Chapter 1

Introduction

In 1920, German botanist Hans Winkler first used the term ‘genome’, reputedly by the fusion of GENe and chromosOME, in order to describe the complex notion of the entire set of chromosomes and all the genes contained within an organism. A great deal of progress has since been made in the elucidation of the complex molecular interactions that underlie cellular functioning and the syntenic relationship between organisms at a nucleotide level.

The basis for these advances was the characterisation of the structure of DNA by Watson and Crick in 1953 (Watson *et al*., 1953), and the realisation that DNA could be decoded to provide a guide to genetic inheritance. This underpinned the concept of genetics and gave scientists the possibility to explore and quantify the nature and extent of the biological information passed on from one generation to the next. The characterisation of biological inheritance permitted the elucidation of what it was that was being encoded and how it could determine biochemical function. Finally, extending from elucidation of the mechanisms behind inheritance of monogenic diseases, scientists are beginning to grasp how sequence is also involved in complex interactions, occasionally under the influence of environmental factors, to contribute to many (but still not all) diseases. Whilst the generation of the complete sequence of the human genome may provide the starting point for the characterisation of all human disease, clinical diagnosis and classification by specialists remains central to the appropriate treatment of individuals suffering genetic disease.

The speed at which the vast amount of human sequence data was generated can be attributed to the evolution of strategies and techniques developed to sequence organisms such, as bacteria (Kohara *et al*.,1987), yeast (Olson *et al*., 1986) and the nematode worm

(Coulson *et al*., 1986). The availability of such an evolutionary diverse collection of sequences, with the addition of mouse (MGSC 2002) and other complex multi-cellular organisms, has also enabled comparisons to be made at a nucleotide level. These inter-species sequence comparisons, in conjunction with direct experimentation and computer based prediction programs, is facilitating the identification of evolutionarily conserved sequences, such as genes, and, to a lesser extent, the motifs that regulate them (Pennacchio *et al*., 2001).

The elucidation of the molecular complexity of gene structure and determining their regulation is, however, only the first step in understanding the intricate networks in which genes interact. Though sequence analysis and homology matching may assist to define a gene on the nucleotide level, determining the structure of the gene's product, the protein, and its function, is a difficult paradigm to resolve. Whilst traditional methods, for example X-ray crystallography, are capable of characterising the structure of a protein their speed of application precludes them from large scale protein analysis. However, *in silico* modelling using previously determined domains or three dimensional structures may provide a means of inferring function and help characterise the networks in which they are involved (Skolnick *et al*., 2000).

The identification of nucleotide differences (polymorphisms) between individuals, by comparison of high quality sequence, will assist our understanding of phenotypic variation. The localisation of single nucleotide polymorphisms (SNPs), particularly within functionally important sequences, such as genes, will contribute to our understanding of the aetiology and susceptibility to human disease and responsiveness to biochemical treatment. The identification of SNPs may also provide the opportunity to partition the human genome into ancestral segments that have undergone minimal evolutionary recombination (haplotype blocks). Haplotype blocks, identified by linkage disequilibrium mapping, can then be used as a means of identifying multiple genes associated with complex phenotypes within unrelated individuals, where family-based studies are impossible because the complexity of the factors which contribute to the phenotype obscure any familial component. It is through approaches such as this that the true impact of the human genome sequence on human health and disease may bear the most fruit.

1.1 Mapping and sequencing model genomes

The first genomes that were characterised were relatively small by current standards, for example bacteriophage ϕX174 (5 kb; Sanger *et al*., 1977, Sanger *et al*., 1978) and bacteriophage λ (48 kb; Sanger *et al*., 1982), but they provided the underlying techniques and strategies that are being used for the more complex organisms currently being studied. Chain termination sequencing, developed by Sanger *et al* was a synthetic method, in which the nested sets of labelled fragments which constituted the sequence ladder were generated *in vitro* by a DNA polymerase reaction. The method was highly sensitive and robust. It was therefore amenable to biochemical optimisation to produce long, accurate sequence reads; and also to automation, which was necessary for large-scale application of the technique. In these respects it differed from the method of Maxam and Gilbert (1977), which necessitated production of all the labelled material prior to chemical degradation to form the sequence ladders of nested fragments. As a result, the synthetic method has remained the technique by which the majority of genomic sequence from a variety of complex organisms is presently being generated, see figure 1.1. Neither method was capable of generating single reads of greater than 2-300 nucleotides, limited in part by the sequence ladder production itself, and partly by the ability to separate the sequence ladder by gel electrophoresis at single-base resolution (even today sequencing read-lengths approaching 1kb are rare). Assembly of larger tracts of DNA therefore required the development of methods to re-assemble a single sequence from multiple individual reads. Two approaches were adopted for this; first, the use of maps of restriction fragments, where multiple enzymes with sequence-specific cleavage activity are used, singly or in combination, to order and orientate segments of the sequence, which could be individually selected for sequencing; second, the use of the information gained from each individual sequence read to order and orient each segment relative to overlapping neighbours. This required the development of advanced computer programs to make the task possible on all but the smallest scale. In a further modification, the random shotgun strategy used by Anderson *et al*., (1981a) to elucidate the mitochondrial genome involved using a random fragmentation process, by partial DNAse I digestion (Anderson *et al*., 1981b). This removed the dependence on sequence-specific restriction enzymes, while still relying on sequence-based assembly of contiguous tracts of overlapping reads.

The random shotgun approach provided the basis of the strategies used to assemble sequences of large inserts cloned in plasmids, lambda phage and cosmid vectors, and also the later bacterial artificial chromosome (BAC) and P1-derived artificial chromosome (PAC) clones. The same strategy was adopted to sequence the 1.8 Mb genome of the bacterium *Haemophilus influenzae* (*H. influenzae*) (Fleischmann *et al*., 1995). Whilst the whole genome shotgun sequencing approach has proven itself to be a successful strategy for the rapid assembly of smaller genomes, there are, however, doubts as to whether this strategy is suitable for assembling the sequence of complex organisms (as discussed in section 1.3).

The generation of a physical map, in which the genome is divided into bacterial clone units of 40 -200 kb and assembled into contiguous stretches (contigs) of overlapping clones, is a process analogous to the sequence contig assembly process. In contrast to sequence assembly, however, the information used to compare individual clones and identify overlaps, for the *C. elegans* (Coulson *et al*., 1986) and *S. cerevisiae* (Olson *et al*., 1986) genome projects, was a one-dimensional fingerprint, prepared by separating restriction fragments from a limit digest of each cloned DNA by electrophoresis. Overlaps between clones were detected on the basis of partially (or completely) shared fingerprint patterns. An alternative approach to identify overlapping relationships between clones was to test clones for the presence of characterised markers. Overlaps between clones could be identified on the basis that they shared a single copy sequence. The presence of the sequence was identified using a specific hybridisation probe or PCR assay.

Given a physical map of overlapping clones, individual clones can then be selected from the map to provide maximum genomic coverage with minimal redundancy. These clones permit specific regions to be targeted for further investigation, and in particular for determination of the complete DNA sequence separately from the other clones. Because the source of the genomic sequence is limited to an individual clone, problems encountered with sequence assemblies are greatly reduced compared to the corresponding whole genome assemblies.

At the time of their inception, the physical maps of the *Caenorhabditis elegans* (*C. elegans*) (Coulson *et al*., 1986) and *Saccharomyces cerevisiae* (*S. cerevisiae*) (Olson *et al*., 1986) genomes were constructed to enhance the molecular genetics of the respective organisms by facilitating the cloning of known genes and to serve as an archive for genomic information. However, the data associated with the construction of the clonal physical maps – even with good alignment to the genetic map – carried only a tiny proportion of information present within the genome. Consequently, a minimum tile path of the 30 kb cosmid and 15 kb lambda clones, used to build the physical maps of the *C. elegans* and *S. cerevisiae*, respectively, were subcloned into M13 phage vectors (1.3-2 kb insert size) and sequenced on a per clone basis. The physical maps of the two genomes, and subsequently of *Escherichia coli* (*E. coli*) (Kohara *et al*., 1987), *Arabidopsis thaliana* (*A. thaliana*) (Arabidopsis Genome Initiative, The, 2000), *Drosophila melanogaster* (*D. melanogaster*) (Hoskins *et al*., 2000) and human (McPherson *et al*., 2000), used restriction enzyme fragments in various ways to overlap clonal units for the construction of genome wide physical maps.

