at room temperature. To remove the antibody cells were rinsed twice with washing buffer, and were then washed for 3 x 10 minutes in washing buffer.

Prior to incubation with the secondary antibody, the cells were incubated in blocking buffer for 15 minutes. The secondary antibody, a goat anti-mouse FITC (stock at 1.1 mg/ml) (Sigma #F2012), was diluted 1:100 in blocking buffer and was incubated with the cells for 45-60 minutes.

Cells were again rinsed twice with washing buffer, and were then washed twice for at least 1 hour. Cells were then stained with DAPI (diluted 1:10,000 in PBS). Finally, cells were washed for 2 x 10 minutes in PBS. Coverslips were mounted using Vectashield (vectorlabs). The FITC fluorescein was typically excited by a 488 nm line of an argon laser, and emission was collected at 530 nm using Nikon Eclipse E800 Microscope confocal microscope (Nikon) and images were captured with a Laser Scanning System ' Radiance 2100fs' (Bio-Rad).

2.26 Computational Analysis

Additional details of the computer programmes used within this thesis can be found at the Wellcome Trust Sanger Institute web site (http://www.sanger.ac.uk/Software/). Frequently used programmes are discussed below.

2.26.1 *Xace*

Human X chromosome data and annotation generated in this thesis were stored in Xace, a chromosome-specific implementation of ACeDB. Other ACeDB implementations were used to store mouse annotation data, as described in Chapter 5. ACeDB was originally developed for the *C. elega*ns genome project (Durbin and Thierry-Mieg, 1996). Documentation code and data available from http://www.acedb.org.

ACeDB works using a system of windows and presents data in different types of windows according to the type of data. All windows are linked in a hypertext fashion, so that clicking on an object will display further information about that object. For example, clicking on a region of a chromosome map will highlight landmarks mapping to that part of the chromosome; clicking on a landmark will display information about that landmark including landmark-clone associations .

In addition to the data generated by the X chromosome mapping group, Xace also contains displays of published X chromosome maps. Genomic sequence data is also displayed in ACeDB along with the collated results from the computational sequence analysis performed by the Sanger Institute Human Sequence Analysis Group.
Xace can be accessed by following the instructions at: http://www.sanger.ac.uk/HGP/ChrX.

2.26.2 *Blixem*

Individual matches identified as a result of similarity searches using the BLAST algorithm, or matches between sequences of cDNA clones or PCR products amplified from genomic DNA generated as part of the project, were viewed in more detail using BLIXEM. BLIXEM, (Blast matches In an X-windows Embedded Multiple alignment) is an interactive browser of pairwise Blast matches displayed as a multiple alignment. Either protein or DNA matches can be viewed in this way at either the amino acid or nucleotide level respectively. BLIXEM contains two main displays: the bottom display panel shows the actual alignment of the matches to the genomic DNA sequence, and the top display shows the relative position of the sequence being viewed within the context of the larger region of genomic DNA. A program “pfetch” retrieves the record from an external database (e.g. EMBL, SWISSPROT).

2.26.3 *RepeatMasker*

Human repeat sequences were masked using RepeatMasker, a program that screens DNA sequence for interspersed repeats and low complexity DNA sequence (Smit & Green RepeatMasker at http://www.repeatmasker.org). The output of the program is a detailed annotation of the repeats that are present in the query sequence and a version of the sequence with repeats masked by "N" characters. Sequence comparisons are performed by the program cross_match, an implementation of the Smith-Waterman-Gotoh algorithm developed by P. Green. The interspersed repeat databases screened by RepeatMasker are based on the repeat databases (Repbase Update Jurka 2000) copyrighted by the Genetic Information Research Institute.

2.26.4 *GAP4*

The quality of DNA sequences generated in this thesis was assessed using Gap4. This sequence analysis software was written to aid the finishing process during genome sequence acquisition. However, several functions of this programme have also been utilised to for smaller scale sequence analysis. The contig editor used phred confidence values to calculate the confidence of the consensus sequence and identifies places requiring visual trace inspection or extra data. Traces were automatically checked for mutation assignments. Vector clip located and tagged vector segments of sequence reads. Additional information about Gap4 can be found at: http://staden.sourceforge.ent/overview.html.

2.26.5 *Perl Scripts*

Computational analysis of large datasets was also performed using customed perl scripts. The scripts, author and function of these scripts are appropriately noted throughout the text.

2.26.6 *Emboss*

EMBOSS is a free open source software analysis package developed by the molecular biology community. It has over 100 applications for sequence analysis which can be accessed via the url - http://www.hgmp.mrc.ac.uk/Software/EMBOSS/. Emboss applications can be run using a webserver, Jemboss, or over the command line when installed locally. Some of the individual programmes that have been used in this thesis are outline in Table 2.8.

2.26.7 *Excel analysis*

All mathematical analysis was carried out using Microsoft Excel.

Statistical analysis was determined using the TTEST function in Microsoft Excel which returns the probability associated with a Student's t-Test. A two-tailed distribution and two-sample unequal variance test was performed.

Table 2.8 Emboss applications used in this study

Programme	Source	Function
Cons	RFCGR	Creates a consensus from multiple sequences
Cutseq	RFCGR	Removes a specified section from a sequence
Diffseq	RFCGR	Finds differences between nearly identical sequences
eprimer3	RFCGR	Picks PCR primers and hybridisation probes
est2genome	Sanger	Aligns EST and genomic DNA sequences
Geecee	Sanger	Calculates the fractional GC content of nucleic acid sequences
Getorf	RFCGR	Finds and extracts open reading frames
Matcher	Sanger	Local alignment of two sequences
Merger	RFCGR	Merge two overlapping sequences
Newseq	RFCGR	Type in a new short sequence
Restrict	RFCGR	Finds restriction enzyme cleavage sites
Revseq	RFCGR	Reverse and complement a sequence
Seqret	Sanger	Reads and writes a sequence
Seqretsplit	RFCGR	Reads writes and returns sequences in individual files
Splitter	RFCGR	Split a sequence into (overlapping) smaller sequences
Transeq	RFCGR	Translates nucleic acid sequences

2.27 Sequence Analysis

The sequence of finished clones was analysed using a standard protocol at the Wellcome Trust Sanger Institute. Briefly, sequence properties of the clones were determined including GC content, and repeat content using RepeatMasker (2.26.3). Prior to repeat masking, the sequence was analysed for other features such as CpG Islands (G. Micklem unpublished), tandem repeats and tRNA genes. Masked sequences were then used for much of the subsequent prediction of potential coding regions. Firstly, a variety of similarity searches against the public domain DNA and protein databases using the BLAST suite of programmes were performed. In addition, *in silico* analysis was also completed using a number of gene and exon prediction programmes, including Genscan (Burge, 1997) and FGenesH (Solovyev *et al.,* 1995). The analysis was performed automatically by the Human Sequence Analysis Group at the Wellcome Trust Sanger Institute and was collated in Xace (section 2.26.1) for manual examination.

2.27.1 *Transcript annotation*

Transcripts were annotated in accordance with the following criteria:

Gene classification	Description
Known	Identical to known cDNAs or protein sequences
Novel CDS	Has an open reading frame and is identical or homologous to cDNAs from human or proteins from all species
Novel transcripts	Similar to above however it cannot be assigned an unambiguous ORF
Putative	Identical or homologous to human spliced ESTs but do not contain an ORF
Predicted	Based on *ab initio* prediction and for which at least one exon is supported by biological data (unspliced ESTs, protein sequence similarity with mouse or tetradon genomes
Processed pseudogenes	Pseudogenes that lack introns and are thought to arise from reverse transcription of mRNA followed by reinsertion of DNA into the genome.
Nonprocessed pseudogenes	Pseudogenes that contain introns and are produced by gene duplications.

Gene structures were annotated onto the genomic sequence using Xace .

2.27.2 *Alignment of nucleic acid and protein sequences*

Nucleic acid and protein sequences were aligned using the program ClustalW (Pearson, 1990; Peasron and Lipman, 1998) via a web-based server at the EBI (http://www.ebi.ac.uk/), or ClustalX installed locally on a PC unless otherwise noted in the text. User-defined parameters were left at their default settings unless directed otherwise in the text. Alignments were then manually edited and presented using the program GeneDoc (Nicholas *et al.,* 1997).

2.27.3 *Calculation of sequence identities and similarities*

Nucleic acid and protein sequences were aligned as described in section 2.24.1. The "statistics report" function of GeneDoc was then used to calculate and display sequence identities and similarities.

2.27.4 *Phylogenetic analysis of protein sequences*

Protein sequence alignments were subjected to various phylogenetic analyses to estimate their order of relationship. In each case, alignments produced as described in section 2.24.1 were manually edited as necessary to minimise the number of gaps, and the most reliably aligned region of the alignment was then

used for the respective phylogenetic analyses. Any columns within the alignment containing gaps were removed prior to phylogenetic analysis.

Phylogenetic analyses were performed using the program Phylo-win (Galtier *et al.,* 1996) installed locally on a PC. This package combines various phylogenetic analysis methodologies in a straightforward interface. In all analyses, 500 bootstrap replicates were selected to assess robustness of the tree produced.

2.28 Comparative sequence analysis

2.28.1 *zPicture*

PIP plots were generated using zPicture as per the authors instructions (Ovcharenko *et al.,* 2004). Text files were generated containing relevant sequences in fasta format, and an annotation file was generated as per the authors instructions. The annotation file was also used to generate an underlay file as per the authors instructions. The base sequence (human unless otherwise specified) was masked for repeats using RepeatMasker. Most program parameters were as default, except sequences were searched on both strands, and chaining was employed. Chaining reports only those matches occurring in the same order in the different species, and avoids build-up of matches due to repetitive sequence occurring throughout the sequences. Chaining assumes that the order of matches should be conserved. The "high sensitivity" setting was also employed.

Chapter 3

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

3.1 Introduction

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

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

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

3.1.1 Xp11.22-p11.3

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

encountered establishing a consensus marker order from these maps - a problem that was later attributed to clonal instability.

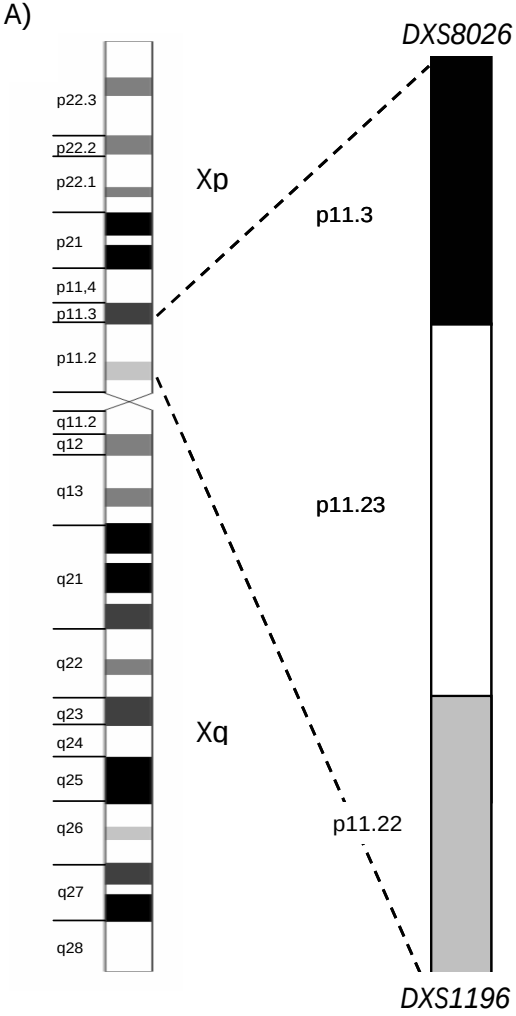
The mapping strategy adopted by the HGP to delineate the clone order of Xp11.22-p11.3 used landmark-based mapping and restriction fingerprinting techniques to generate bacterial clone contigs (Bentley *et al., 2001*). These were then positioned onto physical and genetic maps of the X chromosome. In May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

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

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

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

A)



B)

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

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

A) Ideogram of the human X chromosome. The region analysed in this chapter is also displayed.
B) EMBL accession numbers and Sanger Institute identifiers for clones analysed in this study. GAP = a gap in the physical map at the time of analysis. H – clones annotated by the Human and Vertebrate analysis team (HAVANA).

Results

3.2 Sequence analysis

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

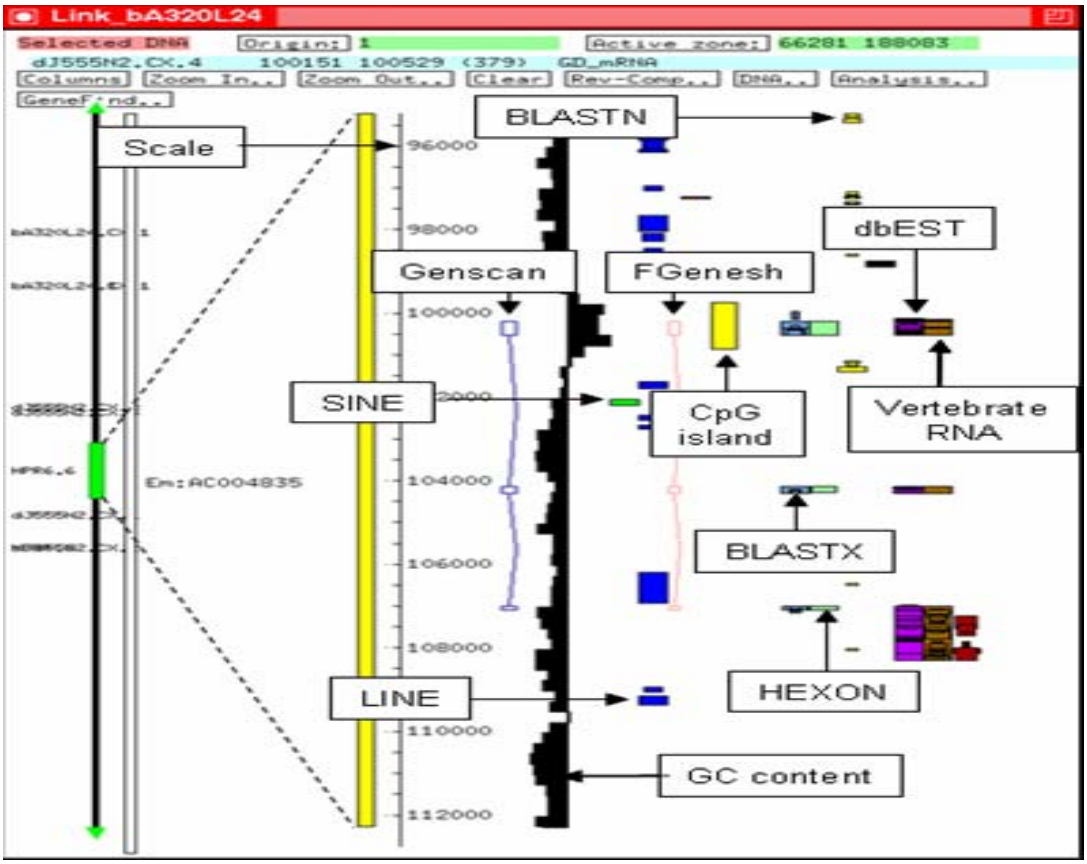


Figure 3.2 Visualisation of sequence data in ACeDB

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

3.2.1 Repeat analysis

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

Table 3.1 Repeat content of human Xp11.23

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

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

* Figures taken from Ross et al, 2005

3.2.2 G+C content

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

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

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

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

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

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

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

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

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

Figure 3.3 Visualisation of sequence alignments using BLIXEM

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

3.3.1 Annotation of known Genes

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

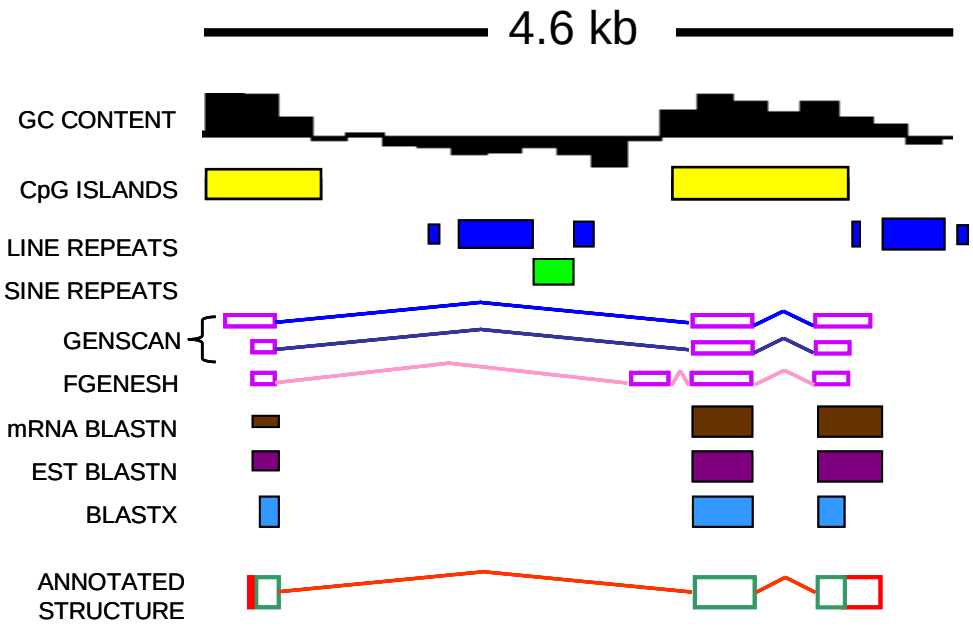


Figure 3.4 Example of a known gene structure, *PCSK1N*.
In this case, the gene was annotated using mRNA accession AF181562 which aligned to the clone AF207550. The diagram shows an ACeDB representation of the gene structure. Displayed from top to bottom are the descriptive analysis of the genome landscape;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

In most cases, the entirety of the mRNA sequence (excluding the polyA tail) matched to the X chromosome sequence. However, the transcripts for the genes *GATA1, CCBN3* and *GRIPAP1* could not be mapped completely to the genome sequence. Subsequent BLAST analysis identified the remainder of *GATA1* in a sequenced clone that had not been analysed (AC115618). It was not possible to identify any genome sequence that contained either a 231 bp internal fragment of *CCBN3* or the first 196 bp of *GRIPAP1*. The gap containing *CCBN3* has not been closed and is now classed as a type 4 gap (a gap in the physical map of the human X chromosome that cannot be closed with existing technologies and resources). A gap was created within the clone AF207550 that was hypothesised to contain the 5' end of *GRIPAP1* and additional mapping was undertaken to close it. The genome sequence that flanked this gap was aligned to end sequences from the fosmid library WIBR2, in the UCSC genome browser (http://www.genome.ucsc.edu/). It was hypothesised that apparently short or orphan fosmids (i.e. only one end

matching the genome sequence) would contain this missing sequence. Eight short/orphan fosmids were identified and were screened using markers designed against the missing mRNA sequence (primer pairs 487051) and the first annotated exon of *GRIPAP1* (primer pair 487049, Table 3.2, Figure 3.5). Three fosmids were identified that contained the 5' end of *GRIPAP1.* The fosmid G248P89409A6 was selected for sequencing as both ends of the clone matched to the genome sequence. Subsequent analysis of its sequence (ACC No: BX530088) confirmed that it contained the missing 196 bp section of *GRIPAP1*.

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

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

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

*results not shown

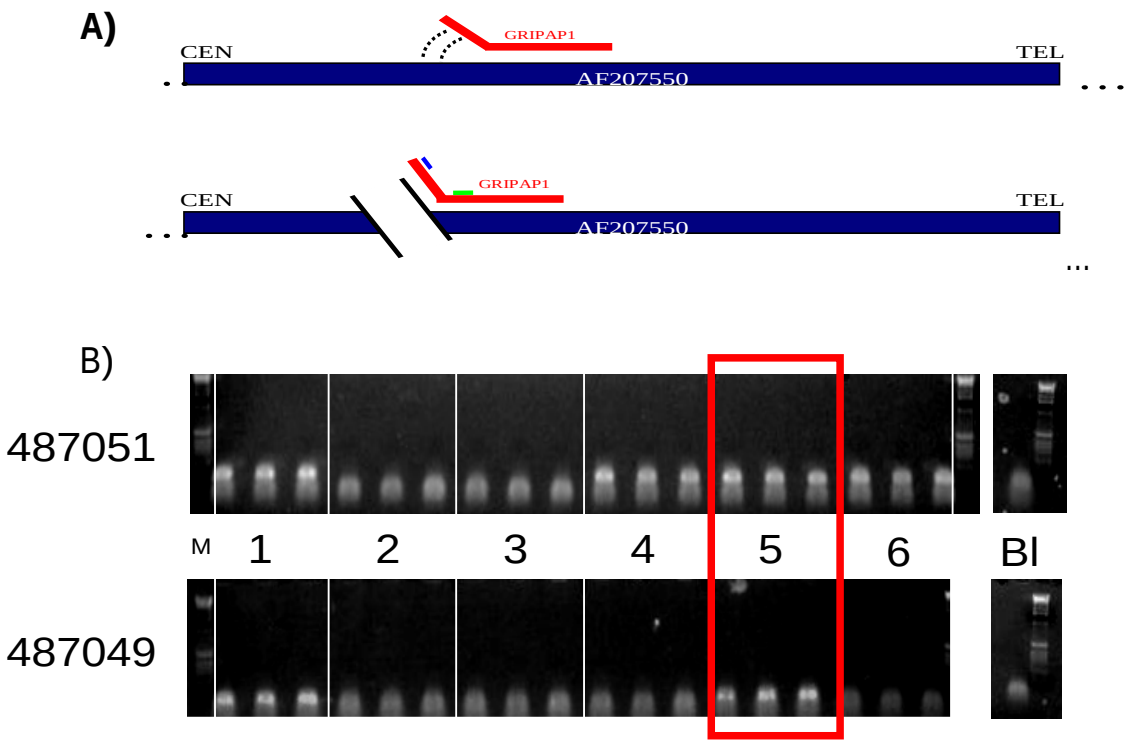


Figure 3.5 Screening of fosmids for 5' end of GRIPAP1.#

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

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

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

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

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

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

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

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

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

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

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

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

3.3.3 Annotation of pseudogenes

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

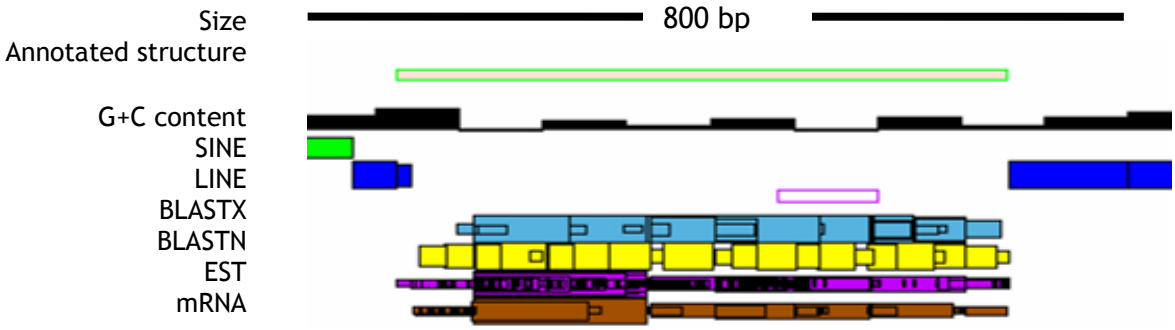


Figure 3.6 Diagram illustrating a "pseudogene" (pseudogene structure), for the locus bA198M15.1.

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

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

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

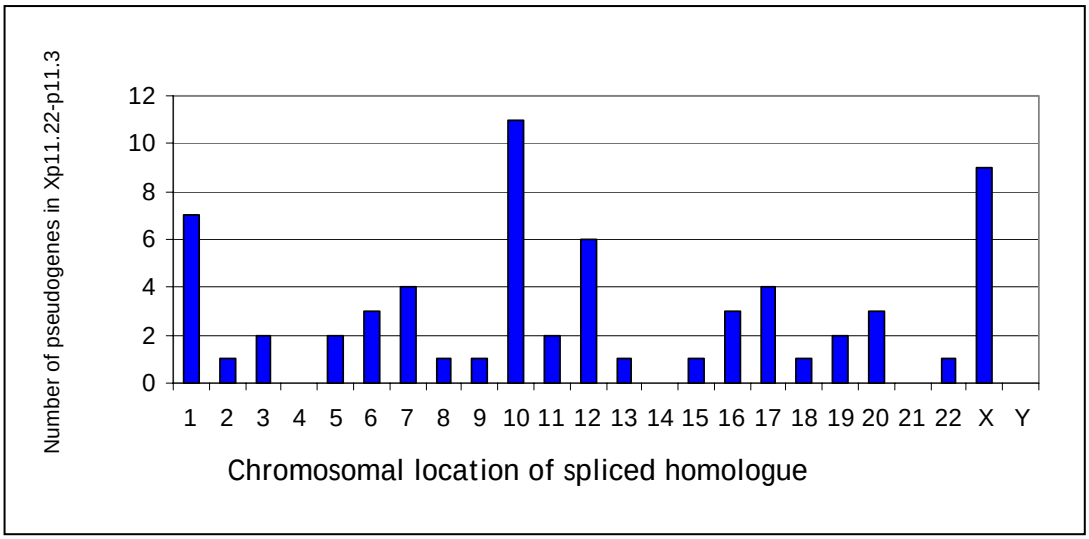


Figure 3.7 Chromosomal origin of pseudogenes identified in human Xp11.22-p11.3.

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

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

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

3.3.4 Distribution of genes

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

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

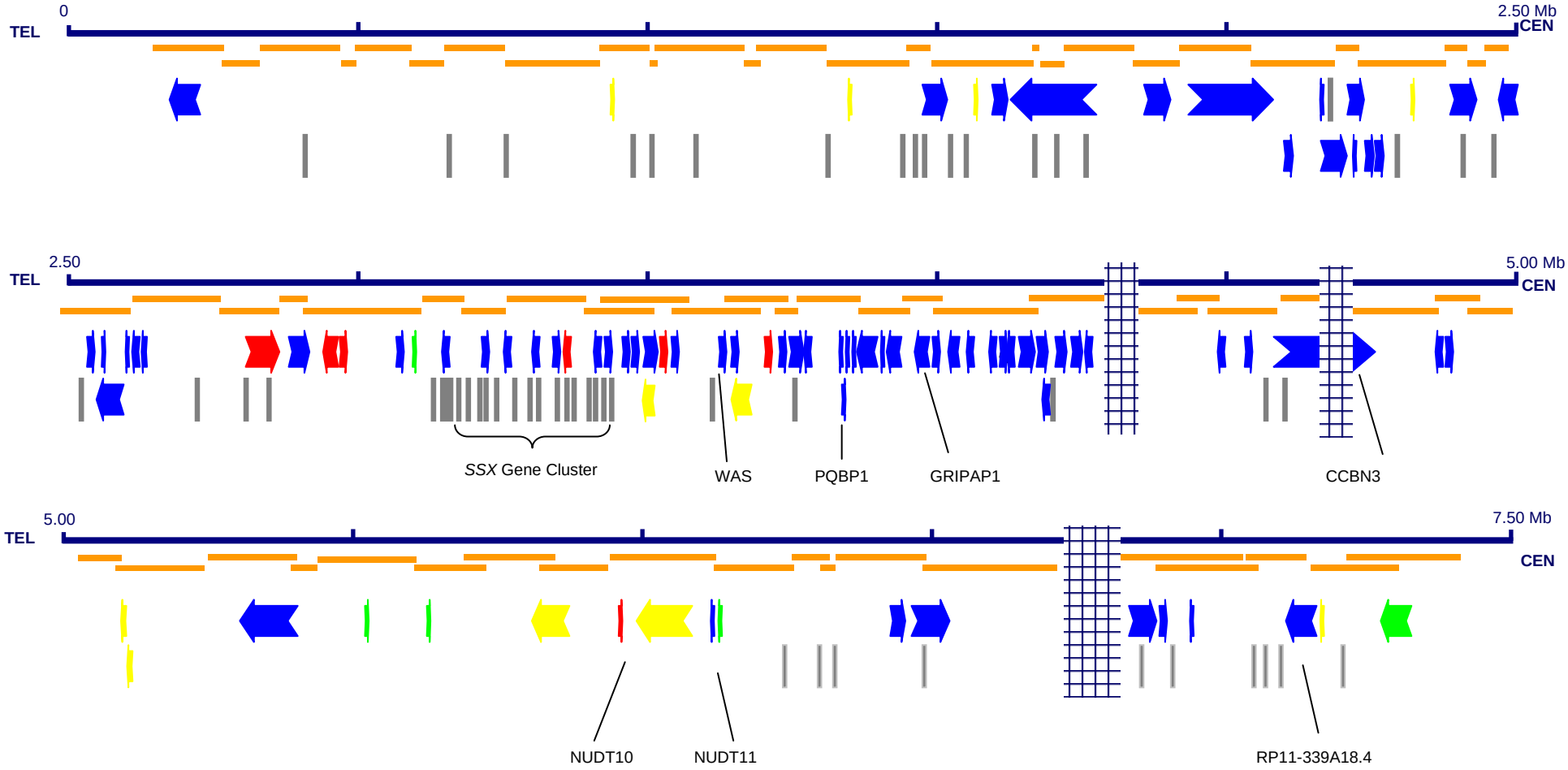


Figure 3.8 Gene annotation of human Xp11.22-p11.3.

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

3.4 Experimental verification of novel and putative genes

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

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

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

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

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

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

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

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

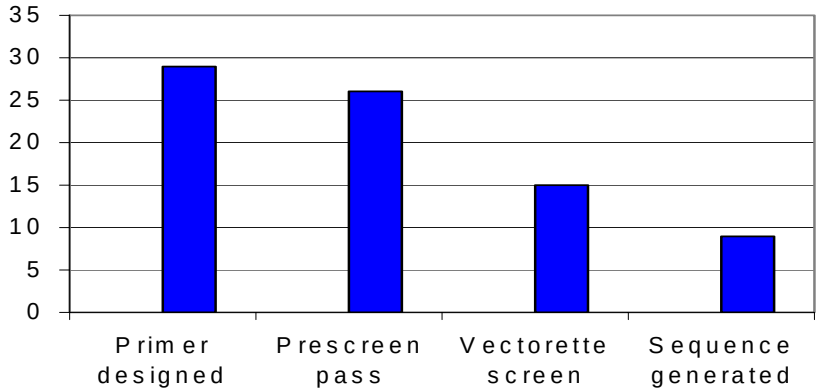


Figure 3.9 Success rates at various stages of analysis to attempt to confirm and extend novel and putative gene structures

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

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

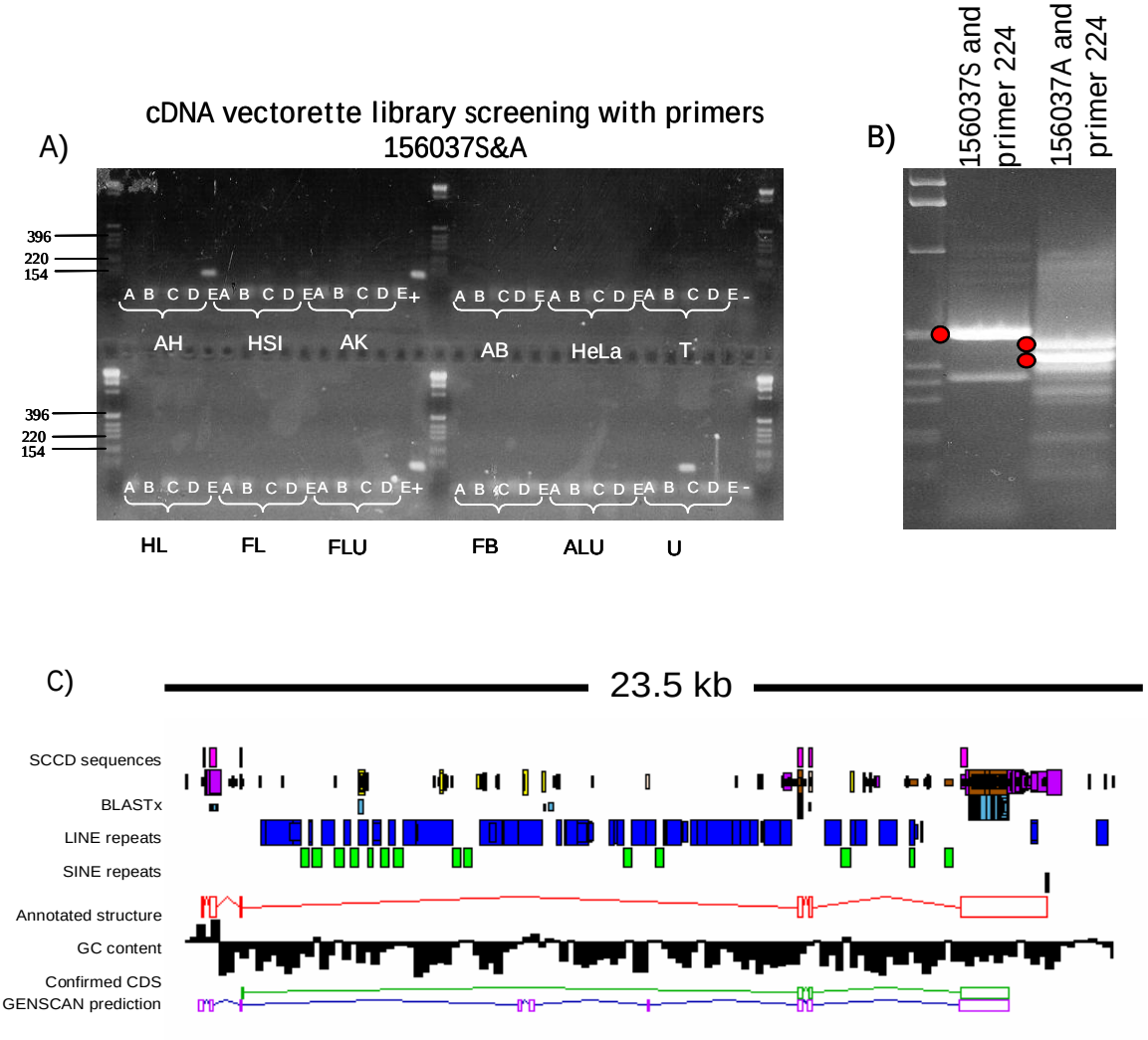


Figure 3.10 Confirmation and extension of novel cds RP11-54B20.4

A) Example of cDNA vectorette library screening for RP11-54B20.4. Products are observed in the vectorette cDNA libraries adult heart (AH) and monocyte (U).
B) Vectorette PCR for the positive superpool AH (E). Bands markers with a dot were excised for sequencing.
C) Gene structures as annotated in Xace. The diagram shows an ACeDB representation of the gene structure, RP11-54B20.4. The confirmed cds structure is then shown. The GENSCAN prediction for this RP11-54B20.4 is also displayed

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

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

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

Table 3.5 continued….

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

Table 3.5 continued….

Putative genes

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

3.4.1 Antisense Transcripts

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

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

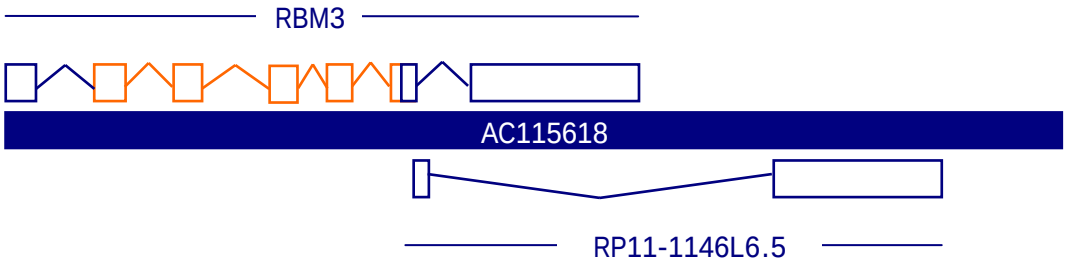


Figure 3.11 Annotation of RBM3 and its antisense gene AC1145L6.5 on the genome sequence, AC115618

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

3.4.2 Transcript features

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

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

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

* longest annotated transcript used
** figures from Ross *et al.*, 2005
[@] CCBN excluded from analysis because it spans a gap in the genomic sequence

3.5 Assessing the completeness of annotated genes

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

3.5.1 Polyadenylation sites

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

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

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

The sequence composition of these signals was assessed in known genes. The highly utilised AATAAA hexanucleotide was present in 64 genes, with its most common

variant being ATTAAA which was identified in five genes. Five other variants were also identified in known genes (Figure 3.12C).

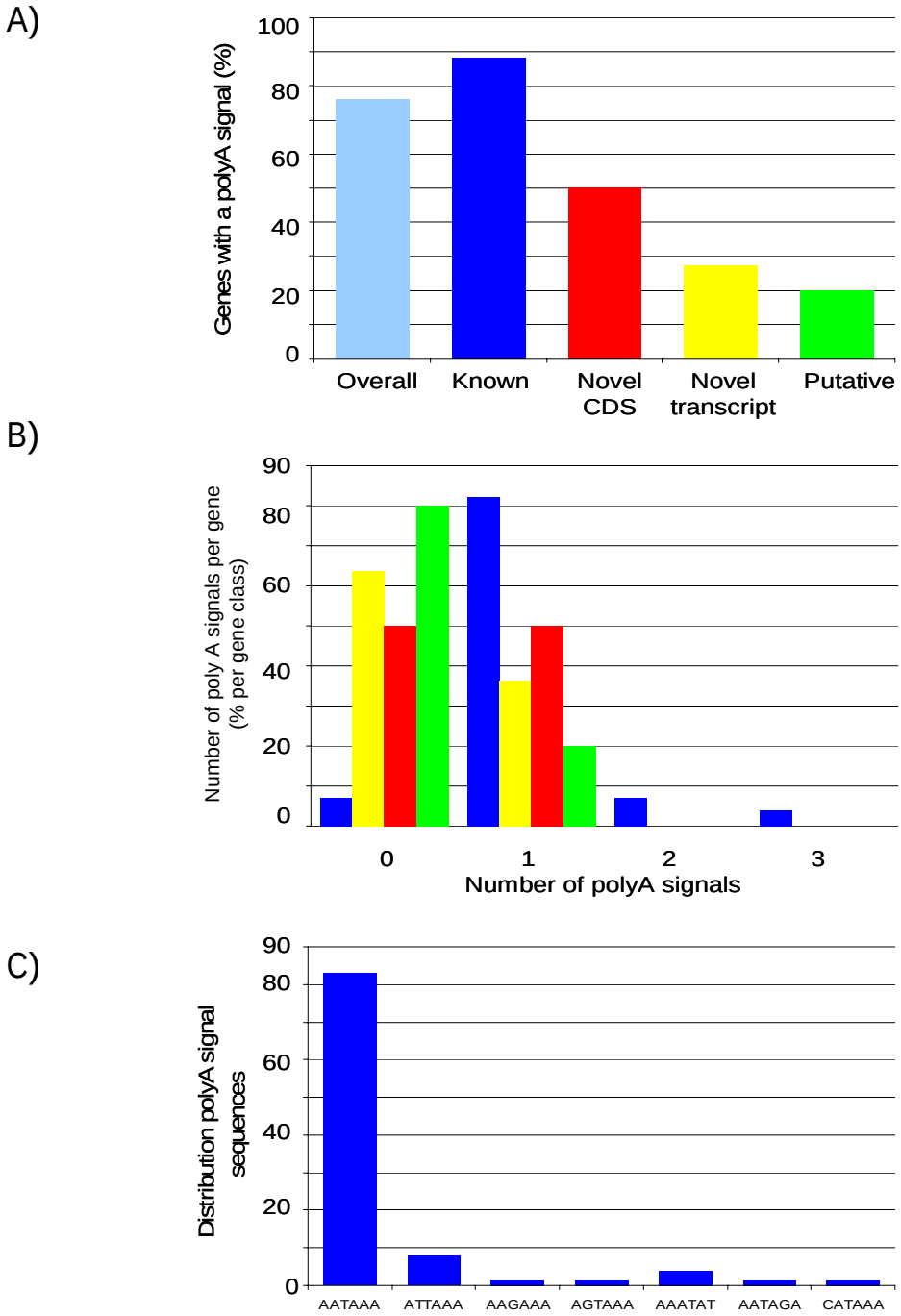


Figure 3.12 Analysis of polyadenylation signals in human Xp11.22-p11.3

A) Proportion of genes with a polyA signal. Known genes (blue), novel CDS (red), novel transcripts (yellow) and putative genes (green).
B) Percentage of polyA signals identified for each gene type: known genes (dark blue, n=77), novel CDSs (red, n=8), novel transcripts(yellow, n=11), and putative genes (green, n=5).
C) Distribution and sequence composition of polyA signals in all genes that have a polyA signal.

3.5.2 Transcription start sites

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

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

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

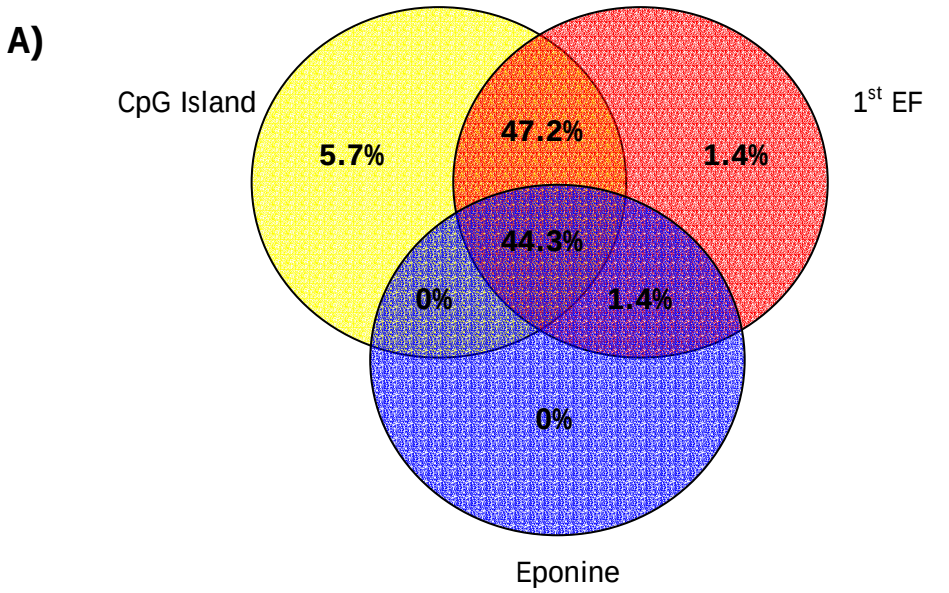
Annotated pseudogenes were also included in the analysis. Here, it was predicted that most pseudogenes would not be transcribed, would not have a functional TSS and therefore would not have any predicted TSSs.

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

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

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

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



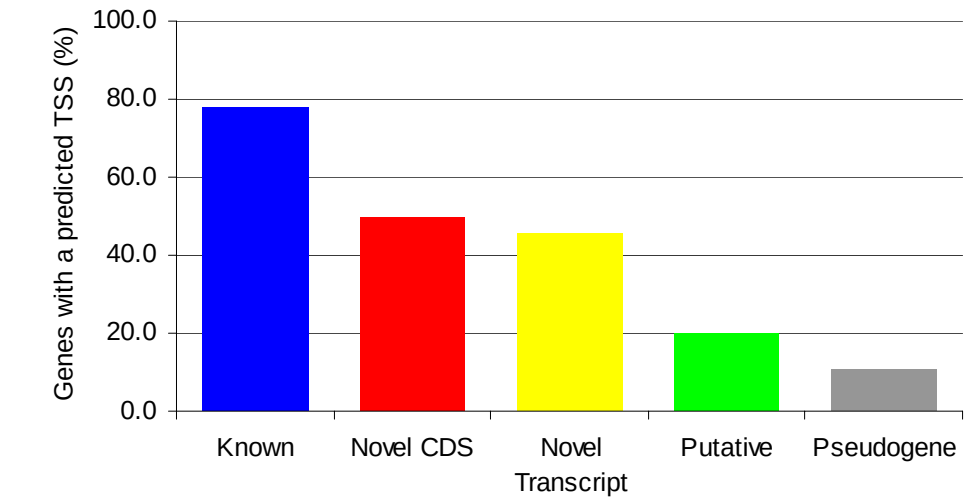


Figure 3.13 Distribution of predicted transcript start sites (TSSs) for genes and pseudogenes in human Xp11.22-p11.3

Three programmes, Eponine, 1st EF and CpGIsland were used to identify TSSs in the genome sequence from human Xp11.22-p11.3.
A) Venn diagram displaying the proportion of all predicted TSSs identified by each programme.
B) Percentage of annotated genes with a predicted TSS in their first exon.

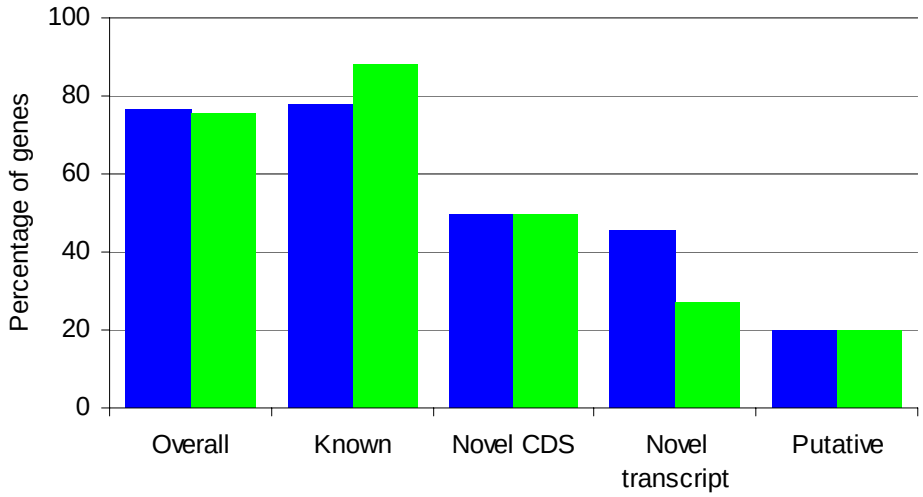


Figure 3.14 State of completion of annotated gene structures
Completion at the 5' end was assumed in the presence of a predicted TSS. Completed 3' ends had a polyA signal and polyA in their transcript sequence (green). The results are displayed for each gene class.

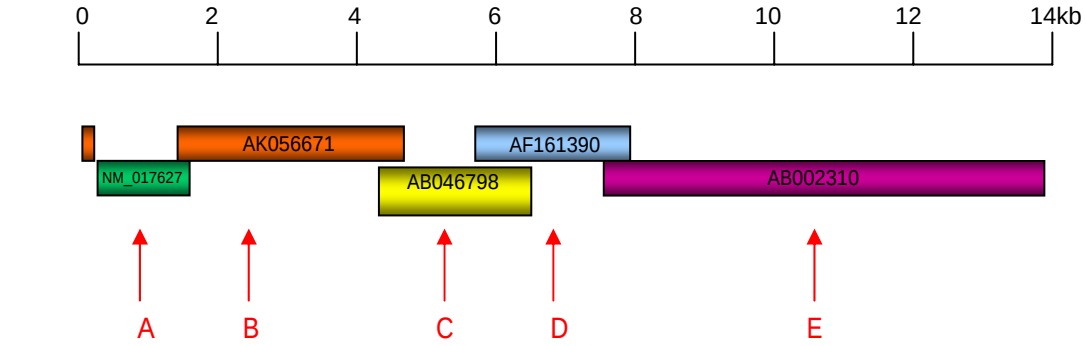
3.5.3 Experimental evidence confirms the transcript size of RP11-339A18.4

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

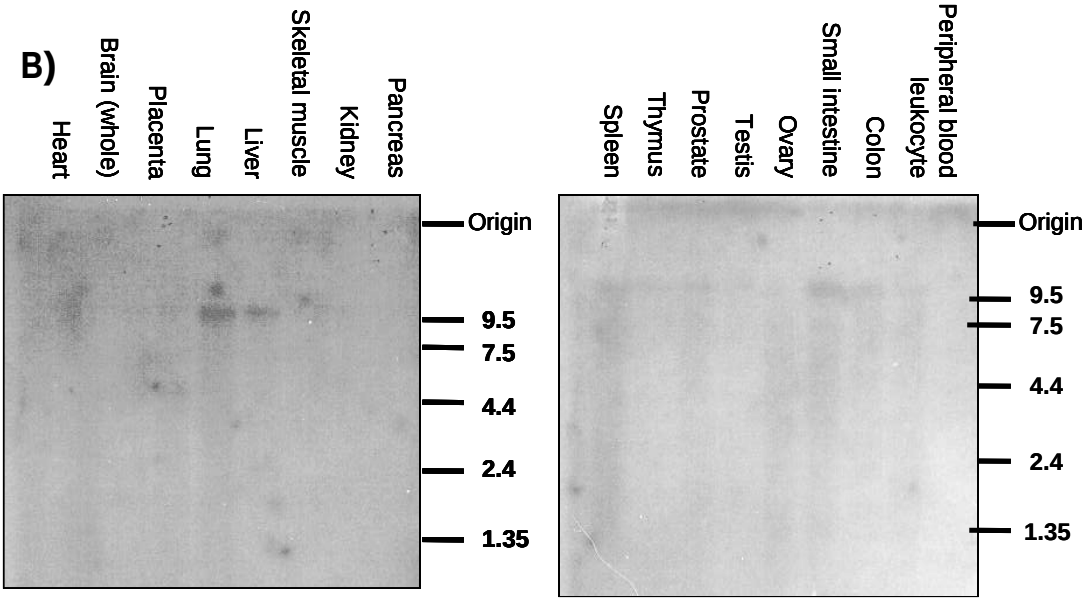
The gene RP11-339A18.4, was analysed because it was annotated from five separate overlapping mRNA transcripts. Northern blot analysis was carried out to confirm that the five RNAs did link up to produce the very large transcript. The gene spans 180 kb, consists of 83 exons and its processed transcript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

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

A)



B)



C)

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

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

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

3.5.4 Alternative splicing

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

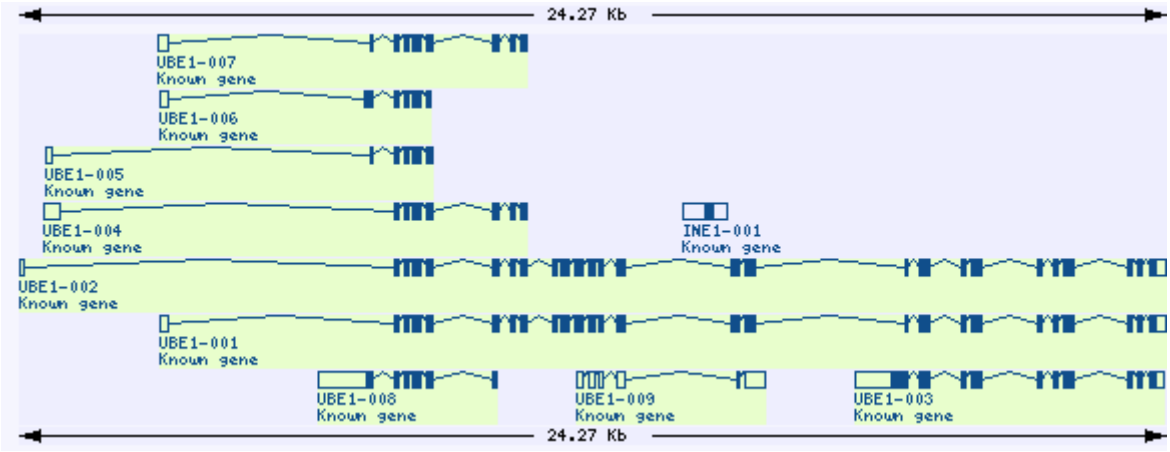


Figure 3.16 Transcript variants of the gene *UBE1*
Open boxes represents UTR sequence, while filled boxes depict coding sequence. This figure was extracted from Vega http://vega.sanger.ac.uk.

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

3.6 Duplication events

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

3.6.1 Duplication of the SSX family

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

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

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

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

A)

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

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

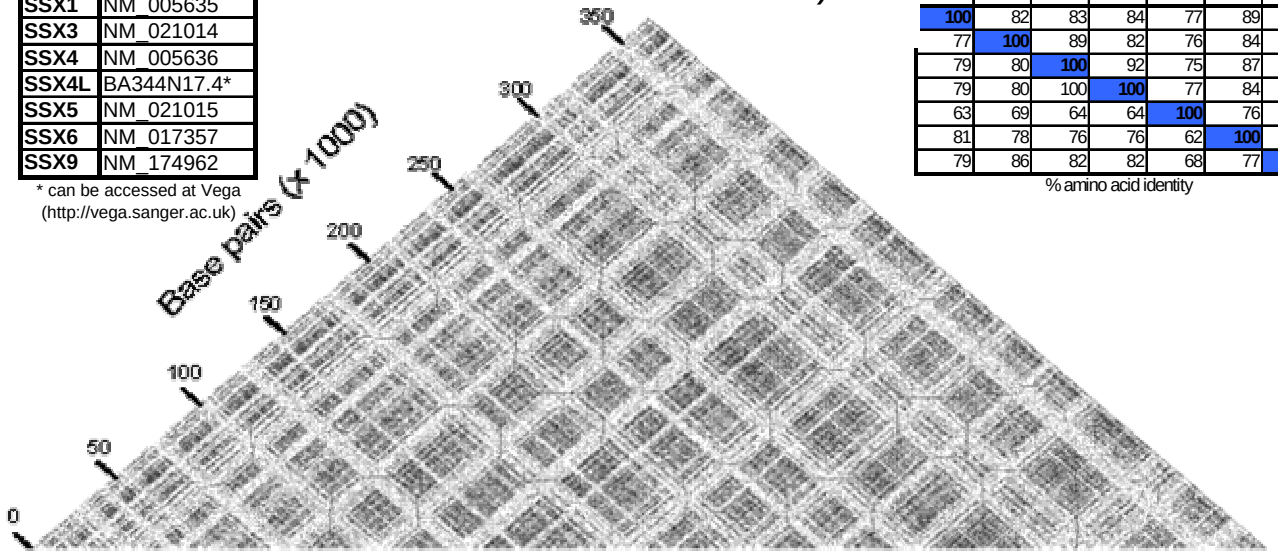
B)

% nucleotide identity

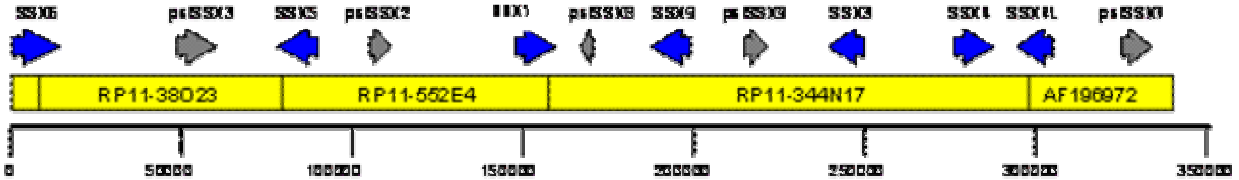
	SSX1	SSX3	SSX4	SSX4L	SSX5	SSX6	SSX9	
SSX1	100	82	83	84	77	89	85	SSX1
SSX3	77	100	89	82	76	84	88	SSX3
SSX4	79	80	100	92	75	87	89	SSX4
SSX4L	79	80	100	100	77	84	83	SSX4L
SSX5	63	69	64	64	100	76	76	SSX5
SSX6	81	78	76	76	62	100	86	SSX6
SSX9	79	86	82	82	68	77	100	SSX9

% amino acid identity

C)



D)



bp

Figure 3.17 Annotation of *SSX* gene family members

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

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

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

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

3.6.2 Genes located in an inverted repeat

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

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

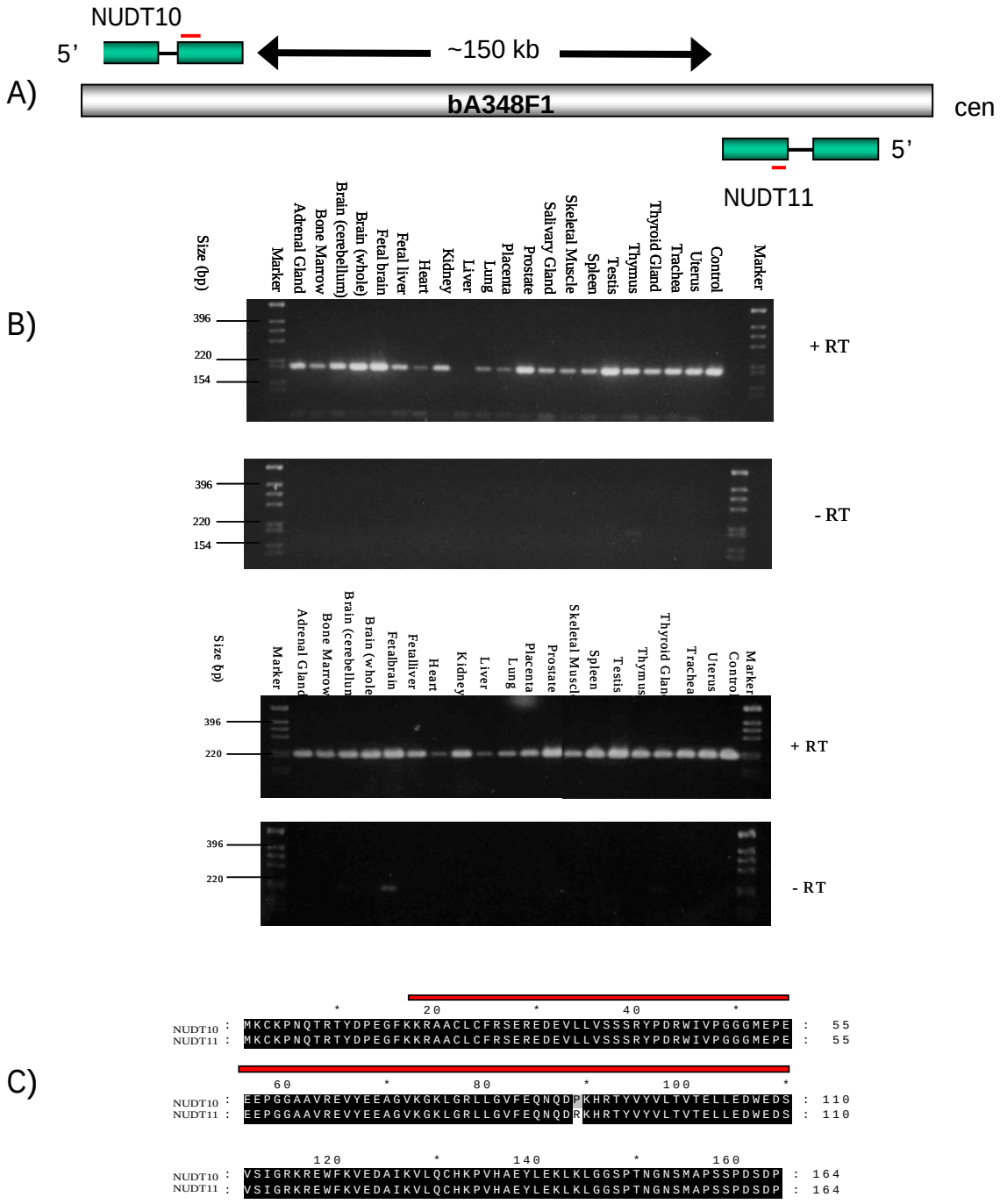


Figure 3.18 Gene duplication in Xp11.23

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

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

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

3.7 Discussion

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Complete genome annotation relies on comprehensive transcriptome characterisation. Apart from the genes that might be expressed in a cell, the additional complexity of a transcriptome is mostly created by three major mechanisms, namely alternative transcription initiation, alternative splicing, and alternative polyadenylation. In particular, one of these features was highlighted by the analysis completed in this chapter - the complexity of the transcriptome generated by alternative splicing. A high degree of variation was observed in the splicing patterns of the annotated genes. Transcript variants can arise as a result

of highly regulated splicing events, and they may serve to enhance the diversity of the proteome that is create more functional products from a limited number of loci. Alternatively, transcript variants may be the result of mis-splicing events, and may not serve any function in the cell. As a preliminary measure to further investigate the diversity of the transcriptome it was decided to assess the presence of alternative transcripts in Xp11.23 in greater detail. The data generated within this chapter provided the basis for the experimental work described in the following chapters.

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

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.

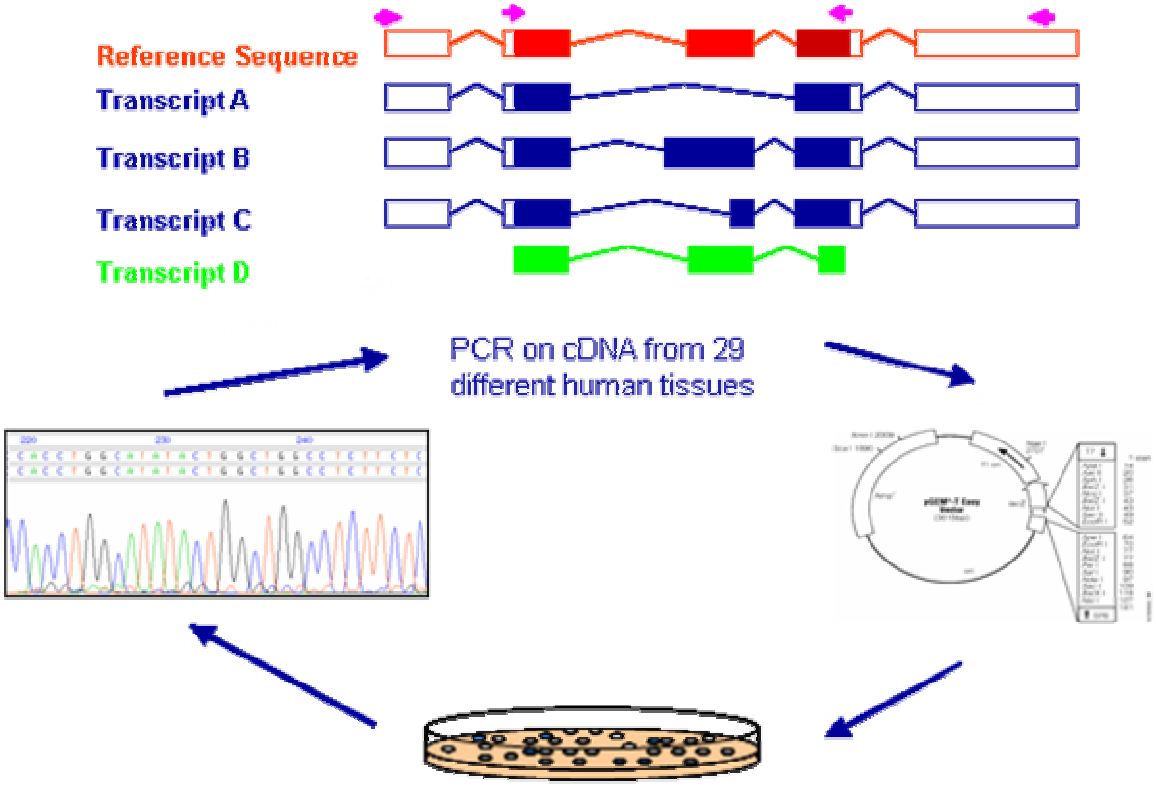


Figure 4.1 Experimental strategy to identify novel transcripts in human Xp11.23
Gene fragments are amplified from a panel of 29 human tissues (primers shown as pink arrows). The PCR products are then cloned into a holding vector and sequenced. The resulting sequences are then aligned against the genome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

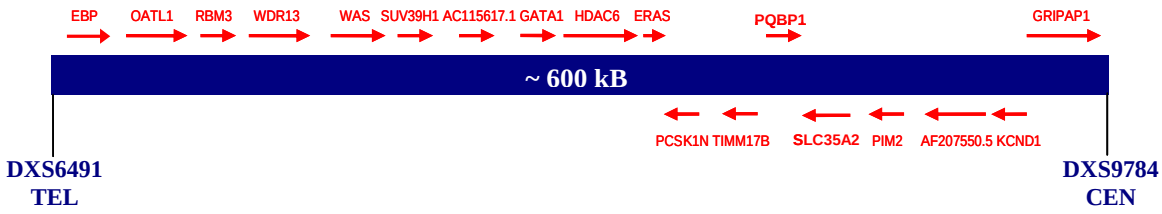


Figure 4.2 Order and orientation of genes involved in this study

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

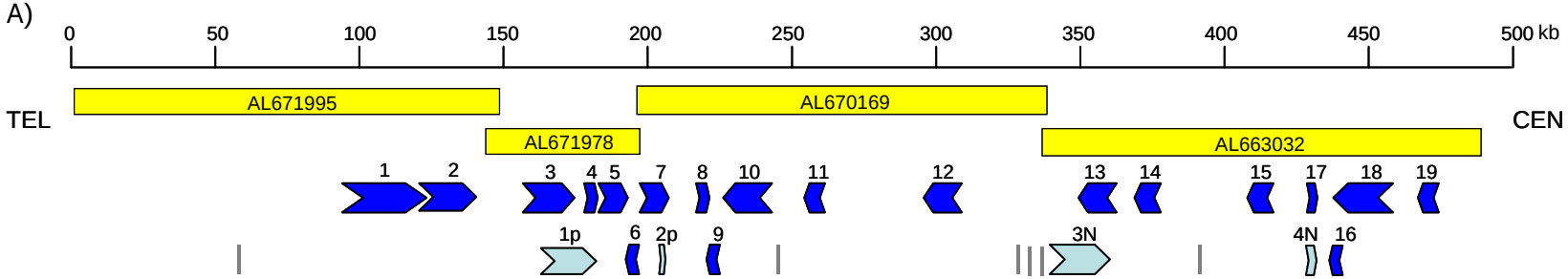
Orthologous genes in mouse for each of the 18 human genes included in this study were identified by BLASTp analysis performed at the National Center for Biotechnology Information (NCBI). Potential orthologues were identified for all 18 genes, and the nucleotide and amino acid sequences were aligned using 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)

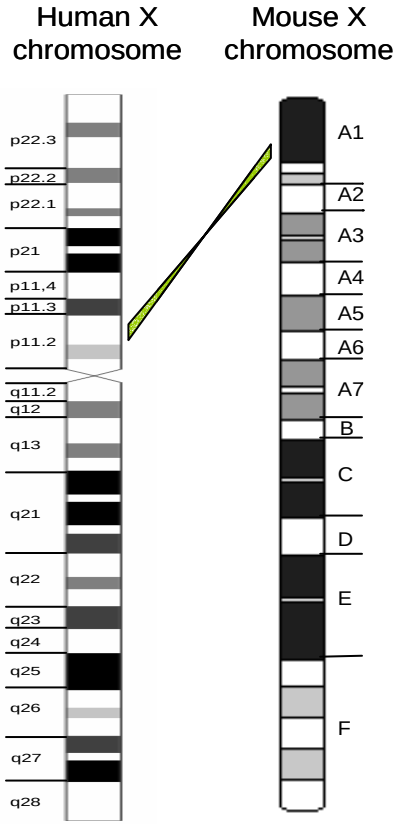


Figure 4.3 Annotation of mouse orthologues

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).

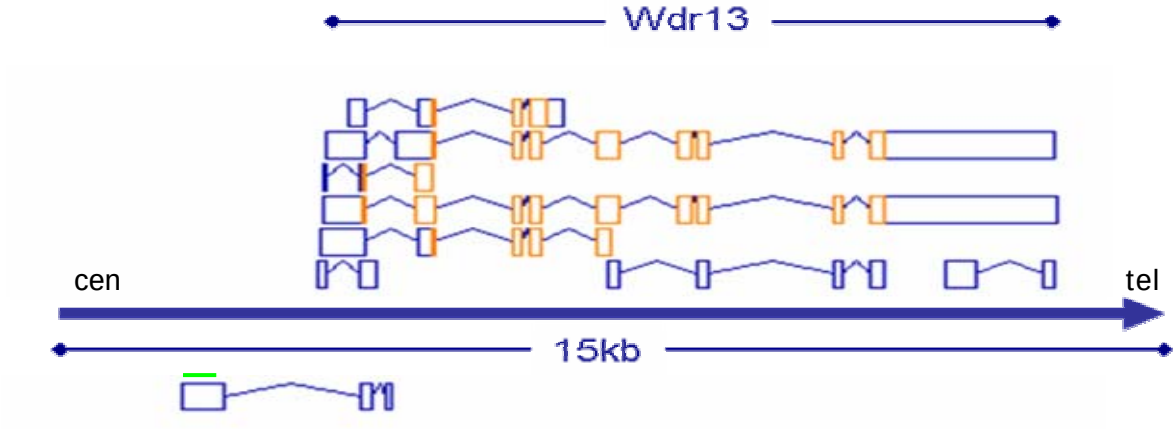


Figure 4.4 Annotation of an antisense transcript to *Wdr13* in the mouse.
The genome sequence is displayed from centromere to telomere. Annotated transcripts on the genome sequence are displayed – orange CDS, blue non-coding. Gene structures displayed above the genome sequence are transcribed from left 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.

