Chapter VI

Discussion

6.1 Summary

This thesis has described structural, comparative and human variation studies for a 10 Mb region on human chromosome 20q12-13.2. The sequence features of the region were investigated and a detailed transcript map was produced. The syntenic mouse region was mapped and sequenced and the generated data were used for systematic human:mouse comparative analyses. Expressed sequences were used to identify human exonic SNPs *in silico*. A set of 2,208 SNPs mapping across the region was used to obtain allele frequencies in three populations and generate a first generation linkage disequilibrium map of 20q12-13.2 in Caucasians.

6.2 Analysis of genomic sequence

The sequence data generated by the HGP is paving the way for the identification of the entire complement of human genes. As described in Chapter I, the vast amount of data produced has prompted the development of fully automated annotation systems, which at the moment have severe limitations. As a result, the various genome sequencing centres prefer to utilise the semi-automatic approach of computational analysis and manual annotation for their clone sequence output.

This thesis described the assembly of a high quality transcript map. Central to this was the availability of a contiguous, finished genomic sequence spanning the entire region.

Expressed sequences and *ab initio* predictions were manually inspected and gene structures were annotated only when supported by experimental evidence. Where necessary, cDNA isolation and sequencing were undertaken to verify exon-intron boundaries and extend the annotation of incomplete gene structures. Although arduous, this approach is necessary to ensure high levels of accuracy. Un-supported gene predictions were nevertheless stored in the 20ace database for future investigations.

In total, 99 coding genes, 30 putative genes and 36 pseudogenes were annotated, and their structural features including splice sites, alternative isoforms and polyadenylation signals were studied in detail. Predicted CpG islands, promoters and transcription start sites were then correlated with the above data suggesting that the structural annotation of most genes is complete. Furthermore, three species sequence comparisons were used to show that very few exons (<2%), and probably no genes remain un-annotated in this region.

The expression pattern of novel genes was experimentally investigated by screening cDNA libraries. Also, protein analysis of the translated ORF of all coding genes revealed that this region is enriched in genes encoding for proteins with particular domains. All generated data has been integrated with other sequence features such as repeats, segmental duplications and recombination in a single map of 20q12-13.2 to provide an advanced tool for future research in this region. In addition, this study can serve as guide of how to proceed with the annotation of the rest of the genome.

6.3 Mouse genomics

Comparative mapping and sequencing of human chromosome 20q12-13.2 in mouse has shown that the syntenic regions share conservation of gene order and content. The benefit of the early data release policy implemented by the Sanger Institute (and other public domain sequencing centres) was also illustrated, as a large amount of information was derived from unfinished mouse genomic sequence.

Even in a region previously subjected to extensive computational and experimental gene annotation, this approach contributes new, although very few, exons. Although no additional human genes were annotated in this study, conserved regions correlated well with gene annotation implying that mouse genomic sequence could provide a powerful tool for gene annotation. In particular, the mouse sequence can be used to identify conserved regions, corresponding to genes with spatially or temporally limited expression patterns. This possibility was examined in this region by identifying 28 human loci that show conservation and are supported by identical predictions from two prediction software. Testing a subset of these by PCR screening of cDNA libraries did not confirm any as being expressed. Confirming any of these regions as being expressed will be challenging and will probably require a high-throughput method. For example, the mouse sequence could be used to construct DNA chips to be used in hybridisation experiments with several mouse cDNA libraries from a large number of tissues.

It is worth noting that although the data from this study strongly suggest that all human genes of 20q12-13.2 are present in the mouse sequence, the total number of mouse genes has not been assessed.

Unlike coding genes, the putative genes were not conserved in the mouse sequence (discussed in Chapters III and IV). This study has demonstrated that these structures differ significantly from coding genes, and that their possible role remains elusive. Preliminary analysis of finished mouse sequence failed to identify any similar gene structures.

The relatively high extent of synteny between human and mouse across intra- and intergenic regions does not allow the systematic identification of non-transcribed gene features. This could be resolved by performing sequence comparisons with the genome of another organism; the region of 20q12-13.2 could be an ideal candidate for a study of this type. As for the mouse, the available resources and annotation could provide an easy means for the rapid construction of comparative maps. In addition, the virtually complete gene annotation would enable the easy discrimination between transcribed gene elements (exons) and non-transcribed gene elements (e.g. promoters, enhancers).

6.4 Human variation and linkage disequilibrium

Unlike most previous studies, the LD map constructed by this study (chapter V) spans a large (~10 Mb), contiguous segment of the genome. The pairwise values for markers obtained using D’ and r^2 confirmed the extensive variability of LD across the region, and that average half-length LD (D'>0.5) extends to ~80 Kb. Inspection of the pattern of LD along the region revealed that tracts of high LD are situated in regions of low recombination, whereas tracts of low LD are situated in regions of high recombination. Stringent criteria were used to identify 141 "LD blocks", which cover approximately 50% of the region. Increasing the SNP coverage of the regions that are currently outside the defined blocks will extend the current blocks, as well as define other, smaller blocks that have been missed by this study. Overall, the generated data will provide the basis for determining the features of haplotype blocks across 20q12-13.2 and accelerate future association studies for common diseases linked to the region.

6.5 Conclusions and future work

This thesis focused on the structural and comparative analysis of 20q12-13.2. More work will be required to elucidate gene expression and function.

The systematic study of the expression profile of the annotated genes using microarray technology is an attractive option. Given that the sequence of the orthologous mouse genes is available, mouse DNA chips could also be used for detailed expression studies of the various developmental stages, as well as responses to all kinds of stimuli (heat/cold shock, starvation, irradiation, addition of compounds with known molecular targets etc) to provide data on cellular and molecular pathways. The DNA chip approach could also be used to test regions conserved between human and mouse, or regions supported by multiple exon predictions for their coding potential.

Our understanding of the region could also be expanded through comparative analysis with other mammals, for example, the dog. The generated data from such studies could shed light on the evolutionary history of the region. In addition, three-species sequence comparisons could streamline the identification of non-coding regulatory regions. Such regions could then be experimentally tested in a systematic fashion.

The analysis and annotation of the generated finished mouse sequence should also be pursued. The features of a detailed mouse map could be compared to the human orthologs and identify similarities and differences between the two species. Such a map would also be a priceless tool for functional studies, such as mouse knockouts.

Overall, the studies described above will provide the basis for further protein analyses. Combined with a refined haplotype map, these resources could lead to the identification of the genes associated with disorders and provide a strong foundation for understanding the molecular basis of the diseases linked to 20q12-13.2.