Chapter I

Introduction

1.1 Introduction

The greatest achievement in biology over the past millennium has been the elucidation of the mechanism of heredity. Organisms encapsulate instructions for creating a member of their species in their gametes; these instructions are passed on to a fertilised egg and are then used to give rise to offspring. Although heredity intrigued philosophers since the time of Hippocrates and Aristotle, the absence of a way to probe the physical nature of these phenomena meant that they could do no more than speculate for more than two millennia.

Milestones in understanding the central dogmas of biology include Mendel's observations of genetic dominance and segregation of traits (1865), the discovery of chromosomes (end of the 19th century), the explanation of the biochemical basis of DNA (1953) and the deciphering of the genetic code (1964). Advances in molecular biology and sequencing sparked the genomics revolution and allowed for the first time the systematic study of individual genes and their environment. The almost "unthinkable" was then proposed in 1985: sequence the whole genome and generate a complete catalogue of every human gene. The Human Genome Project (HGP) was en route.

In the nineties, the HGP became a truly international collaboration involving research centres and funding agencies around the world (Bentley, 2000). Framework maps were generated and used to construct detailed physical and gene maps whereas the introduction of improved sequencing technologies dramatically increased sequence output (Collins,

2001). Model organisms were selected for systematic studies of their genetic make-up and new computational methods have been constantly devised to analyse the huge amount of generated data (Collins and Galas, 1993; Collins *et al*., 1998a). The list of completed genomes is constantly growing and includes *Saccharomyces cerevisiae* (Goffeau *et al*., 1996), *Caenorhabditis elegans* (The C. elegans Sequencing Consortium, 1998) and *Drosophila melanogaster* (Adams *et al*., 2000). In addition, the intermediate goal of the HGP has been achieved with the announcement of the completion and initial analysis of the draft human genome sequence (International Human Genome Sequencing Consortium (IHGSC), 2001).

Besides the tremendous progress that has been made in assembling a human reference sequence, more work is required to extract the full information contained in the genome. A high priority is to establish a complete catalogue of every gene, including all alternative splice forms, and accurately determine the structure of each one of them. Establishing and sequencing a comprehensive collection of full-length cDNAs (IHGSC, 2001) is one way to reach this goal but other approaches also have very strong potential, for example genome sequence comparisons to identify conserved sequences. At the computational level, new algorithms are required to efficiently predict genes and gene-related features. Easy data access should also be ensured through improved, user-friendly genome browsers.

Similar to obtaining the sequence, finding all the genes is an additional layer of genetic information. Many more layers are gradually targeted in a systematic way, from gene expression and regulation to protein localisation and function. For example, the

identification of regulatory regions and their study, in combination with expression array technologies, can identify sets of genes that are co-ordinately regulated. An integral part of all these efforts is the ongoing project that aims to complete the catalogue of human sequence variation and identify the common ancestral haplotypes present in the current population. The generated data from the HGP will serve as a toolbox to probe function in a more systematic manner. Genome data will be combined with improved techniques and databases for the global analysis of RNA and protein expression, protein localisation, protein-protein interactions, and chemical inhibition of pathways.

The generated resources from the HGP had an immediate impact on human health by accelerating the identification of human disease genes (roughly 200 disease-associated genes have been discovered using HGP information, Waterston and Sulston, 1998). In the future, understanding the human genome should illuminate the molecular pathogenesis of disorders that are currently poorly understood, and for which treatments are largely empirical and frequently sub-optimal. Deciphering the pathogenic pathways involved in illnesses could provide the greatest opportunity for the development of targeted therapies since the development of antibiotics (Collins and Guttmacher, 2001).

This chapter provides the background for the work described in the rest of this thesis. Genome organisation and mapping are discussed in section 1.2. Section 1.3 describes the features of the human genome. The computational approaches used for collecting, analysing and representing genomic data are discussed in section 1.4. Mapping and sequencing of model organisms (comparative studies) are discussed in 1.5. Section 1.6 describes the most popular methods employed by functional genomics to study

transcriptomes and proteomes. Human variation is discussed in section 1.7. Finally, chromosome 20 and the aims of this thesis are discussed in sections 1.8 and 1.9 respectively.

1.2 Mapping the human genome

1.2.1 The genome

The size of the haploid nuclear genome is estimated to be circa 3,200 Mb of DNA (Morton, 1991) and two copies are present in human somatic cells (note that the emerging sequence data suggests a smaller size. For example, the reported sizes of all finished chromosomes (20, 21, and 22) are smaller than the Morton estimates). Somatic nuclear DNA is organised into 23 pairs of chromosomes. There are 22 autosomes (numbered 1-22) and two sex chromosomes (X and Y, males are XY and females XX). The shape, relative size and distinctive banding pattern produced by various staining techniques, can identify each chromosome cytogenetically.

The basic shape of a chromosome is defined by the position of the centromere. Metacentric chromosomes have their centromere midway between the ends, submetacentric somewhat closer to one end, and acrocentric close to one end. The banding patterns generated by staining reflect the longitudinal structural heterogeneity of each chromosome (Bickmore and Sumner, 1989).

Giemsa (G) and reverse (R) banding are two of the most frequently used techniques for staining chromosomal regions (reviewed in Craig and Bickmore, 1993). The banding pattern reflects the base composition of a genomic region. Differences in base composition have been correlated with variations in gene density, time of replication, density of repeat sequences and chromatin packaging (Table 1.1) (Holmquist, 1992; Craig and Bickmore, 1993).

Table 1.1: Properties of Giemsa (G) and Reverse (R) bands (reproduced from Gardiner, 1995).

G bands	R bands
Stain strongly with Giemsa and quinacrine	Stain weakly with Giemsa and quinacrine
AT rich	GC rich
Chromomeres	Interchromomeres
DNase insensitive	DNase sensitive
Few breakpoints or rearrangements	Most breakpoints or rearrangements
Gene poor	Gene rich
Alu poor	Alu rich
LINE rich	LINE poor
Replicate late	Replicate early

1.2.2 Genome mapping

Various types of maps can be constructed to provide a more detailed view of a genome. Genetic and physical maps can be used to order landmarks along the length of a chromosome. The landmark-based maps serve as a scaffold for anchoring and orientating overlapping genomic clones (contigs) during clone map construction.

1.2.2.1 Genetic maps

Genetic mapping depends on following the segregation of alleles at two or more loci during meiosis. The unit of the genetic map, the Morgan (M), is defined as the length of chromosomal segment that on average undergoes one exchange per individual chromatid strand during meiosis. Over short chromosomal regions, the recombination fraction is directly proportional to the genetic map distance, so that a recombination fraction of 0.01 corresponds to a genetic map distance of 1 cM.

The construction of genetic maps had been limited by the availability of markers until 1980 when Botstein *et al*. observed that DNA sequence variation could provide a large source of polymorphic markers. These markers, known as Restriction Fragment Length Polymorphisms (RFLPs), could detect DNA polymorphisms when hybridised to restriction digests of an individual's DNA. The collection of RFLP markers was soon supplemented by the more informative minisatellites (Jeffreys *et al*., 1985), or variable number of tandem repeats (VNTR, Nakamura *et al*., 1987), which were used to construct low-resolution maps of the human genome (Donis-Keller *et al*., 1987).

Simple sequence repeats (SSRs), or microsatellites, provide another source of DNA polymorphisms. They are widely dispersed throughout eukaryotic genomes, are highly polymorphic (Weber and May, 1989; Weber, 1990) and when converted to Sequence Tagged Sites (STSs) (Olson *et al*., 1989) they can easily be typed using the polymerase chain reaction (PCR) (Saiki *et al*., 1985, 1988; Litt and Luty, 1989; Tautz, 1989). These technical advances aided the construction of human genetic maps of increasingly higher resolution (Hudson *et al*., 1992; Weissenbach *et al*., 1992; Gyapay *et al*., 1994; Murray *et al*., 1994; Dib *et al*., 1996). Genetic maps have also been constructed for a number of model organisms (section 1.5).

The available genetic map data was extensively used during the construction of physical maps for the human chromosomes. In addition, genetic maps continue to have an instrumental role in identifying disease genes. A new generation of genetic map, able to extract most of the inheritance information from human pedigrees, is under construction. The new, Single Nucleotide Polymorphism (SNP)-based map will contain several-fold more markers and genetic analysis will be performed using high-throughput, automated platforms (section 1.7).

1.2.2.2 Radiation hybrid maps

Based on the original approach by Goss and Harris (1975) and initially modified to study single chromosomes (Cox *et al*., 1990), Radiation Hybrid (RH) mapping was applied to study whole genomes (Walter *et al*., 1994; Gyapay *et al*., 1996).

RH cell lines are constructed by fusing lethally irradiated donor cells to recipient rodent cells deficient in a selectable marker. In RH mapping, the presence or absence of an STS-based marker is tested across a panel of radiation hybrids. The further apart two markers are on the chromosome, the more likely they are to be separated by an irradiation induced break, placing the markers on two separate chromosomal fragments. By estimating the frequency of breakage, and thus the distance between markers, it is possible to determine their order (Cox *et al*., 1990). The method combines aspects of both genetic and physical mapping. In contrast to recombination events, the frequency of breaks induced by the irradiation appears be linearly related to physical distance without cold or hot spots of breakage along the chromosome. The unit of map distance is the centiRay (cR) and represents 1% probability of breakage between two markers for a given radiation dosage (that is, the one used to construct the RH panel). Therefore the correlation between cR units and physical distance in bp will differ from one panel to the other and can only be determined by extrapolation.

In humans, RH mapping has been used to produce high-resolution gene maps by integrating genetic and EST-based markers (Schuler *et al*., 1996; Deloukas *et al*., 1998; http://www.sanger.ac.uk/HGP/Rhmap/). Dense RH framework maps were also constructed to assist the assembly of bacterial clone maps (Mungall *et al*., 1996).

1.2.2.3 Yeast artificial chromosome maps

The construction of clone maps is based on ordering and orientating cloned DNA fragments. As a cloning system, Yeast Artificial Chromosomes (YACs) (Burke *et al*.,

1987) can accommodate large DNA fragments (>1 Mb) and were initially used for the construction of chromosome-specific maps (Chumakov *et al*., 1992; Foote *et al*., 1992). Later, YAC contig maps covering most of the human genome were published (Chumakov *et al*., 1995; Hudson *et al*., 1995). Although YACs allow long-range continuity, they are not optimal substrates for a genome project because they suffer from instability and chimerism (Green *et al*., 1991; Nagaraja *et al*., 1994).

1.2.2.4 Bacterial clone maps

Bacterial (BAC) and P1 (PAC) artificial chromosomes (Shizuya *et al*., 1992; Ioannou *et al*., 1994) can accommodate DNA inserts of approximately 200 Kb (~5 times more than cosmids and fosmids (30-45 Kb), Collins and Hohn, 1978; Kim *et al*., 1992). Both PACs and BACs are stable and very few clones have been found so far to contain rearrangements (Shizuya *et al*., 1992; Ioannou *et al*., 1994). Because of the lower (1-2) copy number per cell and the smaller vector size, BACs are favoured over PACs. The available human PAC and BAC libraries (http://www.chori.org/bacpac/) represent more than 60 genome equivalents and are the preferred resource for sequence-ready map construction.

