Chapter VI Discussion

6.1 Summary

This thesis describes a structural, comparative and functional study of a 3.4 Mb region of human chromosome 22 (22q13.31). Production of a high quality transcript map of the region enabled extensive analysis of the genes encoded within the DNA sequence. Mapping and sequencing of syntenic regions of the mouse genome allowed detailed comparative sequence analysis both of this region and of an upstream syntenic breakpoint junction. The study was extended to include an *in silico* functional characterisation of the proteins encoded within 22q13.31. Additional experimental investigation of the subcellular localisation of a subset of these proteins was also performed.

6.2 Genomic sequence

The work presented in this thesis emphasises the importance of the availability of genomic sequence, as it enables detailed analysis of genes in relation to their genomic environment. The assembly of an experimentally verified transcript map was described in chapter III. In total, 39 genes and 17 pseudogenes were annotated across this region. The high quality transcript data has been integrated with analysis of the surrounding genome to produce a base pair-resolution map that will provide a durable reference for all kinds of future studies.

The genomic sequence provided the basis for a systematic approach to the experimental verification of putative coding sequences. Both positive and negative expression results were referenced against the region of DNA sequence tested, in order to provide a clear picture of sequence transcription. Extensive sequence analysis of the transcribed regions supported earlier studies of the conservation of splice site sequences, but highlighted discrepancies between the sequence context of the putative start codons and the scanning model of translation initiation.

Analysis of these and other features examined in this thesis will soon become possible on a genomic scale with the advent of the finished human sequence. Previously, this type of analysis was only possible on fragmented sequences, often biased towards particular genes or gene families. The availability of the genomic sequence enables a more structured, organised and, once sequencing and annotation are complete, unbiased approach to investigation of both the genomic sequence and its encoded protein products. It will be interesting to determine if the theories previously based on study of fragmented sequences are supported by these studies.

Detailed sequence analysis is reliant on high quality finished sequence. Although the publication and analysis of the draft genome sequence highlighted the utility of unfinished sequence for large-scale analysis of broad features of the human genome, such as GC and repeat content (Lander *et al.*, 2001), resolution of errors and gaps in the draft sequence will enable an unambiguous analysis of these features to be performed. The current imperfect state of the draft genome sequence causes more serious problems in the annotation of human genes. High accuracy is essential in delineation of the protein-coding regions, as ambiguities and errors in unfinished sequence can result in annotation of partial or fragmented genes: predicted genes may also be incorrectly fused or even spurious. Errors leading to alteration of the protein code may affect predictions of function and design of future experiments. The provision of finished sequence will therefore provide a valuable and essential resource for the annotation of human genes.

6.3 Gene annotation

The sequence data generated by the human genome project is paving the way for the identification of the entire complement of human genes. The vast amount of data produced has prompted development of fully automated annotation systems. The Ensembl approach

(Hubbard & Birney, 2000) is based on confirmation of *ab initio* predictions by homology and provides functional annotation via Pfam (Bateman *et al.*, 1999). However, such systems have several limitations. Depending on annotation criteria, if no overlapping similarity information is found, multiple genes may be annotated for what may in fact be a single gene. Artefacts in EST data, arising from unspliced mRNAs, genomic DNA contamination and nongenic transcription (for example from the promoter of a transposable element) detailed both in this project and in the study by Wolfberg and Landman (1997), only confounds this problem as spurious EST data may be used to support incorrect predictions.

A semi-automatic approach, based on sequence similarity information, is utilised in the annotation of the clone-by-clone output of the genome centres. However, this means that genes spanning multiple clones are partially annotated multiple times, a practice that can lead to confusion and redundancy in sequence database entries.

This thesis describes the assembly of a high quality transcript map of human chromosome 22q13.31. Central to the approach used was the availability of high quality, linked, finished genomic sequence spanning nearly the entire region under investigation. Availability of this resource prevented misannotation of genes spanning multiple clones and avoided ambiguities arising from unfinished sequence. All available data, including both expressed sequence evidence and *ab initio* predictions, were manually inspected and gene structures were annotated only when supported by experimental evidence. Where this was absent, additional cDNA sequencing was undertaken to confirm the intron-exon structure. This approach, although arduous, is necessary to ensure high levels of accuracy. Ambiguities generated by inclusion of unsupported gene predictions are thus avoided, although retention of these predictions within the transcript map database (22ace) ensures that this data is easily accessible, if required, together with a record of all cDNA library screens performed. The

transcript map of 22q13.31 therefore provides a strong foundation for future research into this region.

