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. elegans* 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.

De novo gene identification

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).

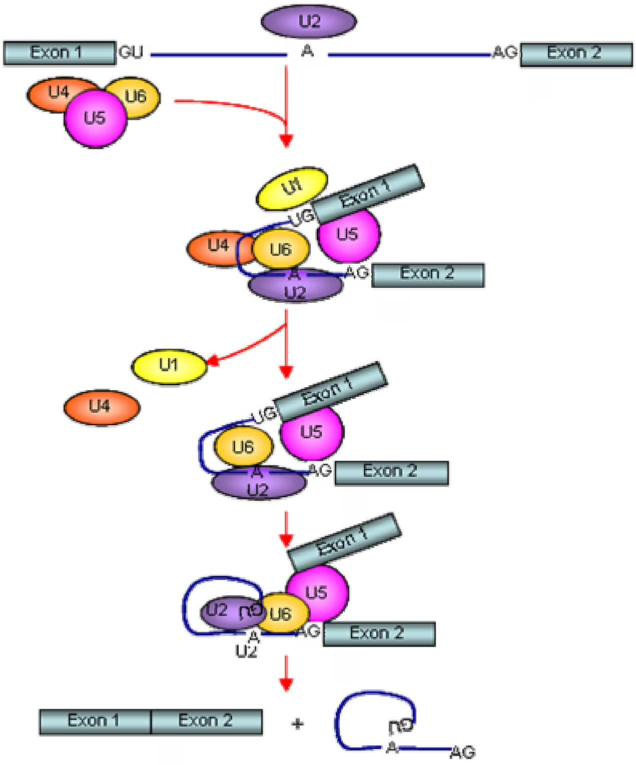


Figure 1.1 mRNA splicing

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.

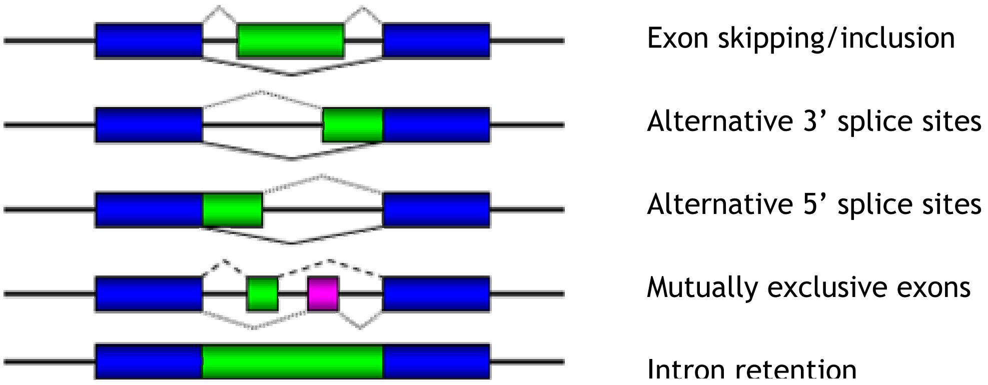


Figure 1.2 Types of splicing variation
Shown are the possible splicing patterns that can occur through alternative splicing. The genome sequence is displayed as a thick black line. Constitutive exons are displayed in blue, while alternative exons are displayed in green (or pink). The exon junctions between 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 AEDB AltExtron	 mRNA EST, mRNA	Manually curated from published transcripts Sequences must conform to either the GT-AG or GC-AG splice sites. Duplicate genes (> 99% un-gapped sequence identity) removed. Hypervariable genes are removed	 1100 entries 37%	Thanaraj *et al.*, 2004
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 α2 (*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).