For the *C. elegans* project, polyacrylamide gel electrophoresis was used to resolve DNA fragments that had been generated by digesting cosmid DNA with two restriction enzymes, *Hin*d III and Sau 3AI. Restriction fragments, of which the *Hin*d III ends had been labelled with a radioactive molecule, were then detected by exposure to autoradiograph film. Digitised cosmid specific 'fingerprints' were analysed by pair-wise comparison to establish contigs of overlapping bacterial clones within a seven fold genomic fingerprint data set (Coulson *et al*., 1986). The mapping of *S. cerevisiae* used a similar contig construction strategy but alternatively, fingerprints were generated by a single restriction digest of lambda DNA and the clone fragments separated on an agarose gel prior to reassembly into contigs (Olson *et al*., 1986). Whilst both methods generated large tracts of genomic contig coverage, gaps remained in the physical maps. To compensate for the regions lacking cloned representation in the *C. elegans* map, fosmids, maintained as single copy within the host cell (Kim *et al*., 1992), and large insert yeast artificial chromosome (YACs), (Burke *et al*., 1987), were incorporated. Fosmids proved to be most useful generating bridging coverage in central, gene rich regions of chromosomes, whilst YACs tended to generate *de novo* map coverage on the more repeat-dense chromosome arms (A Coulson *pers comm*.).

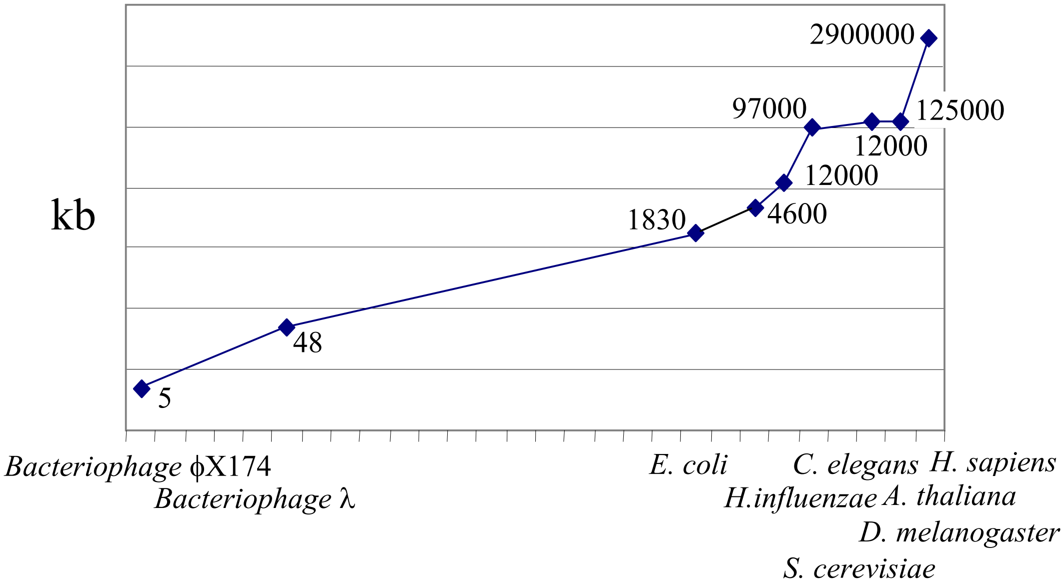


Figure 1.1 A plot of the near logarithmic increase in the complexity of genomic sequencing from the first full genomic sequence of Bacteriophage ϕX174 in 1977 to the anticipated completion in the human genome in April 2003.

1.2 Mapping and sequencing the human genome

The human genome is contained within 22 autosomes (numbered from 1 – 22, largely according to size) and two sex chromosomes, X and Y (female XX and male XY). Chromosomes are punctuated with centromere structures that are either located close to chromosome ends (acrocentric), towards a chromosome end (submetacentric) or centrally between ends (metacentric). The initial size estimate of the genome, 3200 Mb, was based largely upon cytometric measurements (Morton 1991) and has since been revised to 29000 Mb in light of the higher resolution human draft sequence analysis (IHGSC, 2001) supported by observed sizes of completed chromosome sequences, which suggested the earlier figures were over-estimates (Dunham *et al*., 1999, Hattori *et al*., 2000, Deloukas P *et al*., 2001). The construction of a map of the human genome was an important step towards understanding and characterising the sequence contained within it as it provided a means by which all the features could be ordered and partitioned, and the task of detailed characterisation and sequencing could be divided up into manageable segments.

1.2.1 Cytogenetic mapping

The treatment of metaphase chromosome spreads with trypsin digestion and Giemsa staining creates differential chromosome banding patterns. The generation of light (R-bands) and dark (G-bands) bands by Giemsa staining is reliant upon nucleotide content and the staining pattern therefore reflects the base composition and correlates other properties of the different regions (see table 1.1). However, the maximum genome-wide resolution was limited to an 850 genome-wide banding pattern (Bickmore *et al*., 1989). The recognition of characteristic banding patterns of chromosomal regions provided the basis for much of the early characterisation of chromosome aberrations (duplications, deletions and translocations) which were associated with clinical phenotypes (Pinkel *et al.*, 1988, Tkachuk *et al.*, 1990, Dauwerse *et al.*, 1990).

The ability to hybridise labelled probes containing specific sequences and to detect their location on metaphase chromosome by autoradiographic or fluorescent detection techniques (fluorescence *in situ* hybridisation (FISH) (Pinkel *et al.*, 1986)) revolutionised cytogenetic mapping. Initially, the location of the probe relative to the metaphase banding pattern provided an approximate map position for the sequence represented by the probe. Pairs of markers, differentially labelled, could be simultaneously placed relative to the cytogenetic banding, and also ordered with respect to each other. The use of pairs of differentially labelled markers in combination with a third reference marker enabled FISH to be applied to chromosomal DNA in a less condensed state (in interphase nuclei). Although no banding pattern can be obtained in interphase DNA, the decondensed state of the chromatin relative to metaphase chromosomes means that increased levels of resolution could be obtained, as probes were better separated. An inter-probe distance of 1- 5 Mb can be resolved using metaphase FISH, 0.1 – 1.0 Mb by interphase FISH (Wilke *et al.*, 1994), and 5 kb by FISH using mechanical pre-treatment to extend DNA into fibres (Heiskanen *et al.*, 1994).

Table 1.1: Comparison of G-bands and R-bands

G-bands	**R-bands**
Dark staining Giemsa bands	Light Staining Giemsa bands
AT rich	GC rich
Late replicating	Early replicating
Early condensation	Late condensation
DNase insensitive	DNase sensitive
SINE poor, LINE rich	SINE rich, LINE poor
Gene poor	Gene rich
Less frequent recombination	More frequent recombination

Adapted from Bernardi (1989)

1.2.2 Genetic mapping

Genetics maps utilise the likelihood of recombination between adjacent markers during meiosis to calculate inter-marker genetic distances, and from this to infer a physical distance. The closer two landmarks are together on a chromosome, the less likelihood there is of a recombination event occurring between, with the opposite being true for markers that are further apart. The calculation of distance, and therefore the metric upon which the genetic map is based, is the length of the chromosomal segment that, on average, undergoes one exchange with a sister chromatid during meiosis, the Morgan (M). Therefore, a 1% recombination frequency is equivalent to 1 centimorgan (cM), and, since the human genome covers 3000 cM and contains approximately 30000 Mb, 1cM is approximately equivalent to 1 Mb. However, recombination is known to be non-random which can lead to a level of inaccuracy (Dib *et al*., 1996) in inferring physical distances from measurements of genetic recombination.