The initial strategy for constructing bacterial clone maps for whole chromosomes consisted of screening bacterial clones using a high density of STSs (15/Mb on average) obtained from framework maps (The Sanger Institute and Washington University Genome Sequencing Center, 1998). The bacterial clones were then assembled into contigs by comparative restriction fingerprint analysis (Coulson *et al*., 1986; Olson *et al*.,

1986; Gregory *et al*., 1997; Marra *et al*., 1997) and landmark content mapping (Green and Olson, 1990). The generated contigs were then extended and joined by chromosome walking. The strategy was slightly modified once a whole genome fingerprint map was assembled (International Human Genome Mapping Consortium (IHGMC), 2001). Clone maps for all human chromosomes are being constructed (Dunham *et al*., 1999; Hattori *et al*., 2000; Deloukas *et al*., 2001; Bentley *et al*., 2001; Tilford *et al*., 2001; Montgomery *et al*., 2001; Brüls *et al*., 2001; IHGMC, 2001).

1.3 Sequencing and the landscape of the human genome

1.3.1 Construction of sequence maps

Genome sequencing is used to construct maps of base pair resolution. The high quality DNA sequence allows the accurate annotation of all genomic features such as repeats, genes and their control elements. Sequence maps can be assembled either by a hierarchical clone-by-clone sequencing approach or a whole genome shotgun (WGS) approach, or a combination of the two.

The hierarchical approach employed by the IHGSC involves three steps: selecting a set of minimally overlapping clones from the map (tiling path), sequencing each clone and assembling the finished clone sequences into an overall genome sequence map (IHGSC, 2001).

Selected clones (predominantly BACs and PACs) are subjected to shotgun sequencing where the cloned DNA is fragmented and 1.4-2.2 Kb-long fragments are sub-cloned into M13 or plasmid vectors (Bankier *et al*., 1987). The subclones are then sequenced by a modified chain termination method (Sanger *et al*., 1977; IHGSC, 2001). A directed manual editing approach follows the automatic assembly of the generated sequence reads into sequence contigs. Any remaining sequence gaps and problems are resolved during "finishing". Finished sequence contains no gaps and is at least 99.99% accurate. Using this approach, the IHGSC aims to generate finished sequence for the euchromatic regions of all human chromosomes by 2003. Three chromosomes, 20, 21 and 22 (Deloukas *et al*., 2001; Hattori *et al*., 2000; Dunham *et al*., 1999) have already reached this gold standard.

The Whole Genome Shotgun (WGS) approach was first used to determine the sequence of *Haemophilus influenzae* (Fleischmann *et al*., 1995) and is routinely used to sequence small genomes. A hybrid approach was used to sequence the significantly larger and more complex genome of *Drosophila melanogaster* (Adams *et al*., 2000). The authors reported that they determined nearly all of the ~120 Mb euchromatic portion of the Drosophila genome using a whole-genome sequencing strategy supported by extensive clone-based sequence and a high-quality BAC clone map. Venter *et al*. (2001) reported the construction of a human genome sequence map using a WGS approach. Plasmid libraries of various sizes were constructed and sequenced, followed by a whole-genome sequence assembly without any prior mapping information. Sequence data generated by the public effort (using the hierarchical approach) was included in the final assembly; this sparked a debate as to how successful was the WGS approach (Waterston *et al*., 2002; Green, 2002; Myers *et al*., 2002).

A hybrid approach has been adopted for sequencing the mouse genome (Mouse Genome Sequencing Consortium, 2001; http://www.sanger.ac.uk/Info/Press/010508.shtml). The initial whole-genome sequence assembly already provides the scientific community with rapid access to the features of the mouse genome. This will be followed by the time-consuming process of clone-by-clone finishing to generate contiguous sequence and remove errors (such as mis-assemblies) that are common in draft sequence assemblies (Deloukas *et al*., 2001).

1.3.2 The genomic landscape-sequence features

1.3.2.1 GC content and isochores

Equilibrium centrifugation in analytical CsCl-density-gradient shows that DNA preparations from warm-blooded vertebrates are characterised by strong compositional heterogeneity, whereas those from cold-blooded vertebrates exhibit a weak heterogeneity (Thiery *et al*., 1976).

Fractionation of human DNA by equilibrium centrifugation in Cs_2SO_4 density gradients in the presence of sequence-specific DNA ligands (for example Ag^+) showed that human DNA is characterised by a very broad GC range. This analysis also indicated that the human genome is a mosaic of long (>200 Kb) DNA segments that have a homogeneous base composition, the isochores (equal regions) (Bernardi *et al*., 1985, 1989). The L1 and L2 isochore families are GC poor and represent about two-thirds of the genome whereas the H1, H2 and H3 isochore families are GC rich and represent the remaining third

(reviewed in Bernardi, 1989). H3 was later subdivided into three increasingly GC rich sub-fractions, $H3^{-}$, H3* and $H3^{+}$ (Saccone *et al*., 1996).

Saccone *et al*., (1993) used chromosome in situ hybridisation to correlate isochores and chromosomal bands. They showed that G-bands essentially consist of L1 and L2 isochores whereas the GC-richest regions of R-bands consist of H1, H2 and H3 isochores.

Analysis of the draft sequence shows an average GC content of 41% and confirms the variation of GC level across the human genome (for example, regions with GC contents of 33.1% and 59.3% have been identified), but rules out a strict notion of isochores as compositionally homogeneous (IHGSC, 2001). Although the genome does contain large regions of distinct GC content, the substantial variation present at many different scales indicates that a more moderate name such as "GC content domains" would be more appropriate (IHGSC, 2001).

1.3.2.2 CpG islands

The CpG dinucleotides are notable because they are usually methylated on the cytosine base and are greatly under-represented in human DNA, occurring at only about one-fifth of the expected frequency (IHGSC, 2001). A process that can lead to CpG suppression was described by Coulondre *et al*. (1978) who demonstrated that in the *Escherichia coli* lacI gene spontaneous base substitution hotspots occur at 5-methylcytosine residues.

Deamination of these residues to uracil and failure of DNA repair mechanisms can result in G:C→A:T transitions.

CpG islands constitute a distinctive fraction of the genome because they contain the dinucleotide CpG at its expected frequency and in the non-methylated form and their GC content is significantly higher than that of non-island DNA. CpG islands are rich in sites for methyl-sensitive restriction enzymes such as *Hpa*II that recognise unmethylated CpG dinucleotides (Bird, 1986). It is estimated that there are approximately 30,000 CpG islands in the haploid human genome (reviewed in Bird, 1987), but higher numbers have also been reported (Antequera and Bird, 1993; Cross and Bird, 1995).

Restriction enzymes were used to experimentally associate CpG islands and human genes (Lindsay and Bird, 1987). Computational sequence analysis of 375 human genes identified predicted CpG islands at the 5' end of all housekeeping genes, whereas 40% of the genes with tissue-specific or limited expression are also associated with islands. Overall, more than half of the genes analysed were associated with islands (Larsen *et al*., 1992). Analysis of a larger gene set suggests that approximately 58% of human coding genes have a CpG island at their start site (Ponger *et al*., 2001). Because of this association, CpG islands can be used as potential gene markers (Cross *et al*., 1994, 1999, 2000).

Analysis of the draft genome identified 50,267 putative CpG islands, of which at least 21,377 reside in repeats. The density of CpG islands varies substantially across

chromosomes but correlates reasonably well with estimates of relative chromosomal gene densities (IHGSC, 2001).

1.3.2.3 Repeats

An early observation was that genome size does not correlate well with organismal complexity (Lewin, 1994). For example *Homo sapiens* has a genome that is 200 times as large as that of *Saccharomyces cerevisiae*, but 200 times as small as that of *Amoeba dubia*. This phenomenon (known as the C-value paradox) was largely resolved with the recognition that genomes can contain a large quantity of repetitive sequence, far in excess of that devoted to protein-coding genes. Analysis of the draft sequence of the human genome revealed that coding sequences comprise less than 5% of the genome whereas repeat sequences account for at least 50% (IHGSC, 2001).

Generally, repeats fall into five classes:

A. Transposon-derived repeats, often referred to as interspersed repeats.

B. Inactive (partially) retroposed copies of cellular genes (including protein-coding genes and small structural RNAs) usually referred to as processed pseudogenes.

C. Simple Sequence Repeats (SSRs), consisting of direct repetitions of relatively short k-mers such as $(A)_n$, $(CA)_n$ or $(CGG)_n$.

D. Segmental duplications, consisting of blocks of around 10-300 Kb that have been copied from one region of the genome to another region.

E. Blocks of tandemly repeated sequences, such as centromeres, telomeres, the short arms of acrocentric chromosomes and ribosomal gene clusters (such regions are not present in the sequence of 20q12-13.2 and will not be discussed).

A. Transposon-derived repeats

Most of human repeat sequence is derived from transposable elements. 45% of the genome sequence is currently identified as such, and it is believed that much of the remaining "unique" sequence also belongs to this class but has diverged too widely to be recognised (IHGSC, 2001).

In humans there are four main types of transposable elements. These four types fall in two classes: DNA transposons (one type of transposable element) and retrotransposons (three types of transposable elements; Prak and Kazazian, 2000).

A.1 DNA transposons

DNA transposons move by excision and reintegration into the genome ***without*** an RNA intermediate. The main characteristics of this type include the Terminal Inverted Repeats (TIRs) that are 10-500 bp long. DNA transposons encode a transposase gene. When the gene is expressed, the transposase binds specifically to the TIRs and catalyses the cutting and pasting of the element. Integration results in a short, constant length duplication of the target site visible as directed repeats flanking the element (Smit and Riggs, 1996). Although DNA transposition is not replicative, it can result in duplication if (i) it moves from a replicated to a still non-replicated part of the genome (Chen *et al*., 1992) or (ii) the gap resulting from the excision of the transposon is repaired by using the sister chromatid as template (Engels *et al*., 1990).

The human genome contains at least seven major classes of DNA transposons, which can be subdivided into many families with independent origins. DNA transposons cannot exercise a cis-preference. The transposase is produced in the cytoplasm and when it returns to the nucleus it cannot distinguish active from inactive elements. As inactive elements accumulate in the genome, transposition becomes less efficient. This controls the expansion of any DNA transposon family and in due course causes it to die out (IHGSC, 2001).

A.2-4 Retrotransposons

Retrotransposons are duplicated through an RNA intermediate; the original transposon is maintained *in situ*, where it is transcribed. Its RNA transcript is then reverse transcribed into DNA and integrated into a new genomic location.

Long terminal repeat retrotransposons contain partly overlapping regions for group-specific antigen (gag), protease (prt), polymerase (pol) and envelope (env) genes. They are flanked on both ends by Long Terminal Repeats (LTRs) with promoter activity. The transcript is reverse transcribed in a cytoplasmic virus-like particle, primed by a tRNA.

Although a variety of LTR retrotransposons exist, only the vertebrate-specific endogenous retroviruses (ERVs) appear to have been active in the mammalian genome. Mammalian retroviruses fall into three classes (I-III), each comprising many families with independent origins. Most (85%) of the LTR retrotransposon-derived fossils consist only of an isolated LTR, with the internal sequence being lost by homologous recombination between the flanking LTRs (IHGSC, 2001).

Long interspersed elements (LINEs). Three distantly related LINE families are found in the human genome: LINE1, LINE2 and LINE3. Of these only LINE1 is active (IHGSC, 2001). 3,000-4,000 human LINE1s are full length, and of those only 40-60 are estimated to be active (Sassaman *et al*., 1997). In contrast the mouse genome has an estimated 3,000 active LINE1s (DeBerardinis *et al*., 1998).

