Chapter I Introduction

1.1 Introduction

The central tenet of molecular biology, first proposed by Francis Crick in 1957, describes how genes encoded by DNA sequences are copied (transcribed) to messenger RNAs (mRNA), which is then translated into functional proteins. This model became the basis of the colinearity theory, which states that the linear arrangement of subunits in the DNA sequence of a gene corresponds to the amino acid sequence of a protein. Determination of the entire genetic code (Khorana *et al.*, 1966; Nirenberg *et al.*, 1966) enabled prediction of protein sequences by translation of DNA sequences. Ten years later, techniques for rapid DNA sequencing were introduced (Maxam & Gilbert, 1977; Sanger *et al.*, 1977), which led to sequencing of large DNA molecules such as the 16.5 kilobase (kb) human mitochondrial genome (Anderson *et al.*, 1981) and the 40 kb genome of the Lambda bacteriophage (Sanger *et al.*, 1982). Since then, further development and high throughput automation of sequencing techniques has been accomplished and complete sequencing of large genomes is now possible, thus enabling researchers to study the fundamental genetic building blocks of life.

The human genome is the largest genome to be extensively sequenced so far. Preliminary analysis has confirmed that knowledge of the genome sequence will provide valuable insights into human biology. An important goal of current research is the generation of accurate annotation of all the genes encoded within the human genome (section 1.5); this gene index is expected to serve as a ‘periodic table’ for future genetic studies (Lander, 1996). Large-scale studies are being implemented to investigate the function of the genes and proteins identified from this research (section 1.6). Eventual integration of these studies should allow systematic dissection of the circuitry of the human body.

These advances in biological understanding have implications for research into human disease. The human genomic sequence in public databases allows rapid identification *in silico* of potential disease gene candidates and at least thirty disease genes have been identified in research efforts dependent on the genome sequence (Lander *et al.*, 2001). The genome sequence also provides insight into the mechanisms of chromosomal deletion, through homologous recombination and unequal crossing over between large, nearly identical intrachromosomal duplications. Such events are thought to be responsible for several syndromes, including the DiGeorge /velocardiofacial syndrome region on chromosome 22 (Shaikh *et al.*, 2000). Genomic research may lead to the development of new treatments for genetic disease, through the identification of new drug targets and a better understanding of disease mechanisms. Effective approaches to disease prevention may also be developed, as genetic predispositions to disease are recognised.

For the first time, the genomic landscape can be examined from a global perspective. Investigation of the distribution of features such as repetitive elements, GC content, CpG islands and recombination rates, are providing important clues about function and insight into the evolutionary history of the genome (Lander *et al.*, 2001). Comparative genomic data from model organisms also provides a powerful tool for analysis of the human genome, through identification of conserved functional features and novel innovations in different lineages.

This thesis describes the identification and accurate annotation of genes within a 3.4 megabase (Mb) region of human chromosome 22 (HSA22). The availability of the genomic sequence in this region enabled extensive sequence analysis of the gene environment. The utility of genomic sequence from the equivalent region of the mouse genome was explored in comparative analyses of potentially functionally conserved regions and investigation of chromosomal evolution. Finally, the experimentally verified transcript map was used as a basis for preliminary functional analyses of the protein coding genes encoded in this region, using both *in silico* techniques and an experimental approach to determine subcellular protein localisation. The next sections set out the background to the work reported in this thesis.

1.2 Mapping the human genome

1.2.1 Broad Features of the genome

The complete DNA sequence of a human is approximately 3200 Mb (Lander *et al.*, 2001; Morton, 1991). It is contained in 23 pairs of chromosomes: 22 autosomes and 2 sex chromosomes, X and Y. A basic classification of chromosomes is provided by the position of the centromere. In metacentric chromosomes, the centromere is roughly localised in the middle. Acrocentric chromosomes have the centromere close to one end and submetacentric centromeres are in-between these two positions. Chromosomes can be further distinguished by their banding patterns. A variety of treatments involving denaturation and/or enzymatic digestion of chromatin, followed by incorporation of a DNA specific dye, can cause mitotic chromosomes of complex organisms to appear as a series of transverse light and dark staining bands (Craig & Bickmore, 1993). Banding reflects variations in the longitudinal structure of chromatids, where each band differs from adjacent bands in base composition, time of replication, chromatin conformation and in the density of genes and repetitive sequences (see table 1.1). Such banding permits accurate differentiation of chromosomes – previously the only way of doing this was by examining the sizes of the chromosomes and the positions of the centromeres. Additionally chromosome banding allows more accurate definition of translocation breakpoints, subchromosomal deletions and other rearrangements.

Table 1.1: Properties of chromosome bands seen with standard Giemsa staining

Dark bands (G bands)	Pale bands (R bands)
Stain strongly with dyes that bind preferentially to AT-rich regions, such as Giemsa and Quinacrine	Stain weakly with Giemsa and Quinacrine
AT-rich	GC-rich
DNase insensitive	DNase sensitive
Condense early during the cell cycle but replicate late	Condense late during cell cycle but replicate early
Gene poor.	Gene rich
LINE rich, but poor in *Alu* repeats	LINE poor, but enriched in *Alu* repeats

(Adapted from Strachan and Read, 1999 and Lander *et al.*, 2001).

1.2.2 Genome maps

A more detailed delineation of the genome has been achieved by the production of various types of genome maps at increasingly fine scales. These maps have provided a framework of marker orders, established along the length of each chromosome. The frameworks, described briefly below, have been used to orientate and anchor the sequence-ready maps of overlapping cloned genomic segments during the human genome project.

1.2.2.1 Genetic maps

The aim of genetic mapping is to discover how often two loci are separated by meiotic recombination. The further apart two loci are on a chromosome, the more likely it is that a crossover will separate them. Thus the recombination fraction is a measure of the genetic distance between the two loci. Human genetic mapping required the development of genetic markers: Mendelian characters, which are sufficiently polymorphic to give a reasonable chance that a randomly selected person will be heterozygous.

The first human linkage map was published in 1987 (Donis-Keller *et al.*, 1987). The markers for this map were restriction fragment length polymorphisms (RFLPs) (Botstein *et al.*, 1980), whose use was soon replaced by the more informative, highly polymorphic microsatellite repeats (Litt & Luty, 1989; Tautz, 1989; Weber *et al.*, 1991). The microsatellite landmarks have been converted to sequence tagged sites (STSs), (Olson *et al.*, 1989), that can be assayed by the polymerase chain reaction (PCR) (Saiki *et al.*, 1988; Saiki *et al.*, 1985). These technical advances aided the construction of genetic maps at increasingly high resolution (Gyapay *et al.*, 1994; Hudson *et al.*, 1992; Weissenbach *et al.*, 1992), culminating in a 1 cM map (Dib *et al.*, 1996; Murray *et al.*, 1994). Efforts are currently focused on the generation of even more dense maps for the mapping of complex traits, using the most common type of DNA sequence variation: single nucleotide polymorphisms (SNPs).

1.2.2.2 Radiation Hybrid (RH) maps

The original approach by Goss and Harris (1975), where chromosome fragments generated by lethal irradiation of donor cells are rescued with suitable recipient cells, was applied to study whole genomes in 1994 (Walter *et al.*, 1994). The presence or absence of markers within a hybrid can be interpreted to produce a linear map order for the DNA clones (Cox, 1992). This is because the nearer two DNA sequences are on a chromosome, the lower the probability of separating them by the chance occurrence of a breakpoint between them (Cox *et al.*, 1990; Gyapay *et al.*, 1996; Walter *et al.*, 1994). RH mapping has been used to produce high-resolution gene maps by assaying the RH panels with genetic markers and RNA-derived expressed sequence tags (ESTs) by PCR. The ESTs are then ordered relative to the genetic markers (Deloukas *et al.*, 1998; Schuler *et al.*, 1996a). RH maps are also used to order and integrate all chromosome-specific markers to produce a framework map for the construction of bacterial clone maps (Bentley *et al.*, 2001; McPherson *et al.*, 2001; Montgomery *et al.*, 2001; Mungall *et al.*, 1996; Mungall *et al.*, 1997; Tilford *et al.*, 2001).