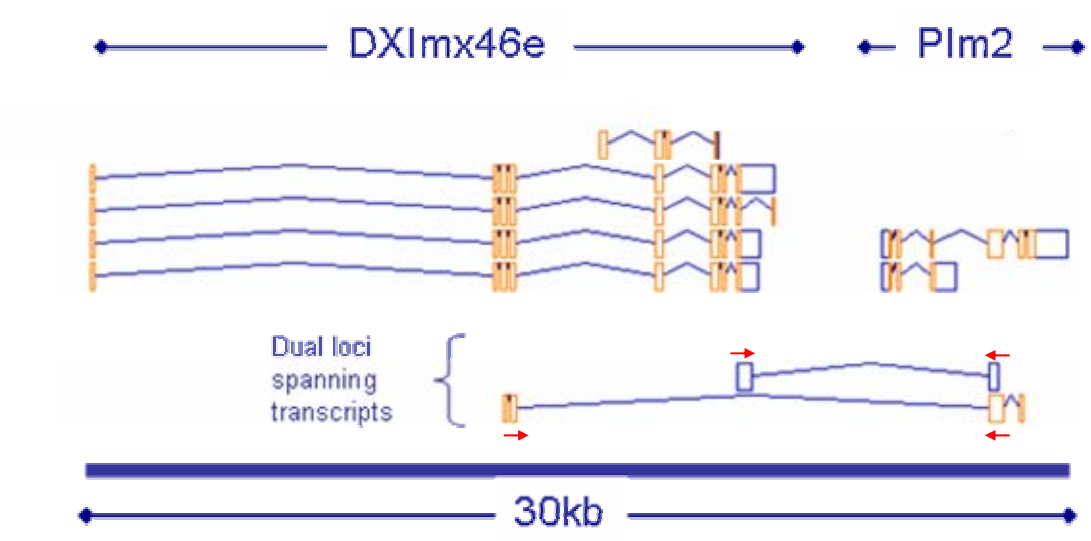


Figure 4.5 Annotation of EST sequences spanning *PIM2* and DXIm46e
The genome sequence is displayed from the centromere to the telomere. Annotated transcripts on the genome sequence are displayed – orange cds, blue non-coding. Gene structures displayed above the genome sequence are transcribed from left 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.

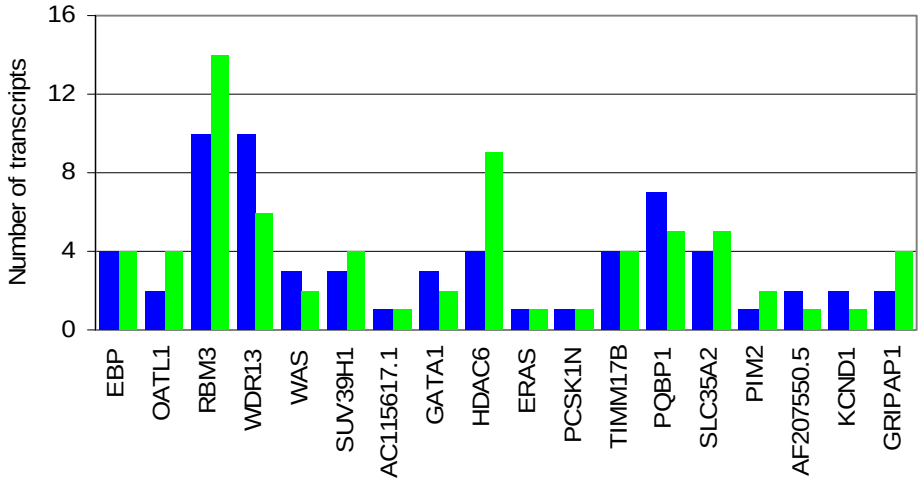


Figure 4.6 Number of alternative transcripts that were annotated for each of the gene pairs.
Human (blue), mouse (green).

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

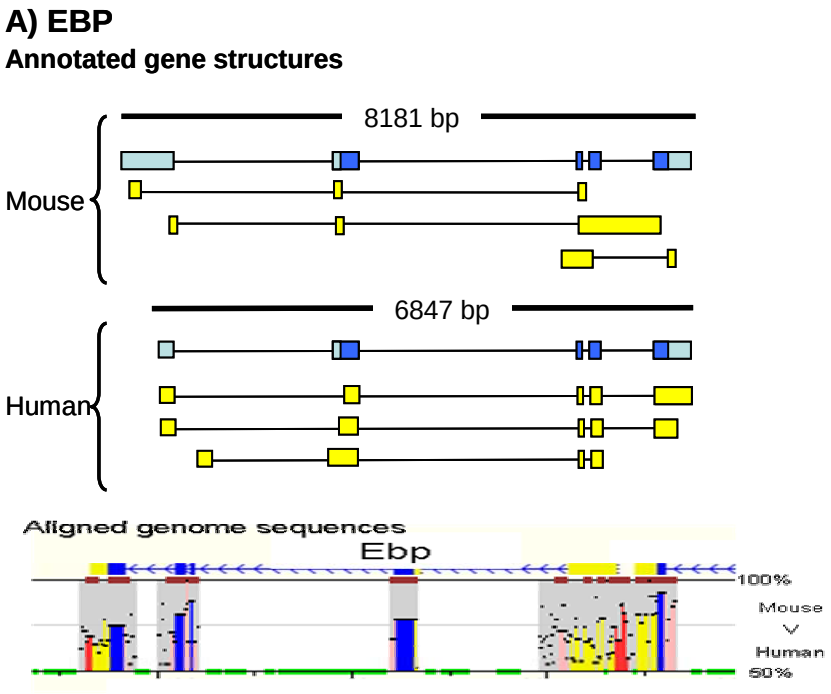


Figure 4.7 Identification of novel transcribed regions using comparative analysis

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)

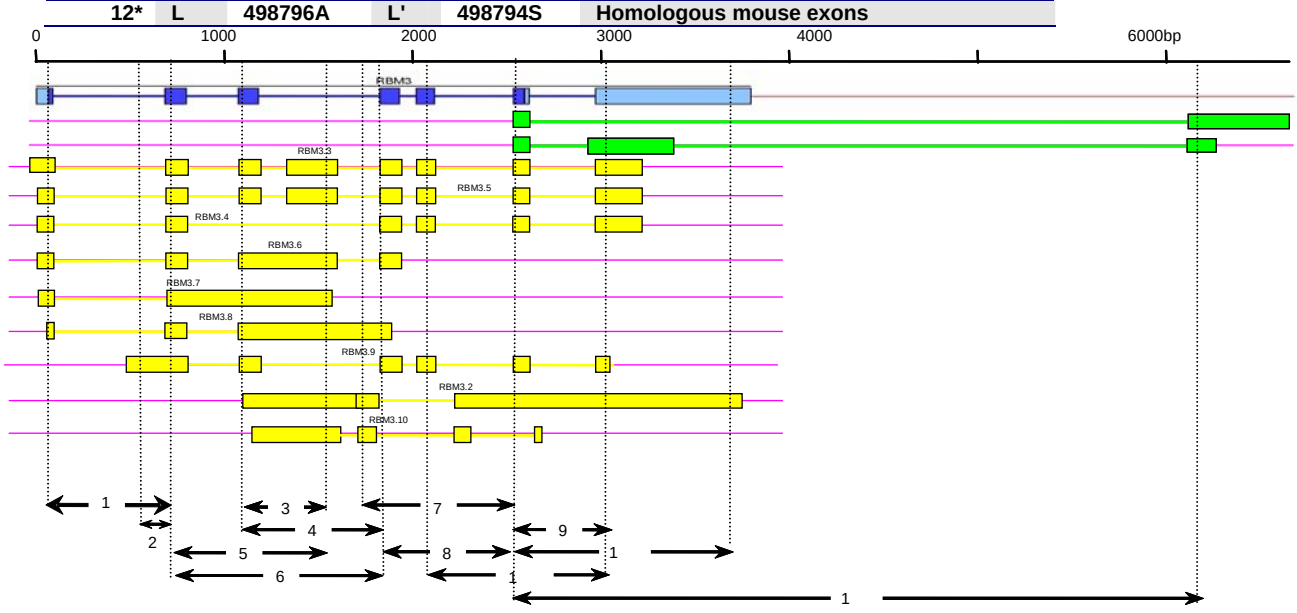


Figure 4.8 Transcript profiling of *RBM3* in 29 different tissues

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)

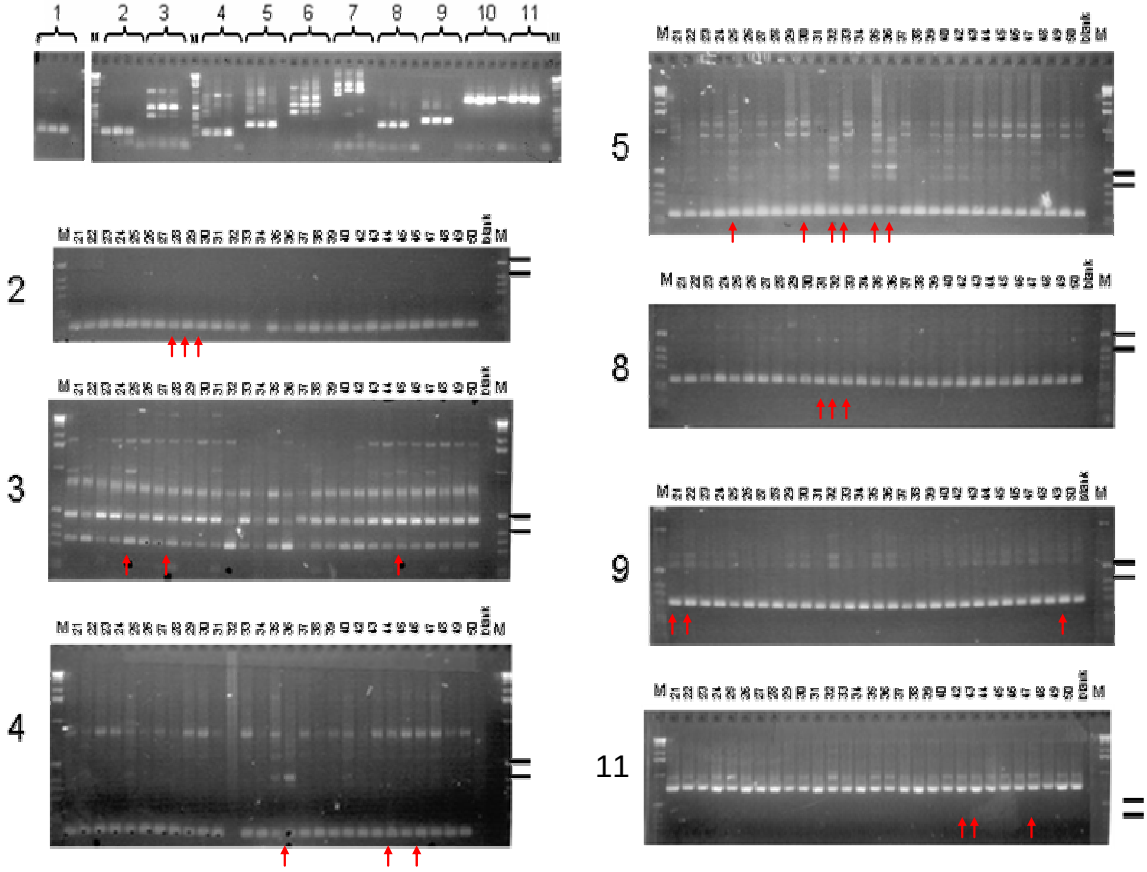
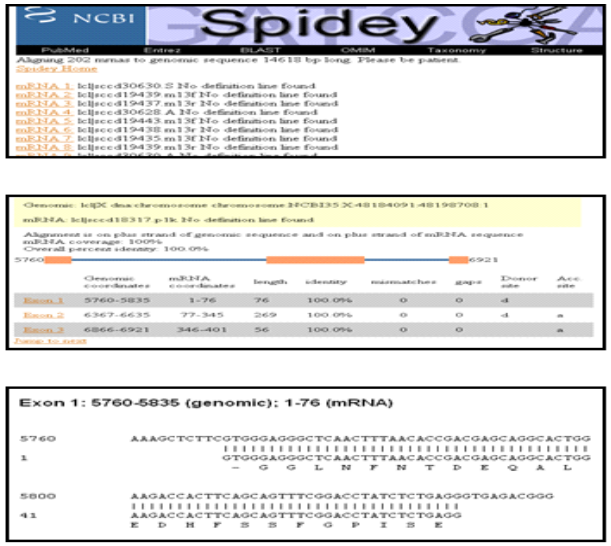


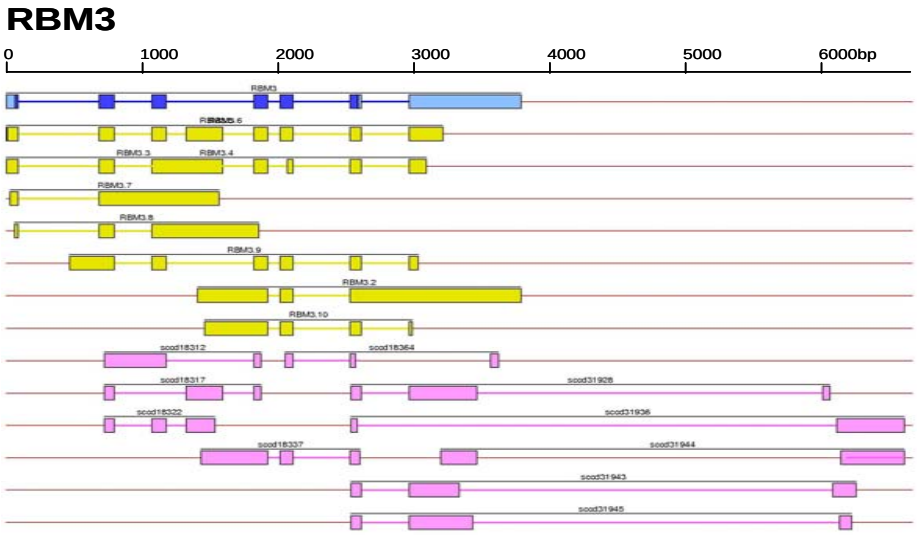
Figure 4.9 cDNA screens of *RBM3*

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).

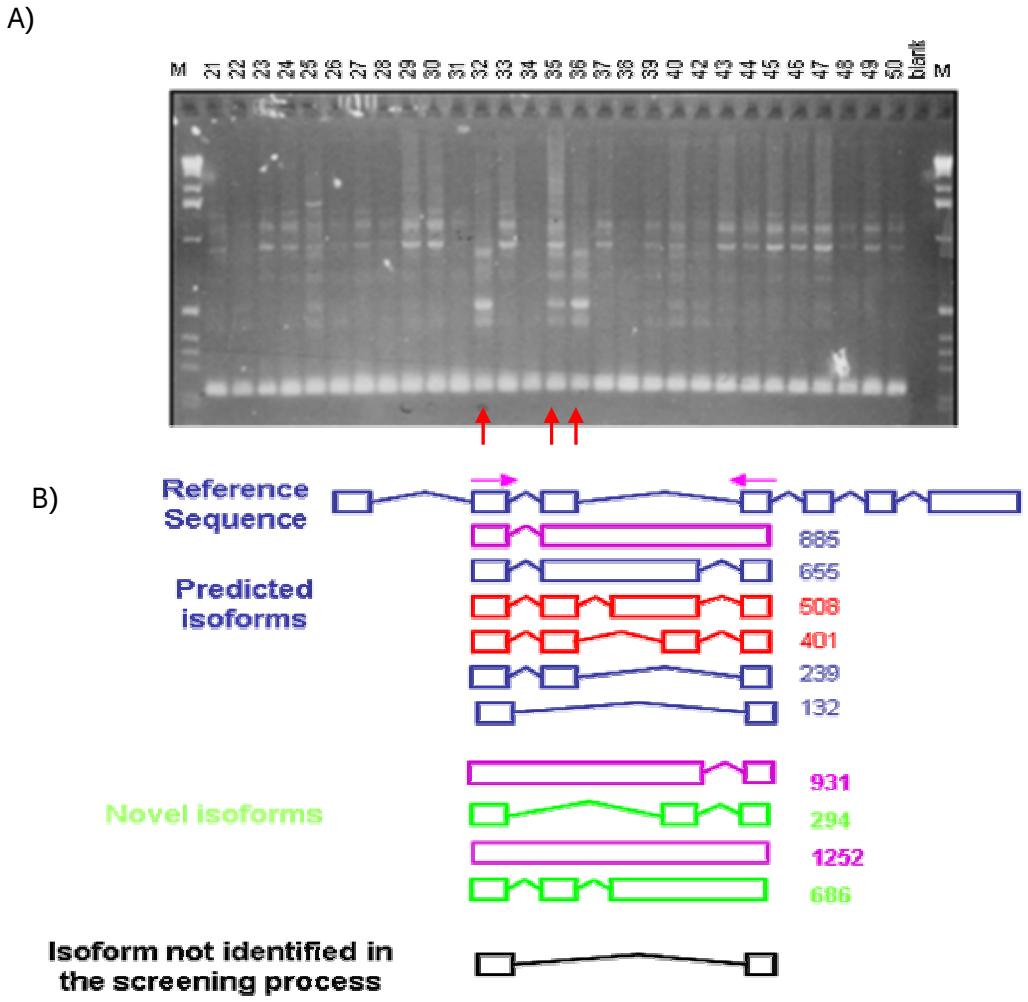


Figure 4.11 Tissue specific expression patterns of *RBM3*

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)

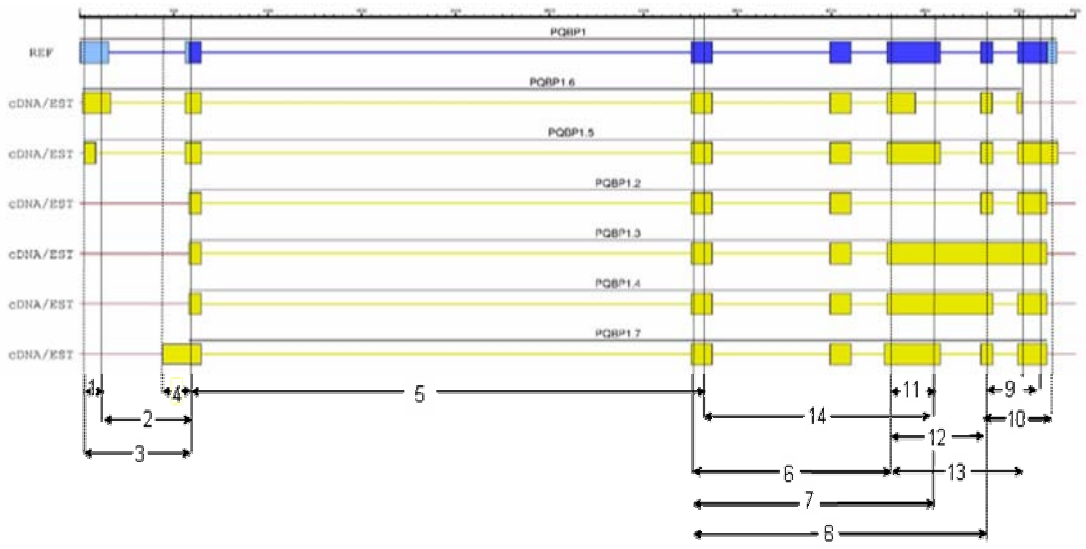


Figure 4.12 The location of primers used to screen for novel *PQBP1* transcripts.

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).

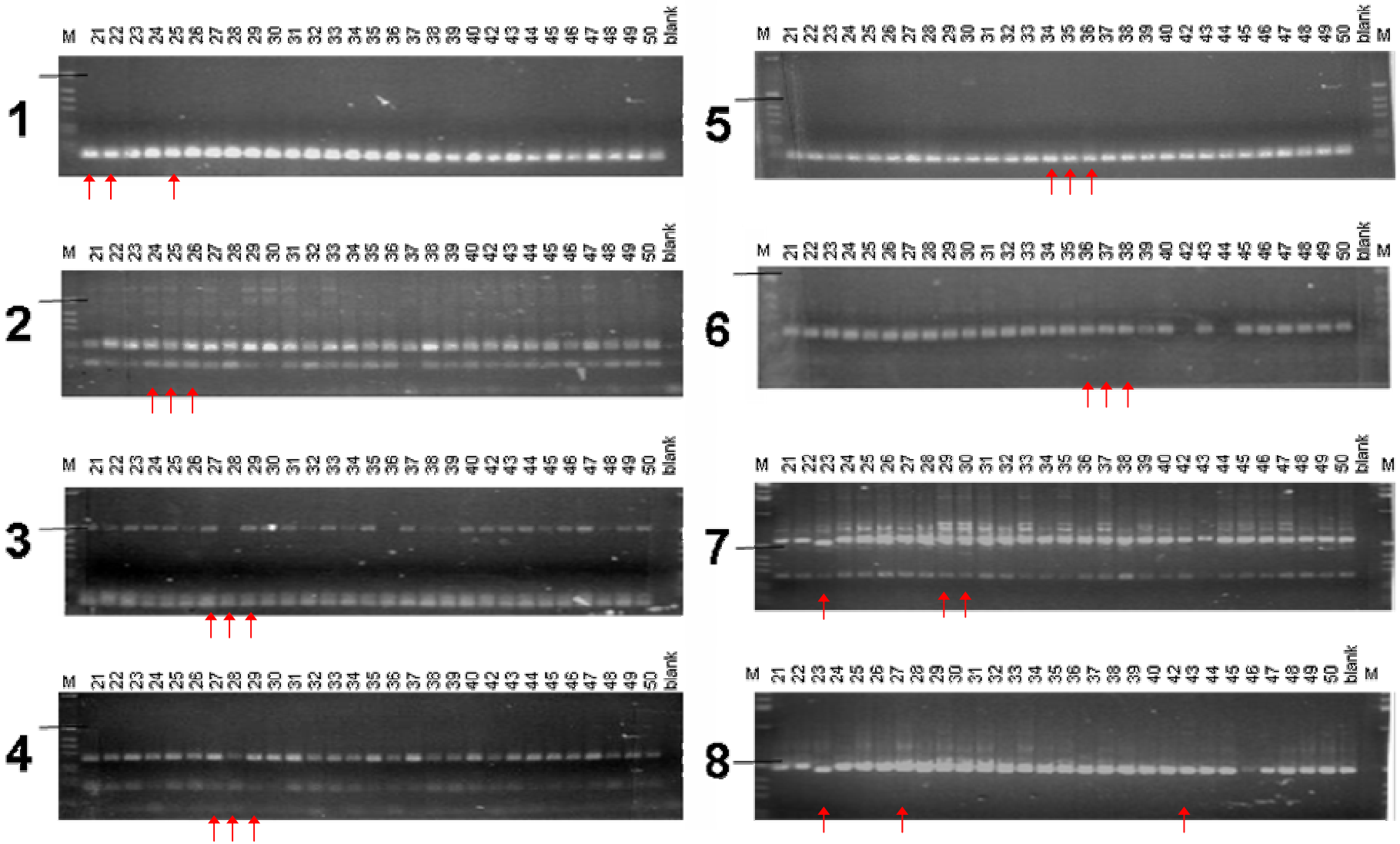


Figure 4.13 Identification of novel transcripts for *PQBP1*- cDNA screens

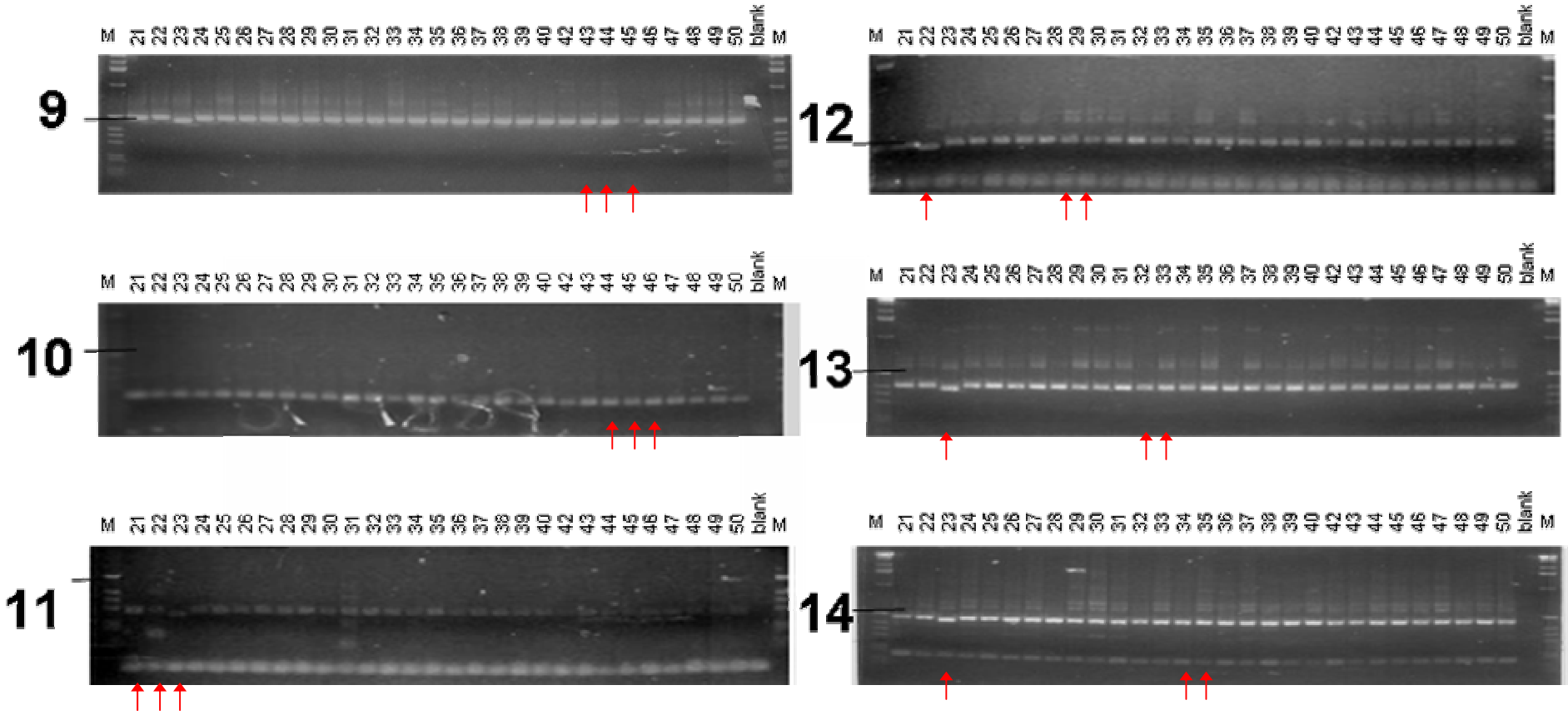


Figure 4.13 Identification of novel transcripts for *PQBP1*- cDNA screens

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.

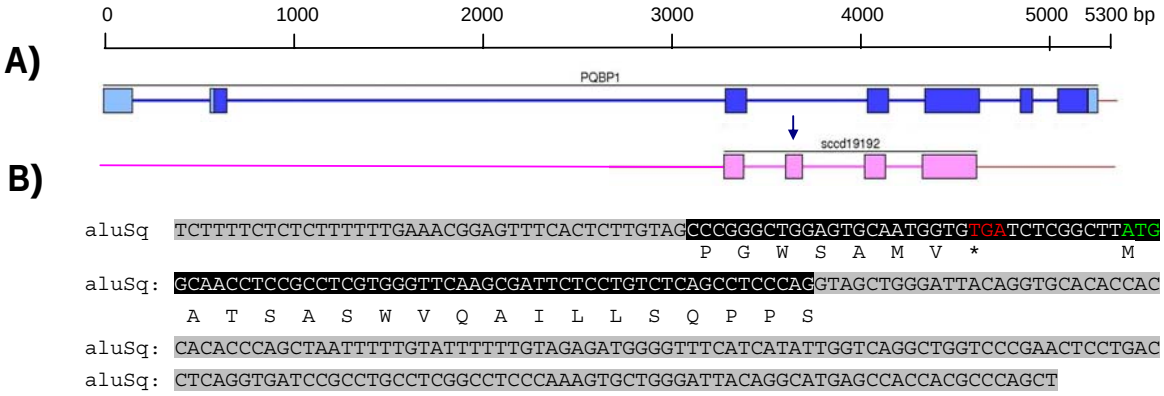


Figure 4.14 The novel exon 2a lies within an *Alu* repeat

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).

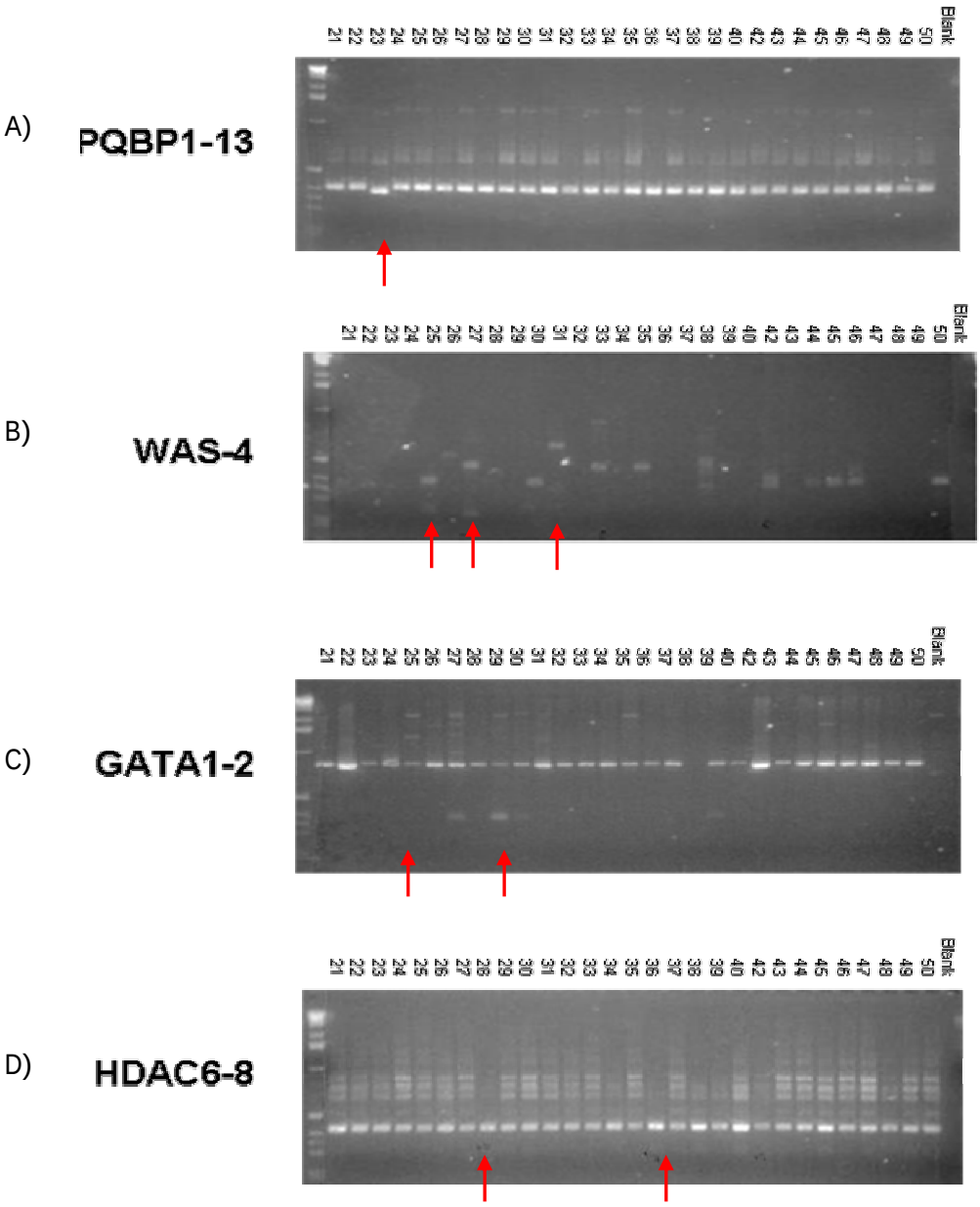


Figure 4.16 Example of cDNA screens that displayed tissue specific expression profiles

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.

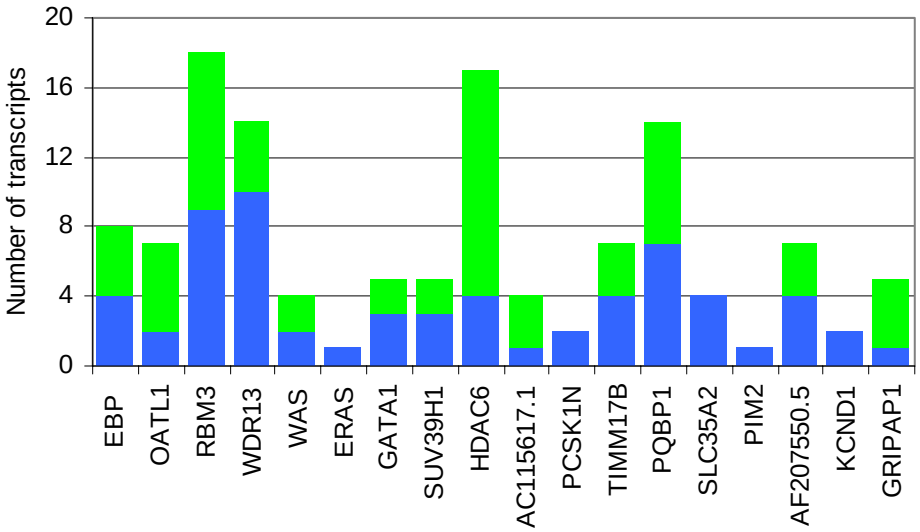


Figure 4.17 Summary of the number of novel transcripts identified by cDNA screening for 18 genes in Xp11.23
The number of transcripts identified from EST and cDNA sequences is displayed in blue while the number of transcripts identified in this study is displayed in green.

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.

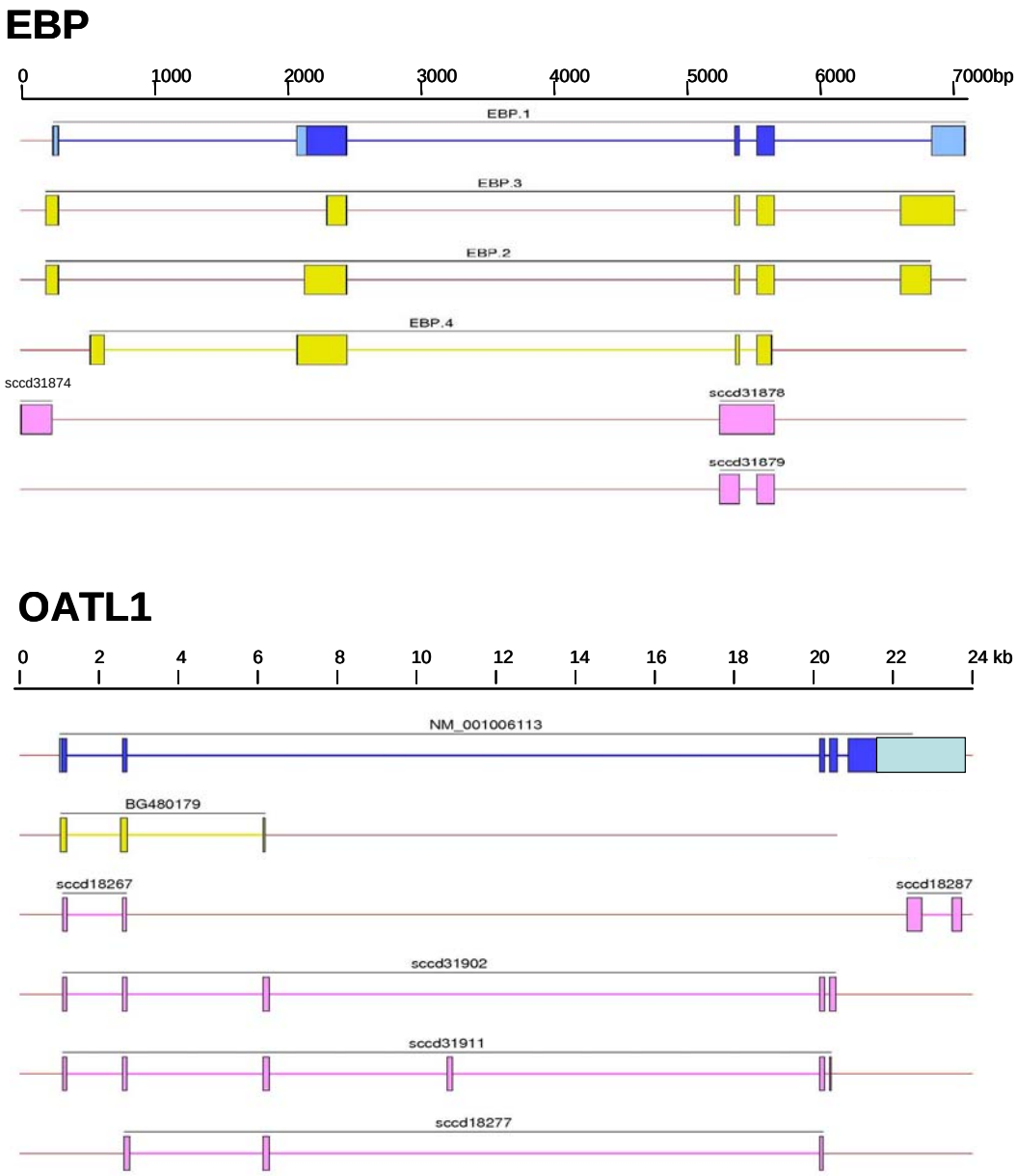


Figure 4.18 Transcript structures for genes in human Xp11.23

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.

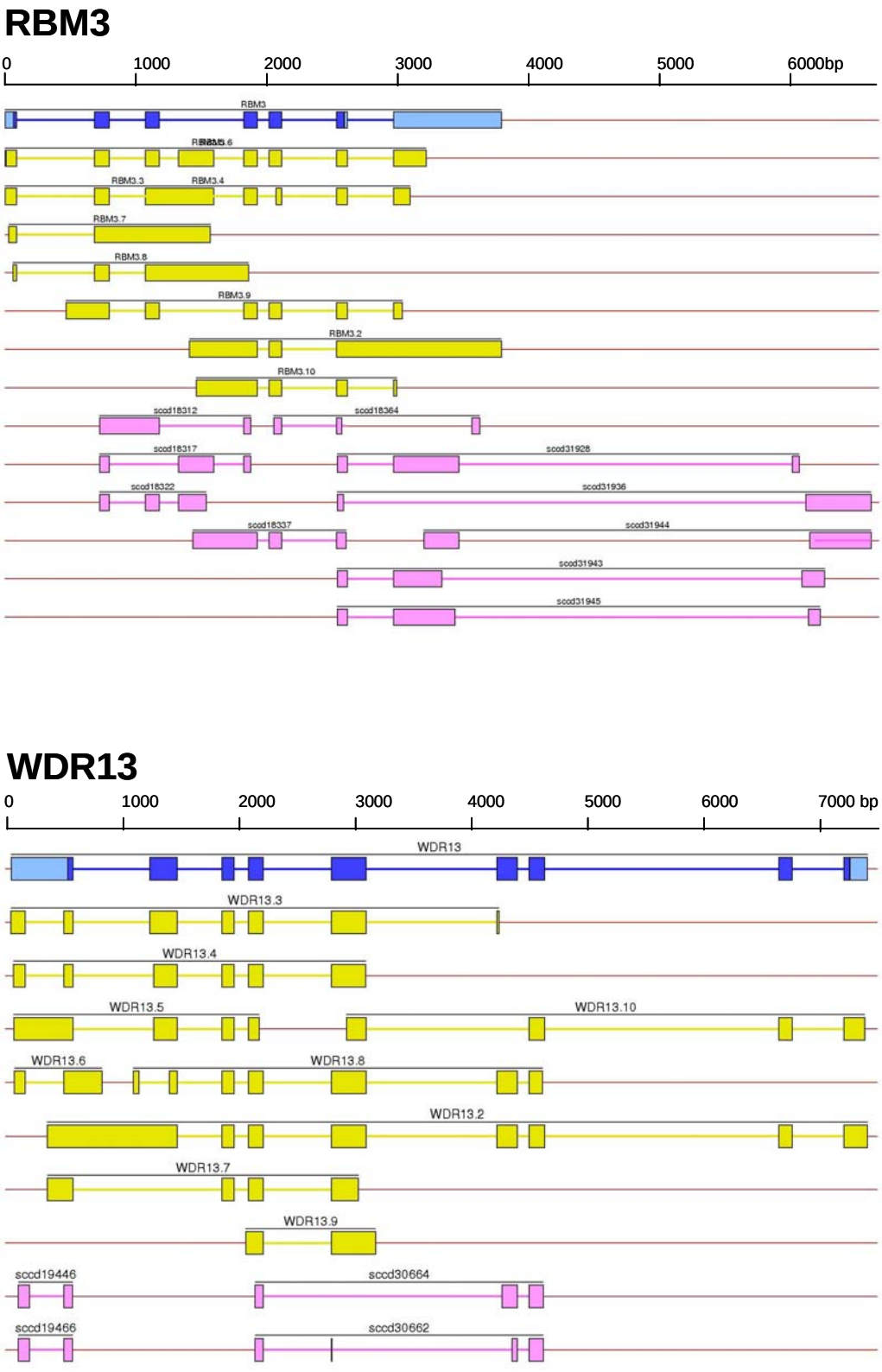


Figure 4.18 Transcript structures for genes in human Xp11.23 – continued overleaf

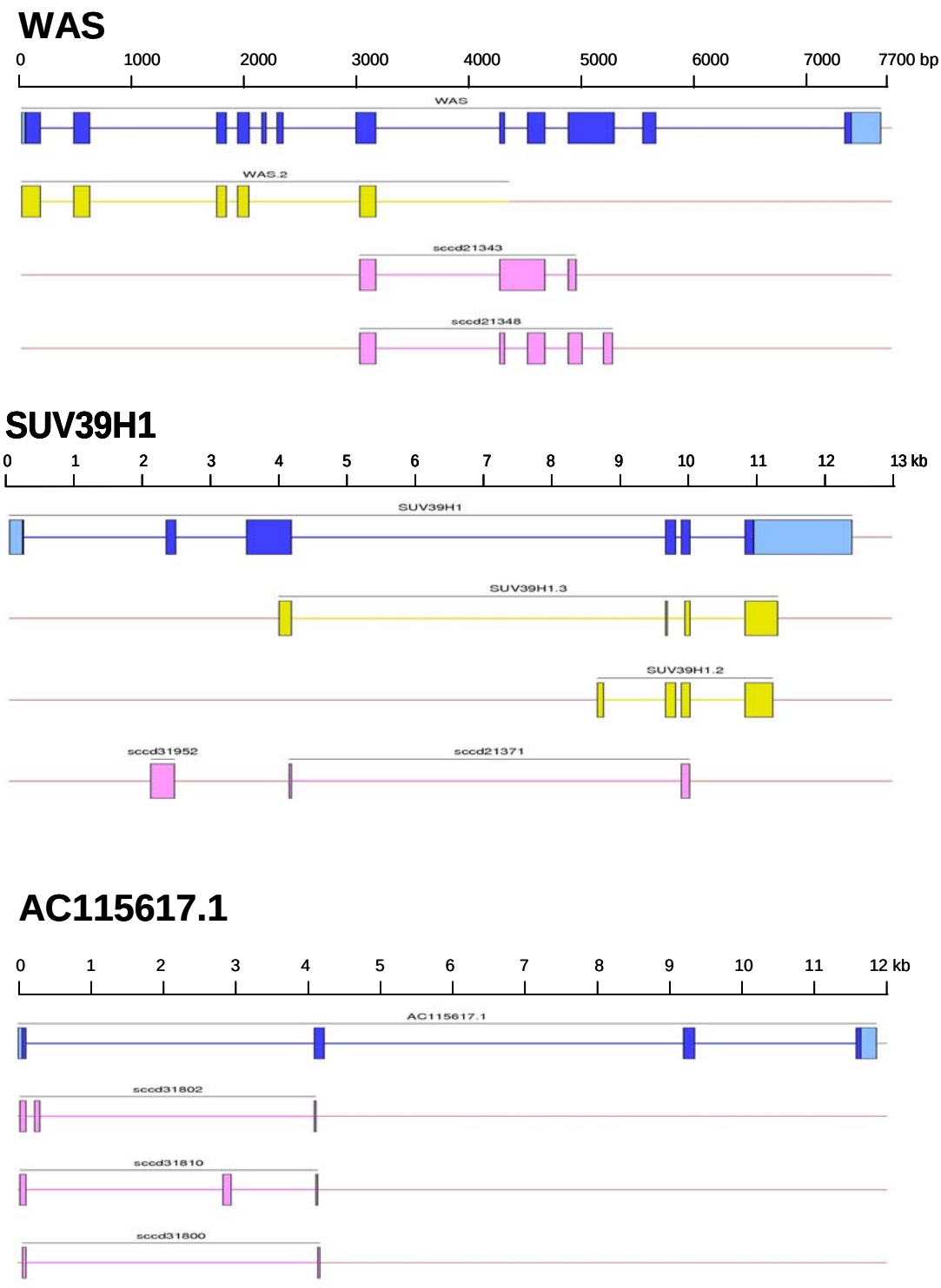


Figure 4.18 Transcript structures for genes in human Xp11.23 – continued overleaf

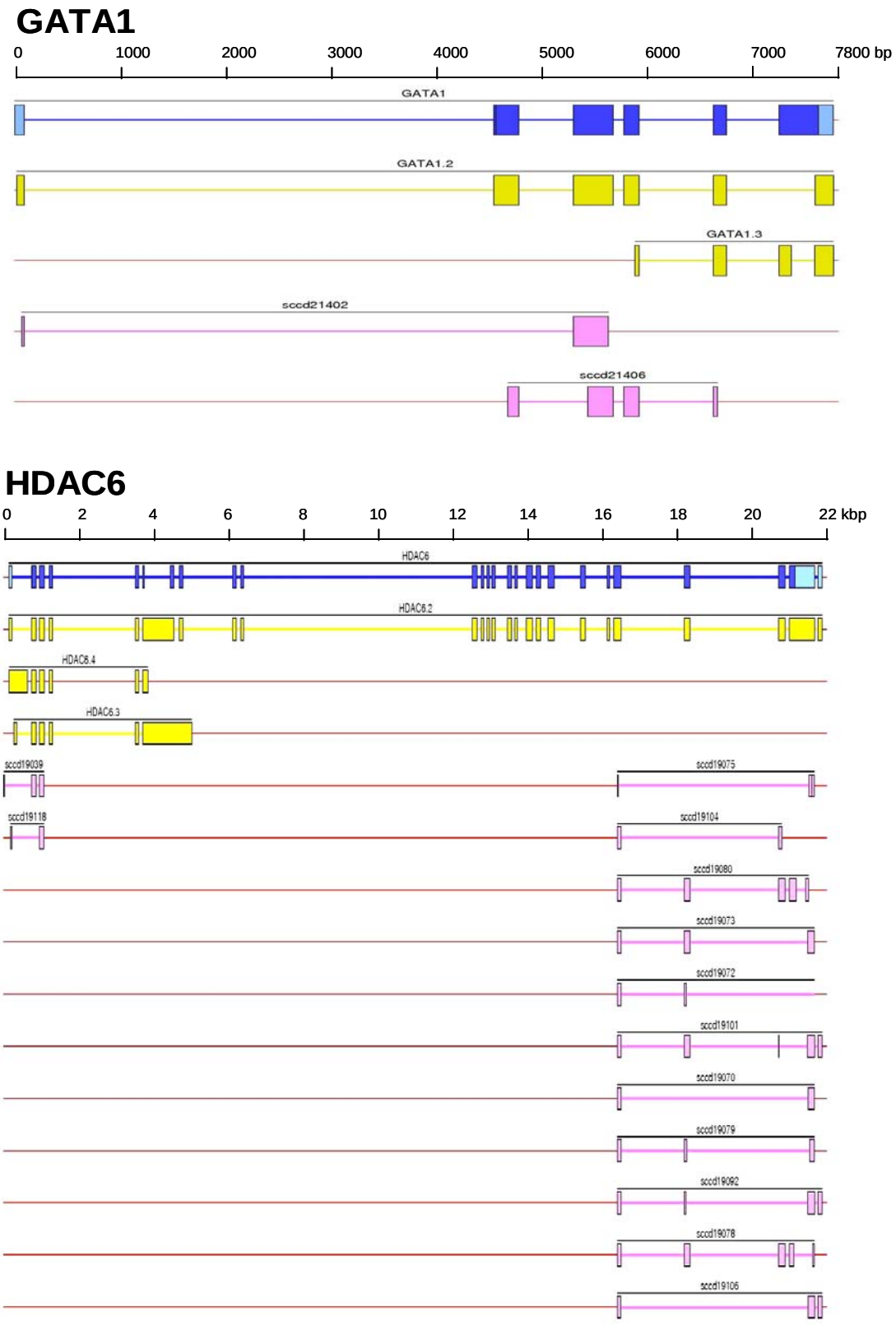


Figure 4.18 Transcript structures for genes in human Xp11.23 – continued overleaf

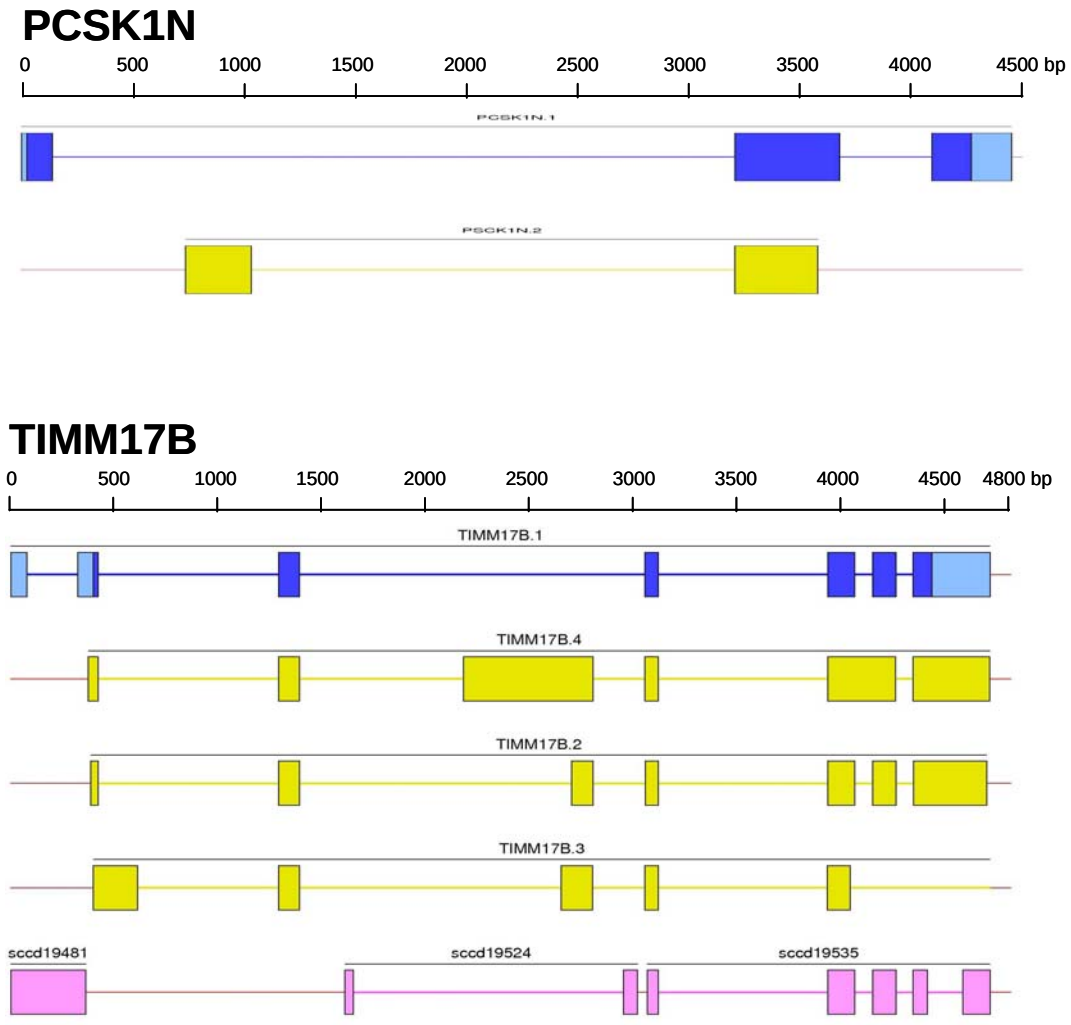


Figure 4.18 Transcript structures for genes in human Xp11.23 – continued overleaf

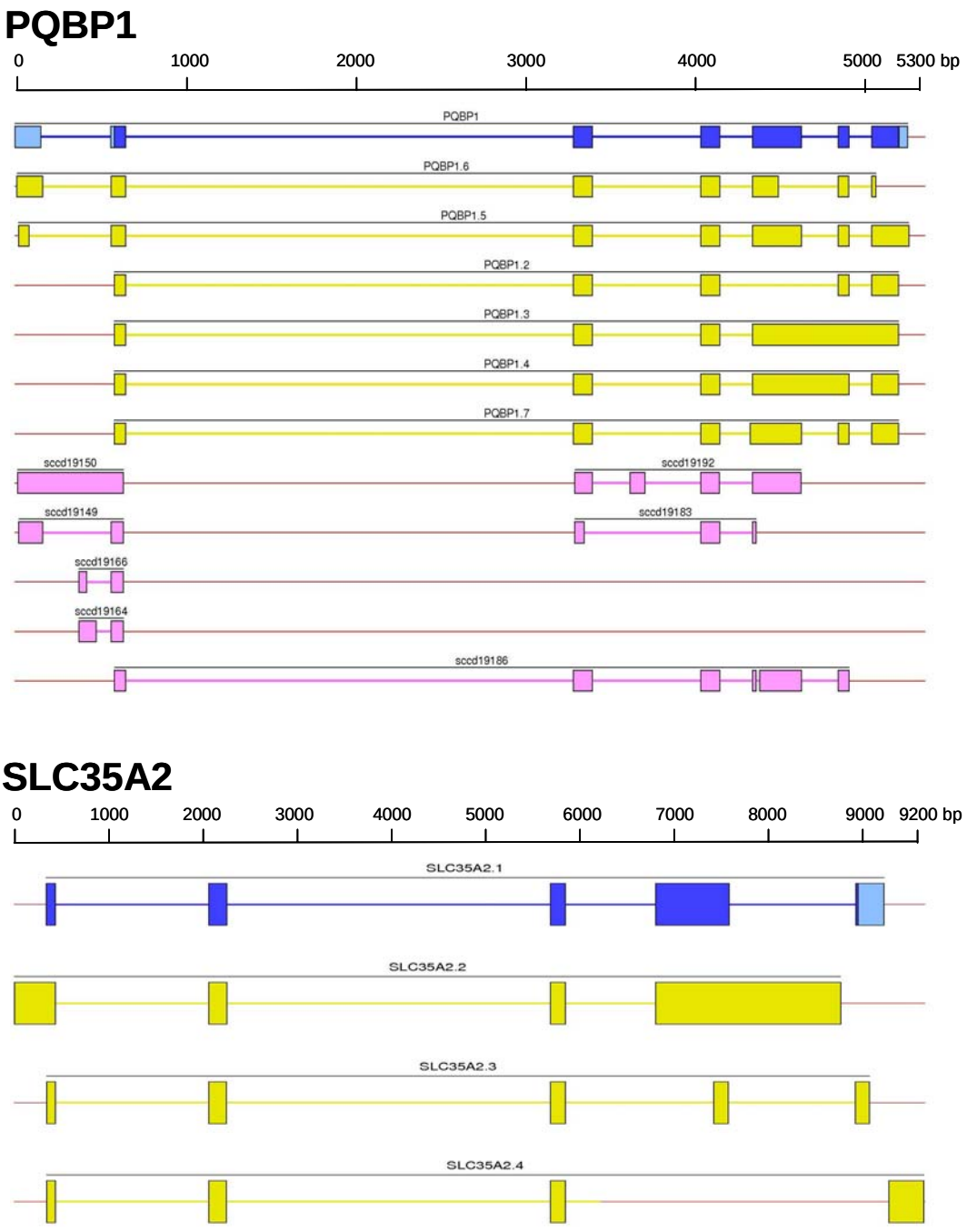


Figure 4.18 Transcript structures for gene in human Xp11.23 - continued overleaf

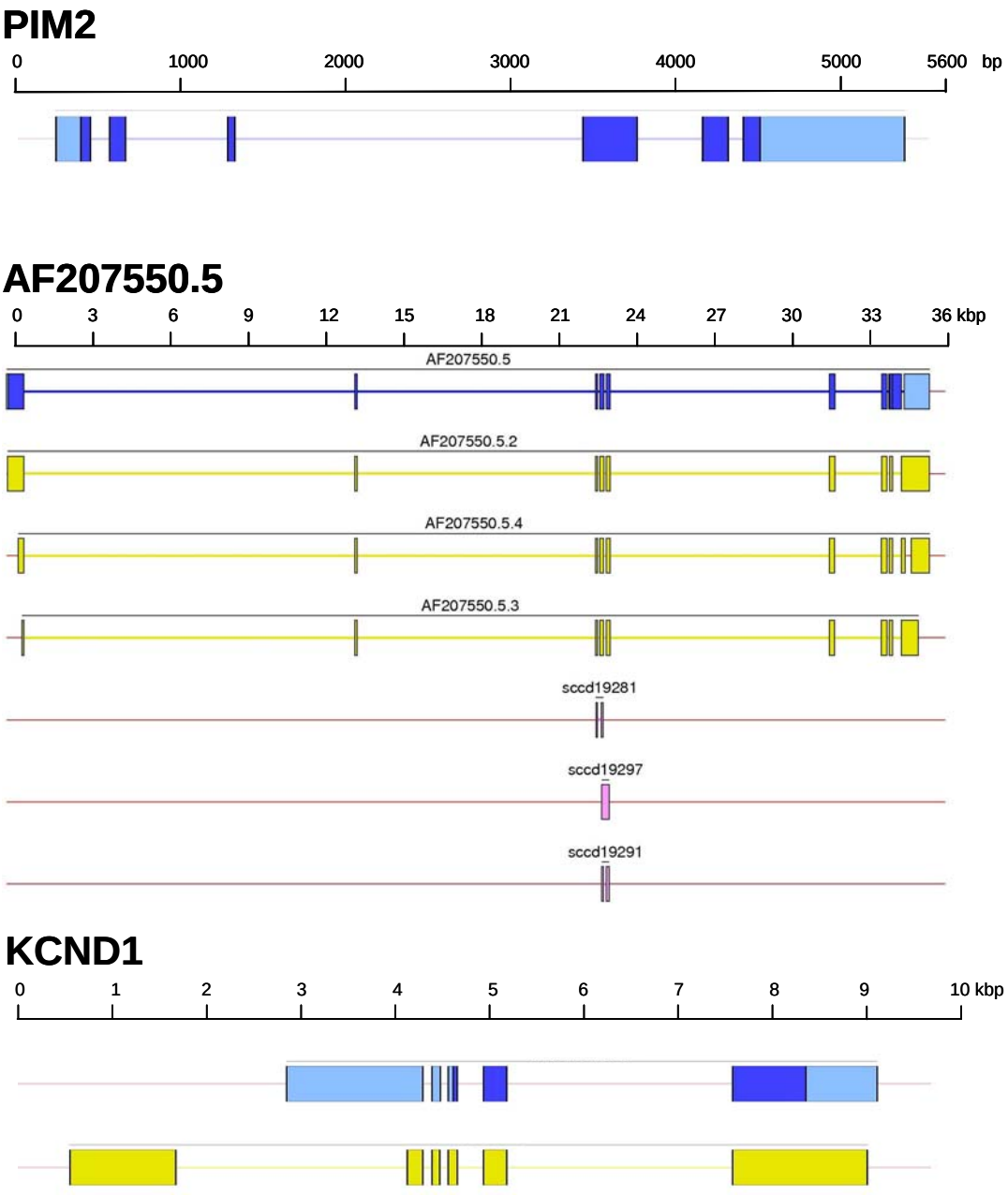


Figure 4.18 Transcript structures for genes in human Xp11.23 - continued overleaf

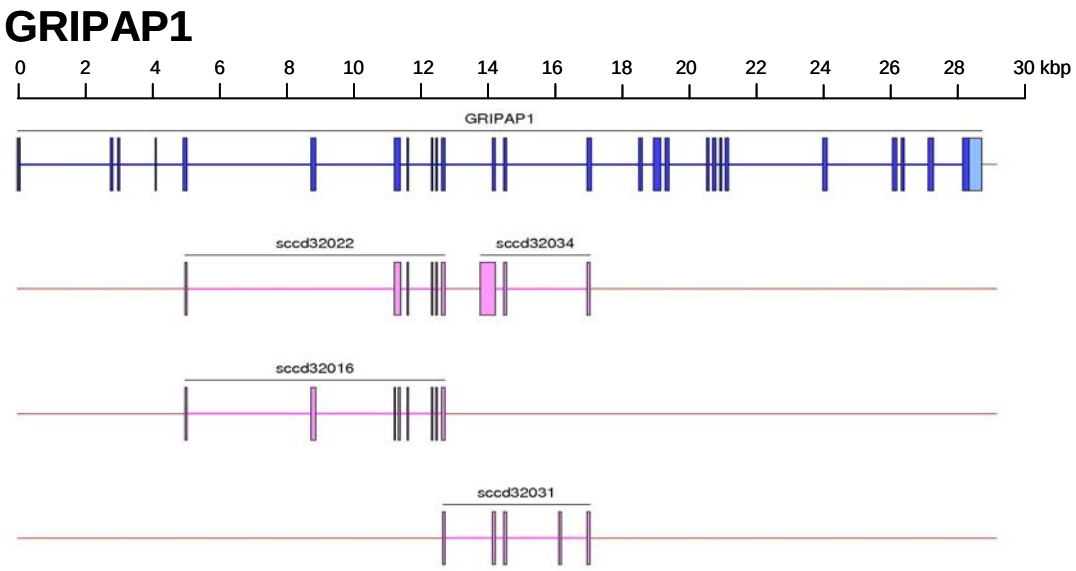


Figure 4.18 Transcript structures for genes in human Xp11.23

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.

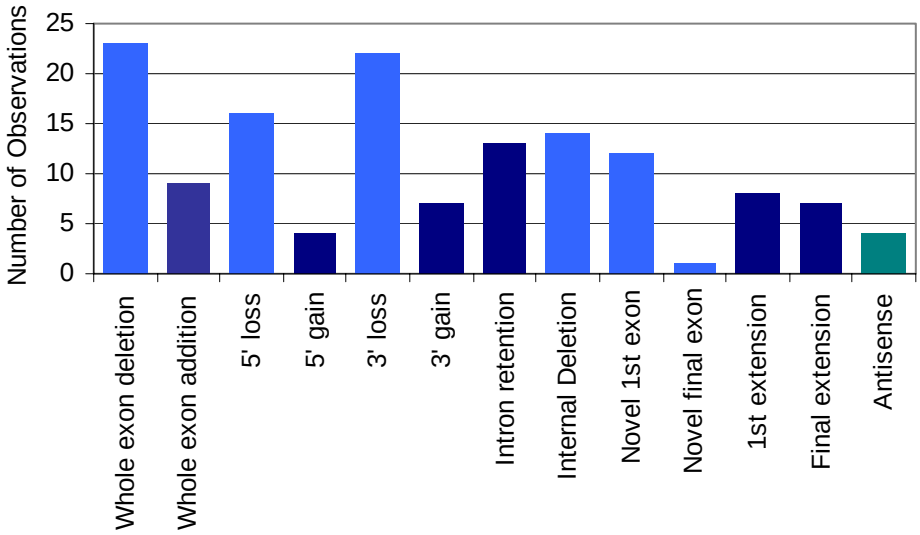


Figure 4.19 Types of alternative splicing events observed in 18 genes from human Xp11.23

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).

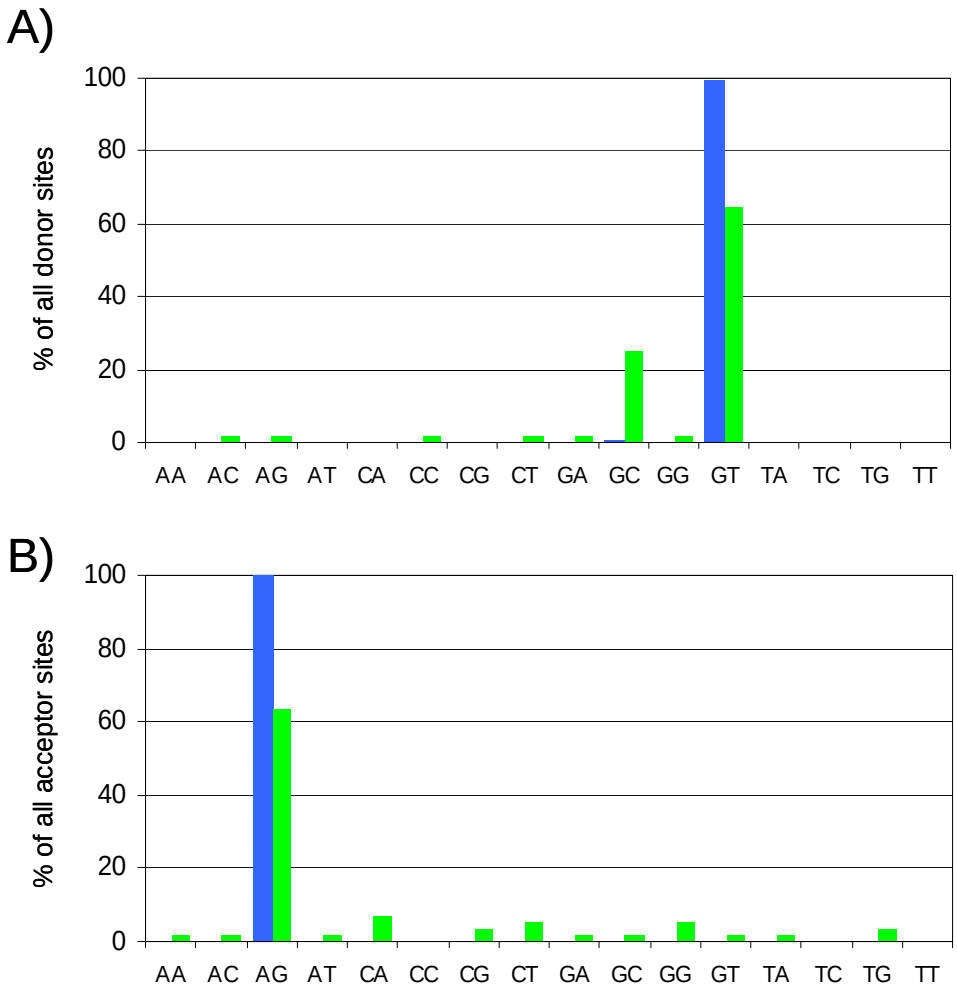


Figure 4.20 Classification of splice site sequences

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.

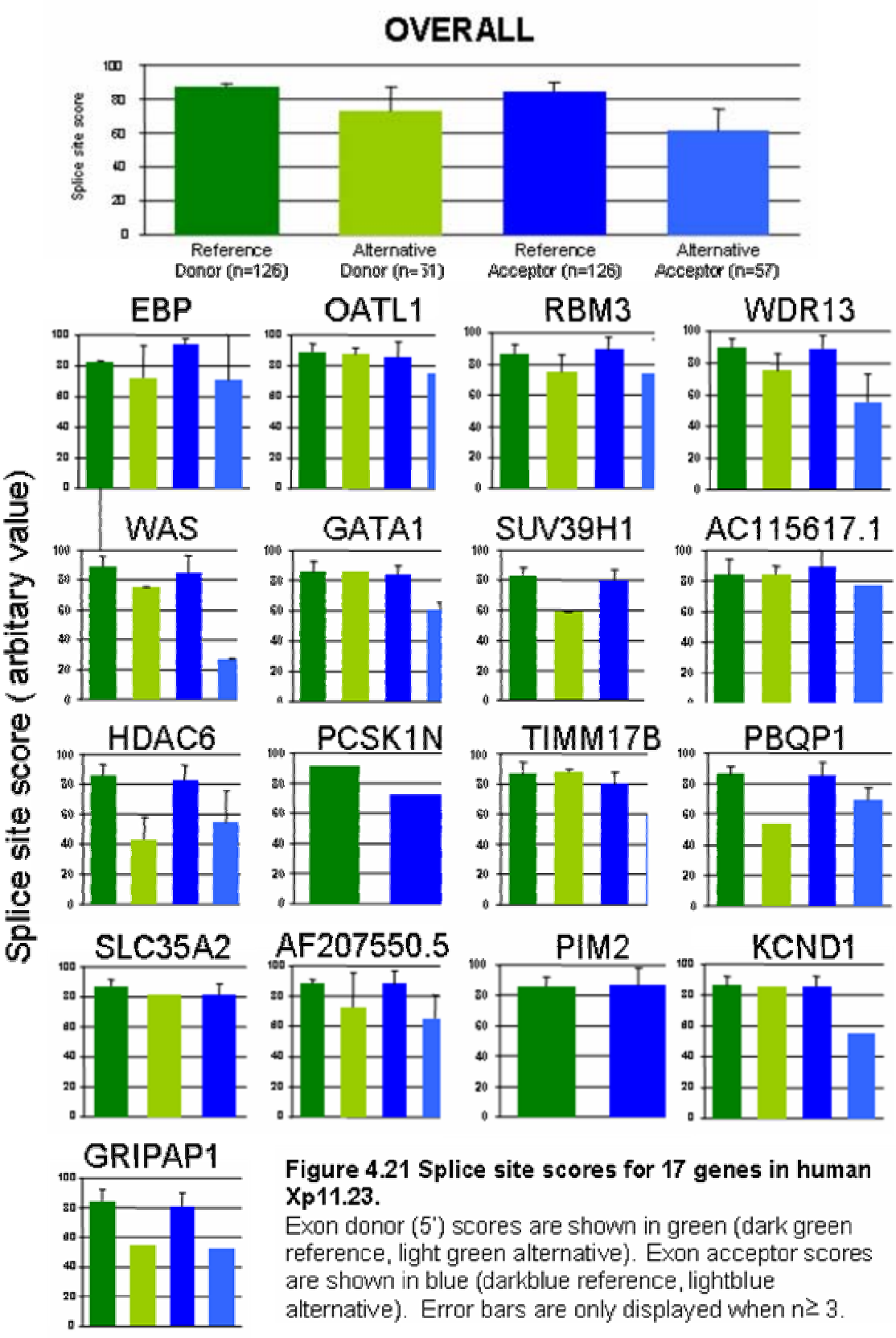


Figure 4.21 Splice site scores for 17 genes in human Xp11.23.
Exon donor (5') scores are shown in green (dark green reference, light green alternative). Exon acceptor scores are shown in blue (darkblue reference, lightblue alternative). Error bars are only displayed when n≥ 3.