A full-length (6.1 Kb) L1 element consists of a 5' untranslated region (5' UTR) that has a promoter activity; 2 open reading frames (ORF1 and ORF2) are separated by an intergenic spacer, followed by a 3'UTR and a polyA tail. ORF1 encodes a 40 KDa protein that binds nucleic acids, whereas ORF2 contains reverse transcriptase (RT), an endonuclease domain (EN) and a cysteine-rich region (C). Genomic L1 insertions are often flanked by 7-20 nucleotide target-site duplications (TSDs) (Prak and Kazazian, 2000). Note that at the stage where the reverse transcriptase uses the nicked DNA to prime reverse transcription from the 3' end of the LINE RNA, the enzyme frequently fails to reach the 5' end, resulting in many truncated, non-functional insertions (Smit, 1996; IHGSC, 2001).

Three distinct families of **short interspersed elements (SINEs)** are found in the human genome: the active Alu family, and the inactive MIR and Ther/MIR3 families. SINEs are characterised by an internal polymerase III promoter that ensures a fair chance for transcriptional activity of new copies (Smit, 1996). SINEs do not code for any proteins.

MIRs (mammalian-wide interspersed repeats) are approximately 260 bp long. Both families are tRNA-derived interspersed repeats that are believed to have spread through

the genome before the mammalian radiation (Jurka *et al*., 1995). MIR elements are the most mammalian-wide interspersed SINEs and are thought to be the most ancient mammalian SINE family (Rogozin *et al*., 2000; Smit and Riggs, 1995).

The most abundant and thoroughly studied SINEs are those belonging to the Alu family. Alus are named after the *Alu*I restriction site they carry (Houck *et al*., 1979; Gu *et al*., 2000). Alu repeats were derived from the signal recognition particle component 7SL (the 300 bp 7SL RNA is an essential component of the Signal Recognition Particle (SRP), which mediates the translocation of secretory proteins across the endoplasmic reticulum (Mighell *et al*., 1997)). A typical human Alu is an approximately 300 bp head-to-tail dimer of 7SL RNA-derived elements. The left monomer has significant similarity with a RNA pol III promoter; an A-rich linker connects right and left monomers. (Rogozin *et al*., 2000).

Based on the presence of diagnostic nucleotide substitutions, Alus are divided into three branches, which are further classified into sub-branches reflecting the age of individual elements from oldest (J), to intermediate (S), to youngest (Y) (Gu *et al*., 2000). The intermediate and older subfamilies have a significant amount of heterogeneity and there are many examples of intermediates between these various subfamilies. Thus, this categorisation is not exhaustive and is simply used to provide a reasonable working nomenclature for these older subfamilies (Batzer *et al*., 1996).

The AluJ repeats are divided into the Jo and Jb sub-branches and it is estimated that they were introduced to the genome 50 to 80 million years ago. The AluS repeats are divided

into the Sq, Sp, Sx, Sc, Sg, and Sg1 sub-branches. It is estimated that they were introduced to the genome 35 million years ago (Jurka and Milosavljevic, 1991; Gu *et al*., 2000).

The AluY repeats (Y, Ya5, Ya8, and Yb8) date back only to 20 million years ago. (Mighell *et al*., 1997; Gu *et al*., 2000). All Alu repeats presently known to retropose differ from the Y subfamily consensus sequence by only a few additional diagnostic mutations. This suggests that the youngest subfamilies of Alu repeats were ancestrally derived from the Y subfamily. Therefore, young subfamily sub-branches are defined as lineages that descended from this gold standard (Batzer *et al*., 1996).

LINE elements have been proposed to be the main generators of Alus. LINEs are thought to mobilise Alus because of the similarity of their target site duplications and the similarity of their insertion sites (the DNA nick for Alu insertions is probably made by L1 endonuclease). This parasitism of LINEs by SINEs remains difficult to reconcile with the observation that LINEs seem to insert preferentially into AT rich regions, whereas SINEs such as Alus accumulate in GC regions. One theory suggests that Alu elements integrate randomly but those that are actively transcribed (and are therefore more likely to reside in GC rich regions of the genome) are more likely to become fixed in the population. This explanation predicts that Alu RNA may have some advantageous function (Smit *et al*., 1999; Prak and Kazazian, 2000).

Possible roles for the transposon-derived repeats include the use of their regulatory elements by genes and regulation of protein translation. LINE-mediated transduction

mobilises DNA sequences around the genome, whereas at least 43 genes probably derived form DNA transposons (IHGSC, 2001). The ability of LINE retrotransposons to cause reverse transcription of genic mRNA can give rise to processed pseudogenes (Esnault *et al*., 2000). The possible role of these elements in the evolution of species suggests that it is important to further understand and evaluate their functional impact (Kass, 2001).

B. Processed pseudogenes

Pseudogenes were first identified in the *Xenopus* laevis genome (Jacq *et al*., 1977). The main characteristics of pseudogenes are the close similarity they have to one or more paralogous genes and the fact that most are non-functional due to failure of either transcription or translation (Mighell *et al*., 2000). Pseudogenes arise either by retrotransposition or duplication of genomic DNA. Pseudogenes that arise by retrotransposition are called processed pseudogenes and their main characteristics include the lack of introns and 5' promoter sequences and the presence of a 3' polyA tail. They are also flanked by short target site repeats (<15 bp) (Maestre *et al*., 1995).

Although pseudogenes are generally considered as defective entities, examples have been described where they retain their open reading frames (Vanin, 1985). Transcribing pseudogenes have been identified and it is postulated that some may also be functional (examples reviewed in Mighell *et al*., 2000; Makalowski, 2000). In addition, pseudogenes can potentially be copied to generate further pseudogenes.

C. Simple sequence repeats

SSRs are a common feature in the human genome. They are perfect or slightly imperfect tandem repeats of a particular k-mer. SSRs with a short repeat unit (n=1-13 bp) are called microsatellites, whereas those with longer repeat units (n=14-500 bp) are called minisatellites. SSRs comprise about 3% of the genome. The biggest contribution is by dinucleotide repeats (0.5% of the genome; IHGSC, 2001).

Of the twelve equivalence classes of triplet repeats (64 classes, twelve taking into account reverse complement and shift), three (CAG, CGG, GAA) have been associated with triplet disease disorders (Baldi and Baisnee, 2000). Expansions of unstable trinucleotide repeats have been associated with at least fifteen inherited neurologic diseases (Lieberman and Fischbeck, 2000).

D. Segmental duplications

Segmental duplications involve the transfer of blocks of sequence (1-200 Kb) to one or more locations of the genome. Interchromosomal duplications involve blocks of sequence duplicated among non-homologous chromosomes whereas intrachromosomal duplications involve blocks of sequence duplicated within a particular chromosome or chromosomal arm. Recombination between duplicons leads to chromosomal rearrangements that could lead to genomic disorders (Ji *et al*., 2000).

1.3.2.4 Genes

The task of nuclear gene transcription is shared by three RNA polymerases (pol I, pol II and pol III). Pol I synthesises rRNA and pol III makes 5S rRNA, tRNA, 7SL RNA, U6 snRNA and a few other small stable RNAs. In contrast, the huge variety of protein-coding genes is transcribed by pol II. So, although pol I and pol III account for 80% of total RNA synthesis (reviewed in Paule and White, 2000), only pol II-transcribed genes will be discussed.

Human protein-coding DNA sequences have complex structures (Figure 1.1), typically segmented by intervening sequences (Tilghman *et al*., 1978a, 1978b), called introns (Gilbert, 1978). A large ribonucleoprotein complex, the spliceosome, recognises sites at the 5' and 3' ends of introns (the donor and acceptor sites respectively), as well as an internal site, the branch point, and removes introns from gene transcripts (Moore and Sharp, 1993). With a few exceptions (Sharp and Burge, 1997; Burset *et al*., 2000), nearly all spliceosomal introns begin with GT and end with AG (donor and acceptor sites, respectively). The retained segments, called exons, form the messenger RNA (mRNA). Differential removal of RNA sequences from gene transcripts of a particular gene (alternative splicing) generates isoforms that can encode for protein variants. EST-based studies estimate that 35-38% of human genes undergo alternative splicing (Mironov *et al*., 1999; Brett *et al*., 2000). Higher numbers of alternatively spliced genes have also been reported (Kan *et al*., 2001; IHGSC, 2001).

The discovery of introns sparked a debate regarding their origin and evolution (reviewed in Long *et al*., 1995; Logsdon *et al*., 1995). The fragmentation of protein-coding genes by

introns may have conferred an advantage, by facilitating the modular shuffling of eukaryotic protein domains in evolutionary time and in real time via alternative splicing (Mattick, 2001).

Other features of protein-coding genes include a translation start site (usually ATG), often contained in an optimal consensus sequence (Kozak, 1987) and in most cases, a highly conserved hexamer (the polyA signal; Kessler *et al*., 1986), which is involved in the polyadenylation of the RNA transcript. A significant fraction of genes display multiple polyadenylation sites (Gautheret *et al*., 1998) and the patterns of variant polyA signals used in this process are currently under study (Beaudoing *et al*., 2000).

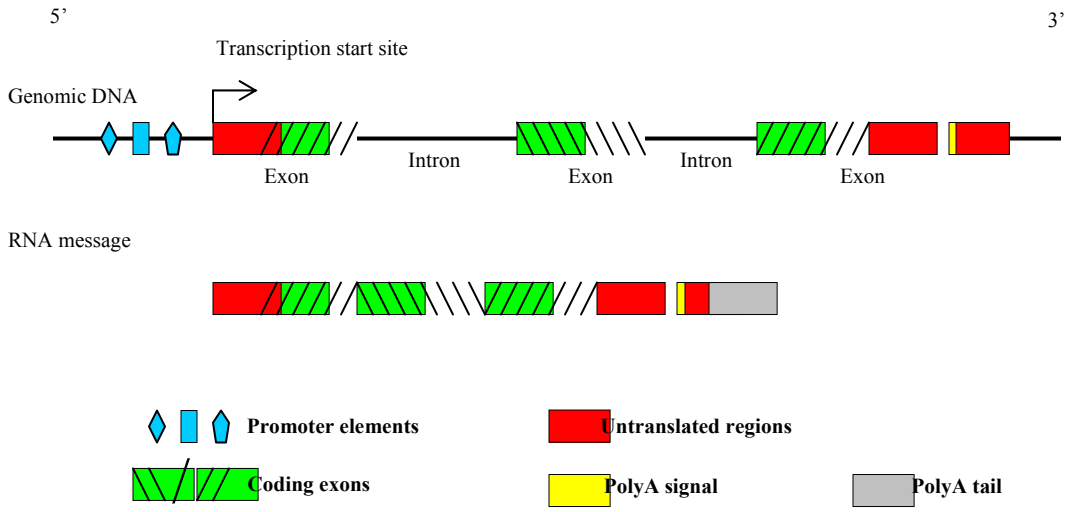


Figure 1.1: Basic gene structure. The organisation of a typical gene locus is shown on top whereas the resulting messenger RNA is shown below. The coloured boxes represent various gene features.

Promoter sequences upstream of the site of transcriptional initiation of protein-coding genes bind the basal transcriptional machinery (and additional potentiating factors) and specify the transcription start (TS) site (reviewed in Fickett and Hatzigeorgiou, 1997; Dillon and Sabbattini, 2000). The complex biochemical processes that govern promoter recognition, binding and control of transcription are currently under intense investigation. In addition to promoters, other cis-acting regulatory sequences could also be present upstream, downstream, or within a gene locus (reviewed in Fraser and Grosveld, 1998; Li *et al*., 1999; Lee and Young, 2000).

