Chapter 1

Introduction

Introduction

The evolution of mapping and sequencing strategies and methodologies is enabling the complete sequence of many genomes to be elucidated. The methods that were used to generate clone maps and sequence of the genomes of smaller model organisms have provided the foundation for analysing the larger and more complex genomes of vertebrates. The DNA sequence of many species, from simple viruses to more complex multicellular organisms such as the nematode worm and the fruitfly, is now available and the finishing of the human genome sequence is scheduled for 2003. The generation of the complete DNA sequence of the mouse and the zebrafish genomes is now well underway. The availability of the genomic sequence facilitates the identification of the biologically significant units encoded within, such as genes and regulatory elements. The genome sequence of different organisms enables comparisons to be made between them in order to both improve the identification of the functional units and to aid in the understanding of their biological significance.

1.1 Mapping and sequencing of model organisms

The ability to generate comprehensive maps and very accurate DNA sequence of large genomes, for example, the human genome, has been made possible because of the pioneering work carried out on smaller genomes. Some of the key organisms for which DNA sequence is now available are listed in Table 1.1.

Table 1.1: *Complete DNA sequence*

Organism	**Size**	**Method**	**Reference**
Bacteriophage ϕX174	5 kb	Plus and Minus	Sanger, F., *et al*., (1977a) Sanger, F., *et al*., (1978)
Bacteriophage λ	48 kb	Random seq	Sanger, F., *et al*., (1982)
H. influenzae	1830 kb	Whole Genome Shotgun	Fleischmann, R. D., *et al*., (1995)
E. coli (K12)	4,600 kb	Clone-based	Blattner, F. R., *et al.,* (1997)
S. cerevisiae	12,000 kb	Clone-based	Goffeau, A., *et al.,* (1996)
C. elegans	97,000 kb	Clone –based	*TCSC (1998)
D. melanogaster	120,000 kb	Whole Genome Shotgun	Adams, M. D., *et al*., (2000)
A. thaliana	125, 000 kb	Clone-based	**TAGI (2000)

* TCSC – C elegans Sequencing Consortium, The
** TAGI – Arabidopsis Genome Initiative, The

The first genome to be sequenced was that of bacteriophage ϕX174 using the plus-minus method based on the elongation of DNA chains with DNA polymerase. However this method was laborious and prone to errors (Sanger, F*., et al.*, 1977a), and so was completely sequenced using chain terminators (Sanger, F*., et al.*, 1977b, Sanger, F*., et al.*, 1978) . Bacteriophage λ was the first organism to be sequenced using the strategy based on sequencing random pieces of DNA (the 'shotgun' approach) (Anderson, S. 1981), in this case restriction fragments (Sanger, F*., et al.*, 1982).The first bacterial genome to be completely sequenced, *Haemophilus influenzae* (*H. influenzae*) (Fleischmann, R. D*., et al.*, 1995), was sequenced using the whole genome shotgun strategy developed for the bacteriophage λ project. In the case of *H. influenzae*, random small insert (2 kb) and large insert (15-20 kb) libraries were constructed and was followed by high throughput DNA sequencing, assembly, sequence editing and annotation.

For the analysis of more complex genomes, a strategy involving the initial generation of a physical map was developed. The discovery of site-specific restriction endonucleases led to the widespread use of restriction mapping for understanding the organisation of DNA fragments. For regions spanning less than 50 kb, restriction maps were constructed routinely. In a few cases, maps as large as 600 kb were constructed, but it proved difficult to extend existing mapping methods beyond that range (Olson, M. V., *et al.*, 1986). For mapping larger regions such as the genomes of *Caenorhabditis elegans* (*C. elegans*) (Coulson, A*., et al.*, 1986) and *Saccharomyces cerevisiae* (*S. cerevisiae*) (Olson, M. V*., et al.*, 1986) two strategies using restriction enzymes were implemented.. For the generation of clone maps covering the *S. cerevisiae* genome, a redundant set of lambda clones was analysed using a single restriction digest, and fragment sizes for each clone compared (Olson, M. V*., et al.*, 1986). For the generation of clone maps covering the *C. elegans* genome, a much larger genome than had been attempted previously, a strategy using cosmids (Collins, J*., et al.*, 1978) was developed. Whole genome restriction digest fingerprinting of cosmids picked at random, and representing a six-fold coverage of the *C. elegans* genome, was carried out. Cosmid DNA was digested with *Hin*d III, the ends of the fragments were labelled with a radioactive molecule and a secondary digest of the labelled *Hin*d III fragments with *Sau*3AI generated fragments of a size that could be resolved on a denaturing polyacrylamide gel. Cosmids with similar fingerprints were interpreted as containing overlapping cloned inserts, and assembled into contigs. In total, 17,500 cosmids were assembled into 700 contigs (Coulson, A*., et al.*, 1986; Coulson, A*., et al.*, 1988). The gaps which persisted between contigs after all the cosmids had been analysed were bridged with yeast artificial chromosomes (YACs) (Burke, D. T*., et al.*, 1987; Coulson, A*., et al.*, 1988). The

YACs had the advantage that DNA fragments greater than that possible for cosmids could be cloned (the YACs were approximately 200 kb in size). The YACs that bridged gaps between contigs accounted for approximately 20% of the nematode genome.

The sequencing of the *C. elegans* genome was carried out on a clone by clone basis and in two phases. In the first phase, each cosmid was subcloned into phage vectors (1.3-2 kb insert size) and the resulting subclones sequenced at random. These random sequences were assembled automatically into sequence contigs. At this point, contigs greater than 2 kb were released into the public domain. The second phase involved a targeted finishing process that enabled contiguous pieces of highly accurate DNA sequence representing the original cosmid to be generated. Finishing involved closing gaps between sequence contigs with targeted resequencing, resolving ambiguities such as GC tracts and improving low quality sequence. Fosmids, a bacterial-based cloning system similar to cosmids but maintained in a single copy number (Kim, U. J*., et al.*, 1992), were incorporated and sequenced in regions where there was no cosmid coverage. Finally, YAC clones were sequenced to close gaps between bacterial clone contigs and in 1998 the completion of the DNA sequence of the nematode worm, the first multicellular organism was announced (C. elegans Sequencing Consortium, The, 1998). Other genomes sequenced by the clone map-based approach included *Escherichia coli* (*E. coli*) (Blattner, F. R*., et al.*, 1997; Kohara, Y*., et al.*, 1987), *S. cerevisiae* (Olson, M. V*., et al.*, 1986; Goffeau, A*., et al.*, 1996) and the first plant *Arabidopsis thaliana* (*A. thaliana*) (Arabidopsis Genome Initiative, The, 2000). *Drosophila melanogaster* (*D. melanogaster*), was sequenced by whole genome shotgun in part as a proof of principal for sequencing the human genome (Adams, M. D*., et al.*, 2000) (discussed later – see Section 1.2).

1.2 Mapping and sequencing the human genome

The advances in the mapping and sequencing of the smaller genomes of organisms such as the nematode worm created the possibility that the complete DNA sequence of the human genome could be generated. The human genome is approximately three gigabases (Gb), which is 30 times larger than that of the worm. It is divided onto twenty-two pairs of autosomal chromosomes and two sex chromosomes (either XY in males or XX in females). One average-sized human chromosome is roughly equivalent to the *C. elegans* genome (100 Mb). The human chromosomes were first characterised by cytogenetic mapping using differential staining and visualisation to identify unique banding patterns for each chromosome. Metaphase chromosome preparations can be treated with trypsin digestion and Giemsa staining and biologically significant differences were observed between the bands (see Table 1.2).