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.

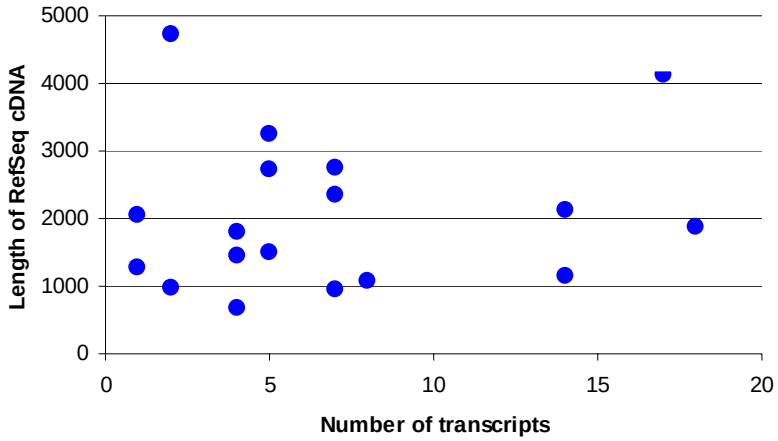
Transcript variation versus gene length

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).

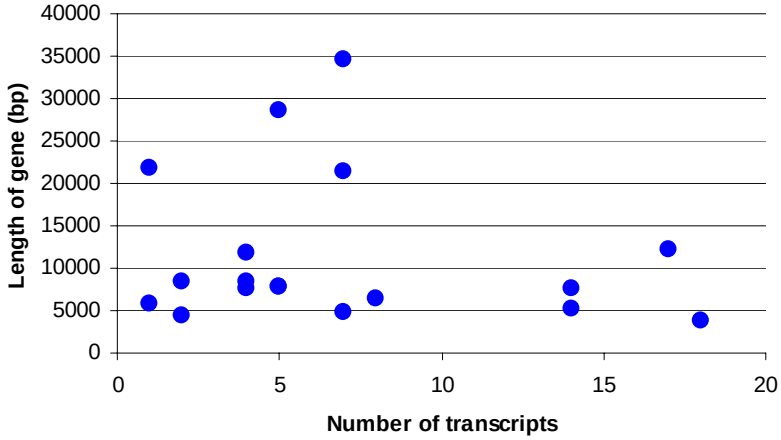
Transcript variation versus size of homologous UniGene cluster

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)

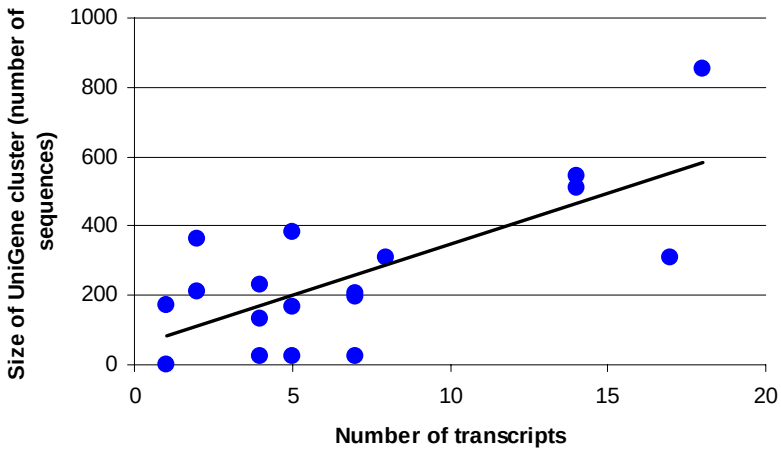


Figure 4.22 Correlation of gene features with transcript numbers

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.

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).

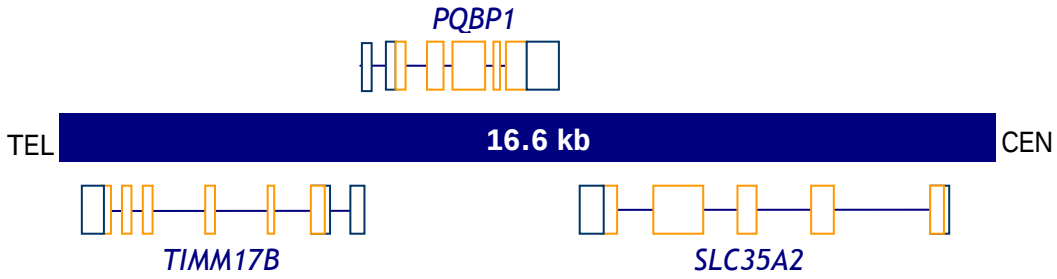


Figure 5.1 *PQBP1* **and its neighbours**
The genome sequence in the centre of the diagram is 16.6 kb long and runs from the telomere to the centromere. Transcript sequences above the genome sequence are transcribed in the forward direction while transcripts beneath the genome sequence are transcribed 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.

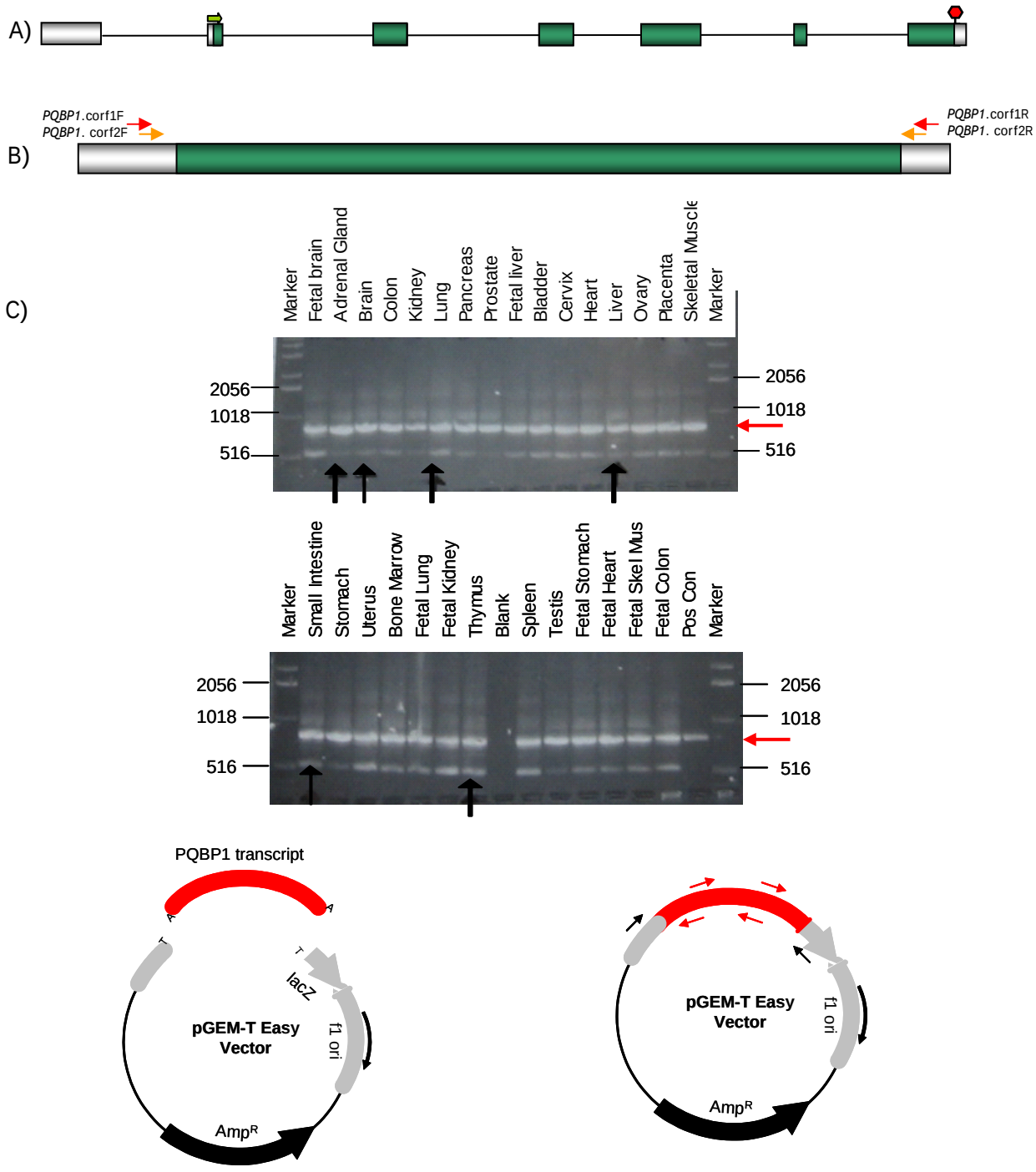


Figure 5.2 Overview of the ORF cloning protocol

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

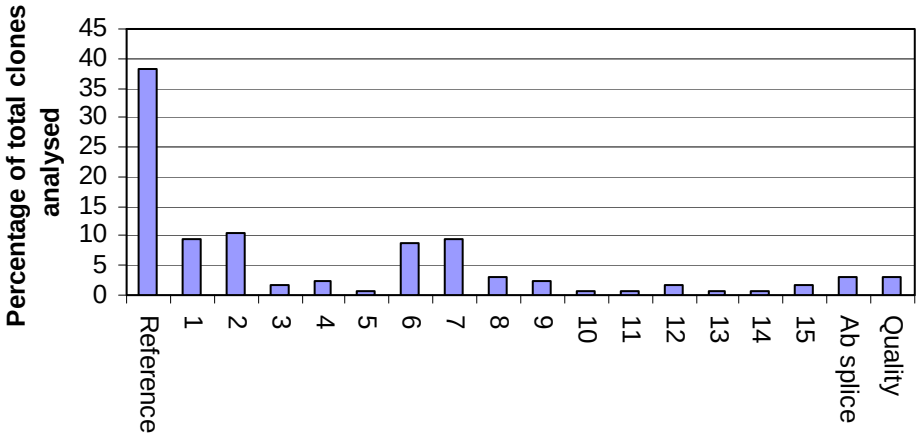


Figure 5.3 Frequency of clones representing different *PQBP1* variants

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

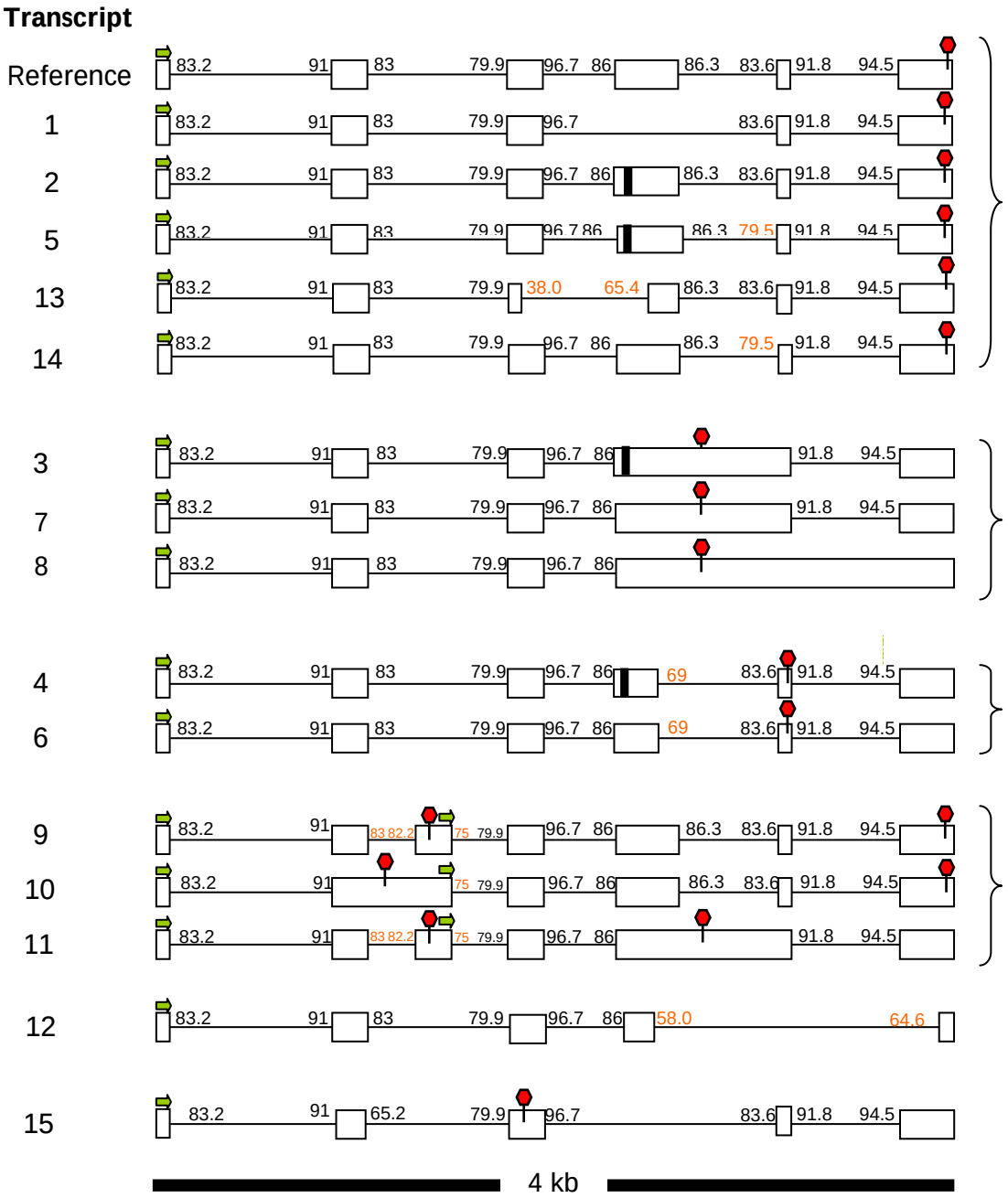


Figure 5.4 Exon/intron structures of *PQBP1* alternative transcripts

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.

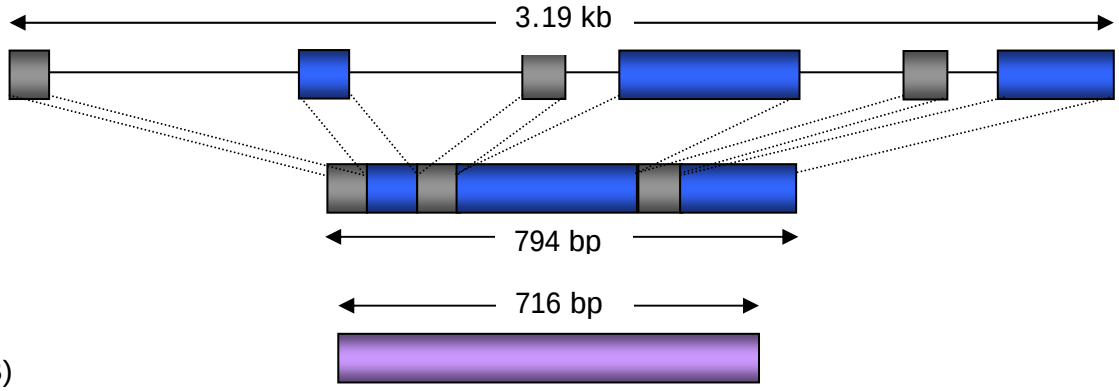
It is therefore suggested that the mouse, rat, dog and opossum have one functional copy and one retroposed non-functional copy of *PQBP1*. An example of the retroposed copy of *PQBP1* identified in the dog is displayed in Figure 5.5.