1.2.2.3 YAC maps

A primary goal of physical mapping is to assemble a comprehensive series of DNA clones with overlapping inserts (clone contigs). This became feasible for larger genomes with the development of yeast artificial chromosomes (YACs) (Burke *et al.*, 1987). The large insert sizes of up to 1500 kb (Chumakov *et al.*, 1995) allow long-range continuity. Many different YAC maps have been published (Bell *et al.*, 1995; Bouffard *et al.*, 1997; Chumakov *et al.*, 1992; Collins *et al.*, 1995; Doggett *et al.*, 1995; Foote *et al.*, 1992; Gemmill *et al.*, 1995; Gianfrancesco *et al.*, 1997; Hudson *et al.*, 1995; Krauter *et al.*, 1995; Nagaraja *et al.*, 1997). However, due to problems with chimaerism and instability (Green *et al.*, 1991; Nagaraja *et al.*, 1994),YACs are not a suitable substrate for sequencing.

1.2.2.4 Bacterial clone maps

Cosmid (Collins & Hohn, 1978) and fosmid (Kim *et al.*, 1992) libraries provided an alternative to YACs, but the disadvantage of these cloning systems is their small insert size (30-45 kb). The development of bacterial clone vectors, which could accommodate larger inserts (up to 200 kb), bacterial and P1 artificial chromosomes (BAC and PAC respectively) (Ioannou *et al.*, 1994; Shizuya *et al.*, 1992), resulted in a number of new clone libraries. These cloning systems are stable, due to the lower copy number replicons and have been shown to contain few rearrangements (Ioannou *et al.*, 1994; Shizuya *et al.*, 1992). For these reasons, these types of library were the chosen resource for sequence ready map construction.

1.2.2.5 Sequence-ready maps and sequencing

Different strategies have been used to construct the sequence ready bacterial clone maps. The Sanger Institute and Washington University Genome Sequencing Center (Lander *et al.*, 2001; McPherson *et al.*, 2001) favour a map-based, hierarchical shotgun method. STSs from previously constructed genetic and physical maps were used to recover BACs and PACs from specific regions. The clones are then assembled into contigs by landmark content mapping (Green *et al.*, 1991) and restriction enzyme fingerprint analysis (Gregory *et al.*, 1997; Marra *et al.*, 1997; Olson *et al.*, 1986) (see chapter IV). A sequence tile path, minimising redundancy from overlapping clones, is then selected for sequencing.

Selected clones were sequenced using a shotgun approach. The cloned genomic insert is fragmented and the 1.4 – 2.2 kb fragments cloned into M13 or plasmid vectors (Bankier *et al.*, 1987). The subclones are then sequenced using the chain termination method (Sanger *et al.*, 1977). This method has been adapted to use two types of fluorescent chemistries: dye labelled primers and terminators (Lee *et al.*, 1992; Prober *et al.*, 1987; Smith *et al.*, 1987). The sequence reads obtained are assembled into contigs, after which a directed approach is used both manually and automatically to edit the sequence. Additional sequence to close any gaps and resolve problems is obtained during ‘finishing’.

An alternative whole genome shotgun (WGS) method was utilised by the biotechnology company Celera Genomics, to produce a second version of the human genome (Venter *et al.*, 2001). Human clone libraries of prescribed insert length were produced from the DNA of five individuals. The ends of clone inserts were sequenced (paired end sequences or mate pairs), generating sequence reads amounting to 5.11-fold coverage of the genome. Sequence generated by the public effort, freely available in public databases, was also used to bring the effective coverage to 8-fold (Venter *et al.*, 2001). Two assembly strategies – a whole-genome assembly and a regional chromosome assembly - were used, each combining sequence data from Celera and the publicly-funded genome effort. Known repeat elements were screened out from the assembly process before sequence overlaps were identified and checked for the presence of repeated elements not removed in the initial screen. Gaps between the assembled contigs could be sized and the contigs orientated, using the mate pair information of sequence reads from opposite ends of the same clone insert. The two assembly strategies yielded very similar results that largely agree with the independent mapping data (Venter *et al.*, 2001).

1.2.2.6 Human genome draft sequence

The public domain sequencing centres published a first draft of the human genome sequence in February 2001. This was generated from a physical map covering more than 96% of the euchromatic part of the human genome and, together with additional sequence in public databases, it covers about 94% of the human genome. A final, accurate draft is promised by 2003 (Lander *et al.*, 2001). On the same day in February, the Celera venture published the second version of the human genome in a rival journal (Venter *et al.*, 2001).

A computational comparison of the two draft sequences (Aach *et al.*, 2001) found that they were overall similar in size, containing comparable numbers of unique sequences and exhibited similar statistics for sample candidate DNA protein-binding motifs. Some differences emerged at a detailed level e.g. contigs in each exhibited different size and gap distributions. However, these differences are expected to diminish as assemblies become more complete and comprehensive.

1.3 Large-scale features of the genome sequence

The availability of the draft genome sequence allows systematic genome wide analysis of the human genome. Analysis has confirmed a variety of large-scale features of the genomic landscape (Lander *et al.*, 2001).

1.3.1 Distribution of GC content

On average the genome is 41% GC but the distribution of base composition varies from 38% to over 55% GC (Lander *et al.*, 2001). Previous studies have indicated that GC-rich and GC-poor regions have different biological properties, such as gene density, composition of repeat sequences and correspondence with cytogenetic bands (Duret *et al.*, 1995; Gardiner, 1996; Hurst & Eyre-Walker, 2000; Saccone *et al.*, 1992; Saccone *et al.*, 1993; Zoubak *et al.*, 1996).

Bernardi and colleagues (1985) proposed that the variation in GC content could reflect that the genome is composed of isochores – local regions of similar GC content. By randomly shearing DNA and fractionating it on CsSO4 gradients, five fractions were identified: two light AT rich fractions L1 and L2 and three increasingly GC rich fractions H1, H2 and H3. The L1 and L2 fractions comprise 62% of the genome, H1 22%, H2 9% and H3 3-4%. The remaining 3-4% consists of satellite and ribosomal DNA. This division was further extended by Saccone *et al.*,(1996) and the H3 isochore was split up into three increasingly GC-rich sub-fractions: $H3^-$, H3* and $H3^+$. Hybridisation *in situ* of the H3 fractions indicates the positions of the most gene-rich bands.

The draft genome sequence was analysed to see if the existence of strict isochores could be verified. However, the average GC content for a variety of different window sizes showed too much variation to be consistent with a homogeneous distribution. Although the genome clearly contains large regions of distinctive GC content, Lander *et al*. (2001) concluded that there is substantial variation at many different scales. However, the existence of isochores in the human genome has been supported through the use of the different, window-less approach of recursive segmentation (Li, 2001). A segmentation point is identified that maximises the base composition difference between the left and right subsequences. Each subsequence is then subdivided into two further subsequences in the same manner, until the resulting domains satisfy a previously determined threshold value. Li proposes that a window-approach may not be able to delineate the borders of relative homogeneous domains accurately enough before carrying out a homogeneity test. The alternative recursive segmentation approach, however, supports the existence of isochores in the human genome.

1.3.2 CpG islands

The CpG dinucleotide occurs at about one fifth of the roughly four percent frequency that would be expected by multiplying the typical fractions of Cs and Gs (0.21x0.21) (Matsuo *et al.*, 1993). The shortfall occurs because CpG dinucleotides are often methylated on the cytosine base and spontaneous deamination of methyl-C residues gives rise to T residues (spontaneous deamination of ordinary cytosine residues gives rise to uracil residues that are readily recognised and repaired by the cell) (Coulondre *et al.*, 1978; Sved & Bird, 1990). However, the genome contains many CpG islands in which the CpG dinucleotides are not methylated and occur at a frequency closer to that predicted by local GC content. One feature of these islands is that they are rich in sites for methyl-sensitive restriction enzymes such as *Hpa*II, which recognise unmethylated CpG dinucleotides (Bird, 1986).

