Chapter 7

Discussion

7.1 Genome Mapping and Sequencing

7.2 The determination of coding features

7.3 Assigning gene function

7.4 Sequence variation

7.5 Conclusion and future work

7.1 Genome Mapping and Sequencing

The generation of genome-wide physical maps of biologically complex organisms with large (i.e. multi-gigabase) genomes has relied upon the application of strategies developed for the characterisation of smaller (~1-100 megabase) genomes. At the time of their construction, the cosmid and lambda clone physical maps of *Caenorhabditis elegans* (*C. elegans*) (Coulson *et al*., 1986) and *Saccharomyces cerevisiae* (*S. cerevisiae*) (Olson *et al*., 1986), respectively, were intended to facilitate the cloning of known genes and to serve as a genomic archive. It was during the construction of these physical maps that concomitant improvements in sequencing technologies enabled a whole genome sequencing strategy to be developed. Clones from the physical map that represented the genomic archive were used as the substrate for subcloning and shotgun sequencing. As size of the clones used to construct these maps was equivalent to large viral genomes, which had already been successfully sequenced using the random shotgun approach (e.g. 48 kb bacteriophage λ (Sanger *et al.,* 1982)), it was clear that the equivalent strategy could be applied to sequencing whole genomes. The limiting factor would be the availability of a complete clone map and sufficient resources to complete the sequencing. For example, results from the *C. elegans* pilot sequencing project were published (Sulston, *et al*., 1992) some 6 years after the physical map was reported to be 60% complete (Coulson *et al*., 1986). The publication reported the assembly of 120 kb of completed sequence by separate shotgun sequencing of 3 selected cosmid clones (two of which overlapped) from the physical map. This report, therefore, confirmed the feasibility of this approach and the value of the map in characterising large genomes.

At this point in time, clone coverage of the human genome was limited to small YAC or bacterial clone contigs which had been constructed during positional cloning projects. The use of YACs in attempts to rapidly construct chromosome specific (Chumakov *et al*., 1992, Foote *et al*., 1992) and genome wide clone maps (Chumakov *et al*., 1995, Hudson *et al*., 1995) demonstrated that larger insert clones could be used to generate a map of the whole genome. Indeed, the adoption of YACs, which have an average insert size of 1 Mb (Chumakov *et al*., 1995) theoretically reduced the complexity of mapping the human genome (1x coverage of 3 GB requiring approximately 3,000 clones) This approximated the complexity of the earlier work where the nematode genome was mapped using 40 kb cosmids (1x coverage of 100 Mb requiring approximately 2,500 clones). However, the instability and chimerism of YACs, in addition to the difficulties of isolating the cloned insert compared to bacterial cloning systems, meant that YACs could not practically provide a sufficiently reliable or convenient resource for the generation of genomic sequence. The development of large insert PACs (Iaonnou *et al.*, 1994) and BACs (Shizuya *et al*., 1992) provided clones with an insert large enough to permit the feasible generation of genomic coverage (1x coverage of the human genome requiring ~25,000 clones) (as described in chapter 4) and a cloning system that was amenable to large-scale sequencing (as described in chapter 5). The development of marker-based maps using genetic, RH or YAC mapping techniques (genetic: Weissenbach *et al*., 1992, Dib *et al*., 1996; RH: Schuler *et al*., 1996; Deloukas *et al*., 1998; YAC: Hudson *et al*., 1995; genome; Foote *et al*., 1992; Collins *et al*., 1995) provided the necessary long-range and independent map information to act as a framework on which bacterial clone coverage could be generated (Olson *et al*., 1993, Bentley *et al*., 2001). In addition, the evolution of fingerprinting techniques (Coulson *et al*., 1986, Olson *et al*., 1986), as described in chapter 3 (Gregory *et al*., 1997, Marra *et* *al*., 1997), provided the tools by which sequence ready bacterial clone maps of large genomes could be constructed.