Identification of transcribing sequences in genomic DNA is a recurrent problem in transcript mapping and positional cloning studies. Traditional techniques employed to address this problem include using CpG islands as positional signposts for the starts of some transcription units (Lindsay and Bird, 1987), 'zoo' blots to detect cross-species conserved genomic sequences that could represent genes (Monaco *et al*., 1986) and hybridisation of entire genomic clones directly to cDNA libraries (Elvin *et al*., 1990; methods reviewed in Gardiner and Mural, 1995).

A variation of the latter is cDNA selection where the genomic sequence of interest is immobilised on a nylon filter and hybridised to an amplified cDNA library. The hybridised cDNA sequences are then eluted and re-amplified using PCR. The process can be repeated to achieve further enrichment (Parimoo *et al*., 1991; Lovett *et al*., 1991). The method was improved by using biotin-labelled genomic DNA and streptavidin-coated magnetic beads to capture the DNA-cDNA hybrids (Korn *et al*., 1992; Morgan *et al*.,

1992). The cDNA selection method was successfully used to identify expressed sequences for specific chromosomes (Touchman *et al*., 1997).

The exon trapping/amplification technique (Duyk *et al*., 1990; Buckler *et al*., 1991) is based on the selection of exonic sequences that are flanked by functional 5' and 3' splice sites. Fragments of genomic DNA are cloned into a vector flanked by 5' and 3' splice sites. The constructs are used to transfect COS-7 cells and the resulting RNA transcripts are processed *in vivo*. Splice sites of exons contained within the inserted genomic fragment are paired with those of the flanking intron. The resulting mRNA contains the previously unidentified exons that can then be PCR-amplified and cloned. Exon trapping has been used to identify candidate disease genes (Vulpe *et al*., 1993; Trofatter *et al*., 1993) and exons from entire chromosomes (Church *et al*., 1993; Trofatter *et al*., 1995). Unfortunately, the methods described above are technically challenging and time-consuming which makes them unsuitable for analysing large genomic regions (Gardiner and Mural, 1995).

Currently, large-scale sequence analysis relies on DNA and protein similarity searches. Extensive collections of Expressed Sequence Tags (ESTs) from human (Adams *et al*., 1991; Wilcox *et al*., 1991), mouse (Marra *et al*., 1999), rat and other model organisms have become available (Boguski *et al*., 1993). Additional resources include the mRNAs of known genes and a large number of anonymous cDNA clone sequences produced by different centres (KIAA collection, http://www.kazusa.or.jp, Nomura *et al*., 1994; RIKEN collection, http://www.riken.go.jp/, The RIKEN Genome Exploration Research Group Phase II Team and the FANTOM Consortium, 2001; Genoscope collection,

http://www.Genoscope.cns.fr; DKFZ collection, http://mbi.dkfz-heidelberg.de/, Wiemann *et al*., 2001). In addition, sequence comparisons between species at the genomic level are utilised to identify conserved regions, which may represent coding exons.

Ab initio methods are also used to complement the homology-based analysis results. These algorithms can predict coding sequences, promoters, transcription start sites and CpG islands (section 1.4). A disadvantage of this approach is the need for experimental confirmation of the various predictions (see chapter III).

Published estimates regarding the total number of human coding genes vary substantially. For example, Crollius *et al*. (2000a) used a human:*Tetraodon nigroviridis* comparative sequence approach and estimated that the human genome contains 28,000-34,000 genes. Human EST analyses by other groups suggest that the total number is approximately 120,000 (Liang *et al*., 2000). The Chromosome 22 Sequencing Consortium estimated a minimum of 45,000 genes based on their annotation of the complete chromosome although their data suggests that there may be additional genes (Dunham *et al*., 1999). Other whole-chromosome studies suggest that the gene number may be closer to 40,000 (Hattori *et al*., 2000), 35,000 (Ewing and Green, 2000) or 31,500 (Deloukas *et al*., 2001). Two independent analyses of the draft genome sequence suggest a total of 39,114 (Venter *et al*., 2001) and 31,778 (IHGSC, 2001) but the minimal overlap between the novel genes of the two sets casts doubts on these gene estimates (Hogenesch *et al*., 2001).

The “gene guessing game” (Fields *et al*., 1994; Dunham, 2000; Aparicio, 2000; Simpson *et al*., 2001) and the recent consensus of there being less than 40,000 genes in the human

genome raises questions regarding the origins of human species complexity (Claverie, 2001). The structure and control architecture of genes are probably at the heart of eukaryotic complexity and phenotypic variation (Mattick, 2001).

1.4 Computational genomics (Bioinformatics)

The "tidal wave of data" (Reichhardt, 1999) caused by the Human Genome Project gave birth to bioinformatics, the computer-assisted data management discipline that helps the gathering, analysis and representation of genomic information (Persidis, 1999, 2000).

1.4.1 Sequence databases

DNA sequences are made publicly available through the International Nucleotide Sequence Databases (INSD) that consist of GenBank (http://www.ncbi.nlm.nih.gov/Genbank/index.html; Benson *et al*., 2002), EMBL (http://www.ebi.ac.uk/embl/; Stoesser *et al*., 2002) and DDBJ (http://www.ddbj.nig.ac.jp; Tateno *et al*., 2002). Data is exchanged between the three sites on a daily basis to ensure that each maintains a comprehensive collection of sequence information. The amount of sequence data stored continues to grow at an exponential rate and more than 105,000 different species are represented in the databases (Benson, 2002; 121,736 as of May 2002, http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/index.cgi?chapter=howmany).

A large number of other DNA databases are also maintained, including the dbEST database (http://www.ncbi.nlm.nih.gov/dbEST/, Boguski *et al*., 1993), the Unigene database (http://www.ncbi.nlm.nih.gov/UniGene/index.html), the Reference Sequence Database (RefSeq; http://www.ncbi.nlm.nih.gov/LocusLink/refseq.html, Pruitt and Maglott, 2001) and the database of Single Nucleotide Polymorphisms (dbSNP; http://www.ncbi.nlm.nih.gov/SNP/, Sherry *et al*., 2001). LocusLink (http://www.ncbi.nlm.nih.gov/LocusLink/, Pruitt and Maglott, 2001) provides summary pages for each gene and links to other relevant databases.

Annotated (curated) protein sequences are stored in SWISS-PROT (http://www.ebi.ac.uk/swissprot/) whereas TrEMBL contains EMBL nucleotide translated sequences (http://www.ebi.ac.uk/trembl/index.html; Bairoch *et al*., 2000). InterPro (http://www.ebi.ac.uk/interpro/index.html; Apweiler *et al*., 2000) provides an integrated documentation resource for protein families, domains and functional sites.

Data is freely accessible to the scientific community via web-based, integrated database retrieval systems such as the Sequence Retrieval System (SRS; http://srs.ebi.ac.uk/; Zdobnov *et al*., 2002) and Entrez (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi; Wheeler *et al*., 2002).

1.4.2 Sequence analysis

In addition to text-based searches, tools such as BLAST (Altschul *et al*., 1990, 1997; Table 1.2) enable sequence-based data mining. Sequence comparisons at the nucleotide

and/or amino acid level are performed to identify homologous sequences. Alignment of such sequences using software tools such as CLUSTAL W (Thompson *et al*., 1994) can be used to predict structural, functional and evolutionary relationships.

Table 1.2: Blast types (reproduced from Brenner, 1998).

Program	Query	Database	Comparison
blastn	DNA	DNA	DNA level
blastp	Protein	Protein	Protein level
blastx	DNA	Protein	Protein level
tblastn	Protein	DNA	Protein level
tblastx	DNA	DNA	Protein level

New homology search tools such as SSAHA (Ning *et al*., 2001), Exonerate (Slater, unpublished) and BLAT (Kent, 2002) can be used to perform fast searches of databases containing multiple gigabases of DNA, whereas software such as PipMaker (Schwartz *et al*., 2000), Vista (Mayor *et al*., 2000), GLASS (Batzoglou *et al*., 2000) and SynPlot (Göttgens *et al*., 2001) are available for performing comparative sequence analyses (reviewed in Miller, 2001).

A program widely used for genomic analysis is RepeatMasker (Smit and Green, unpublished). This program scans sequences to identify full-length and partial members of all known repeat families represented in Repbase (Jurka, 2000). RepeatMasker analysis of the draft genome sequence has shown that at least 50% corresponds to repeat

elements (IHGSC, 2001). So, masking repeats before performing homology searches is an important step that filters out spurious matches.

Software programs that have been developed to predict sequence features are summarised in Table 1.3. Various evaluations of exon/gene prediction programs indicate that each has its individual strengths and weaknesses that are reflected by its sensitivity and specificity scores (Burset and Guigo, 1996; Claverie, 1997; Guigo *et al*., 2000; Rogic *et al*., 2001). Currently, none of the available software can correctly predict all protein-coding genes in a given sequence, although more reliable predictions can be obtained by the combined use of software (Murakami and Takagi, 1998).

Computational analysis of polymerase II promoters can contribute to gene identification. In 1997, Fickett and Hatzigeorgiou reviewed the available prediction programs and concluded that on average they report one false positive per Kb. Although this may not be a problem when analysing short genomic sequences, analysis of a whole chromosome or genome would produce too many false positives. Since then, new software tools have been developed. PromoterInspector (Scherf *et al*., 2000) and Eponine (Down and Hubbard, 2002) have a sensitivity of 43% and 40% respectively and nearly 40% of their predictions correspond to true promoters (Scherf *et al*., 2001; Deloukas *et al*., 2001). Other available software include CPGFIND (Micklem, unpublished) for CpG island prediction (which are associated with the 5' end of genes) and a newly developed algorithm, FirstEF, which according to Davuluri *et al*. (2001), predicts 86% of first exons with only 17% false positives.

Table 1.3: Overview of the main sequence-feature prediction programs.