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---------
```

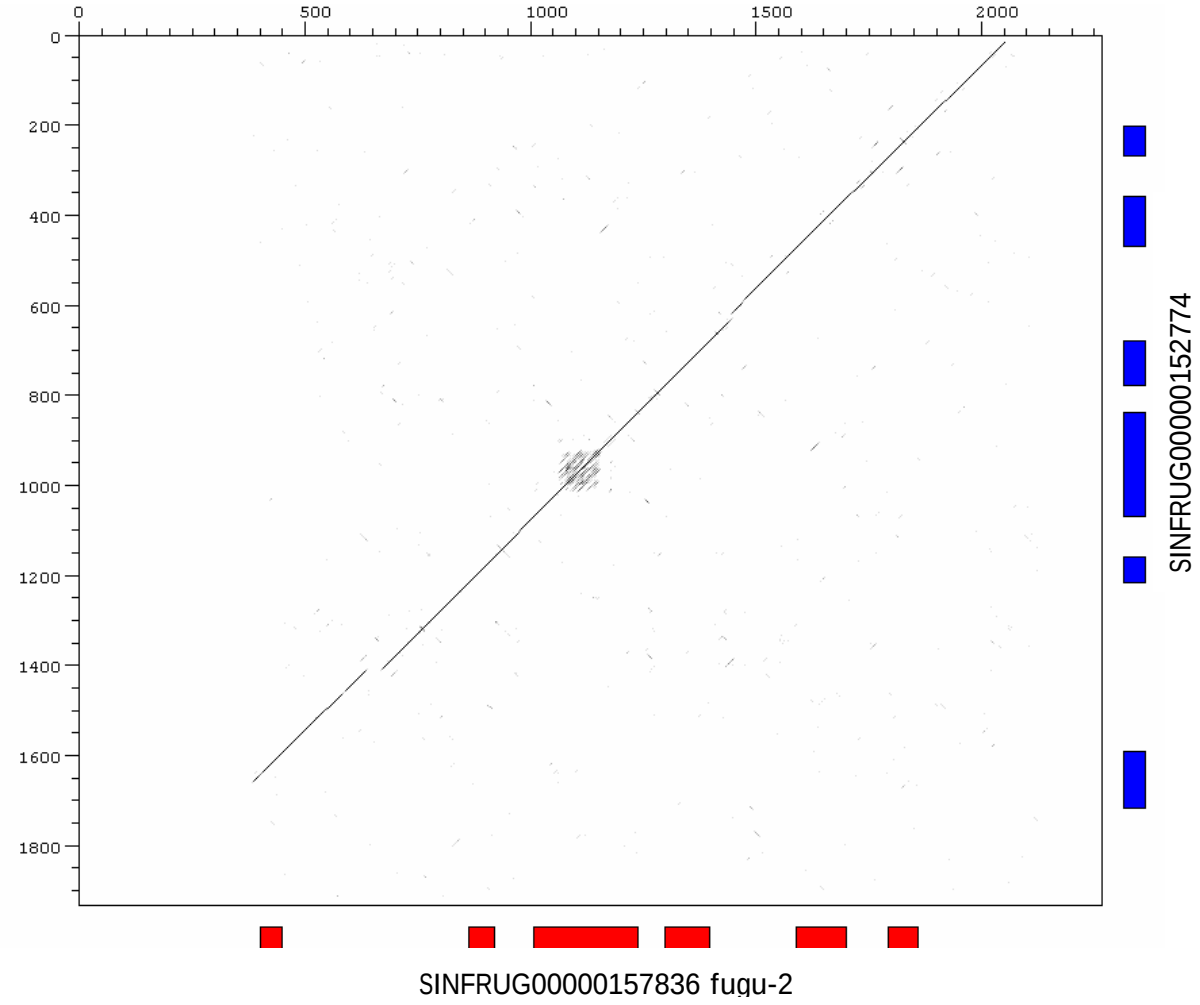


Figure 5.6 Potential duplication of *PQBP1* locus in *F. rubripes*

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.

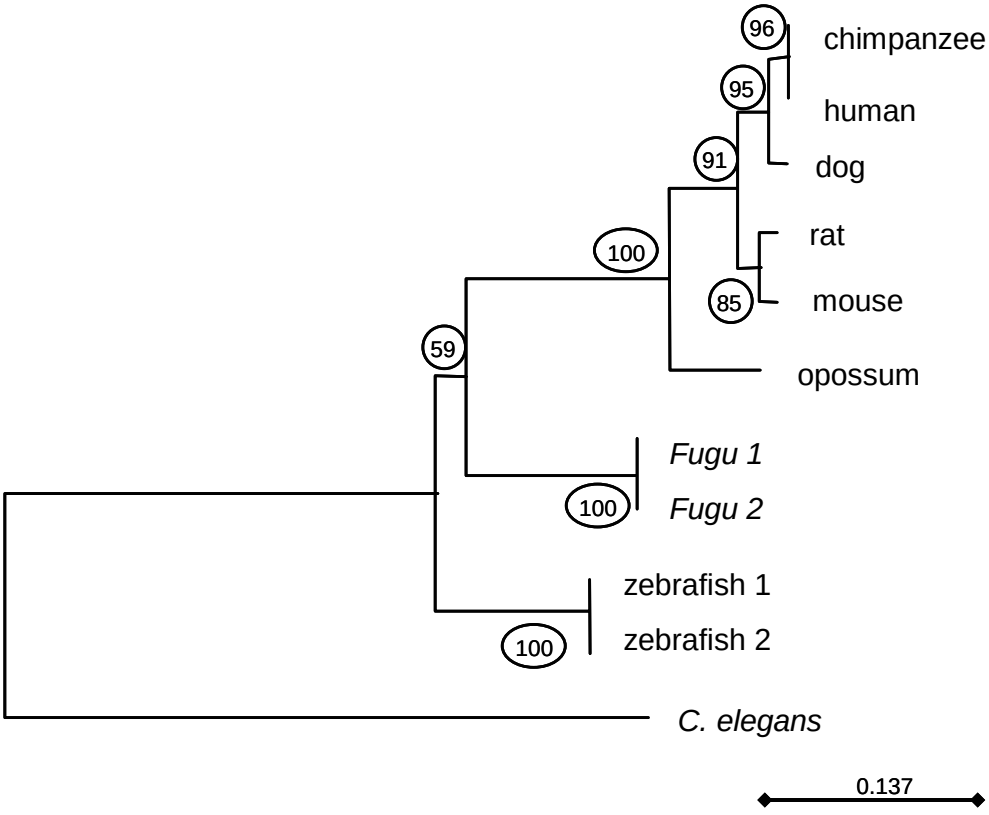


Figure 5.7 Phylogenetic tree of *PQBP1* peptides

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.

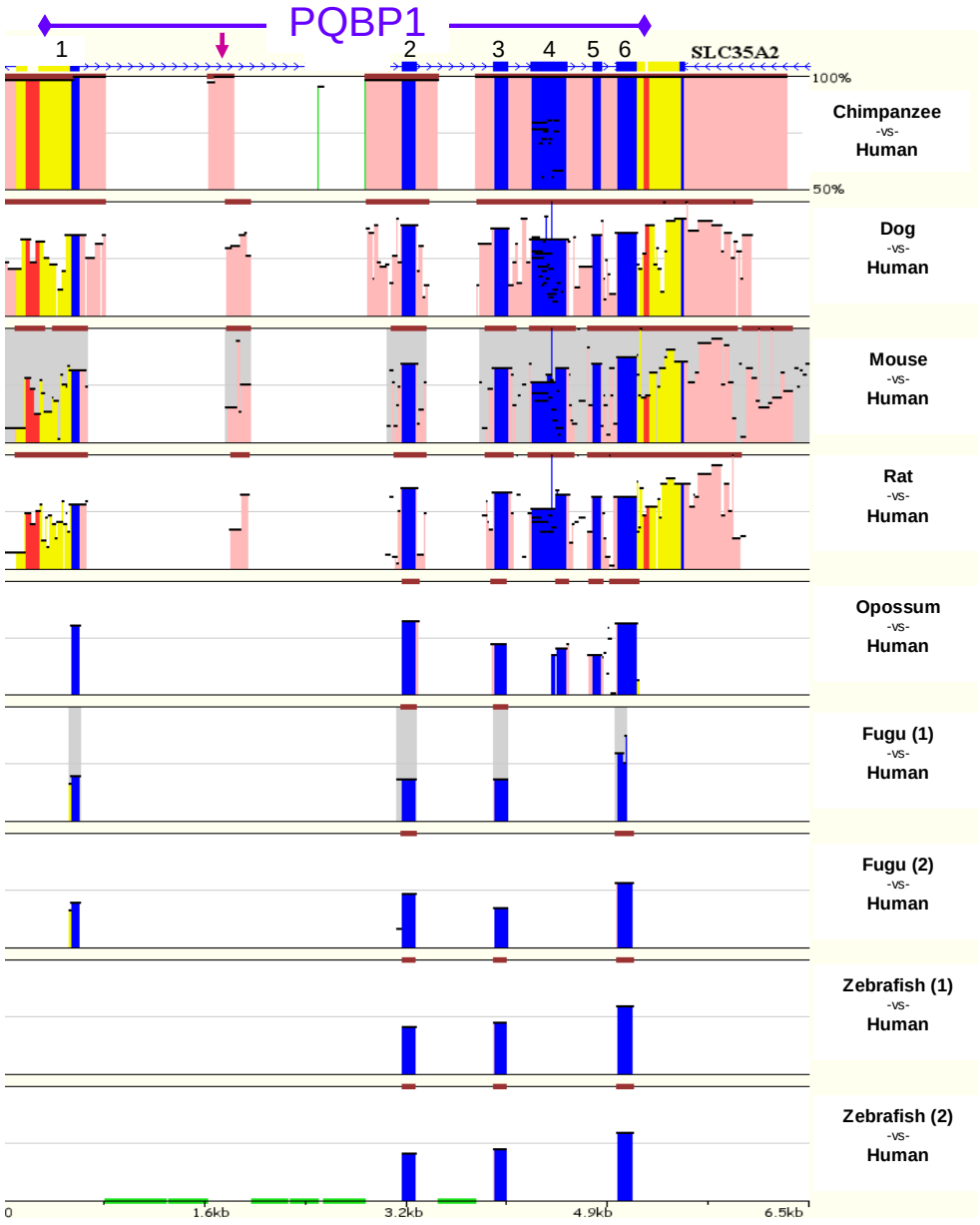


Figure 5.8 Comparative Genome Analysis of the *PQBP1* locus

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.

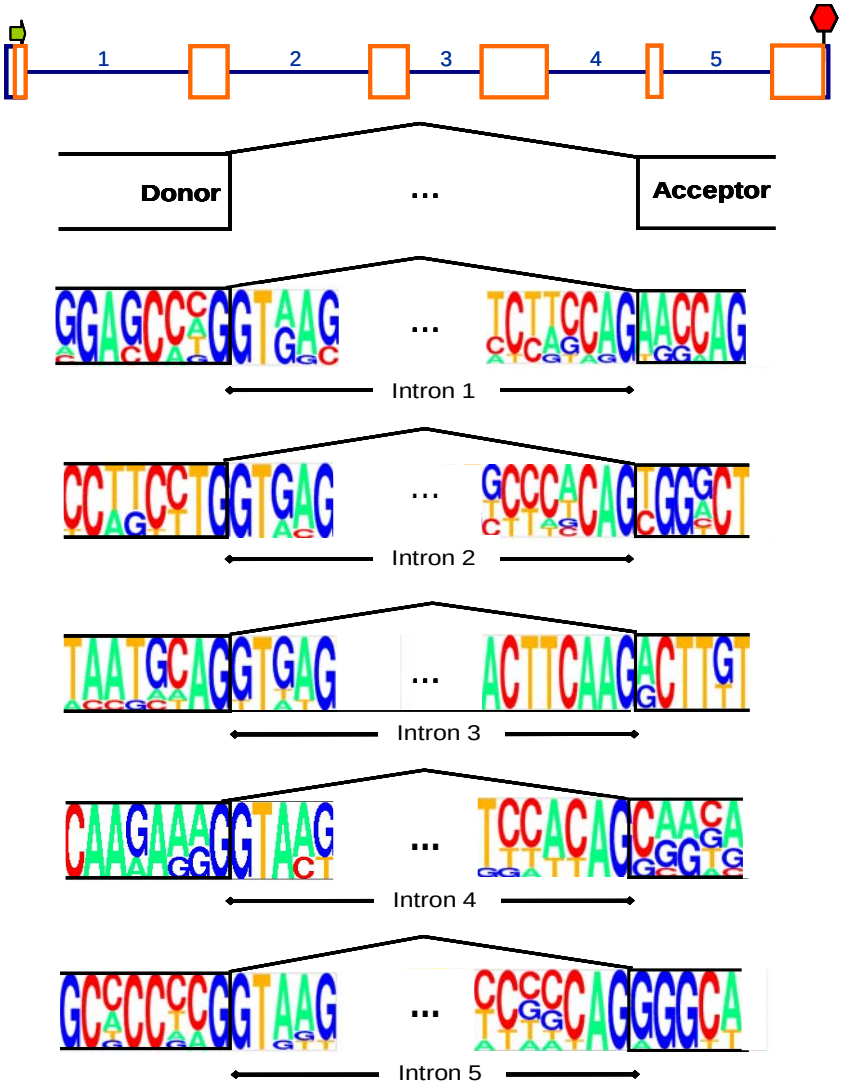


Figure 5.9 Sequence variations around exon/intron junctions in different species

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.

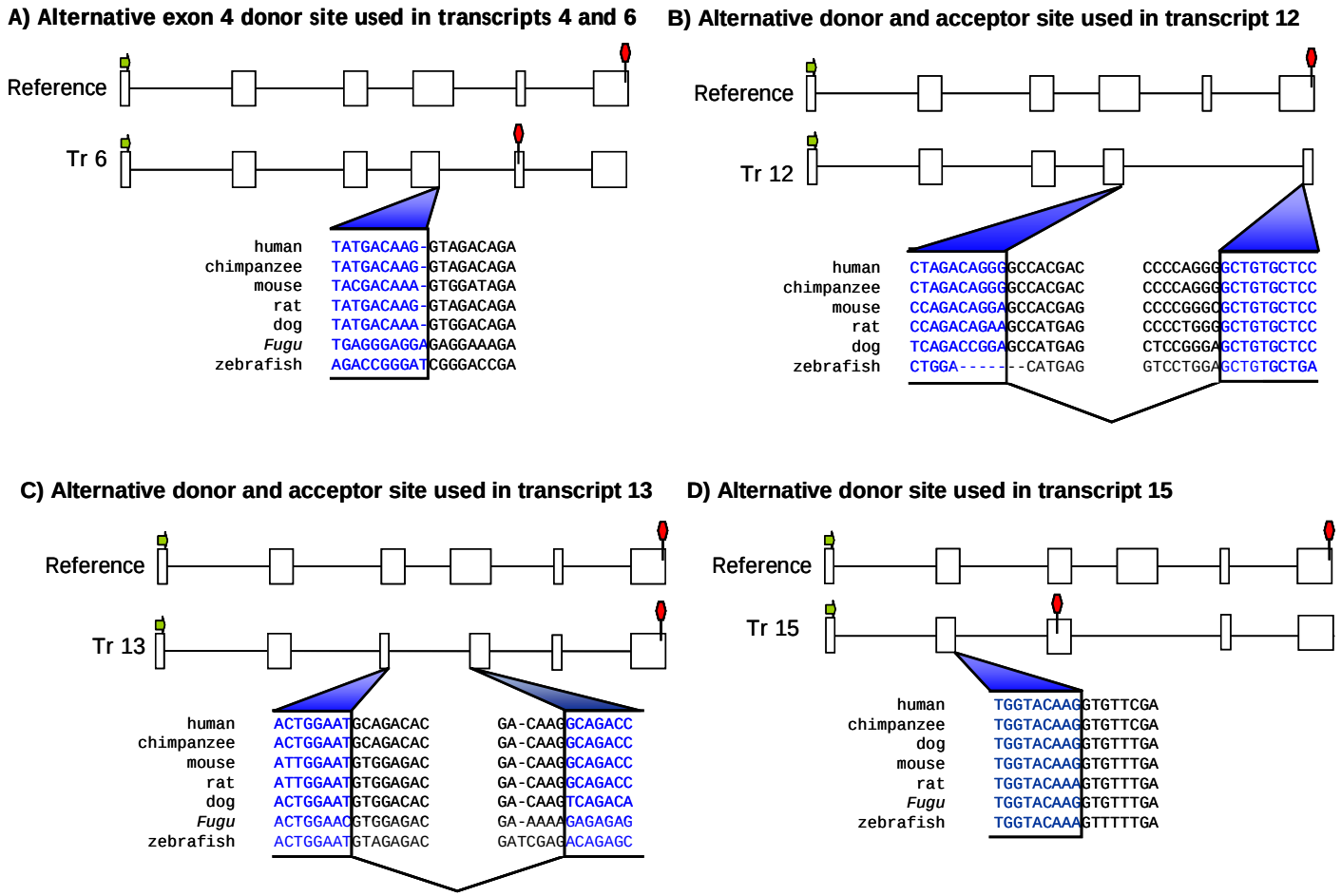


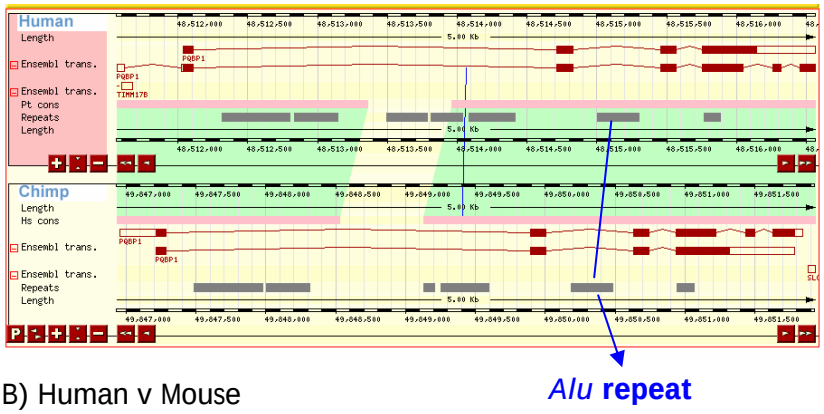
Figure 5.10 Multiple sequence alignments of splice sites used in alternative transcripts

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

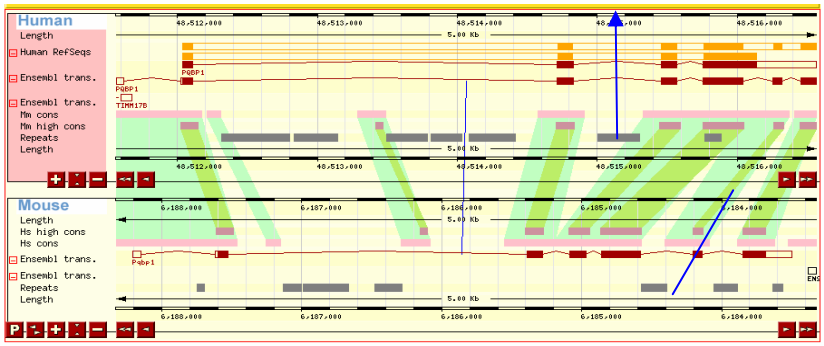


Figure 5.11 Global alignment of genome sequences containing the gene *PQBP1* using *MultiContigView* at Ensembl

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.

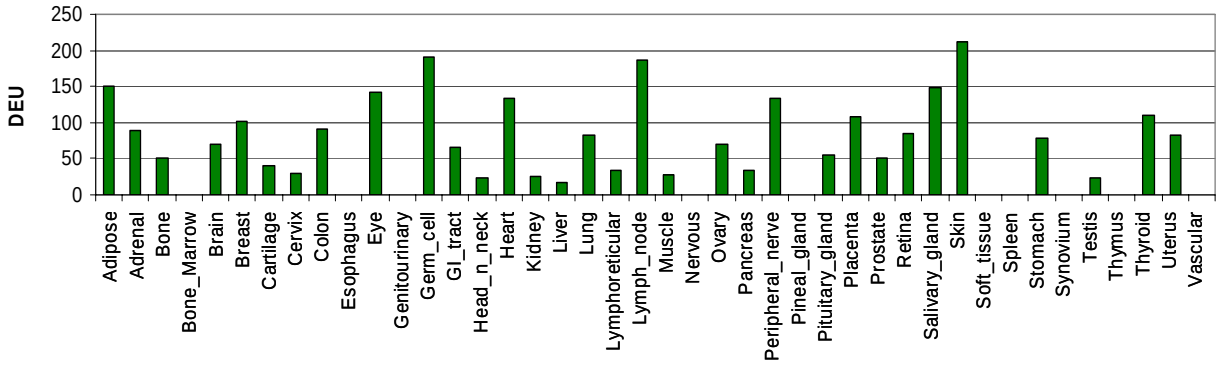


Figure 5.12 EST expression profile for *PQBP1*

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.

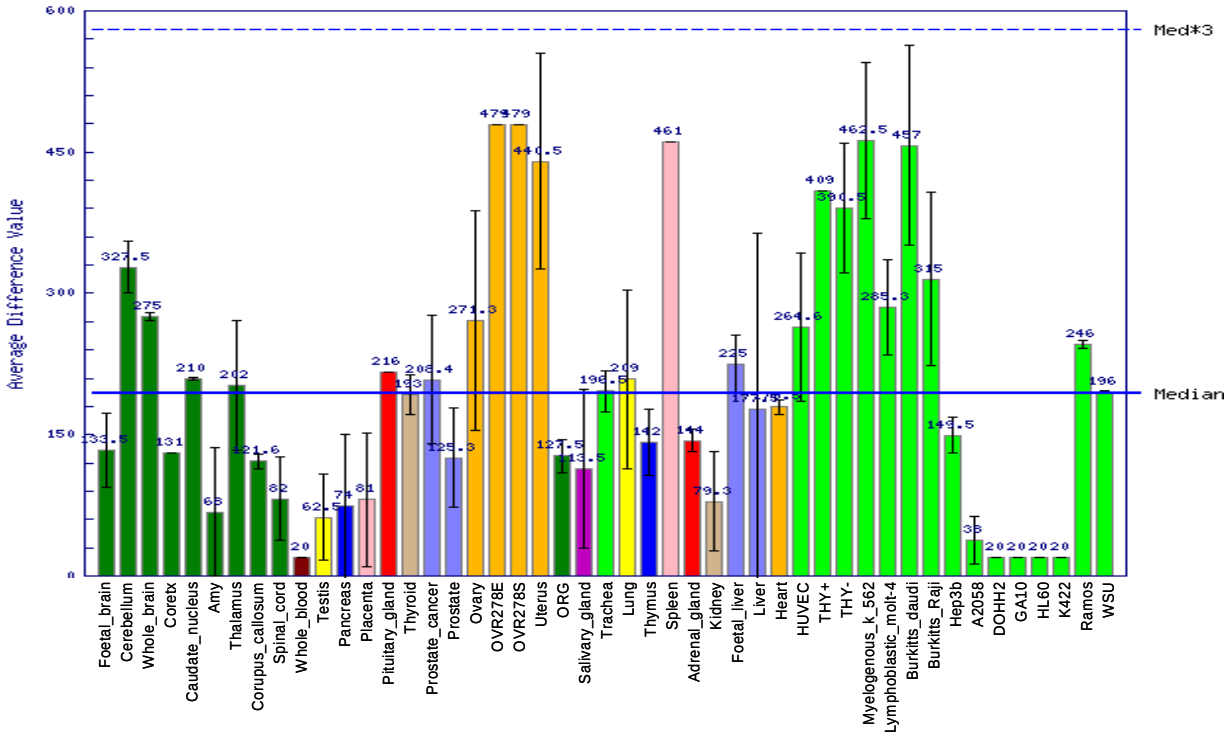


Figure 5.13 Expression profile of *PQBP1* as extracted from the Gene Expression Atlas

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

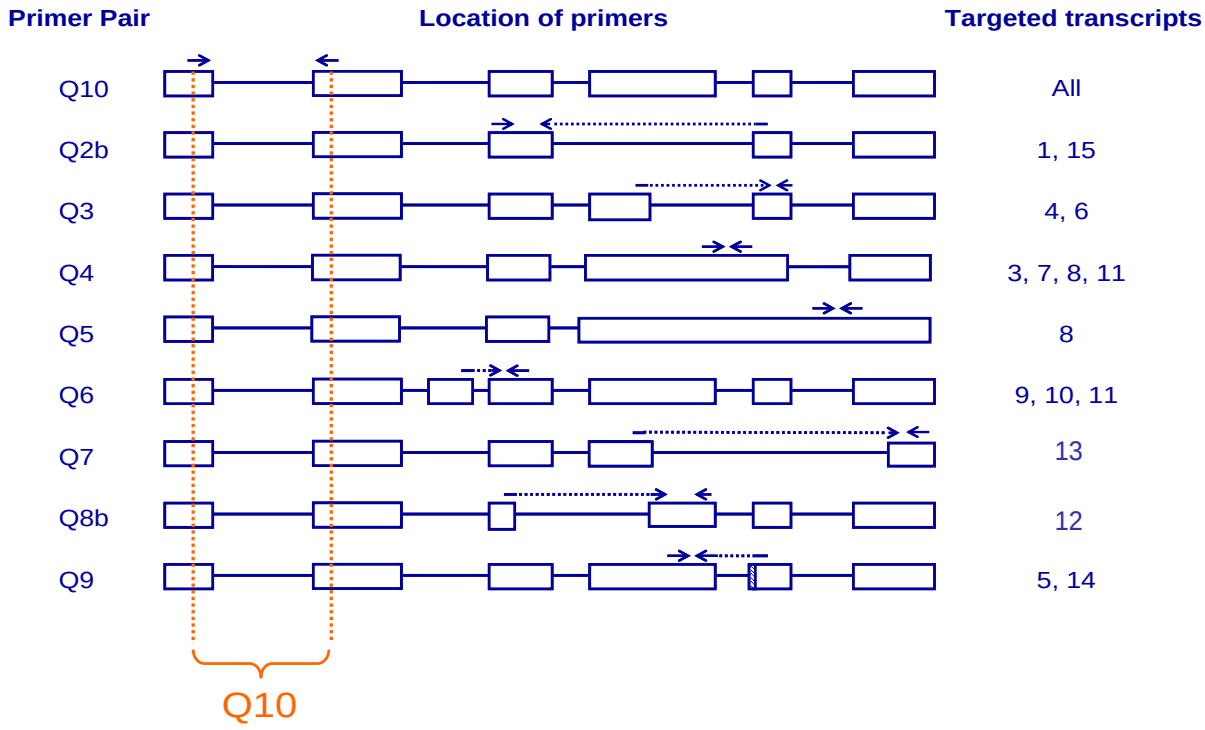
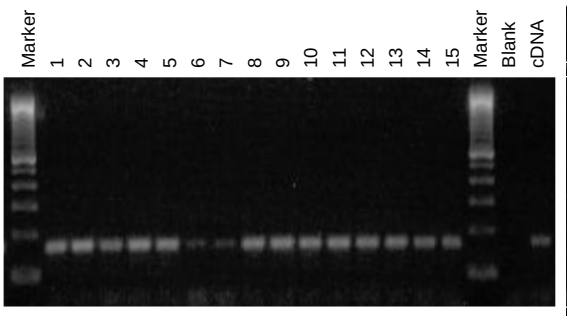


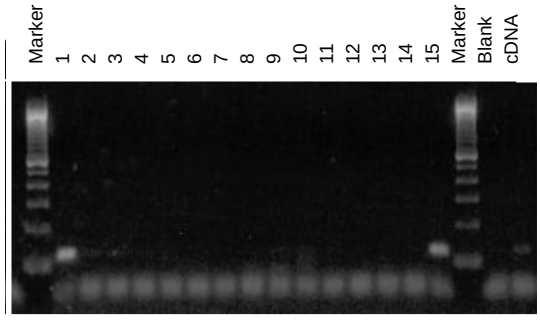
Figure 5.14 Location of primers used to determine the abundance of *PQBP1* alternative transcripts.

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).

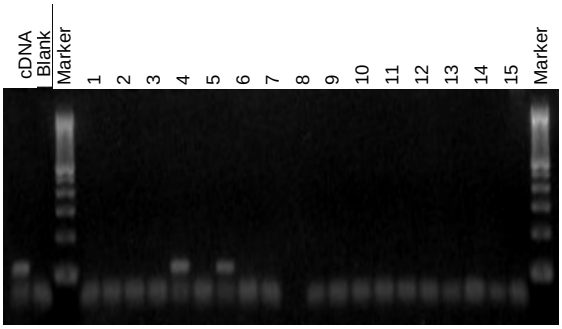
A) Primers *PQBP1*-Q10 (AII)



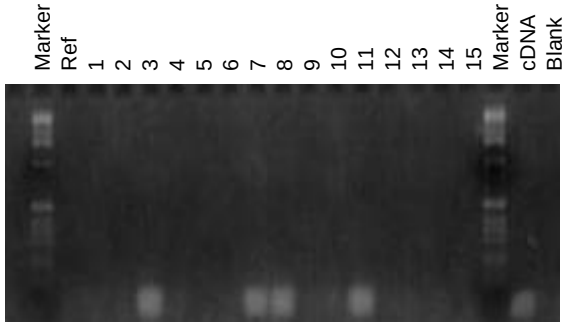
B) Primers *PQBP1*-Q2b (1, 15)



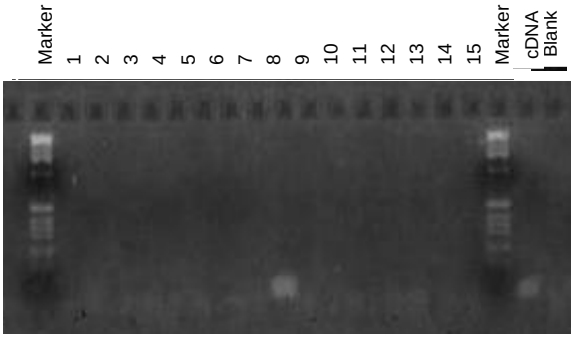
C) Primers *PQBP1*-Q3 (4, 6)



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



E) Primers *PQBP1*-Q5 (8)



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

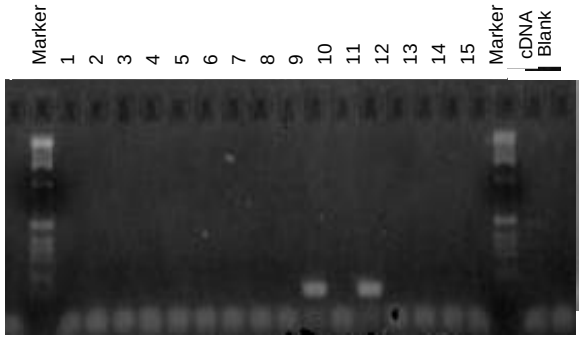
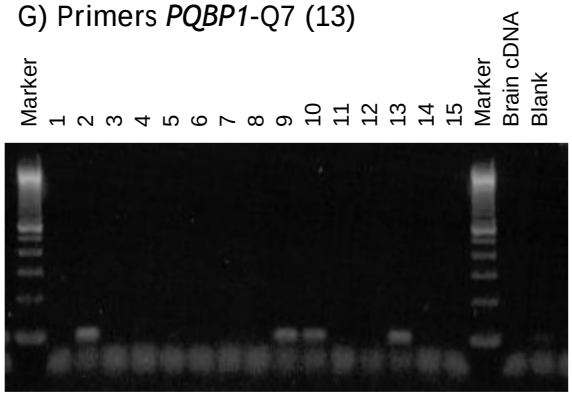


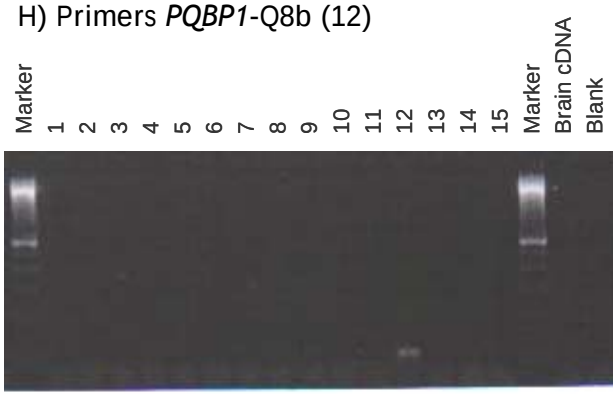
Figure 5.15 Specificity of *PQBP1* alternative transcript primers.

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)

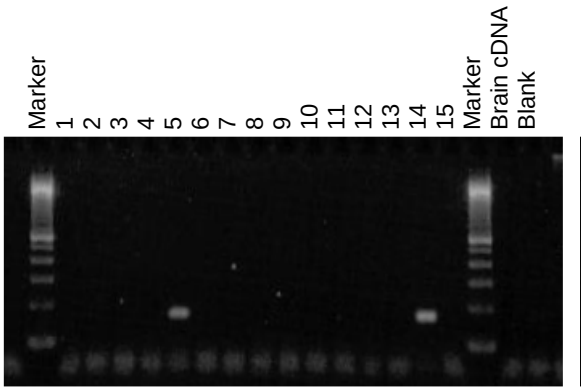


Figure 5.15 continued

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.

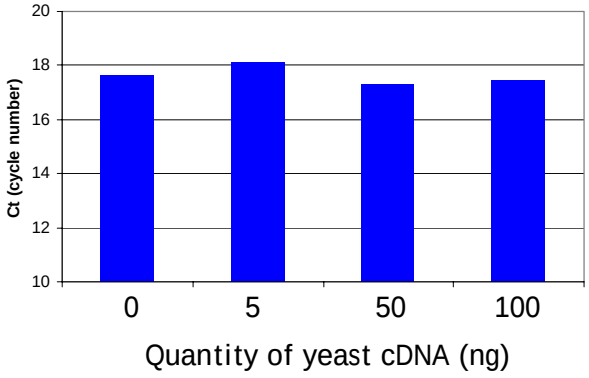


Figure 5.16 Effect of yeast cDNA on real-time PCR amplification efficiency of *PQBP1* transcripts

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.

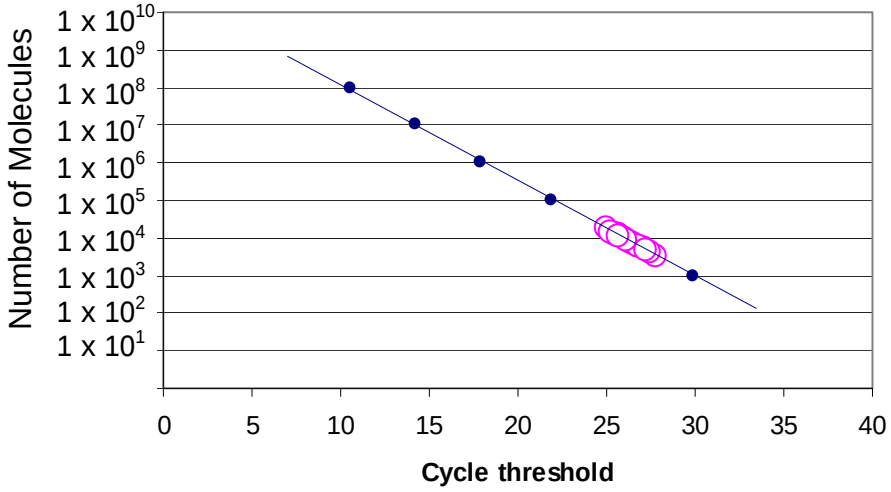


Figure 5.17 Quantification of *PQBP1* alternative transcripts by real-time PCR.
Standard curve for primer pairs *PQBP1*.2b is displayed in blue, while the C_T value of cDNA samples are shown by pink open circles.

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.

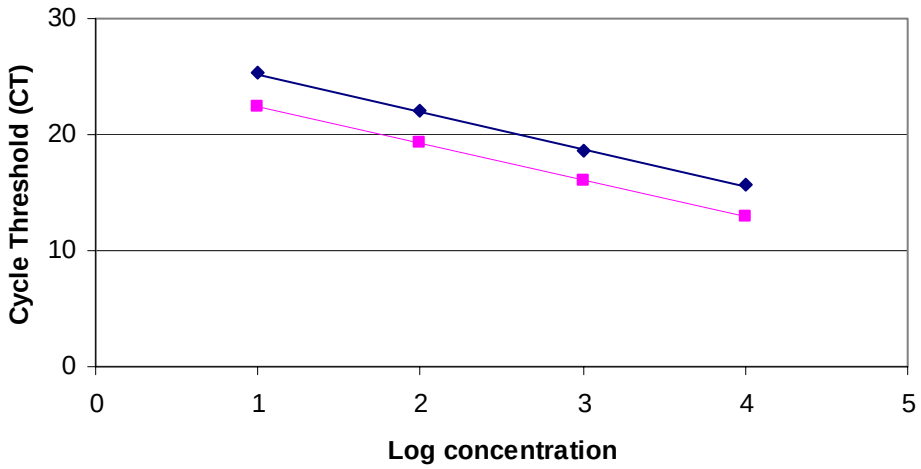


Figure 5.18 Comparison of amplification efficiencies for the primer pairs GAPDH and *PQBP1*.Q10.
Human brain cDNA was serially diluted between the range of 1 and 0.001 (100 ng to 100 pg of cDNA per reaction). The cycle value at which amplification was measured (C_T) and is displayed in pink for the GAPDH 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).

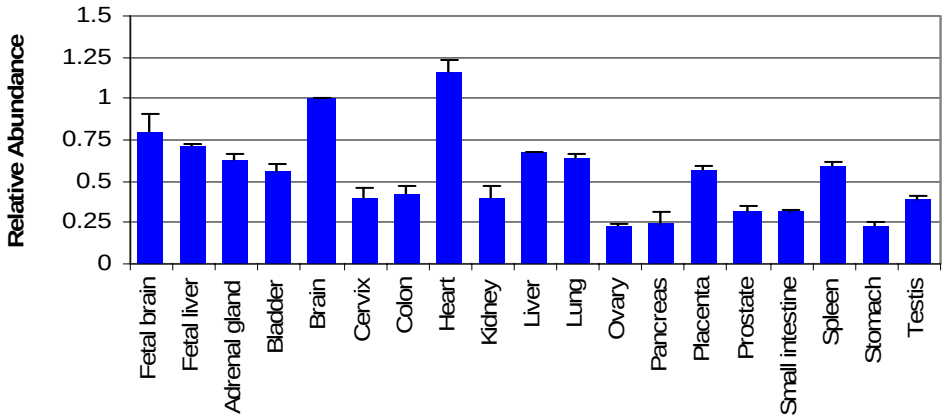


Figure 5.19 Relative abundance of *PQBP1* expression.
The relative abundance of *PQBP1* transcripts was determined for 19 different human tissues. Values are expressed relative to those in the brain and are normalised to the housekeeping gene, GAPDH. Values displayed are the results from duplicate experiments.

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.

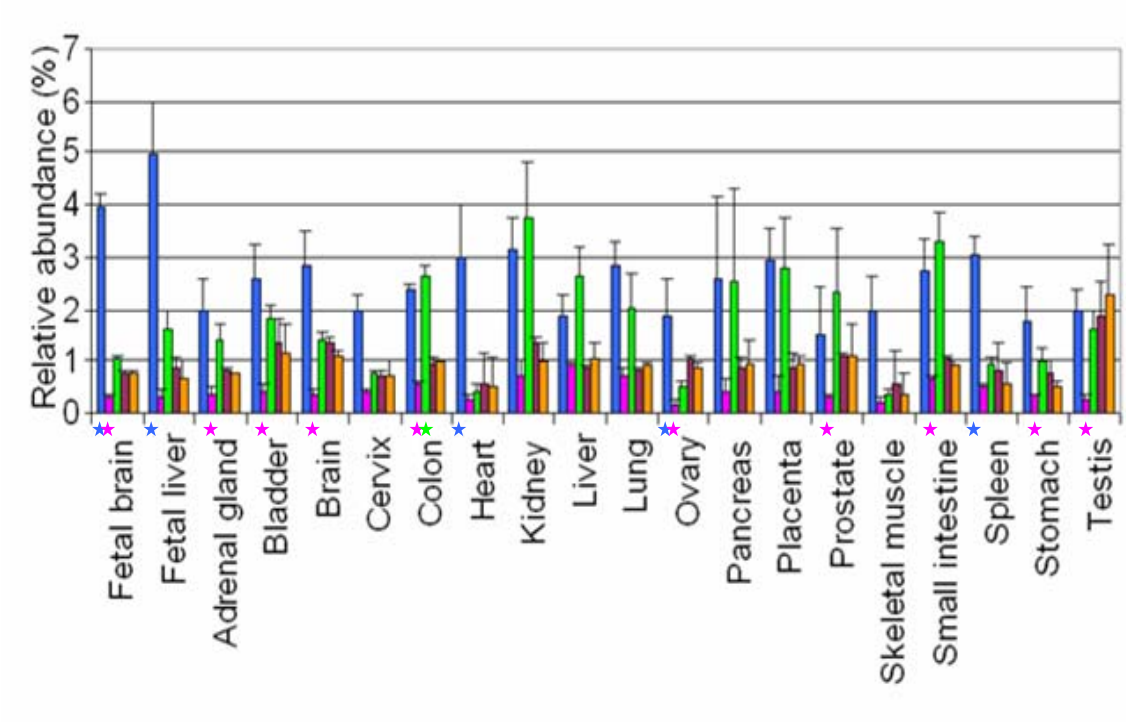


Figure 5.20 Relative abundances of *PQBP1* transcript variants
The abundance of each transcript variant group is expressed as a percentage of the constitutive region of *PQBP1*. Transcripts 1 and 15 (blue), 4 and 6 (pink), 3, 7, 8 and 11 (green), 9 and 11 (brown) and 5 and 14 (orange) are shown. Statistically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

Chapter 6

Functional analysis of

PQBP1 transcript variants

6.1 Introduction

The biological phenomenon of alternative splicing increases the repertoire of mRNA transcripts generated from a single genetic locus. This was illustrated in the previous chapter, where 15 alternative transcripts for the gene *PQBP1* were identified from various human cDNA samples. Having identified and characterised the genetic composition and expression profiles of these variants, it remains to be determined if these transcripts are capable of performing any biological function.

Some questions about the functional influence of alternative splicing can be resolved through analysis of the predicted protein sequences. Alternative splicing events frequently result in the loss or addition of domains which can, in turn, alter a protein's function. For example, alternative splicing of interleukin-4 receptor (*IL-4R*) enables the expression of either a transmembrane or extracellular isoform. The extracellular isoform is encoded by an mRNA that includes a mutually exclusive exon that contains a stop codon, which is inserted before the exons encoding the transmembrane domains (Kruse *et al.,* 1999). However, the introduction of premature termination codons by alternative splicing does not always generate functional proteins and abbreviated proteins are frequently targeted for rapid degradation through the NMD surveillance pathway.

The work described in this chapter further characterises *PQBP1* alternative transcripts in an attempt to differentiate between transcripts that are capable of producing functional proteins and those may be the result of erroneous mRNA processing. Four approaches were used to address these questions, two assessing the predicted protein products and two studying the stability of the variant transcripts. First, the predicted isoforms encoded by the *PQBP1* transcripts were analysed *in silico,* to detect premature termination codons and to predict the encoded protein domain structures. Second, the sub-cellular localisation of each unique protein was determined. Thirdly, the stability of the *PQBP1* transcripts was assessed to determine if the presence of a PTC affected the stability of the mRNA transcript. And finally, evidence was sought concerning the mechanism by which some transcripts are degraded.

Results

6.2 Identification of open reading frames in *PQBP1* alternative transcripts

Potential open reading frames within the *PQBP1* transcripts were identified using orf-finder at the NCBI (http://www.ncbi.nlm.nih.gov/gorf/gorf.html). An alignment of the resulting peptide sequences can be found in Appendix V, while the location of predicted start and stop codons within the exons of *PQBP1* alternative transcripts are shown in Figure 6.1. Termination codons located upstream of the one that encoded the full-length protein, were classed as premature termination codons (PTCs). These were found in 56.25 percent (9/16) of the transcripts.

As discussed in section 5.3, the inclusion of the novel exon 2a in *PQBP1* transcript variants 9 - 11 introduces a premature stop codon. However, these transcripts also contain an additional open reading frame, whose start codon is located 17 bp downstream from the PTC. *In vitro* transcription/translation studies or antibody based hybridisation techniques are required to determine if these alternative proteins can be produced.

Alternative transcripts do not always encode a unique protein. Analysis of deduced amino acid sequence confirmed that the proteins putatively encoded by transcripts 7 and 8 were identical (Table 6.1). Also, the predicted proteins using the reference translation start site were identical for transcripts in 9 and 11, while the putative open reading frames using the novel translation start site where identical in transcripts 9 and 10. For this reason subsequent functional analysis was not completed for isoforms 8, 9b and 11a.

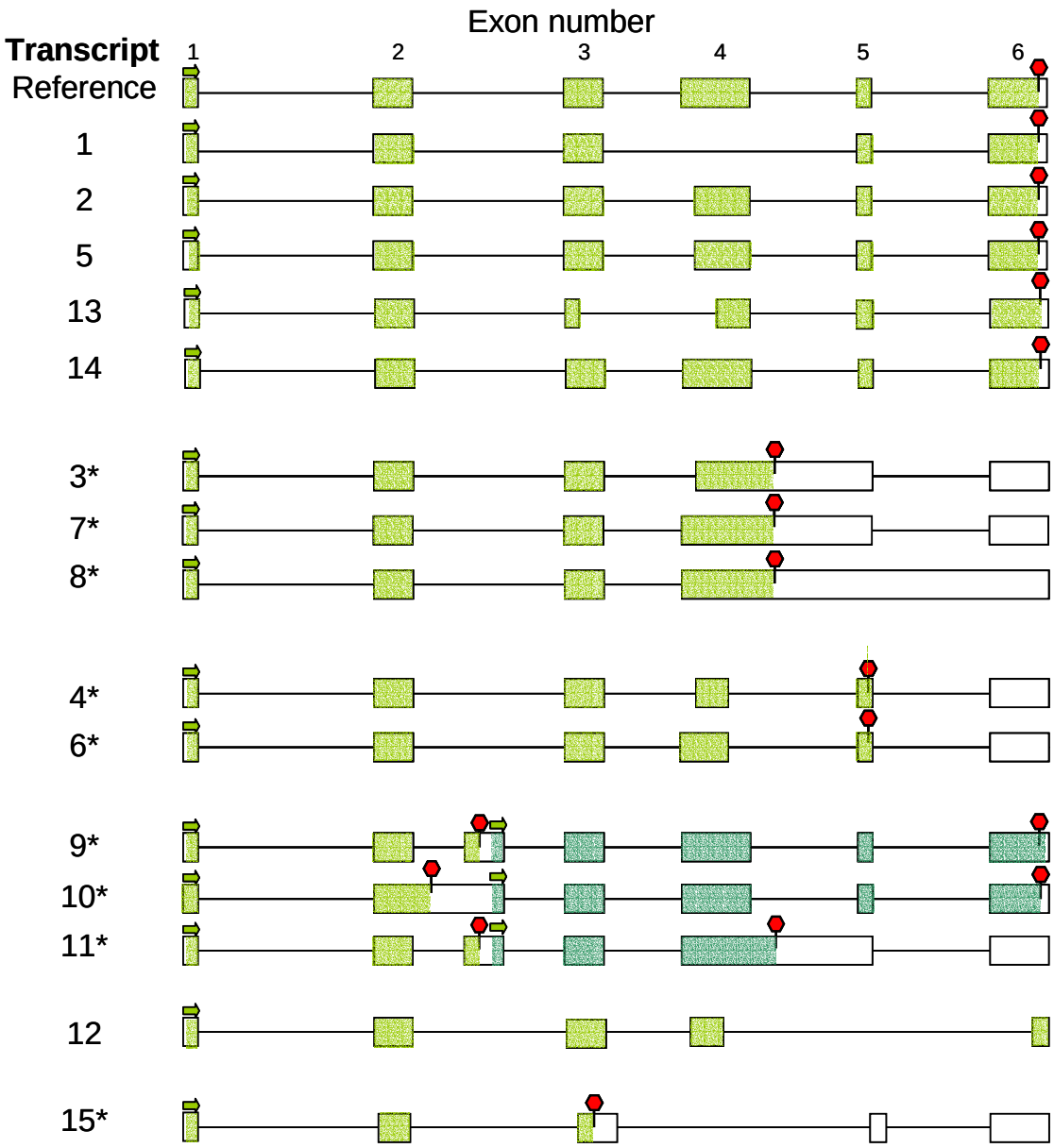


Figure 6.1 Identification of open reading frames in *PQBP1* transcripts
Open reading frames were identified using the programme orf finder. Transcript numbers refer to those outlined in section 5.2. Exons are numbered with respect to the open reading frame, (not the full length cDNA sequence), and these are displayed at the top of the figure. The location of the start codon (green arrow) and stop codon (red hexagon) in context with the exon/intron structure of the gene are shown. Transcripts with a PTC are denoted with an asterisk (*). Open reading frames that utilise the reference start codon are shaded in green. Alternative, downstream open reading frames are shaded in dark green.

6.2.1 *Predicted domains found in PQBP1 variant proteins*

Domains and motifs in all *PQBP1* isoforms were identified at Prosite (Bucher and Bairoch 1994; http://www.expasy.org/prosite/), which identifies conserved patterns (clusters of residue types) and profiles (position-specific amino acid weights) within a protein's amino acid sequence (Figure 6.2). The reference protein of *PQBP1* was predicted to contain a WW domain, an arginine rich region (ARG_RICH) and a nuclear localisation signal. WW domains are found in a number of unrelated proteins and are characterised by two highly conserved tryptophan (W) residues. The WW domain binds to proteins with particular proline-motifs, and/or phosphoserine- phosphothreonine-containing motifs (Chen and Sudol 1995; Macias *et al.,* 2002), and are frequently associated with other domains typical for proteins in signal transduction processes. The c-terminal ARG_RICH domain is rich in arginine residues and has a flexible secondary structure. It is this region of *PQBP1* that binds to homopolymeric glutamine tracts (Kumoro *et al.,* 1999b). *PQBP1* also contains a weak nuclear localisation signal which transports proteins from the cytoplasm to the nucleus.

WW domains were not identified in 25% of the *PQBP1* isoforms. Both the arginine-rich region and the nuclear localisation signal were absent from isoforms 1, 4, 6, 9a, 11a 12 and 15. Isoform 13 was predicted to contain both the WW domain, and the nuclear localisation signal but not the arginine_rich region (Figure 6.2)

Functional information could not be inferred for peptides 9a, 10a, and 15 as Prosite analysis failed to identify any domains in these sequences. Functional proteins might be produced from the alternative open reading frame found in transcripts 9, 10 and 11 (encoding proteins 9b, 10b and 11b) as they retained both the arginine rich region and the nuclear localisation signal. However, transcript 15 encodes a protein without any known domains.

Another predictive algorithm, PSORT (Nakai and Horton 1999; http://psort.nibb.ac.jp/) was employed to predict the subcellular localisation of *PQBP1* isoforms (Figure 6.3). This analysis also suggests that the isoforms encoded by transcripts 1, 4, 6, 9a, 10a 12 and 13 may not be localised to the nucleus, where the predicted likelihood of these transcripts being transported to the nucleus was

less than 50%. Interestingly, PSORT predicted that isoform 1 would be transported to the mitochondria.

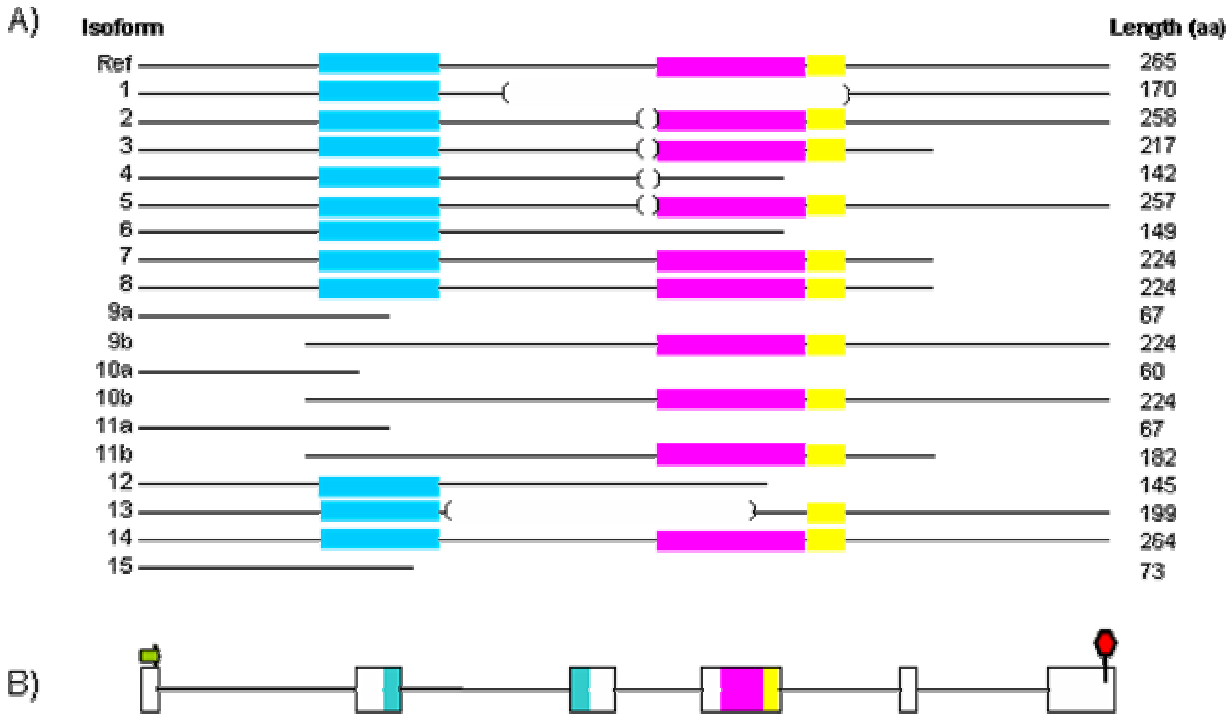


Figure 6.2 Predicted motif patterns in *PQBP1* putative proteins

A) Domain and motifs were predicted using the computational algorithim, Prosite. Deleted regions of the predicted protein sequences are flanked by brackets ( ). The domains/motifs displayed are: WW domain (blue), arginine rich region (pink), nuclear localisation signal (yellow).

B) Location of domains in exon/intron structure of *PQBP1*.

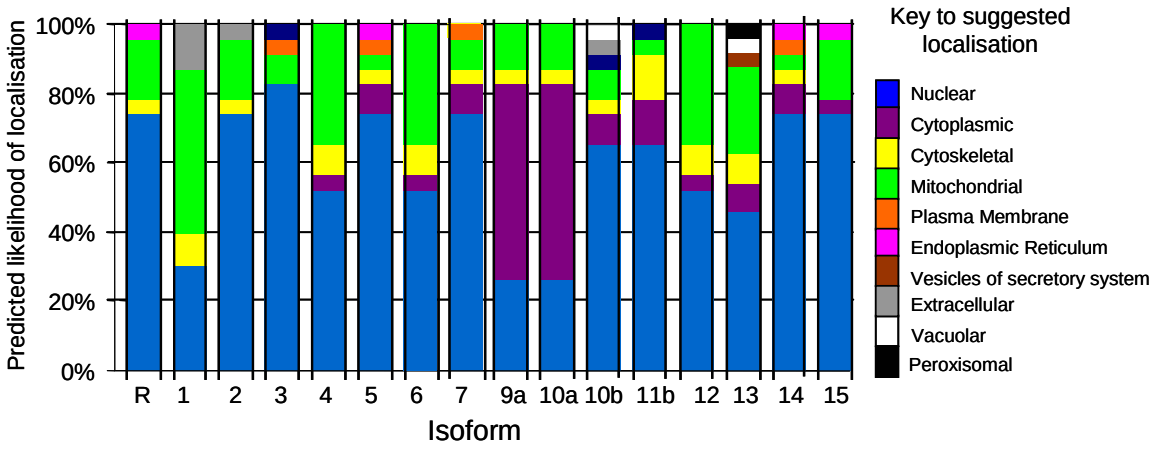


Figure 6.3 Prediction of subcellular localisation of *PQBP1* isoforms

The computational algorithim PSORT was used to predict the subcellular compartment in which the *PQBP1* proteins would be located.

Table 6.1 shows the predictions regarding nuclear localisation of the *PQBP1* isoforms by the two programmes. The majority of the isoforms (11/19) contained a nuclear localisation signal and were predicted to be localised to the nucleus by PSORT. Five isoforms did not contain a nuclear localisation signal and were therefore predicted to be found within a non-nuclear location by PSORT. However, transcripts 4, 6 and 15 did not contain a nuclear localisation signal were predicted to be localised to the nucleus by PSORT. Conversely, transcripts 13 does contain a localisation signal but was not predicted to be localised to the nucleus by PSORT.

Prosite and PSORT use sequence searches to confer similarity to identify known and previously characterised protein domains or localisation signals. Any differences observed in the comparison may be attributed to the different sensitivity and searching patterns of the computational algorithms. However, it is important to remember that these results are merely predictions and require additional experimental analysis to confirm the actual intracellular location of these proteins.

Table 6.1 Predictions regarding the nuclear localisation of *PQBP1* isoforms

Transcript	Prosite	PSORT	Transcript	Prosite	PSORT
Ref	✓	✓	9b	✓	✓
1	×	×	10a	×	×
2	✓	✓	10b	✓	✓
3	✓	✓	11a	×	×
4	×	✓	11b	✓	✓
5	✓	✓	12	×	×
6	×	✓	13	✓	×
7	✓	✓	14	✓	✓
8	✓	✓	15	×	✓
9a	×	×			

6.3 Intracellular localisation of *PQBP1* transcripts

Following *in silico* predictions of sub-cellular localisation, an experimental approach was used to study the localisation of individual isoforms in cultured cells. In order to confirm that the T7 epitope did not influence the cellular distribution of the *PQBP1* isoforms, the distribution was assessed using both amino- and carboxyl-epitope tagged proteins. Fusion proteins carrying the T7 epitope tagged to either the amino or carboxyl – terminal of each *PQBP1* isoform were expressed in both African green monkey kidney cells, cos-7, and Chinese hamster ovary, CHO-K1, cell lines. The subcellular location of the *PQBP1*-T7 isoforms was then detected using an anti-mouse T7 antibody. A schematic of the overall experimental procedure and the sequence of the T7 epitope are displayed in Figure 6.4 and the procedure is described in more detail in sections 2.23 and 2.24.

6.3.1 *Preparation of constructs*

The plasmids used in this analysis were pCDNA.3NT7 and pCDNA.3CT7 (kind gift of Dr J. Collins). These were linearised using the appropriate restriction enzymes for the location of the T7 epitope tag; pCDNA.3NT7 was digested with *Not*I and *XbaI* while pCDNA.3CT7 was digested with *Hind*III and *NheI*.

Primers were designed to amplify each of the cloned open reading frames in the holding vector pGEM®-T Easy. The primers also contained the appropriate restriction enzyme sites to permit their ligation with the pCDNA.3 vector. The primer combinations used to amplify each transcript are listed in Appendix VI. The stop codon was retained in the N-tag constructs while it was avoided in the C-tag constructs. A detailed description of the cloning procedures used in this experiment can be found in section 2.20. Sequence verification of all clones was performed by the Research and Development team at the WTSI.

PQBP1-T7 constructs were prepared for 14/15 amino-tagged isoforms and 11/15 carboxyl-tagged isoforms. Isoform 12 was not cloned as it contained a *Hind*III restriction enzyme site within the ORF which would have also been digested during the preparation for cloning. Alternative vectors of pCDNA-T7 Ntag and pCDNA-T7-Ctag without a *Hind*III site were available to carry out this experiment. These vectors used a *Xmn*I site for ligation between the insert and vector. Attempts made to clone isoform 12 using this vector were unsuccessful. In addition, isoforms 1 and

15 were not successfully cloned using the pCDNA-T7 Ctag vector, as all attempted ligations failed.

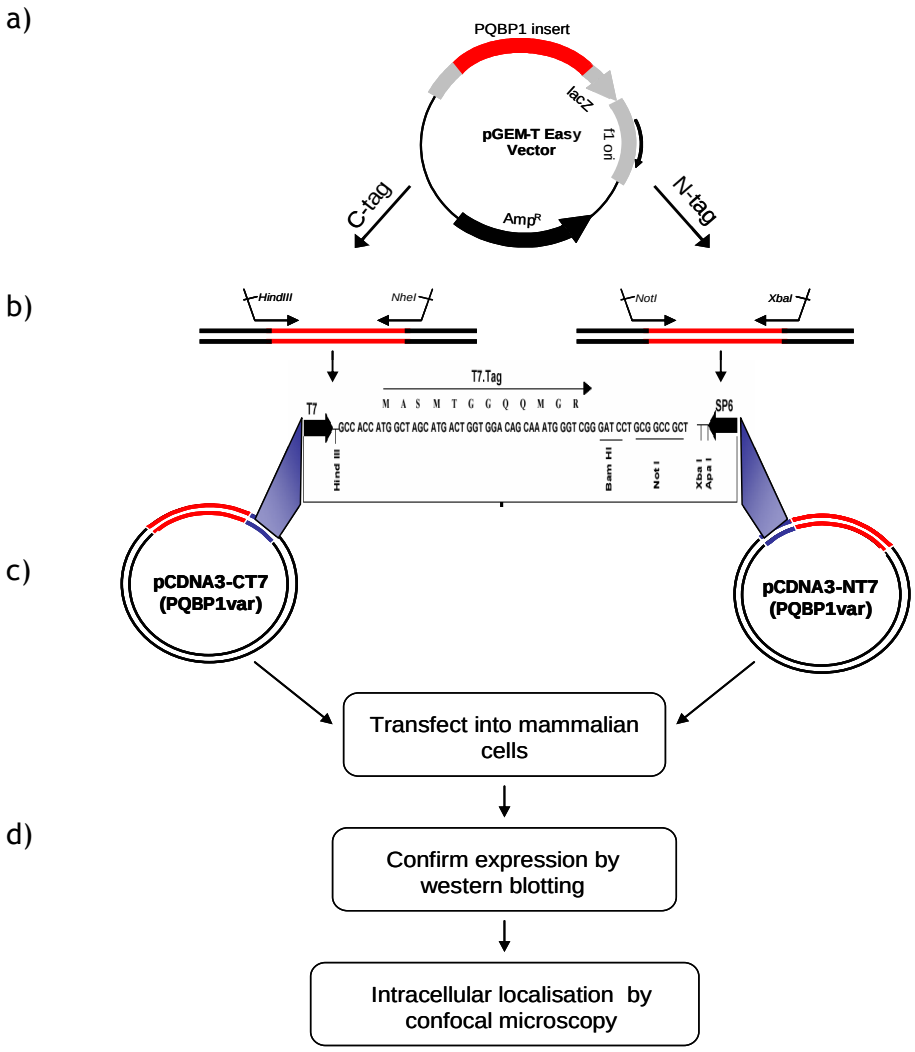


Figure 6.4 Schematic of subcellular localisation protocol
Constructs were prepared in which the T7 epitope was fused to either the N-terminal or C-terminal of *PQBP1* proteins. Following transfection into mammalian cells, the expression of the transcripts was confirmed by Western blotting. Confocal microscopy was also employed to visualise the subcellular localisation of *PQBP1* proteins.

a) Representation of the holding vector containing the cloned ORFs (section 5.2).
b) Amplification of the ORF using primers that contain a restriction enzyme linker suitable for ligation to the pCDNA3 expression vector.
c) The nucleotide and amino acid sequence of the T7 eptiope tag and its location in the pCDNA3 vectors.
d) Overview of experimental protocols carried out to determine the subcellular localisation of *PQBP1* isoforms.

6.3.2 *Western blot analysis of PQBP1-T7 fusion proteins*

Constructs containing the *PQBP1* open reading frames fused to the T7 epitope tag were transiently transfected into Cos-7 and CHO-K1 cell-lines as described in sections 2.12. Two cell lines were used in this analysis to ensure that the genetic background did not influence the cellular distribution of *PQBP1* isoforms and to ensure uniformity between this experiment and the mRNA stability experiment carried out in section 6.4.

The expression of the *PQBP1*-T7 proteins was confirmed by Western blot analysis using an anti-T7 monoclonal antibody (section 2.24). This was carried out to confirm that the predicted size of the *PQBP1* isoforms and to ensure that the isoforms were not exported from the cell. Extracellular proteins were isolated by gentle centrifugation 24 hours after transfection. After the supernatant (the fraction that included the extracellular proteins) had been removed, the cells were disrupted using chemical treatment to release the cellular proteins. Both the extracellular and cellular protein fractions were assayed.

Proteins containing the T7 epitope tag were of the imperically calculated size Table 6.2. Expression of *PQBP1* isoforms could not be confirmed for all constructs. In particular, constructs *PQBP1*-9a, -10a, -10b, 11b and 15 could not be detected. This could be attributed to inefficient transfections, or the constructs may be rapidly degraded with either the mRNA transcript or protein sequence being highly unstable.

Table 6.2 Expected sizes of T7 tagged proteins

Isoform	Predicted size (kDa)	Isoform	Predicted size (kDa)
Reference	30.5	9a	7.6
1	18.8	10a	7.0
2	29.7	10b	25.5
3	25.3	11b	21.1
4	16.4	12	16.3
5	29.6	13	18.7
6	17.2	14	30.3
7	26.2	15	8.4

* molecular weights were predicted using pepstats (emboss)

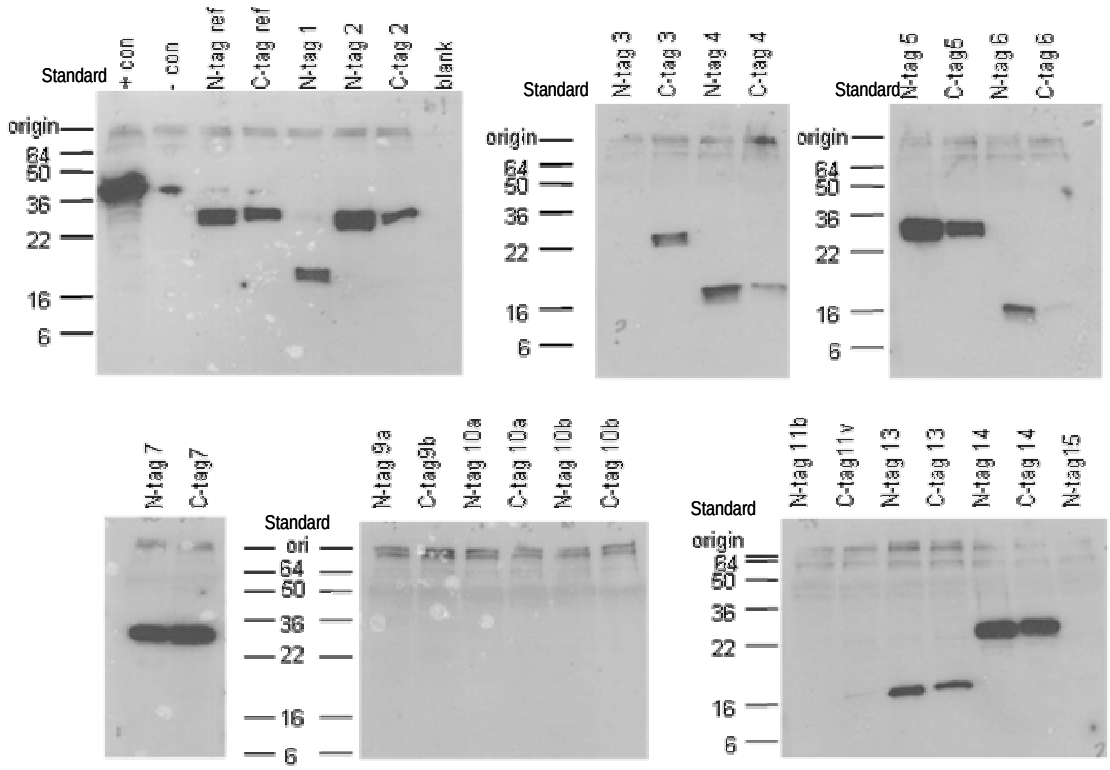


Figure 6.5 Western Blot analysis confirms *PQBP1* expression for some, but not all constructs

Cos-7 cells were transiently transfected with pCDNA3-*PQBP1*(variant)-T7 plasmids. Expression of *PQBP1*-T7 fusion proteins was confirmed using an anti-T7 antibody. Also included in the analysis are a positive control (MAPK) and a negative control (no plasmid added). Standard – protein molecular weight (kDa) standards are also displayed.

6.3.3 *Subcellular localisation of PQBP1 transcripts in cos-7 and CHO cells.*

The distribution of the *PQBP1*-T7 isoforms was assessed by confocal microscopy (Figure 6.6). This was achieved by transfecting the *PQBP1*-T7 tag into African Rhesus monkey, Cos-7, and chinese hamster ovary, CHO-K1 cells (Figure 6.6). All transfections were completed in duplicate on each cell line.

The cellular location of the *PQBP1* transcripts was ascertained for the reference *PQBP1* protein as well as 67% (10/15) of the alternative isoforms. Two isoforms were not assayed because they were not successfully cloned (due to ligation failures). Cellular expression of three other isoforms (isoforms 9, 10 and 11) was not identified by confocal microscopy. Expression of these isoforms also failed to be confirmed by Western blotting.

Visualisation by confocal microscopy localised the reference and isoforms 2, 3, 5, 7, 13 and 14 to the nucleus, while isoforms 1, 4, and 6 were not targeted to any cellular compartment (Figure 6.6). Instead, these isoforms were found to be ubiquitous throughout the cell. These observations were consistent with both the previously published localisation experiments (Okazawa *et al.,* 2001; Kalscheuer *et al.,* 2003), and the computational predictions made in section 6.2.1.

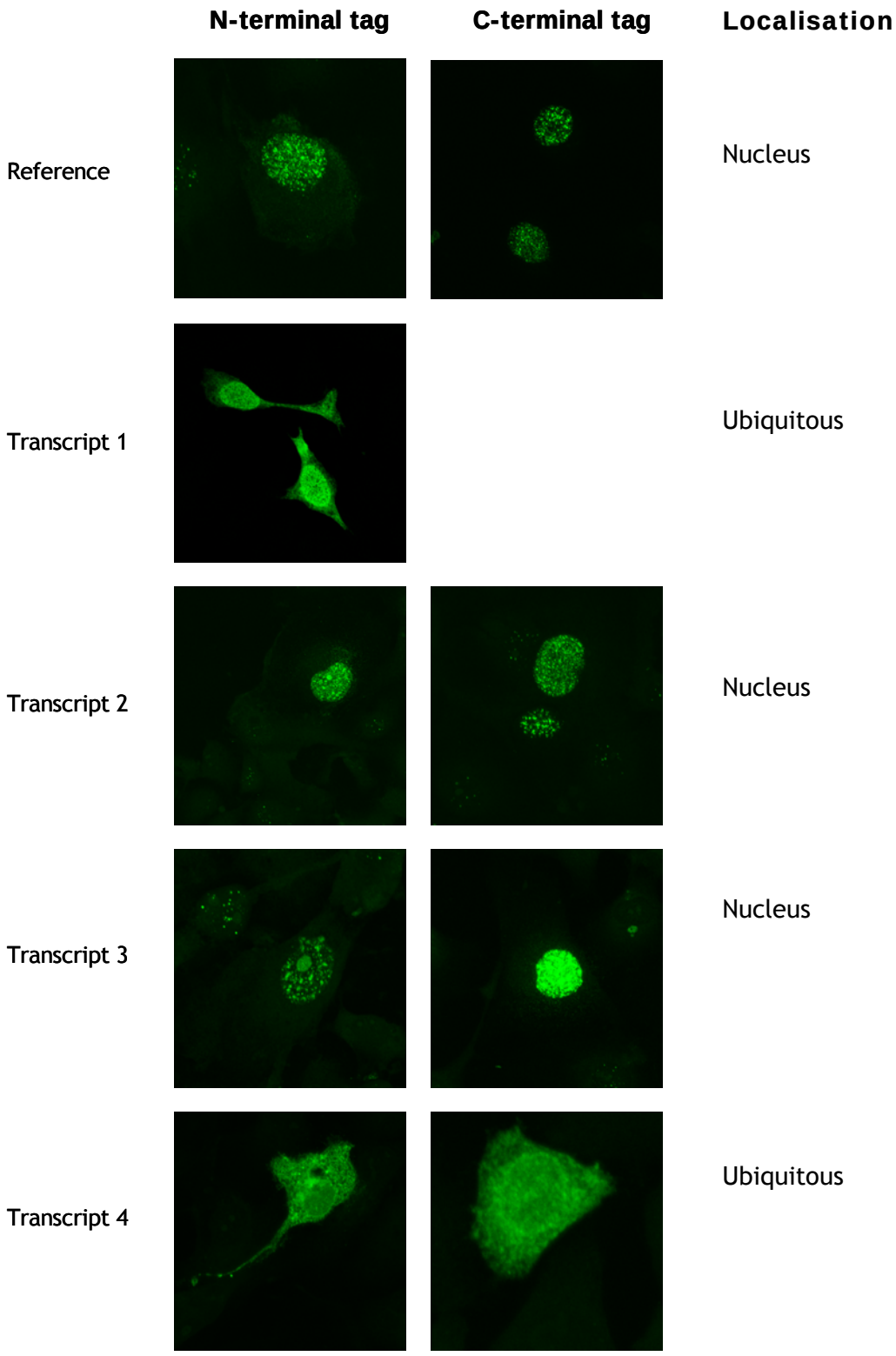


Figure 6.6 Localisation of *PQBP1* alternative isoforms in Cos-7 cells.
Isoforms encoded by *PQBP1* alternative transcripts were expressed as fusion proteins with the T7 epitope tag. The sub-cellular localisation of proteins was assessed by immunofluorescence with an anti-T7 monoclonal antibody and an FITC tagged secondary antibody. Cells were visualised by confocal microscopy. Counter-staining with DAPI confirmed the location of the nucleus (results not shown). Continued overleaf.

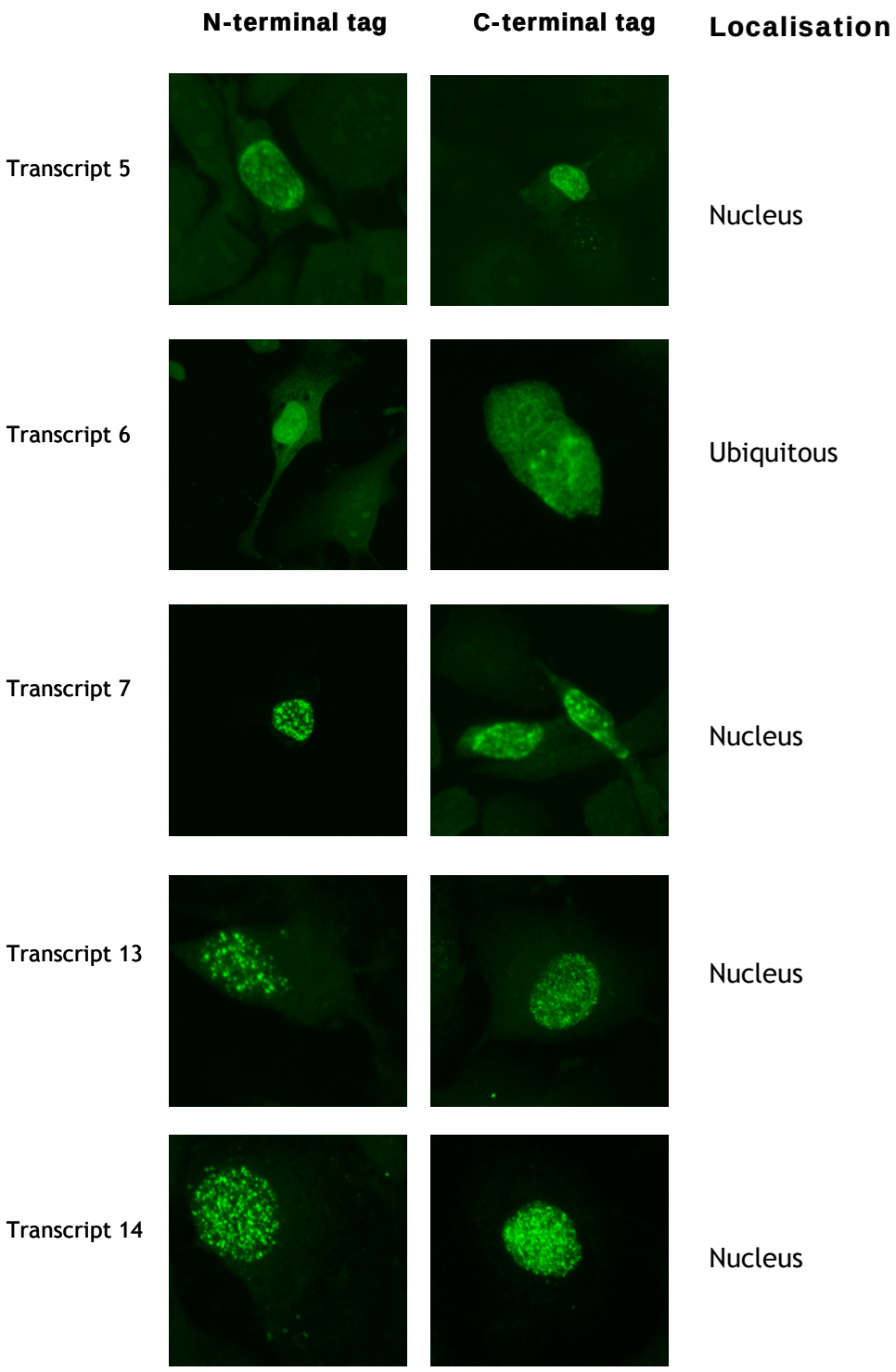


Figure 6.6 continued.

6.4 Analysis of the stability of *PQBP1 transcript variants*

The stability of an alternative transcript may also give an indirect indication of its biological importance. The NMD surveillance mechanism is thought to rid cells of potentially deleterious transcripts that are produced by mutations or errors in mRNA processing. Analysis of the open reading frames within the *PQBP1* transcripts described indicated that 9/16 variants, which represent approximately 2% of total *PQBP1* transcripts, have a PTC and therefore do not have the potential to encode a full length protein. The low expression levels of these transcripts could indicate that they are infrequently produced by the splicing machinery, or that they are removed by the NMD pathway. Furthermore, if these transcripts are rapidly removed, this may be an indication that they have no biological function. Therefore, it was decided to monitor the mRNA decay rates for the *PQBP1* transcripts. The expectation was that a difference in the stability of PQBP1 transcript would be seen in transcripts that contain a PTC and those that do not. As the expression of several *PQBP1* isoforms (isoforms 9-12) was not detected by Western blotting or immunofluorescence, it was conjectured that this may attributed to cellular stability of the relevant mRNA. Analysis of the mRNA decay patterns for these transcripts may shed further light on why expression of these transcripts could not be detected by immunofluorescence.

The method chosen to evaluate mRNA degradation kinetics was a transcriptional quantification strategy using a tetracycline (tet)-regulated promoter. Changes in mRNA levels were monitored over a defined time course, following targeted transcriptional repression with a tetracycline derivative, doxycycline. One advantage of this technique was that expression of the *PQBP1* gene could be tightly regulated without interfering with cellular physiology (Gossen & Bujard, 1992). It was hoped that this specificity would permit *in vivo* decay rates to be more closely reproduced, as pleiotropic effects that are frequently observed when using non-specific transcription/translation inhibitors would be avoided. The BD™ Tet-off system was chosen to regulate the expression of *PQBP1* alternative transcripts, and a description of this expression system follows.

The BD Tet-Off system is dependent on the integration of a tet-response element (TRE) and a regulatory element (pTet-Off) into a mammalian cell line. To control expression the pTet-off regulatory plasmid must first be stably transfected into the

cell line. This plasmid encodes the tet-response transcription activator, which is able to activate transcription by binding to tetracycline response elements (TRE). In this study, the expression of *PQBP1* transcript variants is under the control of the TRE that is found in the plasmid pTRE-TIGHT (Figure 6.6).

Following the introduction of a *PQBP1* transcript into the pTRE-TIGHT response plasmid, the construct is transfected into a CHO-AA8 Tet-Off cell line. Here, the tet-responsive transcriptional activator, expressed from the regulatory plasmid binds to the TRE, thus activating transcription. As doxycyline is added to the culture medium, transcription of the *PQBP1* insert from the TRE is turned off in a dose-dependent manner. This system has previously been used to monitor mRNA stability of the β-globin gene (Couttet and Grange 2004).

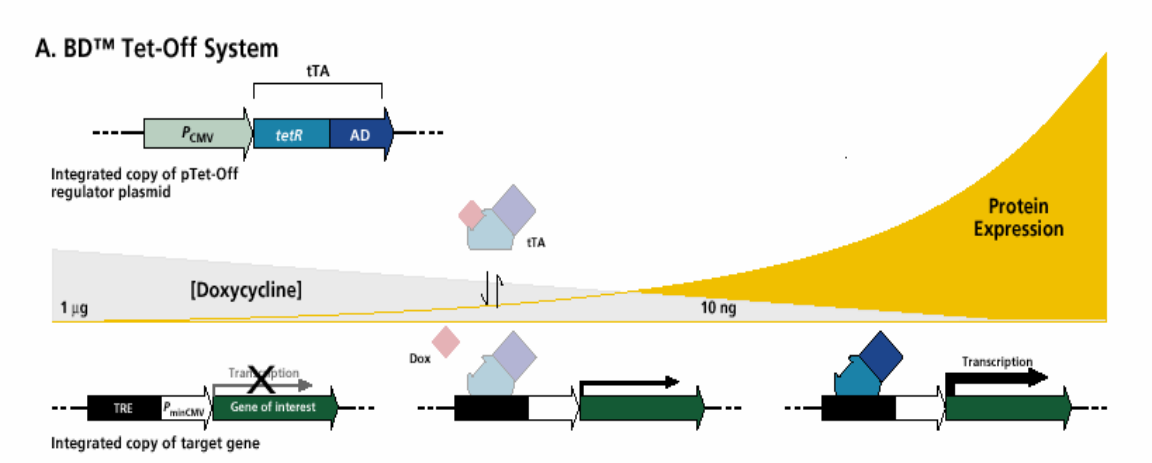


Figure 6.7 Schematic of gene regulation on the BD™ Tet-Off system
This system is dependent on both a regulatory (pTET-off) plasmid and a response plasmid (pTRE-Tight-*PQBP1*). When cells contain both of these plasmids the expression of *PQBP1* transcripts only occurs when the tet regulatory protein (tTA) is bound to the tet regulatory element (TRE). In the tet-off system tTA binds to the TRE and activates transcription in the absence of tetracycline or its analogue, doxycycline.

6.4.1 *Preparation of constructs*

PQBP1 cDNA clones (section 5.2) were subcloned into the response plasmid of the Tet-Off expression system, pTRE-TIGHT. A schematic diagram outlining the methodology employed to prepare the constructs is shown in Figure 6.8.

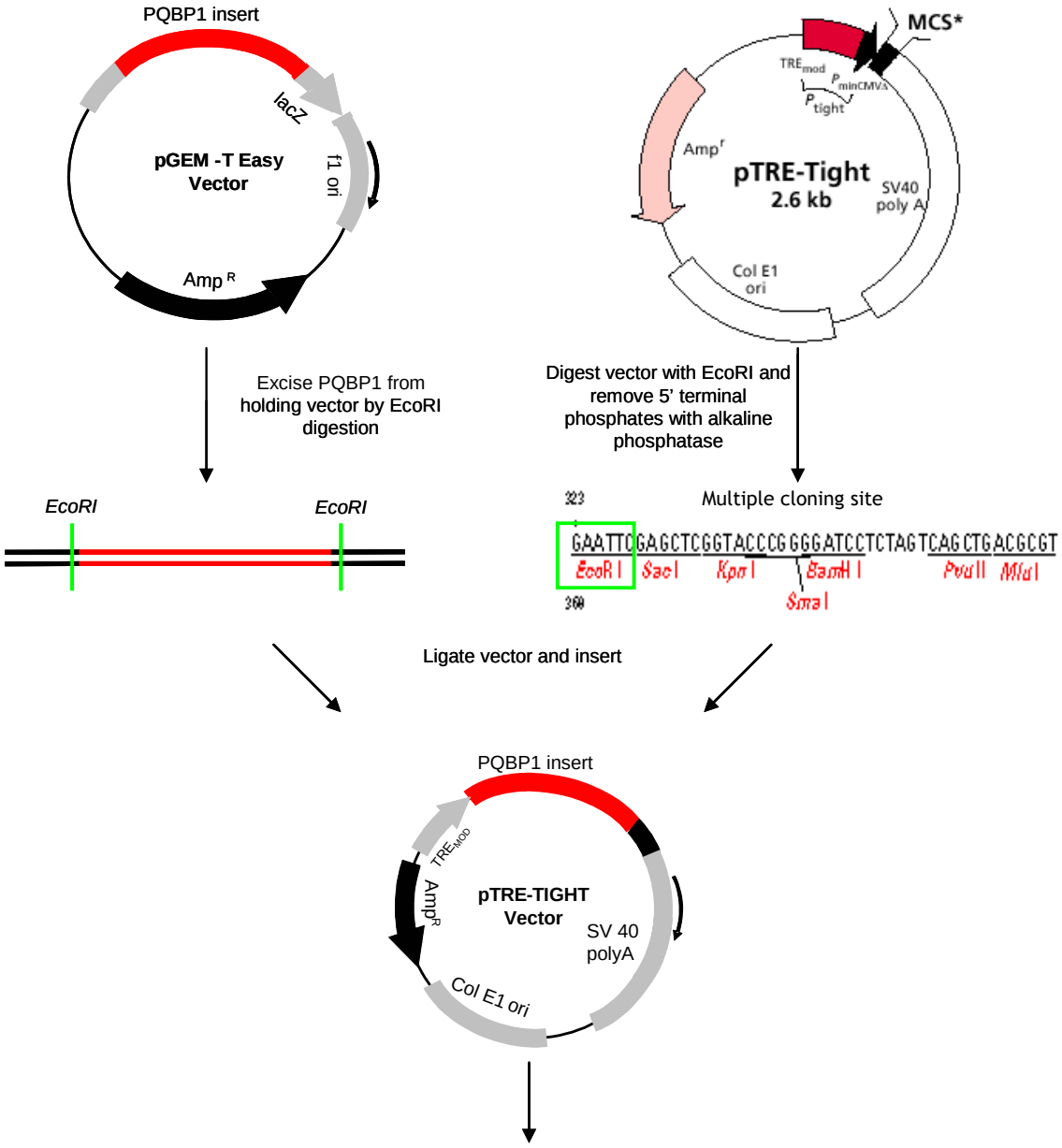


Figure 6.8 Preparation of constructs to study *PQBP1* mRNA decay rates
Cloned cDNAs were excised from the pGEM-T Easy vector by *Eco*RI digestion. The transcripts were ligated to complementary ends of the pTRE-TIGHT vector.

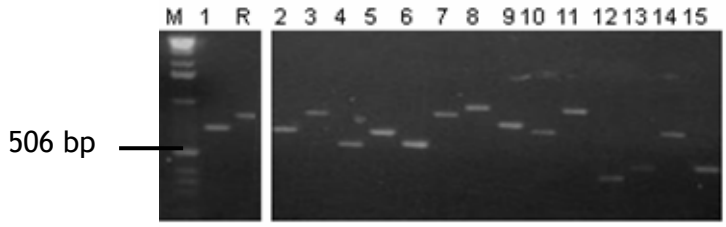


Figure 6.9 Preparation and analysis of pTRE-TIGHT-*PQBP1* plasmids
A) Transcripts (numbered accordingly) were digested from the holding vector by *EcoR*I digestion. Molecular weight marker is denoted (M). Purified plasmids are displayed.

The cloned fragments were excised from the holding vector using the restriction enzyme *EcoR*I and were purified (Figure 6.9), and treated with shrimp alkaline phosphatase prior to ligation with the plasmid pTRE-TIGHT (described in detail in section 2.21). The fidelity of the insert was confirmed by DNA sequencing, which was carried out by the Research and Development Team at the WTSI, and the pTRE-TIGHT-*PQBP1* constructs were purified using a qiagen midi-prep plasmid procedure (section 2.14.6).

6.4.2 *Optimisation of experimental protocol*

Transfection efficiencies

Transient transfection of the response plasmid in to CHO-AA8 tet-off cells was optimised using the vector pTRE-TIGHT-eGFP, where expression of the green fluorescent protein is under the control of the TRE. Transfections were completed varying the volume of transfection reagent used (GeneJuice™ (Novagen) from 1 to 8 µl), cell densities (0.3 x 10^5 to 1 x 10^5 cells per ml) and DNA concentrations (1-2 µg). The efficiency of transfection was monitored by fluorescence microscopy. The number of fluorescing cells in 3 random fields of view were counted for each transfection (results not shown). Optimal transfection results were achieved using 2 µg of plasmid DNA, 4 µl of GeneJuice and 3 x 10^5 cells in a volume of 3 ml per transfection. These conditions were used for all subsequent transfections.

Doxycycline dose response curves

The necessary concentration of antibiotic required to inhibit expression was determined by incubating CHO-AA8-luc cells (Clontech) with various concentrations of dox. The CHO-AA8-luc cells, have both the tet-off TRE and luciferase gene

stably integrated into their genome (Clontech). Expression of the luciferase gene, was under the control of the tet-off TRE and hence was repressed upon the addition of doxycycline. Doxycycline was titrated between the range of 1pg ml^{-1} and 100 ng ml^{-1} and the concentration that most effectively repressed the expression of the luciferase gene was determined using a luciferase activity assay (section 2.22.2). Results for each sample were normalised to the total protein content of each cell as determined by the Bradford assay (section 2.22.3). Approximately 500 fold repression was observed in this experiment and the optimal concentration required effectively to repress the expression of the luciferase gene was 50 ng ml^{-1} (Figure 6.10). This concentration was used in subsequent experiments.

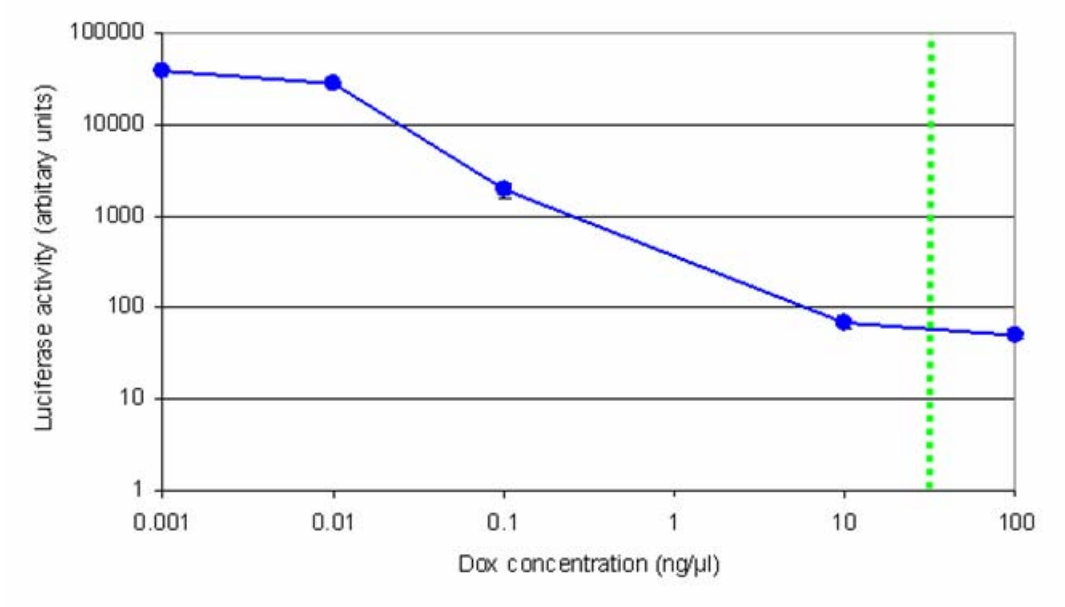


Figure 6.10 Dose response curve for the CHO-AA8-Luc control cell line
Expression of luciferase was repressed using various concentrations of doxycycline. Following an incubation of 4 hours in the presence of doxycycline luciferase activity was detected by the luciferase luminescence assay. The concentration of doxycycline used in all subsequent experiments is also indicated (green – 50 ng/µl).

6.4.3 *Quantitative analysis of mRNA stability*

Messenger RNA decay rates of the *PQBP1* alternative transcripts were determined using the tet-off expression system displayed in Figure 6.11. During transfection cells were grown in the absence of doxycycline, which was added to all cells at a final concentration of 50 ng μl^{-1} at time 0. RNAs were collected at various intervals following transcriptional arrest and cDNA was then synthesised. In total,

13 assays were performed on the following transcripts; reference 1, 2, 4, 6, 7, 8, 9, 10, 12 and 15. Five of these were carried out in duplicate from the transfection stage – they are reference *PQBP1*, 1, 6, 8 and 10. Control assays were also completed to:

- Ensure the specificity of *PQBP1* amplification. Here, CHO-AA8 cells were transfected with plasmid pCMVß to ensure that no amplification was observed using *PQBP1* primers in the CHO genetic background.
- Ensure that gene expression of the tet-regulated gene was not repressed in the absence of doxycycline.

Measures were taken to ensure that the abundance of each *PQBP1* transcript was accurately quantified throughout each experiment. Messenger RNA levels in each sample were normalised to the house keeping gene beta-actin (*Actb)*. This gene was chosen because the mRNA sequence was available in the Chinese Hamster (Embl:U20144**).**

The expression levels of the various *PQBP1* transcripts were determined using the primer pair, *PQBP1*.Q10 (section 5.4.3) which can amplify any of the *PQBP1* transcript variants. To ensure that additional splicing of the *PQBP1* transcript did not take place following its transfection, the full length *PQBP1* transcript was amplified and its size was assessed by end-point PCR using the primers that were used to amplify the entire *PQBP1* transcript (section 5.2).

The pTRE-TIGHT- transcript specific *PQBP1* construct was also co-transfected with the reporter vector, pCMVß in order to correct for varying transfection efficiencies between experiments. In this vector the expression of lacZ was under the control of the strong CMV promoter and is not regulated by doxycycline. The relative abundance of both the *PQBP1* transcript and the lacZ gene were determined by real-time PCR. The abundance of the *PQBP1* transcript was then normalised to the abundance of the LacZ gene.

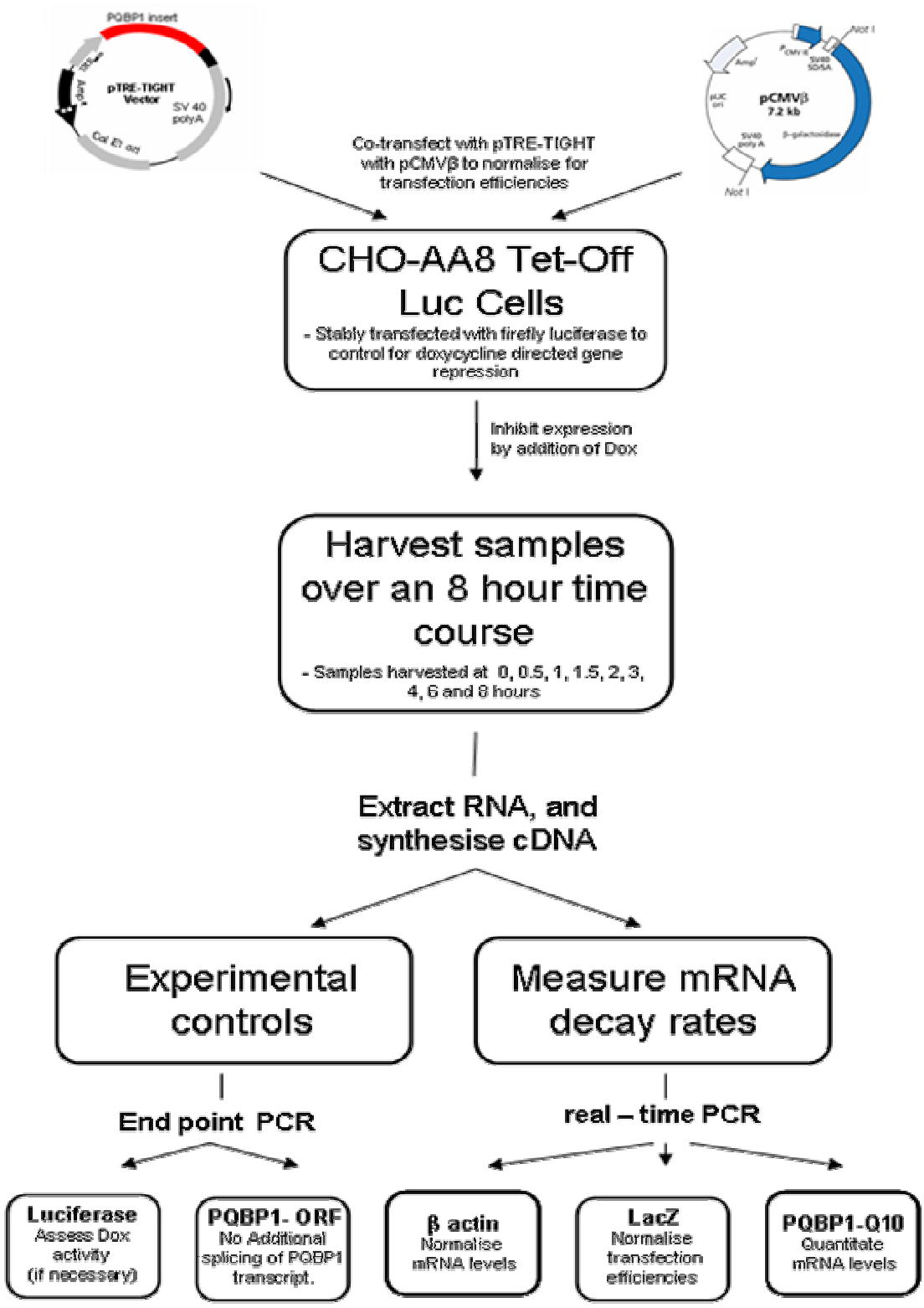


Figure 6.11 Schematic of experimental protocol used to assess mRNA decay rates

Validation of real-time PCR primer combinations

Variations in transcript abundance were determined by real-time PCR using SYBR-Green quantification. As SYBR-Green binds to double stranded DNA in an indiscriminate manner, it was necessary to ensure that all reactions amplified only the product of interest. Therefore, the samples were subjected to the heat dissociation protocol following the final cycle of the PCR. Dissociation of the PCR products consistently produced a single peak for *ActB*, lacZ and *PQBP1* demonstrating the presence of only the expected product in the reaction (data not shown). In addition, *PQBP1* expression was not observed in untransfected cells.

For the real time assay the fold change in lacZ and *PQBP1* mRNA levels were determined using the $2^{-\Delta\Delta CT}$ method (described in section 2.15.6) which assumes similar amplification efficiencies of the target gene and the internal control gene. All three primer pairs had similar amplification efficiencies (data not shown). All quantitative PCRs were performed in triplicate. Moreover, specific mRNA amplification was ensured by the inclusion of a negative control (cDNA prepared without reverse transcriptase) for each sample.

The normalised abundance of the *PQBP1* transcripts was expressed in relation to the amount present at time zero. In all cases, the level of a *PQBP1* transcript increased following the addition of doxycycline, and peaked at either 30 or 60 minutes and then declined. These decay curves are shown in Figure 6.12. The patterns of decay were similar to those described for hexosaminidase A (alpha polypeptide) (*HexA*) which also used SYBR-Green real-time PCR detection (Schmittgen *et al.,* 2000). First order mRNA decay rates were assumed to determine the mRNA half lives. These were calculated by linear regression analysis between the time points 1 and 4 hours. The correlation coefficient and standard error of the estimates were calculated from the linear regression and are shown together with the mRNA half life in Table 6.3 and Figure 6.12.

Owing to time limitations, these experiments were not completed on all *PQBP1* transcript variants. The reference and transcript variants 1, 2, 4, 6, 7, 8, 9, 10, 12, and 15 were assayed.

Table 6.3 Analysis of mRNA decay obtained from *PQBP1* alternative transcripts

Transcript	k (h^{-1})	mRNA half life (h)	r^2	Standard Error
Reference	0.6289	1.10	0.59597	0.491
1	0.4194	1.65	0.7341	0.109
2	0.523	1.27	0.9387	0.309
4	0.8784	0.79	0.9891	0.255
6	1.094	0.63	0.8696	0.156
7	0.5997	1.16	0.602	0.78
8	0.648	1.07	0.8179	0.227
9	1.122	0.62	0.9847	0.04
10	1.1437	0.61	0.8416	0.68
12	0.2088	3.31	0.9009	0.178
15	0.3902	1.78	0.6975	0.65

The amount of each *PQBP1* transcript was determined in relation to *ActB* and LacZ transcript levels for each of the transcripts listed above. Samples were collected over an 8 hour time period following transcription arrest by the addition of doxycycline, and cDNA was prepared. The first-order rate constants for *PQBP1* degradation (k), correlation coefficient (r^2) and standard error were calculated from linear regression analysis of the mRNA decay plots (Figure 6.12).

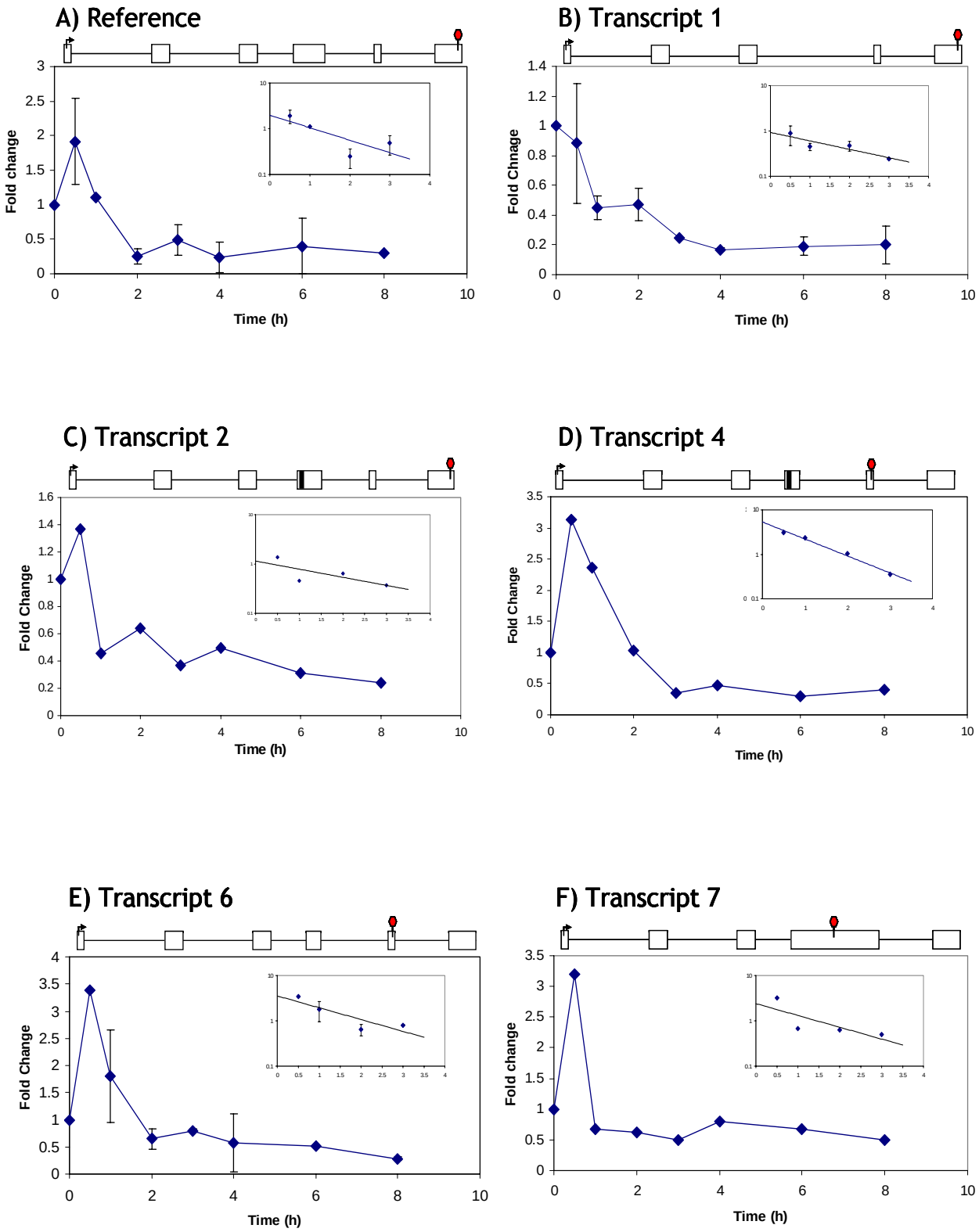


Figure 6.12 Analysis mRNA decay by real-time PCR
Plots show changes in levels of individual *PQBP1* transcripts relative to the level at time=0. mRNA expression levels were normalised to *ActB* levels as described. The exon structure of each transcript is shown, with the start and stop codons denoted by an arrow and red hexagon, respectively. First order decay plots are shown, inset. Error bars represent results from duplicate experiments.

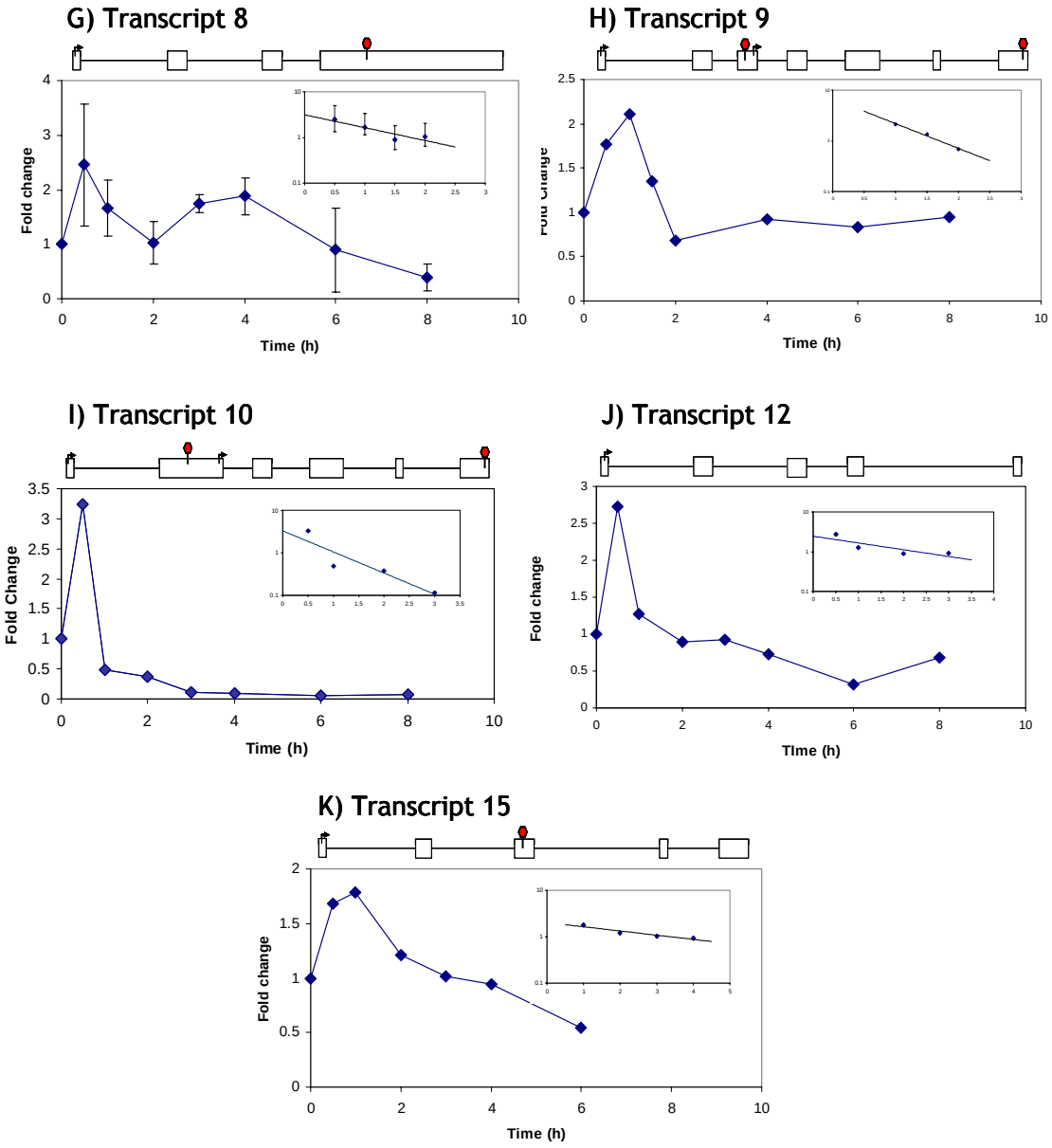


Figure 6.12 continued.

Results from Table 6.3 highlight that the 21 bp deletion identified in transcript 2, does not affect the stability of the mRNA transcripts, reference and transcript 2 being 1.10 and 1.25, hours respectively. This same deletion is also present in transcript 3, 4 and 5 and may not compromise the stability of these transcripts. The results also suggested that the intron retention (as observed in transcripts 7 and 8) did not affect the stability of the mRNA transcripts, despite the fact that this would apparently result in a PTC .

The mRNA half lives of two transcripts lacking exon 4 (transcript 1 and 15) were substantially higher than that of the reference transcript. The half life of transcript 1, where exclusion of exons is the only change, was 50% higher than that of the reference transcript (1.65 v 1.10), while transcript 15 which also has an addition to exon 2 had an mRNA half life that was 59% greater. Together these results suggest that inclusion of exon 4 in a *PQBP1* transcript may decrease the stability of the *PQBP1* transcripts.

Transcript 12 had the most significant increase in the mRNA half-life, (threefold higher than the reference transcript). The splicing pattern of this transcript results in a frame shift that removes the reference transcript's termination codon. Measures were not made to ensure that the open reading frame of the transcript was disrupted upon its introduction into the pTRE-TIGHT vector. However, an in-frame stop codon is located 12 bp downstream from the site of ligation with the pGEM-TEasy vector. This is a significant limitation that must be considered in the interpretation of these results, and it would be beneficial to introduce a stop codon into the transcript when included in the pTRE-TIGHT construct.

Finally, transcripts 4, 6 9 and 10 whose splicing patterns introduced a premature termination codon had shorter half lives than the reference transcripts. The utilisation of an alternative donor site that resulted in a 132 bp deletion in exon 4, which was found in transcripts 4 and 6, was correlated with a 40% decrease in the half life of the mRNA transcript. While the inclusion of a novel exon into the mRNA transcript for *PQBP1* also resulted in a 43% reduction in the mRNA half-lives of transcripts 9 and 10. However, the mRNA decay profile for transcript 9 requires further validation. This is because a result could not be determined for one of the time points (t=3). The inclusion of this time point in the analysis may alter its half-life.

Transcript 8 displayed an unusual, biphasic mRNA decay profile. Over the duration of the time course, two peaks of *PQBP1* transcript abundance were observed, one after 30 minutes and one 4 hours after the addition of doxycycline. While a biological explanation for this observation cannot be offered at this stage, it must be noted that the earlier quantitative analysis in human tissues failed to detect significant levels of this transcript (section 5.4). Transcript 8 was generated through the retention of two introns and may be the result of an aberrant splicing event. Additional experimental analysis is required to characterise the mRNA decay pattern of this transcript further. This could include, monitoring the stability of the mRNA using an NMD depleted system such as reducing Upf1 expression by RNAi (Upf1 is an NMD factor that is thought be to a bridge between the premature termination event and EJC.) (Mendell *et al.,* 2002).In addition, it is possible that the transcribed intron may harbour a secondary RNA structure that may somehow affect its stability. Nucleotide sequence analysis could be employed to predict this (such as mFold), but ideally the structure should be analysed by experimental analysis such as nuclear magnetic resonance.

6.5 Degradation of alternatively spliced *PQBP1* transcripts by nonsense mediated decay

The results obtained in section 6.4 indicated that some transcripts with a PTC (ie transcripts 4, 6, 9 and 10) had shorter half-lives than the reference *PQBP1* transcript and therefore may be targeted for rapid degradation. This experiment however did not identify a cellular mechanism that might be responsible for the rapid degradation of these transcripts.

In order to determine if NMD was involved in the degradation of PTC harbouring *PQBP1* transcripts, the NMD pathway was inhibited using antibiotics that block the translation which is a necessary step in the identification of transcripts harbouring a PTC. Following inhibition, the abundance of native mRNAs was determined over a six hour time course. It was hypothesised that the levels of native mRNAs without PTCs should not be targeted by the inhibited NMD pathway and are predicted to increase in abundance over the duration of the experiment.

HEK293 (human embryonic kidneys) cells were chosen for this assay. These cells were treated with antibiotics that have been shown to inhibit NMD by blocking the pioneering round of translation, a necessary step in the identification of PTCs (Figure 6.1). The antibiotics used were cycloheximde, anisomycin and puromycin and the mechanisms by which they inhibit protein translation are listed in Table 6.4. Cycloheximde and anisomycin work specifically in eukaryotic cells while puromycin blocks translation in both prokaryotes and eukaryotes. The effective doses required to induce translation inhibition without inducing apoptosis were extracted from the scientific literature.

Table 6.4 Antibiotics used to inhibit translation in HEK293 cells

Antibiotic	Concentration used in this study	Mechanism of action	Reference
Cycloheximide	100 µg/ml	Blocks translocation reaction on ribosomes	Harries *et al.*, 2004
Anisomycin	10 µg/ml	Blocks the peptidyl transferase reaction on ribosomes	Caputi *et al.*, 2002; Gatfield *et al.*, 2003; Harries *et al.*, 2004
Puromycin	20 µg/ml	Causes premature release of the nascent polypeptide by its addition to growing chain end.	Gatfield *et al.*, 2003

Unlike the mRNA stability assay carried out in section 6.4, which used cloned cDNAs, all transcripts assayed in this experiment were the native mRNAs produced from the endogenous human *PQBP1* gene. Changes in *PQBP1* expression levels were monitored over a 6 hour time course (see section 2.26). Total RNA was harvested from the cells at time points 0 (addition of antibiotic), 1, 2, 4 and 6 hours and used to synthesise cDNA. The abundance of various *PQBP1* alternative transcripts was quantified by real time PCR using SYBR green detection of double stranded DNA using the primer pairs from section 5.4 (Table 6.5).

All samples were normalised to the house keeping gene *GAPDH* and were expressed as fold-changes from the time point zero. Each real-time PCR was completed in triplicate and each experiment was repeated in duplicate. The results obtained are displayed in Figure 6.14.

Table 6.5 Primer pairs used to quantify alternative *PQBP1* transcripts

Primer Pair	Transcript	PTC	mRNA decay rates*	Predicted NMD**
Q10	All	Some	1.10	No
Q2b	1 15	1 - No 15 - Yes	1.65 1.78	No
Q3	4,6	Yes	0.79, 0.63	Yes
Q4	3,7,8,11	Yes	1.16,1.08	Yes
Q6	9,10,11	yes	0.62, 0.61	Yes

* if known
** based on location of PTC not mRNA decay rates

Throughout the time course experiment, distinct changes in the abundance of some of the *PQBP1* transcript variants were observed, and similar results were obtained with either anisomycin or cycloheximide, but different results were obtained when puromycin was used to block protein translation.

The abundance of all *PQBP1* transcripts amplified by primer pairs Q10 (all *PQBP1* variants identified in section 5.2) remained relatively constant when the cells were exposed to either anisomycin or cycloheximide. A 25% and 40% increase in the abundance of all *PQBP1* transcripts were observed over the six hour time course (anisomycin and cycloheximide respectively). As the reference transcript represents the overwhelming majority of all *PQBP1* transcripts (section 5.4), it is assumed that the expression patterns generated using Q10 primers reflect the stability of the *PQBP1* reference transcript. Slight changes in the expression profile may represent increases in the abundance of minor *PQBP1* transcripts that would normally be degraded by the NMD pathway.

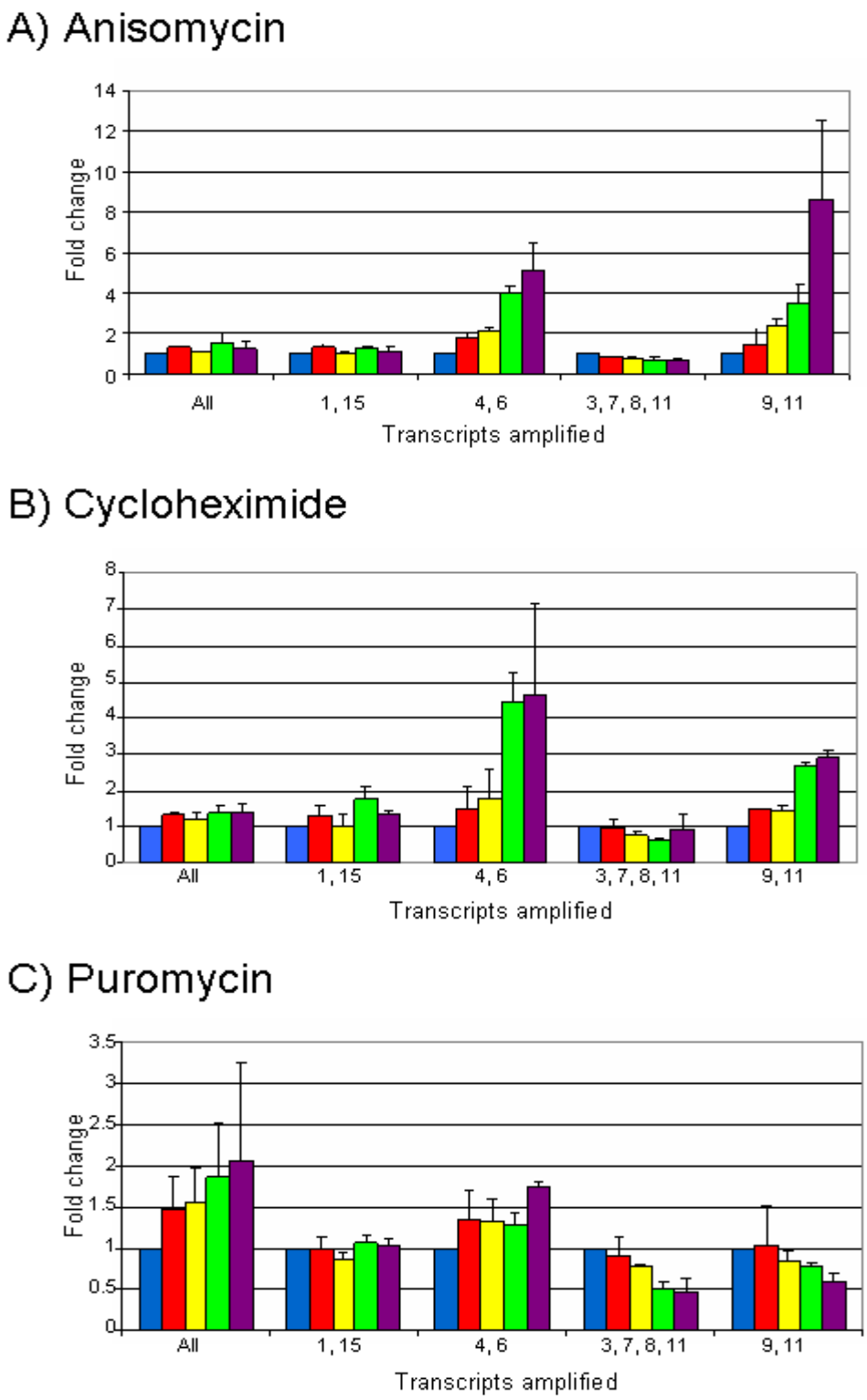


Figure 6.13 Effect on *PQBP1* transcript levels of inhibiting protein translation using anisomycin, cycloheximide and puromycin

Antibiotics were added to HEK293 cells and cell samples were taken over a 6 hour time course. At each time point 0 hr (blue), 1 hour (red), 2 hours (yellow), 4 hours (green) and 6 hours (purple). RNA was harvested and cDNA synthesised. The abundance of cDNA was measured by real-time PCR using SYBR-green fluorescence. The results were normalised to the house-keeping gene *GAPDH*.

PQBP1 transcripts amplified with primer pair Q2b (transcripts 1 and 15) also displayed very little change in their abundance throughout the six hour time course. A 35% and 13% increase in the abundance of transcripts amplified using this primer pair was observed after 6 hours incubation with either anisomycin or cycloheximide, respectively. These results suggest that the *PQBP1* transcripts amplified using primer pair Q2b are not targeted for degradation by the NMD pathway. It was anticipated that transcript 1 would be stable as it does not contain a PTC. Transcript 15 was expected to be degraded by NMD, *a priori,* as it contains a PTC. However, the data from the experiments performed in section 6.4 suggested that transcript 15 is more stable than the reference transcript and therefore may not be targeted for rapid degradation by the NMD pathway. It is possible that transcript 1 is more abundant than transcript 15, and that the changes in the abundance of transcript 15 are masked by the ability of transcript 1 to evade the NMD pathway. Further analysis is required using a primer pair that is able to differentiate between transcripts 1 and 15 in real-time PCR. Only then can the relative abundance of each transcript be described accurately.

PQBP1 transcripts amplified using primer pair Q4 (transcripts 3, 7, 8 and 11) also showed little change in transcripts abundance. The splicing variation shared by these transcripts is the inclusion intron 4. Here, a 9% (anisomycin) and 34% (cycloheximide) decrease in the transcript's abundance was observed. The slight decrease in these transcripts may be attributed to the excision of this intron from the mature mRNA transcript in the course of the experiment.

The abundance of transcripts 4 and 6 (primer pair Q3), 9 and 11 (primer pair Q6) increased in accordance with an increased time of exposure to anisomycin and cycloheximide. The increases after 6 hours were 5.2 fold (anisomycin) and 4.6 fold (cycloheximide) for primer pair Q3 and 8.7 (anisomycin) and 2.9 fold (cycloheximide) for primer pair Q6. Together, these results suggest that the abundance of transcripts 4, 6, 9 and 11 increased as a result of translation inhibition and they may be degraded by the NMD surveillance pathway. All of these transcripts contain PTCs and would be unable to encode a full length protein.

Changes in the abundance of *PQBP1* transcripts that resulted from blocking translation with puromycin had different profiles to those produced by anisomycin and cycloheximide. Here, a doubling in transcript levels after 6 hours was

observed with the Q10 primer pair, which amplifies all known *PQBP1* transcripts. The levels of transcript 1 and 15 (primer pair 2b) and transcripts 3, 7, 8 and 11 (primer pair 3) remained relatively constant over the time course of the experiment. A 3% and 75% increase was observed. Decreases in abundance were observed for transcripts 4 and 6 (primer pair Q4) and transcript 9 and 11 (primer pair Q6) 64% and 40% fold, respectively, at the 6 hour time point. None of the changes observed with puromycin are as striking as some of the changes observed with anisomycin and cycloheximide. Puromycin is known to be capable of inhibiting NMD (Gatfield *et al.,* 2003), but it may be ineffective in HEK293 cells. It is perhaps more likely that the concentration of puromycin used was not sufficient to repress translation effectively in these cells.

6.6 Discussion

This chapter describes further functional analysis of the *PQBP1* alternative transcripts that were identified in Chapter 5. Each isoform was analysed *in silico* to identify any structural alterations that would be predicted to result from alternative splicing. This was complemented by performing sub-cellular localisation of the *PQBP1* isoforms. In addition the *PQBP1* transcript variants were characterised for their ability to be produce stable mRNAs. A summary of the results obtained in this chapter is found in Table 6.6.

Overall, 56% of the *PQBP1* alternative transcripts are predicted to have an abated or altered function, while 44% of transcripts are potential targets of accelerated degradation via the NMD pathway. The functional significance of these observations is discussed in the following sections.

6.6.1 *Splicing patterns affect the sub-cellular location of PQBP1 transcript variants*

In silico analysis suggested that nine of the 15 *PQBP1* predicted isoforms had altered domain structures. One of the most striking observations was the exclusion of a putative nuclear localisation signal from eight of the 15 *PQBP1* isoforms. Sub-cellular localisation of the *PQBP1* isoforms demonstrated that this signal was required for the direction of *PQBP1* isoforms into the nucleus, which is consistent with other published investigations on four *PQBP1* transcripts (Iwamoto *et al.,* 2000).

Neither the expression nor sub-cellular localisation could be detected for isoforms 9, 10, 11 and 15. Several attempts were made to confirm the expression of these peptides and the positive results obtained with the other *PQBP1* isoforms eliminate the possibility of experimental error. The failure to confirm the expression of all *Alu* containing isoforms also suggests that this anomaly is not caused by experimental error but could be the result of the incorporation of an *Alu* fragment into the *PQBP1* transcript which destabilises the transcripts.

However, i*n silico* analysis of the isoforms encoded by transcripts 9, 10, 11 and 15 found that they all lack the WW domain (section 6.2) and it is possible that the absence of this domain may destabilise the protein. Without the WW domain it is plausible that the *PQBP1* isoforms may not fold correctly and this could lead to the

degradation of the unfolded protein rapidly. This domain is also crucial for the correct folding of other human proteins including the protein Yes-associated kinase protein, *h*Yap (Macias *et al.,* 1996). Mutational analysis of this protein has found that shorter constructs of the *h*Yap without the WW domain are either less stable than the wild-type domain or are unable to form any stable structure (Jiang *et al.,* 2001).

Additional analysis is required to determine if the variations in the domain structures of the *PQBP1* isoforms affects their biological function. This could be achieved by completing an assay specific to *PQBP1* function for example, the ability of the protein to bind polyglutamine tracts.

6.6.2 *Comparison of methods used to determine mRNA stability*

Two functionally distinct assays have been used to characterise the stability of the mRNA transcripts: the first utilised cloned cDNAs of *PQBP1* alternative transcripts to monitor mRNA decay rates, while the second interrupted cellular homeostasis by inhibiting translation, and then studied the effect on naturally occurring mRNA transcripts. The advantages of each approach are discussed below.

In the first series of experiments, general trends were observed for most of the transcripts studied, being a rapid increase in the transcript's abundance followed by gradual decay. Additional steps must to be taken to reduce the high level of inter-assay variation that was observed. This could be achieved by creating cell lines where the pTRE-TIGHT-*PQBP1* constructs are stabley integrated into the cell's genome. With this approach, exogenous effects on mRNA transcript abundance such as cell fitness or transfection efficiency would be minimised. Unfortunately, these experiments were not completed due to time constraints.

An obvious advantage of the tetracycline dependent expression system is the ability to arrest only the transcription of the exogenous *PQBP1* gene upon the addition of doxycycline, thus leaving the cell's physiology relatively undisturbed. However, the use of cloned cDNA samples in such assays has been the subject of criticism. Byers (2002) argued that recombinant DNA molecules were not biologically relevant, as they do not resemble naturally occurring spliced mRNA transcripts (Byers 2002). More specifically, cloned and transcribed cDNAs do not

bind proteins that normally bind to exon-exon junctions during splicing. The binding of spliceosomal and regulatory proteins to these junctions has been shown to have a role in identification of transcripts to be targeted by NMD (section 6.5).

Despite this argument, cDNAs containing PTCs have been shown to degrade more quickly than their wild-type counterparts. An exon junction-independent form of NMD has also been shown to degrade PTC-containing transcripts from the human genes β-globin (Couttet and Grange, 2004) Ig-u (Buhler *et al.,* 2004) and TCR-B (Wang *et al.,* 2002a). It is therefore highly probable that NMD in mammalian cells is controlled by multiple pathways that act at different levels with different efficiencies to ensure that PTC-containing mRNAs are eliminated from cells. Using intronless minigenes the transcript levels of *HEXA* harbouring PTCs were shown to be lower than the wild-type transcript (Rajavel and Neufeld 2001). However, these levels were not as low as those obtained for minigenes that contained introns and were therefore capable of binding mRNPs to spliced exon junctions. The results obtained in this chapter provide further support for exon-junction independent NMD, as in most cases transcripts harbouring a PTC degraded more quickly than full length transcripts.

In light of these observations, it was decided to seek additional support for the targeting of some *PQBP1* alternative transcripts that harbour a PTC by NMD by using naturally occurring mRNA species that are bound by mRNPs at exon junctions. Translation inhibitors were used to identify those transcripts that are stabilised when the ribosomal-associated PTC identification mechanism was suppressed. The problem associated with this experimental approach is the potential for pleiotropic effects of general translation inhibition. For example, it is crucial that the appropriate antibiotic concentrations are used to inhibit translation without inducing apoptosis. These figures were obtained from previous experiments on mammalian cells, but further optimisation of the puromycin concentration used in this assay is required to inhibit translation and therefore NMD. NMD was not inhibited for any of the transcripts using a puromycin concentration of 20 ng/µl.

Table 6.6 Summary of results obtained in this chapter

Transcript	Translation confirmed by Western Blotting	Subcellular localisation	mRNA half life (h)	Stabilisation with translation inhibitors	Altered /Abated Function?	NMD candidate?
Reference	Yes	Nucleus	1.10	No	No	No
1	Yes	Ubiquitous	1.65	No	Yes	No
2	Yes	Nucleus	1.27	No	No	No
3	Yes	Nucleus	N.C	No	No	No
4	Yes	Ubiquitous	0.79	Yes	Yes	Yes
5	Yes	Nucleus	N.C	N.C	No	No
6	Yes	Ubiquitous	0.63	Yes	Yes	Yes
7	Yes	Nucleus	1.16	No	No	No
8	N.C.	N.C	1.07	No	No*	No
9	No	Not detected	0.62	Yes	Yes	Yes
10	No	Not detected	0.61	Yes	Yes	Yes
11	No	Not detected	N.C.	Yes	Yes	Yes
12	Yes	Not detected	3.31	N.C	Yes	Yes
13	Yes	Nucleus	N.C	N.C	Yes*	No*
14	Yes	Nucleus	N.C	N.C	No	No
15	No	Not detected	1.78	No (partial)	Yes	Yes*

*Prediction based on *in silico* analysis.
NC - not completed
'Altered means relative to the function of the reference transcript.

The results obtained using cycloheximide and anisomycin supported the results obtained using the tet-off expression system, in that most transcripts containing a PTC were targeted for NMD. A discussion about the stability of the alternative *PQBP1* transcripts follows.

6.6.3 *Not all PQBP1 transcripts containing a PTC are targeted for rapid degradation.*

It has been proposed that susceptibility of transcripts containing PTCs to NMD can be predicted from sequence data alone (Hillman *et al.*, 2004). That is, transcripts containing a PTC 50-55 bp upstream from the ultimate exon junction would be identified by components of the NMD pathway and rapidly degraded. Results obtained in this chapter have demonstrated that a single set of rules cannot be applied to predict the stability of all mRNA transcripts, and that multiple mRNA

surveillance mechanisms probably serve to remove aberrantly spliced transcripts from eukaryotic cells.

Sequence analysis suggested that transcript 7 would be subjected to NMD, as it contains a PTC more than 50 bp upstream from the ultimate exon-junction. However, experiments described here did not provide evidence for decreased mRNA stability; the mRNA half life of the transcript is similar to that of the reference transcript (1.16 v 1.10 hours respectively) and also the transcript was not stabilised by translational inhibitors. *In silico* analysis of the protein encoded by this transcript predicted that it should contain the same structural motifs as the reference protein.

Experimental evidence did, however, demonstrate that on certain occasions the mRNA degradation profile for some of the *PQBP1* alternative transcripts could be predicted from sequence alone. A 132 bp deletion from exon 4 that was identified in transcripts 4 and 6 introduced a frameshift and a PTC. The mRNA half lives of these transcripts were reduced and the transcripts were also stabilised by the addition of anisomycin and cycloheximide to the cell media. Functional changes induced by alternative splicing of these transcripts were also observed experimentally, when the proteins produced from them failed to localise exclusively to the nucleus. Taken together the data suggests that these transcripts are unlikely to have a function and are removed by NMD.

In contrast, transcript 1, which has a deletion of exon 4 was not subjected to NMD. This was predicted as the deletion event did not introduce a PTC into the *PQBP1* variant. However, the functional impact of this alternative splicing event requires further investigation. This functional importance of this observation remains to be solved as the known functional roles of *PQBP1* are performed in the nucleus. Work carried out in chapter 5 suggested that exon 4 may have been acquired after divergence of lineages leading to mammals and fish. If exon 4 was acquired by an already established gene, it is possible that this acquisition introduced an additional nuclear function to the *PQBP1* gene and that non-nuclear *PQBP1* isoforms may also be capable of performing a biological function.

An unexpected result was the increased stability of transcript 15 whose half life was longer than that observed for the reference transcript (1.78 v 1.10 hours), but

whose cognate protein was not detected by either Western blotting or confocal microscopy. The predicted open reading frame encoded by this transcript was significantly shorter than the reference protein, being 76 amino acids in length, and computational analysis failed to identify any structural domains. This suggested that the encoded protein may not be capable of forming structural domains. Together, these results indicate that this transcript appears to be stable and perhaps the problem is more likely to lie with the stability of the encoded protein.

Additional experimental evidence is required to support enhanced stability of transcript 15. This could be obtained on two fronts; firstly the stability of transcript must be confirmed by additional experimental analysis, for example, using transcript specific primers to determine if the transcript is indeed subject to NMD. The transcript stability assay should be repeated and the transcript's stability should be determined using a functionally distinct assay to monitor the mRNA decay, for example by monitoring the rate of mRNA decapping, another indicator of mRNA stability (Couttet and Grange, 2004).

From the results obtained, it appears that the incorporation of an exon located within an *Alu* repeat (transcripts 9-11) has no functional relevance. Computational analysis identified two potential open reading frames for each transcript, but the potential proteins were not detected by Western blotting or sub-cellular localisation. Moreover, both transcript stability assays suggested that these transcripts were less stable than the reference transcript.

It has been hypothesized that the incorporation of *Alu* repeat fragments into mature mRNA species is facilitated by internal sequence motifs that resemble splice sites. Coupled with experimental validation, bioinformatic analysis of exons located within *Alu* repeats have determined that in many cases a point mutation can create the necessary sequence requirements to ensure efficient incorporation the *Alu* fragment into an mRNA transcript (Sorek et al., 2004a). Functional implications of the inclusion of *Alu* in transcripts suggest that apart from contributing to additional primate specific transcript diversity these transcripts may have the potential to cause genetic disorders including Dent's disease (Claverie-Martin *et al.,* 2003), Alzheimer's disease (Clarimon *et al.,* 2003) and Hunter disease (Ricci *et al.,* 2003).

6.6.4 *Conclusions*

This work extends upon expression analysis carried out in chapters 4 and 5, which illustrated that detailed cDNA screening is able to identify transcripts that are found at very low levels in the cell- levels that are much lower than the reference *PQBP1* transcript.

In this chapter it has been shown that some of the *PQBP1* transcript variants are less stable than the reference transcript. At least five of the *PQBP1* transcript variants were degraded rapidly, probably by the nonsense mediated decay pathway. *In silico* analysis of the *PQBP1* isoforms demonstrated that at least 8 isoforms lack domains that appear to the one known function of *PQBP1* (section 6.2). Isoforms without the WW domain failed to produce a protein in a transfection assay (section 6.3), while isoforms without a nuclear localisation signal were not directed to the nucleus following translation. Based on these diverse observations, the picture that seems to be emerging, for *PQBP1* is that much of the transcript diversity that can be detected does not create functional diversity and is more likely to be a result of aberrant splicing.

Chapter 7

Discussion

7.1 Summary

The preparation of this thesis coincided with significant advances in human genomics. The finished human genome sequence was published in October, 2004 (IHGSC, 2004). This achievement has been complemented with more detailed analysis of the sequences of individual chromosomes. At the time of writing, the mapping, sequencing and analysis had been completed for 16 human chromosomes; 2, 4, 5, 6, 7, 9, 10, 13, 14,16,19, 20, 21, 22, X and Y (Dunham *et al.,* 1999; Deloukas *et al.,* 2001; Hattori *et al.,* 2001; Heilig *et al.,* 2003; Hillier *et al.,* 2003; Mungall *et al.,* 2003; Skaletsky *et al.,* 2003; Deloukas *et al.,* 2004; Dunham *et al.,* 2004; Grimwood *et al.,* 2004; Humphray *et al.,* 2004; Martin *et al.,* 2004; Schmutz *et al.,* 2004; Hillier *et al.,* 2005; Ross *et al.,* 2005). It is anticipated that analysis of the remaining chromosomes will be completed in the near future.

The human genome sequence has been used in this thesis to annotate genes which will ultimately enhance our understanding of the complexity and diversity of transcript structures found within a 7.3 Mb region on the human X chromosome. In chapter 3, the human genome sequence was the primary substrate for annotation and preliminary analysis of the gene complement for human Xp11.22-p11.3. This work found that the region contains 77 known genes, 19 novel genes and five putative transcripts, including two antisense loci. In addition, 64 pseudogenes were identified. All gene structures can be accessed from the VEGA database (http://vega.sanger.ac.uk) and it is hoped that the annotated genome sequence will provide a useful resource for future functional studies. The majority of the annotated gene structures (> 65%) extended to a predicted transcription start site and/or a transcription termination site. All gene structures were annotated using evidence from full-length cDNA and EST sequences and during this process it became apparent that many of the genes were alternatively spliced. This observation highlighted the diversity of the human transcriptome, but it was also anticipated that more alternative transcripts remained to be identified as not all cDNA and EST libraries have not been comprehensively sequenced. In order to gain a comprehensive picture of the type and frequency of alternative splicing events in human Xp11.22-p11.3 a decision was made to complete a detailed investigation of transcript variation for a subset of the 101 annotated gene structures.

In chapter 4, a detailed study of transcript variation was carried out for 18 genes located in human Xp11.23. Three different strategies were employed to identify

novel transcripts. Comparative sequence analysis was carried out using genomic sequence from the orthologous region in the mouse where transcripts for each of the 18 orthologous gene pairs were compared with the aim of identifying mouse specific exons. In this way, twenty-one novel human candidate exons were identified, and expression was confirmed for seven of these. Additional transcript variants were also identified using more up-to-date EST and cDNA entries in human nucleotide databases. Finally, a targeted RT-PCR, cloning and sequencing strategy was employed to identify novel transcript fragments. The combination of these approaches, resulted in the identification of 61 novel transcript fragments, which approximately doubled the number found during the initial gene annotation process.

Chapter 5 describes the construction of a cloned open reading frame collection for the gene *PQBP1* which created a resource for functional studies. To create this collection, detailed sampling of cloned PCR products was performed to isolate sixteen variants of *PQBP1*, seven of which were also identified in chapter 4. As expected, more detailed cDNA sampling resulted in more transcript variants being identified. This increase in cDNA sampling also raised the concern that some transcripts may arise from errors in splicing. To distinguish between functional and non-functional transcripts a descriptive analysis of the *PQBP1* transcript variants followed, where the sequence composition and expression patterns of *PQBP1* variants were characterised. It is found that the alternative splice sites all had lower splice site scores than constitutive splice sites, and they appeared to be less conserved throughout vertebrate evolution. Finally, expression analysis confirmed that *PQBP1* is expressed ubiquitously and that the transcript variants represented less than 10% of all *PQBP1* transcripts. These experiments suggested specific splicing events gave rise to the *PQBP1* transcript variants. However, further functional studies would be required to assess the biological relevance of these variants.

Possible functional alterations in *PQBP1* transcript variants were analysed in chapter 6 with the aim of differentiating between transcripts generated through regulated splicing events and those generated through imprecise splicing events. Analysis of the encoded open reading frames found that 66% of the identified transcripts harboured PTCs, while most of the variants had altered domain structures. Transcript variation affected the sub-cellular localisation of at least

three *PQBP1* isoforms, which were not localised exclusively to the nucleus. Finally, mRNA stability assays were performed in order to identify transcripts that may be targeted for rapid degradation. By suppressing the expression of transiently transfected *PQBP1* alternative variants and monitoring the subsequent mRNA decay profiles over 8 hours, it was found that at least three variants have a shorter half-life than the reference *PQBP1* transcript. The results were confirmed by another assay where protein translation and hence NMD were inhibited. Together these results suggest that at least four of the *PQBP1* transcript variants may be targeted for rapid degradation via the NMD pathway.

7.2 The human genome sequence and alternative splicing

The finished human genome sequence is composed of long stretches of contiguous high-quality DNA sequence, which is a suitable framework for gene annotation and many other types of analysis. It has been used as a reference substrate to aid the completion of other mammalian genomes, such as that of the chimpanzee (Watanabe *et al.,* 2004), dog (Parker *et al.,* 2004) and mouse (Gregory *et al.,* 2002; Waterston *et al.,* 2002). The sequence has also been used as the reference substrate to study sequence variation (eg The SNP Consortium, Sachidanandam *et al.,* 2001; The International HapMap project, The International HapMap Consortium, 2003). However, one of the most important applications of the sequence is in the identification of all functional elements by a combination of experimental and computational methods. The Encyclopaedia of DNA elements (ENCODE) project, which was launched in 2003, aims to identify all functional elements in the human genome sequence (the Encode project consortium, 2004) including protein-coding genes, non-protein-coding genes, regulatory elements involved in the control of gene transcription and DNA sequences that mediate chromosomal structure and dynamics. The feasibility of identifying these features is currently being tested on a set of regions representing 1% of the total human genome sequence.

Now that the human genome has been completed the availability of finished sequence is no longer a limiting factor in gene identification. However, as more genes are being identified, the measures required to define the human gene content need to be increasingly sophisticated to ensure that genomic sequence is informatively analysed and that a bottle-neck in the analytical pipeline is not

created. Gene identification approaches commonly use a combination of techniques including sequence similarity searches, and *de novo* analyses to predict novel gene structures. These methods are used in concert to compensate for each other's shortfalls. For example, *de novo* analysis algorithms often over-estimate the number gene structures in the human genome. However, greater confidence can be gained from gene structure predicted by *de novo* analyses when it co-aligns on the genomic sequence with a transcribed sequence. In this capacity, gene identification strategies have benefited from the availability of many full-length cDNA sequences generated in high-throughput sequencing initiatives. Although these initiatives indicate that more human genes remain to be discovered, their utility is decreasing. For example, less than 9% of the 21,243 non-redundant transcript clusters sequenced by Ota and co-workers are novel and have ORFs greater than 300 bp in length (Ota *et al.,* 2004). It is possible that these genes have highly regulated expression patterns that restrict their expression to a short time period, they may be expressed a discrete location or have atypical sequence (e.g., unusual GC richness) are yet to be identified. Unidentified genes may be non-coding, and therefore missed by traditional *de novo* gene prediction programmes which have been trained to identify protein-coding genes. To date, little attention has been given to enhancing the identification of non-coding genes and it is anticipated that many novel non-coding genes need to be identified. Moreover, additional analysis is required to define the functional role of these genes.

It is anticipated that EST sequences will continue to be a valuable resource in defining human genes. Up to 40% of ESTs do not lie in known gene regions (Larsson *et al.,* 2005), some of which will undoubtedly provide evidence for novel gene structures. This may be achieved by combining EST sequence information with advanced gene prediction algorithms.

To identify protein coding genes the ENCODE project will use a targeted approach using computational gene predictions to guide subsequent experimental verification by RT-PCR and RACE analysis (The ENCODE Project Consortium, 2004). Particular effort will be given to genes and transcript variants likely to be underrepresented in the current catalogue of human genes; short and intronless genes, genes undergoing non-canonical splicing, selenoprotein genes (genes translating the TGA stop codon, into a selenocysteine residue), genes with unusual codon composition that may express at very low levels with a very restricted

pattern, human specific genes and genes evolving very rapidly, whose corresponding orthologues either do not exist in other species or are difficult to identify.

The utility of the genome sequence in defining the complexity of human gene complement has been achieved by aligning transcribed sequences to the human genome sequence. Here, it has been demonstrated that alternative splicing of pre-mature mRNA can produce a variety of different transcripts from the one gene. With the completion of the human genome sequence and the accurate annotation of its genic content the revised number of human genes has decreased by approximately 80% from 120,000 to between 20,000 and 25,000 (IHGSC, 2004). The unexpectedly low number of genes revealed through annotating the genome sequence has directed attention towards understanding how post-transcriptional and post-translational mechanisms in the mRNA and protein worlds serve to increase the number of functional products produced from the genome sequence.

Work presented in this thesis has also contributed towards a more comprehensive description of all transcript variants for 18 genes located in human Xp11.23. Eighty-nine percent of the genes targeted for detailed analysis demonstrated the capacity to produce transcript variants. Indeed, 125 transcripts were identified generated from the 18 genes which equates to an average of 6.9 transcripts per gene. If this figure were extrapolated to the entire genome it could be predicted that at least 17,800 human genes could produce 122,820 alternative transcripts (i.e. 17,800 genes x 6.9 transcripts). This number is quite close to the original estimate of 120,000 genes. The frequency of alternative splicing presented in this thesis is greater than other published and unpublished reports. For example, manual annotation of 8,043 known genes (genes with a full-length mRNA transcript and usually a LocusLink identifier) located on 14 finished human chromosomes has been completed by the HAVANA team at the WTSI. Here it has been found that at least 75% genes produce more than one transcript (J. Harrow, personal communication). These genes have an average of 3.8 transcripts per gene. From these figures it is suggested that the unusually high rate of alternative splicing in Xp11.23 may be attributed to the greater depth of sampling.

Random detailed sequencing of *PQBP1* ORFs illustrated the ease with which novel transcripts can be identified and it is anticipated that even more novel transcript would be identified if greater sampling of cDNA from more tissues was completed. However, at least four, and perhaps all, of the *PQBP1* alternative transcripts identified in this study were not functional. Detailed cDNA sequencing and novel transcript identification are intertwined: as more samples are sequenced it is increasingly likely that more novel variants will be identified. The appropriate level of cDNA sampling required to identify all functional transcripts remains to be solved because some known transcript variants are expressed in very low abundance and large amounts of random cDNA sampling would be required to identify them. Detailed sampling also increases the number of spurious transcripts identified. The number of novel transcripts identified is also dependent upon the function of the gene, the frequency at which the alternative splicing event occurs, how the alternative splicing event is regulated (if at all), and the type and number of tissues sampled.

Gene function may influence the plasticity of the loci. Through mechanisms that remain to be characterised, it is possible that the essential, ubiquitous functions of *PQBP1* may influence the diversity of transcripts it produces. This is supported by detailed transcript profiling of two functionally distinct genes DNA polymerase ß (*POLB*) and hypoxanthine phosphoribosyl transferase, (*HPRT*) (Skandalis and Uribe 2004). Approximately 40% of all *POLB* transcripts but only 1% of all *HRPT* transcripts were splice variants (Skandalis and Uribe, 2004). Are the transcripts functional? If not, how and why does the cell tolerate high levels aberrant splicing for certain types of genes?

One of the challenges arising from the apparently high degree of variation in the human transcriptome has been distinguishing between transcript variants generated through highly regulated splicing events and transcript variants produced by imprecise mRNA splicing events, or functional and non-functional mRNAs. Regulated mRNA alternative splicing events have been implicated in a variety of biological processes including cell division, immunological responses and sex determination. For example, in *Drosophila melanogaster* somatic sexual differentiation is accomplished by serial function of the products of sex-

determination genes. Sex-lethal (Sxl), is one such gene that is functionally expressed only in female flies. The sex-specific expression of this gene is regulated by alternative mRNA splicing which results in either the inclusion or exclusion of the translation stop codon containing third exon (Nagoshi *et al.,* 1988).

However, it is possible that transcripts that appear to be non-functional may indeed serve a biological purpose. For example, non-coding transcript variants may have a regulatory function. Although functional roles have been described for numerous non-coding transcripts, an example of a functionally active non-coding transcript variant could not be identified from the scientific literature. The lack of evidence, however, does not discount this possibility.

Alternatively, it is possible that many transcript variants are not functional. Non-functional mRNA transcripts are expected to occur at low frequency in order to minimise the energy expended on the synthesis and removal of aberrant transcripts (Kan *et al.,* 2002). They are often produced when the spliceosome "slips" and identifies cryptic splice sites in addition to, or instead of, normal splice sites. Detailed sampling may increase the chance of identifying imprecise mRNA transcripts which may be generated by a number of different mechanisms including, stochastic spliceosomal errors in splice site recognition, errors in the machinery that regulate mRNA splicing, mis-incorporation errors by RNA pol II during transcription or splice site selection errors due to transcription pausing at DNA lesions.

While it is anticipated that intricate biological processes such as mRNA splicing will have inherent error rates, it is not safe to say that all transcripts produced by unregulated splicing events are non-functional. Although present in small numbers, spurious splicing events may produce transcripts with either the same or novel function and therefore not have any deleterious effect on the cell. It has also been suggested that these splicing events may even represent an evolutionary process where transcript variation may work in parallel with other random mutation processes in an element of trial and error to promote molecular evolution (Kan *et al.,* 2002). Transcript variants generated through spurious splicing events which confer a selective advantage will be selected for.

Alternative pre-mRNA splicing is an important post-transcriptional event that increases protein diversity and may have contributed to the increase in the phenotypic complexity of metzazoans during evolution (Maniatis and Tasic 2002). This phenomenon is confined to higher, more complex eukaryotes and has not been observed in either *S. ceriviseae* or *S. pombe.* The appearance of alternative splicing has followed the appearance of introns and may have originated through a relaxation of the splice site recognition (Ast 2004). The inclusion of novel exons in spliced transcripts usually occurs by alternative splicing (Makalowski 2003), where proteins have the capacity to "test" the biological function of novel domains without compromising the function of the original protein (Gilbert 1978). Sub-optimal splice sites may be used in addition to constitutive splice sites. Two possible examples of this can be obtained from the *PQBP1* transcript variants.

Firstly, eight of the *PQBP1* transcript variants identified in this study lacked the arginine rich domain that binds to homopolymeric glutamine tracts of proteins such as Brain 2 (*BRN2*) and ataxin-1 (*ATXN1*) (Iwamoto *et al.,* 2000). Analysis of the putative *PQBP1* domains carried out in chapter 6 found that this domain is located within exon 4 and evolutionary analysis completed in chapter 5 found that this exon is the least conserved exon in *PQBP1.* The exon was not identified in the zebrafish . It is possible that the zebrafish *PQBP1* orthologue has lost its ability to bind to homopolymeric glutamine tracts. This notion is supported by the observation that zebrafish and Xenopus *Brn*-2 orthologues lack polyglutamine tracts that are conserved in mammalian species (Sumiyama *et al.,* 1996; Nakachi *et al.,* 1997) whereas the WW domain of *PQBP1* has remained conserved throughout evolution and has been identified in the nematode (Komuro *et al.,* 1999). Human transcript variants of *PQBP1* that also lack exon 4 (and hence the arginine rich domain) may represent a form of *PQBP1* that is found only in the zebrafish. Therefore it is possible that the human *PQBP1* transcript variant that lacks exon 4 may still be functional.

An example of the co-option (exaptation) of exons from intronic sequences is illustrated by the exonisation of an Alu repeat located in intron 2 of the *PQBP1* gene (found in transcripts 9-11). Comparative sequence analysis confirmed that this sequence was exclusive to primates. However, analysis of transcripts 9-11 found the inclusion of this exon introduces a PTC and destabilises the mRNA transcript. Exonisation of *Alu* elements has been reported in a number of different

human genes including tumour necrosis factor receptor gene type 2 (*p75TNFR*) (Singer *et al.,* 2004). This event produces a protein with a novel N-terminal domain, and novel function and demonstrates yet another way in which the diversity of the human transcriptome can be increased further.

Millions of years may be required after the integration of transposable elements both to fix these elements in the population and for them to undergo the sequence changes that lead to exonisation events. Given the relatively recent appearance of *Alu* repeats in primates, it is possible that they represent a way in which the size and diversity of primate transcriptomes can be increased. Although the frequency with which *Alu* elements are being incorporated into the human transcriptome remains to be determined, it has been noted that older *Alu* families are over-represented in exonisation events (Sorek *et al.,* 2002). Perhaps this observation could be due to the fact that there has been more time to chance upon the required changes to allow transcription (Sorek *et al.,* 2002). It is possible to speculate that with additional time this exon may acquire the necessary nucleotide substitutions to promote its inclusion in functional *PQBP1* transcripts.

7.3 Future directions

Possible techniques for enhancing the transcript map in human Xp11.23 have been discussed in various chapters throughout this thesis. However, it is pertinent to note that throughout the course of this study volumes of transcript information were deposited into the nucleotide databases that were generated predominantly by large-scale cDNA sequencing projects (Osato *et al.,* 2002; Ota *et al.,* 2004; Sogayar *et al.,* 2004). As a result, many of the novel transcripts identified in chapter 3 may have extended genes structures or may have been assessed by manual curators and classed as known genes. It is very likely that more novel transcripts that map to human Xp11.22-p11.3 have also been sequenced. Re-analysis of Xp11.22-p11.3 should first be completed to ensure that all subsequent work on human Xp11.22-p11.3 utilises all available sequence information. For example, recent annotation of genomic sequence in human Xp11.23 that was not available at the time of analysis identified two novel *MAGE* and seven novel *GAGE* genes that had not been annotated previously (Ross *et al.,* 2005).

It is also predicted that more sophisticated techniques than those employed in this thesis will be required to define and describe the human gene content of human Xp11.23, and the entire human genome, further. For instance, the construction of a DNA tiling array for human Xp11.23 could significantly advance our current understanding of the transcriptome in this region and could be used identify both novel intragenic and intergenic exons. Novel intergenic exons may be then be integrated into existing gene structures by RT-PCR and may represent novel transcription start or termination sites or they may represent novel genes. Much of the transcript analysis carried out in this thesis has focused on the identification and characterisation of transcripts with variable CDS structures. However, the analysis of transcript variation completed in chapter 4 also identified high levels of transcript variation in the untranslated regions of protein-coding genes. The use of an oligonucleotide DNA tiling array in concert with 5' and 3' RACE analysis may identify even more transcript variation in the UTRs. Novel intergenic regions may represent alternative exons or discrete transcriptional units.

Ideally this type of micro-array would be constructed for both strands of DNA in order to aid the identification of sense-antisense gene pairs. Two of these gene pairs in this region have already been identified in human transcripts (chapter 3) while an additional sense-antisense gene pair was also identified in an orthologous mouse gene, *Wdr13* (chapter 4). The figure falls below the predicted frequency of sense:antisense gene pairs of 20% (Chen *et al.,* 2005) and it is likely that more will be identified. Subsequent functional analysis of these transcripts using i*n vitro* coupled transcription-translation studies could also be completed to determine the impact, if any, of bi-directional expression on the human genes.

The superior detail of a gene's structure obtained through the analysis of transcript variation may reveal additional information about a gene's biological function. As discussed, various types of analysis have already been developed to analyse the expression patterns of alternative variants using small amounts of starting material (Shoemaker *et al.,* 2001; Yeakley *et al.,* 2002). Detailed transcript profiling could be completed with a microarray that contains the exon-junctions that were identified in this study. RNA could be extracted from various tissues and cell-types and these samples could then be hybridised to the microarrays. Such analyses

would enhance existing transcript maps and expression profiles for the genes in human Xp11.23.

As more sequence information becomes available it will become increasingly important to distinguish between functional and spurious transcript variants using sequence information alone. This could be further unveiled using comparative sequence analysis. It is proposed that the sequences of transcript variants could be extracted and analysed to identify sequence characteristics that are used in alternative splicing events. For example, a functional splice cite is likely to be conserved in closely related species while a cryptic/spurious splices is likely to have diverged. The completion of several eukaryotic genomes has already been used to advance the understanding of the molecular mechanisms that govern both splice site recognition and the use of alternative splice sites. For example, comparative sequence analysis between the human and mouse has been used to determine some of the sequence characteristics required to convert constitutive exons to alternative exons (reviewed by Ast, 2004). Much analysis completed to date, compared the conservation of splice sites between human and mouse, but the availability of genomic sequence from additional vertebrate species from varying evolutionary distances such as the dog, the cow, opossum or chicken may aid the identification of more novel transcript variants. This type of comparative analysis could be used to assess the conservation of alternative splice sites and the frequency of transcript variation in gene families that are renowned for their heterogeneity.

Work in this thesis has demonstrated how in-depth cDNA sampling can identify both functional and non-functional mRNA transcripts. Another possible avenue of analysis could be to discriminate between spurious and functional transcripts in a high-throughput manner. It is possible that the NMD inhibition experiment completed in chapter 6 (where protein translation and hence NMD was inhibited using antibiotics) could be scaled up to monitor changes in transcript abundance on a larger scale. The abundance of mRNAs transcribed under both normal and NMD inhibited conditions could be compared by labelling the transcripts with two different fluorescent dyes. The labelled mRNAs could then be hybridised to probes of both constitutive and alternative exon-junctions on a micro-array.

Ultimately, the *in vivo* functions of alternative transcripts should be assessed using either gene specific assays or gene knockout experiments. Clearly, testing thousands of transcripts in this fashion is a daunting task. This could be assisted by the use of i*n silico* predictions which can be used to provide clues about the functional implications of alternative splicing, but they cannot be substituted for experimental evidence. These predictions do, however, provide an ideal starting point for large scale analyses as they facilitate the generation of a working hypothesis.

In the future there may be a demand for high-throughput methods, such as RNAi or anti-sense strategies, to knock-out specific isoforms of a gene. When combined with relevant functional assays, it is predicted that this will a suitable means to decipher isoform function. Such RNAi knock-down has been shown to regulate transcription in a transcript specific fashion (Celotto *et al.,* 2005) and this combined with techniques to enable high throughput analysis (Tuschl and Borkhardt, 2002) could permit the characterisation of thousands of transcript variants. Functional alteration of transcript variants could also be tested using high throughput functional genomic techniques, such as reverse transfection, subcellular localisation or phosphorylation assays.

Results from research investigating the utility of other approaches (such as comparative genomics) in the prediction of functional alternative splicing events are only just beginning to emerge in scientific literature. Recently comparative sequence analysis has been used to identify conserved alternative splicing events between the human and mouse (Modrek and Lee 2003; Nurtdinov *et al.,* 2003; Sorek and Ast 2003). This work has mainly focused on exon-skipping events (cassette exons) and is based on the hypothesis the conserved alternative exons are more likely to be functional as they are under selective pressures to remain conserved. In general, these cassette exons maintain an ORF (Thanaraj and Stamm 2003; Sorek *et al.,* 2004) and their length is divisible by three (Sorek *et al., 2004;* Modrek and Lee, 2003). While additional work is required to characterise the conservation of other types of splicing events that can produce transcript variants, such as partial exon additions or deletions, it is hoped that this approach will ultimately be used for *de novo* prediction of an alternative splicing event.

The studies presented in this thesis, combined with published literature, demonstrate the dynamic nature of the human transcriptome. It appears that splicing generates a large number of variants whose function are not known, and it is likely that some of these will be the result of aberrant splicing events. Thus, alternative splicing serves at least two roles in eukaryotic cells (Boue *et al.,* 2003). It is an economical way to create additional diversity and specificity within a cell and can be regulated in either a spatial or temporal fashion. When associated with mRNA surveillance pathways such as NMD, alternative splicing serves to providing a testing ground for the evolution of gene structures. Future work will be required on both a genome-wide level as well as on individual genes to determine the cellular mechanisms that modulate mRNA splicing, and describe the functional consequences of alternative splicing events. It is predicted that a more complete understanding of both functional and aberrant alternative splicing events will contribute towards a greater understanding of human biology and disease.

Bibliography

Aagaard, L., G. Laible, P. Selenko, M. Schmid, R. Dorn, G. Schotta, S. Kuhfittig, A. Wolf, A. Lebersorger, P. B. Singh, G. Reuter and T. Jenuwein (1999). "Functional mammalian homologues of the Drosophila PEV-modifier Su(var)3-9 encode centromere-associated proteins which complex with the heterochromatin component M31." Embo J **18**(7): 1923-38.

Adams, M. D., S. E. Celniker, R. A. Holt, C. A. Evans, J. D. Gocayne, P. G. Amanatides, S. E. Scherer, P. W. Li, R. A. Hoskins, R. F. Galle, R. A. George, S. E. Lewis, S. Richards, M. Ashburner, S. N. Henderson, G. G. Sutton, J. R. Wortman, M. D. Yandell, Q. Zhang, L. X. Chen, R. C. Brandon, Y. H. Rogers, R. G. Blazej, M. Champe, B. D. Pfeiffer, K. H. Wan, C. Doyle, E. G. Baxter, G. Helt, C. R. Nelson, G. L. Gabor, J. F. Abril, A. Agbayani, H. J. An, C. Andrews-Pfannkoch, D. Baldwin, R. M. Ballew, A. Basu, J. Baxendale, L. Bayraktaroglu, E. M. Beasley, K. Y. Beeson, P. V. Benos, B. P. Berman, D. Bhandari, S. Bolshakov, D. Borkova, M. R. Botchan, J. Bouck, P. Brokstein, P. Brottier, K. C. Burtis, D. A. Busam, H. Butler, E. Cadieu, A. Center, I. Chandra, J. M. Cherry, S. Cawley, C. Dahlke, L. B. Davenport, P. Davies, B. de Pablos, A. Delcher, Z. Deng, A. D. Mays, I. Dew, S. M. Dietz, K. Dodson, L. E. Doup, M. Downes, S. Dugan-Rocha, B. C. Dunkov, P. Dunn, K. J. Durbin, C. C. Evangelista, C. Ferraz, S. Ferriera, W. Fleischmann, C. Fosler, A. E. Gabrielian, N. S. Garg, W. M. Gelbart, K. Glasser, A. Glodek, F. Gong, J. H. Gorrell, Z. Gu, P. Guan, M. Harris, N. L. Harris, D. Harvey, T. J. Heiman, J. R. Hernandez, J. Houck, D. Hostin, K. A. Houston, T. J. Howland, M. H. Wei, C. Ibegwam, M. Jalali, F. Kalush, G. H. Karpen, Z. Ke, J. A. Kennison, K. A. Ketchum, B. E. Kimmel, C. D. Kodira, C. Kraft, S. Kravitz, D. Kulp, Z. Lai, P. Lasko, Y. Lei, A. A. Levitsky, J. Li, Z. Li, Y. Liang, X. Lin, X. Liu, B. Mattei, T. C. McIntosh, M. P. McLeod, D. McPherson, G. Merkulov, N. V. Milshina, C. Mobarry, J. Morris, A. Moshrefi, S. M. Mount, M. Moy, B. Murphy, L. Murphy, D. M. Muzny, D. L. Nelson, D. R. Nelson, K. A. Nelson, K. Nixon, D. R. Nusskern, J. M. Pacleb, M. Palazzolo, G. S. Pittman, S. Pan, J. Pollard, V. Puri, M. G. Reese, K. Reinert, K. Remington, R. D. Saunders, F. Scheeler, H. Shen, B. C. Shue, I. Siden-Kiamos, M. Simpson, M. P. Skupski, T. Smith, E. Spier, A. C. Spradling, M. Stapleton, R. Strong, E. Sun, R. Svirskas, C. Tector, R. Turner, E. Venter, A. H. Wang, X. Wang, Z. Y. Wang, D. A. Wassarman, G. M. Weinstock, J. Weissenbach, S. M. Williams, WoodageT, K. C. Worley, D. Wu, S. Yang, Q. A. Yao, J. Ye, R. F. Yeh, J. S. Zaveri, M. Zhan, G. Zhang, Q. Zhao, L. Zheng, X. H. Zheng, F. N. Zhong, W. Zhong, X. Zhou, S. Zhu, X. Zhu, H. O. Smith, R. A. Gibbs, E. W. Myers, G. M. Rubin and J. C. Venter (2000). "The genome sequence of Drosophila melanogaster." Science **287**(5461): 2185-95.

Altieri, M., F. Marini, R. Arban, G. Vitulli and B. O. Jansson (2004). "Expression analysis of brain-derived neurotrophic factor (BDNF) mRNA isoforms after chronic and acute antidepressant treatment." Brain Res **1000**(1-2): 148-55.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman (1990). "Basic local alignment search tool." J Mol Biol **215**(3): 403-10.

Amores, A., A. Force, Y. L. Yan, L. Joly, C. Amemiya, A. Fritz, R. K. Ho, J. Langeland, V. Prince, Y. L. Wang, M. Westerfield, M. Ekker and J. H. Postlethwait (1998). "Zebrafish hox clusters and vertebrate genome evolution." Science **282**(5394): 1711-4.

Anderson, C. L. and C. J. Brown (2002). "Variability of X chromosome inactivation: effect on levels of TIMP1 RNA and role of DNA methylation." Hum Genet **110**(3): 271-8.

Arhondakis, S., F. Auletta, G. Torelli and G. D'Onofrio (2004). "Base composition and expression level of human genes." Gene **325**: 165-9.

Ashurst, J. L., C. K. Chen, J. G. Gilbert, K. Jekosch, S. Keenan, P. Meidl, S. M. Searle, J. Stalker, R. Storey, S. Trevanion, L. Wilming and T. Hubbard (2005). "The Vertebrate Genome Annotation (Vega) database." Nucleic Acids Res **33**(Database issue): D459-65.

Assink, J. J., N. T. Tijmes, J. B. ten Brink, R. J. Oostra, F. C. Riemslag, P. T. de Jong and A. A. Bergen (1997). "A gene for X-linked optic atrophy is closely linked to the Xp11.4-Xp11.2 region of the X chromosome." Am J Hum Genet **61**(4): 934-9.

Ast, G. (2004). "How did alternative splicing evolve?" Nat Rev Genet **5**(10): 773-82.

Avner, P. and E. Heard (2001). "X-chromosome inactivation: counting, choice and initiation." Nat Rev Genet **2**(1): 59-67.

Bajic, V. B., S. L. Tan, Y. Suzuki and S. Sugano (2004). "Promoter prediction analysis on the whole human genome." Nat Biotechnol **22**(11): 1467-73.

Barrass, J. D. and J. D. Beggs (2003). "Splicing goes global." Trends Genet **19**(6): 295-8.

Bass, B. L. (2000). "Double-stranded RNA as a template for gene silencing." Cell **101**(3): 235-8.

Bass, B. L. (2001). "RNA interference. The short answer." Nature **411**(6836): 428-9.

Bauer, M. F., K. Gempel, A. S. Reichert, G. A. Rappold, P. Lichtner, K. D. Gerbitz, W. Neupert, M. Brunner and S. Hofmann (1999). "Genetic and structural characterization of the human mitochondrial inner membrane translocase." J Mol Biol **289**(1): 69-82.

Baytel, D., S. Shalom, I. Madgar, R. Weissenberg and J. Don (1998). "The human Pim-2 proto-oncogene and its testicular expression." Biochim Biophys Acta **1442**(2-3): 274-85.

Beaudoing, E., S. Freier, J. R. Wyatt, J. M. Claverie and D. Gautheret (2000). "Patterns of variant polyadenylation signal usage in human genes." Genome Res **10**(7): 1001-10.

Bentley, D. (2002). "The mRNA assembly line: transcription and processing machines in the same factory." Curr Opin Cell Biol **14**(3): 336-42.

Bentley, D. R., P. Deloukas, A. Dunham, L. French, S. G. Gregory, S. J. Humphray, A. J. Mungall, M. T. Ross, N. P. Carter, I. Dunham, C. E. Scott, K. J. Ashcroft, A. L. Atkinson, K. Aubin, D. M. Beare, G. Bethel, N. Brady, J. C. Brook, D. C. Burford, W. D. Burrill, C. Burrows, A. P. Butler, C. Carder, J. J. Catanese, C. M. Clee, S. M. Clegg, V. Cobley, A. J. Coffey, C. G. Cole, J. E. Collins, J. S. Conquer, R. A. Cooper, K. M. Culley, E. Dawson, F. L. Dearden, R. M. Durbin, P. J. de Jong, P. D. Dhami, M. E. Earthrowl, C. A. Edwards, R. S. Evans, C. J. Gillson, J. Ghori, L. Green, R. Gwilliam, K. S. Halls, S. Hammond, G. L. Harper, R. W. Heathcott, J. L. Holden, E. Holloway, B. L. Hopkins, P. J. Howard, G. R. Howell, E. J. Huckle, J. Hughes, P. J. Hunt, S. E. Hunt, M. Izmajlowicz, C. A. Jones, S. S. Joseph, G. Laird, C. F. Langford, M. H. Lehvaslaiho, M. A. Leversha, O. T. McCann, L. M. McDonald, J. McDowall, G. L. Maslen, D. Mistry, N. K. Moschonas, V. Neocleous, D. M. Pearson, K. J. Phillips, K. M. Porter, S. R. Prathalingam, Y. H. Ramsey, S. A. Ranby, C. M. Rice, J. Rogers, L. J. Rogers, T. Sarafidou, D. J. Scott, G. J. Sharp, C. J. Shaw-Smith, L. J. Smink, C. Soderlund, E. C. Sotheran, H. E. Steingruber, J. E. Sulston, A. Taylor, R. G. Taylor, A. A. Thorpe, E. Tinsley, G. L. Warry, A. Whittaker, P. Whittaker, S. H. Williams, T. E. Wilmer, R. Wooster and C. L. Wright (2001). "The physical maps for sequencing human chromosomes 1, 6, 9, 10, 13, 20 and X." Nature **409**(6822): 942-3.

Berget, S. M. (1995). "Exon recognition in vertebrate splicing." J Biol Chem **270**(6): 2411-4.

Berget, S. M. and P. A. Sharp (1977). "A spliced sequence at the 5'-terminus of adenovirus late mRNA." Brookhaven Symp Biol (**29**): 332-44.

Bird, A. P. (1986). "CpG-rich islands and the function of DNA methylation." Nature **321**(6067): 209-13.

Blanchette, M. and B. Chabot (1999). "Modulation of exon skipping by high-affinity hnRNP A1-binding sites and by intron elements that repress splice site utilization." Embo J **18**(7): 1939-52.

Boguski, M. S. (1995). "The turning point in genome research." Trends Biochem Sci **20**(8): 295-6.

Boguski, M. S. and G. D. Schuler (1995). "ESTablishing a human transcript map." Nat Genet **10(**4): 369-71.

Boguski, M. S., T. M. Lowe and C. M. Tolstoshev (1993). "dbEST--database for "expressed sequence tags"." Nat Genet **4**(4): 332-3.

Botstein, D., R. L. White, M. Skolnick and R. W. Davis (1980). "Construction of a genetic linkage map in man using restriction fragment length polymorphisms." Am J Hum Genet **32**(3): 314-31.

Boue, S., I. Letunic and P. Bork (2003). "Alternative splicing and evolution." Bioessays **25**(11): 1031-4.

Brenner, S. (1990). "The human genome: the nature of the enterprise." Ciba Found Symp **149**: 6-12; discussion 12-7.

Brent, M. R. and R. Guigo (2004). "Recent advances in gene structure prediction." Curr Opin Struct Biol **14(**3): 264-72.

Brett, D., J. Hanke, G. Lehmann, S. Haase, S. Delbruck, S. Krueger, J. Reich and P. Bork (2000). "EST comparison indicates 38% of human mRNAs contain possible alternative splice forms." FEBS Lett **474**(1): 83-6.

Brinkmann, U., G. Vasmatzis, B. Lee, N. Yerushalmi, M. Essand and I. Pastan (1998). "PAGE-1, an X chromosome-linked GAGE-like gene that is expressed in normal and neoplastic prostate, testis, and uterus." Proc Natl Acad Sci U S A **95**(18): 10757-62.

Bristow, J., S. E. Gitelman, M. K. Tee, B. Staels and W. L. Miller (1993). "Abundant adrenal-specific transcription of the human P450c21A "pseudogene"." J Biol Chem **268**(17): 12919-24.Brunkow, M. E., E. W. Jeffery, K. A. Hjerrild, B. Paeper, L. B. Clark, S. A. Yasayko, J. E. Wilkinson, D. Galas, S. F. Ziegler and F. Ramsdell (2001). "Disruption of a new forkhead/winged-helix protein, scurfin, results in the fatal lymphoproliferative disorder of the scurfy mouse." Nat Genet **27**(1): 68-73.

Brunkow,M.E., E.W. Jeffery, K.A. Hjerrild, B. Paeper, L.B. Clark, S.A Yasayko, J.E. Wilkinson, D. Galas, S.F. Ziegler, F. Ramsdell (2001) "Disruption of a new forkhead/winged-helix protein, scurfin, results in the fatal lymphoproliferative disorder of the scurfy mouse." Nat Genet **27**(1):68-73.

Bucher, P. and A. Bairoch (1994). "A generalized profile syntax for biomolecular sequence motifs and its function in automatic sequence interpretation." Proc Int Conf Intell Syst Mol Biol **2**: 53-61.

Buckanovich, R. J., J. B. Posner and R. B. Darnell (1993). "Nova, the paraneoplastic Ri antigen, is homologous to an RNA-binding protein and is specifically expressed in the developing motor system." Neuron **11(**4): 657-72.

Buhler, M., A. Paillusson and O. Muhlemann (2004). "Efficient downregulation of immunoglobulin mu mRNA with premature translation-termination codons requires the 5'-half of the VDJ exon." Nucleic Acids Res **32**(11): 3304-15.

Buratti, E. and F. E. Baralle (2004). "Influence of RNA secondary structure on the pre-mRNA splicing process." Mol Cell Biol **24**(24): 10505-14.

Burge, C. and S. Karlin (1997). "Prediction of complete gene structures in human genomic DNA." J Mol Biol **268**(1): 78-94.

Busch, A., S. Engemann, R. Lurz, H. Okazawa, H. Lehrach and E. E. Wanker (2003). "Mutant huntingtin promotes the fibrillogenesis of wild-type huntingtin: a potential mechanism for loss of huntingtin function in Huntington's disease." J Biol Chem **278**(42): 41452-61.

Byers, P. H. (2002). "Killing the messenger: new insights into nonsense-mediated mRNA decay." J Clin Invest **109**(1): 3-6.

Caccio, S., K. Jabbari, G. Matassi, F. Guermonprez, J. Desgres and G. Bernardi (1997). "Methylation patterns in the isochores of vertebrate genomes." Gene **205**(1-2): 119-24.

Caplen, N. J., S. Parrish, F. Imani, A. Fire and R. A. Morgan (2001). "Specific inhibition of gene expression by small double-stranded RNAs in invertebrate and vertebrate systems." Proc Natl Acad Sci U S A **98(**17): 9742-7.

Caputi, M., R. J. Kendzior, Jr. and K. L. Beemon (2002). "A nonsense mutation in the fibrillin-1 gene of a Marfan syndrome patient induces NMD and disrupts an exonic splicing enhancer." Genes Dev **16**(14): 1754-9.

Caricasole, A., T. Ferraro, J. M. Rimland and G. C. Terstappen (2002). "Molecular cloning and initial characterization of the MG61/PORC gene, the human homologue of the Drosophila segment polarity gene Porcupine." Gene **288**(1-2): 147-57.

Carmel, I., S. Tal, I. Vig and G. Ast (2004). "Comparative analysis detects dependencies among the 5' splice-site positions." RNA **10**(5): 828-40.

Carninci, P. and Y. Hayashizaki (1999). "High-efficiency full-length cDNA cloning." Methods Enzymol **303**: 19-44.

Carninci, P., T. Shiraki, Y. Mizuno, M. Muramatsu and Y. Hayashizaki (2002). "Extra-long first-strand cDNA synthesis." Biotechniques **32**(5): 984-5.

Carninci, P., Y. Shibata, N. Hayatsu, M. Itoh, T. Shiraki, T. Hirozane, A. Watahiki, K. Shibata, H. Konno, M. Muramatsu and Y. Hayashizaki (2001). "Balanced-size and long-size

cloning of full-length, cap-trapped cDNAs into vectors of the novel lambda-FLC family allows enhanced gene discovery rate and functional analysis." Genomics **77**(1-2): 79-90.

Carrel, L. and H. F. Willard (2005). "X-inactivation profile reveals extensive variability in X-linked gene expression in females." Nature **434**(7031): 400-4.

Carrel, L., A. A. Cottle, K. C. Goglin and H. F. Willard (1999). "A first-generation X-inactivation profile of the human X chromosome." Proc Natl Acad Sci U S A **96**(25): 14440-4.

Castle, J., P. Garrett-Engele, C. D. Armour, S. J. Duenwald, P. M. Loerch, M. R. Meyer, E. E. Schadt, R. Stoughton, M. L. Parrish, D. D. Shoemaker and J. M. Johnson (2003). "Optimization of oligonucleotide arrays and RNA amplification protocols for analysis of transcript structure and alternative splicing." Genome Biol **4(**10): R66.

Cederbaum, S. D., H. Yu, W. W. Grody, R. M. Kern, P. Yoo and R. K. Iyer (2004). "Arginases I and II: do their functions overlap?" Mol Genet Metab **81** Suppl 1: S38-44.

Celotto, A.M., J. W. Lee and B.R. Graveley (2005). Exon-specific RNA Interference: A tool to determine the functional relevance of proteins encoded by alternatively spliced mRNAs. Methods Mol Biol **309**:273-82.

Ceulemans, H., W. Stalmans and M. Bollen (2002). "Regulator-driven functional diversification of protein phosphatase-1 in eukaryotic evolution." Bioessays **24**(4): 371-81.

Chappell, S. A., G. C. Owens and V. P. Mauro (2001). "A 5' Leader of Rbm3, a Cold Stress-induced mRNA, Mediates Internal Initiation of Translation with Increased Efficiency under Conditions of Mild Hypothermia." J Biol Chem **276**(40): 36917-22.

Charlesworth, B. (1991). "The evolution of sex chromosomes." Science **251**(4997): 1030-3.

Chelly, J., Z. Tumer, T. Tonnesen, A. Petterson, Y. Ishikawa-Brush, N. Tommerup, N. Horn and A. P. Monaco (1993). "Isolation of a candidate gene for Menkes disease that encodes a potential heavy metal binding protein." Nat Genet **3**(1): 14-9.

Chen, J., M. Sun, W. J. Kent, X. Huang, H. Xie, W. Wang, G. Zhou, R. Z. Shi and J. D. Rowley (2004). "Over 20% of human transcripts might form sense-antisense pairs." Nucleic Acids Res **32**(16): 4812-20.

Chen, M. E., S. H. Lin, L. W. Chung and R. A. Sikes (1998). "Isolation and characterization of PAGE-1 and GAGE-7. New genes expressed in the LNCaP prostate cancer progression model that share homology with melanoma-associated antigens." J Biol Chem **273**(28): 17618-25.

Chen, Y. T., M. J. Scanlan, C. A. Venditti, R. Chua, G. Theiler, B. J. Stevenson, C. Iseli, A. O. Gure, T. Vasicek, R. L. Strausberg, C. V. Jongeneel, L. J. Old and A. J. Simpson (2005). "Identification of cancer/testis-antigen genes by massively parallel signature sequencing." Proc Natl Acad Sci U S A **102**(22): 7940-5.

Chen, H. I. and M. Sudol (1995). "The WW domain of Yes-associated protein binds a proline-rich ligand that differs from the consensus established for Src homology 3-binding modules." Proc Natl Acad Sci U S A **92**(17): 7819-23.

Cheng, J., P. Kapranov, J. Drenkow, S. Dike, S. Brubaker, S. Patel, J. Long, D. Stern, H. Tammana, G. Helt, V. Sementchenko, A. Piccolboni, S. Bekiranov, D. K. Bailey, M. Ganesh, S. Ghosh, I. Bell, D. S. Gerhard and T. R. Gingeras (2005). "Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution." Science **308**(5725): 1149-54.

Chiurazzi, P., B. C. Hamel and G. Neri (2001). "XLMR genes: update 2000." Eur J Hum Genet **9**(2): 71-81.

Choo, K. H., K. G. Gould, D. J. Rees and G. G. Brownlee (1982). "Molecular cloning of the gene for human anti-haemophilic factor IX." Nature **299**(5879): 178-80.

Christoffels, A., E. G. Koh, J. M. Chia, S. Brenner, S. Aparicio and B. Venkatesh (2004). "Fugu genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes." Mol Biol Evol **21**(6): 1146-51.

Clarimon, J., J. Bertranpetit, F. Calafell, M. Boada, L. Tarraga and D. Comas (2003). "Association study between Alzheimer's disease and genes involved in Abeta biosynthesis, aggregation and degradation: suggestive results with BACE1." J Neurol **250**(8): 956-61.

Clark, F. and T. A. Thanaraj (2002). "Categorization and characterization of transcript-confirmed constitutively and alternatively spliced introns and exons from human." Hum Mol Genet **11**(4): 451-64.

Clark, J., P. J. Rocques, A. J. Crew, S. Gill, J. Shipley, A. M. Chan, B. A. Gusterson and C. S. Cooper (1994). "Identification of novel genes, SYT and SSX, involved in the t(X;18)(p11.2;q11.2) translocation found in human synovial sarcoma." Nat Genet **7**(4): 502-8.

Claverie-Martin, F., H. Gonzalez-Acosta, C. Flores, M. Anton-Gamero and V. Garcia-Nieto (2003). "De novo insertion of an Alu sequence in the coding region of the CLCN5 gene results in Dent's disease." Hum Genet 113**(**6): 480-5.

Coleman, M. P., A. H. Nemeth, L. Campbell, C. P. Raut, J. Weissenbach and K. E. Davies (1994). "A 1.8-Mb YAC contig in Xp11.23: identification of CpG islands and physical mapping of CA repeats in a region of high gene density." Genomics **21**(2): 337-43.

Collins, J. E., M. E. Goward, C. G. Cole, L. J. Smink, E. J. Huckle, S. Knowles, J. M. Bye, D. M. Beare and I. Dunham (2003). "Reevaluating human gene annotation: a second-generation analysis of chromosome 22." Genome Res **13**(1): 27-36.

Collins, J. E., C. L. Wright, C. A. Edwards, M. P. Davis, J. A. Grinham, C. G. Cole, M. E. Goward, B. Aguado, M. Mallya, Y. Mokrab, E. J. Huckle, D. M. Beare and I. Dunham (2004). "A genome annotation-driven approach to cloning the human ORFeome." Genome Biol **5**(10): R84.

Coulson, A., Sulston, J., Brenner, S., and J. Karn (1986). "Towards a physical map of the genome of the nematode *C. elegans*. Proc Natl Acad Sci **83:** 7821-7825.

Couttet, P. and T. Grange (2004). "Premature termination codons enhance mRNA decapping in human cells." Nucleic Acids Res **32**(2): 488-94.

Cuppens, H. and J. J. Cassiman (2004). "CFTR mutations and polymorphisms in male infertility." Int J Androl **27**(5): 251-6.

Dahl, H. H., R. M. Brown, W. M. Hutchison, C. Maragos and G. K. Brown (1990). "A testis-specific form of the human pyruvate dehydrogenase E1 alpha subunit is coded for by an intronless gene on chromosome 4." Genomics **8**(2): 225-32.

Davuluri, R. V., I. Grosse and M. Q. Zhang (2001). "Computational identification of promoters and first exons in the human genome." Nat Genet 29(4): 412-7.

Dehal, P., P. Predki, A. S. Olsen, A. Kobayashi, P. Folta, S. Lucas, M. Land, A. Terry, C. L. Ecale Zhou, S. Rash, Q. Zhang, L. Gordon, J. Kim, C. Elkin, M. J. Pollard, P. Richardson, D. Rokhsar, E. Uberbacher, T. Hawkins, E. Branscomb and L. Stubbs (2001). "Human chromosome 19 and related regions in mouse: conservative and lineage-specific evolution." Science **293**(5527): 104-11.

Deloukas, P., L. H. Matthews, J. Ashurst, J. Burton, J. G. Gilbert, M. Jones, G. Stavrides, J. P. Almeida, A. K. Babbage, C. L. Bagguley, J. Bailey, K. F. Barlow, K. N. Bates, L. M. Beard, D. M. Beare, O. P. Beasley, C. P. Bird, S. E. Blakey, A. M. Bridgeman, A. J. Brown, D. Buck, W. Burrill, A. P. Butler, C. Carder, N. P. Carter, J. C. Chapman, M. Clamp, G. Clark, L. N. Clark, S. Y. Clark, C. M. Clee, S. Clegg, V. E. Cobley, R. E. Collier, R. Connor, N. R. Corby, A. Coulson, G. J. Coville, R. Deadman, P. Dhami, M. Dunn, A. G. Ellington, J. A. Frankland, A. Fraser, L. French, P. Garner, D. V. Grafham, C. Griffiths, M. N. Griffiths, R. Gwilliam, R. E. Hall, S. Hammond, J. L. Harley, P. D. Heath, S. Ho, J. L. Holden, P. J. Howden, E. Huckle, A. R. Hunt, S. E. Hunt, K. Jekosch, C. M. Johnson, D. Johnson, M. P. Kay, A. M. Kimberley, A. King, A. Knights, G. K. Laird, S. Lawlor, M. H. Lehvaslaiho, M. Leversha, C. Lloyd, D. M. Lloyd, J. D. Lovell, V. L. Marsh, S. L. Martin, L. J. McConnachie, K. McLay, A. A. McMurray, S. Milne, D. Mistry, M. J. Moore, J. C. Mullikin, T. Nickerson, K. Oliver, A. Parker, R. Patel, T. A. Pearce, A. I. Peck, B. J. Phillimore, S. R. Prathalingam, R. W. Plumb, H. Ramsay, C. M. Rice, M. T. Ross, C. E. Scott, H. K. Sehra, R. Shownkeen, S. Sims, C. D. Skuce, M. L. Smith, C. Soderlund, C. A. Steward, J. E. Sulston, M. Swann, N. Sycamore, R. Taylor, L. Tee, D. W. Thomas, A. Thorpe, A. Tracey, A. C. Tromans, M. Vaudin, M. Wall, J. M. Wallis, S. L. Whitehead, P. Whittaker, D. L. Willey, L. Williams, S. A. Williams, L. Wilming, P. W. Wray, T. Hubbard, R. M. Durbin, D. R. Bentley, S. Beck and J. Rogers (2001). "The DNA sequence and comparative analysis of human chromosome 20." Nature **414**(6866): 865-71.

Deloukas, P., M. E. Earthrowl, D. V. Grafham, M. Rubenfield, L. French, C. A. Steward, S. K. Sims, M. C. Jones, S. Searle, C. Scott, K. Howe, S. E. Hunt, T. D. Andrews, J. G. Gilbert, D. Swarbreck, J. L. Ashurst, A. Taylor, J. Battles, C. P. Bird, R. Ainscough, J. P. Almeida, R. I. Ashwell, K. D. Ambrose, A. K. Babbage, C. L. Bagguley, J. Bailey, R. Banerjee, K. Bates, H. Beasley, S. Bray-Allen, A. J. Brown, J. Y. Brown, D. C. Burford, W. Burrill, J. Burton, P. Cahill, D. Camire, N. P. Carter, J. C. Chapman, S. Y. Clark, G. Clarke, C. M. Clee, S. Clegg, N. Corby, A. Coulson, P. Dhami, I. Dutta, M. Dunn, L. Faulkner, A. Frankish, J. A. Frankland, P. Garner, J. Garnett, S. Gribble, C. Griffiths, R. Grocock, E. Gustafson, S. Hammond, J. L. Harley, E. Hart, P. D. Heath, T. P. Ho, B. Hopkins, J. Horne, P. J. Howden, E. Huckle, C. Hynds, C. Johnson, D. Johnson, A. Kana, M. Kay, A. M. Kimberley, J. K. Kershaw, M. Kokkinaki, G. K. Laird, S. Lawlor, H. M. Lee, D. A. Leongamornlert, G. Laird, C. Lloyd, D. M. Lloyd, J. Loveland, J. Lovell, S. McLaren, K. E. McLay, A. McMurray, M. Mashreghi-Mohammadi, L. Matthews, S. Milne, T. Nickerson, M. Nguyen, E. Overton-Larty, S. A. Palmer, A. V. Pearce, A. I. Peck, S. Pelan, B. Phillimore, K. Porter, C. M. Rice, A. Rogosin, M. T. Ross, T. Sarafidou, H. K. Sehra, R. Shownkeen, C. D. Skuce, M. Smith, L. Standring, N. Sycamore, J. Tester, A. Thorpe, W. Torcasso, A. Tracey, A. Tromans, J. Tsolas, M. Wall, J. Walsh, H. Wang, K. Weinstock, A. P. West, D. L. Willey, S. L. Whitehead, L. Wilming, P. W. Wray, L. Young, Y. Chen, R. C. Lovering, N. K. Moschonas, R. Siebert, K. Fechtel, D. Bentley, R. Durbin, T. Hubbard, L. Doucette-Stamm, S. Beck, D. R. Smith and J.

Rogers (2004). "The DNA sequence and comparative analysis of human chromosome 10." Nature **429**(6990): 375-81.

Deqaqi, S. C., M. N'Guessan, J. Forner, A. Sbiti, C. Beldjord, J. Chelly, A. Sefiani and V. Des Portes (1998). "A gene for non-specific X-linked mental retardation (MRX55) is located in Xp11." Ann Genet **41**(1): 11-6.

D'Errico, I., G. Gadaleta and C. Saccone (2004). "Pseudogenes in metazoa: origin and features." Brief Funct Genomic Proteomic **3**(2): 157-67.

Derry, J. M., U. Jess and U. Francke (1995). "Cloning and characterization of a novel zinc finger gene in Xp11.2." Genomics **30**(2): 361-5.

Diatchenko, L., Y. F. Lau, A. P. Campbell, A. Chenchik, F. Moqadam, B. Huang, S. Lukyanov, K. Lukyanov, N. Gurskaya, E. D. Sverdlov and P. D. Siebert (1996). "Suppression subtractive hybridization: a method for generating differentially regulated or tissue-specific cDNA probes and libraries." Proc Natl Acad Sci U S A **93**(12): 6025-30.

Dominski, Z. and R. Kole (1991). "Selection of splice sites in pre-mRNAs with short internal exons." Mol Cell Biol **11**(12): 6075-83.

Down, T. A. and T. J. Hubbard (2002). "Computational detection and location of transcription start sites in mammalian genomic DNA." Genome Res **12**(3): 458-61.

Dralyuk, I., M. Brudno, M. S. Gelfand, M. Zorn and I. Dubchak (2000). "ASDB: database of alternatively spliced genes." Nucleic Acids Res **28**(1): 296-7.

Dredge, B. K., A. D. Polydorides and R. B. Darnell (2001). "The splice of life: alternative splicing and neurological disease." Nat Rev Neurosci **2**(1): 43-50.

Dresios, J., A. Aschrafi, G. C. Owens, P. W. Vanderklish, G. M. Edelman and V. P. Mauro (2005). "Cold stress-induced protein Rbm3 binds 60S ribosomal subunits, alters microRNA levels, and enhances global protein synthesis." Proc Natl Acad Sci U S A **102**(6): 1865-70.

Dubchak, I. and K. Frazer (2003). "Multi-species sequence comparison: the next frontier in genome annotation." Genome Biol 4(12): 122.

Dube, J. L., P. Wang, J. Elvin, K. M. Lyons, A. J. Celeste and M. M. Matzuk (1998). "The bone morphogenetic protein 15 gene is X-linked and expressed in oocytes." Mol Endocrinol **12**(12): 1809-17.

Dunham, A., L. H. Matthews, J. Burton, J. L. Ashurst, K. L. Howe, K. J. Ashcroft, D. M. Beare, D. C. Burford, S. E. Hunt, S. Griffiths-Jones, M. C. Jones, S. J. Keenan, K. Oliver, C. E. Scott, R. Ainscough, J. P. Almeida, K. D. Ambrose, D. T. Andrews, R. I. Ashwell, A. K. Babbage, C. L. Bagguley, J. Bailey, R. Bannerjee, K. F. Barlow, K. Bates, H. Beasley, C. P. Bird, S. Bray-Allen, A. J. Brown, J. Y. Brown, W. Burrill, C. Carder, N. P. Carter, J. C. Chapman, M. E. Clamp, S. Y. Clark, G. Clarke, C. M. Clee, S. C. Clegg, V. Cobley, J. E. Collins, N. Corby, G. J. Coville, P. Deloukas, P. Dhami, I. Dunham, M. Dunn, M. E. Earthrowl, A. G. Ellington, L. Faulkner, A. G. Frankish, J. Frankland, L. French, P. Garner, J. Garnett, J. G. Gilbert, C. J. Gilson, J. Ghori, D. V. Grafham, S. M. Gribble, C. Griffiths, R. E. Hall, S. Hammond, J. L. Harley, E. A. Hart, P. D. Heath, P. J. Howden, E. J. Huckle, P. J. Hunt, A. R. Hunt, C. Johnson, D. Johnson, M. Kay, A. M. Kimberley, A. King, G. K. Laird, C. J. Langford, S. Lawlor, D. A. Leongamornlert, D. M. Lloyd, C. Lloyd, J. E.

Loveland, J. Lovell, S. Martin, M. Mashreghi-Mohammadi, S. J. McLaren, A. McMurray, S. Milne, M. J. Moore, T. Nickerson, S. A. Palmer, A. V. Pearce, A. I. Peck, S. Pelan, B. Phillimore, K. M. Porter, C. M. Rice, S. Searle, H. K. Sehra, R. Shownkeen, C. D. Skuce, M. Smith, C. A. Steward, N. Sycamore, J. Tester, D. W. Thomas, A. Tracey, A. Tromans, B. Tubby, M. Wall, J. M. Wallis, A. P. West, S. L. Whitehead, D. L. Willey, L. Wilming, P. W. Wray, M. W. Wright, L. Young, A. Coulson, R. Durbin, T. Hubbard, J. E. Sulston, S. Beck, D. R. Bentley, J. Rogers and M. T. Ross (2004). "The DNA sequence and analysis of human chromosome 13." Nature **428**(6982): 522-8.

Dunham, I., N. Shimizu, B. A. Roe, S. Chissoe, A. R. Hunt, J. E. Collins, R. Bruskiewich, D. M. Beare, M. Clamp, L. J. Smink, R. Ainscough, J. P. Almeida, A. Babbage, C. Bagguley, J. Bailey, K. Barlow, K. N. Bates, O. Beasley, C. P. Bird, S. Blakey, A. M. Bridgeman, D. Buck, J. Burgess, W. D. Burrill, K. P. O'Brien and et al. (1999). "The DNA sequence of human chromosome 22." Nature **402**(6761): 489-95.

Durbin, R., and J. T. Mieg (1991-). A C. elegans Database. Documentation, code and data available from anonymous FTP servers at lirmm.lirmm.fr, cele.mrc-lmb.cam.ac.uk and ncbi.nlm.nih.gov.

Duyk, G. M., S. W. Kim, R. M. Myers and D. R. Cox (1990). "Exon trapping: a genetic screen to identify candidate transcribed sequences in cloned mammalian genomic DNA." Proc Natl Acad Sci U S A **87**(22): 8995-9.

Eddy, S. R. (2001). "Non-coding RNA genes and the modern RNA world." Nat Rev Genet **2**(12): 919-29.

Elbashir, S. M., W. Lendeckel and T. Tuschl (2001). "RNA interference is mediated by 21- and 22-nucleotide RNAs." Genes Dev **15**(2): 188-200.

Emanuel, B. S. and T. H. Shaikh (2001). "Segmental duplications: an 'expanding' role in genomic instability and disease." Nat Rev Genet **2**(10): 791-800.

Enokido, Y., H. Maruoka, H. Hatanaka, I. Kanazawa and H. Okazawa (2002). "PQBP-1 increases vulnerability to low potassium stress and represses transcription in primary cerebellar neurons." Biochem Biophys Res Commun **294**(2): 268-71.

Esposito, T., F. Gianfrancesco, A. Ciccodicola, M. D'Esposito, R. Nagaraja, R. Mazzarella, M. D'Urso and A. Forabosco (1997). "Escape from X inactivation of two new genes associated with DXS6974E and DXS7020E." Genomics **43**(2): 183-90.

Estes, P. A., N. E. Cooke and S. A. Liebhaber (1992). "A native RNA secondary structure controls alternative splice-site selection and generates two human growth hormone isoforms." J Biol Chem **267**(21): 14902-8.

Fairbrother, W. G. and L. A. Chasin (2000). "Human genomic sequences that inhibit splicing." Mol Cell **Biol20**(18): 6816-25.

Fairbrother, W. G., R. F. Yeh, P. A. Sharp and C. B. Burge (2002). "Predictive identification of exonic splicing enhancers in human genes." Science **297**(5583): 1007-13.

Falquet, L., M. Pagni, P. Bucher, N. Hulo, C. J. Sigrist, K. Hofmann and A. Bairoch (2002). "The PROSITE database, its status in 2002." Nucleic Acids Res **30**(1): 235-8.

Farris, W., M. A. Leissring, M. L. Hemming, A. Y. Chang and D. J. Selkoe (2005). "Alternative splicing of human insulin-degrading enzyme yields a novel isoform with a decreased ability to degrade insulin and amyloid beta-protein." Biochemistry **44**(17): 6513-25.

Feldmann, H., M. Aigle, G. Aljinovic, B. Andre, M. C. Baclet, C. Barthe, A. Baur, A. M. Becam, N. Biteau, E. Boles and et al. (1994). "Complete DNA sequence of yeast chromosome II." Embo J **13**(24): 5795-809.

Ferrier-Cana, E., C. Macadre, M. Sevignac, P. David, T. Langin and V. Geffroy (2005). "Distinct post-transcriptional modifications result into seven alternative transcripts of the CC-NBS-LRR gene JA1tr of Phaseolus vulgaris." Theor Appl Genet **110**(5): 895-905.

Fichera, M., M. Falco, M. Lo Giudice, L. Castiglia, V. Guarnaccia, F. Cali, A. Spalletta, C. Scuderi and E. Avola (2005). "Skewed X-inactivation in a family with mental retardation and PQBP1 gene mutation." Clin Genet **67**(5): 446-7.

Fiers, W., R. Contreras, G. Haegemann, R. Rogiers, A. Van de Voorde, H. Van Heuverswyn, J. Van Herreweghe, G. Volckaert and M. Ysebaert (1978). "Complete nucleotide sequence of SV40 DNA." Nature **273**(5658): 113-20.

Fire, A., S. Xu, M. K. Montgomery, S. A. Kostas, S. E. Driver and C. C. Mello (1998). "Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans." Nature **391**(6669): 806-11.

Fisher, S. E., A. Ciccodicola, K. Tanaka, A. Curci, S. Desicato, M. D'Urso and I. W. Craig (1997). "Sequence-based exon prediction around the synaptophysin locus reveals a gene-rich area containing novel genes in human proximal Xp." Genomics **45**(2): 340-7.

Fisher, S. E., I. van Bakel, S. E. Lloyd, S. H. Pearce, R. V. Thakker and I. W. Craig (1995). "Cloning and characterization of CLCN5, the human kidney chloride channel gene implicated in Dent disease (an X-linked hereditary nephrolithiasis)." Genomics 29**(**3): 598-606.

Fleischmann, R. D., M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, J. M. Merrick and et al. (1995). "Whole-genome random sequencing and assembly of Haemophilus influenzae Rd." Science **269**(5223): 496-512.

Flicek, P., E. Keibler, P. Hu, I. Korf and M. R. Brent (2003). "Leveraging the mouse genome for gene prediction in human: from whole-genome shotgun reads to a global synteny map." Genome Res **13**(1): 46-54.

Florea, L., G. Hartzell, Z. Zhang, G. M. Rubin and W. Miller (1998). "A computer program for aligning a cDNA sequence with a genomic DNA sequence." Genome Res **8**(9): 967-74.

Franze, A., N. Archidiacono, M. Rocchi, M. Marino and G. Grimaldi (1991). "Isolation and expression analysis of a human zinc finger gene (ZNF41) located on the short arm of the X chromosome." Genomics **9**(4): 728-36.

Frazer, K. A., L. Elnitski, D. M. Church, I. Dubchak and R. C. Hardison (2003). "Cross-species sequence comparisons: a review of methods and available resources." Genome Res **13**(1): 1-12.

Fricker, L. D., A. A. McKinzie, J. Sun, E. Curran, Y. Qian, L. Yan, S. D. Patterson, P. L. Courchesne, B. Richards, N. Levin, N. Mzhavia, L. A. Devi and J. Douglass (2000). "Identification and characterization of proSAAS, a granin-like neuroendocrine peptide precursor that inhibits prohormone processing." J Neurosci **20**(2): 639-48.

Friedman, R. and A. L. Hughes (2001). "Gene duplication and the structure of eukaryotic genomes." Genome Res **11**(3): 373-81.

Fujita, T., J. L. Mandel, T. Shirasawa, O. Hino, T. Shirai and N. Maruyama (1995). "Isolation of cDNA clone encoding human homologue of senescence marker protein-30 (SMP30) and its location on the X chromosome." Biochim Biophys Acta **1263**(3): 249-52.

Galtier, N., Gouy, M. and Gautier, C. (1996) SeaView and Phylo_win, two graphic tools for sequence alignment and molecular phylogeny. Comput. Applic. Biosci., **12**, 543-548.

Gardiner-Garden, M. and M. Frommer (1987). "CpG islands in vertebrate genomes." J Mol Biol **196**(2): 261-82.

Gatfield, D., L. Unterholzner, F. D. Ciccarelli, P. Bork and E. Izaurralde (2003). "Nonsense-mediated mRNA decay in Drosophila: at the intersection of the yeast and mammalian pathways." Embo J **22**(15): 3960-70.

Geraghty, M. T., L. C. Brody, L. S. Martin, M. Marble, W. Kearns, P. Pearson, A. P. Monaco, H. Lehrach and D. Valle (1993). "The isolation of cDNAs from OATL1 at Xp 11.2 using a 480-kb YAC." Genomics **16**(2): 440-6.

Gerhard, D. S., L. Wagner, E. A. Feingold, C. M. Shenmen, L. H. Grouse, G. Schuler, S. L. Klein, S. Old, R. Rasooly, P. Good, M. Guyer, A. M. Peck, J. G. Derge, D. Lipman, F. S. Collins, W. Jang, S. Sherry, M. Feolo, L. Misquitta, E. Lee, K. Rotmistrovsky, S. F. Greenhut, C. F. Schaefer, K. Buetow, T. I. Bonner, D. Haussler, J. Kent, M. Kiekhaus, T. Furey, M. Brent, C. Prange, K. Schreiber, N. Shapiro, N. K. Bhat, R. F. Hopkins, F. Hsie, T. Driscoll, M. B. Soares, T. L. Casavant, T. E. Scheetz, M. J. Brown-stein, T. B. Usdin, S. Toshiyuki, P. Carninci, Y. Piao, D. B. Dudekula, M. S. Ko, K. Kawakami, Y. Suzuki, S. Sugano, C. E. Gruber, M. R. Smith, B. Simmons, T. Moore, R. Waterman, S. L. Johnson, Y. Ruan, C. L. Wei, S. Mathavan, P. H. Gunaratne, J. Wu, A. M. Garcia, S. W. Hulyk, E. Fuh, Y. Yuan, A. Sneed, C. Kowis, A. Hodgson, D. M. Muzny, J. McPherson, R. A. Gibbs, J. Fahey, E. Helton, M. Ketteman, A. Madan, S. Rodrigues, A. Sanchez, M. Whiting, A. Madari, A. C. Young, K. D. Wetherby, S. J. Granite, P. N. Kwong, C. P. Brinkley, R. L. Pearson, G. G. Bouffard, R. W. Blakesly, E. D. Green, M. C. Dickson, A. C. Rodriguez, J. Grimwood, J. Schmutz, R. M. Myers, Y. S. Butterfield, M. Griffith, O. L. Griffith, M. I. Krzywinski, N. Liao, R. Morrin, D. Palmquist, A. S. Petrescu, U. Skalska, D. E. Smailus, J. M. Stott, A. Schnerch, J. E. Schein, S. J. Jones, R. A. Holt, A. Baross, M. A. Marra, S. Clifton, K. A. Makowski, S. Bosak and J. Malek (2004). "The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC)." Genome Res **14(**10B): 2121-7.

Gilbert, S. F. (1978). "The embryological origins of the gene theory." J Hist Biol **11**(2): 307-51.

Glas, R., A. A. De Leo, M. L. Delbridge, K. Reid, M. A. Ferguson-Smith, P. C. O'Brien, M. Westerman and J. A. Graves (1999a). "Chromosome painting in marsupials: genome conservation in the kangaroo family." Chromosome Res **7**(3): 167-76.

Glas, R., J. A. Marshall Graves, R. Toder, M. Ferguson-Smith and P. C. O'Brien (1999b). "Cross-species chromosome painting between human and marsupial directly demonstrates the ancient region of the mammalian X." Mamm Genome **10**(11): 1115-6.

Golabi, M., M. Ito and B. D. Hall (1984). "A new X-linked multiple congenital anomalies/mental retardation syndrome." Am J Med Genet **17**(1): 367-74.

Gossen, M. and H. Bujard (1992). "Tight control of gene expression in mammalian cells by tetracycline-responsive promoters." Proc Natl Acad Sci U S A **89**(12): 5547-51.

Grabowski, P. J. and D. L. Black (2001). "Alternative RNA splicing in the nervous system." Prog Neurobiol 65(3): 289-308.

Graveley, B. R., K. J. Hertel and T. Maniatis (1998). "A systematic analysis of the factors that determine the strength of pre-mRNA splicing enhancers." Embo J **17**(22): 6747-56.

Graves, J. A. (1998). "Evolution of the mammalian Y chromosome and sex-determining genes." J Exp Zool **281**(5): 472-81.

Gregory, S. G., M. Sekhon, J. Schein, S. Zhao, K. Osoegawa, C. E. Scott, R. S. Evans, P. W. Burridge, T. V. Cox, C. A. Fox, R. D. Hutton, I. R. Mullenger, K. J. Phillips, J. Smith, J. Stalker, G. J. Threadgold, E. Birney, K. Wylie, A. Chinwalla, J. Wallis, L. Hillier, J. Carter, T. Gaige, S. Jaeger, C. Kremitzki, D. Layman, J. Maas, R. McGrane, K. Mead, R. Walker, S. Jones, M. Smith, J. Asano, I. Bosdet, S. Chan, S. Chittaranjan, R. Chiu, C. Fjell, D. Fuhrmann, N. Girn, C. Gray, R. Guin, L. Hsiao, M. Krzywinski, R. Kutsche, S. S. Lee, C. Mathewson, C. McLeavy, S. Messervier, S. Ness, P. Pandoh, A. L. Prabhu, P. Saeedi, D. Smailus, L. Spence, J. Stott, S. Taylor, W. Terpstra, M. Tsai, J. Vardy, N. Wye, G. Yang, S. Shatsman, B. Ayodeji, K. Geer, G. Tsegaye, A. Shvartsbeyn, E. Gebregeorgis, M. Krol, D. Russell, L. Overton, J. A. Malek, M. Holmes, M. Heaney, J. Shetty, T. Feldblyum, W. C. Nierman, J. J. Catanese, T. Hubbard, R. H. Waterston, J. Rogers, P. J. de Jong, C. M. Fraser, M. Marra, J. D. McPherson and D. R. Bentley (2002). "A physical map of the mouse genome." Nature **418**(6899): 743-50.

Grimwood, J., L. A. Gordon, A. Olsen, A. Terry, J. Schmutz, J. Lamerdin, U. Hellsten, D. Goodstein, O. Couronne, M. Tran-Gyamfi, A. Aerts, M. Altherr, L. Ashworth, E. Bajorek, S. Black, E. Branscomb, S. Caenepeel, A. Carrano, C. Caoile, Y. M. Chan, M. Christensen, C. A. Cleland, A. Copeland, E. Dalin, P. Dehal, M. Denys, J. C. Detter, J. Escobar, D. Flowers, D. Fotopulos, C. Garcia, A. M. Georgescu, T. Glavina, M. Gomez, E. Gonzales, M. Groza, N. Hammon, T. Hawkins, L. Haydu, I. Ho, W. Huang, S. Israni, J. Jett, K. Kadner, H. Kimball, A. Kobayashi, V. Larionov, S. H. Leem, F. Lopez, Y. Lou, S. Lowry, S. Malfatti, D. Martinez, P. McCready, C. Medina, J. Morgan, K. Nelson, M. Nolan, I. Ovcharenko, S. Pitluck, M. Pollard, A. P. Popkie, P. Predki, G. Quan, L. Ramirez, S. Rash, J. Retterer, A. Rodriguez, S. Rogers, A. Salamov, A. Salazar, X. She, D. Smith, T. Slezak, V. Solovyev, N. Thayer, H. Tice, M. Tsai, A. Ustaszewska, N. Vo, M. Wagner, J. Wheeler, K. Wu, G. Xie, J. Yang, I. Dubchak, T. S. Furey, P. DeJong, M. Dickson, D. Gordon, E. E. Eichler, L. A. Pennacchio, P. Richardson, L. Stubbs, D. S. Rokhsar, R. M. Myers, E. M. Rubin and S. M. Lucas (2004). "The DNA sequence and biology of human chromosome 19." Nature **428**(6982): 529-35.

Gu, J. and Reddy, R (2001). Cellular RNAs: varied roles in *Encyclopedia of Life Sciences* Nature Publishing Group, London,

Gu, J., R. Dubner, A. J. Fornace, Jr. and M. J. Iadarola (1995). "UREB1, a tyrosine phosphorylated nuclear protein, inhibits p53 transactivation." Oncogene **11**(10): 2175-8.

Guigo, R., E. T. Dermitzakis, P. Agarwal, C. P. Ponting, G. Parra, A. Reymond, J. F. Abril, E. Keibler, R. Lyle, C. Ucla, S. E. Antonarakis and M. R. Brent (2003). "Comparison of mouse and human genomes followed by experimental verification yields an estimated 1,019 additional genes." Proc Natl Acad Sci U S A **100**(3): 1140-5.

Guigo, R., P. Agarwal, J. F. Abril, M. Burset and J. W. Fickett (2000). "An assessment of gene prediction accuracy in large DNA sequences." Genome Res **10**(10): 1631-42.

Gumucio, D. L., K. L. Rood, K. L. Blanchard-McQuate, T. A. Gray, A. Saulino and F. S. Collins (1991). "Interaction of Sp1 with the human gamma globin promoter: binding and transactivation of normal and mutant promoters." Blood **78**(7): 1853-63.

Gupta, S., D. Zink, B. Korn, M. Vingron and S. A. Haas (2004). "Genome wide identification and classification of alternative splicing based on EST data." Bioinformatics **20**(16): 2579-85.

Gure, A. O., I. J. Wei, L. J. Old and Y. T. Chen (2002). "The SSX gene family: characterization of 9 complete genes." Int J Cancer **101**(5): 448-53.

Han, K., G. Yeo, P. An, C. B. Burge and P. J. Grabowski (2005). "A combinatorial code for splicing silencing: UAGG and GGGG motifs." PLoS Biol **3(**5): e158.

Handley, P. M., M. Mueckler, N. R. Siegel, A. Ciechanover and A. L. Schwartz (1991). "Molecular cloning, sequence, and tissue distribution of the human ubiquitin-activating enzyme E1." Proc Natl Acad Sci U S A **88**(1): 258-62.

Hanner, M., F. F. Moebius, F. Weber, M. Grabner, J. Striessnig and H. Glossmann (1995). "Phenylalkylamine Ca2+ antagonist binding protein. Molecular cloning, tissue distribution, and heterologous expression." J Biol Chem **270**(13): 7551-7.

Harries, L. W., A. T. Hattersley and S. Ellard (2004). "Messenger RNA transcripts of the hepatocyte nuclear factor-1alpha gene containing premature termination codons are subject to nonsense-mediated decay." Diabetes **53**(2): 500-4.

Harrison, P. M., A. Kumar, N. Lang, M. Snyder and M. Gerstein (2002a). "A question of size: the eukaryotic proteome and the problems in defining it." Nucleic Acids Res **30**(5): 1083-90.

Harrison, P. M., H. Hegyi, S. Balasubramanian, N. M. Luscombe, P. Bertone, N. Echols, T. Johnson and M. Gerstein (2002b). "Molecular fossils in the human genome: identification and analysis of the pseudogenes in chromosomes 21 and 22." Genome Res **12**(2): 272-80.

Hattori M., A. Fujiyama, T.D. Taylor, H. Watanabe, T. Yada, H.S. Park, A. Toyoda, K. Ishii, Y. Totoki, D.K. Choi, Y. Groner, E. Soeda, M. Ohki, T. Takagi, Y. Sakaki, S. Taudien, K. Blechschmidt, A. Polley, U. Menzel, J. Delabar, K. Kumpf, R. Lehmann, D. Patterson, K. Reichwald, A. Rump, M. Schillhabel, A. Schudy, W. Zimmermann, A. Rosenthal, J. Kudoh, K. Schibuya, K. Kawasaki, S. Asakawa, A. Shintani, T. Sasaki, K. Nagamine, S. Mitsuyama, S.E. Antonarakis, S. Minoshima, N. Shimizu, G. Nordsiek, K. Hornischer,P. Brant, M. Scharfe, O. Schon, A. Desario, J Reichelt,G. Kauer, H. Blocker, J. Ramser, A. Beck, S. Klages, S. Hennig, L. Riesselmann, E. Dagand, T. Haaf, S. Wehrmeyer, K. Borzym, K.

Gardiner, D. Nizetic, F. Francis, H. Lehrach, R. Reinhardt, M.L. Yaspo (2000). The DNA sequence of human chromosome 21.Nature 18(405): 311-9.

Heilig, R., R. Eckenberg, J. L. Petit, N. Fonknechten, C. Da Silva, L. Cattolico, M. Levy, V. Barbe, V. de Berardinis, A. Ureta-Vidal, E. Pelletier, V. Vico, V. Anthouard, L. Rowen, A. Madan, S. Qin, H. Sun, H. Du, K. Pepin, F. Artiguenave, C. Robert, C. Cruaud, T. Bruls, O. Jaillon, L. Friedlander, G. Samson, P. Brottier, S. Cure, B. Segurens, F. Aniere, S. Samain, H. Crespeau, N. Abbasi, N. Aiach, D. Boscus, R. Dickhoff, M. Dors, I. Dubois, C. Friedman, M. Gouyvenoux, R. James, A. Madan, B. Mairey-Estrada, S. Mangenot, N. Martins, M. Menard, S. Oztas, A. Ratcliffe, T. Shaffer, B. Trask, B. Vacherie, C. Bellemere, C. Belser, M. Besnard-Gonnet, D. Bartol-Mavel, M. Boutard, S. Briez-Silla, S. Combette, V. Dufosse-Laurent, C. Ferron, C. Lechaplais, C. Louesse, D. Muselet, G. Magdelenat, E. Pateau, E. Petit, P. Sirvain-Trukniewicz, A. Trybou, N. Vega-Czarny, E. Bataille, E. Bluet, I. Bordelais, M. Dubois, C. Dumont, T. Guerin, S. Haffray, R. Hammadi, J. Muanga, V. Pellouin, D. Robert, E. Wunderle, G. Gauguet, A. Roy, L. Sainte-Marthe, J. Verdier, C. Verdier-Discala, L. Hillier, L. Fulton, J. McPherson, F. Matsuda, R. Wilson, C. Scarpelli, G. Gyapay, P. Wincker, W. Saurin, F. Quetier, R. Waterston, L. Hood and J. Weissenbach (2003). "The DNA sequence and analysis of human chromosome 14." Nature **421**(6923): 601-7.

Hidaka, K., J. J. Caffrey, L. Hua, T. Zhang, J. R. Falck, G. C. Nickel, L. Carrel, L. D. Barnes and S. B. Shears (2002). "An adjacent pair of human NUDT genes on chromosome X are preferentially expressed in testis and encode two new isoforms of diphosphoinositol polyphosphate phosphohydrolase." J Biol Chem **277**(36): 32730-8.

Hillier, L. W., R. S. Fulton, L. A. Fulton, T. A. Graves, K. H. Pepin, C. Wagner-McPherson, D. Layman, J. Maas, S. Jaeger, R. Walker, K. Wylie, M. Sekhon, M. C. Becker, M. D. O'Laughlin, M. E. Schaller, G. A. Fewell, K. D. Delehaunty, T. L. Miner, W. E. Nash, M. Cordes, H. Du, H. Sun, J. Edwards, H. Bradshaw-Cordum, J. Ali, S. Andrews, A. Isak, A. Vanbrunt, C. Nguyen, F. Du, B. Lamar, L. Courtney, J. Kalicki, P. Ozersky, L. Bielicki, K. Scott, A. Holmes, R. Harkins, A. Harris, C. M. Strong, S. Hou, C. Tomlinson, S. Dauphin-Kohlberg, A. Kozlowicz-Reilly, S. Leonard, T. Rohlfing, S. M. Rock, A. M. Tin-Wollam, A. Abbott, P. Minx, R. Maupin, C. Strowmatt, P. Latreille, N. Miller, D. Johnson, J. Murray, J. P. Woessner, M. C. Wendl, S. P. Yang, B. R. Schultz, J. W. Wallis, J. Spieth, T. A. Bieri, J. O. Nelson, N. Berkowicz, P. E. Wohldmann, L. L. Cook, M. T. Hickenbotham, J. Eldred, D. Williams, J. A. Bedell, E. R. Mardis, S. W. Clifton, S. L. Chissoe, M. A. Marra, C. Raymond, E. Haugen, W. Gillett, Y. Zhou, R. James, K. Phelps, S. Iadanoto, K. Bubb, E. Simms, R. Levy, J. Clendenning, R. Kaul, W. J. Kent, T. S. Furey, R. A. Baertsch, M. R. Brent, E. Keibler, P. Flicek, P. Bork, M. Suyama, J. A. Bailey, M. E. Portnoy, D. Torrents, A. T. Chinwalla, W. R. Gish, S. R. Eddy, J. D. McPherson, M. V. Olson, E. E. Eichler, E. D. Green, R. H. Waterston and R. K. Wilson (2003). "The DNA sequence of human chromosome 7." Nature **424**(6945): 157-64.

Hillier, L. W., T. A. Graves, R. S. Fulton, L. A. Fulton, K. H. Pepin, P. Minx, C. Wagner-McPherson, D. Layman, K. Wylie, M. Sekhon, M. C. Becker, G. A. Fewell, K. D. Delehaunty, T. L. Miner, W. E. Nash, C. Kremitzki, L. Oddy, H. Du, H. Sun, H. Bradshaw-Cordum, J. Ali, J. Carter, M. Cordes, A. Harris, A. Isak, A. van Brunt, C. Nguyen, F. Du, L. Courtney, J. Kalicki, P. Ozersky, S. Abbott, J. Armstrong, E. A. Belter, L. Caruso, M. Cedroni, M. Cotton, T. Davidson, A. Desai, G. Elliott, T. Erb, C. Fronick, T. Gaige, W. Haakenson, K. Haglund, A. Holmes, R. Harkins, K. Kim, S. S. Kruchowski, C. M. Strong, N. Grewal, E. Goyea, S. Hou, A. Levy, S. Martinka, K. Mead, M. D. McLellan, R. Meyer, J. Randall-Maher, C. Tomlinson, S. Dauphin-Kohlberg, A. Kozlowicz-Reilly, N. Shah, S. Swearengen-Shahid, J. Snider, J. T. Strong, J. Thompson, M. Yoakum, S. Leonard, C. Pearman, L. Trani, M. Radionenko, J. E. Waligorski, C. Wang, S. M. Rock, A. M. Tin-Wollam, R. Maupin, P. Latreille, M. C. Wendl, S. P. Yang, C. Pohl, J. W. Wallis, J. Spieth, T. A. Bieri, N. Berkowicz, J. O. Nelson, J. Osborne, L. Ding, R. Meyer, A. Sabo, Y. Shotland, P. Sinha, P. E. Wohldmann, L. L. Cook, M. T. Hickenbotham, J. Eldred, D. Williams, T. A. Jones, X. She, F. D. Ciccarelli, E. Izaurralde, J. Taylor, J. Schmutz, R. M. Myers, D. R. Cox, X. Huang, J. D. McPherson, E. R. Mardis, S.

W. Clifton, W. C. Warren, A. T. Chinwalla, S. R. Eddy, M. A. Marra, I. Ovcharenko, T. S. Furey, W. Miller, E. E. Eichler, P. Bork, M. Suyama, D. Torrents, R. H. Waterston and R. K. Wilson (2005). "Generation and annotation of the DNA sequences of human chromosomes 2 and 4." Nature **434**(7034): 724-31.

Hillman, R. T., R. E. Green and S. E. Brenner (2004). "An unappreciated role for RNA surveillance." Genome Biol **5**(2): R8.

Holbrook, S. R. (2005). "RNA structure: the long and the short of it." Curr Opin Struct Biol 15(3): 302-8.Hoshino, S., M. Imai, M. Mizutani, Y. Kikuchi, F. Hanaoka, M. Ui and T. Katada (1998). "Molecular cloning of a novel member of the eukaryotic polypeptide chain-releasing factors (eRF). Its identification as eRF3 interacting with eRF1." J Biol Chem **273**(35): 22254-9.

Huang, H. D., J. T. Horng, C. C. Lee and B. J. Liu (2003). "ProSplicer: a database of putative alternative splicing information derived from protein, mRNA and expressed sequence tag sequence data." Genome Biol **4**(4): R29.

Huang, Y. H., Y. T. Chen, J. J. Lai, S. T. Yang and U. C. Yang (2002). "PALS db: Putative Alternative Splicing database." Nucleic Acids Res **30**(1): 186-90.

Hubbard, T., D. Andrews, M. Caccamo, G. Cameron, Y. Chen, M. Clamp, L. Clarke, G. Coates, T. Cox, F. Cunningham, V. Curwen, T. Cutts, T. Down, R. Durbin, X. M. Fernandez-Suarez, J. Gilbert, M. Hammond, J. Herrero, H. Hotz, K. Howe, V. Iyer, K. Jekosch, A. Kahari, A. Kasprzyk, D. Keefe, S. Keenan, F. Kokocinsci, D. London, I. Longden, G. McVicker, C. Melsopp, P. Meidl, S. Potter, G. Proctor, M. Rae, D. Rios, M. Schuster, S. Searle, J. Severin, G. Slater, D. Smedley, J. Smith, W. Spooner, A. Stabenau, J. Stalker, R. Storey, S. Trevanion, A. Ureta-Vidal, J. Vogel, S. White, C. Woodwark and E. Birney (2005). "Ensembl 2005." Nucleic Acids Res **33**(Database issue): D447-53.

Humphray, S. J., K. Oliver, A. R. Hunt, R. W. Plumb, J. E. Loveland, K. L. Howe, T. D. Andrews, S. Searle, S. E. Hunt, C. E. Scott, M. C. Jones, R. Ainscough, J. P. Almeida, K. D. Ambrose, R. I. Ashwell, A. K. Babbage, S. Babbage, C. L. Bagguley, J. Bailey, R. Banerjee, D. J. Barker, K. F. Barlow, K. Bates, H. Beasley, O. Beasley, C. P. Bird, S. Bray-Allen, A. J. Brown, J. Y. Brown, D. Burford, W. Burrill, J. Burton, C. Carder, N. P. Carter, J. C. Chapman, Y. Chen, G. Clarke, S. Y. Clark, C. M. Clee, S. Clegg, R. E. Collier, N. Corby, M. Crosier, A. T. Cummings, J. Davies, P. Dhami, M. Dunn, I. Dutta, L. W. Dyer, M. E. Earthrowl, L. Faulkner, C. J. Fleming, A. Frankish, J. A. Frankland, L. French, D. G. Fricker, P. Garner, J. Garnett, J. Ghori, J. G. Gilbert, C. Glison, D. V. Grafham, S. Gribble, C. Griffiths, S. Griffiths-Jones, R. Grocock, J. Guy, R. E. Hall, S. Hammond, J. L. Harley, E. S. Harrison, E. A. Hart, P. D. Heath, C. D. Henderson, B. L. Hopkins, P. J. Howard, P. J. Howden, E. Huckle, C. Johnson, D. Johnson, A. A. Joy, M. Kay, S. Keenan, J. K. Kershaw, A. M. Kimberley, A. King, A. Knights, G. K. Laird, C. Langford, S. Lawlor, D. A. Leongamornlert, M. Leversha, C. Lloyd, D. M. Lloyd, J. Lovell, S. Martin, M. Mashreghi-Mohammadi, L. Matthews, S. McLaren, K. E. McLay, A. McMurray, S. Milne, T. Nickerson, J. Nisbett, G. Nordsiek, A. V. Pearce, A. I. Peck, K. M. Porter, R. Pandian, S. Pelan, B. Phillimore, S. Povey, Y. Ramsey, V. Rand, M. Scharfe, H. K. Sehra, R. Shownkeen, S. K. Sims, C. D. Skuce, M. Smith, C. A. Steward, D. Swarbreck, N. Sycamore, J. Tester, A. Thorpe, A. Tracey, A. Tromans, D. W. Thomas, M. Wall, J. M. Wallis, A. P. West, S. L. Whitehead, D. L. Willey, S. A. Williams, L. Wilming, P. W. Wray, L. Young, J. L. Ashurst, A. Coulson, H. Blocker, R. Durbin, J. E. Sulston, T. Hubbard, M. J. Jackson, D. R. Bentley, S. Beck, J. Rogers and I. Dunham (2004). "DNA sequence and analysis of human chromosome 9." Nature **429**(6990): 369-74.

Hunkapiller, M. W. (1991). "Advances in DNA sequencing technology." Curr Opin Genet Dev **1**(1): 88-92.

Hutvagner, G. and P. D. Zamore (2002). "A microRNA in a multiple-turnover RNAi enzyme complex." Science **297**(5589): 2056-60.

Ibrahim el, C., T. D. Schaal, K. J. Hertel, R. Reed and T. Maniatis (2005). "Serine/arginine-rich protein-dependent suppression of exon skipping by exonic splicing enhancers." Proc Natl Acad Sci U S A **102**(14): 5002-7.

Imrie, H., B. Vaidya, P. Perros, W. F. Kelly, A. D. Toft, E. T. Young, P. Kendall-Taylor and S. H. Pearce (2001). "Evidence for a Graves' disease susceptibility locus at chromosome Xp11 in a United Kingdom population." J Clin Endocrinol Metab **86**(2): 626-30.

Isbrandt, D., T. Leicher, R. Waldschutz, X. Zhu, U. Luhmann, U. Michel, K. Sauter and O. Pongs (2000). "Gene structures and expression profiles of three human KCND (Kv4) potassium channels mediating A-type currents I(TO) and I(SA)." Genomics **64**(2): 144-54.

Ishida, N., N. Miura, S. Yoshioka and M. Kawakita (1996). "Molecular cloning and characterization of a novel isoform of the human UDP-galactose transporter, and of related complementary DNAs belonging to the nucleotide-sugar transporter gene family." J Biochem (Tokyo) **120**(6): 1074-8.

Ishigaki, Y., X. Li, G. Serin and L. E. Maquat (2001). "Evidence for a pioneer round of mRNA translation: mRNAs subject to nonsense-mediated decay in mammalian cells are bound by CBP80 and CBP20." Cell **106**(5): 607-17.

Iwamoto, K., Y. Huang and S. Ueda (2000). "Genomic organization and alternative transcripts of the human PQBP-1 gene." Gene **259**(1-2): 69-73.

Jackson, I. J. (1991). "A reappraisal of non-consensus mRNA splice sites." Nucleic Acids Res **19**(14): 3795-8.

Jacobson, A. and S. W. Peltz (1999). "Tools for turnover: methods for analysis of mRNA stability in eukaryotic cells." Methods **17**(1): 1-2.

Jaeger, W. (1992). "Horner's law. The first step in the history of the understanding of X-linked disorders." Ophthalmic Paediatr Genet **13**(2): 49-56.

Jasinska, A. and W. J. Krzyzosiak (2004). "Repetitive sequences that shape the human transcriptome." FEBS Lett **567**(1): 136-41.

Jeffreys, A. J., D. L. Neil and R. Neumann (1998). "Repeat instability at human minisatellites arising from meiotic recombination." Embo J **17**(14): 4147-57.

Jeffreys, A. J., K. Tamaki, A. MacLeod, D. G. Monckton, D. L. Neil and J. A. Armour (1994). "Complex gene conversion events in germline mutation at human minisatellites." Nat Genet **6**(2): 136-45.

Jensen, K. B., B. K. Dredge, G. Stefani, R. Zhong, R. J. Buckanovich, H. J. Okano, Y. Y. Yang and R. B. Darnell (2000). "Nova-1 regulates neuron-specific alternative splicing and is essential for neuronal viability." Neuron **25(**2): 359-71.

Jiang, X., J. Kowalski and J. W. Kelly (2001). "Increasing protein stability using a rational approach combining sequence homology and structural alignment: Stabilizing the WW domain." Protein Sci **10**(7): 1454-65.

John, B., A. J. Enright, A. Aravin, T. Tuschl, C. Sander and D. S. Marks (2004). "Human MicroRNA targets." PLoS Biol **2**(11): e363.

Johnson, J. M., S. Edwards, D. Shoemaker and E. E. Schadt (2005). "Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments." Trends Genet **21**(2): 93-102.

Kalnina, Z., P. Zayakin, K. Silina and A. Line (2005). "Alterations of pre-mRNA splicing in cancer." Genes Chromosomes Cancer **42**(4): 342-57.

Kalscheuer, V. M., K. Freude, L. Musante, L. R. Jensen, H. G. Yntema, J. Gecz, A. Sefiani, K. Hoffmann, B. Moser, S. Haas, U. Gurok, S. Haesler, B. Aranda, A. Nshedjan, A. Tzschach, N. Hartmann, T. C. Roloff, S. Shoichet, O. Hagens, J. Tao, H. Van Bokhoven, G. Turner, J. Chelly, C. Moraine, J. P. Fryns, U. Nuber, M. Hoeltzenbein, C. Scharff, H. Scherthan, S. Lenzner, B. C. Hamel, S. Schweiger and H. H. Ropers (2003). "Mutations in the polyglutamine binding protein 1 gene cause X-linked mental retardation." Nat Genet **35**(4): 313-5.

Kampa, D., J. Cheng, P. Kapranov, M. Yamanaka, S. Brubaker, S. Cawley, J. Drenkow, A. Piccolboni, S. Bekiranov, G. Helt, H. Tammana and T. R. Gingeras (2004). "Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22." Genome Res **14**(3): 331-42.

Kan, Z., D. States and W. Gish (2002). "Selecting for functional alternative splices in ESTs." Genome Res **12**(12): 1837-45.

Kan, Z., E. C. Rouchka, W. R. Gish and D. J. States (2001). "Gene structure prediction and alternative splicing analysis using genomically aligned ESTs." Genome Res **11**(5): 889-900.

Kapranov, P., S. E. Cawley, J. Drenkow, S. Bekiranov, R. L. Strausberg, S. P. Fodor and T. R. Gingeras (2002). "Large-scale transcriptional activity in chromosomes 21 and 22." Science **296**(5569): 916-9.

Kashima, T. and J. L. Manley (2003). "A negative element in SMN2 exon 7 inhibits splicing in spinal muscular atrophy." Nat Genet **34**(4): 460-3.

Kawai, S., T. Kato, H. Inaba, N. Okahashi and A. Amano (2005). "Odd-skipped related 2 splicing variants show opposite transcriptional activity." Biochem Biophys Res Commun **328**(1): 306-11.

Kent, W. J. (2002). "BLAT--the BLAST-like alignment tool." Genome Res **12**(4): 656-64.

Kent, W. J., C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler and D. Haussler (2002). "The human genome browser at UCSC." Genome Res **12**(6): 996-1006.

Kim, N., S. Shin and S. Lee (2005). "ECgene: genome-based EST clustering and gene modeling for alternative splicing." Genome Res **15**(4): 566-76.

Kitagawa, H., M. Fujita, N. Ito and K. Sugahara (2000). "Molecular cloning and expression of a novel chondroitin 6-O-sulfotransferase." J Biol Chem **275**(28): 21075-80.

Kiyosawa, H., I. Yamanaka, N. Osato, S. Kondo and Y. Hayashizaki (2003). "Antisense transcripts with FANTOM2 clone set and their implications for gene regulation." Genome Res **13**(6B): 1324-34.

Kleefstra, T., C. E. Franken, Y. H. Arens, G. J. Ramakers, H. G. Yntema, E. A. Sistermans, C. F. Hulsmans, W. N. Nillesen, H. van Bokhoven, B. B. de Vries and B. C. Hamel (2004). "Genotype-phenotype studies in three families with mutations in the polyglutamine-binding protein 1 gene (PQBP1)." Clin Genet **66**(4): 318-26.

Knight, J. C., G. Grimaldi, H. J. Thiesen, N. T. Bech-Hansen, C. D. Fletcher and M. P. Coleman (1994). "Clustered organization of Kruppel zinc-finger genes at Xp11.23, flanking a translocation breakpoint at OATL1: a physical map with locus assignments for ZNF21, ZNF41, ZNF81, and ELK1." Genomics **21**(1): 180-7.

Koenig, M., A. P. Monaco and L. M. Kunkel (1988). "The complete sequence of dystrophin predicts a rod-shaped cytoskeletal protein." Cell **53**(2): 219-26.

Koenig, M., E. P. Hoffman, C. J. Bertelson, A. P. Monaco, C. Feener and L. M. Kunkel (1987). "Complete cloning of the Duchenne muscular dystrophy (DMD) cDNA and preliminary genomic organization of the DMD gene in normal and affected individuals." Cell **50**(3): 509-17.

Komuro, A., M. Saeki and S. Kato (1999a). "Association of two nuclear proteins, Npw38 and NpwBP, via the interaction between the WW domain and a novel proline-rich motif containing glycine and arginine." J Biol Chem **274**(51): 36513-9.

Komuro, A., M. Saeki and S. Kato (1999b). "Npw38, a novel nuclear protein possessing a WW domain capable of activating basal transcription." Nucleic Acids Res **27**(9): 1957-65.

Korf, I., P. Flicek, D. Duan and M. R. Brent (2001). "Integrating genomic homology into gene structure prediction." Bioinformatics **17** Suppl 1: S140-8.

Kornblihtt, A. R., M. de la Mata, J. P. Fededa, M. J. Munoz and G. Nogues (2004). "Multiple links between transcription and splicing." RNA **10**(10): 1489-98.

Kozak, M. (1987). "At least six nucleotides preceding the AUG initiator codon enhance translation in mammalian cells." J Mol Biol **196**(4): 947-50.

Krause, A., S. A. Haas, E. Coward and M. Vingron (2002). "SYSTERS, GeneNest, SpliceNest: exploring sequence space from genome to protein." Nucleic Acids Res **30**(1): 299-300.

Krawczak, M., J. Reiss and D. N. Cooper (1992). "The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences." Hum Genet **90**(1-2): 41-54.

Kriventseva, E. V., I. Koch, R. Apweiler, M. Vingron, P. Bork, M. S. Gelfand and S. Sunyaev (2003). "Increase of functional diversity by alternative splicing." Trends Genet **19**(3): 124-8.

Krogh, A. (1997). "Two methods for improving performance of an HMM and their application for gene finding." Proc Int Conf Intell Syst Mol Biol **5**: 179-86.

Kruse, S., J. Forster, J. Kuehr and K. A. Deichmann (1999). "Characterization of the membrane-bound and a soluble form of human IL-4 receptor alpha produced by alternative splicing." Int Immunol **11**(12): 1965-70.

Kuersten, S. and E. B. Goodwin (2003). "The power of the 3' UTR: translational control and development." Nat Rev Genet **4(**8): 626-37.

Kulp, D., D. Haussler, M. G. Reese and F. H. Eeckman (1996). "A generalized hidden Markov model for the recognition of human genes in DNA." Proc Int Conf Intell Syst Mol Biol **4**: 134-42.

Kwan, S.P., L.A. Sandkuyl, M. Blaese, L.M. Kunkel, G. Bruns, R. Parmley, S.Skarshaug, D.C. Page, J. Ott, F.S. Rosen, (1998) "Genetic mapping of the Wiskott-Aldrich syndrome with two highly-linked polymorphic DNA markers**."** Genomics **3**: 39-43.

Lahn, B. T. and D. C. Page (1999). "Four evolutionary strata on the human X chromosome." Science **286**(5441): 964-7.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blocker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. Smit, E. Stupka, J. Szustakowski, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, A. Patrinos, M. J. Morgan, P. de Jong, J. J. Catanese, K.

Osoegawa, H. Shizuya, S. Choi and Y. J. Chen (2001). "Initial sequencing and analysis of the human genome." Nature **409**(6822): 860-921.

Larsson, T. P., C. G. Murray, T. Hill, R. Fredriksson and H. B. Schioth (2005). "Comparison of the current RefSeq, Ensembl and EST databases for counting genes and gene discovery." FEBS Lett **579**(3): 690-8.

Le Hir, H., D. Gatfield, E. Izaurralde and M. J. Moore (2001). "The exon-exon junction complex provides a binding platform for factors involved in mRNA export and nonsense-mediated mRNA decay." Embo J **20**(17): 4987-97.

Lee, C. J. and K. Irizarry (2003). "Alternative splicing in the nervous system: an emerging source of diversity and regulation." Biol Psychiatry **54**(8): 771-6.

Lee, C., L. Atanelov, B. Modrek and Y. Xing (2003). "ASAP: the Alternative Splicing Annotation Project." Nucleic Acids Res **31**(1): 101-5.

Lee, J. T., L. S. Davidow and D. Warshawsky (1999). "Tsix, a gene antisense to Xist at the X-inactivation centre." Nat Genet **21**(4): 400-4.

Lee, Y., J. Tsai, S. Sunkara, S. Karamycheva, G. Pertea, R. Sultana, V. Antonescu, A. Chan, F. Cheung and J. Quackenbush (2005). "The TIGR Gene Indices: clustering and assembling EST and known genes and integration with eukaryotic genomes." Nucleic Acids Res **33**(Database issue): D71-4.

Leipzig, J., P. Pevzner and S. Heber (2004). "The Alternative Splicing Gallery (ASG): bridging the gap between genome and transcriptome." Nucleic Acids Res **32**(13): 3977-83.

Lemahieu, V., J. M. Gastier and U. Francke (1999). "Novel mutations in the Wiskott-Aldrich syndrome protein gene and their effects on transcriptional, translational, and clinical phenotypes." Hum Mutat **14**(1): 54-66.

Lenski, C., F. Abidi, A. Meindl, A. Gibson, M. Platzer, R. Frank Kooy, H. A. Lubs, R. E. Stevenson, J. Ramser and C. E. Schwartz (2004). "Novel truncating mutations in the polyglutamine tract binding protein 1 gene (PQBP1) cause Renpenning syndrome and X-linked mental retardation in another family with microcephaly." Am J Hum Genet **74**(4): 777-80.

Lev-Maor, G., R. Sorek, N. Shomron and G. Ast (2003). "The birth of an alternatively spliced exon: 3' splice-site selection in Alu exons." Science **300**(5623): 1288-91.

Lewis, B. P., R. E. Green and S. E. Brenner (2003). "Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans." Proc Natl Acad Sci U S A **100**(1): 189-92.

Libri, D., L. Balvay and M. Y. Fiszman (1992). "In vivo splicing of the beta tropomyosin pre-mRNA: a role for branch point and donor site competition." Mol Cell **Biol12**(7): 3204-15.

Lim, L. P., M. E. Glasner, S. Yekta, C. B. Burge and D. P. Bartel (2003). "Vertebrate microRNA genes." Science **299**(5612): 1540.

Livak, K. J. and T. D. Schmittgen (2001). "Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method." *Methods* **25**(4): 402-8.

Lozano, J. C., E. Perret, P. Schatt, C. Arnould, G. Peaucellier and A. Picard (2002). "Molecular cloning, gene localization, and structure of human cyclin B3." Biochem Biophys Res Commun **291**(2): 406-13.

Lynch, K. W. (2004). "Consequences of regulated pre-mRNA splicing in the immune system." Nat Rev Immunol **4**(12): 931-40.

Macchi, P., L. Notarangelo, S. Giliani, D. Strina, M. Repetto, M. G. Sacco, P. Vezzoni and A. Villa (1995). "The genomic organization of the human transcription factor 3 (TFE3) gene." Genomics **28**(3): 491-4.

Macias, M. J., S. Wiesner and M. Sudol (2002). "WW and SH3 domains, two different scaffolds to recognize proline-rich ligands." FEBS Lett **513**(1): 30-7.

Macias, M. J., M. Hyvonen, E. Baraldi, J. Schultz, M. Sudol, M. Saraste and H. Oschkinat (1996). "Structure of the WW domain of a kinase-associated protein complexed with a proline-rich peptide." Nature **382**(6592): 646-9.

Makalowski, W. (2003). "Genomics. Not junk after all." Science **300**(5623): 1246-7.

Maniatis, T. and B. Tasic (2002). "Alternative pre-mRNA splicing and proteome expansion in metazoans." Nature **418**(6894): 236-43.

Maquat, L. E. (2004). "Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics." Nat Rev Mol Cell Biol **5(**2): 89-99.

Marino, M., N. Archidiacono, A. Franze, M. Rosati, M. Rocchi, A. Ballabio and G. Grimaldi (1993). "A novel X-linked member of the human zinc finger protein gene family: isolation, mapping, and expression." Mamm Genome **4(**5): 252-7.

Mark, G. E., T. W. Seeley, T. B. Shows and J. D. Mountz (1986). "Pks, a raf-related sequence in humans." Proc Natl Acad Sci U S A **83**(17): 6312-6.

Martin, J., C. Han, L. A. Gordon, A. Terry, S. Prabhakar, X. She, G. Xie, U. Hellsten, Y. M. Chan, M. Altherr, O. Couronne, A. Aerts, E. Bajorek, S. Black, H. Blumer, E. Branscomb, N. C. Brown, W. J. Bruno, J. M. Buckingham, D. F. Callen, C. S. Campbell, M. L. Campbell, E. W. Campbell, C. Caoile, J. F. Challacombe, L. A. Chasteen, O. Chertkov, H. C. Chi, M. Christensen, L. M. Clark, J. D. Cohn, M. Denys, J. C. Detter, M. Dickson, M. Dimitrijevic-Bussod, J. Escobar, J. J. Fawcett, D. Flowers, D. Fotopulos, T. Glavina, M. Gomez, E. Gonzales, D. Goodstein, L. A. Goodwin, D. L. Grady, I. Grigoriev, M. Groza, N. Hammon, T. Hawkins, L. Haydu, C. E. Hildebrand, W. Huang, S. Israni, J. Jett, P. B. Jewett, K. Kadner, H. Kimball, A. Kobayashi, M. C. Krawczyk, T. Leyba, J. L. Longmire, F. Lopez, Y. Lou, S. Lowry, T. Ludeman, C. F. Manohar, G. A. Mark, K. L. McMurray, L. J. Meincke, J. Morgan, R. K. Moyzis, M. O. Mundt, A. C. Munk, R. D. Nandkeshwar, S. Pitluck, M. Pollard, P. Predki, B. Parson-Quintana, L. Ramirez, S. Rash, J. Retterer, D. O. Ricke, D. L. Robinson, A. Rodriguez, A. Salamov, E. H. Saunders, D. Scott, T. Shough, R. L. Stallings, M. Stalvey, R. D. Sutherland, R. Tapia, J. G. Tesmer, N. Thayer, L. S. Thompson, H. Tice, D. C. Torney, M. Tran-Gyamfi, M. Tsai, L. E. Ulanovsky, A. Ustaszewska, N. Vo, P. S. White, A. L. Williams, P. L. Wills, J. R. Wu, K. Wu, J. Yang, P. Dejong, D. Bruce, N. A. Doggett, L. Deaven, J. Schmutz, J. Grimwood, P. Richardson, D. S. Rokhsar, E. E. Eichler, P. Gilna, S. M. Lucas, R.

M. Myers, E. M. Rubin and L. A. Pennacchio (2004). "The sequence and analysis of duplication-rich human chromosome 16." Nature **432**(7020): 988-94.

Mazzarella, R. and D. Schlessinger (1998). "Pathological consequences of sequence duplications in the human genome." Genome Res **8**(10): 1007-21.

Mendell, J. T., C. M. ap Rhys and H. C. Dietz (2002). "Separable roles for rent1/hUpf1 in altered splicing and decay of nonsense transcripts." Science **298**(5592): 419-22.

Meyer, A. and M. Schartl (1999). "Gene and genome duplications in vertebrates: the one-to-four (-to-eight in fish) rule and the evolution of novel gene functions." Curr Opin Cell Biol **11**(6): 699-704.

Meyerson, M., G. H. Enders, C. L. Wu, L. K. Su, C. Gorka, C. Nelson, E. Harlow and L. H. Tsai (1992). "A family of human cdc2-related protein kinases." Embo J **11**(8): 2909-17.

Mignone, F., C. Gissi, S. Liuni and G. Pesole (2002). "Untranslated regions of mRNAs." Genome Biol **3**(3): REVIEWS0004.

Mironov, A. A. and M. S. Gel'fand (2004). "Prediction and computer analysis of the exon-intron structure of human genes." Genome Biol **38**(1): 82-91.

Modrek, B. and C. J. Lee (2003). "Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss." Nat Genet **34**(2): 177-80.

Modrek, B. and C. Lee (2002). "A genomic view of alternative splicing." Nat Genet **30**(1): 13-9.

Modrek, B., A. Resch, C. Grasso and C. Lee (2001). "Genome-wide detection of alternative splicing in expressed sequences of human genes." Nucleic Acids Res **29**(13): 2850-9.

Mohapatra, B., S. Verma, S. Shankar and A. Suri (1998). "Molecular cloning of human testis mRNA specifically expressed in haploid germ cells, having structural homology with the A-kinase anchoring proteins." Biochem Biophys Res Commun **244**(2): 540-5.

Monaco, A. P., R. L. Neve, C. Colletti-Feener, C. J. Bertelson, D. M. Kurnit and L. M. Kunkel (1986). "Isolation of candidate cDNAs for portions of the Duchenne muscular dystrophy gene." Nature **323**(6089): 646-50.

Moreau-Aubry, A., S. Le Guiner, N. Labarriere, M. C. Gesnel, F. Jotereau and R. Breathnach (2000). "A processed pseudogene codes for a new antigen recognized by a CD8(+) T cell clone on melanoma." J Exp Med **191**(9): 1617-24.

Mulligan, G. J., W. Guo, S. Wormsley and D. M. Helfman (1992). "Polypyrimidine tract binding protein interacts with sequences involved in alternative splicing of beta-tropomyosin pre-mRNA." J Biol Chem **267**(35): 25480-7.

Mungall, A. J., S. A. Palmer, S. K. Sims, C. A. Edwards, J. L. Ashurst, L. Wilming, M. C. Jones, R. Horton, S. E. Hunt, C. E. Scott, J. G. Gilbert, M. E. Clamp, G. Bethel, S. Milne, R. Ainscough, J. P. Almeida, K. D. Ambrose, T. D. Andrews, R. I. Ashwell, A. K. Babbage, C. L. Bagguley, J. Bailey, R. Banerjee, D. J. Barker, K. F. Barlow, K. Bates, D. M. Beare, H.

Beasley, O. Beasley, C. P. Bird, S. Blakey, S. Bray-Allen, J. Brook, A. J. Brown, J. Y. Brown, D. C. Burford, W. Burrill, J. Burton, C. Carder, N. P. Carter, J. C. Chapman, S. Y. Clark, G. Clark, C. M. Clee, S. Clegg, V. Cobley, R. E. Collier, J. E. Collins, L. K. Colman, N. R. Corby, G. J. Coville, K. M. Culley, P. Dhami, J. Davies, M. Dunn, M. E. Earthrowl, A. E. Ellington, K. A. Evans, L. Faulkner, M. D. Francis, A. Frankish, J. Frankland, L. French, P. Garner, J. Garnett, M. J. Ghori, L. M. Gilby, C. J. Gillson, R. J. Glithero, D. V. Grafham, M. Grant, S. Gribble, C. Griffiths, M. Griffiths, R. Hall, K. S. Halls, S. Hammond, J. L. Harley, E. A. Hart, P. D. Heath, R. Heathcott, S. J. Holmes, P. J. Howden, K. L. Howe, G. R. Howell, E. Huckle, S. J. Humphray, M. D. Humphries, A. R. Hunt, C. M. Johnson, A. A. Joy, M. Kay, S. J. Keenan, A. M. Kimberley, A. King, G. K. Laird, C. Langford, S. Lawlor, D. A. Leongamornlert, M. Leversha, C. R. Lloyd, D. M. Lloyd, J. E. Loveland, J. Lovell, S. Martin, M. Mashreghi-Mohammadi, G. L. Maslen, L. Matthews, O. T. McCann, S. J. McLaren, K. McLay, A. McMurray, M. J. Moore, J. C. Mullikin, D. Niblett, T. Nickerson, K. L. Novik, K. Oliver, E. K. Overton-Larty, A. Parker, R. Patel, A. V. Pearce, A. I. Peck, B. Phillimore, S. Phillips, R. W. Plumb, K. M. Porter, Y. Ramsey, S. A. Ranby, C. M. Rice, M. T. Ross, S. M. Searle, H. K. Sehra, E. Sheridan, C. D. Skuce, S. Smith, M. Smith, L. Spraggon, S. L. Squares, C. A. Steward, N. Sycamore, G. Tamlyn-Hall, J. Tester, A. J. Theaker, D. W. Thomas, A. Thorpe, A. Tracey, A. Tromans, B. Tubby, M. Wall, J. M. Wallis, A. P. West, S. S. White, S. L. Whitehead, H. Whittaker, A. Wild, D. J. Willey, T. E. Wilmer, J. M. Wood, P. W. Wray, J. C. Wyatt, L. Young, R. M. Younger, D. R. Bentley, A. Coulson, R. Durbin, T. Hubbard, J. E. Sulston, I. Dunham, J. Rogers and S. Beck (2003). "The DNA sequence and analysis of human chromosome 6." Nature **425**(6960): 805-11.

Mural, R. J., M. D. Adams, E. W. Myers, H. O. Smith, G. L. Miklos, R. Wides, A. Halpern, P. W. Li, G. G. Sutton, J. Nadeau, S. L. Salzberg, R. A. Holt, C. D. Kodira, F. Lu, L. Chen, Z. Deng, C. C. Evangelista, W. Gan, T. J. Heiman, J. Li, Z. Li, G. V. Merkulov, N. V. Milshina, A. K. Naik, R. Qi, B. C. Shue, A. Wang, J. Wang, X. Wang, X. Yan, J. Ye, S. Yooseph, Q. Zhao, L. Zheng, S. C. Zhu, K. Biddick, R. Bolanos, A. L. Delcher, I. M. Dew, D. Fasulo, M. J. Flanigan, D. H. Huson, S. A. Kravitz, J. R. Miller, C. M. Mobarry, K. Reinert, K. A. Remington, Q. Zhang, X. H. Zheng, D. R. Nusskern, Z. Lai, Y. Lei, W. Zhong, A. Yao, P. Guan, R. R. Ji, Z. Gu, Z. Y. Wang, F. Zhong, C. Xiao, C. C. Chiang, M. Yandell, J. R. Wortman, P. G. Amanatides, S. L. Hladun, E. C. Pratts, J. E. Johnson, K. L. Dodson, K. J. Woodford, C. A. Evans, B. Gropman, D. B. Rusch, E. Venter, M. Wang, T. J. Smith, J. T. Houck, D. E. Tompkins, C. Haynes, D. Jacob, S. H. Chin, D. R. Allen, C. E. Dahlke, R. Sanders, K. Li, X. Liu, A. A. Levitsky, W. H. Majoros, Q. Chen, A. C. Xia, J. R. Lopez, M. T. Donnelly, M. H. Newman, A. Glodek, C. L. Kraft, M. Nodell, F. Ali, H. J. An, D. Baldwin-Pitts, K. Y. Beeson, S. Cai, M. Carnes, A. Carver, P. M. Caulk, A. Center, Y. H. Chen, M. L. Cheng, M. D. Coyne, M. Crowder, S. Danaher, L. B. Davenport, R. Desilets, S. M. Dietz, L. Doup, P. Dullaghan, S. Ferriera, C. R. Fosler, H. C. Gire, A. Gluecksmann, J. D. Gocayne, J. Gray, B. Hart, J. Haynes, J. Hoover, T. Howland, C. Ibegwam, M. Jalali, D. Johns, L. Kline, D. S. Ma, S. MacCawley, A. Magoon, F. Mann, D. May, T. C. McIntosh, S. Mehta, L. Moy, M. C. Moy, B. J. Murphy, S. D. Murphy, K. A. Nelson, Z. Nuri, K. A. Parker, A. C. Prudhomme, V. N. Puri, H. Qureshi, J. C. Raley, M. S. Reardon, M. A. Regier, Y. H. Rogers, D. L. Romblad, J. Schutz, J. L. Scott, R. Scott, C. D. Sitter, M. Smallwood, A. C. Sprague, E. Stewart, R. V. Strong, E. Suh, K. Sylvester, R. Thomas, N. N. Tint, C. Tsonis, G. Wang, G. Wang, M. S. Williams, S. M. Williams, S. M. Windsor, K. Wolfe, M. M. Wu, J. Zaveri, K. Chaturvedi, A. E. Gabrielian, Z. Ke, J. Sun, G. Subramanian, J. C. Venter, C. M. Pfannkoch, M. Barnstead and L. D. Stephenson (2002). "A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome." Science **296**(5573): 1661-71.

Myers, E. W., G. G. Sutton, A. L. Delcher, I. M. Dew, D. P. Fasulo, M. J. Flanigan, S. A. Kravitz, C. M. Mobarry, K. H. Reinert, K. A. Remington, E. L. Anson, R. A. Bolanos, H. H. Chou, C. M. Jordan, A. L. Halpern, S. Lonardi, E. M. Beasley, R. C. Brandon, L. Chen, P. J. Dunn, Z. Lai, Y. Liang, D. R. Nusskern, M. Zhan, Q. Zhang, X. Zheng, G. M. Rubin, M. D. Adams and J. C. Venter (2000). "A whole-genome assembly of Drosophila." Science **287**(5461): 2196-204.

Nagase, T., K. Ishikawa, N. Miyajima, A. Tanaka, H. Kotani, N. Nomura and O. Ohara (1998). "Prediction of the coding sequences of unidentified human genes. IX. The complete sequences of 100 new cDNA clones from brain which can code for large proteins in vitro." DNA Res **5**(1): 31-9.

Nagoshi, R. N., M. McKeown, K. C. Burtis, J. M. Belote and B. S. Baker (1988). "The control of alternative splicing at genes regulating sexual differentiation in D. melanogaster." Cell **53**(2): 229-36.

Nakachi, Y., T. Hayakawa, H. Oota, K. Sumiyama, L. Wang and S. Ueda (1997). "Nucleotide compositional constraints on genomes generate alanine-, glycine-, and proline-rich structures in transcription factors." Mol Biol Evol **14**(10): 1042-9.

Nakai, K. and P. Horton (1999). "PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization." Trends Biochem Sci **24**(1): 34-6.

Nakanishi, T., R. Kekuda, Y. J. Fei, T. Hatanaka, M. Sugawara, R. G. Martindale, F. H. Leibach, P. D. Prasad and V. Ganapathy (2001). "Cloning and functional characterization of a new subtype of the amino acid transport system N." Am J Physiol Cell Physiol **281**(6): C1757-68.

Nam, J. W., K. R. Shin, J. Han, Y. Lee, V. N. Kim and B. T. Zhang (2005). "Human microRNA prediction through a probabilistic co-learning model of sequence and structure." Nucleic Acids Res **33**(11): 3570-81.

Nicholas, K.B., Nicholas H.B. Jr., and Deerfield, D.W. II. 1997 GeneDoc: Analysis and Visualization of Genetic Variation, EMBNEW.NEWS **4:**14

Ning, Z., A. J. Cox and J. C. Mullikin (2001). "SSAHA: a fast search method for large DNA databases." Genome Res **11**(10): 1725-9.

Nolan, K. F., S. Kaluz, J. M. Higgins, D. Goundis and K. B. Reid (1992). "Characterization of the human properdin gene." Biochem J **287** (Pt 1): 291-7.

Numata, M. and J. Orlowski (2001). "Molecular cloning and characterization of a novel (Na+,K+)/H+ exchanger localized to the trans-Golgi network." J Biol Chem **276**(20): 17387-94.

Nurtdinov, R. N., Artamonova, II, A. A. Mironov and M. S. Gelfand (2003). "Low conservation of alternative splicing patterns in the human and mouse genomes." Hum Mol Genet **12**(11): 1313-20.

Ohno, S. (1967). Sex chromosomes and sex-linked genes. Berlin; New York [etc.], Springer-Verlag.

Okada, T., T. Fujii, N. Tanuma, S. Mitsuhashi, T. Urano, Y. Araki, H. Shima and K. Kikuchi (2004). "Analysis of isoform specific function of PP1 catalytic subunits in mammalian cells using siRNA." Int J Oncol **25**(5): 1383-8.

Okazawa, H., M. Sudol and T. Rich (2001). "PQBP-1 (Np/PQ): a polyglutamine tract-binding and nuclear inclusion-forming protein." Brain Res Bull **56**(3-4): 273-80.

Okazawa, H., T. Rich, A. Chang, X. Lin, M. Waragai, M. Kajikawa, Y. Enokido, A. Komuro, S. Kato, M. Shibata, H. Hatanaka, M. M. Mouradian, M. Sudol and I. Kanazawa (2002). "Interaction between mutant ataxin-1 and PQBP-1 affects transcription and cell death." Neuron **34**(5): 701-13.

Okuda, T., H. Hattori, S. Takeuchi, J. Shimizu, H. Ueda, J. J. Palvimo, I. Kanazawa, H. Kawano, M. Nakagawa and H. Okazawa (2003). "PQBP-1 transgenic mice show a late-onset motor neuron disease-like phenotype." Hum Mol Genet **12**(7): 711-25.

Oliva, M. M., T. C. Wu and V. W. Yang (1993). "Isolation and characterization of a differentiation-dependent gene in the human colonic cell line HT29-18." Arch Biochem Biophys **302**(1): 183-92.

Olson, M. V., J. E. Dutchik, M. Y. Graham, G. M. Brodeur, C. Helms, M. Frank, M. MacCollin, R. Scheinman and T. Frank (1986). "Random-clone strategy for genomic restriction mapping in yeast." Proc Natl Acad Sci U S A **83**(20): 7826-30.

Osato, N., M. Itoh, H. Konno, S. Kondo, K. Shibata, P. Carninci, T. Shiraki, A. Shinagawa, T. Arakawa, S. Kikuchi, K. Sato, J. Kawai and Y. Hayashizaki (2002). "A computer-based method of selecting clones for a full-length cDNA project: simultaneous collection of negligibly redundant and variant cDNAs." Genome Res **12**(7): 1127-34.

Ota, T., Y. Suzuki, T. Nishikawa, T. Otsuki, T. Sugiyama, R. Irie, A. Wakamatsu, K. Hayashi, H. Sato, K. Nagai, K. Kimura, H. Makita, M. Sekine, M. Obayashi, T. Nishi, T. Shibahara, T. Tanaka, S. Ishii, J. Yamamoto, K. Saito, Y. Kawai, Y. Isono, Y. Nakamura, K. Nagahari, K. Murakami, T. Yasuda, T. Iwayanagi, M. Wagatsuma, A. Shiratori, H. Sudo, T. Hosoiri, Y. Kaku, H. Kodaira, H. Kondo, M. Sugawara, M. Takahashi, K. Kanda, T. Yokoi, T. Furuya, E. Kikkawa, Y. Omura, K. Abe, K. Kamihara, N. Katsuta, K. Sato, M. Tanikawa, M. Yamazaki, K. Ninomiya, T. Ishibashi, H. Yamashita, K. Murakawa, K. Fujimori, H. Tanai, M. Kimata, M. Watanabe, S. Hiraoka, Y. Chiba, S. Ishida, Y. Ono, S. Takiguchi, S. Watanabe, M. Yosida, T. Hotuta, J. Kusano, K. Kanehori, A. Takahashi-Fujii, H. Hara, T. O. Tanase, Y. Nomura, S. Togiya, F. Komai, R. Hara, K. Takeuchi, M. Arita, N. Imose, K. Musashino, H. Yuuki, A. Oshima, N. Sasaki, S. Aotsuka, Y. Yoshikawa, H. Matsunawa, T. Ichihara, N. Shiohata, S. Sano, S. Moriya, H. Momiyama, N. Satoh, S. Takami, Y. Terashima, O. Suzuki, S. Nakagawa, A. Senoh, H. Mizoguchi, Y. Goto, F. Shimizu, H. Wakebe, H. Hishigaki, T. Watanabe, A. Sugiyama, M. Takemoto, B. Kawakami, M. Yamazaki, K. Watanabe, A. Kumagai, S. Itakura, Y. Fukuzumi, Y. Fujimori, M. Komiyama, H. Tashiro, A. Tanigami, T. Fujiwara, T. Ono, K. Yamada, Y. Fujii, K. Ozaki, M. Hirao, Y. Ohmori, A. Kawabata, T. Hikiji, N. Kobatake, H. Inagaki, Y. Ikema, S. Okamoto, R. Okitani, T. Kawakami, S. Noguchi, T. Itoh, K. Shigeta, T. Senba, K. Matsumura, Y. Nakajima, T. Mizuno, M. Morinaga, M. Sasaki, T. Togashi, M. Oyama, H. Hata, M. Watanabe, T. Komatsu, J. Mizushima-Sugano, T. Satoh, Y. Shirai, Y. Takahashi, K. Nakagawa, K. Okumura, T. Nagase, N. Nomura, H. Kikuchi, Y. Masuho, R. Yamashita, K. Nakai, T. Yada, Y. Nakamura, O. Ohara, T. Isogai and S. Sugano (2004). "Complete sequencing and characterization of 21,243 full-length human cDNAs." Nat Genet **36**(1): 40-5.

Otterlei, M., T. Haug, T. A. Nagelhus, G. Slupphaug, T. Lindmo and H. E. Krokan (1998). "Nuclear and mitochondrial splice forms of human uracil-DNA glycosylase contain a complex nuclear localisation signal and a strong classical mitochondrial localisation signal, respectively." Nucleic Acids Res **26**(20): 4611-7.

Ovcharenko, I., G. G. Loots, R. C. Hardison, W. Miller and L. Stubbs (2004). "zPicture: dynamic alignment and visualization tool for analyzing conservation profiles." Genome Res **14**(3): 472-7.

Pagani, F., E. Buratti, C. Stuani, R. Bendix, T. Dork and F. E. Baralle (2002). "A new type of mutation causes a splicing defect in ATM." Nat Genet **30**(4): 426-9.

Pan, L., J. Xu, R. Yu, M. M. Xu, Y. X. Pan and G. W. Pasternak (2005). "Identification and characterization of six new alternatively spliced variants of the human mu opioid receptor gene, Oprm." Neuroscience **133**(1): 209-20.

Parker, H. G., L. V. Kim, N. B. Sutter, S. Carlson, T. D. Lorentzen, T. B. Malek, G. S. Johnson, H. B. DeFrance, E. A. Ostrander and L. Kruglyak (2004). "Genetic structure of the purebred domestic dog." Science **304**(5674): 1160-4.

Parra, G., P. Agarwal, J. F. Abril, T. Wiehe, J. W. Fickett and R. Guigo (2003). "Comparative gene prediction in human and mouse." Genome Res **13**(1): 108-

Pearson, W. R. (1990). "Rapid and sensitive sequence comparison with FASTP and FASTA." Methods Enzymol **183**: 63-98.

Pearson, W.R. and D. J. Lipman (1988) Improved Tools for Biological Sequence Comparison.PNAS **85**:2444- 2448.

Pennacchio, L. A., M. Olivier, J. A. Hubacek, J. C. Cohen, D. R. Cox, J. C. Fruchart, R. M. Krauss and E. M. Rubin (2001). "An apolipoprotein influencing triglycerides in humans and mice revealed by comparative sequencing." Science **294**(5540): 169-73.

Perez, C., J. Vandesompele, I. Vandenbroucke, G. Holtappels, F. Speleman, P. Gevaert, P. Van Cauwenberge and C. Bachert (2003). "Quantitative real time polymerase chain reaction for measurement of human interleukin-5 receptor alpha spliced isoforms mRNA." BMC Biotechnol **3**(1): 17.

Phylactides, M., R. Rowntree, H. Nuthall, D. Ussery, A. Wheeler and A. Harris (2002). "Evaluation of potential regulatory elements identified as DNase I hypersensitive sites in the CFTR gene." Eur J Biochem **269**(2): 553-9.

Pintard, L., D. Kressler and B. Lapeyre (2000). "Spb1p is a yeast nucleolar protein associated with Nop1p and Nop58p that is able to bind S-adenosyl-L-methionine in vitro." Mol Cell Biol **20**(4): 1370-81.

Pold, M., J. Zhou, G. L. Chen, J. M. Hall, R. A. Vescio and J. R. Berenson (1999). "Identification of a new, unorthodox member of the MAGE gene family." Genomics **59**(2): 161-7.

Ponger, L., L. Duret and D. Mouchiroud (2001). "Determinants of CpG islands: expression in early embryo and isochore structure." Genome Res **11**(11): 1854-60.

Porteous M.E., A. Curtis, S. Lindsay, O. Williams, D. Goudie, S. Kamakari and S.S. Bhattacharya (1992). "The gene for Aarskog syndrome is located between DXS255 and DXS566 (Xp11.2-Xq13)." Genomics **14**(2): 298-301.

Proudfoot, N. J., A. Furger and M. J. Dye (2002). "Integrating mRNA processing with transcription." Cell **108**(4): 501-12.

Rajavel, K. S. and E. F. Neufeld (2001). "Nonsense-mediated decay of human HEXA mRNA." Mol Cell Biol **21**(16): 5512-9.

Rao, V. N., K. Huebner, M. Isobe, A. ar-Rushdi, C. M. Croce and E. S. Reddy (1989). "elk, tissue-specific ets-related genes on chromosomes X and 14 near translocation breakpoints." Science **244**(4900): 66-70.

Renpenning, H., J. W. Gerrard, W. A. Zaleski and T. Tabata (1962). "Familial sex-linked mental retardation." Can Med Assoc J **87**: 954-6.

Ricci, V., S. Regis, M. Di Duca and M. Filocamo (2003). "An Alu-mediated rearrangement as cause of exon skipping in Hunter disease." Hum Genet **112**(4): 419-25.

Rinn, J. L., G. Euskirchen, P. Bertone, R. Martone, N. M. Luscombe, S. Hartman, P. M. Harrison, F. K. Nelson, P. Miller, M. Gerstein, S. Weissman and M. Snyder (2003). "The transcriptional activity of human Chromosome 22." Genes Dev **17**(4): 529-40.

Roberts, R. G. (1995). "Dystrophin, its gene, and the dystrophinopathies." Adv Genet **33**: 177-231.

Rocques, P. J., J. Clark, S. Ball, J. Crew, S. Gill, Z. Christodoulou, R. H. Borts, E. J. Louis, K. E. Davies and C. S. Cooper (1995). "The human SB1.8 gene (DXS423E) encodes a putative chromosome segregation protein conserved in lower eukaryotes and prokaryotes." Hum Mol Genet **4**(2): 243-9.

Ross, M. T., D. V. Grafham, A. J. Coffey, S. Scherer, K. McLay, D. Muzny, M. Platzer, G. R. Howell, C. Burrows, C. P. Bird, A. Frankish, F. L. Lovell, K. L. Howe, J. L. Ashurst, R. S. Fulton, R. Sudbrak, G. Wen, M. C. Jones, M. E. Hurles, T. D. Andrews, C. E. Scott, S. Searle, J. Ramser, A. Whittaker, R. Deadman, N. P. Carter, S. E. Hunt, R. Chen, A. Cree, P. Gunaratne, P. Havlak, A. Hodgson, M. L. Metzker, S. Richards, G. Scott, D. Steffen, E. Sodergren, D. A. Wheeler, K. C. Worley, R. Ainscough, K. D. Ambrose, M. A. Ansari-Lari, S. Aradhya, R. I. Ashwell, A. K. Babbage, C. L. Bagguley, A. Ballabio, R. Banerjee, G. E. Barker, K. F. Barlow, I. P. Barrett, K. N. Bates, D. M. Beare, H. Beasley, O. Beasley, A. Beck, G. Bethel, K. Blechschmidt, N. Brady, S. Bray-Allen, A. M. Bridgeman, A. J. Brown, M. J. Brown, D. Bonnin, E. A. Bruford, C. Buhay, P. Burch, D. Burford, J. Burgess, W. Burrill, J. Burton, J. M. Bye, C. Carder, L. Carrel, J. Chako, J. C. Chapman, D. Chavez, E. Chen, G. Chen, Y. Chen, Z. Chen, C. Chinault, A. Ciccodicola, S. Y. Clark, G. Clarke, C. M. Clee, S. Clegg, K. Clerc-Blankenburg, K. Clifford, V. Cobley, C. G. Cole, J. S. Conquer, N. Corby, R. E. Connor, R. David, J. Davies, C. Davis, J. Davis, O. Delgado, D. Deshazo, P. Dhami, Y. Ding, H. Dinh, S. Dodsworth, H. Draper, S. Dugan-Rocha, A. Dunham, M. Dunn, K. J. Durbin, I. Dutta, T. Eades, M. Ellwood, A. Emery-Cohen, H. Errington, K. L. Evans, L. Faulkner, F. Francis, J. Frankland, A. E. Fraser, P. Galgoczy, J. Gilbert, R. Gill, G. Glockner, S. G. Gregory, S. Gribble, C. Griffiths, R. Grocock, Y. Gu, R. Gwilliam, C. Hamilton, E. A. Hart, A. Hawes, P. D. Heath, K. Heitmann, S. Hennig, J. Hernandez, B. Hinzmann, S. Ho, M. Hoffs, P. J. Howden, E. J. Huckle, J. Hume, P. J. Hunt, A. R. Hunt, J. Isherwood, L. Jacob, D. Johnson, S. Jones, P. J. de Jong, S. S. Joseph, S. Keenan, S. Kelly, J. K. Kershaw, Z. Khan, P. Kioschis, S. Klages, A. J. Knights, A. Kosiura, C. Kovar-Smith, G. K. Laird, C. Langford, S. Lawlor, M. Leversha, L. Lewis, W. Liu, C. Lloyd, D. M. Lloyd, H. Loulseged, J. E. Loveland, J. D. Lovell, R. Lozado, J. Lu, R. Lyne, J. Ma, M. Maheshwari, L. H. Matthews, J. McDowall, S. McLaren, A. McMurray, P. Meidl, T. Meitinger, S. Milne, G. Miner, S. L. Mistry, M. Morgan, S. Morris, I. Muller, J. C. Mullikin, N. Nguyen, G. Nordsiek, G. Nyakatura, C. N. O'Dell, G. Okwuonu, S. Palmer, R. Pandian, D. Parker, J. Parrish, S. Pasternak, D. Patel, A. V. Pearce, D. M. Pearson, S. E. Pelan, L. Perez, K. M. Porter, Y. Ramsey, K. Reichwald, S. Rhodes, K. A. Ridler, D. Schlessinger, M. G. Schueler, H. K. Sehra, C. Shaw-Smith, H. Shen, E. M. Sheridan, R. Shownkeen, C. D. Skuce, M. L. Smith, E. C. Sotheran, H. E. Steingruber, C. A. Steward, R. Storey, R. M. Swann, D. Swarbreck, P. E. Tabor, S. Taudien, T. Taylor, B. Teague, K. Thomas, A. Thorpe, K. Timms, A. Tracey, S. Trevanion, A. C. Tromans, M. d'Urso, D. Verduzco, D. Villasana, L. Waldron, M. Wall, Q. Wang, J. Warren, G. L. Warry, X. Wei, A. West, S. L. Whitehead, M. N. Whiteley, J. E. Wilkinson, D.

L. Willey, G. Williams, L. Williams, A. Williamson, H. Williamson, L. Wilming, R. L. Woodmansey, P. W. Wray, J. Yen, J. Zhang, J. Zhou, H. Zoghbi, S. Zorilla, D. Buck, R. Reinhardt, A. Poustka, A. Rosenthal, H. Lehrach, A. Meindl, P. J. Minx, L. W. Hillier, H. F. Willard, R. K. Wilson, R. H. Waterston, C. M. Rice, M. Vaudin, A. Coulson, D. L. Nelson, G. Weinstock, J. E. Sulston, R. Durbin, T. Hubbard, R. A. Gibbs, S. Beck, J. Rogers and D. R. Bentley (2005). "The DNA sequence of the human X chromosome." Nature **434**(7031): 325-37.

Ruiz-Echevarria, M. J., C. I. Gonzalez and S. W. Peltz (1998). "Identifying the right stop: determining how the surveillance complex recognizes and degrades an aberrant mRNA." Embo J **17**(2): 575-89.

Saccone, S. and G. Bernardi (2001). "Human chromosomal banding by in situ hybridization of isochores." Methods Cell Sci **23**(1-3): 7-15.

Sachidanandam, R., D. Weissman, S. C. Schmidt, J. M. Kakol, L. D. Stein, G. Marth, S. Sherry, J. C. Mullikin, B. J. Mortimore, D. L. Willey, S. E. Hunt, C. G. Cole, P. C. Coggill, C. M. Rice, Z. Ning, J. Rogers, D. R. Bentley, P. Y. Kwok, E. R. Mardis, R. T. Yeh, B. Schultz, L. Cook, R. Davenport, M. Dante, L. Fulton, L. Hillier, R. H. Waterston, J. D. McPherson, B. Gilman, S. Schaffner, W. J. Van Etten, D. Reich, J. Higgins, M. J. Daly, B. Blumenstiel, J. Baldwin, N. Stange-Thomann, M. C. Zody, L. Linton, E. S. Lander and D. Altshuler (2001). "A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms." Nature **409**(6822): 928-33.

Saha, S., A. B. Sparks, C. Rago, V. Akmaev, C. J. Wang, B. Vogelstein, K. W. Kinzler and V. E. Velculescu (2002). "Using the transcriptome to annotate the genome." Nat Biotechnol **20**(5): 508-12.

Sanger, F., A. R. Coulson, G. F. Hong, D. F. Hill and G. B. Petersen (1982). "Nucleotide sequence of bacteriophage lambda DNA." J Mol Biol **162**(4): 729-73.

Sanger, F., A. R. Coulson, T. Friedmann, G. M. Air, B. G. Barrell, N. L. Brown, J. C. Fiddes, C. A. Hutchison, 3rd, P. M. Slocombe and M. Smith (1978). "The nucleotide sequence of bacteriophage phiX174." J Mol Biol **125**(2): 225-46.

Sanger, F., G. M. Air, B. G. Barrell, N. L. Brown, A. R. Coulson, C. A. Fiddes, C. A. Hutchison, P. M. Slocombe and M. Smith (1977). "Nucliotide sequence of bacteriophage phi X174 DNA." Nature **265**(5596): 687-95.

Schadt, E. E., S. W. Edwards, D. GuhaThakurta, D. Holder, L. Ying, V. Svetnik, A. Leonardson, K. W. Hart, A. Russell, G. Li, G. Cavet, J. Castle, P. McDonagh, Z. Kan, R. Chen, A. Kasarskis, M. Margarint, R. M. Caceres, J. M. Johnson, C. D. Armour, P. W. Garrett-Engele, N. F. Tsinoremas and D. D. Shoemaker (2004). "A comprehensive transcript index of the human genome generated using microarrays and computational approaches." Genome Biol **5**(10): R73.

Scheper, W., R. Zwart and F. Baas (2004). "Alternative splicing in the N-terminus of Alzheimer's presenilin 1." Neurogenetics **5**(4): 223-7.

Scherf, M., A. Klingenhoff and T. Werner (2000). "Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach." J Mol Biol **297**(3): 599-606.

Schmittgen, T. D., B. A. Zakrajsek, A. G. Mills, V. Gorn, M. J. Singer and M. W. Reed (2000). "Quantitative reverse transcription-polymerase chain reaction to study mRNA decay: comparison of endpoint and real-time methods." Anal Biochem **285**(2): 194-204.

Schmucker, D. and J. G. Flanagan (2004). "Generation of recognition diversity in the nervous system." Neuron **44**(2): 219-22.

Schmutz, J., J. Martin, A. Terry, O. Couronne, J. Grimwood, S. Lowry, L. A. Gordon, D. Scott, G. Xie, W. Huang, U. Hellsten, M. Tran-Gyamfi, X. She, S. Prabhakar, A. Aerts, M. Altherr, E. Bajorek, S. Black, E. Branscomb, C. Caoile, J. F. Challacombe, Y. M. Chan, M. Denys, J. C. Detter, J. Escobar, D. Flowers, D. Fotopulos, T. Glavina, M. Gomez, E. Gonzales, D. Goodstein, I. Grigoriev, M. Groza, N. Hammon, T. Hawkins, L. Haydu, S. Israni, J. Jett, K. Kadner, H. Kimball, A. Kobayashi, F. Lopez, Y. Lou, D. Martinez, C. Medina, J. Morgan, R. Nandkeshwar, J. P. Noonan, S. Pitluck, M. Pollard, P. Predki, J. Priest, L. Ramirez, J. Retterer, A. Rodriguez, S. Rogers, A. Salamov, A. Salazar, N. Thayer, H. Tice, M. Tsai, A. Ustaszewska, N. Vo, J. Wheeler, K. Wu, J. Yang, M. Dickson, J. F. Cheng, E. E. Eichler, A. Olsen, L. A. Pennacchio, D. S. Rokhsar, P. Richardson, S. M. Lucas, R. M. Myers and E. M. Rubin (2004). "The DNA sequence and comparative analysis of human chromosome 5." Nature **431**(7006): 268-74.

Schroer, A., S. Schneider, H. Ropers and H. Nothwang (1999). "Cloning and characterization of UXT, a novel gene in human Xp11, which is widely and abundantly expressed in tumor tissue." Genomics **56**(3): 340-3.

Schwahn, U., S. Lenzner, J. Dong, S. Feil, B. Hinzmann, G. van Duijnhoven, R. Kirschner, M. Hemberger, A. A. Bergen, T. Rosenberg, A. J. Pinckers, R. Fundele, A. Rosenthal, F. P. Cremers, H. H. Ropers and W. Berger (1998). "Positional cloning of the gene for X-linked retinitis pigmentosa 2." Nat Genet **19**(4): 327-32.

Schwartz, S., W. J. Kent, A. Smit, Z. Zhang, R. Baertsch, R. C. Hardison, D. Haussler and W. Miller (2003). "Human-mouse alignments with BLASTZ." Genome Res **13**(1): 103-7.

Schwerk, C. and K. Schulze-Osthoff (2005). "Regulation of Apoptosis by Alternative Pre-mRNA Splicing." Mol Cell **19**(1): 1-13.

Shapiro, M. B. and P. Senapathy (1987). "RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression." Nucleic Acids Res **15**(17): 7155-74.

Sharp, A., D. O. Robinson and P. Jacobs (2001). "Absence of correlation between late-replication and spreading of X inactivation in an X;autosome translocation." Hum Genet **109**(3): 295-302.

Sharp, A., D. O. Robinson and P. Jacobs (2001). "Absence of correlation between late-replication and spreading of X inactivation in an X;autosome translocation." Hum Genet **109**(3): 295-302.

Sharp, P. A. (1994). "Split genes and RNA splicing." Cell **77**(6): 805-15.

Shen, H., J. L. Kan and M. R. Green (2004). "Arginine-serine-rich domains bound at splicing enhancers contact the branchpoint to promote prespliceosome assembly." Mol Cell **13**(3): 367-76.

Shibata, Y., P. Carninci, K. Sato, N. Hayatsu, T. Shiraki, Y. Ishii, T. Arakawa, A. Hara, N. Ohsato, M. Izawa, K. Aizawa, M. Itoh, K. Shibata, A. Shinagawa, J. Kawai, Y. Ota, S. Kikuchi, N. Kishimoto, M. Muramatsu and Y. Hayashizaki (2001). "Removal of polyA tails from full-length cDNA libraries for high-efficiency sequencing." Biotechniques **31**(5): 1042, 1044, 1048-9.

Shoemaker, D. D., E. E. Schadt, C. D. Armour, Y. D. He, P. Garrett-Engele, P. D. McDonagh, P. M. Loerch, A. Leonardson, P. Y. Lum, G. Cavet, L. F. Wu, S. J. Altschuler, S. Edwards, J. King, J. S. Tsang, G. Schimmack, J. M. Schelter, J. Koch, M. Ziman, M. J. Marton, B. Li, P. Cundiff, T. Ward, J. Castle, M. Krolewski, M. R. Meyer, M. Mao, J. Burchard, M. J. Kidd, H. Dai, J. W. Phillips, P. S. Linsley, R. Stoughton, S. Scherer and M. S. Boguski (2001). "Experimental annotation of the human genome using microarray technology." Nature **409**(6822): 922-7.

Simmons, D.L. (1993). Cloning cell surface molecules by transient expression in mammalian cells. In Cellular Interactions and Development . (D. Hartley, ed.). IRL Press at Oxford University Press, Oxford, pp93-127.

Singer, S. S., D. N. Mannel, T. Hehlgans, J. Brosius and J. Schmitz (2004). "From "junk" to gene: curriculum vitae of a primate receptor isoform gene." J Mol Biol **341**(4): 883-6.

Singh, B. N., A. Suresh, G. UmaPrasad, S. Subramanian, M. Sultana, S. Goel, S. Kumar and L. Singh (2003). "A highly conserved human gene encoding a novel member of WD-repeat family of proteins (WDR13)." Genomics **81**(3): 315-28.

Singh, N. N., E. J. Androphy and R. N. Singh (2004). "The regulation and regulatory activities of alternative splicing of the SMN gene." Crit Rev Eukaryot Gene Expr **14**(4): 271-85.

Singh, R., J. Valcarcel and M. R. Green (1995). "Distinct binding specificities and functions of higher eukaryotic polypyrimidine tract-binding proteins." Science **268**(5214): 1173-6.

Sironi, M., G. Menozzi, L. Riva, R. Cagliani, G. P. Comi, N. Bresolin, R. Giorda and U. Pozzoli (2004). "Silencer elements as possible inhibitors of pseudoexon splicing." Nucleic Acids Res **32**(5): 1783-91.

Skaletsky, H., T. Kuroda-Kawaguchi, P. J. Minx, H. S. Cordum, L. Hillier, L. G. Brown, S. Repping, T. Pyntikova, J. Ali, T. Bieri, A. Chinwalla, A. Delehaunty, K. Delehaunty, H. Du, G. Fewell, L. Fulton, R. Fulton, T. Graves, S. F. Hou, P. Latrielle, S. Leonard, E. Mardis, R. Maupin, J. McPherson, T. Miner, W. Nash, C. Nguyen, P. Ozersky, K. Pepin, S. Rock, T. Rohlfing, K. Scott, B. Schultz, C. Strong, A. Tin-Wollam, S. P. Yang, R. H. Waterston, R. K. Wilson, S. Rozen and D. C. Page (2003). "The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes." Nature **423**(6942): 825-37.

Skandalis, A. and E. Uribe (2004). "A survey of splice variants of the human hypoxanthine phosphoribosyl transferase and DNA polymerase beta genes: products of alternative or aberrant splicing?" Nucleic Acids Res **32**(22): 6557-64.

Skordis, L. A., M. G. Dunckley, B. Yue, I. C. Eperon and F. Muntoni (2003). "Bifunctional antisense oligonucleotides provide a trans-acting splicing enhancer that stimulates SMN2 gene expression in patient fibroblasts." Proc Natl Acad Sci U S A **100**(7): 4114-9.

Smit, A.F., R. Hubley & P. Green RepeatMasker at http://repeatmasker.org

Smith, C. L. and C. R. Cantor (1986). "Approaches to physical mapping of the human genome." Cold Spring Harb Symp Quant Biol **51** Pt 1: 115-22.

Smith, Z. E. and D. R. Higgs (1999). "The pattern of replication at a human telomeric region (16p13.3): its relationship to chromosome structure and gene expression." Hum Mol Genet **8**(8): 1373-86.

Sogayar, M. C., A. A. Camargo, F. Bettoni, D. M. Carraro, L. C. Pires, R. B. Parmigiani, E. N. Ferreira, E. de Sa Moreira, D. d. O. L. M. do Rosario, A. J. Simpson, L. O. Cruz, T. L. Degaki, F. Festa, K. B. Massirer, M. C. Sogayar, F. C. Filho, L. P. Camargo, M. A. Cunha, S. J. De Souza, M. Faria, Jr., S. Giuliatti, L. Kopp, P. S. de Oliveira, P. B. Paiva, A. A. Pereira, D. G. Pinheiro, R. D. Puga, S. d. S. JE, D. M. Albuquerque, L. E. Andrade, G. S. Baia, M. R. Briones, A. M. Cavaleiro-Luna, J. M. Cerutti, F. F. Costa, E. Costanzi-Strauss, E. M. Espreafico, A. C. Ferrasi, E. S. Ferro, M. A. Fortes, J. R. Furchi, D. Giannella-Neto, G. H. Goldman, M. H. Goldman, A. Gruber, G. S. Guimaraes, C. Hackel, F. Henrique-Silva, E. T. Kimura, S. G. Leoni, C. Macedo, B. Malnic, B. C. Manzini, S. K. Marie, N. M. Martinez-Rossi, M. Menossi, E. C. Miracca, M. A. Nagai, F. G. Nobrega, M. P. Nobrega, S. M. Oba-Shinjo, M. K. Oliveira, G. M. Orabona, A. Y. Otsuka, M. L. Paco-Larson, B. M. Paixao, J. R. Pandolfi, M. I. Pardini, M. R. Passos Bueno, G. A. Passos, J. B. Pesquero, J. G. Pessoa, P. Rahal, C. A. Rainho, C. P. Reis, T. I. Ricca, V. Rodrigues, S. R. Rogatto, C. M. Romano, J. G. Romeiro, A. Rossi, R. G. Sa, M. M. Sales, S. C. Sant'Anna, P. L. Santarosa, F. Segato, W. A. Silva, Jr., I. D. Silva, N. P. Silva, A. Soares-Costa, M. F. Sonati, B. E. Strauss, E. H. Tajara, S. R. Valentini, F. E. Villanova, L. S. Ward and D. L. Zanette (2004). "A transcript finishing initiative for closing gaps in the human transcriptome." Genome Res **14**(7): 1413-23.

Solovyev, V.V., A. A. Salamov and C.B. Lawrence (1995). Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. Nucleic Acids Res **11(**22) 5156-63.

Sorek, R. and G. Ast (2003). "Intronic sequences flanking alternatively spliced exons are conserved between human and mouse." Genome Res **13**(7): 1631-7.

Sorek, R., G. Ast and D. Graur (2002). "Alu-containing exons are alternatively spliced." Genome Res **12**(7): 1060-7.

Sorek, R., G. Lev-Maor, M. Reznik, T. Dagan, F. Belinky, D. Graur and G. Ast (2004a). "Minimal conditions for exonization of intronic sequences: 5' splice site formation in alu exons." Mol Cell **14**(2): 221-31.

Sorek, R., R. Shemesh, Y. Cohen, O. Basechess, G. Ast and R. Shamir (2004b). "A non-EST-based method for exon-skipping prediction." Genome Res **14**(8): 1617-23.

Spencer, J. A., J. M. Watson and J. A. Graves (1991). "The X chromosome of marsupials shares a highly conserved region with eutherians." Genomics **9**(4): 598-604.

Stassar, M. J., G. Devitt, M. Brosius, L. Rinnab, J. Prang, T. Schradin, J. Simon, S. Petersen, A. Kopp-Schneider and M. Zoller (2001). "Identification of human renal cell carcinoma associated genes by suppression subtractive hybridization." Br J Cancer **85**(9): 1372-82.

Stevenson, R. E., C. W. Bennett, F. Abidi, T. Kleefstra, M. Porteous, R. J. Simensen, H. A. Lubs, B. C. Hamel and C. E. Schwartz (2005). "Renpenning syndrome comes into focus." Am J Med Genet A **134**(4): 415-21.

Stoltzfus, A., D. F. Spencer, M. Zuker, J. M. Logsdon, Jr. and W. F. Doolittle (1994). "Testing the exon theory of genes: the evidence from protein structure." Science **265**(5169): 202-7.

Strathdee, G., A. Sim and R. Brown (2004). "Control of gene expression by CpG island methylation in normal cells." Biochem Soc Trans **32**(Pt 6): 913-5.

Strausberg, R. L., E. A. Feingold, L. H. Grouse, J. G. Derge, R. D. Klausner, F. S. Collins, L. Wagner, C. M. Shenmen, G. D. Schuler, S. F. Altschul, B. Zeeberg, K. H. Buetow, C. F. Schaefer, N. K. Bhat, R. F. Hopkins, H. Jordan, T. Moore, S. I. Max, J. Wang, F. Hsieh, L. Diatchenko, K. Marusina, A. A. Farmer, G. M. Rubin, L. Hong, M. Stapleton, M. B. Soares, M. F. Bonaldo, T. L. Casavant, T. E. Scheetz, M. J. Brownstein, T. B. Usdin, S. Toshiyuki, P. Carninci, C. Prange, S. S. Raha, N. A. Loquellano, G. J. Peters, R. D. Abramson, S. J. Mullahy, S. A. Bosak, P. J. McEwan, K. J. McKernan, J. A. Malek, P. H. Gunaratne, S. Richards, K. C. Worley, S. Hale, A. M. Garcia, L. J. Gay, S. W. Hulyk, D. K. Villalon, D. M. Muzny, E. J. Sodergren, X. Lu, R. A. Gibbs, J. Fahey, E. Helton, M. Ketteman, A. Madan, S. Rodrigues, A. Sanchez, M. Whiting, A. Madan, A. C. Young, Y. Shevchenko, G. G. Bouffard, R. W. Blakesley, J. W. Touchman, E. D. Green, M. C. Dickson, A. C. Rodriguez, J. Grimwood, J. Schmutz, R. M. Myers, Y. S. Butterfield, M. I. Krzywinski, U. Skalska, D. E. Smailus, A. Schnerch, J. E. Schein, S. J. Jones and M. A. Marra (2002). "Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences." Proc Natl Acad Sci U S A **99**(26): 16899-903.

Su, A. I., M. P. Cooke, K. A. Ching, Y. Hakak, J. R. Walker, T. Wiltshire, A. P. Orth, R. G. Vega, L. M. Sapinoso, A. Moqrich, A. Patapoutian, G. M. Hampton, P. G. Schultz and J. B. Hogenesch (2002). "Large-scale analysis of the human and mouse transcriptomes." Proc Natl Acad Sci U S A **99**(7): 4465-70.

Sudhof, T. C. and J. Rizo (1996). "Synaptotagmins: C2-domain proteins that regulate membrane traffic." Neuron **17**(3): 379-88.

Sudhof, T. C., F. Lottspeich, P. Greengard, E. Mehl and R. Jahn (1987). "The cDNA and derived amino acid sequences for rat and human synaptophysin." Nucleic Acids Res **15**(22): 9607.

Sugahara, Y., P. Carninci, M. Itoh, K. Shibata, H. Konno, T. Endo, M. Muramatsu and Y. Hayashizaki (2001). "Comparative evaluation of 5'-end-sequence quality of clones in CAP trapper and other full-length-cDNA libraries." Gene **263**(1-2): 93-102.

Sumiyama, K., K. Washio-Watanabe, N. Saitou, T. Hayakawa and S. Ueda (1996). "Class III POU genes: generation of homopolymeric amino acid repeats under GC pressure in mammals." J Mol Evol **43**(3): 170-8.

Sureau, A., R. Gattoni, Y. Dooghe, J. Stevenin and J. Soret (2001). "SC35 autoregulates its expression by promoting splicing events that destabilize its mRNAs." Embo J **20**(7): 1785-96.

Sutherland, G. R., A. K. Gedeon, E. A. Haan, P. Woodroffe and J. C. Mulley (1988). "Linkage studies with the gene for an X-linked syndrome of mental retardation, microcephaly and spastic diplegia (MRX2)." Am J Med Genet **30**(1-2): 493-508.

Takahashi, K., K. Mitsui and S. Yamanaka (2003). "Role of ERas in promoting tumour-like properties in mouse embryonic stem cells." Nature **423**(6939): 541-5.

Taylor, J. S., I. Braasch, T. Frickey, A. Meyer and Y. Van de Peer (2003). "Genome duplication, a trait shared by 22000 species of ray-finned fish." Genome Res **13**(3): 382-90.

Tennyson, C. N., H. J. Klamut and R. G. Worton (1995). "The human dystrophin gene requires 16 hours to be transcribed and is cotranscriptionally spliced." Nat Genet **9**(2): 184-90.

Thanaraj, T. A. and F. Clark (2001). "Human GC-AG alternative intron isoforms with weak donor sites show enhanced consensus at acceptor exon positions." Nucleic Acids Res **29**(12): 2581-93.

Thanaraj, T. A. and S. Stamm (2003). "Prediction and statistical analysis of alternatively spliced exons." Prog Mol Subcell Biol **31**: 1-31.

Thanaraj, T. A., S. Stamm, F. Clark, J. J. Riethoven, V. Le Texier and J. Muilu (2004). "ASD: the Alternative Splicing Database." Nucleic Acids Res **32**(Database issue): D64-9.

The C. elegans sequencing consortium (1998). "Genome sequence of the nematode C. elegans: a platform for investigating biology." Science 282(5396): 2012-8.

The ENCODE project consortium (2004). "The ENCODE (ENCyclopedia Of DNA Elements) Project." Science **306**(5696): 636-40.

The International HapMap Project Consortium (2003). "The International HapMap Project." Nature **426**(6968): 789-96.

The International Human Genome Sequencing Consortium (2004). "Finishing the euchromatic sequence of the human genome." Nature **431**(7011): 931-45.

The International SNP Map Working Group (2003). "The International HapMap Project." Nature **426**(6968): 789-96.

The Yeast Sequencing Consortium (1997). "The yeast genome directory." Nature **387**(6632 Suppl): 5.

Thiselton, D. L., J. McDowall, O. Brandau, J. Ramser, F. d'Esposito, S. S. Bhattacharya, M. T. Ross, A. J. Hardcastle and A. Meindl (2002). "An integrated, functionally annotated gene map of the DXS8026-ELK1 interval on human Xp11.3-Xp11.23: potential hotspot for neurogenetic disorders." Genomics **79**(4): 560-72.

Thompson, J. D., D. G. Higgins and T. J. Gibson (1994). "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice." Nucleic Acids Res **22**(22): 4673-80.

Tinsley, J. H., S. Y. Yuan and E. Wilson (2004). "Isoform-specific knockout of endothelial myosin light chain kinase: closing the gap on inflammatory lung disease." Trends Pharmacol Sci **25**(2): 64-6.

Tian, B., J. Hu, H. Zhang, C.S. Lutz (2005). A large-scale analysis of mRNA polyadenylation of human and mouse genes.Nucleic Acids Res **12**(33):201-12.

Toder, R. and J. A. Graves (1998). "CSF2RA, ANT3, and STS are autosomal in marsupials: implications for the origin of the pseudoautosomal region of mammalian sex chromosomes." Mamm Genome **9**(5): 373-6.

Torrents, D., M. Suyama, E. Zdobnov and P. Bork (2003). "A genome-wide survey of human pseudogenes." Genome Res **13**(12): 2559-67.

Trofatter, J. A., M. M. MacCollin, J. L. Rutter, J. R. Murrell, M. P. Duyao, D. M. Parry, R. Eldridge, N. Kley, A. G. Menon, K. Pulaski and et al. (1993). "A novel moesin-, ezrin-, radixin-like gene is a candidate for the neurofibromatosis 2 tumor suppressor." Cell **75**(4): 826.

Tuschl, T. and A. Borkhardt (2002). Small interfering RNAs: a revolutionary tool for the analysis of gene function and gene therapy. Mol Interv **2**(3):158-67.

Ureta-Vidal, A., L. Ettwiller and E. Birney (2003). "Comparative genomics: genome-wide analysis in metazoan eukaryotes." Nat Rev Genet **4**(4): 251-62.

Van den Eynde, B., O. Peeters, O. De Backer, B. Gaugler, S. Lucas and T. Boon (1995). "A new family of genes coding for an antigen recognized by autologous cytolytic T lymphocytes on a human melanoma." J Exp Med **182**(3): 689-98.

Velculescu, V. E., L. Zhang, B. Vogelstein and K. W. Kinzler (1995). "Serial analysis of gene expression." Science **270**(5235): 484-7.

Venables, J. P. (2004). "Aberrant and alternative splicing in cancer." Cancer Res **64**(21): 7647-54.

Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R. R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y. H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigo, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y. H. Chiang, M.

Coyne, C. Dahlke, A. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh and X. Zhu (2001). "The sequence of the human genome." Science **291**(5507): 1304-51.

Verma, I. M., G. F. Temple, H. Fan and D. Baltimore (1972). "In vitro synthesis of DNA complementary to rabbit reticulocyte 10S RNA." Nat New Biol **235**(58): 163-7.

Vinogradov, A. E. (2003). "Isochores and tissue-specificity." Nucleic Acids Res **31**(17): 5212-20.

Wagner, E. and J. Lykke-Andersen (2002). "mRNA surveillance: the perfect persist." J Cell Sci **115**(Pt 15): 3033-8.

Wagner, E. G. and R. W. Simons (1994). "Antisense RNA control in bacteria, phages, and plasmids." Annu Rev Microbiol **48**: 713-42.

Wagner, R. W. (1994). "Gene inhibition using antisense oligodeoxynucleotides." Nature **372**(6504): 333-5.

Wahl, M. B., U. Heinzmann and K. Imai (2005). "LongSAGE analysis significantly improves genome annotation: identifications of novel genes and alternative transcripts in the mouse." Bioinformatics **21**(8): 1393-400.

Wang, J., V. M. Vock, S. Li, O. R. Olivas and M. F. Wilkinson (2002a). "A quality control pathway that down-regulates aberrant T-cell receptor (TCR) transcripts by a mechanism requiring UPF2 and translation." J Biol Chem **277**(21): 18489-93.

Wang, J., Y. F. Chang, J. I. Hamilton and M. F. Wilkinson (2002b). "Nonsense-associated altered splicing: a frame-dependent response distinct from nonsense-mediated decay." Mol Cell **10**(4): 951-7.

Wang, L., L. Duke, P. S. Zhang, R. B. Arlinghaus, W. F. Symmans, A. Sahin, R. Mendez and J. L. Dai (2003). "Alternative splicing disrupts a nuclear localization signal in spleen tyrosine kinase that is required for invasion suppression in breast cancer." Cancer Res **63**(15): 4724-30.

Wang, Q., J. A. Blackford, Jr., L. N. Song, Y. Huang, S. Cho and S. S. Simons, Jr. (2004). "Equilibrium interactions of corepressors and coactivators with agonist and antagonist complexes of glucocorticoid receptors." Mol Endocrinol **18**(6): 1376-95.

Waragai, M., C. H. Lammers, S. Takeuchi, I. Imafuku, Y. Udagawa, I. Kanazawa, M. Kawabata, M. M. Mouradian and H. Okazawa (1999). "PQBP-1, a novel polyglutamine tract-binding protein, inhibits transcription activation by Brn-2 and affects cell survival." Hum Mol Genet **8**(6): 977-87.

Waragai, M., E. Junn, M. Kajikawa, S. Takeuchi, I. Kanazawa, M. Shibata, M. M. Mouradian and H. Okazawa (2000). "PQBP-1/Npw38, a nuclear protein binding to the polyglutamine

tract, interacts with U5-15kD/dim1p via the carboxyl-terminal domain." Biochem Biophys Res Commun **273**(2): 592-5.

Watanabe, H., A. Fujiyama, M. Hattori, T. D. Taylor, A. Toyoda, Y. Kuroki, H. Noguchi, A. BenKahla, H. Lehrach, R. Sudbrak, M. Kube, S. Taenzer, P. Galgoczy, M. Platzer, M. Scharfe, G. Nordsiek, H. Blocker, I. Hellmann, P. Khaitovich, S. Paabo, R. Reinhardt, H. J. Zheng, X. L. Zhang, G. F. Zhu, B. F. Wang, G. Fu, S. X. Ren, G. P. Zhao, Z. Chen, Y. S. Lee, J. E. Cheong, S. H. Choi, K. M. Wu, T. T. Liu, K. J. Hsiao, S. F. Tsai, C. G. Kim, O. O. S, T. Kitano, Y. Kohara, N. Saitou, H. S. Park, S. Y. Wang, M. L. Yaspo and Y. Sakaki (2004). "DNA sequence and comparative analysis of chimpanzee chromosome 22." Nature **429**(6990): 382-8.

Waterston, R. H., K. Lindblad-Toh, E. Birney, J. Rogers, J. F. Abril, P. Agarwal, R. Agarwala, R. Ainscough, M. Alexandersson, P. An, S. E. Antonarakis, J. Attwood, R. Baertsch, J. Bailey, K. Barlow, S. Beck, E. Berry, B. Birren, T. Bloom, P. Bork, M. Botcherby, N. Bray, M. R. Brent, D. G. Brown, S. D. Brown, C. Bult, J. Burton, J. Butler, R. D. Campbell, P. Carninci, S. Cawley, F. Chiaromonte, A. T. Chinwalla, D. M. Church, M. Clamp, C. Clee, F. S. Collins, L. L. Cook, R. R. Copley, A. Coulson, O. Couronne, J. Cuff, V. Curwen, T. Cutts, M. Daly, R. David, J. Davies, K. D. Delehaunty, J. Deri, E. T. Dermitzakis, C. Dewey, N. J. Dickens, M. Diekhans, S. Dodge, I. Dubchak, D. M. Dunn, S. R. Eddy, L. Elnitski, R. D. Emes, P. Eswara, E. Eyras, A. Felsenfeld, G. A. Fewell, P. Flicek, K. Foley, W. N. Frankel, L. A. Fulton, R. S. Fulton, T. S. Furey, D. Gage, R. A. Gibbs, G. Glusman, S. Gnerre, N. Goldman, L. Goodstadt, D. Grafham, T. A. Graves, E. D. Green, S. Gregory, R. Guigo, M. Guyer, R. C. Hardison, D. Haussler, Y. Hayashizaki, L. W. Hillier, A. Hinrichs, W. Hlavina, T. Holzer, F. Hsu, A. Hua, T. Hubbard, A. Hunt, I. Jackson, D. B. Jaffe, L. S. Johnson, M. Jones, T. A. Jones, A. Joy, M. Kamal, E. K. Karlsson, D. Karolchik, A. Kasprzyk, J. Kawai, E. Keibler, C. Kells, W. J. Kent, A. Kirby, D. L. Kolbe, I. Korf, R. S. Kucherlapati, E. J. Kulbokas, D. Kulp, T. Landers, J. P. Leger, S. Leonard, I. Letunic, R. Levine, J. Li, M. Li, C. Lloyd, S. Lucas, B. Ma, D. R. Maglott, E. R. Mardis, L. Matthews, E. Mauceli, J. H. Mayer, M. McCarthy, W. R. McCombie, S. McLaren, K. McLay, J. D. McPherson, J. Meldrim, B. Meredith, J. P. Mesirov, W. Miller, T. L. Miner, E. Mongin, K. T. Montgomery, M. Morgan, R. Mott, J. C. Mullikin, D. M. Muzny, W. E. Nash, J. O. Nelson, M. N. Nhan, R. Nicol, Z. Ning, C. Nusbaum, M. J. O'Connor, Y. Okazaki, K. Oliver, E. Overton-Larty, L. Pachter, G. Parra, K. H. Pepin, J. Peterson, P. Pevzner, R. Plumb, C. S. Pohl, A. Poliakov, T. C. Ponce, C. P. Ponting, S. Potter, M. Quail, A. Reymond, B. A. Roe, K. M. Roskin, E. M. Rubin, A. G. Rust, R. Santos, V. Sapojnikov, B. Schultz, J. Schultz, M. S. Schwartz, S. Schwartz, C. Scott, S. Seaman, S. Searle, T. Sharpe, A. Sheridan, R. Shownkeen, S. Sims, J. B. Singer, G. Slater, A. Smit, D. R. Smith, B. Spencer, A. Stabenau, N. Stange-Thomann, C. Sugnet, M. Suyama, G. Tesler, J. Thompson, D. Torrents, E. Trevaskis, J. Tromp, C. Ucla, A. Ureta-Vidal, J. P. Vinson, A. C. Von Niederhausern, C. M. Wade, M. Wall, R. J. Weber, R. B. Weiss, M. C. Wendl, A. P. West, K. Wetterstrand, R. Wheeler, S. Whelan, J. Wierzbowski, D. Willey, S. Williams, R. K. Wilson, E. Winter, K. C. Worley, D. Wyman, S. Yang, S. P. Yang, E. M. Zdobnov, M. C. Zody and E. S. Lander (2002). "Initial sequencing and comparative analysis of the mouse genome." Nature **420**(6915): 520-62.

Wei, C. L., P. Ng, K. P. Chiu, C. H. Wong, C. C. Ang, L. Lipovich, E. T. Liu and Y.Ruan (2004). "5' Long serial analysis of gene expression (LongSAGE) and 3' LongSAGE for transcriptome characterization and genome annotation." Proc Natl Acad Sci U S A **101**(32): 11701-6.

Wheelan, S. J., D. M. Church and J. M. Ostell (2001). "Spidey: a tool for mRNA-to-genomic alignments." Genome Res **11**(11): 1952-7.

Wickens, M. P., G. N. Buell and R. T. Schimke (1978). "Synthesis of double-stranded DNA complementary to lysozyme, ovomucoid, and ovalbumin mRNAs. Optimization for full length second strand synthesis by Escherichia coli DNA polymerase I." J Biol Chem **253**(7): 2483-95.

Wieringa, B., E. Hofer and C. Weissmann (1984). "A minimal intron length but no specific internal sequence is required for splicing the large rabbit beta-globin intron." Cell **37**(3): 915-25.

Wilcox, S. A., J. M. Watson, J. A. Spencer and J. A. Graves (1996). "Comparative mapping identifies the fusion point of an ancient mammalian X-autosomal rearrangement." Genomics **35**(1): 66-70.

Wu, Q. and A. R. Krainer (1999). "AT-AC pre-mRNA splicing mechanisms and conservation of minor introns in voltage-gated ion channel genes." Mol **Cell Biol 19**(5): 3225-36.

Xie, H., W. Y. Zhu, A. Wasserman, V. Grebinskiy, A. Olson and L. Mintz (2002). "Computational analysis of alternative splicing using EST tissue information." Genomics **80**(3): 326-30.

Xing, Y. and C. J. Lee (2004). "Negative selection pressure against premature protein truncation is reduced by alternative splicing and diploidy." Trends Genet **20**(10): 472-5.

Xu, Q., B. Modrek and C. Lee (2002). "Genome-wide detection of tissue-specific alternative splicing in the human transcriptome." Nucleic Acids Res **30**(17): 3754-66.

Yan, S. D., J. Fu, C. Soto, X. Chen, H. Zhu, F. Al-Mohanna, K. Collison, A. Zhu, E. Stern, T. Saido, M. Tohyama, S. Ogawa, A. Roher and D. Stern (1997). "An intracellular protein that binds amyloid-beta peptide and mediates neurotoxicity in Alzheimer's disease." Nature **389**(6652): 689-95.

Yang, Y. Y., G. L. Yin and R. B. Darnell (1998). "The neuronal RNA-binding protein Nova-2 is implicated as the autoantigen targeted in POMA patients with dementia." Proc Natl Acad Sci U S A **95**(22): 13254-9.

Ye, D., M. Wei, M. McGuire, K. Huang, G. Kapadia, O. Herzberg, B. M. Martin and D. Dunaway-Mariano (2001). "Investigation of the catalytic site within the ATP-grasp domain of Clostridium symbiosum pyruvate phosphate dikinase." J Biol Chem **276**(40): 37630-9.

Yeakley, J. M., J. B. Fan, D. Doucet, L. Luo, E. Wickham, Z. Ye, M. S. Chee and X. D. Fu (2002). "Profiling alternative splicing on fiber-optic arrays." Nat Biotechnol **20**(4): 353-8.

Yeo, G. W., E. Van Nostrand, D. Holste, T. Poggio and C. B. Burge (2005). "Identification and analysis of alternative splicing events conserved in human and mouse." Proc Natl Acad Sci U S A **102**(8): 2850-5.

Yeo, G., D. Holste, G. Kreiman and C. B. Burge (2004). "Variation in alternative splicing across human tissues." Genome Biol **5**(10): R74.

Yong, J., T. J. Golembe, D. J. Battle, L. Pellizzoni and G. Dreyfuss (2004). "snRNAs contain specific SMN-binding domains that are essential for snRNP assembly." Mol Cell Biol **24**(7): 2747-56.

Zavolan, M. and T. B. Kepler (2001). "Statistical inference of sequence-dependent mutation rates." Curr Opin Genet Dev **11**(6): 612-5.

Zhang, L., H. H. Lu, W. Y. Chung, J. Yang and W. H. Li (2005). "Patterns of segmental duplication in the human genome." Mol Biol Evol **22**(1): 135-41.

Zhang, L., V. Pavlovic, C. R. Cantor and S. Kasif (2003). "Human-mouse gene identification by comparative evidence integration and evolutionary analysis." Genome Res **13**(6A): 1190-202.

Zhang, X. H. and L. A. Chasin (2004). "Computational definition of sequence motifs governing constitutive exon splicing." Genes Dev **18**(11): 1241-50.

Zhang, Y., D. A. Eberhard, G. D. Frantz, P. Dowd, T. D. Wu, Y. Zhou, C. Watanabe, S. M. Luoh, P. Polakis, K. J. Hillan, W. I. Wood and Z. Zhang (2004). "GEPIS-quantitative gene expression profiling in normal and cancer tissues." Bioinformatics **20**(15): 2390-8.

Zhang, Y., T. Lindblom, A. Chang, M. Sudol, A. E. Sluder and E. A. Golemis (2000). "Evidence that dim1 associates with proteins involved in pre-mRNA splicing, and delineation of residues essential for dim1 interactions with hnRNP F and Npw38/PQBP-1." Gene **257**(1): 33-43.

Zhang, Y. and M. A. Frohman (1997). "Using rapid amplification of cDNA ends (RACE) to obtain full-length cDNAs." Methods Mol Biol **69**: 61-87.

Zhang, Z. and M. Gerstein (2003). "Of mice and men: phylogenetic footprinting aids the discovery of regulatory elements." J Biol **2**(2): 11.

Zhang, Z., P. M. Harrison, Y. Liu and M. Gerstein (2003). "Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome." Genome Res **13**(12): 2541-58.

Zheng, Z. M., M. Huynen and C. C. Baker (1998). "A pyrimidine-rich exonic splicing suppressor binds multiple RNA splicing factors and inhibits spliceosome assembly." Proc Natl Acad Sci U S A **95**(24): 14088-93.

Zhu, J., A. Mayeda and A. R. Krainer (2001). "Exon identity established through differential antagonism between exonic splicing silencer-bound hnRNP A1 and enhancer-bound SR proteins." Mol Cell **8**(6): 1351-61.

Zhu, J., J. Shendure, R. D. Mitra and G. M. Church (2003). "Single molecule profiling of alternative pre-mRNA splicing." Science **301**(5634): 836-8.

Zinn, A. R., V. S. Tonk, Z. Chen, W. L. Flejter, H. A. Gardner, R. Guerra, H. Kushner, S. Schwartz, V. P. Sybert, D. L. Van Dyke and J. L. Ross (1998). "Evidence for a Turner syndrome locus or loci at Xp11.2-p22.1." Am J Hum Genet **63**(6): 1757-66.

Appendices

APPENDIX I Primers used to screen vectorette cDNA libraries for the expression of novel genes

GENE	stSG number	Sense (5' → 3')	Antisense (5' → 3')
FOXP3	stSG156841	CATCATCCGACAAGGGCTC	AATGTGCTGTTTCCATGGCT
UBE1	stSG156840	TCTGGAGTCACGAACAGGTG	CTTCTTCAGGGGAGGTTGTG
PHF16	stSG156758	CTTTCCTCTTTTGGCTGCAC	GGAGCTGGAGGAGGAGTTCT
N.A.	stSG156806	CATGACTCCAGAGCAGCAAG	GGGCTGAGCTCATAGTCTGG
bA637B23.1	stSG156802	GAGGATTGCCCAAAAATCA	GGTGCTTGTCATCTTCAGCA
RP11-1145B22.1	stSG388756	AGGATTGGTGAACATGGAGG	TCCTAGCACCCCTATCCACA
RP11-54B20.4	stSG156037	GGGCTAGTGACATTTGAAGATG	GACCAAGTTGCTGTAGGTCTCC
RP11-38O23.2	stSG187910	AAGTGATTCGAAGGGAGTGC	CTCTCGTGAATCTTCTCAGAGCTA
AC115617.1	stSG486758	CCTCCATCACGAGGAGATG	CATCTCCTCGTGATGGAGG
AC115617.1	stSG486759	TGATGACGGTGAAGAGCATC	GATGCTCTTCACCGTCATCA
AC115617.1	stSG486760	TTTGGTCCCAAACCTGTCTT	AAGACAGGTTTGGGACCAAA
AC115617.1	stSG529890	GGACATATGTCTGATCACAG	CTGTGATCAGACATATGTCC
RP11-348F1.1	stSG156807	AAGTGCAAACCCAACCAGAC	GACTCCCGCTTCTTCGTACA
RP11-348F1.1	stSG172512	AAACATTGATGTGAACCCAGTG	CAACAAAAGGCTATAGGTGCTG
AF196971.4	stSG156785	TGCCTGAGTACAAGGCTGT	TGTCCAGGGTCAACTCCTT
AF196971.4	stSG498780	TGCCTGAGTACAAGGCTGTG	AGCACACCATCACAGACAGC
RP11-54B20.3	stSG156035	TCTCTTCATATGCACACAGCTG	GCCTTCGATGATATTGCCAA
AF238380.5	stSG156846	GCTTTAGCCCCTTGGAGAAA	GATCCTATCTTCACCCACGC
RP11-1148L6.5	stSG486773	TCTGCTTTCCGTCTCGCTAT	TGTGGGAAGAACATGAAGAATG
RP11-805H4.2	stSG156812	ATCAACTCCAAAAGTGGCGA	CTTCAGGTCCACCCTTCAAT
AF196970.3	stSG380014	TGGGTAGAAGGAAACATGCC	AATTACACCACAGGCAGCG
AF196970.3	stSG380015	CGGACACTGACAGTAGAAGCC	AACTCTGGACACTCTGGCCT
AF196970.3	stSG172510	GCTTTGGCTACTGGAGATCG	CTGTGCCAGCCCTTGAGTA
AF235097.6	stSG380006	CACTTCTCCCTCCTGTGCTC	GCTAACACTTACGGGGGTGA
RP11-258C19.5	stSG156811	GAGGAAGAGGAAGACGCTCA	TTCTGCATAGAAAATGGCCC
RP11-576P23.4	stSG156810	CAGGCCTAGAGGCACTGCT	CACCTCCTCCTGCTTTCATC
RP1-30G7.2	stSG156783	ATGAGTGTTCATGTCGTGG	TGTCTTGTGAGGCTCTGGAA
RP5-1158E12.2	stSG156757	GCACATTCACACGGAACACT	CTGAGGGCTGATACTGCACA
bA104D21.3	stSG156804	ACCTCCAAAGCAGACCTTCA	ATACTGGGAATACTCGGCCA
RP1-71L6.2	stSG400506	GAGGTTTCAGGCTGTCGGCA	CGACTAAGTGGGGAAGCCGA
AF196972.9	stSG387162	AATGACCCACAGCACCTACC	ATGGAAACAATCAAGTCCGC
bA56H2.2	stSG156798	GGGAGACGTTCAGGAAACTG	CCGATAGGTGACGATGCTTT
RP11-1158E12.1	stSG156871	AAGAGGTGGCTGGGAAAGTT	CCCACAGTTTCCTGAACGTC
bA258C19.6	stSG156809	GCTGGAGCTGAACAACATCA	GTTCTGGTACTGGCTCTCGC

APPENDIX II Primers used to identify transcript variants

GENE	stSG number	Sense (5'→3')	Antisense (5'→3')
EBP			
A	487013	GCTCCGTAAGGCAAGAGAAC	GGCTCCGGGCTCTCTTTAG
B	487014	CCTATACACACGCAGCCATC	TCACGACTAAGACCCCTGTG
C	487015	GTTGTCCCATTGGGGACTT	GAACCAGCCCTCGATCACC
D	487016	GGTGGTGATCGCCTTTCTC	CCACAGAGACCACAAGCTGTA
E	487017	GCCAAGAGCAAGAAGAACTGA	CAAACTGTGGGAGCAGGACT
F	498791	TTGAGGGTCTTCATGCCAAT	GCAGCGGTCATTCTCTCTTC
OATL1			
A	487003	CTCCGGAGCCAGACAAGT	ACTTGTCTGGCTCCGGAG
B	487004	TGCTGTAGATCCCCAGATCA	AAAGGCTCGGATGAGGATGT
C	487005	TCGTCAGTGATGACCGTTTC	AATCAGCCCCAAAGATGTCA
D	487006	CAAGTGCTGAGCTGGTCGTA	TCGTGGTTCAGGTACGTGTG
E	487007	GCGCTGATGACCTCTTCTTC	GAAGAAGAGGTCATCAGCG
F	487008	TGGGTCTGGGGAGTTCAAT	TATTGAACTCCCCAGACCCA
G	487009	CCCAGGCCTAAGTATGTGGA	TCCACATACTTAGGCCTGGG
H	487010	AAGACAGGTGCTGTGAGCTG	CAGCTCACAGCACCTGTCT
I	487011	TTGAGCCCCCTAGACACAGA	TCTGTGTCTAGGGGGCTCAA
J	487012	GGATCTTGGCCTCTCTGTGA	TGCACGGTACCTGTCTGAAG
WDR13			
A	486663	CGGCGACAGTGGTCTCAGG	CCTGAGACCACTGTCGCCG
B	486664	TCGCGTCCACTGCTAAGACT	AGTCTTAGCAGTGGACGCGA
C	486665	GGCAGTACCTGAGGCTTCG	TGCTATAGGCACGAGCACTG
D	486666	GACACGTCACTGAGCGAGAA	GTAGCACGCGAAGCACTGT
E	486667	ATCCTCGTGTCCACCTCACT	GCAGAGCAGTTCAGCGCTAT
F	486668	ACCGTGGCAGTGTCTTCTCT	CTTGTTGAGGCAAGCATTGA
G	486669	AGAGAAGCTTCCCCATCGAG	GACATGAGGGGACAGAAGATG
H	486670	ACATGTGCGTGCACTTCTTT	AAAGAAGTGCACGCACATGT
I	486671	AAGTTTCGGTGGTCATGCTG	CAGCATGACCACCGAAACTT
RBM3			
A	486772	CCTCCGAGCTCGCTGTTC	GAACAGCGAGCTCGGAGG
B	486773	TCTGCTTTCCGTCTCGCTAT	TGTGGGAAGAACATGAAGAATG
C	486774	GTGGGAGGGCTCAACTTTAAC	CCTGCTCGTCGGTGTTAAAG
D	486775	TGTCAAGGACCGGGAGACT	CTCTCATGGCAACTGAAGCA
E	486776	CGGAGAGGACCTCGTGAGTT	TCCTCTAAGCTGCCCGACTA
F	486777	AAGTCTGCTCGGGGAACC	GTTCCCCGAGCAGACTTG
G	486778	GACAGTCGACCTGGAGGGTA	CTCTGGAACGTCCATATCCAT
H	486779	GGGTGGTTATGACCGCTACT	TGTGCATGTCTCATTTCAGTTG
I	486780	TATGGGACGTTTGTAGAACCTG	CAGGTTCTACAAACGTCCCATA
J	486781	CAAGAACATGATTATCCAGGGG	CCCCTGGATAATCATGTTCTTG
K	457763	GGGGAAGCGTCTTTGGGATTAGTG	TGGGGTACTGACTGGTCCACATTG
L	498793	TCAGTGGGAATATGCATACAGTT	TGGCATGAACTGCTTTTAACAAT

GENE	stSG number	Sense (5'→3')	Antisense (5'→3')
WAS			
A	486751	AGATGCTTGGACGAAAATGC	GCATTTTCGTCCAAGCATCT
B	486752	GGAACAGGAGCTGTACTCACA	TGTGAGTACAGCTCCTGTTCC
C	486753	ACTTTGCAGACGAGGACGAG	CTCGTCCTCGTCTGCAAAGT
D	486754	CGACAGTGGACATCCAGAAC	GTTCTGGATGTCCACTGTCG
E	486755	GGGGTAACAAGGGTCGTTCT	GCCGGATTTGATCCAAAAG
F	486756	GAGAAGCAGAGCCATCCACT	AGTGGATGGCTCTGCTTCTC
G	486757	CAACCCCTCCAATGCTGTTA	TAACAGCATTGGAGGGGTTG
H	486762	AGAACGACCCTTGTTACCCC	CTTTTGGATCAAATCCGGC
AC115617.1			
A	486758	CCTCCATCACGAGGAGATG	CATCTCCTCGTGATGGAGG
B	486759	TGATGACGGTGAAGAGCATC	GATGCTCTTCACCGTCATCA
C	486760	TTTGGTCCCAAACCTGTCTT	AAGACAGGTTTGGGACCAAA
GATA1			
A	486656	GAACCCTCCGCAACCACCAG	CTGGTGGTTGCGGAGGGTTC
B	486657	TTAATCCCCAGAGGCTCCAT	ATAGGCTGCTGAATTGAGGG
C	486658	CCACTACCTATGCAACGCCT	GGCAGTTGGTGCACTGAGTA
D	486659	GCTTTGAAGGTTCAAGCCA	TGGCTTGAACCTTCAAAGC
E	486660	GCTCAACTGTATGGAGGGGA	TCCCCTCCATACAGTTGAGC
F	486661	GCGGAAGGATGGTATTCAGA	TCTGAATACCATCCTTCCGC
G	511595	TGCTCCTGCAGTTACAATCG	TCAGATCATGTTTATTGTGGGG
SUV39H1			
A	487030	TGCTAAGCCAGGAAGCAGAAT	ATTCTGCTTCCTGGCTTAGCA
B	487031	CTGCTCGTGCTTACAGTGCT	ACAGCTCTTAGGGCCAACAC
C	487032	GAGCTCACCTTTGATTACAA	TTGTAATCAAAGGTGAGCTC
D	367547	ATTCGCAAGAACAGCTTCGT	TTGTAATCAAAGGTGAGCTC
E	367548	TAAGAAGCGGGTCCGTATT	AACCAGATTCAGTCCATGC
F	433850	CTAGGCCCGAATGTCGTTAG	CGGGAGTTGCACTCGTAGAT
G	498794	CTGGTGACCGTTACCCATTC	CGCACAGGTACTCGACTTCA
H	498806	CTTAGAGGTCTGCTGCTCCA	CTGGGACATCCCAAAGACTG
HDAC6			
A	486638	GGGGCTGCGTCCAATGAGTG	CACTCATTGGACGCAGCCCC
B	486639	GGCTGGTTGAAACGCTAGGG	CCCTAGCGTTTCAACCAGCC
C	486640	TGATTATCCCGGGAGATAGG	CCTATCTCCCGGGATAATCA
D	486641	ACCAGGCAGCGAAGAAGTAG	CCTTGCAGTCCCACGATTAG
E	486642	GCTGAAAAGGAAGAGCTGATG	CATCAGCTCTTCCTTTTCAGC
F	486643	TGGCTATTCCTTCAGGTTGAC	GTCAACCTGAAGGAATAGCCA
G	486644	GGGAGCCAGAAAGGGAGTAA	TTACTCCCTTTCTGGCTCCC
H	486645	GGCTGAGATCCGGAATGG	CCATTCCGGATCTCAGCC
I	486646	AGTTCACCTTCGACCAGGAC	TCCAAGGCACATTGATGGTA
J	486647	CAGCTAACCCACCTGCTCAT	AGAAGGGTGTGGAGCGAAG
K	486648	GATCATGTGCCGTCTGGAG	TGTGCACAGGCGAAGGTA
L	486649	CACAGGCTTCACCGTCAAC	TAAGGATAATGCGGCCACTG
M	486650	GGATGACCACACGAGAAAAGA	TCTTTTCTCGTGTGGTCATCC
N	486651	GACCTTGGAGCTAGGCAGC	GCTGCCTAGCTCCAAGGTC
O	486652	CCCCATTTGGTGGCAGTAT	GGCTGACAGGTCGATGTAGC
P	486653	CACCACTACTCCAGCCCAGAA	TTCTGGGCTGGAGTAGTGGTG

GENE	stSG number	Sense (5'→3')	Antisense (5'→3')
PCSK1N			
A	484537	GTCGGCCTTTTGGTGCTG	CAGCACCAAAAGGCCGAC
B	484538	CCGTGTGCACATTCCATAGT	TGCTCCTGCAGTTACAATCG
C	484539	CCTAAGCGCAGCGTCTCC	GGAGACGCTGCGCTTAGG
D	484540	GGACCCCGAGCTGTTGAG	GTCCACGTCGGGTGTCTC
E	484541	TACTTGCTGGGACGGATTCT	GGTAAGTTGCTCTGGACGTG
PIM2			
A	487044	ACCAGTTTCTCTGCTTTCCAC	GCCCACTGAACCCGCTAA
B	487045	AGTCACATGCCCACTCGAAG	CATAGCAGTGCGACTTCGAG
C	487046	AAGCTGCTTACCTGCCTCAG	GAAGCCTTGGGGTAACAACA
D	487047	TTCGAGGCCGAGTATCGACT	AGTCGATACTCGGCCTCGAA
E	487048	GGTACATCCTCGGCTGGTGT	CGACCCTCACTGGAAGAGAT
F	530013	CCGGACAAACAACTACCAGC	GCGACCTTCTGTCGAAGACT
G	499718	GAGCTTTCCGCCTATTACCG	CGGTAATAGGCGGAAAGCTC
TIMM17B			
B	483753	GGTTGAGAGGTAGTGCCAATG	CATTGGCACTACCTCTCAACC
C	483754	GGGAGGGAACCGAGAAGTAG	CTACTTCTCGGTTCCCTCCC
D	483755	CCACCCCAACGTTACCAATA	TATTGGTAACGTTGGGGTGG
H	485476	CAGCGCCATGGAGGAGTAC	GTACTCCTCCATGGCGCTG
I	485477	TAAAGCCCGTTTTGAAATGC	GCATTTCAAAACGGGCTTTA
J	485479	TAATTGGCAGACGGAAGACAG	CTGTCTTCCGTCTGCCAATTA
PQBP1			
A	483740	GAAGAGATCATTGCCGAGGA	TCCTCGGCAATGATCTCTTC
B	483741	GGAAGGGTCGAACACCTTGTA	TACAAGGTGTTCGACCCTTCC
C	483742	TAGACCCCATGGACCCTAGC	GCTAGGGTCCATGGGGTCTA
D	483743	ATGCTGAAGAAAAGTTGGACCG	CGGTCCAACTTTTCTTCAGCAT
E	483744	CTTCTTGCTCTTGGGATAGGG	CCCTATCCCAAGAGCAAGAAG
F	483745	GGGCAGGATCACCAGAAAG	CTTTCTGGTGATCCTGCCC
G	483746	CAATCCTGCTGCTTGGTTC	GAACCAAGCAGCAGGATTG
H	483747	AAGACTGGCGCTGACACC	GGTGTCAGCGCCAGTCTT
I	483748	CCTTCCACTACCTGCACGA	TCGTGCAGGTAGTGGAAGG
J	483749	GTGAAGGCCTCGTTGAGAGA	TCTCTCAACGAGGCCTTCAC
K	483750	TGGGAGTTGCGATGATATTG	CAATATCATCGCAACTCCCA
L	483751	CCAAGAGAGGCATCCTCAAA	TTTGAGGATGCCTCTCTTGG
M	499719	CCTTCACAACTCCTTGCCCCA	TGGGGCAAGGAGTTGTGAAGG
N	498795	TCTTCGCACTCTGTTTATTCG	GTAACGTTGGGGTGGGTGCAC
SLC35A2			
A	482243	TCTTCCTCCATGAGGCTGTC	GAGGTGGGTCATGAGAGAGC
C	482245	CTCATCACGGAGCCCTTTCT	AGAAAGGGCTCCGTGATGA
D	482244	GATCCTCAAAGGCAGCTCA	TGGCATTGGATATCCTGACA
G	482277	AGCGTGTCCACATACTGCAC	TTTCCGCGGGTGCATTGGAG
H	482278	CTCCAATGCACCCGCGGAAA	AATCCCGGGGCTATTCATAC
J	486761	TCACGTGCAGTGCCTAGAAC	ATTCAGCATGAGCACGGAG

GENE	stSG number	Sense (5'→3')	Antisense (5'→3')
AF207550.5			
A	487022	CTACGGCACCACCCACAC	CTCAAGGCCCGCTACCAG
B	487023	GCAACCACATTGAGATGCAG	ACCTCCACAGGACGGTTGTA
C	487024	ACGAACCCATTCGTGTTAGC	TGGCCTTGTTAGGATTCACC
D	487025	GACCGGCTAGTCCACTCCTC	CGGGATCAGGAGAAACAGG
E	487026	TGATGATGAGATCCTAGCTTCG	CCTGGGTCTCCATCAACTCT
F	487027	CTCAGACGTTCTTGGGGGTA	GGGTGCCGAAGATAGACAAG
G	487028	TGCTCCCACACTCCCTACAC	GGGGTCATCTTTGGGGATTA
H	487029	AGGTGTATGGAGACCAGGACA	TAGTCCATGCAATGCTTTCG
I	498796	GGGCCCTGTAACAGTGAGAA	GGTTTGTTGGAGGGGAGAAT
KCND1			
A	487033	GACACACCCAACTCCCTTTC	ATCTAGACGGGCCAACAAAG
B	487034	TGTCTAGGGCAGATGCTGTG	CAACTGTGAGACATGGGCAG
C	487035	CAGGCTCTTTGTGTCAGGAAC	GAGGTGCTACGGCTGGTG
D	487036	TGGGTGACCTTCCAAAACTC	TCTTTGTCAGCCACCAGCTC
E	487037	TCCCTTTCCAGATCCATGAG	GAAGTGGGGGTTCTAGAGGG
F	487038	ATTCTCCTAAACGCCACCCT	ATCTCCTCGAGATGCCTTCA
G	487039	GCTGCCTTGAAGAGTATCGG	GATGAAGAAGCCGGTCACAT
H	487040	GTTCTGATTGGTTCTGCGGT	GACCCCTTTTCATTCCCCTA
I	487041	AGGGCACAAACAAGACCAAC	GAAAATCTTGCCAGCAATGG
J	487042	GACTTCGTGGCTGCCATTAT	GATCTTGACAGTCTCGGGGA
K	487043	CTTGGGCTCTCAGATGAAGG	CTCCAAAGCTCCATATCCCA
GRIPAP1			
A	530014	ACAGCCAGGAGGAGGACTTC	GAAGTCCTCCTCCTGGCTGT
B	530015	AGCTGATCACAAAGCCCAGT	CTGATTGGCAGCCTCAAGTT
C	530016	ACAGGATGTACGGGATCAGC	GCTTCAGGCTGCTAATGGTC
D	530017	CTTATCCAGCAGCCCTCAAG	GCTGAAAGAGCTCAGCCACT
E	530018	GCAGCGTCCTGAGAGACCTA	TCTTGTTCATCTCCCGAAGG
F	487049	CAGCAAAGACCCAAGAACTG	CTCGTACTCAGCCTGCTGCT
M13 F & R		GTAAAACGACGGCCAGTG	GGAAACAGCTATGACCATG

APPENDIX III Primer pairs used to screen human cDNA samples for novel transcript variants

EBP

Number	Primer 1		Primer 2		Product Size				
1	A	487013S	B'	487014A	328	273			
2	B	487014S	C'	487015A	282				
3	A	487013S	C'	487015A	444	389	220		
4	C	487015S	E'	487017A	613				
5	F	498791S	B'	487014A	287				

OATL1

Number	Primer 1		Primer 2		Product Size	
1	A	487003A	B'	487004A	257	209
2	B	487004S	C'	487005A	317	
3	C	487005S	E'	487007S	802	
4	D	487006S	G'	487009A	1671	
5	D	487006S	I'	487001S	3019	

RBM3

Number	Primer 1		Primer 2		Product size				
1	A	486772S	C'	486774A	566	237			
2	B	486773S	C'	486774A	228				
3	D	486775S	E'	486776A	480	329	219		
4	D	486775S	F'	486777A	646	425	318	156	
5	C	486774S	F'	486777A	655	508	401	239	132
6	C	486774S	E'	486776A	840	562	416	309	
7	K'	457763A	H'	486779A	323				
8	F	486777S	H'	486779A	233				
9	H	486779S	I'	486780A	654	299			
10	H	486779S	J'	486781A	1096	733			
11	G	486778S	J'	486781A	1155	800			
12	L	498796A	K	498794S	1104	1000			

WDR13

Number	Primer 1		Primer 2		Product Size		
1	A	486663S	B'	486664S	482	139	
2	B	486664A	C'	486665A	897	218	185
3	C	486665S	D'	486666A	486		
4	D	486666S	F'	486668A	693		
5	F	486668S	H'	486670A	339		
6	F	486668S	I'	486671A	674		

WAS

Number	Primer 1		Primer 2		Product Size		
1	A	486751S	C'	486753A	290		
2	B	486752S	D'	486754A	352	220	
3	C	486753S	D'	486754A	233	101	
4	D	486754S	E'	486755A	725	656	
5	E	486755S	E'	486755A	320	251	
6	H'	486762A	G'	486757A	350		

SUV39H1

Number	Primer 1		Primer 2		Product Size		
1	F	433850S	F'	433850A	706		
2	D	367547S	D'	367547A	400	308	134
3	A	487030S	D'	367547A	248	74	
4	C	487032S	E'	367548A	749		

AC115617.1

Number	Primer 1		Primer 2		Product Size
1	A'	486758A	C	486760S	200
2	B	486759S	D'	529890A	150
3	C	4869760S	D'	529890A	230

GATA1

Number	Primer 1		Primer 2		Product size		
1	A	486656S	B'	486657A	599		
2	B	486657S	C'	486658A	802		
3	C	486658S	D'	486659A	695	475	354

HDAC6

Number	Primer 1		Primer 2		Product size		
1	A	486638S	D'	486641A	311		
2	B	486639S	D'	486641A	290	559	
3	C	486640S	D'	486641A	280		
4	B	486639S	C'	486640A	160		
5	D	486641S	E'	486642A	396		
6	E	486642S	H'	486645A	1052	858	162
7	H	486645S	I'	486646A	260		
8	I	486646S	J'	486647A	427		
9	J	486647S	K'	486648A	647		
10	K	486648S	L'	486649A	803		
11	L	486649S	N'	486651A	1093		
12	M	486650S	O'	486652A	947		
13	O	486652S	P'	486653A	547		
14	L	486649S	P'	486653A	1799	253	

PCSK1N

Number	Primer 1		Primer 2		Product Size
1	A	484537S	C'	484539A	101
2	B	484538S	C'	484539A	256
3	C	484539S	E'	484541A	754 451
4	D	484540S	E'	484541A	310
5	A	484537S	D'	484540A	545

TIMM17B

Number	Primer 1		Primer 2		Product Size
1	D'	483755A	H'	485476A	115
2	G'	483757A	H'	485476A	93
3	H	485476S	B	483753S	167 317 788
4	B'	483753A	C'	483754A	423 509
5	B'	483753A	I	485477S	548 745

PQBP1

Number	Primer 1		Primer 2		Product Size
1	J	483749S	I'	476748A	144
2	J	486749S	L'	483751A	203 135
3	I	483748S	L'	483751A	99
4	K	483750S	L'	483751A	260
5	L	483751S	B	483741S	143
6	A	483740S	D	483743A	239
7	A	483740S	C'	483742A	251 285 546 763
8	A	483740S	E	483744S	501
9	C	483742S	G	483746S	198 328
10	C	483742S	F	483745S	216
11	D	483743S	E	483744S	284
12	D	483743S	C'	483742A	329 543
13	D	483743S	H'	483747A	329 621 753
14	B'	483741A	C'	483742A	170 204 465 682

SLC35A2

Number	Primer 1		Primer 2		Product Size
1	A	482243S	C'	482245A	272 731 1600 620
2	D	482244S	C'	482245A	492
3	C	482245S	F	482244A	1135
4	D	482244S	F	482244A	1707 560
5	D	482244S	B	482243A	732
6	C	482245S	B	482243A	260
7	H	482277S	G	482277A	282
8	I	482278S	G	482277A	542

PIM2

Number	Primer 1		Primer 2		Product Size
1	A	487044S	B'	487045A	336
2	B	487045S	C'	487046A	1262

AF207550.5

Number	Primer 1		Primer 2		Product Size
1	A	487022A	B'	487023A	694
2	B	487023S	C'	487024A	205 180
3	C	487024S	D'	487025S	391
4	D	487025A	E'	487026A	494
5	E	487026S	F'	487027S	596 370
6	E	487026S	G'	487028S	977 851

KCND1

Number	Primer 1		Primer 2		Product Size
1	A	487033S	I'	487041A	2154
2	I	4870141S	C'	487035A	507
3	C	487035S	D'	487036A	1715
4	B	487034S	I'	487041A	1188

GRIPAP1

Number	Primer 1		Primer 2		Product Size
1	A	530013S	B'	530014A	248
2	B	530014S	C'	530015A	622
3	C	530015S	G'	487049A	465
4	G	487049S	D'	530016A	708 630
5	D	530016S	E'	530017A	445
6	E	530017S	F'	530018A	316

APPENDIX IV Transcript variants identified for 18 genes in human Xp11.23
(Blue – reference transcript, pink transcripts identified in this study)

Variant	Sequence	Identified in screen	Description	Size change (+/- bp)	Location	Coding
EBP.1	NM_006579		n.a.	n.a.	n.a.	n.a.
EBP.2	AL583368		3' gain	-54	exon 2	5' UTR
EBP.3	BE300348		3' gain	-223	exon 2	5' UTR
EBP.4	BI224033		novel first exon	108	exon 1a	5' UTR
EBP.5	AJ973496	*EBP*- 2	internal deletion	-81	exon 2	5' UTR
EBP.6	AJ973494	*EBP*- 5	first exon extension	408	exon 1	5' UTR
EBP.7	AJ973495 AJ973491	*EBP*- 6	first exon extension	124	exon 1	5' UTR
EBP.8	AJ973492 AJ973493	*EBP*- 6	first exon extension	323	exon 1	5' UTR
OATL1.1	NM_001006113		n.a.	n.a.	n.a.	n.a.
OATL1.2	BG480179		3' gain	48	exon 2	CDS
OATL1.3	AJ973528 AJ973590	OATL1-6	5' loss 3' loss	-57	exon 1,2	CDS
OATL1.4	AJ973529 AJ973530	OATL1-6	whole exon addition	30	exon 4a	CDS
OATL1.5	AJ973532	OATL1-2	5' gain	72	exon 2	CDS
OATL1.6	AJ973532	OATL1-5	internal deletion	768	exon 6	3' UTR
OATL1.7	AJ973532	OATL1-1	5' gain	12	exon 1	CDS
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

Variant	Sequence	Identified in screen	Description	Size change (+/- bp)	Location	Coding
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 AJ973565	RBM3-12	internal deletion 5' loss final exon extension	-616	exon 8d	3' UTR
WDR13.1	AF329819		n.a.	n.a.	n.a.	n.a.
WDR13.2	AF158978		first exon extension	1149	exon 1	5' UTR
WDR13.3	BM791753		internal deletion 3' loss	-375	exon 1	5' UTR
WDR13.4	AL552544		internal deletion	-363	exon 1	5' UTR
WDR13.5	AL544291		3' loss	-34	exon 1	5' UTR
WDR13.6	BM193349		internal deletion first exon extension	256, -343	exon 1	5' UTR
WDR13.7	BM921240		whole exon deletion	-241	exon 2	CDS
WDR13.8	BE253921		novel first exon 3' loss	52	exon 1	5' UTR
WDR13.9	BU158146		3' gain	85	exon 4	CDS
WDR13.10	AU408073		whole exon deletion	181	exon 6	CDS
WDR13.11	AJ973583	WDR13-3	3' loss	-56	exon 6	CDS

Variant	Sequence	Identified in screen	Description	Size change (+/- bp)	Location	Coding
WDR13.12	AJ973579	WDR13-4	whole exon deletion 3' loss	-423	exon 5, 6	CDS
WDR13.13	AJ973577 AJ973578	WDR13-2	antisense	80,98	n.a.	n.a.
WDR13.14	AJ973581	WDR13-5	antisense	18585	n.a.	n.a.
WAS.1	NM_000377		n.a.	n.a.	n.a.	n.a.
WAS.2	BI833034		whole exon deletion	-132	exons 5,6	CDS
WAS.3	AJ973571	WAS-4	intron retention	1099	intron 7	CDS
WAS.4	AJ973572	WAS-4	internal deletion	90	exon 9	CDS
ERAS	NM_181532		n.a.	n.a.	n.a.	n.a.
GATA1.1	NM_002049		n.a.	n.a.	n.a.	n.a.
GATA1.2	BC009797		3' loss	-341	exon 6	3' UTR
GATA1.3	AI022182		internal deletion	-217	exon6	3' UTR
GATA1.4	Sccd21402	GATA1-1	whole exon deletion	-239	exon2	CDS
GATA1.5	AJ973499	GATA1-2	5' loss	-137	exon 3	CDS
SUV39H1.1	NM_003173		n.a.	n.a.	n.a.	n.a.
SUV39H1.2	AL549893		novel first exon	93	exon 2a	5' UTR
SUV39H1.3	BG285236		5' loss, 3' loss	-174	exon 3,4	CDS
SUV39H1.4	AJ973548	SUV39H1.1	whole exon deletion	-147	exon 2	5' UTR
SUV39H1.5	AJ973549	SUV39H1.5	first exon extension	352	exon 2	5' UTR
AC115617.1.	t.b.c.		n.a.	n.a.	n.a.	n.a.
AC115617.1.2	AJ973481	AC115617.1-1	whole exon addition	69	exon 1a	CDS
AC115617.1.3	AJ973482	AC115617.1-1	whole exon addition 5' loss	118, 47	exon 1b	CDS
AC115617.1.4	AJ973484	AC115617.1-1	5' loss	47	exon 3	CDS
HDAC6.1	AK024083		n.a.	n.a.	n.a.	n.a.
HDAC6.2	AL137696		intron retention	696	intron 6	CDS
HDAC6.3	BC005872		novel first exon	59	exon 1a	5'
HDAC6.4	BC011498		novel first exon final exon extension	83, 1313	exon 6	CDS

Variant	Sequence	Identified in screen	Description	Size change (+/- bp)	Location	Coding
HDAC6.5	AJ973509	HDAC6-1	novel first exon	24	exon 1b	CDS
HDAC6.6	AJ973527	HDAC6.10	whole exon deletion 3' loss	925	exons 22-24	CDS
HDAC6.7	AJ973511	HDAC6.14	whole exon deletion	150	exon 22	CDS
HDAC6.8	AJ973513	HDAC6.10	internal deletion	229	exon 25	CDS
HDAC6.9	AJ973515	HDAC6.10	whole exon deletion 3' loss	486	exons 23-24	CDS
HDAC6.10	AJ973516	HDAC6.10	whole exon deletion 3' loss	374	exons 23-24	CDS
HDAC6.11	AJ973526	HDAC6.13	whole exon deletion	123	exon 2	CDS
HDAC6.12	AJ973517 AJ973518	HDAC6.14	5' loss 3' loss	633	exons 23-24	CDS
HDAC6.13	AJ973519 AJ973520	HDAC6.14	whole exon deletion 3' loss	24	exons 22-23	CDS
HDAC6.14	AJ973521	HDAC6.10	3' loss	396	exon 25	CDS
HDAC6.15	AJ973523	HDAC6.14	whole exon deletion whole exon deletion 5' loss 3' loss	837	exons 22-25	CDS
HDAC6.16	AJ973524 AJ973525	HDAC6.10	whole exon deletion 5' loss 3' loss	797	exons 23-25	CDS
HDAC6.17	AJ973507 AJ973508	HDAC6.10	whole exon deletion 5' loss 3' loss	818	exons 23-25	CDS
TIMM17B.1	NM__005854		n.a.	n.a.	n.a.	n.a.
TIMM17B.2	AL529917		whole exon addition	103	exon 4a	CDS
TIMM17B.3	BF530129		whole exon addition 3' gain	150, 188	exon 4b ,exon 1	CDS
TIMM17B.4	BC028017		whole exon deletion intron retention	621, 87	exon 4c, intron 5	CDS

Variant	Sequence	Identified in screen	Description	Size change (+/- bp)	Location	Coding
TIMM17B.5	AJ973566 AJ973567	TIMM17B-1	intron retention	246	intron 2	5' UTR
TIMM17B.6	AJ973568 AJ973569	TIMM17B-5	internal deletion	-170	exon 6	3' UTR
TIMM17B.7	AJ973570	TIMM17B-4	antisense	69, 71	n.a.	Anti
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
SLC35A2.1	D88146		n.a.	n.a.	n.a.	
SLC35A2.2	BI820134		final exon extension	1871	exon4	3' UTR
SLC35A2.3	D84454		3' loss	-590	exon 4	3' UTR
SLC35A2.4	BE902730		3' loss final exon	165	exon 4	3' UTR
PIM2.1	NM_002648		n.a.	n.a.	n.a.	n.a.
PCSK1N.1	NM_013271		n.a.	n.a.	n.a.	n.a.
PSCK1N.2	AW163271		novel first exon	297	5' UTR	5' UTR
AF207550.5.1	NM_017602		n.a.	n.a.	n.a.	n.a.
AF207550.5.2	AL137509		5' loss	-15	exon 4	CDS

Variant	Sequence	Identified in screen	Description	Size change (+/- bp)	Location	Coding
AF207550.5.3	AK026260		novel final exon	-414	exon9	3' UTR
AF207550.5.4	BC011738		internal deletion	-211	exon 9	3' UTR
AF207550.5.5	AJ973485	AF207550.5-2	3' loss	-60	exon 4	CDS
AF207550.5.6	AJ973487	AF207550.5-3	intron retention	84	intron 4	CDS
AF207550.5.7	AJ973489 AJ973490	AF207550.5-3	5' gain 3' loss	-72	exons4,5	CDS
KCND1.1	NM_004979		n.a.	n.a.	n.a.	n.a.
KCND1.2	AJ005898		novel first exon	1278	5' UTR	5' UTR
GRIPAP1.1	NM_020137		n.a.	n.a.	n.a.	n.a.
GRIPAP1.2	AJ973501	GRIPAP1-3	3' loss	-97	exon 7	CDS
GRIPAP1.3	AJ973502	GRIPAP1-3	whole exon addition	75	exon 13a	CDS
GRIPAP1.4	AJ973504	GRIPAP1-2	intron retention	1413	intron 11	CDS
GRIPAP1.5	AJ973505	GRIPAP1-2	whole exon deletion	-151	exon 6	CDS

APPENDIX V Multiple sequence alignment of *PQBP1* transcripts

Alternating exon sequences are displayed in blue and black. Retained intron sequences are shown in aqua, deletions located within exons are displayed in red. The start and stop codons are also highlighted in green.