The chromosome specific, clone by clone approach that the Human Genome Project utilised to generate sequence of the human genome provided a means by which the work of many separate groups could be coordinated and a method by which highly accurate genomic sequence (>99.99%) could be produced. The principles demonstrated in the nematode project (Coulson *et al*., 1986) proved effective, and essential for the successful production of the human genome sequence. Problem solving (in particular resolution of the sequence of repeats, or technically difficult regions) was facilitated by the modular clone-based nature of the project; while long-range map information added weight to the clone order.

Celera, a private company that published a draft version of the human genome (Venter *et al*., 2001) in parallel to the publicly funded project (IHGSC 2001), purported to have assembled the human genome exclusively by a whole genome shotgun (WGS) strategy. However, they too ultimately relied upon a clone based map by incorporating a "perfect tiling path" of the HGP data into their sequence assembly. Specifically, 10 million ‘faux’ reads in a regular ordered pattern were generated by dividing up the public assembly into overlapping segments of 550 bp each, thus capturing the overall synthesis of all the information used to assemble the public domain sequence. The public sequence data used by Celera was thus derived from an accurately assembled clone based physical map and effectively representing complete genomic coverage (Waterston *et al*., 2002). Projects to sequence large genomes using a WGS strategy have also proved impossible to finish with high accuracy. This was illustrated by the reliance upon clone based finishing of the WGS assembly for the *Drosophila* genome (Celniker *et al*., 2002). Similarly, both the human and the mouse genome sequences are being finished using a large-insert bacterial clone based approach.

The next phase in the evolution of physical map construction was driven by the availability of ordered genomic sequence. Conservation of sequence and long range order, between organisms that are sufficiently closely related, means that the genome of one species can act as the template upon which a physical map of another can built and, in doing so, elucidate the homologous relationship between them (Thomas *et al*., 2002). The success of the comparative physical mapping approach was demonstrated by construction of a clone map of the mouse genome using the assembled human genome sequence as a template. In this study, human genomic sequence was used to align stringently assembled BAC fingerprint contigs by matching mouse BAC end sequences (BESs) to their corresponding locations in the human genome. Ordered and orientated contigs (previously assembled by fingerprinting) were subsequently joined following further fingerprint analysis and addition of available genetic and radiation hybrid markers. The availability of BESs from a highly redundant fingerprint assembly of BAC clones, and using the strategy outlined above, greatly simplified the process of contig assembly, as the majority of the 7,500 contigs generated in the first fingerprinting phase were juxtaposed correctly relative to each other on the basis of homology between the two genomes. As a result, 7,500 x 7,500 possible joins (more than 56 million) was reduced to analysis of <10,000 putative joins. This permitted the construction of a physical map covering 98% of the 2500 Mb mouse genome, contained within 296 contigs, in approximately 12 months (see accompanying paper). The same approach could be adopted for any genome where there is sufficient sequence homology to allow alignment of BESs (or equivalent sequence tags), plus sufficient homology between the template genome and the genome under study. The approach has important applications both for genomes where the full genome sequence is anticipated, and also (perhaps even more importantly) is a cost-effective way to provide access to regions of a genome for which there are no plans to generate genomic sequence on any scale.