Table 1.2: *Comparison of G-bands and R-bands*

G-bands (Paleogenome)	R-bands (Neogenome)
Dark Staining	Light Staining
Late replicating	Early replicating
Early condensation	Late condensation
DNase insensitive	DNase sensitive
Less frequent recombination	More frequent recombination

There are up to 850 bands that can be visualised by staining (Bickmore, W. A*., et al.*, 1989) and the unique banding enables the identification of each chromosome in the human genome.

At the time the sequencing of the nematode worm began in 1989, small regions of the human genome were being studied because of a specific interest in that region such as association with a particular disease or phenotype. The discovery and characterisation of polymorphic markers within the genome and the development of methods for analysing linkage between them in pedigree studies led to the development and formalisation of genetic mapping. Markers on a genetic map must exist in two or more forms (or alleles) and are said to be polymorphic. This allows for a distinction between different alleles on different chromosomes in a population or different individuals. Naturally occurring DNA polymorphisms are present throughout the human genome and have been utilised for the production of genetic maps.

Genetic maps relate the distance between markers to the likelihood of recombination occuring during meiosis. The closer together two landmarks are on a chromosome, the less likelihood there is of a recombination occuring at meiosis. The reverse is true for markers that are further apart. Distance on a genetic map is measured in centimorgans (cM) and is a measure of the frequency of recombination between two markers. Recombination frequency can be used as an approximate measure of physical distance and is based on the assumption that recombination is random, i.e. there is an equal chance that it will occur at any position in the genome. On this basis, one centimorgan is equivalent to a 1% recombination frequency. The human genome covers 3000 cM and is 3000 Mb in size, therefore 1 cM is corresponds on average to 1 Mb. However recombination is known to be non-random so this is subject to inaccuracy (Dib, C.*, et al.*, 1996).

The first genetic map covering the human genome was based on restriction fragment length polymorphisms (RFLP) (Donis-Keller, H.*, et al.*, 1987), but the low frequency of their occurrence, and the maximum heterozygosity of 50% of RFLPs limited their usefulness. The discovery of hypervariable regions in DNA showing mulitallelic variation within the population (Wyman, A. R.*, et al.*, 1980) provided a new source of markers with major advantages over RFLPs. Both mini-satellite markers (11-60bp repeats) and micro-satellite markers (di-, tri- and tetra-nuceotide repeats) have been utilised in genetic mapping (Nakamura, Y.*, et al.*, 1987), (Weissenbach, J.*, et al.*, 1992), although mini-satellites have an irregular distribution, being more prevalent near telomeres and therefore are less suitable for genetic mapping. The first high resolution, genetic map of the human genome was produced in 1992 (Weissenbach, J.*, et al.*, 1992) using micro-satellites. Subsequent maps followed in 1994 (Buetow, K. H.*, et al.*, 1994; Murray, J. C.*, et al.*, 1994; Gyapay, G., *et al*., 1994) culminating in the version published in 1996 (Dib, C.*, et al.*, 1996) contained 5,264 short tandem repeats with a mean heterozygosity of 70%. The sex-averaged distance covered equals 3,699 cM and the average distance between markers is 1.6 cM. However, a subset of 2032 markers is ordered at high odds (1000:1 or better odds against an alternative order), and these maps have allowed localisation of many monogenic disorders.

The ability to order landmarks across the human genome was becoming increasingly important with the growing interest in genome-wide analysis. A method for marker ordering, similar in some respects to genetic mapping, is hybrid mapping. In the same way that genetic maps rely upon breaks in chromosomes during meiosis, hybrid mapping is based on either naturally occurring physical breaks, such as translocations or deletions, or breaks induced artificially, for example by irradiation.

Naturally occuring chromosomal abnormalities such as translocations, deletions or inversions are useful mapping tools. Somatic cell hybrids have been generated from fusions between human cells containing abnormal chromosomes and rodent cells (e.g hypoxanthine phosphoribosyltransferase (HPRT) negative), where the fusion cell containing the aberrant human chromosome has been recovered using selectable markers (e.g - HPRT, which was used as a selectable marker for incoming human DNA). Panels of somatic cell hybrids have been generated containing different regions of the same chromosome. DNA from each hybrid can be tested for the presence or absence of particular landmarks. Combining the data from different hybrids allows the landmarks to be placed in intervals defined by overlapping chromosome segments, for example in chromosome 22 (Bell, C. J*., et al.*, 1995).

A panel of somatic cell hybrids containing aberrant chromosomes is limited to naturally occurring abnormalities. However techniques are available to induce chromosome breaks randomly. Radiation hybrids (RH) were originally developed by Goss and Harris in 1975, where they fragmented chromosomes by irradiation-induced breakage (Goss, S. J*., et al.*, 1975). Fragments were then rescued by fusion to a rodent cell and hybrids isolated by selection. Multiple cell lines each contain pieces of the human genome, together in overlapping fragments. Individual breaks define order where the closer two markers are together the more likely they are to be maintained on a single fragment. For the analysis of two markers, a matrix is generated and the separation of markers is measured by the number of differences between the two matrices. Markers can be screened against a panel of these hybrids to determine their position in the genome relative to each other and other markers whose location is known. The greater the amount of radiation used, the greater the number of fragments generated, and subsequently the greater the short-range resolution of the panel. Similarly to genetic mapping, the RH map includes a set of framework markers that are ordered at high odds (greater than 1000:1 against a different order). Other markers have then been positioned in 'bins' relative to these framework markers. For instance, radiation panels derived from the whole human genome, such as the GB4 panel and the G3 panel (Gyapay, G*., et al.*, 1996; Stewart, E. A*., et al.*, 1997) have been used to position more than 30,000 ESTs and other genetic and physical markers to produce a gene map of the human genome (Deloukas, P*., et al.*, 1998, updated electronically 1999, see http://www.ncbi.nih.nlm.gov/genemap99).

The plans to sequence the human genome relied upon the construction of clone-based maps. The YAC system was the largest possible cloning system and provided potentially the most efficient way to tackle the problem of assembling long-range clone maps of the larger human genome. YAC maps could be constructed using the markers that had been positioned on both the available genetic and hybrid maps. YAC libraries of the human genome (Anand, R*., et al.*, 1990; Imai, T*., et al.*, 1990; Larin, Z*., et al.*, 1991; Chumakov, I. M*., et al.*, 1992b; Albertsen, H. M*., et al.*, 1990) were generated and improvements in the cloning systems led to the construction of libraries containing clones with an insert size that exceeded 1 Mb (Chumakov, I. M*.,* *et al.*, 1995). The first human YAC contigs were built using shared STSs across disease-associated loci such as the cystic fibrosis gene (Green, E. D*., et al.*, 1990) and the dystrophin gene (Coffey, A. J*., et al.*, 1992) and subsequently larger regions were covered e.g the euchromatic portions of chromosomes Y, 21 and 22 (Foote, S*., et al.*, 1992), (Chumakov, I*., et al.*, 1992b), (Collins, J. E*., et al.*, 1995). The success of these maps lay in the combination of applying unique STSs to map YACs by STS content mapping and in the vectorette end-rescue of YACs (Riley, J*., et al.*, 1990) to isolate new markers from the ends of clones which could be used to develop probes for walking.