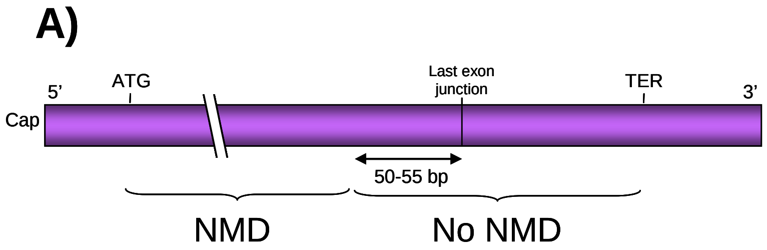
On account of NMD, PTC harbouring transcripts do not generally participate in the synthesis of truncated proteins, which could have a dominant negative effect on the organism. In this manner, NMD is a surveillance mechanism that recognises and degrades aberrant transcripts resulting from erroneous mRNA 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).

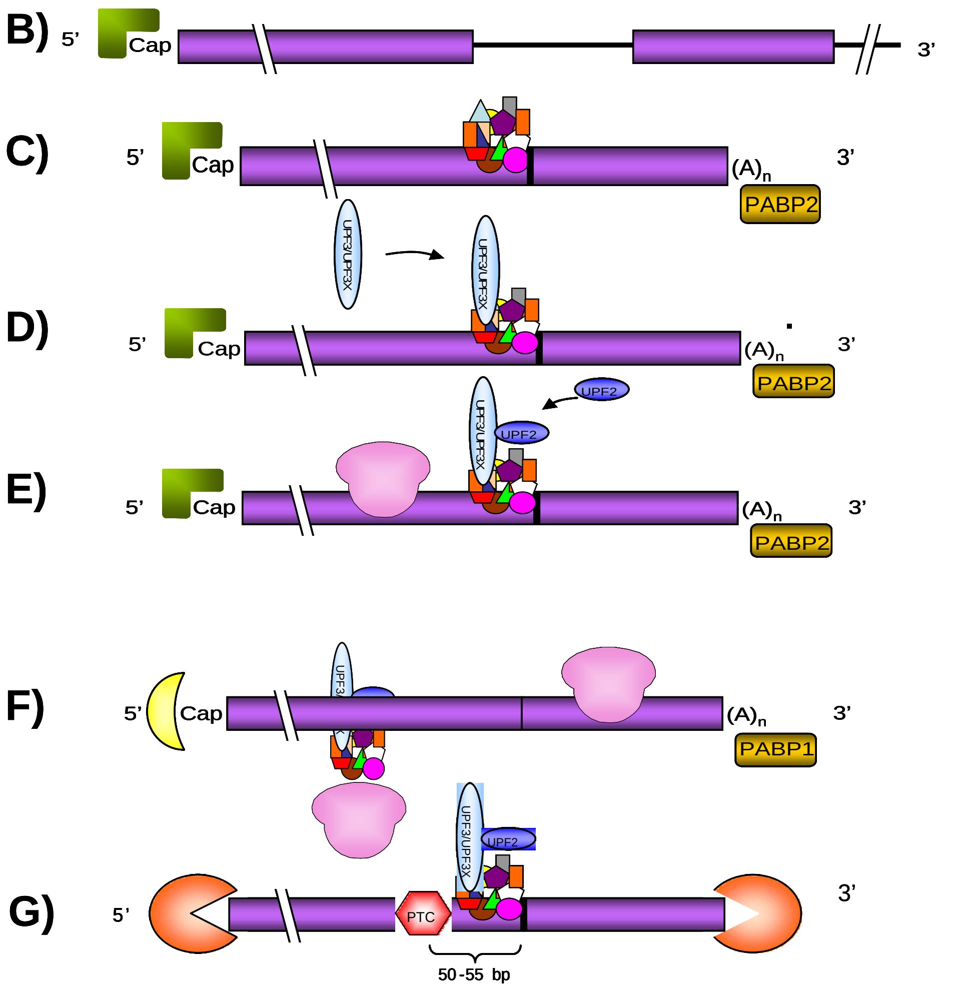
The biological importance of the NMD pathway is confirmed by its evolutionary conservation. Sequence homology for a defined set of proteins involved in the NMD pathway has been confirmed in a diverse range of eukaryotes such as yeast, *C. elegans,* and humans. Seven different proteins involved in NMD 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.