Whilst a clone by clone approach proved successful for the generation of human sequence, the possible contribution of WGS data to large projects has continued to be evaluated. The main advantages of WGS are that the production of data is very rapid, can be highly automated, avoids cloning biases of BAC systems, and is very cost-effective. The assembly inherent from the sequence alignment also provides important mapping information which is unbiased by additional experimental mapping systems or procedures. While it remains true that WGS in isolation has disadvantages which prevent completion of either the map or finished sequence of a large genome, the possibility of combining the advantages of both approaches has been explored. A hybrid strategy emerged from the *Drosophila* project, and has since been adopted for the mouse genome. Seven fold WGS coverage was generated from sub-cloned plasmids of varying sizes which, when assembled with BESs, generated 96% coverage of the euchromatic portion of the mouse genome. This estimate was derived by assessing the amount of WGS coverage provided which matched 187 Mb of finished mouse sequence. For a second, independent estimate, a genomic alignment of a curated collection of cDNAs to the WGS assembly was also used. This alignment included 96.4% of cDNA bases. Paired-end reads from large insert plasmids and BACs provided the scaffold upon which the assembled whole genome shotgun sequence was ordered and orientated, and simultaneously integrated BAC clones into the sequence. A tiling path of BAC clones from the physical map is currently being used for directed finishing of the draft genomic sequence. The physical map helped to assemble the sequence scaffold, whilst the WGS data increased the rate of clone based finishing (MGSC 2002).

If WGS sequence data can accurately place BACs via their BESs within the sequence assembly, is a restriction fingerprint database actually required? The answer is probably yes. Whilst BES localisation within a whole genome shotgun assembly facilitates a more optimal minimum tiling path selection, overlaps within fingerprinting contigs can link sequence assemblies (as reported by the assembly of the mouse WGS sequence (MGSC 2002)). Three hundred and seventy-seven anchored ‘supercontigs’ were generated by assembling plasmid and BAC end sequences in the WGS assembly. This number was reduced to 88 when two or more sequence supercontigs were localised within a single restriction fingerprint contig. The overlaps generated by fingerprint analysis may also be able to resolve errors in the genomic assembly where, for example, low copy repeats may have resulted in a compression of the sequence assembly. The proven success of assembling genome wide physical maps, the cost of constructing a >15 fold genomic BAC library and the ease with which genome-wide fingerprint databases can be assembled, has lead to the construction of several genomic fingerprint databases, table 7.1. Whilst genome-wide fingerprint maps will facilitate the large-scale characterisation of many varied species, the construction of small region specific sequence-ready maps will continue to be important for detailed inter-species sequence comparisons (Thomas *et al*., 2002).

Table 7.1: Organisms for which genome-wide fingerprint databases have or are being constructed.

Organism	**Reference**
A. thaliana	Marra et al 1999
Rice	Tao et al. 2000
H. sapiens	McPherson et al 2001
M. musculus	Gregory et al 2002
R. rattus	http://www.bcgsc.ca/lab/mapping
C. neoformans	http://www.bcgsc.ca/lab/mapping
Bovine	http://www.bcgsc.ca/lab/mapping
Porcine	http://www.nps.ars.usda.gov/
D. rario	http://www.sanger.ac.uk/Projects/D_rerio/
Soybean	http://hbz.tamu.edu/soybean.html

7.2 The identification of coding features

The availability of human genomic sequence provides the framework upon which the structure of genes and their regulatory elements can be placed. The existence of genes as protein-coding genes, pseudogenes (Harrison *et al*., 2002), non-protein-coding RNA transcripts (Mattick *et al*., 2001) and genes arising from genomic duplications (Bailey *et al*., 2002) requires a multifaceted approach to gene discovery. The three main methods used for large-scale genome analysis, *in silico* prediction of gene structures, sequence alignment of expressed and genomic sequences and identification of conserved sequences between different species, are described in chapter 5. These approaches are preferentially used in combination to assist correction for the shortcomings of each method. For example, *in silico* gene identification can lead to over prediction and false negatives (Guigo *et al*., 2000). EST and cDNA sequence alignment can be confounded by artefactual or unprocessed clones in cDNA libraries. Cross species sequence comparison, for example between human and mouse, can also identify homologies outside genes and regulatory elements (Deloukas *et al*., 2001, Kondrashov and Shabalina *et al*., 2002).