The inherent limitation of primary genetic maps was the lack of availability of polymorphic markers between which genetic distances could be calculated. This was ameliorated in part by the suggested use of restriction fragment polymorphisms (RFLPs), identified by Kan and Dozy (1978), for the construction of a genome wide genetic linkage map (Botstein *et al*., 1980).The first such map (Donis-Keller *et al*., 1987) was limited in its usefulness, however, due to RFLPs having a maximum heterozygosity of 50% and the low level of resolution of the 403 characterised polymorphic markers, including 393 RFLPs, covering the genome. The identification of hypervariable regions, which showed multi-allelic variation (Wyman and White 1980), provided a new source of markers for genetic mapping. The variable regions contained short 11 to 60 bp variable number tandem repeats which showed allelic variation. However, these minisatellite markers (Jeffreys *et al*., 1985) and variable number tandem repeats (VNTRs) (Nakamura *et al*., 1987) were shown to cluster at chromosome arms and were not inherently stable (Royle *et al*., 1988). The identification of microsatellite markers (containing di-, tri- or tetra nucleotide repeats) greatly facilitated the generation of genetic maps. They were proven to be widely distributed throughout the genome, showed allelic variation (Litt and Luty 1989, Weber and May, 1989), were amenable to PCR amplification (Saiki *et al*., 1988) by sequence-tagged-site screening (Olson *et al*., 1989). In a relatively short period of time a number of genetic maps were published with increasing marker density and resolutions, culminating in the most recent deCODE genetic map which contains 5136 markers genotyped across 1257 meioses (Kong *et al*., 2002), table 1.2.

Table 1.2: A comparison of marker content within genetic maps.

No. of Markers	Reference
100	Hudson *et al*., 1992
813	Weissenbach *et al.,* 1992
2066	Gyapay *et al*., 1994
5840	Murray *et al*., 1994
5264	Dib *et al*., 1996
5136	Kong *et al*., 2002

1.2.3 Radiation hybrid mapping

The utilisation of somatic cell hybrids to maintain human genomic fragments, such as whole chromosomes or chromosomal regions, permits the generation of another form of mapping resource to be generated, the radiation hybrid map. The modification of a technique that fragmented human chromosomes by irradiation and which were then rescued by fusion to rodent cells (Goss and Harris 1975) prompted Cox *et al*., (1990) to propose that radiation hybrid (RH) mapping could be applied to the construction of long range maps of mammalian chromosomes.

The premise of the technique is similar to that of the genetic map, i.e. the more closely related two markers are related within the genome the less likelihood there is of a radiation induced break in between them in a reference panel of cell lines, and hence the less likely is their segregation to different chromosomal locations based on association of the markers to different sets of fragments. As the presence of two markers within a radiation fragment gives no indication to their physical distance a panel of radiation hybrids was required. By estimating the frequency of breakage, and thus the distance between two markers, it is possible to determine their order. The unit of map distance is the centiRay (cR) and represents 1% probability of breakage between two markers for given a radiation dose. Unlike the level of information garnered from a genetic marker, that may or may not be informative within a varying number of meioses, the radiation hybrid marker is either positive or negative for a DNA fragment, effectively digitising PCR results. Any amplifiable single copy sequence can therefore be placed in a radiation hybrid map. The radiation hybrid mapping technique has been used for the construction of high resolution gene maps (Schuler *et al*., 1996, Deloukas *et al*., 1998) and has also been used to supplement the construction of chromosome physical maps (Mungall *et al*., 1996, Deloukas *et al*., 2001, chapter 4).

1.2.4 Physical mapping

The generation of a physical map relies upon the construction of an ordered and orientated set of clone based contigs. The term “contig” was coined by Staden (Staden 1980) to refer to a contiguous set of overlapping segments which together represent a consensus region. These segments can be sequence, or clones, whose overlapping relationship is defined by information in common to each pair of overlapping segments. The overlaps are identified by performing a pair wise comparison of the dataset associated with each segment. Similarities that are statistically significant indicate the presence, and sometimes the extent, of overlap. Bacterial clone contigs are the most convenient route for the sequence generation of larger genomes. They presenting a means of coordinating physical mapping and, because of the way in which they are constructed, provide an optimal set of clones (the tile path) for sequencing.

1.2.4.1 YAC Maps

The main benefit of using YACs for the constructing of a physical map is that the insert size (up to 2 Mb) results in coverage of large regions of the genome with relatively few clones. Green and Olson (1990) utilised YACs to construct a physical map across the cystic fibrosis region on human chromosome 7 by overlapping YACs by STS content data. Chromosome specific (Chumakov *et al*., 1992, Foote *et al*., 1992) and genome wide YAC maps have also been published (Chumakov *et al*., 1995, Hudson *et al*., 1995). Though STS content mapping is the most frequent method used to generate YAC contigs, techniques such as repeat mediated fingerprinting, either by *Alu*-PCR (Coffey *et al*., 1992) or by repeat content hybridisation (Cohen *et al*., l993), have also been used.

The advantages of using YACs are, however, offset by the relative difficulty of constructing YAC libraries and of analysing the cloned DNA, compared to the use of bacterial cloning systems. Many YAC clones have also been found to be chimeric, that is, to contain fragments derived from non-contiguous parts of genomic DNA being cloned (Green *et al*., 1991, Bates *et al*., 1992, Slim *et al*., 1993). Rather than being used as a primary sequence resource, YACs became more generally used to support the construction of detailed landmark maps, and to underpin sequence ready bacterial clone maps (Collins *et al*., 1995; Bouffard *et al*., 1997). Recently, YACs have been used to facilitate gap closure in the bacterial clone maps by linking contigs (Coulson *et al*., 1995). The links are identified by STS content mapping. In these cases the YACs have been sequenced directly.

1.2.4.2 Bacterial Clone Maps

In contrast to YACs, bacterial clone libraries are easier to make and the cloned DNA is more easily manipulated. Chimerism is low (Shizuya *et al*., 1992, Ioannou *et al.*, 1994), and the supercoiled recombinant DNA can be purified readily from the host DNA. An important factor influencing the construction of bacterial clone contigs is the available genomic resources. Whilst the *C. elegans* and *S. cerevisiae* maps utilised total genomic 30 kb cosmid and 15 kb lambda libraries, in 7- and 5-fold coverage respectively, current bacterial clone contig construction utilises large insert P1-derived artificial chromosome (PAC) (Iaonnou *et al.*, 1994) and bacterial artificial chromosome (BAC) (Shizuya *et al*., 1992) libraries. Each BAC or PAC clone typically contains an insert of 100 – 300 kb and maps have been constructed from a >15 fold genomic clone coverage.

1.3 Generating human genomic sequence

The elucidation of the all genic and other features contained within the human genome is reliant upon the generation of high quality sequence. Two different strategies were adopted to produce human genomic sequence; the first , utilised by the International Human Genome Sequencing Consortium (IHGSC), is a hierarchical approach utilising bacterial clones from a well characterised sequence ready map as the basis for generating genomic sequence; the second strategy is a whole genome shotgun (WGS) approach, adopted by Celera, which relies upon the assembly of a consensus sequence from randomly derived genomic sequence reads (Venter *et al*., 2001).