However, YACs have two major disadvantages. They are prone to deletions and rearrangements, possibly because of the size of the cloned insert and because of the recombinogenic background of the commonly used yeast hosts. Their other major problem is chimaerism, which is the occurrence of insert DNA from two non-contiguous regions of the human genome. Chimaerism may result either from co-ligation of two fragments to form a single insert, or recombination between two independent recombinants co-transformed in the same host cell (Monaco, A. P*., et al.*, 1994). The STS content approach to mapping YACs substantially overcame these disadvantages for constructing long-range maps. The absence of an STS or a group of STSs in a single YAC indicated the presence of a deletion in the YAC clone. In the case of chimaerism, each STS generated from the end of a YAC was checked by PCR as being from the region of interest where possible, or at least on the same chromosome (by screening a chromosome-specific hybrid cell line).

The lack of a sufficient density of markers to generate YAC maps across the whole genome led Bellane-Chantelot C., *et al* (1992) to develop YAC fingerprinting to compensate for the landmark deficiency (Bellanne-Chantelot, C*., et al.*, 1992). Subsequently, an attempt to generate a YAC-based map covering the human genome using a combination of landmark mapping and YAC fingerprinting was published in 1993 (Cohen, D*., et al.*, 1993). Fingerprints for 33,000 YAC clones were generated by detecting fragments containing medium-repeat sequences and assembled into contigs on the basis of similarly sized fragments. More than 2000 genetic markers, and 5322 novel STSs generated from sequencing PCR products amplified between *Alu* elements, were also used to screen the YACs and detect overlaps between clones. Finally, approximately 500 YACs containing genetically mapped polymorphic STSs (one every 7.4 cM) were positioned on metaphase chromosomes using FISH. The method attempted to mirror the successful approach developed for mapping the nematode worm, using large scale fingerprinting to assemble clones into contigs but with additional organisation from positioning the contigs in the genome with markers, including the polymorphic markers placed on the genetic map. An updated version of the map was published in 1995 (Chumakov, I. M*., et al.*, 1995). However, even though small regions of the YAC map were assembled correctly, the inherent problem of chimaerism assembled non-contiguous portions of the human genome within the same contigs. Combined with the false negative and positive data, this caused major misassemblies of the data, and resulted in a poor final map. Given the inconsistencies that still existed in the YAC maps across the whole genome, construction of YAC maps across single chromosomes continued as groups had more confidence in the data they were generating and resulted in YAC maps covering chromosomes 3 (Gemmill, R. M*., et al.*, 1995), 7 (Bouffard, G. G*., et al.*, 1997), 12

(Krauter, K.*, et al.*, 1995), 16 (Doggett, N. A.*, et al.*, 1995), 21 (Chumakov, I. M.*, et al.*, 1992b), (Nizetic, D.*, et al.*, 1994), 22 (Collins, J. E.*, et al.*, 1995), X (Nagaraja, R.*, et al.*, 1997) and Y (Foote, S.*, et al.*, 1992).

A more successful attempt to generate a physical map across the human genome was carried out using a combination of RH mapping and STS-based YAC mapping (Hudson, T. J.*, et al.*, 1995). The physical map contained 15,086 STSs, of which 10,850 were screened against YACs to produce an integrated map anchored by RH and genetic maps. This approach to YAC mapping was more successful than the previous attempts by Cohen, D., *et al.*, (1993) as it did not rely on fingerprinting to assemble YACs into contigs.

Although some of the YACs that formed part of the clone map covering the nematode worm were being sequenced, the inherent problems of chimaerism, deletions and the difficulty in purifying the YAC DNA from the host DNA, made YAC clones a less suitable substrate for large-scale sequencing than bacterial clones. The majority of the map covering the nematode genome was constructed using the bacterial clones. Initial sequence-ready bacterial clone contig construction in human utilised some of the techniques developed for the nematode worm. For instance, for the early work on sequencing human chromosome 22, although cosmids were identified using the YACs and STSs ordered on the YAC map as a framework, the cosmids were assembled into contigs using restriction digest fingerprinting (see Section 1.1). The development of larger insert bacterial cloning systems (P1 artificial chromosomes - PACs (Ioannou, P. A.*, et al.*, 1994) and bacterial artificial chromosomes - BACs (Shizuya, H.*, et al.*, 1992), with insert sizes of between 130-

150 kb for PACs, and up to 300 kb for BACs, improved the efficiency of map construction. Combined with the density of markers available from the YAC maps bacterial clone contigs could be generated directly from the markers covering the whole of chromosome 22. The availability of a high density of STSs from YAC maps also enabled bacterial clone contig construction to proceed on other chromosomes at an earlier stage, notably the X and Y chromosomes, and chromosome 7 and 21.

Sequencing of the bacterial clones in the maps such as those on chromosome 22 also utilised the procedures that had been developed during the sequencing of smaller genomes such as the nematode worm (see Section 1.1). A minimally overlapping set of bacterial clones were chosen for sequencing on a clone by clone basis, and a random shotgun phase was followed by a targeted finishing process for each clone. The complete sequence of chromosome 22, the first human chromosome to be finished, was published in 1999 (Dunham, I*., et al.*, 1999).

The early construction of bacterial clone contigs on chromosomes such as 7, 16, and 22, and the X and Y chromosome relied upon the high density of ordered markers available from the chromosome-specific YAC maps. For the chromosomes for which no YAC map was available, a sufficient density of markers was achieved from a combination of existing STSs, positioned on available genetic and radiation hybrid maps, and in some cases, novel STSs generated from chromosome-specific sequences (Ross, M*., et al.*, 1997). Radiation hybrid mapping of these STSs produced maps with a sufficient density (average 15 markers per Mb) of ordered STSs that could be used as a framework for map construction using PACs or BACs, and this eliminated the need to generate YAC maps. Chromosome mapping projects that were among the first to benefit from this increased density of markers included chromosomes 1, 2, 6, 19, 20 and 21. The clone map of chromosome 6 is now complete and sequencing is well advanced. The finished sequence of chromosome 20 was recently announced (Deloukas, P.*, et al.*, 2001) with the sequence of chromosome 6 expected to be finished in the spring of 2002.

Other human chromosomes were not covered by sequence-ready maps and in order to generate contigs covering the entire human genome, human genome BAC libraries (initially RPCI-11 and RPCI-13) were fingerprinted using a restriction digest fingerprinting method similar to that developed for the nematode worm. The method consisted of a HindIII digest of BAC DNA, with the fragments resolved on an agarose gel (closely resembling the approach taken to map the *S. cerevisiae* genome, (Olson, M. V.*, et al.*, 1986). Analysis of the fingerprints was carried out using IMAGE and FPC, as was the case for fluorescent fingerprinting. Incorporation of the mapping data already available enabled minimum tiling sets of bacterial clones to be identified representing more than 90% of the human genome.