An important facet of fully identifying all coding features within any organism is the availability of high quality finished genomic sequence. Annotation of genes using draft sequence, which may contain regions of low quality, incomplete or unordered data, may lead to inaccurate annotation of genes and possibly errors in inference of the protein products derived from them. Another important consideration is reanalysis of existing annotation. The continual emergence of new sequence data from independent studies, and reanalysis of genomic sequence on a regular basis (or on demand), is required to ensure incorporation of all available evidence for genes and other features. The genomic alignment of larger sets of non-redundant ESTs and cDNAs, derived from a wider range of tissues or of sequence from related organisms, may assist to fully define gene structures, identify novel coding features or regulatory regions. Recent publications detailing the re-annotation of human chromosome 22 (Collins *et al*., 2003) and *Drosophila* genomes (Misra *et al*., 2002), which list structural changes to genes previously identified, indicate that the characterisation of all coding features will require several iterations of automatic analysis, manual annotation and directed laboratory work. Other examples of comparative analysis have also revealed new regulatory elements or genes (Pennacchio *et al*., 2001, Gottgens *et al*., 2002). The identification of genic features such as the 5' and 3' ends of genes, splice variants and even the functional determination of non-coding genes such as anti-sense RNAs (Green *et al*., 1986), will take many years of painstaking study.

7.3 Assigning gene function

Characterisation of the functional product of a gene is not achieved directly by the identification of translated amino acid sequence contained within the open reading frame of the coding sequence. Within a genomic context, the final sequence and structure of an mRNA and the encoded protein may be influenced by priming from multiple promoters, splice variation within the coding exons or the existence of alternative polyadenylation sites, as discussed in chapter 5. Post-translational processing can also result in modification of a protein product. Whilst *in vivo* and *in vitro* studies within model organisms, by chemical mutagenesis and gene targeting can identify gene function by generating an observable phenotype it may not always be clear how a disruption of the target gene has given rise to a particular effect within a complex network of gene interactions. Alternatively, protein function may be predicted by *in silico* structural analysis. The assignment of new function to a novel protein at a nucleotide sequence level comparison, however, may fail as BLAST analysis within the Protein Data Bank (PDB) was shown to only find 10% of the known relationships (Brenner *et al*., 1998). Whilst iterative PSI-BLAST (Altschul *et al*., 1997) is more sensitive, relationships are still missed.

An alternative approach is the sequence-to-function method which uses pair-wise sequence or motif alignment to derive significant homologies between proteins and hence suggest similarity of function. Whilst these methods are powerful, they are not ideally suited to identify loss or gain of function during protein evolution and encounter difficulties when assigning function as protein databases become more diverse (Skolnick and Fetrow 2000). Alternatively, the possible function of a protein may be suggested by comparison of three-dimensional structure to proteins of known function. Since the tertiary structure of proteins of common function is likely to be more conserved than their primary structures (amino acid sequences), attempts have been made to classify groups of proteins based on structural and phylogenetic relationships, e.g. SCOP (Murzin *et al*., 1995). A second approach describes proteins according to their structural characteristics, such as class of architecture and fold type. In practice both approaches are used, initially grouping proteins according to their sequence homology and then by their structural descriptors (Thornton *et al*., 1999).

The application of the sequence to structure to function approach aims to determine the structure of a protein and then to identify the functionally important residues, as described in chapter 6. *Ab initio* folding can be used to predict a native structure based on domains contained within the protein. A process known as threading utilises a known structure as a template upon which proteins of up to 500 residues can be moulded. These three dimensional structures can then be used to infer function by analysis of internal or external residues, the shape and molecular composition of the protein or the juxtaposition of individual groups. The prediction of protein folds, their 3D structure and function is, however, primarily reliant upon experimental evidence, either as a basis for modelling or as support for a prediction. X-ray crystallography and nuclear magnetic resonance spectroscopy are methods by which these proteins structures have been experimentally determined.