The IHGSC used an approach in which a sequence ready bacterial clone map was constructed by utilising high density panel of markers (15 markers / Mb) (Olson 1993, Bentley *et al*., 2000) derived from genetic and radiation hybrid maps. Large insert bacterial clones, primarily PACs (Ioannou *et al*., 1994) and BACs (Shizuya *et al*., 1992), were identified by hybridisation and overlapped by restriction digest fingerprinting (Gregory *et al*., 1997, Marra *et al*., 1997), with STS content data (Olson *et al*., 1989) supporting ambiguous overlaps. DNA generated from each one of a minimally overlapping set of clones from the physical map (tile path) was fragmented into 1.4 - 2.2 kb units, subcloned into M13 or plasmid vectors (Bankier *et al*., 1987) and then sequenced by modified chain termination sequencing (see chapter 5). One of the benefits of generating sequence on a clone by clone basis is that if the consensus sequence generated by the shotgun phase is not contiguous, the bacterial clone from which the sequence was derived can be then used for directed finishing experiments. The production of finished sequence, which is >99.99% accurate and without the presence of gaps, is one of the main differences between the WGS and hierarchical approaches. Whilst the private WGS strategy was declared to be complete following the assembly of the shotgun sequence, The public hierarchical approach, provided an initial draft covering 90% of the genome in 2000; and its production of highly accurate sequence, which will serve as a long-term reference, is now 87% done and due to be completed in April 2003. The results of the private WGS strategy were also announced in 2000 following the assembly of the shotgun sequence, incorporating a representative set of sequences from the public domain draft; however no subsequent finishing was undertaken on this product.

The elucidation of complete genomic sequence by the generation of whole genome shotgun data, as used by Celera, is a relatively simple approach that was first used for the characterisation of the mitochondrial genome (Anderson *et al*., 1981). In 2001, Venter *et al.,* (2001) published their applied strategy to the elucidation of the human genome by assembling sequence data from complete inserts or ends of 1 – 2 kb and 50 kb subcloned plasmids, respectively, and by the incorporation of BAC end sequences. The publication of the human genome sequence in, on average, 100kb scaffolds proved that a considerable amount of human sequence could be generated and assembled in a relatively short period of time. However, some doubts remain as to the success of the approach as a sole means of assembling a complex genome (Waterston *et al*., 2002, Myers *et al*., 2002, Green *et al*., 2002). Without the incorporation of the publicly available sequence data, and the inherent map information associated with it, highly repetitive motifs and low copy duplications introduce errors into the assembly of the sequence.

1.4 Interpreting the human genome landscape

The complete characterisation of the human genome will not be restricted to experimental or *in silico* identification of coding features and the elements affecting their expression. A full description of the long range sequence composition will be necessary to, amongst other things, facilitate an understanding of the coordinated regulation of genes, identify the under lying reasons for repeat mediated genomic rearrangements and to provide a basis for the identification of variation between populations.

1.4.1 Sequence composition

The human genome has previously been shown to contain considerable variation in its nucleotide composition. Separation of mammalian genomic DNA on caesium gradients indicated compositional heterogeneity (Thiery *et al*., 1976) whilst fractionation of human DNA on Cs_2SO_4-Ag^+ gradients showed a broad range of GC based separation (Bernardi *et al*., 1985). Regional partitioning of genomic sequence based on GC content can be used as a low resolution means of determining biological properties such as cytogenetic banding (Hurst and Eyre-Walker 2000), repeat composition and gene density (Zoubak *et al*., 1996, Gardiner *et al*., 1996) and structure (Oliver *et al*., 1996). Analysis of the human draft sequence which, at the time of publication covered an estimated 94% of the genome, revealed GC content could vary by as much as 59.3% to 33%, from the genome average of 41%, within a 300 kb region (IHGSC, 2001). It was also estimated that 98% of large insert clones mapping to the darkest Giemsa staining bands contained below average GC content, whilst 80% of the clones mapping to light bands contained higher than average GC content.

The fractionation of human DNA on Cs_2SO_4-Ag^+ led Bernardi *et al.,* (1985) to suggest that DNA fractions could be classified based on their relatively homogeneous GC content (isochores). GC-poor isochore family members, L1 and L2, contained <38% and 38-42%, GC respectively, whilst heavy (GC rich) family members contain 42-47%, 47-52% and >52%, GC respectively. However, the debate continues as to whether isochores provide a useful means of partitioning the sequence contained within the human genome (IHGSC, 2001, Oliver http://genomebiology.com/)

1.4.2 CpG island identification

CpG islands are short stretches of hypomethylated DNA that are rich in GC nucleotides and have a CpG:GpC ratio approaching or exceeding 1:1. Their occurrence within human genomic sequence is one fifth of the expected 4% frequency, by multiplying the relative fractions of G's and C's within the genome (0.21 X 0.21)) (IHGSC, 2001) due to the spontaneous deamination of the methyl-C residue which give rise to a thymine (Coulondre *et al*., 1978). By contrast, deamination of non-methylated cytosine produces uracil residues which are recognised and repaired as cytosine by the cell. The identification of CpG islands within human sequence is significant because they are often associated with the 5' ends of genes (Bird *et al*., 1986, Gardiner-Garden and Frommer 1987). It has been estimated that 56% of human genes are associated with CpG islands (Antequera and Bird 1993), including constitutively expressed genes, and approximately 40% of genes with tissue specific patterns of expression (Larsen *et al*., 1992). Analysis of the draft human sequence identified 28,890 putative CpG islands (putative as no information on the methylation status was obtained) within sequence from which repeats had been removed by RepeatMasker (Smit and Green, unpublished) (IHGSC, 2001).

Whilst variation in the length (up to 36,619 bp), and density (2.9 - 22 CpG islands / Mb) of putative CpG islands was observed, their estimated number closely matched previous estimates (Antequera and Bird 1993).

1.4.3 Repeat content

Analysis of the human draft sequence revealed that repetitive motifs account for at least 50% of the genomic content (IHGSC, 2001). These repeat motifs can be roughly divided into five different classes,

1) transposon-derived (interspersed) repeats
2) inactive retroposed copies of genes (processed pseudogenes)
3) simple sequence repeats (e.g. $(A)_n$, $(CA)_n$, $(CGG)_n$)
4) segmental duplications (10 – 300 kb duplicated within the genome)
5) tandemly repeated sequences (centromeres / telomeres / RNA gene clusters).

The transposon-derived interspersed repeats account for more than 90% of all repeats currently identified (IHGSC, 2001). This class of repeats, which includes short-interspersed-elements (SINES), long interspersed elements (LINES), LTR retrotransposons and DNA transposons, accounts for 13%, 20%, 8% and 3% of the draft sequence, respectively. In general the SINE and LINE elements, as has previously been reported (Soriano *et al.*, 1983), show an inversely proportional distribution by genomic GC content. LINE elements are found at a higher density in AT-rich, GC-poor, regions whilst SINEs are found in AT-poor, GC-rich regions. The prevalence of LINE elements in AT-rich regions, where gene density is lower, is logical from the perspective that these genomic parasites would not present a mutational burden to their host. Whereas SINEs, that utilise LINE transposon machinery for replication (Jurka *et al*., 1997), do not co-localise with LINEs and are in fact found in GC-rich regions at a four fold higher density than in GC-poor region. It has been proposed that SINEs may somehow target GC-rich regions for insertion or that they are co-distributed similarly to LINEs, in gene-poor regions, but their distribution is subsequently reshaped by evolutionary forces (IHGSC, 2001) or that they are randomly integrated but are then fixed preferentially in GC-rich DNA (Smit *et al*., 1999).

1.5 Gene identification

It is estimated that 5% of the human genome contains coding sequence (IHGSC, 2001), with as little 1 – 2% encoding for protein (Green *et al*., 2001). It is therefore important that the structures of genes are clearly and accurately defined above the background of apparent non-coding and repetitive sequence. It is known that human genes encode RNAs that, for example, facilitate the expression of protein-coding genes by the excision of introns by the spliceosome (small nuclear ribonucleoproteins -snRNA), participate in translational machinery (ribosomal RNA - rRNA and transfer RNA - tRNA), or act as the template for the transcription of messenger RNA (mRNA) which is subsequently translated into a protein product. The majority of the protein-coding genes will most likely be identified within human genomic sequence by *in silico* gene prediction with the support of cDNA and EST sequence alignment. However, techniques that were developed to identify genes from the transcript rather than the sequence will be central to the elucidation of human genes that are difficult to characterise, and to assist identification of genes within other organisms for which large amounts of sequence data will not be available.