Using the definition proposed by Gardiner-Garda and Frommer (1987), a search of the repeat masked draft genome sequence highlighted 28,890 possible CpG islands. CpG islands are of particular interest, because many are associated with the 5’ ends of genes (Bird *et al.*, 1985; Bird, 1986; Chan *et al.*, 2000) and may also contain promoter sequences (Cross *et al.*, 2000). Analysis of the draft genome sequence showed that the relative density of CpG islands correlated reasonably well with estimates of relative gene density on chromosomes.

1.4 Coding and non-coding sequence

An important distinction that can be made between the different compartments of the genome is coding versus non-coding sequence. The function of non-coding DNA remains to be fully understood (Gardiner, 1995). There are a number of features that distinguish these two fractions of the genome, which are discussed below.

1.4.1 Non-coding features

Genomes can contain a large quantity of repetitive sequence, far in excess of that devoted to protein-coding genes. Analysis of the draft human genome sequence showed that repeats account for at least 50% of the genome (Lander *et al.*, 2001). Several different classes of repeats have been described.

1.4.1.1 Transposon-derived repeats

About 45% of the genome sequence consists of repeats derived from one of four types of transposable element, of which three transpose through RNA intermediates (LINEs, SINEs and LTR retrotransposons) and one transposes as DNA (DNA transposons).

In humans, full length LINEs are about 6 kb long, contain an internal polymerase II promoter and encode two ORFs. Three LINE families, LINE1, LINE2 and LINE3, are found in the human genome; only LINE1 is still active. The transcribed LINE RNA and translated proteins move to the nucleus, where an encoded endonuclease activity makes a 3' single stranded nick from which the reverse transcriptase is primed. This frequently fails to proceed to the 5' end, resulting in many truncated, non-functional insertions. The LINE machinery is believed to be responsible for most reverse transcription in the genome, including SINE retrotransposition.

SINEs are on average between 100bp and 400bp long, harbour an internal polymerase III promoter but encode no proteins. They are thought to use the LINE machinery for transposition and have been noted to share the 3' end with LINE elements (Okada & Hamada, 1997). The human genome contains three families of SINEs: the active *Alu* and the inactive MIR and Ther2/MIR3.

LTR retrotransposons are flanked by long terminal direct repeats, which contain all of the necessary transcriptional regulatory elements. Mammalian retroviruses fall into three classes (I-III), each comprising of many families. Homologous recombination between flanking LTRs can result in loss of the internal sequence.

DNA transposons have terminal inverted repeats and encode a transposase enzyme. The human genome contains at least seven major classes of DNA transposon, each containing many families. Transposons have been indirectly responsible for many evolutionary innovations in the genome. Over forty human genes have been recognised as probably derived from transposons (Jurka & Kapitonov, 1999; Lander *et al.*, 2001; Smit, 1999).

LINE1 activity can also bring about exon reshuffling by co-transcription of neighbouring DNA. They can also cause reverse transcription of mRNA, which typically results in non-functional processed pseudogenes, but can occasionally give rise to functional processed genes. There are at least eight human genes that may be derived from this origin (Brosius, 1999).

1.4.1.2 Simple Sequence Repeats (SSRs)

Human satellite DNA is comprised of very large arrays of tandemly repeated DNA, often from 100 kb to several megabases in length. The repeat unit can range from 5 base pairs (bp) in length to over 170 bp (centromeric alphoid DNA). Repeated DNA of this type makes up the bulk of the heterochromatic genome regions, approximately 5-10% of the total sequence.

SSRs with a short repeat unit (n = 1-13 bases), often spanning less than 150 bp in total, are termed microsatellites, whilst these with longer repeat units (n = 14 – 500 bases) and spanning within a range of ~0.1 – 2.0 kb, are termed minisatellites. Slippage during DNA replication is thought to result in the production of SSRs (Kruglyak *et al.*, 1998; Toth *et al.*, 2000). SSRs comprise about 3% of the euchromatic human genome, with the greatest single contribution coming from dinucleotide repeats (0.5%) (Lander *et al.*, 2001).

SSRs have been used in human genetic studies (section 1.2.2.1). The microsatellites and especially the expansion of triplet repeats have also been implicated in neurodegenerative disorders. Since the cause of fragile X was shown to be repeat expansion (Yu *et al.*, 1991; Verkerk *et al.*, 1991; Kremer *et al.*, 1991) the list of diseases caused by repeat expansion has continued to grow. Triplet repeat expansions are associated with non-B DNA structures: these structures may account for the expansion and instability and therefore the disease-causing feature of the triplet repeats (Sinden, 1999). Interestingly, one of the genes analysed in this thesis (E46L) has been implicated in the causation of spinocerebellar ataxia 10, through polymorphism of an unstable pentanucleotide repeat in intron IX (Matsuura *et al.*, 2000).

1.4.1.3 Segmental duplications

Analysis of the draft sequence shows that the human genome seems likely to consist of about 5% segmental duplication. Intrachromosomal duplications occur within a particular chromosome. Interchromosomal duplications are defined as segments that are duplicated among non-homologous chromosomes. Regions near the centromere and telomeres are composed almost entirely of interchromosomal duplicated segments. It is hypothesised that chromosomal breakage products are preferentially inserted here by an unknown mechanism, in order to limit possible damage caused by insertion into more gene-rich regions (Lander *et al.*, 2001).

1.4.2 Coding genome features

1.4.2.1 Non-coding RNA (ncRNA) genes

Less than 5% of the human genome is thought to encode genes (Lander *et al.*, 2001). The majority of human genes ultimately specify polypeptides that carry out numerous diverse functions. However, a smaller minority instead specify a mature RNA product. In addition to the many genes involved in protein synthesis (rRNA genes, tRNA genes), there are other RNA genes that process and modify rRNA in the nucleolus (snoRNAs), spliceosomal RNAs and other ncRNA genes such as telomerase RNA and the 7S signal recognition particle RNAs. ncRNAs do not have translated open reading frames (ORFs), are often small and are not polyadenylated. Accordingly, novel ncRNAs are hard to find by experimental sequencing, but attempts are being made using computational techniques that exploit their secondary structural characteristics (Rivas *et al.*, 2001).

1.4.2.2 Protein coding genes

Human protein coding genes have a complex structure (figure 1.1). Whereas in simple organisms, such as yeast, the genes are simply single ORFs, in complex organisms, such as human, the ORF is segmented with the protein-coding exons being separated by introns. Nuclear pre-mRNA introns are excised from the primary transcript by a large ribonucleoprotein complex, known as the spliceosome (reviewed in Moore & Sharp, 1993), which recognises sites at the 5' and 3' ends of the intron (the donor and acceptor sites respectively) as well as an internal site known as the branch point. With a few exceptions (Sharp & Burge, 1997) nearly all spliceosomal introns begin with GT and end with AG.

Protein coding genes contain a translational start site (usually ATG), often contained in an optimal consensus sequence (Kozak, 1987). Some genes also contain a polyadenylation signal, most commonly an AATAAA hexamer sequence followed by a more complex signal (not yet completely characterised) located 20-30bp downstream (Beaudoing *et al.*, 2000; Gautheret *et al.*, 1998). Less is known about the identity of regulatory sequences that could be present in the 5' and 3' Untranslated regions (UTRs) and introns (section 1.5.2).