```
                    *        20         *        40         *        60         *        80         *       100         *       120
Reference     : TCTCTGCTTCAGCTATGCCGCTGCCCGTTGCGCTGCAGACCCGCTTGGCCAAGAGAGGCATCCTCAAACATCTGGAGCCTGAACCAGAGGAAGAGATCATTGCCGAGGACTATGACGATG :  120
Transcript 1  : TCTCTGCTTCAGCTATGCCGCTGCCCGTTGCGCTGCAGACCCGCTTGGCCAAGAGAGGCATCCTCAAACATCTGGAGCCTGAACCAGAGGAAGAGATCATTGCCGAGGACTATGACGATG :  120
Transcript 2  : TCTCTGCTTCAGCTATGCCGCTGCCCGTTGCGCTGCAGACCCGCTTGGCCAAGAGAGGCATCCTCAAACATCTGGAGCCTGAACCAGAGGAAGAGATCATTGCCGAGGACTATGACGATG :  120
Transcript 3  : TCTCTGCTTCAGCTATGCCGCTGCCCGTTGCGCTGCTGACCCGCTTGGCCAAGAGAGGCATCCTCAAACATCTGGAGCCTGAACCAGAGGAAGAGATCATTGCCGAGGACTATGACGATG :  120
Transcript 4  : TCTCTGCTTCAGCTATGCCGCTGCCCGTTGCGCTGCAGACCCGCTTGGCCAAGAGAGGCATCCTCAAACATCTGGAGCCTGAACCAGAGGAAGAGATCATTGCCGAGGACTATGACGATG :  120
Transcript 5  : TCTCTGCTTCAGCTATGCCGCTGCCCGTTGCGCTGCAGACCCGCTTGGCCAAGAGAGGCATCCTCAAACATCTGGAGCCTGAACCAGAGGAAGAGATCATTGCCGAGGACTATGACGATG :  120
Transcript 6  : TCTCTGCTTCAGCTATGCCGCTGCCCGTTGCGCTGCAGACCCGCTTGGCCAAGAGAGGCATCCTCAAACATCTGGAGCCTGAACCAGAGGAAGAGATCATTGCCGAGGACTATGACGATG :  120
Transcript 7  : TCTCTGCTTCAGCTATGCCGCTGCCCGTTGCGCTGCAGACCCGCTTGGCCAAGAGAGGCATCCTCAAACATCTGGAGCCTGAACCAGAGGAAGAGATCATTGCCGAGGACTATGACGATG :  120
Transcript 8  : TCTCTGCTTCAGCTATGCCGCTGCCCGTTGCGCTGCAGACCCGCTTGGCCAAGAGAGGCATCCTCAAACATCTGGAGCCTGAACCAGAGGAAGAGATCATTGCCGAGGACTATGACGATG :  120
Transcript 9  : TCCCTGCTTCAGCTATGCCGCTGCCCGTTGCGCTGCAGACCCGCTTGGCCAAGAGAGGCATCCTCAAACATCTGGAGCCTGAACCAGAGGAAGAGATCATTGCCGAGGACTATGACGATG :  120
Transcript 10 : TCTCTGCTTCAGCTATGCCGCTGCCCGTTGCGCTGCAGACCCGCTTGGCCAAGAGAGGCATCCTCAAACATCTGGAGCCTGAACCAGAGGAAGAGATCATTGCCGAGGACTATGACGATG :  120
Transcript 11 : TCTCTGCTTCAGCTATGCCGCTGCCCGTTGCGCTGCAGACCCGCTTGGCCAAGAGAGGCATCCTCAAACATCTGGAGCCTGAACCAGAGGAAGAGATCATTGCCGAGGACTATGACGATG :  120
Transcript 12 : TCTCTGCTTCAGCTATGCCGCTGCCCGTTGCGCTGCAGACCCGCTTGGCCAAGAGAGGCATCCTCAAACATCTGGAGCCTGAACCAGAGGAAGAGATCATTGCCGAGGACTATGACGATG :  120
Transcript 13 : TCTCTGCTTCAGCTATGCCGCTGCCCGTTGCGCTGCAGACCCGCTTGGCCAAGAGAGGCATCCTCAAACATCTGGAGCCTGAACCAGAGGAAGAGATCATTGCCGAGGACTATGACGATG :  120
Transcript 14 : TCTCTGCTTCAGCTATGCCGCTGCCCGTTGCGCTGCAGACCCGCTTGGCCAAGAGAGGCATCCTCAAACATCTGGAGCCTGAACCAGAGGAAGAGATCATTGCCGAGGACTATGACGATG :  120
Transcript 15 : TCTCTGCTTCAGCTATGCCGCTGCCCGTTGCGCTGCAGACCCGCTTGGCCAAGAGAGGCATCCTCAAACATCTGGAGCCTGAACCAGAGGAAGAGATCATTGCCGAGGACTATGACGATG :  120

