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.