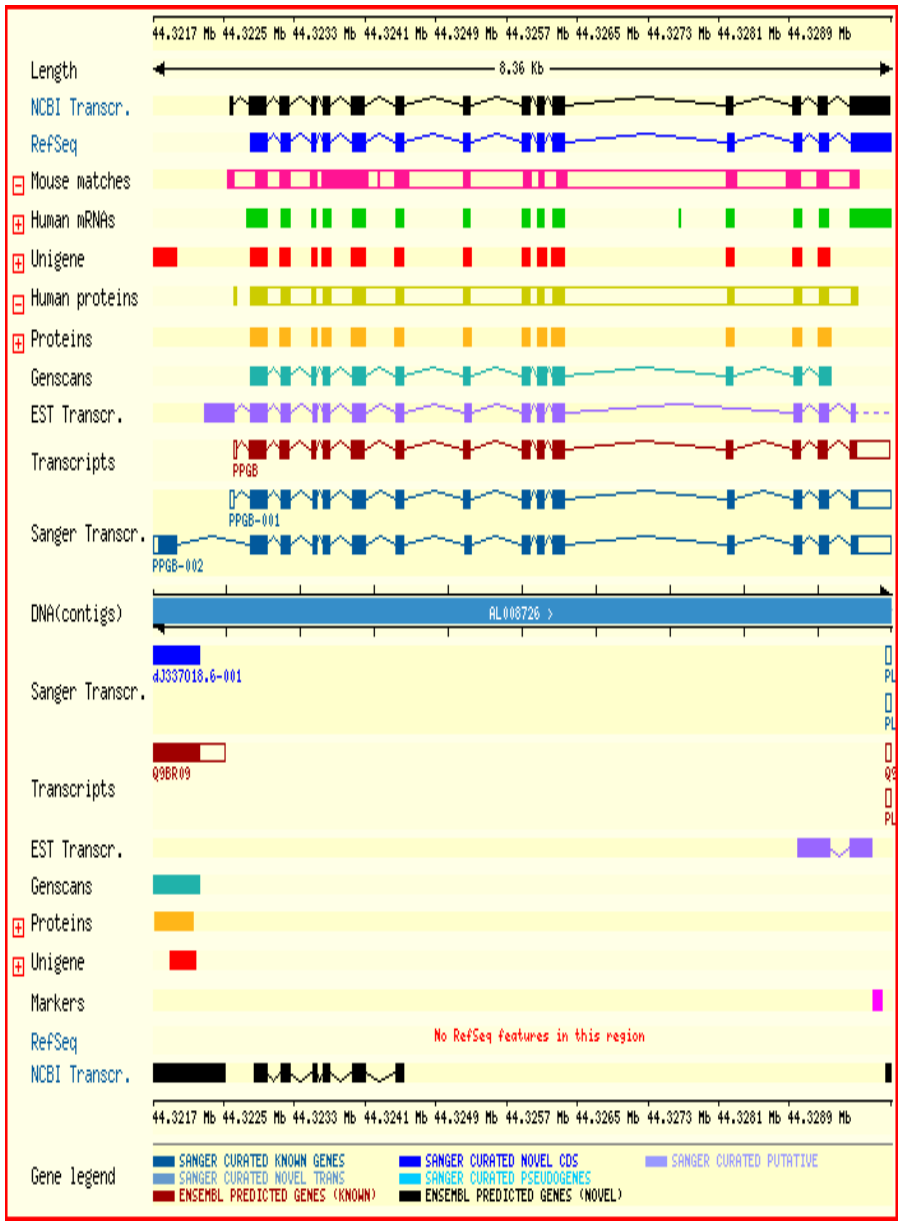
Program	Description	Reference
Genscan	Gene prediction	Burge and Karlin, 1997
FGENESH	Gene prediction	Salamov and Solovyev, 2000
Hexon	Exon prediction	Solovyev *et al*., 1994
GRAIL	Exon prediction	Uberbacher *et al*., 1996
Xpound	Exon prediction	Thomas and Skolnick, 1994
PromoterInspector	Promoter prediction	Scherf *et al*., 2000
Eponine	TS site prediction	Down and Hubbard, 2002
FirstEF	First exon prediction	Davuluri *et al*., 2001
CPGFIND	CpG island prediction	Micklem, unpublished
RepeatMasker	Repeat sequences prediction	Smit and Green, unpublished

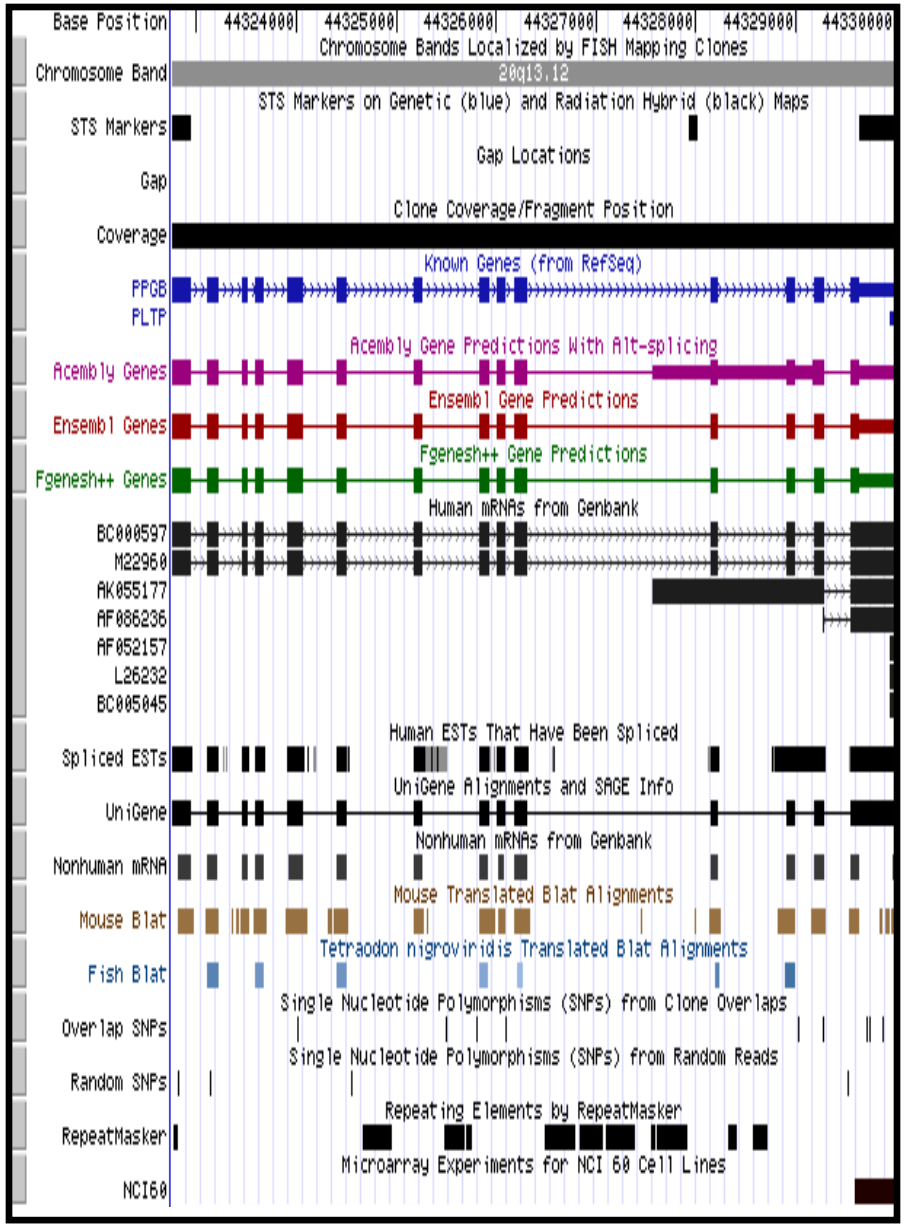
1.4.3 Viewing genomic information

Individual laboratories use different software to represent genomic information. For example, ACeDB, which was originally developed for the *Caenorhabditis elegans* community (A C. elegans DataBase; Durbin and Thierry-Mieg, 1994), is extensively used at the Sanger Institute. ACeDB is used for graphical display and browsing of biological data such as DNA/peptide sequence and annotation, map data and hybridisation data (Kelley, 2000). Several of the human chromosome projects rely on ACeDB for data management (http://www.sanger.ac.uk/HGP/Humana/).

Ensembl (http://www.ensembl.org/) is one of the leading sources of automated genome sequence annotation. Ensembl provides a bioinformatics framework to organise biology around the sequences of large genomes. Annotation of gene features is performed automatically using confirmed gene predictions that have been integrated with external data resources (Hubbard *et al*., 2002). The other main genome browser is based at the University of California Santa Cruz (UCSC) (Kent *et al*., 2002; http://genome.ucsc.edu/). Like Ensembl, the UCSC genome browser displays a variety of information, including assembly contigs and gaps, mRNA and expressed sequence tag alignments, gene predictions, cross-species homologies, single nucleotide polymorphisms, sequence tagged sites and repeat elements. An example of how the sequence features are displayed by the Ensembl and the UCSC genome browser is shown in Figure 1.2.

Figure 1.2 (next page): The Ensembl (red box) and UCSC (black box) genome browsers. In this example both browsers display the sequence features of the PPGB locus.


44.3217 Mb 44.3225 Mb 44.3233 Mb 44.3241 Mb 44.3249 Mb 44.3257 Mb 44.3265 Mb 44.3273 Mb 44.3281 Mb 44.3289 Mb
Length
8.36 Kb
NCBI Transcr.
RefSeq
Mouse matches
Human mRNAs
Unigene
Human proteins
Proteins
Genscans
EST Transcr.
Transcripts
PPGB
PPGB-001
Sanger Transcr.
PPGB-002
DNA(contigs)
AL008726 >
Sanger Transcr.
dJ337018.6-001
Transcripts
Q9BR09
EST Transcr.
Genscans
Proteins
Unigene
Markers
RefSeq
No RefSeq features in this region
NCBI Transcr.
Gene legend
SANGER CURATED KNOWN GENES
SANGER CURATED NOVEL TRANS
ENSEMBL PREDICTED GENES (KNOWN)
SANGER CURATED NOVEL CDS
SANGER CURATED PSEUDOGENES
ENSEMBL PREDICTED GENES (NOVEL)
SANGER CURATED PUTATIVE


Base Position
44324000 44325000 44326000 44327000 44328000 44329000 44330000
Chromosome Bands Localized by FISH Mapping Clones
Chromosome Band
20q13.12
STS Markers on Genetic (blue) and Radiation Hybrid (black) Maps
STS Markers
Gap Locations
Gap
Clone Coverage/Fragment Position
Coverage
Known Genes (from RefSeq)
PPGB
PLTP
Acembly Gene Predictions With Alt-splicing
Acembly Genes
Ensembl Gene Predictions
Ensembl Genes
Fgenesh++ Gene Predictions
Fgenesh++ Genes
Human mRNAs from Genbank
BC000597
M22960
AK055177
AF086236
AF052157
L26232
BC005045
Human ESTs That Have Been Spliced
Spliced ESTs
UniGene Alignments and SAGE Info
UniGene
Nonhuman mRNAs from Genbank
Nonhuman mRNA
Mouse Translated Blat Alignments
Mouse Blat
Tetraodon nigroviridis Translated Blat Alignments
Fish Blat
Single Nucleotide Polymorphisms (SNPs) from Clone Overlaps
Overlap SNPs
Single Nucleotide Polymorphisms (SNPs) from Random Reads
Random SNPs
Repeating Elements by RepeatMasker
RepeatMasker
Microarray Experiments for NCI 60 Cell Lines
NCI60

1.5 Comparative genomics

Much of the power of molecular genetics arises from the ability to isolate and study genes from one species based on knowledge about related genes in other species. Comparisons between genomes that are distantly related provide insight into the universality of biological mechanisms and identify experimental models for studying complex processes. Furthermore, comparisons between genomes that are closely related provide unique insights into the details of gene structure and function (Collins *et al*., 1998a).

Five model organisms, *Escherichia coli*, *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster* and *Mus musculus* were targeted by the first five-year plan for the Human Genome Project (1993-1998; Collins and Galas, 1993). The genomes of these organisms can provide vital insights regarding the function and organisation of the human genome. For example, the genome of a unicellular organism could reveal a minimal set of proteins required for life.

Genomic analysis of multicellular organisms can help unravel the complex developmental pathways, whereas the mouse genomic sequence can be used to highlight conserved functional features such as genes and regulatory elements (Watson, 1990; Collins *et al*., 1998a). Of the five model organisms initially proposed, four have been fully sequenced so far (table 1.4), whereas the mouse genome sequencing project is still in progress.

Table 1.4: Genome sizes of the model organisms initially proposed. Current protein numbers and a summary of InterPro analysis are also shown (as of January 2002).

Organism	Genome size (Mb)	Number of proteins	Proteins with InterPro matches
E. coli[1]	4.6	4,363	71.2 %
S. cerevisiae[2]	12	6,123	65.6 %
C. elegans[3]	97	18,940	67.9 %
D. melanogaster[4]	120	13,958	71.6 %
M. musculus	3,000 (est)	15,856	72.6 %

[1]Blattner *et al*., 1997
[2]Goffeau *et al*., 1996
[3]The C. elegans sequencing consortium, 1998
[4]Adams *et al*., 2000

Initial sequence analysis of the *Escherichia coli* genome revealed that protein-coding genes account for approximately 88% of the genome. Insertion sequence (transposable) elements, phage remnants and many other patches of unusual composition were also identified, indicating genome plasticity through horizontal transfer (Blattner *et al*., 1997).