                    *       140         *       160         *       180         *       200         *       220         *       240
Reference     : ATCCTGTGGACTACGAGGCCACCAGGTTGGAGGGCCTACCACCAAGCTGGTACAAGGTGTTCGACCCTTCCTG---------------------------------------------- :  193
Transcript 1  : ATCCTGTGGACTACGAGGCCACCAGGTTGGAGGGCCTACCACCAAGCTGGTACAAGGTGTTCGACCCTTCCTG---------------------------------------------- :  193
Transcript 2  : ATCCTGTGGACTACGAGGCCACCAGGTTGGAGGGCCTACCACCAAGCTGGTACAAGGTGTTCGACCCTTCCTG---------------------------------------------- :  193
Transcript 3  : ATCCTGTGGACTACGAGGCCACCAGGTTGGAGGGCCTACCACCAAGCTGGTACAAGGTGTTCGACCCTTCCTG---------------------------------------------- :  193
Transcript 4  : ATCCTGTGGACTACGAGGCCACCAGGTTGGAGGGCCTACCACCAAGCTGGTACAAGGTGTTCGACCCTTCCTG---------------------------------------------- :  193
Transcript 5  : ATCCTGTGGACTACGAGGCCACCAGGTTGGAGGGCCTACCACCAAGCTGGTACAAGGTGTTCGACCCTTCCTG---------------------------------------------- :  193
Transcript 6  : ATCCTGTGGACTACGAGGCCACCAGGTTGGAGGGCCTACCACCAAGCTGGTACAAGGTGTTCGACCCTTCCTG---------------------------------------------- :  193
Transcript 7  : ATCCTGTGGACTACGAGGCCACCAGGTTGGAGGGCCTACCACCAAGCTGGTACAAGGTGTTCGACCCTTCCTG---------------------------------------------- :  193
Transcript 8  : ATCCTGTGGACTACGAGGCCACCAGGTTGGAGGGCCTACCACCAAGCTGGTACAAGGTGTTCGACCCTTCCTG---------------------------------------------- :  193
Transcript 9  : ATCCTGTGGACTACGAGGCCACCAGGTTGGAGGGCCTACCACCAAGCTGGTACAAGGTGTTCGACCCTTCCTG---------------------------------------------- :  193
Transcript 10 : ATCCTGTGGACTACGAGGCCACCAGGTTGGAGGGCCTACCACCAAGCTGGTACAAGGTGTTCGACCCTTCCTGGTGAGCCTGGGTGAGGGGGAGCTAACTTCTGGCTTCACCCTTCCTGT :  240
Transcript 11 : ATCCTGTGGACTACGAGGCCACCAGGTTGGAGGGCCTACCACCAAGCTGGTACAAGGTGTTCGACCCTTCCTG---------------------------------------------- :  193
Transcript 12 : ATCCTGTGGACTACGAGGCCACCAGGTTGGAGGGCCTACCACCAAGCTGGTACAAGGTGTTCGACCCTTCCTG---------------------------------------------- :  193
Transcript 13 : ATCCTGTGGACTACGAGGCCACCAGGTTGGAGGGCCTACCACCAAGCTGGTACAAGGTGTTCAACCCTTCCTG---------------------------------------------- :  193
Transcript 14 : ATCCTGTGGACTACGAGGCCACCAGGTTGGAGGGCCTACCACCAAGCTGGTACAAGGTGTTCGACCCTTCCTG---------------------------------------------- :  193
Transcript 15 : ATCCTGTGGACTACGAGGCCACCAGGTTGGAGGGCCTACCACCAAGCTGGTACAAG----------------------------------------------------------- :  176
```