Prior to the availability of large tracts of human genomic sequence, the identification of protein-coding genes was largely reliant upon the isolation of a transcript from tissue specific mRNA or cDNA clones. These early studies used mRNA enrichment coupled with biological assays of *in vitro* translated products, or cDNAs library hybridisation, using plasmids containing coding fragments, or oligonucleotide mixtures based on peptide sequence, to obtain clones for sequencing to characterise the genes of interest. This highly targeted approach has successfully identified a number of genes, including the rabbit (Rabbitts *et al*., 1976) and human α- β- γ- globins (Little *et al*., 1978), human interferon, IFN (Taniguchi *et al*., 1980) and factor IX (Choo *et al*., 1982).

Genomic fragments, in the form of cloned DNA (cosmids and YACs), have also been used to identify cDNA clones by direct screening of cDNA libraries. This technique proved successful for the identification of the neurofibromatosis type 1 (Wallace *et al*., 1990) and Menkes genes (Chelly *et al*., 1993). Whilst the identification of candidate genes by genomic fragment hybridisation is successful, it relies upon the correct hybridisation kinetics between the exons in the genomic DNA and transcribed sequence. Therefore, genomic hybridisation may not be the most productive method of identifying the gene of interest where the gene or exons may be very short in the genomic sequence.

Direct selection, which can be used to enrich cDNA libraries for genes encoded by large genomic regions, can result in a 1000-fold amplification of target cDNAs. The technique has successfully been applied to the characterisation of genes within a 300 kb region around the G6PD on Xq28 (Sedlacek *et al*., 1993) and a 6.5 Mb region on Xq21, which led to the identification of the X-linked agammaglobulinemia gene (XLA) (Vetrie *et al*., 1993).

Another technique of gene discovery, which utilises the genomic sequence of a gene is exon trapping. Though the technique has successfully been used in several positional cloning projects (Walker *et al*., 1993, Trofatter *et al*., 1993, The Huntington's Disease Collaborative Research Group 1993) its technical complexity precludes it from large-scale application.

The technique of amplifying multiple genes by oligo dT priming (Verma *et al.,* 1972, Wickens *et al*., 1978) or with anchored oligo(dT) priming (Khan *et al*., 1991), together with advances in sequencing technologies, led to the development of a strategy to generate single pass sequence of all human cDNAs. It was suggested that this approach would obviate the need to map and sequence the entire genome (Brenner *et al*., 1990). Large-scale sequencing projects were initiated to generate gene fragments by single pass sequencing from the 5' and 3' ends of cDNA clones (Adams *et al*., 1991, Okubo *et al*., 1992, Adams *et al*., 1993, Sudo *et al*., 1994, Hillier *et al*., 1996). One of the difficulties encountered with this strategy was the representation of genes by multiple expressed sequence tags (ESTs). Consequently, a database (UniGene) was established that rationalised EST data by clustering the sequences into a non-redundant data set (Boguski and Schuler 1995, Schuler *et al*., 1996) (http://www.ncbi.nlm.nih.gov/UniGene/clust.cgi?ORG=Hs). UniGene currently contains (31st October 2002) 3,911,348 EST sequences which have been condensed into 110743 clusters. Almost half of the clusters, 50476, contain a single EST entry, whilst the average cluster contains 5 – 8 ESTs, with the largest cluster containing 16385-32768

ESTs. Another problem encountered with the generation of EST sequence was the contamination of the cDNA libraries with genomic DNA. Poly(A) sequences within the genomic DNA act as templates for the oligo(dT) primer and false cDNA sequence is generated. Additionally, some transcripts have proven to be incomplete, for example, large transcripts that are sometimes not represented if reverse transcription of the mRNA has terminated prematurely. A further problem with the cDNA based approach is that some transcripts may be absent if the gene is not expressed in the tissues used to construct any of the cDNA libraries used for EST sequencing.

Whilst the integration of ESTs within existing radiation hybrid and genetic maps has facilitated the identification of putative candidates for a number of diseases by positional cloning (APOE; Pericak-Vance *et al*., 1991, Strittmatter *et al*., 1993, RET; Mulligan *et al*., 1993) it is their localisation within the human genome sequence that ESTs will prove their greatest worth. Alignment of ESTs to genomic sequence have facilitated the correct annotation of gene structure by assisting to define exon – intron boundaries and helped to identify splice variants, which are present in at least 35% of all human genes (IHGSC, 2001).

Detailed sequence analysis of previously characterised genes has permitted conclusions to be drawn their about generalised genomic structure. Regulatory elements, such as those residing at the 5’ ends of genes, which act as a template upon which transcription factors assemble prior to the initiation of RNA synthesis, have been inherently difficult to identify. This is due to the absence of consistently shared motifs within the genomic sequence. However, the combined occurrence of specific elements, CpG islands and TATA boxes flanked by regions of C-G, permits *in silico* prediction of at least a fraction of the core promoter region to be made. Whilst Eponine (Down and Hubbard 2002) and PromoterInspector (Scherf *et al*., 2000) currently have approximately 50% sensitivity (detecting known TSSs) and 75% specificity (predictions supported by known TSSs) their accuracy is likely to increase when trained upon larger data sets of elucidated promoters.

Unlike prokaryotic organisms, in which genes are located as a single tract of DNA, complex organisms, from yeast onwards, contain genes that are usually segmented by introns of non-coding sequence (Tilghman *et al*., 1978; Gilbert *et al*., 1978). The spliceosome recognises sequence motifs within the intron which leads to their excision, usually a GT di-nucleotide at the 5' end of an intron (splice donor), an AG di-nucleotide at the 3' end (splice acceptor) and an internal branch point (Moor and Sharp 1993).

Localising the site of translation initiation within genomic sequence, unlike the TSS, has been facilitated by the identification of a consensus motif within the genomic sequence. The translation start site, usually an ATG, is typically found in a consensus, GCC^{A}/$_{G}$CC**ATG**G, which includes the two bases which exert the strongest effect, a G at the first base after the translation start, ATG, and a purine (preferably an A) which is located three nucleotides upstream (Kozak *et al*., 1987). The identification of stop codons (TGA, TAA or TAG) and polyadenylation signals (Kessler *et al*., 1986), most commonly as AATAAA (Beaudoing *et al*., 2000), have helped define the 3' ends of genes within genomic sequence.

1.6 Computational Genomics

1.6.1 *In silico* gene prediction

The *in silico* prediction of genes within genomic sequence utilises characteristic sequence motifs associated with genes mentioned above. Whilst initial computer programs were developed to identify single exons, for example GRAIL (Uberbacher *et al*., 1996), and HEXON (Solovyev *et al*., 1994), more recently developed programs attempt to identify complete gene structures. These programs, GENSCAN (Burge *et al*., 1997) and FGENESH (Solovyev *et al*., 1995) predict individual exons based on codon usage and sequence signals and assemble these putative exons into candidate gene structures. The greatest problem associated with using *in silico* gene prediction programs to identify coding structures within genomic sequence is the number of over predictions that are generated. Whilst all genes may be identified by setting low prediction thresholds, the gene structures will be generated with low specificity. Guigo *et al* (Guigo *et al*., 2000) assayed the accuracy of gene prediction methods using an artificially generated data set. They found that GENSCAN accurately predicted 90% of coding nucleotides with 70% of the exons being predicted correctly. However, there was also significant over prediction of 30% based upon predictions in simulated intergenic sequences. The study concluded that it is not currently viable to use computational methods alone to accurately identify the exonic structure of every gene in the human genome (Guigo *et al*., 2000).

The estimated total number of genes contained within the human genome has varied considerably according to the technique that was used to evaluate it and the data that was available at the time. Estimates have been variously based upon average genome and gene size, the number of observed CpG islands (and the proportion associated with genes), redundant and non-redundant EST sequences and, latterly, upon chromosome specific totals, comparative analyse and genome-wide sequence analysis (table 1.3).