All of the human sequence generated by early in 2000 had been carried out using the clone-based strategy. Previous sequencing of whole genomes had centred around two strategies, clone-based or whole genome shotgun (see Section 1.1). However, the increased complexity of the human genome led to a debate as to whether the whole genome shotgun strategy was suitable for sequencing such a genome (Green, P., 1997, Venter, J. C.*, et al.*, 1998, Weber, J. L.*, et al.*, 1997). The clone-based approach makes full use of the map information to verify each clone to be sequenced and allows efficient co-ordination to minimise the effort and duplication. For complex genomes, it also avoids major assembly problems due to genome-wide repeats. The whole genome shotgun approach circumvents the construction of clone-based maps and involves generating and assembling random sequence reads from cloned fragments varying in size from 1-2 kb upto 50 kb into sequence contigs or scaffolds, and integration of sequences generated at the ends of BAC clones. In spite of the concerns, a whole genome shotgun strategy was implemented to generate human genomic sequence in a useable form in under two years. Celera, in collaboration with the Berkley *Drosophila* Genome project had used the whole genome shotgun strategy to generate sequence representing approximately 120 Mb of the euchromatic portion of the *D. melanogaster* genome (Adams, M. D*., et al.*, 2000).

Celera (the private effort) was to make the human sequence available in the form of a database which researchers could subscribe to for a fee. The ethos of the Human Genome Project (HGP - the public effort) was to make all the human sequence freely available in the public domain as it was produced. The increasing interest and demand in providing as much sequence of the human genome as quickly as possible in the public domain to support new research led to the formalisation of the existing intermediate, i.e. the assembled shotgun sequence data of each clone, as a defined product, and it was termed the ‘working draft sequence’. Draft sequence mostly contains very accurate sequence (an error rate of less than 1 in 10,000 bases), with some gaps and missassemblies. In February of 2001, the completion of the draft sequence, the first stage of sequencing the human genome, was announced (Lander, E. S*., et al.*, 2001) and provided researchers with human sequence covering up to

90% of the genome, two or more years earlier than the anticipated production of finished sequence.

At the same time Celera announced the completion of the whole genome shotgun sequence assembly of the human genome in 100 kb scaffolds (Venter, J. C.*, et al.*, 2001). A comparison showed that although there was more raw sequence in the Celera data (99% as opposed to 94%), there were fewer contigs in the HGP data and they contained a higher percentage of assembled sequence. One major advantage of the clone-based approach is the ability to carry out a targeted finishing process to produce a high quality product with as few a gaps as possible, and this effort is continuing. To date, the complete sequence of chromosomes 20 (Deloukas, P.*, et al.*, 2001), 21 (Hattori, M.*, et al.*, 2000) and 22 (Dunham, I.*, et al.*, 1999) has been determined, with other chromosomes such as chromosome 6 and chromosome 14 soon to be announced. The completion of the human genome sequence is planned for spring of 2003.

1.3 Interpreting the human genome sequence

The complete DNA sequence of the human genome contains all the information required for the correct development, structure and function of a human being. The identification of the features embodied in the human genome sequence will be one step towards the understanding of the processes that generate and sustain life. As an essential part of this process, the human genome sequence will provide the basis for the identification of all human genes, their organisation and physical position within the genome. The activation and suppression of transcription of these genes are controlled by molecules binding to promoter sequences (usually located immediately upstream of genes) and regulatory elements (which can be located within, nearby or far away from the gene or genes they influence).

Current estimates suggest that at least 50 % of the human genome is made up of repeat sequences (analysis carried out on the draft sequence, (IHGSC, 2001)). These can be identified in human sequence by comparing genomic sequence to databases of prototypic sequences representing repetitive DNA from primates (see http://ftp.genome.washington.edu/cgi-bin/RepeatMasker and http://www.girinst.org/Repbase_Update.html).These can be broadly divided into five classes, (1) transposon-derived interspersed repeats, (2) inactive retroposed copies of genes (processed pseudogenes), (3) simple sequence repeats, (4) segmental duplications and (5) tandemly repeated sequences such as those seen at centromeres and telomeres, which are involved in the maintenance of chromosomes. By far the most common is the transposon-derived interspersed repeat which accounts for more than 80% of all repeats currently recognised, and includes short interspersed elements (SINES) and long interspersed elements (LINES). Analysis of the draft sequence showed segmental duplications at the pericentromic and subtelomeric regions that are present elsewhere in the genome.

It is not fully understood what the remainder of the human genome encodes for, but will include features such as origins of replication and may include sequences such as those specific to chromosome packaging, but an important first goal is to identify all the genes encoded with the human genome.

1.3.1 Gene identification

A gene can be defined as a region of genomic DNA that is transcribed to form a functional RNA molecule. Some of these RNA molecules are processed to form messenger RNAs (mRNAs) and translated to form proteins. Others function within the cell as RNA molecules, such as those that are associated with the ribosome (ribosomal RNAs, rRNAs or transfer RNAs, tRNAs), or the spliceosome (e.g snoRNAs). All genes will have a transcription start site, and those mRNAs that code for proteins will also have a translation start site and a translation termination site. Individual genes of simple organisms such as prokaryotes are located within a single stretch of genomic DNA, but genes of the more complex organisms, from yeast onwards, are composed of exons interrupted by introns which are removed from the RNA molecule during processing steps that occur after transcription. The process of gene identification utilises these features involved in the correct functioning of the genes notably codon usage and splice junctions, as they can be identified within the genomic sequence, along with other features associated with genes such as CpG islands.

Estimates for the number of genes in the human genome have varied widely depending on the type and the interpretation of information used. These are summarised in Table 1.3.

Table 1.3: *Genes in the human genome*

Data set	Gene Number	Date	Reference or source
Hypothetical	100,000	1992	Gilbert, W., *et al.,*1992
CpG Islands	80,000	1993	Antequera, F., *et al*., 1993
EST clusters	60,000-70,000	1994	Fields, C., *et al*., 1994
Unigene clusters	92,000	1996	Schuler, G. D*., et al*., 1996
Gene sequences	140,000	1999	*IncyteGenomics
Chr. 22 seq.	43,000-61,000	1999	Dunham, I., *et al*., 1999
Chrs 22, 21 seq.	44,000	2000	Hattori, M*., et al*., 2000
Tetraodon seq.	28,000-34,000	2000	Roest-Crollius, H., *et al*., 2000
ESTs in dbEST	120,000	2000	Liang, F., *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

Estimates for the number of genes encoded in the human genome have varied depending upon the type of information available to calculate the figure. An early rough estimate suggested a genome size of 3,000 Mb could contain 300,000 non-overlapping units of 10 kb, or 100,000 units of 30 kb (Gilbert, W., 1992). A later estimate of gene number by Antequera and Bird (1993) used the fact that the 5' end of approximately 56% of genes are associated with CpG islands, which are regions of non-methlyated DNA that contain the dinucleotide CG at the expected frequency. Analysis of CpG islands present in small amounts of sequence available at the time, suggested the human genome contains approximately 80,000 genes (Antequera, F*., et al*., 1993). Fields *et al* (1994) clustered the available EST data being generated as part of an EST sequencing project to suggest that there may be between 60,000 and 70,000 genes in the human genome (Fields, C*., et al.*, 1994). There have been two estimates, based on cDNA sequence (Incyte Genomics) and EST clusters in dbEST that the human gene number exceeds 100,000 (Liang, F*., et al.*, 2000). However, more recent estimates based on larger datasets or more refined analyses tend to converge on a lower figure, in the region of 30,000, as follows. Analysis of the finished sequence of human chromosome 22, which accounts for 1.1% of the human genome and was predicted to contain a minimum of 679 genes, would suggest there are between 43,000 and 61,000 genes (Dunham, I.*, et al.*, 1999). This figure may be artificially inflated by the incorporation of 134 pseudogenes, which assuming they are not expressed, are not likely to be represented in the other estimates. In combination with the analysis of human chromosome 21, the figure is quoted to be approximately 44,000, but this figure assumes chromosomes 21 and 22 represent an average gene density similar to that observed across the genome (Hattori, M.*, et al.*, 2000). This is slighty higher than a study carried out which looked at conserved sequences between human and tetraodon and suggested there may be as few as 28,000 genes present (Roest-Crollius, H.*, et al.*, 2000). The most recent prediction, based on the analysis of the human draft sequence suggests the human genome may contain 31,000 genes (IHGSC, 2001).