```
                    *       260         *       280         *       300         *       320         *       340         *       360
Reference     : ------------------------------------------------------------------------------------------------------------------------ :    -
Transcript 1  : ------------------------------------------------------------------------------------------------------------------------ :    -
Transcript 2  : ------------------------------------------------------------------------------------------------------------------------ :    -
Transcript 3  : ------------------------------------------------------------------------------------------------------------------------ :    -
Transcript 4  : ------------------------------------------------------------------------------------------------------------------------ :    -
Transcript 5  : ------------------------------------------------------------------------------------------------------------------------ :    -
Transcript 6  : ------------------------------------------------------------------------------------------------------------------------ :    -
Transcript 7  : ------------------------------------------------------------------------------------------------------------------------ :    -
Transcript 8  : ------------------------------------------------------------------------------------------------------------------------ :    -
Transcript 9  : ------------------------------------------------------------------------------------------------------------------------ :    -
Transcript 10 : GCTGACCTTGGTTGTGAGGTTTAGGGGGACACAAGGGAGGGAGCCTCGGGTGGAGGGTGTTGGCATTAGGTATCTGCAGGACTCAAGTTGCTGCCTGCTGGGGCCTGGCTCCTCTGGGGT :  360
Transcript 11 : ------------------------------------------------------------------------------------------------------------------------ :    -
Transcript 12 : ------------------------------------------------------------------------------------------------------------------------ :    -
Transcript 13 : ------------------------------------------------------------------------------------------------------------------------ :    -
Transcript 14 : ------------------------------------------------------------------------------------------------------------------------ :    -
Transcript 15 : ------------------------------------------------------------------------------------------------------------------------ :    -