Table 1.3: Genes in the human genome

Data set	Gene Number	Date	Reference or source
Hypothetical	100,000	1992	Gilbert *et al*., 1992
CpG Islands	80,000	1993	Antequera *et al*., 1993
EST clusters	60,000-70,000	1994	Fields *et al*., 1994
Unigene clusters	92,000	1996	Schuler *et al*., 1996
Gene sequences	140,000	1999	*IncyteGenomics
Chr. 22 sequence	43,000-61,000	1999	Dunham *et al*., 1999
Chrs 22, 21 sequence	44,000	2000	Hattori *et al*., 2000
Tetraodon seq.	28,000-34,000	2000	Roest-Crollius *et al*., 2000
ESTs in dbEST	120,000	2000	Liang *et al*., 2000
EST and mRNA	35,000	2000	Ewing, B*., et al*., 2000
Draft Sequence	31,000	2001	IHGSC, 2001

*press release available at http://incyte.com/company/news/1999/genes.shtml

1.6.2 Sequence Analysis

Whilst computer programs have been written to predict a number of features associated with coding sequences (table 1.4) the alignment of experimental data is critical to the validation of predicted structures.

Table 1.4: Prediction programs used to identify gene features

Program	Description	Reference
RepeatMasker	Repeat Sequence prediction	Smit and Green, unpublished
CPGFIND	CpG island prediction	Micklem, unpublished
PromoterInspector	Promoter Prediction	Scherf *et al*., 2000
Eponine	TSS prediction	Down and Hubbard, 2002
Hexon	Exon prediction	Solovyev *et al*., 1994
Grail	Exon prediction	Uberbacher *et al*., 1991
GENSCAN	Gene prediction	Burge *et al*., 1997
FGENESH	Gene prediction	Solovyev *et al*., 1995

Programs such as CLUSTALW (Thompson *et al*., 1994), which is used to align multiple nucleotide or protein sequences, and DOTTER, which utilises pair-wise local sequence alignment strategy (Sonnhammer *et al*., 1995), are useful for inferring structural and functional conservation by sequence homology. Sequence homology searching can also be performed using SSAHA (Ning *et al*., 2001), Exonerate (Slater, unpublished) and BLAT (Kent *et al*., 2002) which permits homology sequences to be identified within gigabases of DNA. BLAST (Basic Local Alignment Search Tool), which measures the local similarity between two sequences (Altschul *et al*., 1990, 1997), is the primary method for identifying protein and DNA sequence similarities prior to incorporation of the features into project specific ACeDB databases (Durbin and Thierry-Meig 1994) (http://www.acedb.org/) or genome browsers. One of the major advantages of using BLAST for sequence alignment is the flexibility with which nucleotide and amino acid sequences can be aligned (table 1.5)

Table 1.5: Sequence queries available using BLAST alignment

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

From Brenner (1998)

Whilst BLAST alignment of human mRNA, EST and protein sequences to predicted coding structures provides a primary level of support, predicted features can also be supported by alignments with sequence from other organisms for comparison (comparative sequence analysis). The identification of sequences that are conserved between species is important because sequences that contain elements that are potentially functional are more likely to retain their sequence than non-functional segments, under the constraints of natural selection during evolution. The evolutionary distance between species is an important consideration. Sequence comparisons between closely related species may facilitate the identification of gene structures and regulatory elements but, if the evolutionary distance between the species is relatively small, these sequences may be obscured by non-functional sequence conservation. Therefore a variety of species, including more distantly related species may be required to identify potential functional sequences using the comparative approach.

The identification of conserved sequences by comparative analysis has focused on the identification of non-coding regions (Hardison *et al*., 1993; Koop *et al*., 1994; *et al*., 1997; Hardison *et al*., 1997) and protein coding regions (Makalowski *et al*., 1996; Ansari-Lari *et al*., 1998; Jang *et al*., 1999) between human and mouse genomes. The alignment of sequence from multiple organisms has also been used to identify upstream regions that may affect gene expression. Gottgens *et al* (Gottgens *et al*., 2000) used the alignment of human, mouse and chicken sequences to identify a novel neural enhancer element in their elucidation of the human stem cell leukaemia (SCL) gene region. Whilst comparative sequence analysis may not identify all control regions associated with a gene, conserved regions may be identified that would be candidates for further experimental investigation (Pennacchio *et al*., 2001). Large scale sequencing and comparative analyses is progressing on a number of different organisms (table 1.6).

Table 1.6: A list of the large scale comparative organisms sequencing projects

Organism	**Genome Size (Mb)**	**Reference**
E. coli	4.6	Blattner *et al.*, 1997
S cerevisiae	12	Goffeau *et al.*, 1996
C. elegans	97	The C. elegans sequencing consortium, 1998
D. melanogaster	120	Adams *et al.*, 2000
M. musculus	2600	The Mouse Genome Sequencing Consortium, 2002*
D. rerio	1600	on going
F. rubripes	400	on going
R. rattus	2800	on going
T. nigroviridis	350	on going

* draft sequence publication

The availability of large tracts of human genomic sequence has necessitated the development of databases (genome browsers) that provide a framework upon which the enormous amount of data associated with the human genome can be stored and displayed. The two main databases, Ensembl, developed at the Sanger Institute and the European Bioinformatics Institute (http://www.ensembl.org/Homo_sapiens/), and the University of California Santa Cruz (UCSC) genome browser. (http://genome.cse.ucsc.edu/) contain information pertaining to physical maps, chromosome specific sequence assemblies, aligned mRNAs and ESTs, cross species homologies, SNPs and repeat elements. The development of generic genome browsers, as such as those hosted by Ensembl, makes possible the rapid identification of homologous sequences between comparative organisms and in doing so assist to identify conserved features that may be of some functional significance.

1.7 Allelic variation

Most differences between individuals, at the nucleotide level, can be attributed to allelic sequence variation. The characterisation of sequence differences and comprehension of how these genomic variations affect the expression and function of genes will be crucial for the study of molecular alterations in human disease. Whilst sequence variation has previously been used for genome-wide linkage and positional cloning studies (leading to the identification of many disease causing genes (see 1.2.2)), the association of single nucleotide polymorphisms (SNPs) with genes, either by mapping or as causal sequence variants, promises to be a valuable method in the future for identifying genes involved in complex diseases.

Approximately 90% of the allelic differences existing within the human genome can be attributed to SNPs, the remainder being insertions or deletions (Collins 1998b). A comparison of any two diploid genomes is estimated to identify one SNP per 1.3 kb which has an allele frequency of > 1% (ISNPMWG, 2001). The prevalence of SNPs in the genome, their existence as bi-allelic variants and their stability through inheritance makes them amenable to large-scale high through-put analyses. SNPs will, therefore, be applied to several research areas, including 1) large-scale genome analysis of linkage disequilibrium and haplotype patterns, 2) genetic analysis of simple and complex disease states, and 3) genetics and diversity of human populations.

1.7.1 SNP discovery

De novo candidate SNPs were initially identified by the alignment of STSs and ESTs to available genomic sequence (Wang *et al*., 1998, Picoult-Newberg *et al*., 1999, Irizarry *et al*., 2000, Deutsch *et al*., 2001). The clone based strategy used by the Human Genome Project for the large-scale production of human genomic sequence contributed to a dramatic increase in the SNP numbers by allowing identification of novel SNPs within sequence overlaps between minimum tile path clones (Taillon-Miller *et al*., 1998, Dawson *et al*., 2001). A directed approach to SNP discovery was initiated by sequencing DNA from population specific individuals (Mullikin *et al*., 2000, Altshuler *et al*., 2000). Two to five fold redundant shotgun sequence coverage was generated from 1.5 kb small insert library clones and the resultant sequences were aligned to each other in clusters. As the Human Genome Project progressed, these assemblies and additional shotgun sequence data were aligned to available genomic sequence to identify more SNPs. The total number of SNPs identified using the strategies outlined above culminated in the International SNP Map Working Group (ISNPMWG) constructing a SNP map of the human genome which contained 1.42 million candidate SNPs (ISNPMWG 2001). A proportion of the candidate SNPs were validated experimentally during the project, confirming that >90% were real SNPs. The SNPs identified by The SNP Consortium (TSC) (http://snp.cshl.org, Marshall *et al*., 1999) and the HGP had generated a SNP density of 1 SNP per ~1.9 kb of available sequence.