Prior to large-scale sequencing of the human genome, gene identification relied upon the identification of an mRNA representing a particular gene. The ability to generate complementary DNA (cDNA) from an mRNA molecule using reverse transcriptase enabled the cloning of the cDNA molecule for further investigation. For instance, the cloning and partial sequencing of the cDNAs representing the α−, β−, and γ−globin genes enabled characterisation of the structure of the globin gene cluster (Little, P. F.*, et al.*, 1979; Rabbitts, T. H., 1976). cDNA libraries containing clones representing the complement of mRNA molecules from individual tissues were also generated (Gubler, U.*, et al.*, 1983).

Individual or multiple genes of interest could be sequenced, primarily by using a poly(dT) primer to prime from the polyA tail present at the 3' end of the cDNA clones. The advances in the sequencing technologies including the anchored poly(dT) primer (Khan, A. S*., et al.*, 1991) and the improvements in the cDNA library preparation led to the development of strategies for large scale single-pass sequencing and mapping of human cDNAs. It was suggested the sequencing of cDNA clones would be a more efficient strategy for the identification of all human genes, rather than mapping and sequencing the entire human genome (Brenner, S., 1990). Projects involving the partial sequencing of cDNA clones, primarily representing the 3' end of genes, were generating large amounts of novel expressed sequences, such as that described by Adams, M.D., *et al*. (1991) who generated brain-specific expressed sequence tags (ESTs) representing more than 300 novel genes (Adams, M. D*., et al.*, 1991). In 1992 it was reported that global cDNA sequencing may have identified as many as 10,000 human genes (Khan, A. S*., et al.*, 1992). By 1996, there were approximately 450,000 ESTs residing in Genbank (Hillier, L. D*., et al.*, 1996). These ESTs were clustered into non-redundant sets (Unigene, Boguski, M. S., 1995; Schuler, G. D*., et al.*, 1996), Merck Gene Index (Aaronson, J. S*., et al.*, 1996), TIGR Human cDNA collection (Adams, M. D*., et al.*, 1995). Each cluster includes all sequence information, as well as any expression and mapping data that are available. There are currently 1.2 million ESTs in 84,000 clusters in the Unigene collection (December 2001). Sequence alignment programs are used to identify sequences matching at greater than 97%. This is likely to be an over estimate of the overall gene number as one gene could be represented by more than one EST cluster, particularly for the genes with large mRNAs which may be represented by an EST cluster at both the 5' and 3' end.

The EST sequencing projects which used a wide variety of cDNA libraries and carried out single pass sequencing from both the 5' end and 3' end of cDNA clones, proved extremely useful for gene identification but as a stand alone attempt to identify all the genes in the human genome had major shortcomings. cDNA sequence alone gives no indication of the structure and organisation of a gene such as the positioning and size of introns, nor does it provide sequence of the elements involved in regulation of gene expression. There is also evidence that the cDNA libraries used for EST sequencing may have contained contaminating genomic DNA, from which priming for sequencing could have occurred. Also, some mRNAs contained intronic sequences which were subsequently incorporated in to the EST data (Burglin, T. R.*, et al.*, 1992). Single pass sequencing from the 5' end and the 3' end of each cDNA clone may not generate the entire sequence coverage particularly for larger cDNA clones. The generation of the cDNA from the mRNA using reverse transcription was not always 100% efficient and so the cDNA libraries did not necessarily represent the full length mRNA molecule. Sequencing from the 5' end of a truncated cDNA clone can give a false indication of the true 5' end of a gene. If a gene is not expressed in any of the cDNA libraries used for EST sequencing, the gene will never be identified from these datasets.

Human genome sequence enables a whole new set of information that, although previously known, could not be applied, such as sequence motifs, homology searches and computational predictions and provides the basis for a more systematic approach to gene identification. The identification of genes in genomic sequence can be divided into four main categories: (1) direct evidence of transcription based on EST and cDNA sequence, (2) comparative protein analysis as proteins or parts of proteins look like other genes in both human and non-human sequence, (3) *ab initio* gene prediction, (4) comparing genome sequences of different organisms on the assumptions that regions that have conserved function since the divergence from a common ancestor will remain conserved at the sequence level.

(1) Direct evidence

The comparison of cDNA sequence with the corresponding genomic sequence is a more powerful method of identifying genes and their structures, than cDNA sequence alone which gives no information regarding gene structure and organisation. The availability of the genomic sequence and the ability to align the vast amounts of both full length and partial cDNA sequences enables not only the identification of the genes, but also their exon/intron structure. However, in order to identify all genes by this manner, a cDNA representing each gene would need to be sequenced. The importance of this strategy led to the establishment of a number of full length cDNA sequencing efforts to complement the genomic sequencing efforts, for example Mammalian Gene Project (MGP – http://mgc.nci.nih.gov/) and RIKEN (http://genome.gsc.riken.go.jp/home.html). Unlike single pass cDNA sequencing which provided partial information, these more recent cDNA sequencing projects are working to produce sequence covering the entire length of the cDNA clones.

Full-length cDNA sequencing is still limited by the quality and diversity of the cDNA libraries used. Improvements have been made in the technologies used to generate full-length cDNA clones, including the efficiency of the reverse transcriptase enzymes used and the procedure for selecting full length clones. For instance, the cap-trapper method has been developed based on the introduction of a biotin group into the diol residue of the cap structure of the mRNA, followed by RNase I treatment to select full-length cDNA clones (Carninci, P*., et al.*, 2001). Clones that are not full length will loose the biotin group and not be trapped using streptavidin-coated magnetic beads. However, the cDNA sequence may still not represent the entire length of the gene, identifying only part of the gene, and if the gene is not expressed, or expressed at a low level, in the cDNA libraries used, the gene will not be identified at all by this approach.

(2) Comparative protein analysis

Genes encode proteins which are made up of discrete domains that can be defined on the basis that they are members of recognisable families with specific structure and/or function, which work in combination to contribute to the overall functioning of the protein. For instance the prothrombin prescusor contains a gla, two kringle domains, and one trypsin domain (Bateman, A*., et al.*, 2002) although in some cases a protein may only contain a single domain, for example *SH2D1A* which contains a single SH2 domain (Coffey, A. J*., et al.*, 1998).