The identification of well conserved, functionally important residues within primary and tertiary structures of some protein families can assist to predict the function of novel proteins. Based upon the conservation of an Asp-Thr(Ser)-Gly amino acid sequence at their active sites Pearl and Taylor (1987a, 1987b) concluded that retroviral proteases could belong to aspartic protease family. Modelling and subsequent structural comparisons between aspartic proteases, which contain 300 residues and an active site in each of its two domains, and retroviral proteases, that do not exceed 130 residues and have only one active site, led Pearl and Taylor to hypothesise that the retroviral enzymes may be dimeric aspartic proteases. This prediction was subsequently proven by the expression of the retroviral proteases in *E. coli* (Meek *et al*., 1989) and by the determination of it crystal structure (Navia *et al*., 1989)

In many cases, however, even in these well characterised families, the catalytic component may be recognisable, but the specific substrate binding properties may be difficult to determine. Additional protein domains (encoded by separate exons) which are required for function, but localise to other regions of the protein, are less readily elucidated by homology alone. For this, direct experimental approaches are required to determine the substrate and products in the appropriate biochemical pathway. For example protein binding assays, using yeast two hybrid systems, can identify interacting binding proteins. Knockouts, or natural mutants, may be investigated to determine the biochemistry of the altered phenotype in some detail. For example, a defective enzyme may result in accumulation of abnormally high levels of substrate, and comparison of normal vs. mutant systems will reveal candidates as possible substrates.

Additional information can be gained by determining the cellular and tissue localisation of the protein. The co-localisation of proteins in a highly tissue-specific pattern may provide evidence for some level of protein interaction. The fusion of the sequence encoding a novel protein to the sequence of a reporter molecule in a shuttle vector can be used to determine cellular localisation if the construct can be introduced into a physiologically relevant cell line. This work may be followed up, for example. by manipulation of the construct and introduction into embryonic stem cells in order to create a transgenic animal model where the gene is under control of the endogenous promoter. This would enable investigation of the expression of the gene presumably in response to physiologically natural intracellular and extracellular signals. The cellular distribution of the signal molecule should, therefore, reflect the distribution of the natural gene product. Data from co-localisation experiments may be correlated with protein-protein interaction studies, and possibly analysis (e.g. by mass spectrometry) of the components of co-purified complexes, to build a picture of the interactions between specific proteins

7.4 Sequence variation

The key to understanding the effects of sequence variation within protein-coding genes is the identification of genuine sequence variants from accurately annotated gene structures and determination of the functional effect that a variant has upon the encoded protein, or on expression of the gene. Chapter 6 describes the generation of exon specific sequences from a collection of twelve medically important genes, including five closely related members of a gene family. The design of primers which are uniquely localised within the genome, in conjunction with detailed analysis of the sequence flanking the SNP, permitted identification of novel locus specific polymorphisms within highly homologous genes and the unique placement of SNPs which previously had been given multiple localisations within Ensembl and dbSNP. The functional effect a minor coding SNP may have upon protein structure was predicted using protein modelling, as outlined in chapter 5. Whilst *in silico* prediction would always require experimental support (see also discussion above), it can be used to suggest the effects a SNP may have upon a protein structure and function. A novel non-synonymous, non-conservative coding SNP was identified within GSTT1, as part of this study.

The majority of genetic variation which contributes to cancer are somatic mutations caused by exposure to environmental carcinogens rather than inherited variants in susceptibility genes. However, genes that encode enzymes involved in the metabolism of carcinogens can be polymorphic and these polymorphisms may be related to elevated risk of cancer susceptibility. This mechanism illustrates the importance of genetic background and the effect on interaction between the individual and the environment. GSTT1 is a phase II enzyme that mediates the detoxification of xenobiotic chemicals by conjugation of glutathione which changes the polarity of the chemical and makes them more readily excreted. GSTT1 is a five exon gene which localises to 22q11.2 and is known to be absent from 38% of the population (Pemble *et al*., 1994). The null genotype of GSTT1 has been implicated with increased risk of myelodysplastic syndromes (MDS) (Chen *et al*., 1996), aplastic anemia (Lee *et al*., 2001) and, in conjunction with a CYP1A1 mutation, has effects on live birth weight (Wang *et al*., 2002). These studies indicate that the gene, by its absence, may influence the aetiology of disease. The identification of a novel non-synonymous non-conservative SNP, which may induce a change to conformation and function of the protein, as described in chapter 6, may also influence an individual's susceptibility to cancer.