6.4 Mouse genomics

Mapping and sequencing of three regions of the mouse genome with conserved synteny to both 22q13.31 and to a synteny breakpoint junction on 22q13.1, demonstrated the potential utility of mouse genomic sequence for gene annotation and analysis of chromosome evolution. The study also illustrated the justification of the early data release policy implemented by the Sanger Institute and other public domain sequencing centres, as a large amount of information was derived from unfinished mouse genomic sequence.

Although comparative analysis of the mouse sequence with 22q13.31 did not result in the annotation of any further genes, conserved regions were found to correlate closely with the gene annotation. This implies that mouse genomic sequence could provide a powerful tool for gene annotation in less well-studied areas of the human genome. Identification of functionally conserved coding regions could also be useful in the identification of genes that are not represented in the available RNA or cDNA resources, perhaps because of a spatially or temporally limited expression pattern. As this study was largely based on unfinished mouse sequence, identification of potentially conserved coding regions outside of the existing annotation was not considered strong enough grounds for inclusion. However, the increasing availability of finished mouse sequence should permit a more detailed examination of these regions, including the use of sequence similarity searches and gene prediction programs.

Over 30 putative regulatory sequences were identified within the conserved sequences upstream of four annotated transcription start sites. Increased specificity in this study could perhaps be achieved by the inclusion of a third genomic sequence from a vertebrate organism

in the comparison. Recent studies (Frazer *et al.*, 2001; Göttgens *et al.*, 2000) have shown that identification of a conserved non-coding sequence in three vertebrate genomes increases confidence that the putative regulatory sequence is not a false positive. Experimental assays, such as gel retardation, DNase footprinting or methylation interference, could then be carried out to identify protein-binding sites within the region of interest.

This analysis also redefined the boundaries of the syntenic junction of mouse chromosomes 8 and 15 on human chromosome 22q13.1, to a 50 kb region between two adjacent human genes. No potentially causal similarity could be discerned from comparison with breakpoints previously described at the sequence level (Lund *et al.*, 2000; Pletcher *et al.*, 2000). However, as finished mouse sequence for this, and other, syntenic junctions becomes available, a clearer picture of mammalian chromosomal evolution may develop.

An area of finished mouse sequence spanned a ~50 kb gap in the sequence of human chromosome 22, providing a picture of what the equivalent 'unclonable' human sequence may contain. The region includes the 3' exons of the murine orthologue of C22orf1. The putative human exon sequences could be used to design experiments to screen genomic libraries in an effort to close the gap in the human sequence. The comparison of mouse and human GC profiles showed that the mouse GC profile is very similar to that of the human region, but has a lower GC content overall. Interestingly, GC content is raised throughout the murine 'gap' region, and thus possibly exaggerated in the equivalent human region. The high GC content could cause the region to be deletion-prone though frameshift mutagenesis or other unknown cellular mechanism (Bichara *et al.*, 1995, 2000) and thus difficult to clone.

Possibly the most valuable contribution that the mouse genome sequence will make will be to the functional characterisation of orthologous human genes. This thesis illustrates several examples where identification of a functionally characterised murine orthologue permitted

more efficient characterisation of the human protein. As more mapped mouse genomic sequence becomes available, the identification of murine orthologues will become easier. The high quality transcript map of human chromosome 22q13.31, and the comparative sequencing and annotation of the equivalent murine region of conserved synteny described in this thesis, provides an excellent resource for the study of known and potential genetic diseases in this region. This may include further study of spinocerebellar ataxia type 10 (Matsuura *et al.*, 2000), which is caused by the expansion of a pentanucleotide repeat in an intron of the gene E46L. The accurate annotation of this gene onto the genomic sequence and the near availability of finished mouse sequence surrounding the orthologous gene will provide a vital resource for future study of this disorder, both in human populations and in model populations of the laboratory mouse.

6.5 Functional studies

Annotated sequence is now available for much of the human genome, but in the vast majority of cases, the question of gene and protein function remains unsolved. The determination of function is being addressed in a growing number of ways by the emerging field of functional genomics.