Genes can be clustered into families based on sequence similarity at the nucleotide and amino acid level, but also based on the similarity of the domains they contain. The important genetic event in generating a gene family can be either; (1) The divergence of a common ancestral organism to form two or more species with related genes or orthologues which are free to evolve separately but will remain similar at the sequence level if their functions remain similar, (2) The duplication of an ancestral gene to form paralogues within a species, and possible expansion to form gene clusters to produce related genes, so they can evolve independently for new biological functions.

Non-homologous recombination between repeated elements is one method by which gene duplication can occur. After gene duplication, alterations in the nucleotide sequence may lead to either an altered function of one gene or a silencing of one gene to form a pseudogene (Papadakis, M. N*., et al.*, 1999). An example of a gene family is the globin genes. The α-globin genes, of which there are three functional copies and two pseudogenes located in a cluster on chromosome 16, are highly similar at the nucleotide sequence level, and are suggested to have arisen from a single ancestral α-globin gene by tandem duplication (Papadakis, M.N. *et al*., 1999).

Gene and protein sequences can be analysed to identify novel members of existing families or novel families, on the basis of the protein domains the proteins contain. A variety of different protein family databases are available, for example PFAM (see http://www.sanger.ac.u/software/Pfam and Bateman, A., *et al*., 2002), where the available protein sequence data available in Swissprot, a database of well characterised protein sequences, and TrEMBL, a database of less well characterised translated nucoleotide sequences is analysed for protein domains. In the case of PFAM, sequences from each domain are aligned and a hidden markov model (HMM) is generated which can be used to search for other proteins containing the domain of interest. The proteins can then be characterised on the basis of similarity to these known protein domains. In PFAM, there are currently 3071 protein families which match 69% of proteins in Swissprot and TrEMBL (Bateman, A., *et al*., 2002).

The ability to compare all known nucleotide and protein sequences with genomic sequence enables novel members of previously characterised gene families to be identified. Genomic sequence can be analysed at the protein level by putative six-frame translation (Gish, W., *et al*., 1993). This approach will identify those portions of proteins which are similar to previously identified proteins (e.g. specific encoded domains as opposed to whole proteins), but may not identify the entire protein and will not identify the portion of the untranslated region of the gene which encodes the protein. Also, those proteins that are not members of known protein families, or are members of as yet unidentified protein families, will not be identified by comparative protein analysis.

(3) Ab initio gene prediction

The two approaches described above, cDNA sequencing and comparative protein analysis will not identify all the genes in the human genome, and those that are identified are not necessarily going to be complete. Therefore, methods were devised to predict regions of the genome likely to contain genes. These methods took advantage of gene-specific signals such as CpG islands, codon usage and splice sites. CpG islands are associated with approximately 56% of genes (Antequera, F., *et al*., 1993) and the availability of the genome sequence allows computer programs such as CpGfinder (written by Gos Micklen, The Sanger Institute) to predict the location of CpG islands and using this information infer the possible presence of the 5' end of gene. Using CpG island information alone does not provide any information about the structure of a gene and precludes those genes that are not associated with CpG islands but is a useful strategy that can be used in conjunction with other predictive methods.

Codon usage was first recognised as being useful for gene prediction by Staden, R., *et al*., in 1982. The triplet codon for each amino is degenerative, but there is a tendency that one particular codon is to code for each amino acid. This leaves a signature of protein coding regions in sequence when compared with non-coding areas. Other signals that are present are acceptor and donor splice sites (often AG and GC respectively – approximately 0.5% of splice sites (Thanaraj, T. A*., et al.*, 2001), translation start sites (commonly ATG), polyadenylation signals (such as AATAAA in 60% of genes) and stop codons (TGA, TAA, TAG).

The first computer prediction programs for gene identification predicted the presence of single exons using primarily codon usage statistics and characteristic sequence signals such as acceptor and donor splice sites e.g GRAIL (Uberbacher, E. C*., et al.*, 1991), and MZEF (Zhang, M. Q., 1997). More recently prediction programs that attempt to predict entire gene structures were developed. These programs (e.g GENSCAN (Burge, C*., et al.*, 1997)) and FGENESH (Solovyev, V. V*., et al.*, 1995)) differ in underlying algorithms used but have the same basic premise: prediction of individual exons based on codon usage and sequence signals, followed by assembly of these putative exons into candidate gene structures. *Ab initio* gene prediction methods are not capable of accurately predicting all the genes in the human genome without overprediction. If the thresholds for prediction were set very low they may capture all genes but with low specificity. In a recent study (Guigo, R., et al., 2000) that looked at the accuracy of gene prediction methods using an artificially generated data set, GENSCAN was shown to accurately predict 90% of coding nucleotides and 70% of the exons were predicted correctly. The study concluded that it is not yet possible to use computational methods alone to accurately identify the exonic structure of every gene in the human genome (Guigo, R.*,* *et al.*, 2000).

(4) Comparative Genome Analysis

Segments of DNA that have function are more likely to retain their sequence than non-functional segments, as they are under the restraints of natural selection during evolution. Therefore, DNA segments that are conserved between species are more likely to have function. The ideal species to compare with human are those whose form, physiology and behaviour are similar, but the non-functional sequence has diverged sufficiently, thus maximising the possibility to detect the differences between conserved and non-conserved sequences. In practice, there is no single species that provides all the answers for human annotation, as different genes and regulatory regions evolve at different rates. Comparisons with more closely related species will provide information such as gene structures and regulatory elements but these sequences may be obscured by non-functional conservation of sequence and therefore a variety of species, including more distantly related species may be required.

The predicted number of genes in the human genome currently stands at 31,000 (based on draft sequence analysis, see Table 1.3). This is only twice the number needed to make a worm, a fruitfly, or a plant. Sixty percent of predicted human proteins have sequence similarity with proteins predicted from finished genome sequence of other organisms. Sixty-one percent of all fruitfly proteins, 43% of those from worm and 46% of yeast proteins show sequence similarity to predicted human proteins (Rubin, G. M.*,* *et al.*, 2000). Those proteins that appear specific to a particular species may have either a similar function to proteins from other organisms but the sequence has diverged, or novel species-specific function. More than 90% of the domains in human proteins are present in the worm and the fruitfly. Therefore vertebrate evolution has required the invention of very few domains.

Comparative genome sequence analysis is very informative, and so very accurate tools are required to carry it out. Two main strategies for comparing two sequences are available, local alignment and global alignment. One comparison tool that is capable of comparing sequences is DOTTER and utilises a local alignment strategy (Sonnhammer, E. L*., et al.*, 1995). The concept uses graphical matrix plots where one sequence is drawn on the horizontal axis and the other along the vertical axis of a coordinate system and a dot is drawn where two residues match. Regions of similarity that are co-linear between sequences will result in a diagonal row of dots. Spurious matches give rise to a background of dots. A second example of a commonly-used local alignment tool is BLAST (Basic Local Aligment Search Tool) which measures the local similarity between two sequences (Altschul, S. F*., et al.*, 1990), and has become widely used for searching protein and DNA databases for sequence similarities. An adaptation of BLAST, gapped BLAST was developed that is three times faster than BLAST and is more sensitive for comparing cross-species sequences as it allows for gaps in the alignment and can identify weak similarities (Altschul, S. F*., et al.*, 1997). Where BLAST may have found three similarities between two sequences that individually may not be significant, gapped-BLAST is able to link them together by introducing gaps in the alignment to produce a significant alignment. PIPMAKER is an alignment program that is capable of comparing long sequences (Schwartz, S*., et al.*, 2000). The alignments are generated using BLASTZ, a derivative of gapped-BLAST, and viewed on a percentage identity plot (pip). An example of a commonly-used global alignment tool to compare two sequences is VISTA (VISualisation Tool for Alignments) which uses GLASS (Global Alignment SyStem) to generate the alignment (Dubchak, I*., et al.*, 2000). Initially a rough alignment map finds the long segments that match exactly and flanking regions that have high similarity. The process is repeated on the intervening regions using successively smaller matching segments. The remaining regions are aligned using standard alignment techniques. Conserved regions are identified by calculating the percentage of identical nucleotides within a window (for example 100 nucleotides) moved in smaller nucleotide increments (for example 25 nucleotides).