In *Saccharomyces cerevisiae* only 4% of protein-coding genes contain introns (Goffeau *et al*., 1996) and so Open Reading Frames (ORFs) can be easily identified. The functions of individual ORFs can be evaluated readily because it is relatively easy to disrupt them *in vivo* (Rothstein, 1991; Burns *et al*., 1994). A variety of approaches are used to analyse the yeast sequence on a genome-wide scale (Bassett *et al*., 1996; Ross-MacDonald *et al*., 1999). LacZ/transposon insertion-based approaches were used to perform large-scale analysis of gene expression, protein localisation and gene disruption (Burns *et al*., 1994; Ross-MacDonald *et al*., 1999). A PCR-based approach was also developed to perform targeted ORF deletions (Baudin *et al*., 1993). This was later refined by adding two unique

20 bp sequences in each deletion construct that can serve as molecular barcodes for the strain carrying them (Shoemaker *et al*., 1996). Other methods used to study the yeast genome and proteome include genetic footprinting to assess the phenotypic effects of induced Ty1 transposon insertions (Smith *et al*., 1996), serial analysis of gene expression (SAGE; Velculescu *et al*., 1997), microarray analysis (DeRisi *et al*., 1997), two-dimensional (2D) gel/protein identification analysis (Maillet *et al*., 1996), two-hybrid system analysis (Fromont-Racine *et al*., 1997) and genome-wide protein tagging (Martzen *et al*., 1999).

The worm *Caenorhabditis elegans* is an ideal animal to study basic gene functions. It is fully transparent at all stages of its life, allowing cell divisions, migrations, and differentiation to be monitored in live animals. Its anatomy is simple, yet the 959 somatic cells of the adult represent most major differentiated tissue types. In addition it is the only animal for which the complete neuronal wiring pattern is known (reviewed in Ahringer, 1997). On average, the amino acid (aa) similarity and identity between aligned human and *Caenorhabditis elegans* orthologous gene products are 69.3% and 49.1% respectively, and the nucleotide identity is 49.8% (Wheelan *et al*., 1999). *Caenorhabditis elegans* genes can be studied by generating knockouts using a PCR/Tc1-transposon-based approach (Collins *et al*., 1987; Mori *et al*., 1988; Rushforth *et al*., 1993; Zwaal *et al*., 1993). Gene expression patterns can be determined using high resolution FISH to visualise mRNA distributions at the cellular and sub-cellular level (Birchall *et al*., 1995), and manipulated using double-stranded RNA interference (Fire *et al*., 1998).

RNA interference (RNAi) describes the use of double-stranded RNA (dsRNA) to target mRNAs for degradation. When dsRNA is injected into worms the RNAi machinery uses the sequence information in the dsRNA to generate a protein-RNA complex that destroys the corresponding mRNA. This new approach led to the development of new strategies for blocking gene function, which have been successfully applied to silence worm and fly genes (reviewed in Schmid *et al*., 2002; Bargmann, 2001; Zamore, 2001).

In terms of evolutionary sequence conservation, the fruit fly *Drosophila melanogaster* is the closest of the invertebrate model organisms to humans (Sidow and Thomas, 1994). It has been studied for more than 80 years and numerous publications are dedicated to describing its genetics and hundreds of individual genes (reviewed in Rubin, 1996). The fruit fly provides a powerful system to study the function of conserved genes because any of its ORFs can be mutated and subjected to detailed functional analysis within the context of an intact organism. An ongoing gene-disruption project establishes mutant strains that each contains a single, genetically engineered P transposable element in a defined genomic region (Spradling *et al*., 1995; Deák *et al*., 1997).

The mouse has served over the past century as an excellent experimental system for studying mammalian genetics and physiology (Dietrich *et al*., 1995). A number of detailed mouse genetic and physical maps have been constructed (discussed in section 4.1.1), whilst in May 2001, the Mouse Sequencing Consortium reported a first sequence draft of the mouse genome (http://www.sanger.ac.uk/Info/Press/010508.shtml). Detailed human:mouse comparative studies at distinct genomic regions are used to identify coding regions and regulatory elements (discussed in section 4.1.2). Methods such as Mb-long

chromosomal rearrangements (Mills and Bradley, 2001) and tagged random mutagenesis (Zambrowicz *et al*., 1998) in embryonic stem cells are used to study gene function and generate mammalian models for developmental processes and cancer.

In 1998(a), Collins *et al*. suggested the selection of additional model organisms. For example, the vertebrate fish *Danio rerio* (zebrafish) is an excellent genetic system for the identification and functional analysis of genes that control pattern formation and organogenesis. Zebrafish complements other experimental systems, since information gained from analyses of its functionally important genes can readily be extended to homologous systems in mice and other organisms (Driever *et al*., 1994; Talbot and Hopkins, 2000). Genetic screens have identified mutations in hundreds of genes with essential functions in the zebrafish embryo (Dreiever *et al*., 1996; Haffter *et al*., 1996). Genetic linkage (Shimoda *et al*., 1999; Gates *et al*., 1999; Kelly *et al*., 2000) and radiation hybrid (Geisler *et al*., 1999) maps have been constructed, followed by a whole-genome human:zebrafish comparative map (Woods *et al*., 2000). A hybrid approach to sequence the zebrafish genome is underway at the Sanger Institute.

The genome of the pufferfish *Fugu rubripes* (Fugu) is only four times larger than that of *Caenorhabditis elegans* and it was shown that a random genomic sequence is 7.5 times more likely to be coding than a random human sequence (Brenner *et al*., 1993). This compact genome provides a simple and economic approach to compare sequence data from mammals and fish and could enable the identification of essential conserved elements because of the large evolutionary divergence (Elgar *et al*., 1996). The presence (or not) of synteny between human and Fugu sparked a debate on whether Fugu is a good

model organism (Gilley *et al*., 1997; Aparicio and Brenner, 1997; Elgar *et al*., 1997). Miles *et al*. (1998) reported extensive conservation of synteny between a 1.5 Mb region of human chromosome 11 and <100 Kb of the Fugu genome, but a comparative study of seven genes on human chromosome 9 revealed extensive gene order differences within regions of conserved synteny (Gilley and Fried, 1999). A recent study based on chromosome 20 genes identified considerable conservation of synteny, but not good conservation of gene order (Smith *et al*., 2002).

Tetraodon nigroviridis is a freshwater pufferfish 20-30 million years distant from *Fugu rubripes* (Roest Crollius *et al*., 2000b). It has been suggested that studying a species related to Fugu but distant by 20-30 million years would enable the identification of functionally important sequences that appeared after the human/teleostan divergence (Crnogorac-Jurcevic *et al*., 1997). Human:*Tetraodon nigroviridis* comparative analysis was used to provide an estimate for the total human gene number (Crollius *et al*., 2000a), whilst similar analysis was used to estimate the completion of annotation of human chromosome 20 (Deloukas *et al*., 2001). No genetics are available in *Tetraodon nigroviridis*.

The laboratory rat is one of the most important animal models for the genetic mapping of complex phenotypes (James and Lindpaintner, 1997). Over the past ten years approximately 150 genetic loci controlling multifactorial traits have been identified in rats (http://ratmap.gen.gu.se). Rat genomic resources include genetic, RH and EST maps (Bihoreau *et al*., 1997; Watanabe *et al*., 1999; McCarthy *et al*., 2000; Scheetz *et al*., 2001; Bihoreau *et al*., 2001), as well as rat:human:mouse comparative maps (Summers *et*

al., 2001; Kwitek *et al*., 2001). Shotgun sequencing of the rat genome is undertaken at the Baylor College of Medicine (http://www.hgsc.bcm.tmc.edu/projects/rat/) in collaboration with other institutes. DNA sequence alignments of transcribing human:rat orthologous sequences indicate that untranslated exons share on average >68% identity. The mean aligned identity of human/rat coding sequences is 85.9% and the mean aligned identity of human/rat proteins is 88% (Makalowski and Boguski, 1998).

1.6 Functional genomics

Although genome sequence analyses provide a wealth of information on predicted gene products, the majority of these have no known function (Blackstock and Weir, 1999). The generated resources and data can be used to develop and apply global (genome-wide) experimental approaches to assess gene function. This approach (referred to as functional genomics; Hieter and Boguski, 1997) promises to rapidly narrow the gap between sequence and function and provide a molecular understanding of biological processes (Thornton, 2001).

A number of powerful approaches are available to monitor the RNA (transcriptome) and protein (proteome) molecules in a cell. For example, differential-display-reverse transcription PCR (DDRT-PCR) can be used to identify and compare mRNAs during different cell processes. In DDRT-PCR, RNAs isolated from different cell populations are reverse transcribed with a set of degenerate, anchored oligo(dT) primers to generate cDNA pools. This is followed by PCR amplification and labelling using the original

primer and a degenerate one. The products can be visualised on a sequencing gel and differentially expressed molecules can be isolated for further studies (Liang and Pardee, 1992; refinements and modifications reviewed in Liang and Pardee, 1995; Matz and Lukyanov, 1998; Sturtevant, 2000).

Serial analysis of gene expression (SAGE; Velculescu *et al*., 1995) allows the quantitative and simultaneous analysis of a large number of transcripts. Double-stranded cDNA is digested, ligated to linkers and amplified. The linkers contain a restriction site for a type II restriction enzyme that cuts DNA twenty nucleotides away from the recognition site, which is used to digest the cDNA pool. 13-20 bp-long tagged cDNAs are then ligated to generate a library of clones, each representing twenty or more tagged genes. The expression profile can be obtained by sequencing each clone.

Microarray technology provides a format for the simultaneous measurement of the expression level of thousands of genes in a single hybridisation assay. Each array consists of a reproducible pattern of thousands of different DNAs (primarily PCR products or oligonucleotides) attached to a solid support, usually glass. Fluorescently labelled DNA or RNA is hybridised to complementary DNA on the array and signals are detected by laser scanning. Hybridisation intensities for each arrayed DNA sequence are determined using an automated process and converted to a quantitative read-out of relative gene expression levels. The data can then be further analysed to identify expression patterns and variation and to correlate with cellular development, physiology and function (reviewed in Harrington *et al*., 2000; Noordwier and Warren, 2001; Lee and Lee, 2000).

Work describing microarray studies of gene expression across the entire yeast genome were first reported in 1997 (De Risi *et al*., 1997). Since then, numerous transcription profiles in various conditions have been generated (Devaux *et al*., 2001). A similar approach has been used to create a reference database of 300 full-genome expression profiles in yeast corresponding to mutations in ORFs, as well as treatments with compounds with known molecular targets (Hughes *et al*., 2000). The generated profiles identified many co-regulated sets of genes, allowing dissection of transcriptional responses and isolation of candidate genes for many cellular processes.

cDNA microarrays were used to characterise gene expression patterns in 49 mouse tissues. Clustering genes coding for known enzymes into metabolic pathways revealed coordination of expression within each pathway among different tissues (Miki *et al*., 2001). Other applications of the DNA microarray technology include the identification of genome-wide locations and functions of DNA binding proteins (Ren *et al*., 2000; White, 2001) and the experimental annotation of the human genome (Shoemaker *et al*., 2001).