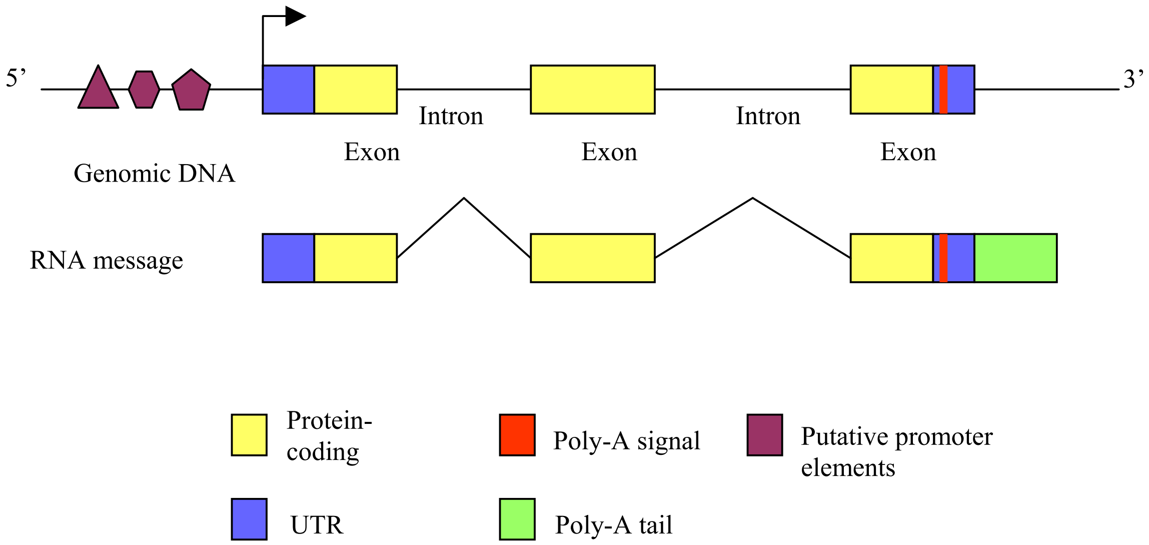

Figure 1.1: Protein-coding gene features. The genomic layout is shown at the top of the figure and the transcribed RNA message below. The colours identify the different features discussed in the text.

1.5 Gene Identification

Genes represent the major functional elements of the genome and are thus the main focus of interest of genome researchers. In principle, three major features permit the DNA of genes to be distinguished from DNA that does not have a coding function:

1. Expression: all active genes are capable of making an RNA product, usually mRNA. Mammalian genes usually contain introns, so the initial RNA transcript undergoes splicing.
2. Sequence conservation: because genes execute important cellular functions, mutations that alter the sequence of the product will often be deleterious and eliminated by natural selection. The sequence of coding DNA and important regulatory sequences is therefore more strongly conserved in evolution than that of non-coding DNA.
3. CpG islands: many vertebrate genes are associated with CpG islands (Bird *et al.*, 1995). Identification of these sites can aid identification of the adjacent gene.

1.5.1 Traditional approaches

Several techniques have been developed that rely on sequence conservation to find genes. For example, a zoo blot (Monaco *et al.*, 1986) involves the hybridisation of a DNA clone to a Southern blot of genomic DNA samples from a variety of animal species. Conserved sequences, which are likely to be genes, are thus identified.

CpG islands usually contain multiple rare-cutter restriction sites (Cross & Bird, 1995). These can be identified by restriction mapping (DNA clones are hybridised against Southern blots of genomic DNA, cut with *Sac*II, *Eag*I or *Bss*HII, to identify clustering of rare cutter sites) (Sargent & Bennett, 1986) or by island-rescue PCR (PCR amplification between islands and neighbouring *Alu* repeats) (Valdes *et al.*, 1994). The identified fragments can be tested for expression by hybridisation to Northern blots containing RNA isolated from a range of different tissues. If a transcript is identified, the corresponding complementary DNA (cDNA) can be isolated from the appropriate library. Alternatively, entire genomic clones can be hybridised against a Northern blot or against appropriate cDNA libraries

More efficient approaches can be used to construct a transcript map of a large region (Gardiner & Mural, 1995). Exon trapping uses a functional assay for splice sites in genomic DNA. The DNA is shotgun cloned into a vector containing a functional splice donor site, an intervening sequence and a selectable marker (Buckler *et al.*, 1991; Duyk *et al.*, 1990). This method has been used to identify the genes for a number of diseases (Trofatter *et al.*, 1993; Vulpe *et al.*, 1993). The technique has also been used to isolate exons from entire chromosomes (Church *et al.*, 1993; Trofatter *et al.*, 1995).

cDNA selection or capture involves repeated purification of a subset of cDNAs that hybridise to a given genomic region. cDNAs that hybridise specifically to genomic fragments immobilised on nylon membranes can be eluted and amplified by PCR. The process is repeated two or three times before the eluted cDNAs are cloned. This results in highly specific and enriched sub libraries for expressed sequences of the genomic region (Lovett *et al.*, 1991; Parimoo *et al.*, 1991). These methods have been improved by using biotin–labelled genomic DNA and streptavidin-coated magnetic beads to capture the genomic DNA-cDNA hybrids (Korn *et al.*, 1992; Morgan *et al.*, 1992). This approach has also been used to generate specific chromosome-enriched libraries (Touchman *et al.*, 1997).

1.5.2 Post-genomic era

The availability of large amounts of genomic sequence, defining the 'post-genomic era', facilitates sequence analysis as a method for gene identification. In small prokaryotic genomes, finding the encoded genes is largely a matter of identifying all the long open reading frames. Ambiguities arise if long ORFs overlap on opposite strands – the true coding region must then be investigated. Genes are found using a computer program that carries out six-frame translation, identifying ORFs longer than a chosen threshold (such as 500 bp (Burge & Karlin, 1998)). However, smaller genes could be missed.

Finding genes in eukaryotes becomes considerably harder as the signal:noise ratio increases. For example, the 8 Mb prokaryotic genome of *H. influenzae* contains 85% coding sequence, whereas more complex eukaryotic genomes, such as those of the fly and worm, are less than 25% coding. The human genome contains an estimated 3% coding sequence (Duret *et al.*, 1995), recently confirmed for chromosome 22 (Dunham *et al.*, 1999). Gene annotation in these more complex organisms is complicated further by splicing and alternative splicing. The arrangement of genes in genomes is also prone to exceptions. Although usually separated with an intergenic region, there are examples of genes nested within each other (Dunham *et al.*, 1999); that is, one gene located in an intron of another gene or overlapping genes on the same (Ashburner *et al.*, 1999; Schulz & Butler, 1989) or opposite (Cooper *et al.*, 1998) DNA strands. The presence of pseudogenes (non-functional sequences resembling real genes), which are distributed in numerous copies throughout the genome, further complicates the identification of true protein coding genes. Current approaches to gene identification approaches include computer prediction packages and homology searches.

1.5.2.1 *Ab initio* prediction packages

The most natural way to find genes computationally would be to mimic as closely as possible the processes of transcription and RNA processing (e.g. splicing and polyadenylation) that define genes biologically. A number of important signals related to transcription, translation and splicing are now sufficiently well characterised as to be useful in computer predictions of the location and exon-intron organisation of genes. The genomic elements that researchers seek include splice sites, start and stop codons, branching points, promoters and terminators of transcription, polyadenylation sites, ribosomal binding sites, topoisomerase II binding sites, topoisomerase I cleavage sites and various transcription factor binding sites (reviewed by Gelfand, 1995). These conserved elements are used by *ab initio* prediction programs: gene prediction from sequence data without the use or prior knowledge about similarities to other genes. *Ab initio* gene prediction programs are discussed in more detail in chapter III.

1.5.2.2 Sequence similarity

The similarity of a region of the genome to a sequence that is already known to be transcribed, is a powerful predictor of whether or not a sequence is part of a gene. Similarity-based methods rely on matches to DNA and protein databases with the genomic sequence under investigation using, for instance, the Basic Local Alignment Search Tool (BLAST) (Altschul *et al.*, 1997). This type of search has become very powerful due to increased EST availability (Adams *et al.*, 1991; Wilcox *et al.*, 1991) (section 1.8).