Akin to the rationalisation that was required to establish a unique set of ESTs, a database was established to generate a non-redundant collection of candidate SNPs (dbSNP) (Sherry *et al*., 2001, http://www.ncbi.nlm.nih.gov/SNP/index.html). dbSNP currently contains 4.8 million entries which have been condensed into a non-redundant set of 3.0 million SNPs, 522,072 of which have been validated to date (build 110, 13th January 2003). Localisation of these unique SNPs within a recent the human sequence assembly (build 30) yields a SNP density of 1 per 1.2 kb. Additional validation of SNPs was carried out on a subset of TSC and HGP candidate SNPs by Marth *et al.*, (2001) by screening 1200 SNPs across 30 individuals from 3 difference populations. Results indicated that 80% of the SNPs were polymorphic in the populations tested, and that 50% had allele frequencies of greater than 20%. Data generated by Kruglyak and Nickerson (Kruglyak and Nickerson 2001) suggests that the number of non-redundant SNPs currently present within dbSNP, in which the minor allele is present in > 1%, comprise 11 - 12% of all single nucleotide sequence variants. The identification of >95% of available SNPs will require analysis of 96 haploid genomes, many more than the number from which the candidate SNPs have been derived so far.

There are many different platforms that have been developed for SNP analysis but which are based upon four basic allele-specific assays types, 1) hybridisation with allele-specific probes, 2) oligonucleotide ligation, 3) single nucleotide primer extension and 4) enzymatic cleavage. Many of the techniques have been developed further and automated in commercial systems. The range of formats used include colorimetric microtitre-plate based assays (Taqman by Applied BioSystems or Invader assay by Third Wave Technologies) or fluorometric methods of detecting SNP alleles that have been separated by gel electrophoresis (Applied Biosystems), fluorometric assay of target hybridised to oligonucleotides immobilised in a microarray chip format (Affymetrix), or immobilised via beads on the ends of arrays light sensing glass fibres (Illumina).

1.7.2 Utilising SNPs

SNPs may be utilised for population genetic studies in order to identify an association between a SNP allele and a specific phenotype. Ultimately the goal of such a study is to identify the causal variant, the mechanism by which the variant has its functional effect. The functional variant will have maximal predictive value in future individual tests, and the gene involved may encode a target protein or mRNA for possible therapeutic intervention. Functional variants may be assayed for using two approaches. The direct approach requires prior availability of a candidate functional variant (e.g. a SNP which alters the encoded protein sequence in a non-conservative way, thus affecting function). The variant is then tested by genotyping a population of defined phenotype and comparing the frequency of one allele with the frequency in a population of matched controls (a case-control study). In the absence of a candidate functional variant, the indirect approach can be taken, in which available SNPs within specific genes (candidate gene association studies) or throughout the genome (genome-wide association studies) can be used to test the same populations.

An aid to the indirect approach is to the identification of allele specific sequence variants and generation of a map of common combinations of specific alleles (or haplotype patterns) that have been largely conserved during the recent population expansions. Among other factors, it is believed that the regions of conserved local haplotype patterns have been maintained by the absence of ancestral recombination within each region. Identification of these conserved segments is facilitated by the availability of SNPs identified by the ISNPMWG. Pairs of alleles can be statistically quantified to determine whether recombination has occurred between them, in which case they are said to be in equilibrium, or if the alleles share evolutionary co-segregation and are therefore in linkage disequilibrium (LD). The generation of an LD map does not require the analysis of related individuals (by comparison to the genetic map), only that they share a common evolutionary history (although inclusion of pedigrees allows direct determination of the phase between SNP alleles and facilitates definition of long range haplotypes). It is hoped that the generation of a map of common haplotype patterns (HapMap) will facilitate the identification of common diseases by indirect association studies as described above (Couzin *et al*., 2002, Harris *et al*., 2002).

The availability of genome sequence with annotated gene structures provides the means to search for candidate functional variants. Build 110 in dbSNP (13th January 2003) contains 60541 SNPs that have been localised to exons, untranslated regions or non-coding regions adjacent to genes (introns or flanking sequence), table 1.7.

Table 1.7: SNP totals contained within or adjacent to coding features.

SNP Count	FUNCTIONAL CLASSIFICATION
20851	Gene region
5228	Synonymous
5220	Non-synonymous
13462	untranslated region
15651	Intron
129	Splice site

SNPS localising within coding a feature (cSNPs) have the greatest potential to affect the structure and function of the gene. The characterisation of allelic variants enables conclusions to be drawn as to whether a specific allele may have an effect upon the amino acid sequence. Slightly more than half of the SNPs localising to coding sequences result in a synonymous change (no change in the amino acid sequence because of codon redundancy) whilst the remaining SNPs result in a non-synonymous change. Non-synonymous changes are further classified as to whether the resultant amino acid has similar biological properties to the ‘normal’ allele, in which case the change is conservative, or if the biological properties of the amino acid are different, then the change is non-conservative. Whilst the molecular significance cSNPs have upon protein structure and function has previously been reported (Chasman *et al*., 2001, Sunyaev *et al*., 2001, Wang and Moult 2001), the effects that SNPs have in non-coding sequence such as splice junctions (Pan *et al*., 2002, Khan *et al*., 2002), folding of mRNAs (Shen *et al*., 1999) and promoter function (Knight *et al*., 1999, Hijikata *et al*., 2000) have also been described.

The identification of SNPs that show an allelic influence on the functioning of proteins, particularly of drug metabolising enzymes, promises a bright future for the optimisation of clinical therapeutics. Associating inherited variations with pharmacological responsiveness provides a basis for the possible development of personalised medicine which will improve the efficacy of drug treatments and decrease the side effects experienced by the individual (Roses *et al*., 2000, 2002, Pfost *et al*., 2000).

1.8 Chromosome 1

Chromosome 1, the largest human chromosome, is estimated to be 263 Mb in length (Morton 1991), which represents 8% of a 3200 Mb human genome. The chromosome is submetacentric and contains a large block of heterochromatin adjacent to the centromere on the long arm which, together with tandemly repeat sequences contained within the centromere and telomeres, reduces the euchromatic size of the chromosome to 214 Mb

(IHGSC, 2001). The chromosome has an average GC content of 43%, compared to the genome average of 41%, with the 40 Mb telomeric region of the short arm containing 47.1% GC (IHGSC, 2001). Draft sequence analysis indicates that chromosome 1 contains slightly more (~11 / Mb) than the genome average of 10.5 CpG islands / Mb, whilst the estimated gene content is noticeably higher (~15 genes / Mb) when compared to the genome average of ~ 11.5 genes / Mb (IHGSC, 2001).

A hierarchical strategy was used to map and sequence chromosome 1, as outlined in section 1.3. To date (15th of February 2003), the sequence ready map of chromosome 1 is contained within 9 bacterial clone contigs from which 2244 minimum tile path clones have been selected for sequencing (including 5 cosmids and 4 YACs). Currently, 95% of the chromosome is contained within finished sequence clones. The chromosome 1 mapping and sequencing project has directly facilitated the elucidation of a number of genetic disease genes, table 1.8.

Table 1.8: Genes that are associated with disease that have been elucidated as a result of the Sanger Institute chromosome 1 mapping and sequencing project.