Initial strategies using comparative analysis focused on using human and mouse sequence to identify non-coding regions (e.g. (Hardison, R*., et al.*, 1993; Koop, B. F*., et al.*, 1994; Duret, L*., et al.*, 1997; Hardison, R*., et al.*, 1997), but once more substantial data sets were available it became clear that comparing two sequences as closely related as human and mouse was also valuable for protein coding regions (Makalowski, W*., et al.*, 1996; Ansari-Lari, M. A*., et al.*, 1998; Jang, W*., et al.*, 1999). Realistic analysis of the effectiveness of alternative gene prediction methods suggested the need for improved prediction accuracy (Dunham, I*., et al.*, 1999; Guigo, R*., et al.*, 2000). In order to complement human gene identification, the mouse sequencing efforts have been accelerated which has necessitated the need for novel comparison tools (Miller, W., 2001).

Proteins that control gene expression bind to DNA but these binding sites are difficult to predict computationally. Analysing genes that have similar patterns of expression may aid the process of identification as these genes may be expected to share regulatory elements. Aligning sequences upstream of these genes may identify conservation between similarly expressed genes. DNA containing the stem cell leukaemia (SCL) gene from human, mouse and chicken was sequenced and compared. Regions of conservation between the three species were identified that corresponded with exons of the SCL gene. However other regions of conservation were also shown to coincide with regions of known SCL enhancers. One particular region, (+23 region) was shown using a transgenic xenopus reporter assay demonstrated the region contained a novel neural enhancer (Gottgens, B*., et al.*, 2000).

The identification of control regions using comparative genome sequence analysis has been shown (Brickner, A. G*., et al.*, 1999; Gottgens, B*., et al.*, 2000). Comparative genome sequence analysis alone does not identify control regions, but it does identify conserved regions that are candidates for further experimental investigation.

In order to identify all the features encoded in the human genome a systematic analysis of the sequence is required. In order to identify all genes a combination of large-scale computational analysis, manual interpretation and experimental investigation will be required. Already projects are underway to identify and catalogue as many human genes as possible, such as ENSEMBL (http://www.ensembl.org), UCSC (http://genome.cse.ucsc.edu/cgi-bin/hgGateway) and NCBI (http://www.ncbi.nlm.nih.gov/genome/guide/human). As the DNA sequence of individual human chromosomes is finished, systematic analysis of the genes and other features provides insights into the features encoded within.

1.4 The human X chromosome

The human X chromosome is estimated to be 164 Mb in size and accounts for one-twentieth of the human genome. Females have two X chromosomes and males have one and the highly degenerate Y chromosome which carries the testis-determining factor. The hemizygous state of the male reveals the phenotypic effect of recessive mutations directly and this, along with X-linked dominant disorders, accounts for large numbers of X-linked genetic diseases, which have a characteristic pattern of inheritance (McKusick, V. A., 1998). Mutation analysis for X-linked disorders in males is easier than for autosomal diseases as they only carry the affected chromosome and this can be analysed directly, whereas a mutation on an autosome is masked by the affected sequence if the individual is heterozygous.

An unique feature of the human X chromosome, shared by mammalian homologues including marsupials, is the inactivation of one of the two chromosomes in females. X chromosome inactivation (XCI) is a mechanism for dosage compensation and was first hypothesied by Mary Lyon in 1961 (Lyon, M. F., 1961). Either the paternally- or the maternally- derived X chromosome is inactivated at random but the same X chromosome is inactivated in future generations of each cell. In the extra-embryonic tissues of mouse and marsupials it is always the paternally-derived X chromosome that is inactivated. The inactive X chromosome replicates late during S phase of meiosis (Takagi, N., 1974) and is associated with hypoacetylation of the histone proteins H2A, H3 and H4 (Jeppesen, P.*, et al.*, 1993; Belyaev, N.*, et al.*, 1996). In interphase FISH, the inactive X chromosome appears as a condensed mass or Barr body (Barr, M. L.*, et al.*, 1949; Barr, M. L.*, et al.*, 1961). In its condensed state, the inactive X chromosome is highly stable and is only reversed in female germ cells at meiosis (Chapman, V. M., 1986).

Initiation of XCI occurs in the early embryo and originates in the X-inactivation centre (XIC) in Xq13, and then propagates along the length of the chromosome (Rastan, S., 1994). The *Xist* (X inactive specific transcript) gene is located within the XIC and is transcribed on the inactive X chromosome only (Borsani, G.*, et al.*, 1991; Brockdorff, N.*, et al.*, 1991). The *Xist* gene produces a 15-17 kb non-coding mRNA that coats the inactive chromosome (Willard, H. F., 1996; Brockdorff, N., 1998; Panning, B.*, et al.*, 1998). The 5' end of *Xist* is essential for initiation of X inactivation, and the 3' end of is essential for chromosome counting – i.e. ensuring one chromosome remains active (Brockdorff, N., 1998, Clerc, P.*, et al.*, 1998).

The mechanism for X chromosome inactivation is still largely unknown. Coating of the inactive X chromosome begins at the XIC and spreads to the whole chromosome. Segments of X chromosome without an XIC, through deletion or translocation, remain active. Spreading of inactivation from the X chromosome to regions of translocated autosomal chromosomes occurs but much less efficiently than on the X chromosome portions (Rastan, S., 1983). This observation that all DNA is susceptible to the initial coating of XIST RNA but there is something unique about the X chromosome that promotes the coating led to the hypothesis that interspersed repeat elements, particularly L1s may play a role (Lyon, M. F., 1998). Recent analysis of the human genome sequence showed that the X chromosome is in fact richer in L1s than any other chromosome, 26% as compared to 13% (Bailey, J. A.*, et al.*, 2000).

X inactivation is thought to have arisen early in the evolution of mammals. Monotremes show late replication of part of the X chromosome, which may be a rudimentary form of X inactivation. The transfer of genes from autosomes to the X chromosome and from the X chromosome to autosomes is thought to be limited because of dosage compensation. Therefore genes that are X-linked in one mammal are likely to be X-linked in others (Ohno '67). There are exceptions to this (Graves, J. A., 1996) but it is assumed that these did not disturb the mechanism of dosage compensation. To date, the majority of X-linked genes in humans are shown to be present on the X chromosome in mice. The observation that genes on the short arm of the X chromosome in humans are present on an autosome in monotremes and marsupials suggests that the short arm was of autosomal origin and was added to the X chromosome in eutherian mammals.