Currently, proteomic analysis relies on a limited number of techniques. These include 2D-gel electrophoresis and mass spectrometry (MS) to separate and analyse thousands of proteins, and ICAT (isotope-coded affinity tag)/MS technology for qualitative and quantitative comparisons of complex protein mixtures. Applications such as the yeast two-hybrid system and phage display are also used to study protein-protein interactions (proteomic analysis methods reviewed in Pandey and Mann, 2000; Dutt and Lee, 2000; Legrain and Selig, 2000; Martin and Nelson, 2001; Yaspo, 2001; Lee, 2001).

1.7 Human variation

The central aim of genetics is to correlate specific molecular variation with phenotypic changes by exploiting the polymorphic nature of the genome. The human genome sequence is not one sequence but rather many variations on a common theme, each of which alters the inherent molecular circuitry, and thus, consequent phenotypes, in a specific manner (Chakravarti, 1999).

Since the mid-1980s sequence variations such as RFLPs (Botstein *et* al., 1980), minisatellites (Jeffreys *et* al., 1985) and microsatellites (Weber and May, 1989) (also see section 1.2.2.1) have been successfully used in genome-wide linkage and positional cloning analyses to identify hundreds of genes for human diseases (Collins, 1995; http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM). Unfortunately, most successes in finding genes that contribute to disease risk have been for highly penetrant diseases, caused by single genes. The identification of genes involved in complex diseases requires the use of dense genetic maps that could be obtained through the systematic discovery, analysis and characterisation of SNPs (Collins *et al*., 1998b; International SNP Map Working Group (ISNPMWG), 2001).

SNPs are single base-pair substitutions in genomic DNA at which different alleles exist with a frequency of at least 1% in one or more populations (Brookes, 1999; Table 1.5). About 90% of sequence variants in humans are SNPs (Collins *et al*., 1998b) and comparison of two haploid genomes yields one SNP per 1,331 bp (ISNPMWG, 2001).

Analysis of all chromosomes from 40 individuals is expected to identify 17 million SNPs, of which 500,000 will map within coding regions (Collins *et al*., 1998b).

Table 1.5: Essential SNP facts (Risch, 2000; Taylor *et al*., 2001; Collins *et al*., 1998b).

- Defined by a frequency of >1% in at least one population
- Stable inheritance
- Building block of haplotypes
- Estimated density of 1 in ~2 Kb (when 2 chromosomes are compared)
- Bi-allelic, suitable for high-throughput analysis
- Topological classification
  Coding amino acid change (non-synonymous, non-conservative aa change)
  Coding amino acid change (non-synonymous, conservative aa change)
  No change in amino acid (synonymous coding)
  Non-coding (5' or 3' UTR)
  Other non-coding (intra- and inter-genic regions)
- More stable, more numerous and easier to score than microsatellite repeat variants

1.7.1 SNP identification

SNP discovery and characterisation is a multi-step process (Kwok and Gu, 1999). Sequences such as random STS and/or EST sequences can be used for identifying candidate SNPs (Wang *et al*., 1998a; Gu *et al*., 1998; Deutsch *et al*., 2001; Picoult-Newberg *et al*., 1999; Irizarry *et al*., 2000). More systematic approaches were centred on the emerging sequence of the human genome. For example, candidate SNPs were identified by analysing the sequence overlaps from the clone tile paths (Dawson *et al*., 2001; Taillon-Miller *et al*., 1998), or by reduced representation shotgun (RRS) sequencing (Altshuler *et al*., 2000; Mullikin *et al*., 2000). In the RRS approach, pooled

DNA from a group of individuals is digested with restriction enzyme(s) and size fractionated on an agarose gel. Fragments approximately 1.5 Kb in size are isolated and used to construct a small insert library that is shotgun sequenced to 2-5-fold redundant coverage. The sequence traces are then aligned and compared for mismatches (SNPs). The sequence reads can also be aligned to the reference sequence of the human genome to identify additional SNPs.

In February 2001 the ISNPMWG published a map of the human genome containing 1.42 million SNPs (one SNP every ~1.9 Kb of available sequence). The SNP Consortium (TSC, http://snp.cshl.org/; Marshall, 1999) and the HGP identified over 95% of the SNPs by using the RRS sequencing approach and the sequence from overlapping clones, respectively (ISNPMWG, 2001).

The most comprehensive public repository of SNPs is the dbSNP database (http://www.ncbi.nlm.nih.gov/SNP/index.html; Sherry *et al*., 2001; Barnes, 2002). Established in 1998, dbSNP currently contains 4.2 million SNPs (Build 104 - May 2002) that can be grouped into a non-redundant set of 2.6 million SNPs (RefSNPs; http://www.ncbi.nlm.nih.gov/SNP/snp_summary.cgi). dbSNP is currently integrated with other large public variation databases such as the NCI CGAP-GAI database of EST-derived SNPs (http://lpgws.nci.nih.gov/; Masood, 1999), the TSC (http://snp.cshl.org/; Masood, 1999) and HGBASE (http://hgbase.cgr.ki.se; Brookes *et al*., 2000).

Validation studies on a small subset of TSC and HGP candidate SNPs showed that more than 80% are polymorphic and that 50% have a minor allele frequency of ≥20% (Marth

et al., 2001; ISNPMWG, 2001). In a recent publication Kruglyak and Nickerson (2001) suggest that the 2001 SNP map comprises 11-12% of all human sequence variations. In addition, they estimate that obtaining a nearly complete catalogue (95%) of human polymorphic sites (allele frequency >1%) would require a comparison between 96 haploid genomes and highlight the fact that the biggest challenge lying ahead is not SNP discovery, but SNP analysis (genotyping) (Kruglyak and Nickerson, 2001).

1.7.2 SNP analysis

Most current SNP analysis methods rely on PCR amplification of the sequence of interest, which is then tested for the presence, or absence of the polymorphism using an assay system. The multitude of assay systems in use are described in Landegren *et al*. (1998), Gut (2001), Jenkins and Gibson (2002), Syvänen (2001), and summarised in Figure 1.3. Some examples are also discussed below.

The most straightforward gel-based assay exploits the introduction/removal of a restriction enzyme site at the polymorphic site. Primers designed on either site of the SNP are used to PCR amplify the sequence of interest. Restriction digest and differential migration on an agarose gel by electrophoresis of the PCR product indicates the presence or absence of the restriction site in the DNA sample tested (Dawson *et al*., 2001). Alternatively, the presence of a SNP in a PCR product can be determined by the Sanger method of sequencing (Wang *et al*., 1998a; Taillon-Miller *et al*., 1998; Mullikin *et al*., 2000; Deutsch *et al*., 2001; ISNPMWG, 2001).

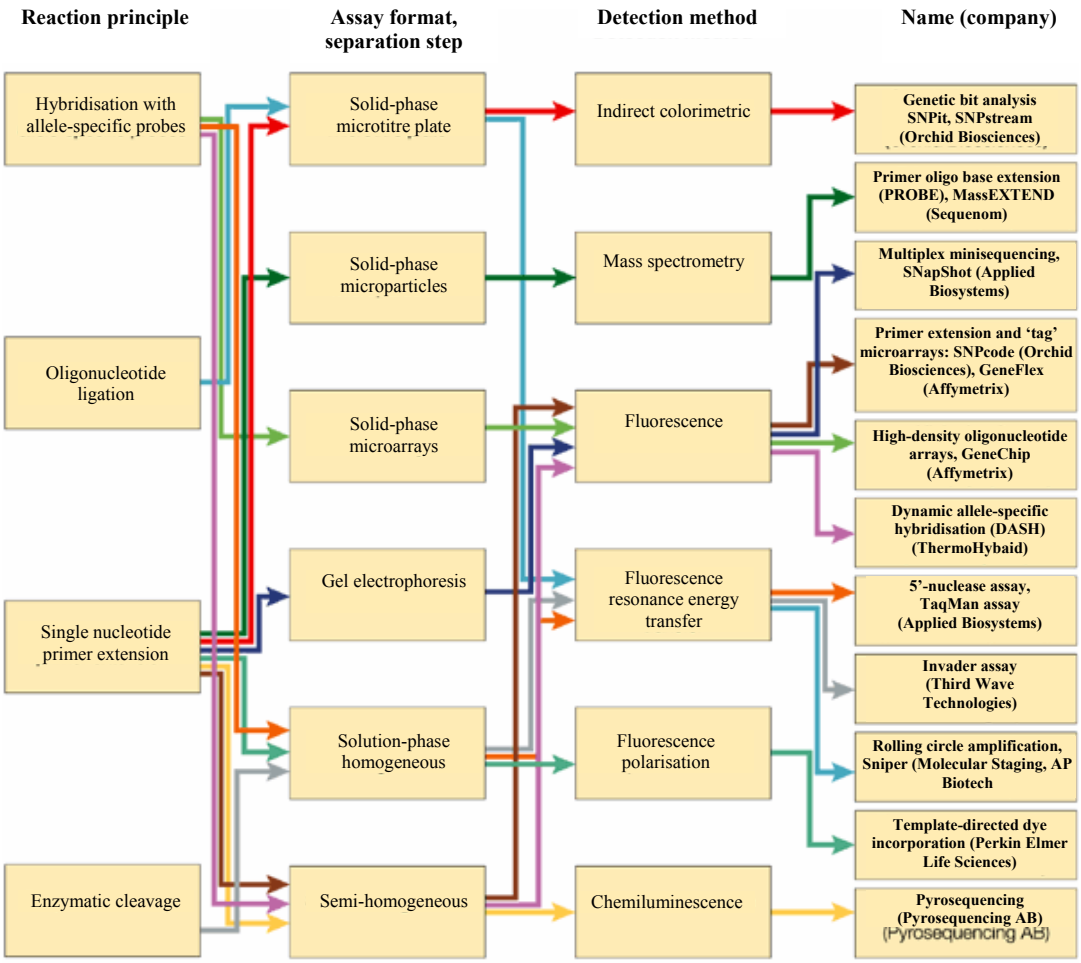


Figure 1.3 (reproduced from Syvänen, 2001): "Modular" design of some of the assays for SNP genotyping. Coloured arrows are used to show the reaction principles, assay format and detection methods that make up a particular genotyping method. For example, the TaqMan™ assay involves hybridisation with allele-specific oligonucleotides, a solution-phase assay and detection by fluorescence resonance energy transfer. The figure illustrates principles for assay design, and the list of assays is not intended to be comprehensive.

The Invader assay was used to systematically analyse SNPs during a whole-chromosome-22 study (Dawson *et al*., 2002). This method utilises the invasive cleavage of oligonucleotide probes (Lyamichev *et al*., 1999) in a two-stage reaction set-up. The first step involves the hybridisation of two target-specific hybridisation oligonucleotides, an allele-specific signalling probe with a 5'-region that is non-complementary to the target sequence and an upstream Invader oligonucleotide. When the allele-specific probe is

perfectly matched at the SNP, the three-dimensional structure formed by the oligonucleotides and the target sequence at the SNP is recognised and is cleaved by a 5'-endonuclease, called FLAP endonuclease, which is specific for this structure. The cleavage releases the 5'-sequence of the signalling probe, which can be detected using fluorescence- or mass spectrometry-based detection.

The TaqMan™ and Molecular Beacon approaches employ Allele-Specific Oligonucleotide (ASO) hybridisation coupled to fluorescence detection. They are both based on an energy transfer principle in which fluorescence is detected as a result of a change in physical distance between a reported fluorophore and a quencher molecule.