Although the ESTs generally cover only a segment of the gene, their utility for gene identification was immediately recognized. In a pilot project, Adams *et al.*(1991) performed automated partial sequencing of more than 600 randomly selected cDNAs from human brain. Of the generated ESTs, 337 represented new genes and 48 had significant similarity to genes from other organisms. Since then, a large number of publications have generated increasing sets of ESTs and their analysis (Adams *et al.*, 1993; Hillier *et al.*, 1996; Houlgatte *et al.*, 1995; Khan *et al.*, 1992; Okubo *et al.*, 1992). All public domain ESTs are deposited in dbEST, a subdivision of GenBank/EMBL/DDBJ (Boguski *et al.*, 1993). An important key development for the widespread use of ESTs, was the formation of the IMAGE consortium (Lennon *et al.*, 1996) (http://bbrp.llnl.gov/bbrp/image/), to ensure that collections of clones as well as sequences would be accessible by the biomedical research community.

A large amount of redundancy exists in the large EST collections, owing to repeated sequencing of the same widely expressed genes in different or the same tissues. Different clustering methods have been devised to address the redundancy. Examples include Unigene (Boguski & Schuler, 1995) and the GeneExpress project (Houlgatte *et al.*, 1995).

There are many applications for which partial sequences are not adequate. For example, accurately predicting the function or structure of a gene product or isolating the protein product, requires a full-length sequence. The Mammalian Gene Collection (MGC) project represents an ongoing effort by the National Institute of Health (NIH) to generate a full-length cDNA resource, eventually representing all human genes (Strausberg *et al.*, 1999).

Protein databases provide a further resource for gene annotation (section 1.8). Notable examples include the SwissProt database (Bairoch & Apweiler, 2000), which is a database of protein sequences derived from translations of DNA sequences from the EMBL nucleotide sequence database, adapted from the Protein Identification Resource (PIR) collection, extracted from the literature and directly submitted by researchers. SwissProt is a curated database, containing high-quality annotation, is non-redundant, and cross-referenced to several other databases. TrEMBL is a computer-annotated protein sequence database, which supplements SwissProt. TrEMBL contains the translations of all coding sequences (CDS) present in the EMBL nucleotide sequence database not yet integrated into SwissProt. TrEMBL can be considered as a preliminary section of SwissProt. Annotation of genes through similarity searches is discussed in chapter III.

1.5.3 Comparative studies

Conserved genetic linkage groups have been documented in a variety of vertebrate species (for a summary, see Jones *et al.*, 1997). Genomic sequence from a range of species is now becoming available (section 1.7) and a wide variety of cross-species comparative studies are being undertaken to identify conserved and novel functional features, to elucidate the mechanisms that have acted during genome evolution and to study gene and protein function in model systems.

1.5.3.1 Identification of conserved functional regions

In order to exploit genomic comparison as an analytical tool for identification of functional regions, species must be selected of great enough evolutionary divergence to permit identification of functionally conserved regions from the rest of the genomic background, yet small enough that comparison of conserved syntenic lineage is meaningful (Lundin, 1993). Such comparisons can allow identification of genes and possible regulatory regions in both genomes with no previous knowledge of the gene content of either.

BLAST (Altschul *et al.*, 1997) is one of a range of alignment algorithms that can also be used to compare genomic sequences with homologous genomic sequences from closely related organisms such as mouse, chicken or pufferfish. For example, the ‘Exofish’ algorithm (Crollius *et al.*, 2000), utilises a specific implementation of BLAST (TBLASTX) to conduct homology searches of human sequence with available *T. nigroviridis* sequence. Exofish has already proved useful in the identification of human genes and has demonstrated the potential of comparative genomics using the pufferfish genome. Additional algorithms are being developed for the specific purpose of species sequence comparison and gene identification.

The benefits of using mouse genomic sequence for identification of gene and regulatory regions have been illustrated by a large number of small-scale studies and is reviewed in more detail in chapter IV. Additionally, genome-wide alignments of human and mouse sequence are becoming available from the public genome project (Meisler, 2001). As more finished mouse sequence is added, this resource will aid identification of genes and candidate regulatory regions within the human genome.

1.5.3.2 Evolutionary chromosomal rearrangements

Chromosomal rearrangements such as inversions and translocations have played an important role in defining genome organisation in existing mammals. The number of rearrangements that have occurred since divergence from the primordial mammal has been modest and the distribution of these rearrangements among chromosomes appears random (Eppig & Nadeau, 1995). As a result, each mammalian species has a unique arrangement of conserved and disrupted chromosomal segments as compared to other mammalian species. Genes provide excellent markers for these chromosomal segments as their homologies can be detected in highly divergent species (Eppig & Nadeau, 1995; Nadeau & Sankoff, 1998).

The mouse genome represents the most thoroughly studied of all non-human vertebrate genomes. However, rodent chromosomes have undergone an unusually high number of genomic rearrangements per unit of evolutionary time (Graves, 1996). Nevertheless, the degree to which gene content and order is conserved is considerable (Carver & Stubbs, 1997). As the resolution of the physical maps of the human and mouse genomes increases from cytogenetic bands to nucleotide sequence (section 1.2, chapter IV), breakpoints in the comparative map can be mapped more precisely and their characteristics examined. Chromosomal painting by fluorescence *in situ* hybridisation (FISH) with chromosome-specific libraries is an easy way to compare the location of homologous chromosomes in different mammalian species (for example Rettenberger *et al.*, 1995; Scherthan *et al.*, 1994). This method is used to identify the location and approximate size of homologous segments and to estimate the number of rearrangements that have occurred since the divergence of the lineages leading up to the species being compared. Finer mapping of a syntenic breakpoint was provided by the first sequence-level analysis of a conserved synteny breakpoint between mouse chromosome 10 (MMU10) and HSA21 and HSA22, recently described by Pletcher *et al*.(2000). Examination of the structural features of this segment, and comparison with other breakpoints, should provide insight as to whether particular DNA sequences contain structural features that are predisposed to ancestral chromosomal rearrangements (chapter IV).

Data from comparative maps is also used in functional studies to identify candidate disease genes and to characterise the genetic basis for complex traits (section 1.7).

1.6 Functional genomics

The initial interest in the human genome was precipitated by a desire to identify the cause of observable gene phenotypes: in 1986, Dulbecco stated the wish to 'sequence the whole genome of a whole animal species for the purpose of finding genes involved in cancer'. However, even once the entire complement of genes is established, the function for most of them will remain unknown (Blackstock & Weir, 1999). The emerging field of functional genomics is addressing these problems. This not only involves the determination of gene function, but also the determination of expression patterns and both spatial and temporal analysis of the proteins. Functional genomics is characterised by high throughput or large-scale experimental methodologies combined with statistical and computational analysis of the results. The underlying strategy is to expand the scope of biological investigation from studying single genes or proteins to studying all genes and proteins at once in a systematic fashion. Computational biology will perform a critical and expanding role in this area (Hieter & Boguski, 1997).

1.6.1 Expression studies

The first small-scale approaches to identifying and cloning differentially expressed genes were primarily based on subtractive hybridisation (Hess *et al.*, 1998). Although several genes were cloned using this method, subtractive hybridisation turned out to have some crucial disadvantages: it reveals only a small fraction of the overall changes in gene expression; it requires large amounts of RNA; and it is time-consuming and laborious. In 1992, differential display PCR (DD-PCR) was introduced to compare, identify and isolate differentially expressed transcripts (Liang & Pardee, 1992). In principle, the method utilises reverse transcription (RT)-PCR amplification of two different mRNA populations and separate the resulting fragments side-by-side on a denaturing polyacrylamide gel. Bands expressed at different levels are isolated and cloned.

Large-scale gene expression studies have been revolutionised by microarray technology. This takes advantage of the fact that increasingly complete sets of cDNA clones and sequences representing all human and mouse genes are becoming available for high throughput surveys of gene expression. DNA microarrays consist of genes, gene fragments or oligonucleotides covalently attached in a high-density array on a glass slide. The DNA on the array is selected from databases such as Unigene (Boguski & Schuler, 1995). Arrays can also be produced using photolithography to synthesise specific oligonucleotides *in situ* on the array (Fodor *et al.*, 1991).