This thesis illustrates a selection of these techniques in a preliminary functional characterisation of 27 protein coding genes encoded within 22q13.31. This study represents the first functional analysis of 15 novel genes identified in this region. The importance of a high quality transcript map was demonstrated by this work. Confidence in the gene annotation enabled the generation of cDNA clones containing the full, unambiguous ORF. Discrepancies in the clone insert sequences were easily identified and were systematically assessed to determine of they were due to PCR error, or were an accurate representation of genomic polymorphisms. Inaccurate gene annotation could have led to these discrepancies being

missed, which could potentially have altered the results of functional analyses. Accuracy in gene annotation was also vital for *in silico* investigations of protein function. Much of this work was based upon sequence similarity searches: errors in unverified transcripts could again have potentially altered functional predictions.

A range of software was used to undertake an *in silico* analysis of secondary structure, domain content and subcellular localisation. This type of investigation is amenable to high throughput analysis, but comparison of the results of different algorithms and experimental verification of subcellular localisation highlighted shortcomings in individual programs. However, as the amount of gene data in the public domain increases as a result of the human genome project, such analysis software is likely to improve.

Identification of a previously characterised orthologue was found to be an effective method of attaching a putative function to a protein. This type of analysis is currently less amenable to large-scale study, as current database search techniques cannot distinguish between orthologous genes or merely paralogous matches. The problem is exacerbated by redundancy and examples of poor description of submitted sequences. Some of these problems will be relieved by the increasing amounts of mapped genomic data emanating from the mouse and other model organism sequencing projects, enabling chromosomal location to be taken into account during orthologue identification. Even so, the example of PACSIN2 discussed in this study, where determination of the subcellular localisation of the protein contradicted previous findings from the mouse and chicken orthologues, emphasises the importance of experimental as well as computational investigation in this field.

This thesis illustrates one experimental approach that could be adapted for the high throughput analysis of protein subcellular localisation. Additional high throughput studies are being developed, or are already underway, in order to accumulate information about DNA sequence,

regulatory regions, mRNA profiles, protein expression and interaction and metabolite concentration. Model organisms are also used in functional studies to manipulate the orthologous gene and observe the functional affect. The ultimate aim of functional genomics is to integrate information from all these 'levels' in order to generate effective models of biological systems. The mass of data generated from these projects will necessitate a combination of bioinformatic and experimental approaches. The future challenge to the bioinformatics community will be the integration and finding of patterns in the combined datasets, such as the linking of expression data to genotype and the deduction of genetic pathways from available functional information and expression data.

The generation of a capable and reliable bioinformatic infrastructure is essential to ensure success in this future work to define the functions of the human genome. The beginnings of this infrastructure are already in place through the development of sequence databases and, more recently, whole genome browsers such as Ensembl (Hubbard & Birney, 2000). However, in order to provide a firm foundation for higher-level, interconnecting databases containing functional information, it is necessary to ensure that complete, non-redundant gene and protein information is accurately catalogued. Extensive curation of the existing sequence databases is required to 'clean-up' the thousands of redundant entries that have been generated by the continuous release of genomic sequence over the past few years. A large amount of functional information is already available, both from small-scale investigations of individual genes described in the literature and from high throughput studies. Integration of the existing data into a readily accessible bioinformatic infrastructure will greatly enhance the utility of the previous research and enable an accurate assessment of current functional understanding. Further data can easily, and usefully, be derived using existing *in silico* approaches. For example, a large-scale bioinformatic analysis to establish a database of orthologous protein

relationships across whole genomes, will provide important preliminary information for future research.

It is unlikely that high throughput projects alone will provide all the answers. The functional characterisation of just 38 protein-coding genes described in this thesis illustrates that group analysis generates many different avenues for further study of individual proteins. In this case, further investigation could include individual biochemical assays of the functional domains identified in the protein sequences, experimental confirmation that the functional characteristics of particular orthologous proteins are retained in the human version, or further analysis of the possible post-transcriptional modifications noted from the experimental expression of the protein in COS-7 cells. The accurately annotated human genome sequence and preliminary functional studies described in this thesis, provide an excellent resource for future functional characterisation of these genes.

6.6 Conclusion

This thesis demonstrates the utility of the human genomic sequence in the generation of a high quality transcription map. The availability of the genomic sequence enabled extensive sequence analysis of the annotated genes and their environment. The value of comparative sequence analysis was illustrated through investigation of regions of the mouse genome syntenic to human chromosome 22.The study also illustrates the utility of these genomic resources for functional analysis in the post-genomic era. Both the transcript map and comparative mouse data will provide a valuable tool for future research to further characterise the proteins encoded within 22q13.31.