In the TaqMan™ (or 5' nuclease allelic discrimination) method, the region flanking the polymorphism is amplified in the presence of allele-specific probes (labelled with a different fluorophore). Probes have a fluorophore, called reporter, at the 5' end and a quencher at the 3' end that absorbs fluorescence from the reporter (Livak *et al*., 1995). During the extension phase of PCR, the DNA polymerase encounters probes specifically base-paired with its target and unwinds them. The 5'→3' exonuclease activity of the DNA polymerase degrades the partially unwound probes and liberates the reporter fluorophore from the quencher, thereby increasing net fluorescence. Mismatched probes are displaced from the target without degradation (Livak, 1999; Holloway *et al*., 1999; Ranade *et al*., 2001).

Molecular Beacon probes consist of sequence that is complementary to the target sequence and a short stretch of self-complementary 5' and 3' nucleotides with a

fluorophore at the 5' end and a quencher at the 3' end. When not hybridised to a target sequence the probes adopt a stem-loop conformation, bringing the fluorophore and quencher pair close together, thereby extinguishing the donor fluorescence. When the probes hybridise to a perfectly matched target during the primer-annealing phase of PCR, the stem-loop structure opens, and the distance between the quencher and fluorophore increases, resulting in a 900-fold increase in fluorescence. Fluorescence from mismatched probes is quenched because they readily adopt the stem-loop structure, enabling allele discrimination (Tyagi and Kramer, 1996; Tyagi *et al*., 1998; Marras *et al*., 1999).

Arrays of oligonucleotides (DNA chips) can also be used to detect human variation in a number of different ways (Hacia and Collins 1999). For example SNP detection can be achieved by carrying out multiplex ASO reactions on microarrays that carry several probes for each SNP to be analysed (Hacia *et al*., 1998). Alternatively, hybridised targets can be used as templates for the extension of the immobilised probes (mini-sequencing; Pastinen *et al*., 1997; Pastinen *et al*., 2000).

The mass spectrometry-based approaches are discussed in section 5.1.

1.7.3 Utilising SNP data

There are great hopes that SNP data can be used to improve biological understanding and to advance medicine (Isaksson *et al*., 2000). Shen *et al*. (1999) used structural mapping and structure-based targeting strategies to show that SNPs can have marked effects on the structural folds of mRNAs. These results suggest that phenotypic consequences of SNPs

could arise from mechanisms that involve allele-specific structural motifs in mRNA, which could influence a diverse range of events such as mRNA splicing, processing, translational control and regulation.

SNPs that map in coding exons can also affect protein function, if they result in amino acid substitutions (non-synonymous cSNPs). The impact of such an amino acid replacement on the three-dimensional structure and function of a protein can be predicted computationally with methods that rely on protein structure information from previous research, and/or the structural context of known disease-causing mis-sense mutations and/or the amino acid similarity with homologous proteins (Chasman and Adams, 2001; Sunyaev *et al*., 2001; Wang and Moult, 2001; Ng and Henikoff, 2002). Such analyses predict that (i) the SNP discovery efforts have discovered several hundred of nsSNPs (ii) in the genome of the average human there are hundreds of nsSNPs that have a direct effect on protein function (a subset of which could impact on human health).

Pharmacogenetics (the study of how genetic differences influence the variability in patients' responses to drugs) holds great promise for the optimisation of new drug development and the individualisation of clinical therapeutics. It is predicted that the use of pharmacogenetics in pre-marketing clinical trials will enable a greater percentage of those trials to produce significant results, because patients whose genetic profile suggests that the drug will be harmful or ineffective to them will be intentionally excluded (Pfost *et al*., 2000). In addition, physicians will be able to use genetic testing to predict the patient's response to a drug, which can aid in individual dosing of medications or avoidance of side effects (Roses, 2000; Pfost *et al*., 2000; Chakravarti, 2001).

The most challenging role envisaged for SNPs is their use for the identification of genetic factors in common disease. For example, SNPs can be used as genetic markers in linkage equilibrium studies (Kruglyak, 1997; Carlson *et al*., 2001). Simulations by Kruglyak (1997) indicate that a map of 700-900 moderately polymorphic SNPs is equivalent to, and a map of 1,500-3,000 superior to, a 300-400 microsatellite marker set. More importantly, they can also be used in association studies of complex diseases to test whether a SNP is enriched in patients compared to suitable controls (Risch and Merikangas, 1996; Risch, 2000; Gray *et al*., 2000; Nowotny *et al*., 2001).

Marker (SNP) selection is critical to the success of such studies. The densities of SNPs to be used depend on the haplotype features of the region they map to. Haplotype information is obtained by determining the combinations of SNPs that are inherited together on the same DNA strand using linkage disequilibrium (LD) analysis.

LD analysis measures the degree of association between two genetic markers. The extent of LD in the human genome is still under debate, mainly because of lack of experimental evidence. The emerging experimental evidence suggests that the human genome can be parsed objectively into haplotype blocks: sizeable regions over which there is little evidence of historical recombination, and within which only a few common haplotypes are observed. It is also suggested that the boundaries of blocks and specific haplotypes they contain are highly correlated across populations (Patil *et al*., 2001; Olivier *et al*., 2001; Gabriel *et al*., 2002). LD will be discussed in more detail in section 5.1.

1.8 Chromosome 20

Chromosome 20 is a metacentric chromosome and represents ~2.2% of the human genome (Morton, 1991). The clone map, which was assembled by fingerprinting and STS content analysis (Bentley *et al*., 2001), is in six contigs of which one spans the entire p arm. The four gaps in the q arm have all been sized by fibre FISH and together account for less than 320 Kb of DNA. 59,421,637 bp of non-redundant sequence generated from 629 overlapping clones represents 99.4% of euchromatic DNA (Deloukas *et al*., 2001).

The finished sequence was analysed on a clone-by-clone basis using a combination of similarity searches and *ab initio* gene predictions, followed by manual annotation of 895 gene features. Excluding pseudogenes, chromosome 20 has a gene density of 12.18 genes/Mb, which is intermediate to 6.71 (low) and 16.31 (high) reported for chromosome 21 and 22, respectively. 32,763 unique SNPs have been identified on chromosome 20. Comparative analysis of chromosome 20 to mouse and *Tetraodon nigroviridis* genomic sequences indicates that the current analysis may account for over 95% of all coding exons and almost all genes (Deloukas *et al*., 2001). The above analysis included data generated by this study (discussed in chapter III).

The best-known disorders linked to chromosome 20 (in total 46 are reported by OMIM (May 2002), http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM) are Creutzfeldt-Jacob disease (Collinge *et al*., 1996) and severe combined immunodeficiency (Wiginton *et al*., 1986). The resources generated by the Chromosome 20 Project accelerated the cloning of disease genes such as those involved in the Alagille (JAG1; Li *et al*., 1997),

the McKusick-Kaufman (MKKS; Stone *et al*., 2000), the ICF (DNMT3B; Hansen *et al*., 1999) and the Hallervorden-Spatz (PANK2; Zhou *et al*., 2001) syndromes. In addition, the candidate disease loci for myeloproliferative disorders, myelodyplastic syndromes and congenital hereditary endothelial dystrophy were refined (Bench *et al*., 2000; Ebenezer *et al*., unpublished). Other disorders linked to chromosome 20 for which the underlying genetic defect is still unknown include type 2 diabetes, cataract, obesity and Graves disease.

The studies described in this thesis (chapters III, IV, and V) are centred on the region of 20q12-13.2, which is linked to many characterised and uncharacterised candidate disease loci. Deletion of the long arm of chromosome 20 represents the most common chromosomal abnormality associated with the myeloproliferative disorders (MPD) and is also found in other malignancies including myelodysplastic syndromes (MDS) and acute myeloid leukaemia (AML) (Bench *et al*., 2000). Earlier studies identified a common deleted region in 20q11.2-13.1 spanning circa 11 Mb (Wang *et al*., 1998b; Bench *et al*., 1998; Bench *et al*., 2000; Wang *et al*., 2000). MDS/AML and MPD associated with a del(20q) are distinct diseases, raising the possibility that multiple target genes on 20q are involved in the pathogenesis of these neoplastic disorders (Wang *et al*., 2000). Acute lymphoblastic leukaemia (ALL) is the most common malignancy in childhood and has also been mapped to 20q (Chambon-Pautas *et al*., 1998; Couque *et al*., 1999).

Autoimmune disease occurs when the immune system mounts a response directed against self (Gough, 2000). The autoimmune thyroid diseases (AITDs) including Graves disease (GD) and Hashimoto's thyroiditis (HT) are the commonest human autoimmune diseases

and are responsible for significant morbidity in pre-menopausal women. The pathogenesis of both diseases appears to develop a result of a complex interaction between predisposing genes and environmental triggers (Tomer and Davies, 1997; Tomer *et al*, 1999). A whole genome linkage study of families with AITD was used to identify three loci that are linked with GD, one of which (GD2) was identified on 20q11.2 (Tomer *et al*., 1998; Tomer *et al*., 1999; Barbesino *et al*., 1998; Pearce *et al*., 1999). GD2 showed the strongest evidence for linkage to GD and it was fine-mapped to a 1 cM interval between markers D20S107 and D20S108 (Tomer *et al*., 1999).

Diabetes is a heterogeneous disorder characterised by a chronic elevation of plasma glucose and its mode of inheritance remains unclear (Rich, 1990). Several studies reported evidence for linkage on chromosome 20q for type 2 diabetes (Ji *et al*., 1997; Bowden *et al*., 1997; Zouali *et al*., 1997; Ghosh *et al*., 1999; Ghosh *et al*., 2000).

The interest in identifying and characterising disease-causing genes prompted the construction of several physical maps of the region. Physical YAC/PAC/BAC clone maps (such as those described in Bench *et al*., 1998; Wang *et al*., 1998b; Price *et al*., 1999; Wang *et al*., 2000) were constructed and used to position genes utilising the available information from resources such as the Gene Map (Deloukas *et al*., 1998). The chromosome 20 mapping effort (Bentley *et al*., 2001) produced a PAC/BAC contiguous map across the region of 20q12-13.2. A set of 111 minimally overlapping clones was selected and sequenced as part of the chromosome 20 sequencing project (Deloukas *et al*., 2001).

1.9 This thesis

The raw DNA sequence is the output of large-scale genome sequencing. This thesis aims to show ways of increasing our knowledge about the genetic information content of this product in order to maximise the impact of the Human Genome Project on genome biology. A 10 Mb region on human chromosome 20q12-13.2 (representing 1/6th of the whole chromosome) provided the basis for the following studies:

- Computational and experimental annotation of sequence features (chapter III). Three experimental approaches were used to confirm and/or extend gene structures and the advantages and disadvantages of each approach are discussed. All annotated features were studied in terms of total sequence coverage (e.g. exon sizes). Splice sites and isoforms, as well as first-pass expression data obtained for the studied transcripts are also discussed. Estimates for gene annotation completion were obtained using different computational approaches and the translated annotated coding features were compared to the proteome of other species. The sequence environment of gene features was also investigated.

- Human:mouse comparative studies (chapter IV). This chapter describes the construction of a bacterial clone map for a mouse chromosome 2 region syntenic to human 20q12-13.2, using a homology-based, gene-orientated approach. The mouse map was used to select a sequence tile path and the annotation and analysis of the generated mouse sequence is described. The human:mouse comparative sequence analysis findings are also discussed.

- Human variation across 20q12-13.2 (chapter V). In this final results chapter, I discuss the analysis of human variation across the region. The use of a data mining approach to identify more than a hundred novel cSNPs is described. Circa 2,000 candidate SNPs were selected for genotyping across three populations (Caucasians, Asians and African Americans). The genotypes obtained from 95 Caucasian individuals (from twelve CEPH families) were then used to generate a first generation linkage disequilibrium map of the region.