The arrays can be used to record differences in expression between a reference and a test RNA population. Each RNA population is used as a template in a reverse transcription PCR reaction, incorporating a distinguishing fluorescently labelled dinucleotide. The fluorescently labelled cDNAs are then hybridised to the array. The relative fluorescence intensity measured for the two different fluors at each array element provides a reliable measure of the relative abundance of the corresponding mRNA in the two RNA populations. This technique has been used to assay expression in inflammatory disease (Heller *et al.*, 1997), the diauxic shift in *S. cerevisiae* (DeRisi *et al.*, 1997) and tumorigenic versus non-tumorigenic cell lines. Microarrays have also been used in other techniques to assay expression patterns (section 1.6.3.2).

Serial Analysis of Gene Expression (SAGE) also takes advantage of the possibility of a human gene index. SAGE is based on short nucleotide tags of 9 to 10 bp, derived from the complete mRNA pool of a cell population. These tags contain enough information to identify a transcript. Concatenation of short tags allows the simultaneous analysis of multiple transcripts by sequencing of many tags (10-50) within a single clone. The advantage of this method is that it is possible to count the number of distinct mRNA molecules in a given cell population for each condition and, from this accounting, a particular mRNA would be described as differentially expressed if its frequency is significantly greater in one condition versus another.

A current limitation of microarray and SAGE technologies is that there is not yet a comprehensive and accurate index of human genes. cDNAs representing each gene also need to be collected both for the human genome, and for all other species of interest. The second limitation is the need for sufficiently powerful mathematical and visualisation tools for whole-genome expression studies to analyse the mass of new data. This is currently one of the major challenges faced by bioinformatics (section 1.8).

1.6.2 Control of gene expression

Our understanding of how regulatory information is encoded by a DNA sequence is still very fragmented. Large-scale genome sequencing projects currently determine hundreds of megabases every year. Thus, one of the major challenges that biologists face is to identify the regulatory elements within the bulk genome.

Wet laboratory approaches to the identification of regulatory regions includes the use of reporter genes and deletion analyses to assess how deletion of different segments of DNA upstream of a gene, or occasionally on the first intron, affects gene expression. Gel retardation, DNAse footprinting and methylation interference assays can identify protein-binding sites on a DNA molecule. However, the amount of experimental work that would be required to systematically analyse these non-coding sequences could exceed current research capacity. There is therefore a need for experimental and computational tools that identify potential regulatory elements more quickly to allow focusing of experimental design.

Two main kinds of computational approaches can be distinguished. The first one includes methods that rely on biological knowledge of transcription factor binding sites, promoters, enhancers, locus control regions etc. to set up rules to predict regulator elements. However, the major obstacle to this approach is that the sequence motifs corresponding to these features contain too little diagnostic information for them to be distinguished from chance occurrences (Audic & Claverie, 1998). Experimentalists have shown that transcription factors, facing the same problem in the cell, find their physiological targets mainly via interactions with other factors bound to neighbouring sites. Chromatin structure modulated accessibility may also play a role. Computational biologists now agree that predictive algorithms should do the same thing: using sequence contextual features (e.g. predicted neighbouring elements) in order to distinguish between functional and biologically irrelevant sites.

The second type of approach relies on comparative analysis of homologous sequences. This approach has recently gained considerable interest, thanks to projects intended to sequence large regions of model vertebrate genomes (section 1.7). Tagle *et al.*(1988) proposed the term “phylogenetic footprinting” to describe the phylogenetic comparisons that reveal evolutionary conserved functional elements in homologous genes. However, regulatory elements that have been acquired very recently in evolution may not be detectable by this method. In addition, the conserved feature may not reside in the primary sequence structure, but rather in the spatial structure or a compositional property of the DNA or RNA that is subject to selective pressure.

As well as computational approaches, laboratory protocols are also being developed for the large-scale identification of putative regulatory regions. Frazer *et al*.(2001) recently described the most extensive human/mouse comparison available to date, through hybridisation of orthologous BAC contigs to an oligonucleotide array representing four 25-mers for each nucleotide in 16.6 Mb of non-repetitive DNA from human chromosome 21. The sequence of conserved elements could be determined from the hybridisation pattern. In the human-mouse comparison, 3400 conserved elements ranging in length from 30 bp to > 1 kb were identified, corresponding to 1.6% of the tested sequence. Only 44% of the conserved elements corresponded to known exons. 2.6 Mb of orthologous dog DNA was also hybridised to the oligo arrays, in order to estimate the proportion of conserved sequence resulting solely from common origin in the absence of active selection for function. Only half of the human-mouse noncoding elements were conserved in the dog sequence, indicating that it is worthwhile to extend comparisons beyond two species before initiating functional tests of putative regulatory elements.

1.6.3 Proteomics

The methods described above are critically important for a detailed understanding of the regulation of biological systems; however, such methods provide no information about post-translational control of gene expression. Indeed, experimental evidence suggests that there is no obvious correlation between mRNA expression levels and protein levels either in human liver cells (Anderson & Seilhamer, 1997) or in yeast (Gygi *et al.*, 1999). An emerging field for the analysis of biological systems is therefore the study of the complete protein complement of the genome, the ‘proteome’.

1.6.3.1 Two-dimensional gel electrophoresis and mass spectroscopy

One of the most widely used tools for proteome analysis is two-dimensional protein electrophoresis (2.G.E) (O'Farrell, 1975). This technique resolves complex mixtures of proteins first by isoelectric point and then by size. The resulting gel images form a two-dimensional ‘barcode’ of the proteome of a particular biological system. A comparison of two or more such barcodes might help to identify differences in protein expression that result in particular phenotypes. A protein spot of interest can also be purified and further analysed (by direct amino-acid sequencing, amino acid analysis and mass spectroscopy) to relate the protein to the underlying gene.

The need to characterise spots has fostered an increasing use of mass spectroscopy (MS) for protein characterisation. The two most commonly used approaches for spot characterisation involve peptide mass mapping and tandem MS of a proteolytic digest of a 2.G.E spot. The masses of resulting peptides from a proteolytic digest can be measured using MS. These masses can be compared with *in silico* digests of protein databases or six-frame translations of nucleic acid databases to help characterise the spot. In a tandem MS experiment, peptide mixtures are studied in an initial MS scan and particular peptides can be fragmented during a second step to generate amino acid sequence information. The sequence information is derived by an attempt to match mass spectra from fragmentation patterns with *in silico* spectra, obtained from databases, or by matching amino acid sequence information with available databases.

1.6.3.2 Chip-based methods for proteome analysis

On array-based methods, protein spots are immobilised onto glass slides. Such arrays can then be used to screen complex protein mixtures for particular binding affinities or other interactions (Walter *et al.*, 2000). Antibodies can be arrayed using bacteria that express recombinant antibodies. These arrays can be probed for specific antibody-antigen binding interactions, using a filter-based enzyme-linked immunosorbent assay (ELISA) technique.

Ziauddin and Sabatini (2001) have produced microarrays of cells expressing defined cDNAs. Arrays are printed with sets of cDNAs cloned in expression vectors. Mammalian cells are cultured on the glass slide and cells growing on the printed areas take up the DNA, creating spots of localised transfection within a lawn of non-transfected cells. Two uses for this approach have been identified so far: as an alternative to protein microarrays for the identification of drug targets and as an expression cloning system for the discovery of gene products that alter cellular physiology.

A commercial device, the ProteinChipSystem (Senior, 1999), combines chip-based techniques with MS to selectively capture proteins from biological systems using surface-enhanced laser desorption and ionisation (SELDI) technology. Protein mixtures are incubated with a variety of available chips that probe Lewis acid/base interactions or hydrophobic, electrostatic and coordinate covalent bonding. The surfaces of these chips are precoded with chromatographic affinity surfaces that extract, structurally modify or amplify a particular protein. ProteinChipArrays have been used to identify disease markers (Xiao *et al.*, 2000). For example, prostate-specific membrane antigen, a protein thought to indicate prostate cancer tumour progression, can be detected in blood sera and quantified, based on a normalised peak, via ProteinChip technology.