                    *       380         *       400         *       420         *       440         *       460         *       480
Reference     : ------------------------------------------------------------------------------------------------------------------------ :    -
Transcript 1  : ------------------------------------------------------------------------------------------------------------------------ :    -
Transcript 2  : ------------------------------------------------------------------------------------------------------------------------ :    -
Transcript 3  : ------------------------------------------------------------------------------------------------------------------------ :    -
Transcript 4  : ------------------------------------------------------------------------------------------------------------------------ :    -
Transcript 5  : ------------------------------------------------------------------------------------------------------------------------ :    -
Transcript 6  : ------------------------------------------------------------------------------------------------------------------------ :    -
Transcript 7  : ------------------------------------------------------------------------------------------------------------------------ :    -
Transcript 8  : ------------------------------------------------------------------------------------------------------------------------ :    -
Transcript 9  : ----------------------------------------------------CCCGGGCTGGAGTGCAATGGTGTGATCTCGGCTTATGGCAACCTCCGCCTCGTGGGTTCAAGCGATTCTC :  263
Transcript 10 : TGGAAGACTGTCTTTTCTCTCTTTTTTGAAACGGAGTTTCACTCTTGTAGCCCGGGCTGGAGTGCAATGGTGTGATCTCGGCTTATGGCAACCTCCGCCTCGTGGGTTCAAGCGATTCTC :  480
Transcript 11 : ----------------------------------------------------CCCGGGCTGGAGTGCAATGGTGTGATCTCGGCTTATGGCAACCTCCGCCTCGTGGGTTCAAGCGATTCTC :  263
Transcript 12 : ------------------------------------------------------------------------------------------------------------------------ :    -
Transcript 13 : ------------------------------------------------------------------------------------------------------------------------ :    -
Transcript 14 : ------------------------------------------------------------------------------------------------------------------------ :    -
Transcript 15 : ------------------------------------------------------------------------------------------------------------------------ :    -
```

```
                    *       500         *       520         *       540         *       560         *       580         *       600
Reference     : -----------------CGGGCTCCCTTACTACTGGAATGCAGACACAGACCTTGTATCCTGGCTCTCCCCACATGACCCCAACTCCGTGGTTACCAAATCGGCCAAGAAGCTCAGAAGC :  296
Transcript 1  : -----------------CGGGCTCCCTTACTACTGGAATGCAGACACAGACCTTGTATCCTGGCTCTCCCCACATGACCCCAACTCCGTGGTTACCAAATCGGCCAAGAAGCTCAGAAGC :  296
Transcript 2  : -----------------CGGGCTCCCTTACTACTGGAATGCAGACACAGACCTTGTATCCTGGCTCTCCCCACATGACCCCAACTCCGTGGTTACCAAATCGGCCAAGAAGCTCAGAAGC :  296
Transcript 3  : -----------------CGGGCTCCCTTACTACTGGAATGCAGACACAGACCTTGTATCCTGGCTCTCCCCACATGACCCCAACTCCGTGGTTACCAAATCGGCCAAGAAGCTCAGAAGC :  296
Transcript 4  : -----------------CGGGCTCCCTTACTACTGGAATGCAGACACAGACCTTGTATCCTGGCTCTCCCCACATGACCCCAACTCCGTGGTTACCAAATCGGCCAAGAAGCTCAGAAGC :  296
Transcript 5  : -----------------CGGGCTCCCTTACTACTGGAATGCAGACACAGACCTTGTATCCTGGCTCTCCCCACATGACCCCAACTCCGTGGTTACCAAATCGGCCAAGAAGCTCAGAAGC :  296
Transcript 6  : -----------------CGGGCTCCCTTACTACTGGAATGCAGACACAGACCTTGTATCCTGGCTCTCCCCACATGACCCCAACTCCGTGGTTACCAAATCGGCCAAGAAGCTCAGAAGC :  296
Transcript 7  : -----------------CGGGCTCCCTTACTACTGGAATGCAGACACAGACCTTGTATCCTGGCTCTCCCCACATGACCCCAACTCCGTGGTTACCAAATCGGCCAAGAAGCTCAGAAGC :  296
Transcript 8  : -----------------CGGGCTCCCTTACTACTGGAATGCAGACACAGACCTTGTATCCTGGCTCTCCCCACATGACCCCAACTCCGTGGTTACCAAATCGGCCAAGAAGCTCAGAAGC :  296
Transcript 9  : CTGTCTCAGCCTCCCAGCGGGCTCCCTTACTACTGGAATGCAGACACAGACCTTGTATCCTGGCTCTCCCCACATGACCCCAACTCCGTGGTTACCAAATCGGCCAAGAAGCTCAGAAGC :  383
Transcript 10 : CTGTCTCAGCCTCCCAGCGGGCTCCCTTACTACTGGAATGCAGACACAGACCTTGTATCCTGGCTCTCCCCACATGACCCCAACTCCGTGGTTACCAAATCGGCCAAGAAGCTCAGAAGC :  600
Transcript 11 : CTGTCTCAGCCTCCCAGCGGGCTCCCTTACTACTGGAATGCAGACACAGACCTTGTATCCTGGCTCTCCCCACATGACCCCAACTCCGTGGTTACCAAATCGGCCAAGAAGCTCAGAAGC :  383
Transcript 12 : -----------------CGGGCTCCCTTACTACTGGAATGCAGACACAGACCTTGTATCCTGGCTCTCCCCACATGACCCCAACTCCGTGGTTACCAAATCGGCCAAGAAGCTCAGAAGC :  296
Transcript 13 : -----------------CGGGCTCCCTTACTACTGGAATGCAGAC---------------------------------------------------------------------- :  221
Transcript 14 : -----------------CGGGCTCCCTTACTACTGGAATGCAGACACAGACCTTGTATCCTGGCTCTCCCCACATGACCCCAACTCCGTGGTTACCAAATCGGCCAAGAAGCTCAGAAGC :  296
Transcript 15 : -----------------CGGGCTCCCTTACTACTGGAATGCAGACACAGACCTTGTATCCTGGCTCTCCCCACATGACCCCAACTCCGTGGTTACCAAATCGGCCAAGAAGCTCAGAAGC :  279


                    *       620         *       640         *       660         *       680         *       700         *       720
Reference     : AGTAATGCAGATGCTGAAGAAAAGTTGGACCGGAGCCATGACAAGTCGGACAGGGGCCATGACAAGTCGGACCGCAGCCATGAGAAACTAGACAGGGGCCACGACAAGTCAGACCGGGGC :  416
Transcript 1  : AGTAATGCAG---------------------------------------------------------------------------------------------------- :  306
Transcript 2  : AGTAATGCAGATGCTGAAGAAAAGTTGGACCGGAGCCATGACAAGTC----------------------GGACCGCAGCCATGAGAAACTAGACAGGGGCCACGACAAGTCAGACCGGGGC :  395
Transcript 3  : AGTAATGCAGATGCTGAAGAAAAGTTGGACCGGAGCCATGACAAGTC----------------------GGACCGCAGCCATGAGAAACTAGACAGGGGCCACGACAAGTCAGACCGGGGC :  395
Transcript 4  : AGTAATGCAGATGCTGAAGAAAAGTTGGACCGGAGCCATGACAAGTC----------------------GGACCGCAGCCATGAGAAACTAGACAGGGGCCACGACAAGTCAGACCGGGGC :  395
Transcript 5  : AGTAATGCAGATGCTGAAGAAAAGTTGGACCGGAGCCATGACAAGTC----------------------GGACCGCAGCCATGAGAAACTAGACAGGGGCCACGACAAGTCAGACCGGGGC :  395
Transcript 6  : AGTAATGCAGATGCTGAAGAAAAGTTGGACCGGAGCCATGACAAGTCGGACAGGGGCCATGACAAGTCGGACCGCAGCCATGAGAAACTAGACAGGGGCCACGACAAGTCAGACCGGGGC :  416
Transcript 7  : AGTAATGCAGATGCTGAAGAAAAGTTGGACCGGAGCCATGACAAGTCGGACAGGGGCCATGACAAGTCGGACCGCAGCCATGAGAAACTAGACAGGGGCCACGACAAGTCAGACCGGGGC :  416
Transcript 8  : AGTAATGCAGATGCTGAAGAAAAGTTGGACCGGAGCCATGACAAGTCGGACAGGGGCCATGACAAGTCGGACCGCAGCCATGAGAAACTAGACAGGGGCCACGACAAGTCAGACCGGGGC :  416
Transcript 9  : AGTAATGCAGATGCTGAAGAAAAGTTGGACCGGAGCCATGACAAGTCGGACAGGGGCCATGACAAGTCGGACCGCAGCCATGAGAAACTAGACAGGGGCCACGACAAGTCAGACCGGGGC :  503
Transcript 10 : AGTAATGCAGATGCTGAAGAAAAGTTGGACCGGAGCCATGACAAGTCGGACAGGGGCCATGACAAGTCGGACCGCAGCCATGAGAAACTAGACAGGGGCCACGACAAGTCAGACCGGGGC :  720
Transcript 11 : AGTAATGCAGATGCTGAAGAAAAGTTGGACCGGAGCCATGACAAGTCGGACAGGGGCCATGACAAGTCGGACCGCAGCCATGAGAAACTAGACAGGGGCCACGACAAGTCAGACCGGGGC :  503
Transcript 12 : AGTAATGCAGATGCTGAAGAAAAGTTGGACCGGAGCCATGACAAGTCGGACAGGGGCCATGACAAGTCGGACCGCAGCCATGAGAAACTAGACAGGGGC---------------------- :  395
Transcript 13 : ---------------------------------------------------------------------------------------------------- :    -
Transcript 14 : AGTAATGCAGATGCTGAAGAAAAGTTGGACCGGAGCCAAGACAAGTCGGACAGGGGCCATGACAAGTCGGACCGCAGCCATGAGAAACTAGACAGGGGCCACGACAAGTCAGACCGGGGC :  416
Transcript 15 : AGTAATGCAG---------------------------------------------------------------------------------------------------- :  289
```

```
                    *       740         *       760         *       780         *       800         *       820         *       840
Reference     : CACGACAAGTCTGACAGGGATCGAGAGCGTGGCTATGACAAGGTAGACAGAGAGAGAGAGAGCGAGACAGGGAACGGGATCGGGACCGCGGGTATGACAAGGCAGACCGGGAAGAGGGCAAA :  536
Transcript 1  : ------------------------------------------------------------------------------------------------------------------------ :    -
Transcript 2  : CACGACAAGTCTGACAGGGATCGAGAGCGTGGCTATGACAAGGTAGACAGAGAGAGAGAGAGCGAGACAGGGAACGGGATCGGGACCGCGGGTATGACAAGGCAGACCGGGAAGAGGGCAAA :  515
Transcript 3  : CACGACAAGTCTGACAGGGATCGAGAGCGTGGCTATGACAAGGTAGACAGAGAGAGAGAGAGCGAGACAGGGAACGGGATCGGGACCGCGGGTATGACAAGGCAGACCGGGAAGAGGGCAAA :  515
Transcript 4  : CACGACAAGTCTGACAGGGATCGAGAGCGTGGCTATGACAAG------------------------------------------------------------------------------ :  437
Transcript 5  : CACGACAAGTCTGACAGGGATCGAGAGCGTGGCTATGACAAGGTAGACAGAGAGAGAGAGAGCGAGACAGGGAACGGGATCGGGACCGCGGGTATGACAAGGCAGACCGGGAAGAGGGCAAA :  515
Transcript 6  : CACGACAAGTCTGACAGGGATCGAGAGCGTGGCTATGACAAG------------------------------------------------------------------------------ :  458
Transcript 7  : CACGACAAGTCTGACAGGGATCGAGAGCGTGGCTATGACAAGGTAGACAGAGAGAGAGAGAGCGAGACAGGGAACGGGATCGGGACCGCGGGTATGACAAGGCAGACCGGGAAGAGGGCAAA :  536
Transcript 8  : CACGACAAGTCTGACAGGGATCGAGAGCGTGGCTATGACAAGGTAGACAGAGAGAGAGAGAGCGAGACAGGGAACGGGATCGGGACCGCGGGTATGACAAGGCAGACCGGGAAGAGGGCAAA :  536
Transcript 9  : CACGACAAGTCTGACAGGGATCGAGAGCGTGGCTATGACAAGGTAGACAGAGAGAGAGAGAGCGAGACAGGGAACGGGATCGGGACCGCGGGTATGACAAGGCAGACCGGGAAGAGGGCAAA :  623
Transcript 10 : CACGACAAGTCTGACAGGGATCGAGAGCGTGGCTATGACAAGGTAGACAGAGAGAGAGAGAGCGAGACAGGGAACGGGATCGGGACCGCGGGTATGACAAGGCAGACCGGGAAGAGGGCAAA :  840
Transcript 11 : CACGACAAGTCTGACAGGGATCGAGAGCGTGGCTATGACAAGGTAGACAGAGAGAGAGAGAGCGAGACAGGGAACGGGATCGGGACCGCGGGTATGACAAGGCAGACCGGGAAGAGGGCAAA :  623
Transcript 12 : ------------------------------------------------------------------------------------------------------------------------ :    -
Transcript 13 : ----------------------------------------------------------------------------------------------------CGGGAAGAGGGCAAA :  236
Transcript 14 : CACGACAAGTCTGACAGGGATCGAGAGCGTGGCTATGGCAAGGTAGACAGAGAGAGAGAGAGCGAGACAGGGAACGGGATCGGGACCGCGGGTATGACAAGGCAGACCGGGAAGAGGGCAAA :  536
Transcript 15 : ------------------------------------------------------------------------------------------------------------------------ :    -


                    *       860         *       880         *       900         *       920         *       940         *       960
Reference     : GAACGGCGCCACCATCGCCGGGAGGAGCTGGCTCCCTATCCCAAGAGCAAGAAGG---------------------------------------------------------------- :  591
Transcript 1  : ------------------------------------------------------------------------------------------------------------------------ :    -
Transcript 2  : GAACGGCGCCACCATCGCCGGGAGGAGCTGGCTCCCTATCCCAAGAGCAAGAAGG---------------------------------------------------------------- :  570
Transcript 3  : GAACGGCGCCACCATCGCCGGGAGGAGCTGGCTCCCTATCCCAAGAGCAAGAAGGGTAAGCTGGGCAGAATGGGGCTCGGTGAGACCAACAAGGTGCAGGGTGCACTGCGTGAGGAAGCC :  635
Transcript 4  : ------------------------------------------------------------------------------------------------------------------------ :    -
Transcript 5  : GAACGGCGCCACCATCGCCGGGAGGAGCTGGCTCCCTATCCCAAGAGCAAGAAGG---------------------------------------------------------------- :  570
Transcript 6  : ------------------------------------------------------------------------------------------------------------------------ :    -
Transcript 7  : GAACGGCGCCACCATCGCCGGGAGGAGCTGGCTCCCTATCCCAAGAGCAAGAAGGGTAAGCTGGGCAGAATGGGGCTCGGTGAGACCAACAAGGTGCAGGGTGCACTGCGTGAGGAAGCC :  656
Transcript 8  : GAACGGCGCCACCATCGCCGGGAGGAGCTGGCTCCCTATCCCAAGAGCAAGAAGGGTAAGCTGGGCAGAATGGGGCTCGGTGAGACCAACAAGGTGCAGGGTGCACTGCGTGAGGAAGCC :  656
Transcript 9  : GAACGGCGCCACCATCGCCGGGAGGAGCTGGCTCCCTATCCCAAGAGCAAGAAGG---------------------------------------------------------------- :  678
Transcript 10 : GAACGGCGCCACCATCGCCGGGAGGAGCTGGCTCCCTATCCCAAGAGCAAGAAGG---------------------------------------------------------------- :  895
Transcript 11 : GAACGGCGCCACCATCGCCGGGAGGAGCTGGCTCCCTATCCCAAGAGCAAGAAGGGTAAGCTGGGCAGAATGGGGCTCGGTGAGACCAACAAGGTGCAGGGTGCACTGCGTGAGGAAGCC :  743
Transcript 12 : ------------------------------------------------------------------------------------------------------------------------ :    -
Transcript 13 : GAACGGCGCCACCATCGCCGGGAGGAGCTGGCTCCCTATCCCAAGAGCAAGAAGG---------------------------------------------------------------- :  291
Transcript 14 : GAACGGCGCCACCATCGCCGGGAGGAGCTGGCTCCCTATCCCAAGAGCAAGAAGG---------------------------------------------------------------- :  591
Transcript 15 : ------------------------------------------------------------------------------------------------------------------------ :    -
```

```
                    *       980         *      1000         *      1020         *      1040         *      1060         *      1080
Reference     : ---------------------------------------------------------------------------------------------------------------------- :     -
Transcript 1  : ---------------------------------------------------------------------------------------------------------------------- :     -
Transcript 2  : ---------------------------------------------------------------------------------------------------------------------- :     -
Transcript 3  : TTCCCTCAAAAAGATGCCTGGACCTGGGGCTAGAGGAGGGTGCTGTGGTACATGGCAGCCAGGGGCTTCATTTCTTCTTGTGGGGTGGGGCTCAGTGATCAGGGGCTCCTGGTGCCTCTA :   755
Transcript 4  : ---------------------------------------------------------------------------------------------------------------------- :     -
Transcript 5  : ---------------------------------------------------------------------------------------------------------------------- :     -
Transcript 6  : ---------------------------------------------------------------------------------------------------------------------- :     -
Transcript 7  : TTCCCTCAAAAAGATGCCTGGACCTGGGGCTAGAGGAGGGTGCTGTGGTACATGGCAGCCAGGGGCTTCATTTCTTCTTGTGGGGTGGGGCTCAGTGATCAGGGGCTCCTGGTGCCTCTA :   776
Transcript 8  : TTCCCTCAAAAAGATGCCTGGACCTGGGGCTAGAGGAGGGTGCTGTGGTACATGGCAGCCAGGGGCTTCATTTCTTCTTGTGGGGTGGGGCTCAGTGATCAGGGGCTCCTGGTGCCTCTA :   776
Transcript 9  : ---------------------------------------------------------------------------------------------------------------------- :     -
Transcript 10 : ---------------------------------------------------------------------------------------------------------------------- :     -
Transcript 11 : TTCCCTCAAAAAGATGCCTGGACCTGGGGCTAGAGGAGGGTGCTGTGGTACATGGCAGCCAGGGGCTTCATTTCTTCTTGTGGGGTGGGGCTCAGTGATCAGGGGCTCCTGGTGCCTCTA :   863
Transcript 12 : ---------------------------------------------------------------------------------------------------------------------- :     -
Transcript 13 : ---------------------------------------------------------------------------------------------------------------------- :     -
Transcript 14 : ---------------------------------------------------------------------------------------------------------------------- :     -
Transcript 15 : ---------------------------------------------------------------------------------------------------------------------- :     -
```

```
                    *      1100         *      1120         *      1140         *      1160         *      1180         *      1200
Reference     : ----------------------------CAGTAAGCCGAAAGGATGAAGAGTTAGACCCCATGGACCCTAGCTCATACTCAGACGCCCCCCGG------------------------- :   656
Transcript 1  : ----------------------------CAGTAAGCCGAAAGGATGAAGAGTTAGACCCCATGGACCCTAGCTCATACTCAGACGCCCCCCGG------------------------- :   371
Transcript 2  : ----------------------------CAGTAAGCCGAAAGGATGAAGAGTTAGACCCCATGGACCCTAGCTCATACTCAGACGCCCCCCGG------------------------- :   635
Transcript 3  : TTGAAGACTTTGCCCTGCCACTTCCACAGCAGTAAGCCGAAAGGATGAAGAGTTAGACCCCATGGACCCTAGCTCATACTCAGACGCCCCCCGG------------------------- :   849
Transcript 4  : ----------------------------CAGTAAGCCGAAAGGATGAAGAGTTAGACCCCATGGACCCTAGCTCATACTCAGACGCCCCCCGG------------------------- :   502
Transcript 5  : -------------------------------TAAGCCGAAAGGATGAAGAGTTAGACCCCATGGACCCTAGCTCATACTCAGACGCCCCCCGG------------------------- :   632
Transcript 6  : ----------------------------CAGTAAGCCGAAAGGATGAAGAGTTAGACCCCATGGACCCTAGCTCATACTCAGACGCCCCCCGG------------------------- :   523
Transcript 7  : TTGAAGACTTTGCCCTGCCACTTCCACAGCAGTAAGCCGAAAGGATGAAGAGTTAGACCCCATGGACCCTAGCTCATACTCAGACGCCCCCCGG------------------------- :   870
Transcript 8  : TTGAAGACTTTGCCCTGCCACTTCCACAGCAGTAAGCCGAAAGGATGAAGAGTTAGACCCCATGGACCCTAGCTCATACTCAGACGCCCCCCGGTAAGTGACAACCCCTCTTGACTCAGT :   896
Transcript 9  : ----------------------------CAGTAAGCCGAAAGGATGAAGAGTTAGACCCCATGGACCCTAGCTCATACTCAGACGCCCCCCGG------------------------- :   743
Transcript 10 : ----------------------------CAGTAAGCCGAAAGGATGAAGAGTTAGACCCCATGGACCCTAGCTCATACTCAGACGCCCCCCGG------------------------- :   960
Transcript 11 : TTGAAGACTTTGCCCTGCCACTTCCACAGCAGTAAGCCGAAAGGATGAAGAGTTAGACCCCATGGACCCTAGCTCATACTCAGACGCCCCCCGG------------------------- :   957
Transcript 12 : -------------------------------------------------------------------------------------------------------------------- :     -
Transcript 13 : ----------------------------CAGTAAGCCGAAAGGATGAAGAGTTAGACCCCATGGACCCTAGCTCATACTCAGACGCCCCCCGG------------------------- :   356
Transcript 14 : -------------------------------TAAGCCGAAAGGGTGAAGAGTTAGACCCCATGGACCCTAGCTCATACTCAGACGCCCCCCGG------------------------- :   653
Transcript 15 : ----------------------------CAGTAAGCCGAAAGGATGAAGAGTTAGACCCCATGGACCCTAGCTCATACTCAGACGCCCCCCGG------------------------- :   354
```

```
                      *      1220         *      1240         *      1260         *      1280         *      1300         *      1320
Reference     : ----------------------------------------------------------------------------------------------------------------GGCACGTGGTCAAC :  670
Transcript 1  : ----------------------------------------------------------------------------------------------------------------GGCACGTGGTCAAC :  385
Transcript 2  : ----------------------------------------------------------------------------------------------------------------GGCACGTGGTCAAC :  649
Transcript 3  : ----------------------------------------------------------------------------------------------------------------GGCACGTGGTCAAC :  863
Transcript 4  : ----------------------------------------------------------------------------------------------------------------GGCACGTGGTCAAC :  516
Transcript 5  : ----------------------------------------------------------------------------------------------------------------GGCACGTGGTCAAC :  646
Transcript 6  : ----------------------------------------------------------------------------------------------------------------GGCACGTGGTCAAC :  537
Transcript 7  : ----------------------------------------------------------------------------------------------------------------GGCACGTGGTCAAC :  884
Transcript 8  : ACGTGGACACCATCCTCCGGCCTCCTTCCTCCATTCCTCATTGGGACCAGGTGGGCTGTGTCCGCCACATCACCCATCCCCATCCCCTGACTCTTTCACCGGCAGGGGCACGTGGTCAAC : 1016
Transcript 9  : ----------------------------------------------------------------------------------------------------------------GGCACGTGGTCAAC :  757
Transcript 10 : ----------------------------------------------------------------------------------------------------------------GGCACGTGGTCAAC :  974
Transcript 11 : ----------------------------------------------------------------------------------------------------------------GGCACGTGGTCAAC :  971
Transcript 12 : ----------------------------------------------------------------------------------------------------------------               :    -
Transcript 13 : ----------------------------------------------------------------------------------------------------------------GGCACGTGGTCAAC :  370
Transcript 14 : ----------------------------------------------------------------------------------------------------------------GGCACGTGGTCAAC :  667
Transcript 15 : ----------------------------------------------------------------------------------------------------------------GGCACGTGGTCAAC :  368
```

```
                      *      1340         *      1360         *      1380         *      1400         *      1420         *      1440
Reference     : AGGACTCCCCAAGCGGAATGAGGCCAAGACTGGCGCTGACACCACAGCAGCTGGGCCCCTCTTCCAGCAGCGGCCGTATCCATCCCCAGGGGCTGTGCTCCGGGCCAATGCAGAGGCCT :  789
Transcript 1  : AGGACTCCCCAAGCGGAATGAGGCCAAGACTGGCGCTGACACCACAGCAGCTGGGCCCCTCTTCCAGCAGCGGCCGTATCCATCCCCAGGGGCTGTGCTCCGGGCCAATGCAGAGGCCT :  504
Transcript 2  : AGGACTCCCCAAGCGGAATGAGGCCAAGACTGGCGCTGACACCACAGCAGCTGGGCCCCTCTTCCAGCAGCGGCCGTATCCATCCCCAGGGGCTGTGCTCCGGGCCAATGCAGAGGCCT :  768
Transcript 3  : AGGACTCCCCAAGCGGAATGAGGCCAAGACTGGCGCTGACACCACAGCAGCTGGGCCCCTCTTCCAGCAGCGGCCGTATCCATCCCCAGGGGCTGTGCTCCGGGCCAATGCAGAGGCCT :  982
Transcript 4  : AGGACTCCCCAAGCGGAATGAGGCCAAGACTGGCGCTGACACCACAGCAGCTGGGCCCCTCTTCCAGCAGCGGCCGTATCCATCCCCAGGGGCTGTGCTCCGGGCCAATGCAGAGGCCT :  635
Transcript 5  : AGGACTCCCCAAGCGGAATGAGGCCAAGACTGGCGCTGACACCACAGCAGCTGGGCCCCTCTTCCAGCAGCGGCCGTATCCATCCCCAGGGGCTGTGCTCCGGGCCAATGCAGAGGCCT :  765
Transcript 6  : AGGACTCCCCAAGCGGAATGAGGCCAAGACTGGCGCTGACACCACAGCAGCTGGGCCCCTCTTCCAGCAGCGGCCGTATCCATCCCCAGGGGCTGTGCTCCGGGCCAATGCAGAGGCCT :  656
Transcript 7  : AGGACTCCCCAAGCGGAATGAGGCCAAGACTGGCGCTGACACCACAGCAGCTGGGCCCCTCTTCCAGCAGCGGCCGTATCCATCCCCAGGGGCTGTGCTCCGGGCCAATGCAGAGGCCT : 1003
Transcript 8  : AGGACTCCCCAAGCGGAATGAGGCCAAGACTGGCGCTGACACCACAGCAGCTGGGCCCCTCTTCCAGCAGCGGCCGTATCCATCCCCAGGGGCTGTGTTCCGGGCCAATGCAGAGGCCT : 1136
Transcript 9  : AGGACTCCCCAAGCGGAATGAGGCCAAGACTGGCGCTGACACCACAGCAGCTGGGCCCCTCTTCCAGCAGCGGCCGTATCCATCCCCAGGGGCTGTGCTCCGGGCCAATGCAGAGGCCT :  876
Transcript 10 : AGGACTCCCCAAGCGGAATGAGGCCAAGACTGGCGCTGACACCACAGCAGCTGGGCCCCTCTTCCAGCAGCGGCCGTATCCATCCCCAGGGGCTGTGCTCCGGGCCAATGCAGAGGCCT : 1093
Transcript 11 : AGGACTCCCCAAGCGGAATGAGGCCAAGACTGGCGCTGACACCACAGCAGCTGGGCCCCTCTTCCAGCAGCGGCCGTATCCATCCCCAGGGGCTGTGCTCCGGGCCAATGCAGAGGCCT : 1090
Transcript 12 : ----------------------------------------------------------------------------------------TGTGCTCCGGGCCAATGCAGAGGCCT :  421
Transcript 13 : AGGACTCCCCAAGCGGAATGAGGCCAAGACTGGCGCTGACACCACAGCAGTTGGGCCCCTCTTCCAGCAGCGGCCGTATCCATCCCCAGGGGCTGTGCTCCGGGCCAATGCAGAGGCCT :  489
Transcript 14 : AGGACTCCCCAAGCGGAATGAGGCCAAGACTGGCGCTGACACCACAGCAGCTGGGCCCCTCTTCCAGCAGCGGCCGTATCCATCCCCAGGGGCTGTGCTCCGGGCCAATGCAGAGGCCT :  786
Transcript 15 : AGGACTCCCCAAGCGGAATGAGGCCAAGACTGGCGCTGACACCACAGCAGCTGGGCCCCTCTTCCAGCAGCGGCCGTATCCATCCCCAGGGGCTGTGCTCCGGGCCAATGCAGAGGCCT :  487
```

```
Reference     : CCCGAACCAAGCAGCAGGATTGAAGCTTC       :  817
Transcript 1  : CCCGAACCAAGCAGCAGGATTGAAGCTTC       :  533
Transcript 2  : CCCGAACCAAGCAGCAGGATTGAAGCTTC       :  797
Transcript 3  : CCCGAACCAAGCAGCAGGATTGAAGCTTC       : 1011
Transcript 4  : CCCGAACCAAGCAGCAGGATTGAAGCTTC       :  665
Transcript 5  : CCCGAACCAAGCAGCAGGATTGAAGCTTC       :  794
Transcript 6  : CCCGAACCAAGCAGCAGGATTGAAGCTTC       :  685
Transcript 7  : CCCGAACCAAGCAGCAGGATTGAAGCTTC       : 1032
Transcript 8  : CCCGAACCAAGCAGCAGGATTGAAGCTTC       : 1165
Transcript 9  : CCCGAACCAAGCAGCAGGATTGAAGCTTC       :  905
Transcript 10 : CCCGAACCAAGCAGCAGGATTGAAGCTTC       : 1122
Transcript 11 : CCCGAACCAAGCAGCAGGATTGAAGCTTC       : 1119
Transcript 12 : CCCGAACCAAGCAGCAGGATTGAAGCTTC       :  450
Transcript 13 : CCCGAACCAAGCAGCAGGATTGAAGCTTC       :  520
Transcript 14 : CCCGAACCAAGCAGCAGGATTGAAGCTTC       :  815
Transcript 15 : CCCGAACCAAGCAGCAGGATTGAAGCTTC       :  516
```

APPENDIX VI Primers used to quantify *PQBP1* alternative transcripts

NAME	Targeted transcript	Sense (5' 3')	Antisense (5' → 3')
PQBP1.Q1	all	CCTCAAACATCTGGAGCCTGAAC	TCGTCATAGTCCTCGGCAATG
PQBP1.Q2	1, 15	CCCAACTCCGTGGTTACCAA	GGCTTACTGCTGCATTACTGCTT
PQBP1.Q2b	1, 1,5	GGGCTCCCTTACTACTGGAA	GGCTTACTGCTGCATTACTGCTT
PQBP1.Q3	4 ,6	GACCGCAGCCATGAGAAACTAGA	CTTTCGGCTTACTGCTTGTCATAGC
PQBP1.Q4	3, 7, 8, 11	GGCTCCTGGTGCCTCTATTG	CATCCTTTCGGCTTACTGCTG
PQBP1.Q5	8	CCTTCCTCCATTCCTCATTGG	GGTGATGTGGCGGACACAG
PQBP1.Q6	9, 10, 11	CTGCCCGGGCTGGAGTG	GAGAGCCAGGATACAAGGTCTGTGT
PQBP1.Q7	12	GCTGAAGAAAAGTTGGACCGG	GGAGCACAGCCCCTGTCTAGT
PQBP1.Q8	13	GGAATGCAGACCGGGAAGAG	CCTCCCGGCGATGGTG
PQBP1.Q8b	13	GGAATGCAGACCGGGAAGAG	CTTTCGGCTTACTGCCTTCT
PQBP1.Q9	5, 14	ATCGAGAGCGTGGCTATGG	CACCCTTTCGGCTTACCTTCT
PQBP1.Q10	all	CAGACCCGCTTGGCCAAGA	TGGTGGTAGGCCCTCCAA
PQBP1.Q11	all	CTATGCCGCTGCCCGTTG	TTGTACCAGCTTGGTGGTA

APPENDIX VII Multiple sequence alignment of *PQBP1* peptides

Predicted open reading frames from the *PQBP1* alternative transcripts were identified and the amino acid sequences extracted from orf-Finder at NCBI. Transcript numbers refer to those outlined in section 5.2. Sequences were aligned using clustalw and edited in GeneDoc. Alternating exons are shown in grey, amino acids resulting from retained introns are shown in yellow, deleted regions are shown in red while alterative translation start sites are shown in green.

```
Ref  : MPLPVALQTRLAKRGILKHLEPEPEEEIIAEDYDDDPVDYEATRLEGLPP :
Tr  1: MPLPVALQTRLAKRGILKHLEPEPEEEIIAEDYDDDPVDYEATRLEGLPP :
Tr  2: MPLPVALQTRLAKRGILKHLEPEPEEEIIAEDYDDDPVDYEATRLEGLPP :
Tr  3: MPLPVALQTRLAKRGILKHLEPEPEEEIIAEDYDDDPVDYEATRLEGLPP :
Tr  4: MPLPVALQTRLAKRGILKHLEPEPEEEIIAEDYDDDPVDYEATRLEGLPP :
Tr  5: MPLPVALQTRLAKRGILKHLEPEPEEEIIAEDYDDDPVDYEATRLEGLPP :
Tr  6: MPLPVALQTRLAKRGILKHLEPEPEEEIIAEDYDDDPVDYEATRLEGLPP :
Tr  7: MPLPVALQTRLAKRGILKHLEPEPEEEIIAEDYDDDPVDYEATRLEGLPP :
Tr 9a: MPLPVALQTRLAKRGILKHLEPEPEEEIIAEDYDDDPVDYEATRLEGLPP :
Tr10a: MPLPVALQTRLAKRGILKHLEPEPEEEIIAEDYDDDPVDYEATRLEGLPP :
Tr10b: ------------------------------------------------- :
Tr11b: ------------------------------------------------- :
Tr 12: MPLPVALQTRLAKRGILKHLEPEPEEEIIAEDYDDDPVDYEATRLEGLPP :
Tr 13: MPLPVALQTRLAKRGILKHLEPEPEEEIIAEDYDDDPVDYEATRLEGLPP :
Tr 14: MPLPVALQTRLAKRGILKHLEPEPEEEIIAEDYDDDPVDYEATRLEGLPP :
Tr 15: MPLPVALQTRLAKRGILKHLEPEPEEEIIAEDYDDDPVDYEATRLEGLPP :


Ref  : WLSPHDPNSVVTKSAKKLRSSNADAEEKLDRSHDKSDRGHDKSDRSHEKL :
Tr  1  WLSPHDPNSVVTKSAKKLRSSNA--------------------------- :
Tr  2: WLSPHDPNSVVTKSAKKLRSSNADAEEKLDRS-------HDKSDRSHEKL :
Tr  3: WLSPHDPNSVVTKSAKKLRSSNADAEEKLDRS-------HDKSDRSHEKL :
Tr  4: WLSPHDPNSVVTKSAKKLRSSNADAEEKLDRS-------HDKSDRSHEKL :
Tr  5: WLSPHDPNSVVTKSAKKLRSSNADAEEKLDRS-------HDKSDRSHEKL :
Tr  6: WLSPHDPNSVVTKSAKKLRSSNADAEEKLDRSHDKSDRGHDKSDRSHEKL :
Tr  7: WLSPHDPNSVVTKSAKKLRSSNADAEEKLDRSHDKSDRGHDKSDRSHEKL :
Tr 9a: ------------------------------------------------- :
Tr10a: ------------------------------------------------- :
Tr10b: WLSPHDPNSVVTKSAKKLRSSNADAEEKLDRSHDKSDRGHDKSDRSHEKL :
Tr11b: WLSPHDPNSVVTKSAKKLRSSNADAEEKLDRSHDKSDRGHDKSDRSHEKL :
Tr 12: WLSPHDPNSVVTKSAKKLRSSNADAEEKLDRSHDKSDRGHDKSDRSHEKL :
Tr 13: SWYKVFNPSCGLPYYWN---------------------------------- :
Tr 14: WLSPHDPNSVVTKSAKKLRSSNADAEEKLDRSHDKSDRGHDKSDRSHEKL :
Tr 15: SWYKRAPLLLECRHRPCILALPT-------------------------- :


Ref  : DRG-----------------HDKSDRGHDKSDRDRERGYDKVDRERERD :
Tr  1: ------------------------------------------------- :
Tr  2: DRG-----------------HDKSDRGHDKSDRDRERGYDKVDRERERD :
Tr  3: DRG-----------------HDKSDRGHDKSDRDRERGYDKVDRERERD :
Tr  4: DRG-----------------HDKSDRGHDKSDRDRERGYDKQ------- :
Tr  5: DRG-----------------HDKSDRGHDKSDRDRERGYDKVDRERERD :
Tr  6: DRG-----------------HDKSDRGHDKSDRDRERGYDKQ------- :
Tr  7: DRG-----------------HDKSDRGHDKSDRDRERGYDKVDRERERD :
Tr 9a: ------------------------------------------------- :
Tr10a: ------------------------------------------------- :
Tr10b: DRG-----------------HDKSDRGHDKSDRDRERGYDKVDRERERD :
Tr11b: DRG-----------------HDKSDRGHDKSDRDRERGYDKVDRERERD :
Tr 12: DRGCAPGQCRGLPNQAAGLKL---------------------------- :
Tr 13: ---------------------------------------------------- :
Tr 14: DRG-----------------HDKSDRGHDKSDRDRERGYGKVDRERERD :
Tr 15: ------------------------------------------------- :
```

```
Ref  : RERDRDRGYDKADREEGKERRHHRREELAPYPKSKK---------- :
Tr  1: ------------------------------------------------ :
Tr  2: RERDRDRGYDKADREEGKERRHHRREELAPYPKSKK---------- :
Tr  3: RERDRDRGYDKADREEGKERRHHRREELAPYPKSKKGKLGRMGLGETNKV :
Tr  4: ------------------------------------------------ :
Tr  5: RERDRDRGYDKADREEGKERRHHRREELAPYPKSKK---------- :
Tr  6: ------------------------------------------------ :
Tr  7: RERDRDRGYDKADREEGKERRHHRREELAPYPKSKKGKLGRMGLGETNKV :
Tr 9a: ------------------------------------------------ :
Tr10a: ------------------------------------------------ :
Tr10b: RERDRDRGYDKADREEGKERRHHRREELAPYPKSKK---------- :
Tr11b: RERDRDRGYDKADREEGKERRHHRREELAPYPKSKKGKLGRMGLGETNKV :
Tr 12: ------------------------------------------------ :
Tr 13: ----------ADREEGKERRHHRREELAPYPKSKK---------- :
Tr 14: RERDRDRGYDKADREEGKERRHHRREELAPYPKSKK---------- :
Tr 15: ------------------------------------------------ :


Ref  : ----------------AVSRKDEELDPMDPSSYSDAPRGTWSTGLPKR :
Tr  1: ----------------AVSRKDEELDPMDPSSYSDAPRGTWSTGLPKR :
Tr  2: ----------------AVSRKDEELDPMDPSSYSDAPRGTWSTGLPKR :
Tr  3: QGALREEAFPQKDAWTWG------------------------------ :
Tr  4: ------------------------------------------------ :
Tr  5: ----------------VSRKDEELDPMDPSSYSDAPRGTWSTGLPKR :
Tr  6: ------------------------------------------------ :
Tr  7: QGALREEAFPQKDAWTWG------------------------------ :
Tr 9a: ------------------------------------------------ :
Tr10a: ------------------------------------------------ :
Tr10b: ----------------AVSRKDEELDPMDPSSYSDAPRGTWSTGLPKR :
Tr11b: QGALREEAFPQKDAWTWG------------------------------ :
Tr 12: ------------------------------------------------ :
Tr 13: ----------------AVSRKDEELDPMDPSSYSDAPRGTWSTGLPKR :
Tr 14: ----------------VSRKGEELDPMDPSSYSDAPRGTWSTGLPKR :
Tr 15: ------------------------------------------------ :


Ref  : NEAKTGADTTAAGPLFQQRPYPSPGAVLRANAEASRTKQQD--------- :
Tr  1: NEAKTGADTTAAGPLFQQRPYPSPGAVLRANAEASRTKQQD--------- :
Tr  2: NEAKTGADTTAAGPLFQQRPYPSPGAVLRANAEASRTKQQD--------- :
Tr  3: ------------------------------------------------ :
Tr  4: ------------------------------------------------ :
Tr  5: NEAKTGADTTAAGPLFQQRPYPSPGAVLRANAEASRTKQQD--------- :
Tr  6: ------------------------------------------------ :
Tr  7: ------------------------------------------------ :
Tr 9a: ------------------------------------------------ :
Tr10a: ------------------------------------------------ :
Tr10b: NEAKTGADTTAAGPLFQQRPYPSPGAVLRANAEASRTKQQD--------- :
Tr11b: ------------------------------------------------ :
Tr 12: ------------------------------------------------ :
Tr 13: NEAKTGADTTAVGPLFQQRPYPSPGAVLRANAEASRTKQQD--------- :
Tr 14: NEAKTGADTTAAGPLFQQRPYPSPGAVLRANAEASRTKQQD--------- :
Tr 15: ------------------------------------------------ :
```

APPENDIX VIII Primer combinations and sequences used in the preparation of T7 epitope::*PBQP1* (variant) pCDNA.3 constructs

Primer combinations

Transcript	N-tag primers		C-tag primers	
Reference	N1S	N1A	C1S	C1A
1	N1S	N1A	C1S	C1A
2	N1S	N1A	C1S	C1A
3	N1S	N3A	C1S	C3A
4	N1S	N2A	C1S	C2A
5	N1S	N1A	C1S	C1A
6	N1S	N2A	C1S	C2A
7	N1S	N3A	C1S	C3A
9a	N1S	N4A	C1S	C4A
10a	N1S	N5A	C1S	C5A
10b	N2S	N1A	C2S	C1A
11b	N2S	N3A	C2S	C3A
12	X1S	X1A	X1S	X1A
13	N1S	N1A	C1S	C1A
14	N1S	N1A	C1S	C1A
15	N1S	N6A	C1S	C6A

Primer Sequences

Direction	Name	Sequence (5' → 3')
Sense	N1S	GGCCAAGCTTTAGGCTATGCCGCTGCCCGTTG
	N2S	GGCCAAGCTTTAGCTTATGGCAACCTCCGCCT
	C1S	GGCCGCGGCCGCTATGCCGCTGCCCGTGG
	C2S	GGCCGCGGCCGCTATGGCAACCTCCGCCT
	X1S	ATGCCGCTGCCCGTTGCGCT
Antisense	N1A	GGCCGCTAGCATCCTGCTGCTTGGTTCGGGAG
	N2A	GGCCGCTAGCCTGCTTGTCATAGCCACGCTCT
	N3A	GGCCGCTAGCGCCCCAGGTCCAGGCATCTTTT
	N4A	GGCCGCTAGCCACCATTGCACTCCAGCCCGGG
	N5A	GGCCGCTAGCTGTGGGGAGAGCCAGGATACAA
	N6A	GGCCGCTAGCGGAGGAAGGGTCGAACACCTTG
	C1A	GGCCTCTAGATCAATCCTGCTGCTTGG
	C2A	GGCCTCTAGATTACTGCTTGTCATAGC
	C3A	GGCCTCTAGACTAGCCCCAGGTCCAGG
	C4A	GGCCTCTAGATCACACCATTGCACTCC
	C5A	GGCCTCTAGATCATGTGGGGAGAGCCA
	C6A	GGCCTCTAGATCACCAGGAAGGGTCGA
	X1A	CCAAGCTTCAATCCTGCTGCTT

APPENDIX IX Control primers used in real-time PCR analysis

Gene	Species	Sense (5' → 3')	Antisense (5' → 3')
GAPDH	Human	GAAGGTGAAGGTCGGAGTC	GAAGATGGTGATGGGATTTC
ActB	Chinese Hamster	ACCAACTGGGACGACATGGAGAAGA	TACGACCAGAGGCATACAGGGACAA
LacZ		TTGAAAATGGTCTGCTGCTG	TATTGGCTTCATCCACCACA
Luc		TGCAGAGATCCTGTGTTTGG	GTACCAGCAACGCACTTTGA