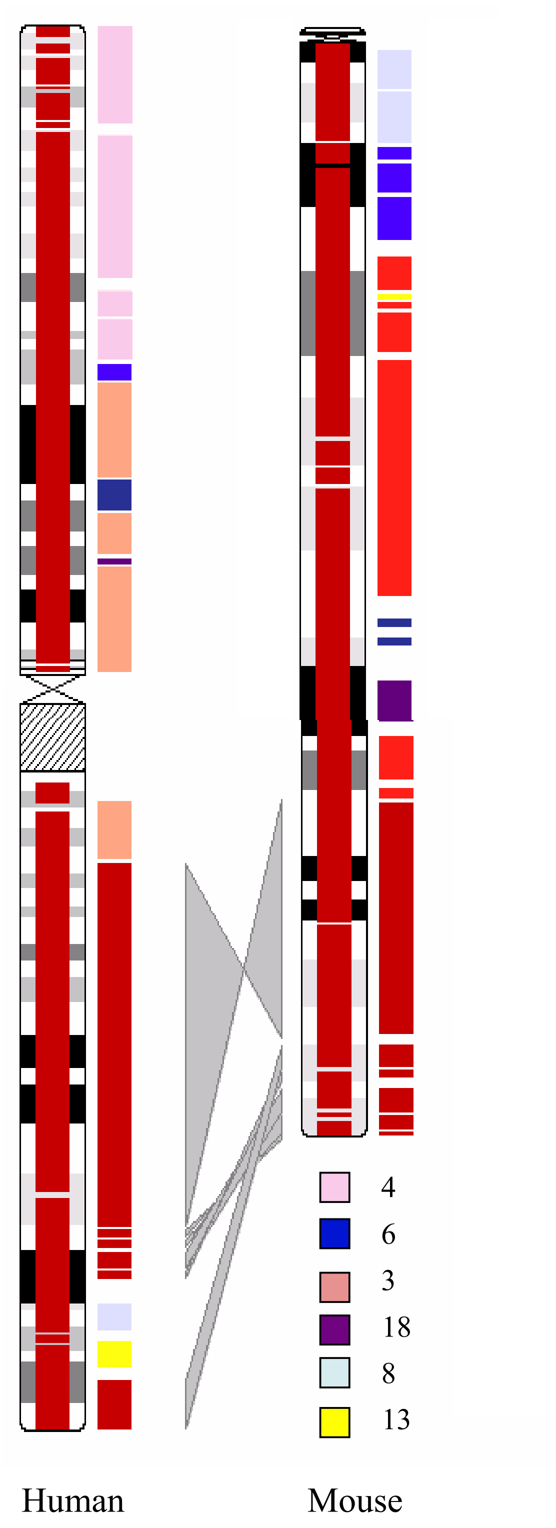
Gene	**Disease**	**Publication**
CACP	Camptodactyly-arthropathy-coxa vara-pericarditis	Marcelino *et al*., 1999
SLC19A2	Thiamine-responsive megaloblastic anaemia	Labay *et al*., 1999
HPC1	Prostate cancer	Carpten *et al*., 2002
LMNA	Partial lipodystrophy	Shackleton *et al*., 2000
CIAS1	Muckle-Wells syndrome	Hoffman *et al*., 2001
IRF6	Van der Woude syndrome	Kondo *et al*., 2002
TBCE	HRD/Autosomal recessive Kenny–Caffey syndrome	Parvari *et al* *et al*., 2002

Included among the 157 genetic diseases localised to chromosome 1 within OMIM (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM) (16th of January 2003) are: Alzheimer disease-4 (Levy-Lahad *et al*., 1995), Chediak-Higashi syndrome (Kritzler *et* *al*., 1964), Fanconi anemia (Loetscher *et al*., 1987), Gaucher disease (Hsia *et al*., 1959) and Usher syndrome, type 2 (Kimberling *et al*., 1990)

Alterations of chromosome 1 are amongst the most common chromosomal abnormalities in human neoplasia. These abnormalities include translocations and other structural rearrangements involving chromosome 1 and other regions of the genome, as well as region specific deletions and amplifications. In general, regions on 1p (the short arm) seem to be most frequently lost, indicating the possible presence of tumour suppressor genes, whereas regions on 1q (the long arm) are more often amplified. Loss of 1p has been detected in meningiomas and oligodendrogliomas (Bello *et al*., 2000, 2002), endometrial hyperplasia (Kiechle *et al*., 2000) and primary gastrointestinal tumors (El-Rifai *et al*., 2000). Gains of 1q have previously been implicated in neoplasia such as, hepatoblastoma (Nagata *et al*., 1999), lymphoma (Rao *et al*., 1999) and sarcoma (Forus *et al*., 1995).

The generation of continuous stretches of genomic sequence has permitted higher resolution synteny maps to be drawn than those based on mapping of orthologous genes or genetic markers. The construction of a physical map (Gregory *et al*., 2002) and the draft sequence (MGSC 2002) has enabled accurate localisation of syntenic boundaries between mouse and human genomes. The availability of more genomic sequence from a broader range of organisms will improve upon our understanding of chromosomal evolution. Figure 1.2 is a representation of the alignment of human and mouse chromosome 1 (also shown are the mouse syntenic blocks contained within human chromosome 1).

A number of physical maps have previously been constructed on chromosome 1. The clones used in their construction have evolved with the availability of new cloning systems, i.e. earlier physical maps were constructed using YACs and cosmids whilst later maps have utilised PACs and BACs. Comprehensive lists of chromosome 1 physical maps have been published within chromosome 1 workshop reviews (Gregory *et al*., 1998, White *et al*., 1999, Schutte *et al*., 2001). Though constructed primarily for the elucidation of disease genes maps have also been generated in difficult to clone regions, such as the telomere on 1q (Xiang *et al*., 2001).



(From Gregory *et al.*, 2002)

Figure 1.2 The alignment of syntenic region between human and mouse chromosomes 1. Bacterial clone coverage on each chromosome is represented by red boxes on the ideograms of each chromosome, whilst chromosomes syntenic to human and mouse chromosome 1 are individually coloured and listed.

1.9 Aims of this thesis

The construction of physical maps, prior to the commencement of this thesis, primarily relied upon the assembly of YAC contigs by STS content hybridisation and, to a lesser extent, cosmid assembly into contigs by radioactive restriction digest fingerprinting. If the goal to generate sequence of the human genome was to be realised, techniques would require development that would enable physical maps to be constructed rapidly and safely. The first part of this thesis describes the development of a fluorescence based restriction digest fingerprinting technique that could be applied to the generation of sequence ready maps. The second aim of this thesis was to apply this fingerprinting, in combination with large-insert bacterial clone library screening, to the generation of a 12 Mb contig within 1pcen – 1p13. Minimum tile path clones from the contig would then selected for sequencing and to assist in elucidating disease causing genes that had been localised to the interval (table 1.9)

. **Table 1.9: Disease loci mapping to 1pcen -1p13**

Disease	**Reference**	**Localisation**
Radiation induced meningioma	Zattara-Cannoni *et al.*, 2001	1p11
Autosomal recessive tachycardia	Lahat *et al.*, 2001	1p11 - 1p13.3
Acute megakaryoblastic leukaemia	Mercher *et al.*, 2001	1p13
Hypothyriodism	Dracopoli *et al.*, 1986	1p13
Achromatopsia	Kohl *et al.*, 2002	1p13
Adrenal Hyperplasia II	Zachmann *et al.*, 1979	1p13.1
Colorectal Cancer	Nitta *et al.*, 1987	1p13.2

The third aim was to use the sequence data generated from the first two sections to characterise the genomic landscape of the interval and to identify as many coding features as possible by *in silico* prediction and experimental support. A family of genes, annotated during the course of this section, and seven other genes of medical interest, form the basis for the final part of the thesis.

The aim was to assemble exon specific sequences from 47 unrelated individuals, of the 12 target genes, to confirm the presence of known, or identify novel, single nucleotide polymorphisms. SNPs identified by this strategy would then be categorised according their occurrence in coding sequence and, where found, further analysis carried out to predict what possible affect they may have upon the structure and function of the resultant protein.