The ultimate aim of functional genomics is to integrate information from various ‘levels’ including DNA sequence, mRNA profiles, protein expression and metabolite concentrations, as well as information about dynamic spatio-temporal changes in these molecules to form effective models of biological system. Attempts have already been made to create whole cell computer simulation models (Tomita, 2001). Rapid accumulation of biological data from genome, proteome, transcriptome and metabolome projects could advance efforts to construct virtual cells in silico. A solid foundation of accurate and complete gene annotation, together with a high quality index of encoded proteins, is a prerequisite for all of this future research.

1.7 Model organisms

1.7.1 Model organism genome projects

The first proposal of the HGP included the study of five model organisms (Watson, 1990). Thus far, the genomes of four of the five initially proposed organisms have been fully sequenced (table 1.2).

Table 1.2: the model organisms initially proposed for genome sequencing (Watson, 1990)

Organism	Genome size	Estimated no. of genes	Sequenced?
Escherichia coli[a]	4.2×10^6	4000	Y
Saccharomyces cerevisiae[b]	1.5×10^7	6000	Y
Caenorhabditis elegans[c]	1.0×10^8	13000	Y
Drosophila melanogaster[d]	1.2×10^8	10000	Y
Mus musculus	3.0×10^9	30000-100000	Nearly

[a] (Blattner *et al.*, 1997)
[b] (Goffeau , 1996)
[c] (Coulson, 1996)
[d] (Adams , 2000)

In addition, projects are underway to sequence the genome of the rat, zebrafish, and the pufferfish *T nigroviridis* and *T. rubripes*. Plans are also under consideration for sequencing additional primate and other organisms that will help define key developments along the vertebrate and non-vertebrate lineages.

The utility of the genomic sequence of model organisms in comparative sequence analysis is reviewed in section 1.5.3 and chapter IV. Model organisms also play an important role in elucidation of protein function and investigation of human disease (see below).

1.7.2 Functional studies in model organisms

Characterisation of an orthologous gene product through experimental assays in a model organism can offer valuable insights into function. Comparative maps provide the basis for identification of orthologous genes in a variety of organisms. For example, gene mapping placed the murine Pax3 gene on proximal mouse chromosome 1 and made it a candidate for the 'Splotch' mutation (Goulding *et al.*, 1991). When the locus for Waardenburg syndrome type I (WS1) was mapped to the homologous portion of the human genome, 2q37 (Foy *et al.*, 1990), Pax3 became a candidate for WS1 as well. Subsequent molecular studies showed mutations in the Pax3 gene in ‘Splotch’ mice (Epstein *et al*., 1991) and individuals with WS1 (Tassabehji *et al.*, 1992).

The nucleotide resolution of the physical maps resulting from the human and model organism genome projects (section 1.7.1) greatly enhances the efficiency of the identification process of candidate disease genes. This may be particularly important in the study of more complex disease traits, such as epilepsy, diabetes and hypertension, which involve more than one locus. These traits are difficult to study in humans because of limited family material, genetic heterogeneity, and complex environmental interactions. In experimental species such as the laboratory mouse and rat, planned crosses can be used, large numbers of progeny can be obtained and environmental factors can be controlled. Experiments can be replicated and gene-gene and gene-environment interactions studied. For example, genetic factors involved in inherited susceptibility to hypertension have been mapped to rat chromosomes 1, 2, 10 and 16 (Deng & Rapp, 1994). In these cases, knowledge of the genomic sequence will provide important clues to identifying homologous susceptibility genes in humans.

The genome projects have resulted in the identification of thousands of novel genes. Many large-scale experiments are underway to systematically study gene function in a more general way using knockouts. In yeast, where each of the ~6000 ORFs is likely to specify a protein product (Goffeau *et al.*, 1996), systematic knockouts of all the ORFs are in progress, either by insertion of transposons (Burns *et al.*, 1994; Ross-Macdonald *et al.*, 1999; Smith *et al.*, 1996), or total deletion of the ORF using a PCR based approach (Baudin *et al.*, 1993). This last method has been refined to include a molecular barcode where each deletion strain is tagged by a unique 20bp sequence (Shoemaker *et al.*, 1996).

A large project is also underway to disrupt all the genes annotated within the *Drosophila* genome using P transposable elements (Deak *et al.*, 1997; Spradling *et al.*, 1995). Similarly, availability of the genome sequence of *C. elegans* has increased the utility of this organism for functional studies. Genes identified within the sequence are easily knocked out using transposons (Collins *et al.*, 1987; Mori *et al.*, 1988; Rushforth *et al.*, 1993; Zwaal *et al.*, 1993) or by double stranded RNA interference (RNAi) (reviewed by Bargmann, 2001).

The increasing availability of mouse genomic sequence is also being exploited by the extensive array of genetic techniques available in mouse. For example, mice can be constructed with pre-determined genetic modifications to the germline by transgenic technology and gene targeting in embryonic stem cells. Recently, RNAi has been demonstrated in mouse cells (Yang *et al.*, 2001). Functional studies in mouse should therefore help to elucidate gene and protein function in humans.

1.8 Bioinformatics

The different large-scale genome related projects have produced a ‘tidal wave’ of data (Reichhardt, 1999). Bioinformatics has emerged as a science of recent creation, that uses biological data and knowledge stored in computer databases, complemented by computational methods, to aid interpretation of, and derive new, biological knowledge. The development of the World Wide Web (WWW) has accelerated this field as it allows easy access and sharing of data.

Initially, data is collected into databases. Large public domain databases are available for different types of information, for example, EMBL (Baker *et al.*, 2000), TrEMBL for translated

DNA sequences, GenBank for nucleotide and protein sequences and SwissProt for protein sequences (section 1.5.2.2). Individual labs can also use locally maintained databases to store data. For example, ACeDB, a *C. elegans* database, originally developed for the data generated by the *C. elegans* community, has been adapted for data management for many of the human sequencing projects.

A number of tools are available to analyse data. Data retrieval methods can be divided into text based or sequence based retrieval. Examples of text-based retrieval systems are Entrez (Schuler *et al.*, 1996b) which allows access to the data collections at the National Center for Biotechnology (NCBI) and the Sequence Retrieval System (SRS) (Etzold *et al.*, 1996) which allows the exploration of virtually all existing molecular biology databases. Most sequence-based searches are based on pairwise sequence-sequence comparison using algorithms such as BLAST. Similarities and differences are analysed at the nucleotide and/or amino acid level, with the aim to infer structural, functional and evolutionary relationships (Schuler, 1998).

If the sequence of interest contains protein-coding regions, analysis is more sensitive at the protein level as the DNA code is degenerate and, because of selective pressure, protein coding regions are more highly conserved. In general, a set of aligned sequences can be organised into an emerging family to define a ‘profile’. Such profiles aim at capturing the key functionally constrained features of the protein family. As a result, profile-sequence comparisons are a more powerful search tool than sequence-sequence comparisons. Profile-profile comparisons increase the possibility of detecting remotely related family members. In rare cases, discovery of similarities in 3D structure, when it is available, without apparent sequence similarity, can lead to the unification of functional families.

Bioinformatic techniques are also used to find genes. Both comparison-based and predictive methods were discussed previously (section 1.5).

Increasingly, sequence analysis is faced with problems of scale. New sequences become available every day from the various genome projects and processing them by hand is too slow. The flood of new sequence data can be handled only by automation. The genome browser Ensembl (Hubbard & Birney, 2000) provides one example. Ensembl is a software system that produces and maintains automatic annotation on eukaryotic genomes. Currently, *H. Sapiens*, M. *musculus* and *D. melanogaster* Ensembl servers exist. The data can be searched to identify genes, SNPs, proteins and protein families. Currently Ensembl provides identification of 90% of known human genes in the genome sequence and predicts 10,000 additional genes, all with supporting evidence.

Genome sequence provides a static picture. New high-throughput techniques of transcript and protein profiling will soon provide massive amounts of dynamic data, complementing static genome sequences. Database groups are now faced with the challenge of integrating genome sequence data with other data emanating from large-scale molecular biology.