7.5 Conclusions and future work

This thesis describes the application of a novel restriction fingerprinting technique to the generation of a sequence ready map of 1pcen – 1p13, the elucidation of the genic features within the interval and the characterisation of sequence variation within a selected number of these genes. In the short term, all these areas of research could be extended. The annotation of genes within the transcript map of the interval will require further work, as shown by the gain in information by the iterative annotation of *Drosophila* (Misra *et al*., 2002) and human chromosome 22 (Collins *et al*., 2003). Alignment of novel cDNAs and ESTs to genomic sequence may assist to further define gene structures, specifically the 5' and 3' ends, as well as identifying new genes and splice variants. Comparative sequence analyses, using species from a variety of evolutionary distances, will also help to better characterise the genes in the interval and also to identify regions that control their regulation.

Comparative sequence analysis could also be used to characterise the structural evolution of chromosome 1. Regions either side of the chromosome 1 centromere are contained within a contiguous homologous region on mouse chromosome 3. The investigation of gene order, number and orientation may assist to elucidate the number of chromosomal rearrangements that have taken place during the evolution of human chromosome 1 from the organisation of the ancestral karyotype.

The functional consequence of the novel non-synonymous, non-conservative SNP within exon 3 requires further investigation. The frequency of the non-synonymous, non-conservative GSTT1*B allele within other ethic groups (besides the Northern

European population used here) could be investigated. In addition, other phenotype to genotype correlations of the Thr104 – Pro SNP within large prospective DNA collections, such as the Avon Longitudinal Study of Parents and Children, could be determined.

In the longer term, it is anticipated that substantial efforts will result in all of the genes in the human genome being investigated in detail. These studies will result in a fuller understanding of the specificity and range of biochemical structures and functions that are encoded in the human genome sequence. In general there is likely to remain a distinction between the study of functions encoded at the DNA level, which affect gene expression via transcriptional control, and the study of functions reflected at the protein level following translation, taking into account post-translational modifications (processes which are largely genetically determined). Without a genic catalogue, functional studies are necessarily limited to the investigation of a specific target – a gene, a protein, or a disease. These approaches are an essential part of fully interpreting the genome as they provide a means by which hypotheses can be experimentally tested and which produce valid and supplementary results. However, the production of a complete gene catalogue (if completion can indeed be measured or achieved) will provide the raw material for modelling whole systems. The extensive use of computational biology to suggest how such complex systems are made up of their interacting components will, in itself, enable predictions to be made of the system model. These predictions can be tested, both to determine the validity of the modelled system, and also to test the success of the methods used to derive the system.

A more complete knowledge of biochemical processes will yield a better understanding of complex disease and how it should be treated. At present, our knowledge is primarily based on monogenic diseases. As the problem is reduced to a single gene, hypotheses for function can be tested by biochemical assays, protein structural studies, experimental knock-outs, or the study of naturally occurring mutants. The approach to complex disease centres on a similar approach, i.e. trying to identify the one or few dominant genetic factors which contribute the most significant effect to the overall phenotype. However, there is a realisation that these genetic factors may not fully explain the observed phenotype and that a proportion of the remaining factors may not be identified. In these instances, a comprehensive knowledge of the systems involved will be more informative than the approach of complex disease genetics, both in how the phenotype arises, and how it might be possible to intervene more effectively. This is potentially a true long-term value of the genome sequence, and its interpretation in a biochemical, biological and genetic context, for the advancement of medicine in the future.