There are two blocks on the X chromosome that escape X chromosome inactivation, termed the pseudo-autosomal regions (PARs). These pair with the Y chromosome during meiosis and a varied recombination frequency is observed between males and females. A greater degree of recombination is seen in males due to the obligatory exchange of material within the 2.5 Mb PARs during male meiosis. This higher recombination rate is also seen in other regions of the X chromosome (DMD; Abbs, S*., et al.*, 1990), FRAXA; Richards, R. I*., et al.*, 1991).

There has always been great interest in the X chromosome because of the high proportion of X-linked disorders. It was the first chromosome to have a genetic map, based on RFLPs (Aldridge, J*., et al.*, 1984; Drayna, D*., et al.*, 1985). The most recent genetic map contains 216 polymorphic markers and is estimated to be 216 cM in size (Dib, C*., et al.*, 1996). The published gene map (Deloukas, P*., et al.*, 1998) (updated electronically in 1999 – http://www.ncbi.nlm.nih.gov/GeneMap99) indicated the X chromosome may be relatively gene poor, containing only 50% of the number of ESTs as those expected based on its size and assuming a random distribution. It is not known whether the relative paucity of genes on the X chromosome is due to its role in sex determination or the mechanism arose independently. The intense interest in the X chromosome led to rapid progression of the YAC map and successive X chromosome workshops produced consensus landmark maps on regional efforts. This culminated at the most recent workshop in the generation of YAC maps covering virtually the entire X chromosome (7th X chromosome Workshop (7XCW) – The Sanger Centre 1995; Nagaraja, R*., et al.*, 1997).

Also at the 7XCW, responsibility for bacterial clone mapping and sequencing of the X chromosome was divided up between centres. This underlies the international collaborative effort involved in sequencing the human X chromosome. Currently there are 26 bacterial clone contigs covering the X chromosome and 125 Mb of finished sequence and 65 Mb of draft sequence is available (see http://www.sanger.ac.uk/ChrX). Efforts are continuing to close the remaining gaps and finish the sequence.

1.4.1 Xq22

The region in Xq22 under study in chapter 3, between DXS366 and DXS1230 encompasses approximately 7 Mb on the long arm of the X chromosome. This region has been shown to contain a number of genes involved in genetic disorders, some of which have already been isolated. These include the genes involved in Fabry disease (Bernstein, H. S*., et al.*, 1989), Pelizaeus Merzbacher disease (Hudson, L. D*., et al.*, 1989; Trofatter, J. A*., et al.*, 1989), and X-linked agammaglobulinaemia (Tsukada, S*., et al.*, 1993; Vetrie, D., 1993). Other disease loci mapping to the region, for which candidate genes have not yet been identified include X-linked megalocornea (Chen, J. D*., et al.*, 1989) and X-linked deafness 2 (DFN2; Tyson, J*., et al.*, 1996). The majority of the region between DXS366 and DXS1230 is thought to lie within a light band hence is expected to contain many genes (discussed earlier, see Section 1.2). At the start of this work, physical mapping had concentrated at the resolution of the YAC map (Vetrie, D*., et al.*, 1994; Kendall, E*., et al.*, 1997). The generation of sequence-ready bacterial clone contigs covering the region, described in chapter 3, will provide the first step towards the complete genomic sequence, and a fully annotated version containing all the genes and other biologically relevant information.

1.4.2. Xq23-24

The region in Xq23-24 under study in chapter 4, between DXS7598 and DXS7333 encompasses approximately 8 Mb of the long arm of the X chromosome. This region has been shown to contain a number of genes including ANT2 (Nagaraja, R., *et al.*, 1998, Schiebel, K., *et al.*, 1994, Steingruber, H. E., *et al.*, 1999) and LAMP2 (Manoni, M., *et al.*, 1991, Nagaraja, R., *et al.*, 1998, Steingruber, H. E., *et al.*, 1999). Xq23 is a dark staining R-band and is expected to contain few genes whereas Xq24 is a light staining G-band and is thought to be gene rich. Initial evidence that Xq24 is gene rich was shown by a cluster of CpG islands mapping to the region (Maestrini, E., *et al.*, 1990). The production of sequence-ready maps and the generation of genomic sequence will enable a systematic approach for gene identification to be undertaken.

1.4.3 Non-specific X-linked mental retardation

One of the diseases whose critical region is contained within Xq23-24, MRX23 (Gregg, R. G., *et al.*, 1996), is one of many non-specific mental retardation (NSMR) disorders mapping to the human X chromosome. NSMR includes all those disorders whose only consistent clinical manifestation is mental retardation and includes X-linked mental retardation (XLMR) (Neri, G., *et al*., 1999). It has been known for a long time that there is an excess (25-30%) of males among the mental retardation patients, particularly with a mild to moderate phenotype (Lehrke, R. G., 1974).

XLMR represents approximately 5% of all mental retardation and corresponds to a prevalence of 1 in 600 males in the general population (Crow, Y. J*., et al.*, 1998). Regional assignment along the X chromosome of different families with XLMR has shown that at least fifty MRX families exist (Toniolo, D*., et al.*, 2000; Neri, G*., et al.*, 1999). By convention, each family represents a locus and is designated by the acronym MRX and by a progressive number. A database listing XLMR disorders has been developed where the XLMR have been divided into two groups, the syndromic and the non-specific XLMR (Cabezas, D. A*., et al.*, 1999).

Eight genes for NSMR have been identified to date and form a heterogenous group encoding diverse proteins ranging from transmembrane proteins to transcription factors (Toniolo, D*., et al.*, 2000). However, all the genes identified so far are either directly or indirectly involved in signalling pathways. Further study of these genes and the identification of more NSMR genes are required before a full understanding of NSMR and the development of cognitive function are achieved.

1.5 Aims of this thesis

At the time this thesis began, the majority of physical maps covering large portions of the human genome were in the form of YAC maps. However, the plans to generate human genome sequence necessitated the construction of physical maps in bacterial clones, a more suitable substrate for sequencing. The aim of the first part of this thesis was to generate sequence-ready bacterial clone contigs across a 7 Mb portion of the long arm of the X chromosome in Xq22.

The generation of the sequence of the human genome is only the first step in its complete characterisation. An important subsequent step is the identification of all the genes and other biologically relevant information encoded within. The second aim of this thesis was to construct a transcript map in Xq23-24, identifying and experimentally confirming as many genes as possible using the resources available at the time. Gene identification using sequence similarity searches in combination with *ab initio* gene prediction to predict genes that are then confirmed by cDNA isolation and sequence has two major limitations: not all genes will be identified by this method and not all of those predicted genes will be confirmed experimentally.

Comparative genome analysis is playing an important role in the elucidation of all the features within the human genome and in the understanding of their function. Generating DNA sequence from syntenic portions in other species allows functionally conserved units to be identified at the sequence level. This analysis provides additional data that can be used to support previously identified genes as well as identifying potential novel functional units. The final aim of this thesis was to construct bacterial clone contigs in mouse and zebrafish, sytnenic to a region in human Xq24. The sequence from all three species will allow further analysis of the features previously identified in the human sequence and identify potential novel functional units.