1.9 Chromosome 22

Chromosome 22 is the second smallest of the human autosomes, comprising of 1.6 – 1.8% of the genomic DNA. It is an acrocentric chromosome: the p arm encodes the tandemly repeated rRNA genes and a series of other tandem repeat sequence arrays. There is no evidence to indicate the presence of any protein coding genes on 22p.

There are a number of genetic diseases located to this chromosome, listed in table 1.3.

Table 1.3: Syndromes linked to chromosome 22 genes.

Syndrome	Gene	Position	OMIM
Cat eye syndrome	CECR, CES	22q11	115470
Conotruncal cardiac anomalies	CTHM	22q11	217095
DiGeorge syndrome chromosome region (velocardiofacial syndrome)	DGCR, DGS, VCF	22q11	188400
Thrombophilia due to heparin cofactor II deficiency	HCF2, HC2	22q11	142360
Schindler disease; Kanzaki disease; NAGA deficiency, mild	NAGA	22q11	104170
Rhabdoid tumors; Rhabdoid predisposition syndrome, familial	SMARCB1, SNF5, INI1, RDT	22q11	601607
Breast cancer, t(11-22) associated	?	22q11	600048
Epilepsy, partial, with variable foci	FPEVF	22q11-q12	604364
Epstein syndrome	EPSTS	22q11-q13	153650
Schizophrenia 4	SCZD4	22q11-q13	600850
Glutathioninuria	GGT1, GTG	22q11.1-q11.2	231950
Gamma-glutamyltransferase, familial high serum	GGT2	22q11.1	137181
Bernard-Soulier syndrome, type B; giant platelet disorder, isolated	GP1BB	22q11.2	138720
May-Hegglin anomaly; Fechtner syndrome; Sebastian syndrome; Deafness, autosomal dominant 17	MYH9, MHA, FTNS, DFNA17	22q11.2	160775
Opitz G syndrome, type II	OGS2, BBBG2, GBBB2	22q11.2	145410
Hyperprolinemia, type I	PRODH	22q11.2	239500
Cataract, cerulean, type 2	CRYBB2, CRYB2	22q11.2-q12.2	123620
Agammaglobulinemia, autosomal recessive	IGLL1, IGO, IGL5	22q11.21	146770
Transcobalamin II deficiency	TCN2, TC2	22q11.2-qter	275350
Leukemia, chronic myeloid, Leukemia, acute lymphocytic	BCR, CML, PHL, ALL	22q11.21	151410
Ewing sarcoma; Neuroepithelioma	EWSR1, EWS	22q12	133450
Heme oxygenase-1 deficiency	HMOX1	22q12	141250
Colon cancer (deletions)	?	22q12-qter	
Li-Fraumeni syndrome	CHEK2, RAD53, CHK2, CDS1	22q12.1	604373
Sorsby fundus dystrophy	TIMP3, SFD	22q12.1-q13.2	188826
Neurofibromatosis, type 2; Meningioma, NF2-related, sporadic;Schwannoma, sporadic; Neurolemmomatosis; Malignantmesothelioma, sporadic	NF2	22q12.2	101000
Schwannomatosis	?	22q12.2	162091
Pulmonary alveolar proteinosis	CSF2RB	22q12.2-q13.1	138981
Meningioma	LARGE	22q12.3-q13.1	603590
Meningioma, SIS-related; Dermatofibrosarcoma protuberans; Giant-cell fibroblastoma	PDGFB, SIS	22q12.3-q13.1	190040
Neutrophil immunodeficiency syndrome	RAC2	22q12.3-q13.2	602049
Meningioma	MGCR, MN1	22q12.3-qter	156100
Colorectal cancer	EP300	22q13	602700
Megakaryoblastic leukemia, acute	MKL1, AMKL, MAL	22q13	606078
Spinocerebellar ataxia-10	SCA10	22q13	603516
Cardioencephalomyopathy, fatal infantile, due to cytochrome coxidase deficiency	SCO2	22q13	604272
Waardenburg-Shah syndrome; Yemenite deaf-blindhypopigmentation syndrome	SOX10, WS4	22q13	602229

Male infertility due to acrosin deficiency	ACR	22q13-qter	102480
Ovarian cancer (deletions)	?	22q13.1	
Adenylosuccinase deficiency; Autism, succinylpurinemic	ADSL	22q13.1	103050
Parkinsonism, susceptibility to; Debrisoquine sensitivity	CYP2D, P450C2D	22q13.1	124030
Glucose/galactose malabsorption	SLC5A1, SGLT1	22q13.1	182380
Chromosome 22q13.3 deletion syndrome	PSAP2, PROSAP2, KIAA1650	22q13.3	606230
Metachromatic leukodystrophy	ARSA	22q13.31-qter	250100
Methemoglobinemia, type I; Methemoglobinemia, type II	DIA1	22q13.31-qter	250800
Myoneurogastrointestinal encephalomyopathy syndrome	ECGF1	22q13.32-qter	131222
Megalencephalic leukoencephalopathy with subcortical cysts	MLC1, LVM, VL	22qter	605908

Adapted from OMIM Gene Map (**http://www.ncbi.nlm.gov/htbin-post/Omim**).

Chromosome 22 was the first human chromosome to be completely sequenced by a consortium of labs, providing 33.4Mb of sequence of the euchromatic portion of chromosome 22 in 12 contiguous segments (Dunham *et al.*, 1999). The sequence has been subjected to exhaustive computational analysis (Dunham *et al.*, 1999) and has served as a benchmark for new computational and experimental methods of analysis (for example, de Souza *et al.*, 2000; Mullikin *et al.*, 2000; Roest Crollius *et al.*, 2000; Salamov & Solovyev, 2000; Scherf *et al.*, 2001; Shoemaker *et al.*, 2001).

1.10 This thesis

The huge impact of the human genome project and the availability of genomic sequence can alter approaches to finding and identifying genes. Accurate annotation of genes within the genomic sequence is essential: the genome project will be an important reference for future genetic research and errors in the gene annotation at this early stage could adversely affect future studies of gene and protein function (see section 1.6).

The aim of this project was therefore to generate a highly accurate transcript map of a region of human chromosome 22. The utility of mouse genomic sequence for gene annotation and study of chromosomal evolution was also addressed. The generated human transcript map was then used as a basis for further study of the function of the annotated genes. A variety of bioinformatic and experimental methods were explored to provide a preliminary functional characterisation of the genes encoded within the region of interest.

The thesis will discuss:

1. Sequence analysis of a 3.4 Mb region of chromosome 22 (22q13.31) and the annotation of 39 experimentally verified genes and 17 pseudogenes. Screening of *ab initio* gene predictions in cDNA libraries was carried out to ensure completeness of the transcript map. Northern blotting and creation and screening of a 32 –tissue human cDNA panel, confirmed expression of the annotated gene features. Extensive sequence analysis of the genes and surrounding genomic sequence permitted investigation of translational start sites, polyadenylation sites, splice sites and predicted promoter regions. The correlation of each type of sequence evidence used to generate the transcript map was assessed against the final version, to determine the level of annotation accuracy each approach provided.
2. Comparative study of approximately 3.0 Mb in 22q13.31 and an additional 1.9 Mb region of 22q13.1. This chapter describes the construction of three mouse BAC clone contigs, spanning orthologous regions of mouse chromosomes 15 and 8. The available sequence was used to perform comparative analyses of coding and non-coding regions conserved in both mouse and human genomic sequence. The correlation of the conserved regions with both the existing annotation and with *ab initio* gene predictions

was assessed, but no further genes or exons were identified. Additionally, the study resulted in the refinement of a synteny breakpoint junction on human chromosome 22q13.1 and mouse chromosomes 15 and 8.

3. Preliminary functional characterisation of 27 complete genes identified from the transcript map of 22q13.31. *In silico* analyses were used to identify potential secondary structure and domain features within the predicted protein sequences. Phylogenetic analysis was also utilised to identify orthologous proteins from different species. Additionally, the subcellular localisation of a subset of the proteins was investigated through cloning and expression in a mammalian cell line.

This work was carried out as part of the ongoing work on chromosome 22 at the Sanger Institute.