# Structural, functional and comparative studies of human chromosome 22q13.31

**Melanie Elizabeth Anne Goward**

**A thesis submitted in partial fulfilment of the requirements of Cambridge University for the degree of Doctor of Philosophy**

**Trinity Hall, Cambridge University**

**January 2002**

**This dissertation is the result of my own work and includes nothing that is the outcome of work done in collaboration. The dissertation does not exceed the length limit set by the Biology Degree Committee.**

## Abstract

As the human genome project nears completion, there is a need to identify and accurately annotate the genes contained within the genomic sequence. The next challenge is the functional analysis of these genes. The aim of this project was to utilise and evaluate different approaches to human gene annotation through analysis of a region of the genomic sequence of human chromosome 22 and then to carry out initial functional studies of the genes identified.

The thesis describes the assembly of a transcript map across a 3.4 Mb region of human chromosome 22 (22q13.31). Candidate gene structures were identified from publicly available expressed sequence evidence and *ab initio* gene predictions, then experimentally verified and extended. This analysis resulted in the annotation of 39 gene and 17 pseudogene structures. Expression of the annotated genes was investigated by Northern blot analysis and RT-PCR screening of RNA isolated from 32 human tissues. The tissue distribution of EST hits to the cDNA sequences were also analysed. The majority of genes demonstrated expression in a wide range of tissues, but the expression of four genes was shown to be limited to reproductive tissues only. Computational analysis of transcription and translation start sites, splice sites and polyadenylation signals showed strong conservation of the sequence contexts necessary for correct transcription and translation. One exception was noted in the gene NUP50, whose features do not correlate with those required by the scanning model of translation initiation.

The contribution that mouse genomic sequence can make, both to human gene annotation and understanding of genome evolution, was evaluated through the construction of bacterial clone maps across a region of mouse chromosome 15, orthologous to human chromosome 22q13.31

and also across a nearby conserved synteny breakpoint between human chromosome 22 and mouse chromosomes 15 and 8. Comparison of available mouse sequence from the mapped clones to the orthologous human regions showed strong conservation of gene order and content, but no conservation of human pseudogenes was noted within the mouse sequence. The analysis of the mouse genomic sequence did not result in extension of the annotation of 22q13.31, but enabled finer mapping of the synteny breakpoint from a 160 kb region on human chromosome 22, to one of 50 kb flanked by adjacent conserved genes.

Functional characterisation was carried out using BLASTP searches to identify protein homologues. The Interpro database was searched to identify protein domains within the amino acid sequences. These results allowed preliminary functional categorisation of the proteins. The localisation of 16 gene products was experimentally determined, by cloning the genes and expressing the encoded proteins in mammalian cells in conjunction with a short peptide tag that conferred antibody specificity. Both N- and C- terminals of each protein were individually tagged. The majority of proteins were distributed in the cytoplasm, with a subset also localised to the cell membrane. An endoplasmic reticular and an unidentified protein localisation pattern were also observed.

Through sequence analysis of regions of human chromosome 22, this project demonstrates and evaluates the contributions that different types of evidence can provide to annotation and analysis of the human genome sequence. It also presents a potential high-throughput approach to determination of protein localisation, which could contribute to the determination of the function of human genes found within the genome.

## Acknowledgements

Many people have kindly provided me with their help and advice throughout the course of this project. The most important of these is Ian Dunham, who has given me guidance and encouragement since the beginning of my time at the Sanger Institute. Thank you for all the time you have spent in helping me to complete this project.

I would like especially to thank the past and present members of the chromosome 22 group, Dave Beare, Charlotte Cole, John Collins, Elisabeth Dawson, Carol Edwards, Owen McCann, Andy Mungall, Tamsin Tarling and Charmain Wright, for their encouragement and practical assistance. I would also like to thank Begoña Aguado, Meera Mallya, David Vetrie and Clare East for their help and advice with the protein and RNA work. Many people from the Sanger Institute have provided me with invaluable assistance and I would particularly like to thank Richard Evans, George Stavrides, Rhian Gwilliam, Karen Halls, Elisabeth Huckle, Carol Carder and Paul Hunt and, from informatics, Carol Scott, Sarah Hunt and Kate Rice.

I would like to also acknowledge the helpful discussions and critical reading of this manuscript by Begoña Aguado, John Collins, Ian Dunham, Richard Evans, Brian Goward and Luc Smink. The fold out diagrams were created with the help of Ewan Birney and printed by Richard Summers.

Finally, I would like to thank my family and friends for their support and encouragement throughout this project.

# Table of contents

## List of tables

## List of figures

## List of abbreviations

| | |
|---|---|
| 22ace | Chromosome 22 implementation of ACeDB |
| aa | Amino Acid |
| ACeDB | A C. elegans DataBase |
| AUM | Asymmetric Unit Membrane |
| BAC | Bacterial Artificial Chromosome |
| BLAST | Basic Local Alignment Search Tool |
| bp | Base pair(s) |
| BSA | Bovine Serum Albumin |
| cDNA | Complementary DNA |
| CDS | CoDing Sequence |
| CM | Cytoplasm and cell membrane |
| Cy | Cytoplasm |
| DMEM | Dulbecco's Modified Eagle Medium |
| DNA | DeoxyriboNucleic Acid |
| EMBL | European Molecular Biology Laboratoriums |
| ePCR | Electronic PCR |
| ER | Endoplasmic Reticulum |
| EST | Expressed Sequence Tag |
| FBS | Fetal Bovine Serum |
| FISH | Fluorescent In Situ Hybridisation |
| FPC | FingerPrinting Contigs |
| gff | Genome Feature Format |
| GFP | Green Fluorescent Protein |
| HSA22 | Homo Sapiens chromosome 22 |
| iATG | Translation Initiation site |
| kb | kilo base pairs |
| LINE | Long INterspersed repeat Element |
| LTR | Long Terminal Repeat |
| Mb | mega base pairs |
| MGC | Mouse Genome Consortium |
| MGD | Mouse Genome Database |
| MGSC | Mouse Genome Sequencing Consortium |
| Mi | Mitochondria |
| MIR | Mammalian-wide Interspersed Repeat |
| MMU8 | Mus Musculus chromosome 8 |
| mRNA | Messenger RNA |
| MS | Mass Spectroscopy |
| NCBI | National Center for Biotechnology Information |
| ncRNA | Non Coding RNA |
| NIH | National Institute of Health |
| NJ | Neighbour-Joining |
| nt | Nucleotide |
| Nu | Nucleus |
| ORF | Open Reading Frame |

| | |
|---|---|
| PAC | P1 Artificial Chromosome |
| PBS | Phosphate Buffered Saline |
| PCR | Polymerase Chain Reaction |
| PIP | Percentage Identity Plot |
| R | Purine |
| RFLP | Restriction Fragment Length Polymorphism |
| RH | Radiation Hybrid |
| RNA | Ribonucleic acid |
| RNAi | RNA Interference |
| rRNA | Ribosomal RNA |
| RT-PCR | Reverse Transcription PCR |
| SINE | Short INterspersed repeat Element |
| Sn | Sensitivity |
| snRNA | Small Nuclear RNA |
| SNP | Single Nucleotide Polymorphism |
| Sp | Specificity |
| SSR | Simple Sequence Repeat |
| STS | Sequence Tagged Site |
| tRNA | Transfer RNA |
| upATG | ATG upstream of the iATG |
| UTR | UnTranslated Region |
| WGS | Whole Genome Shotgun |
| WS1 | Waardenburg Syndrome type 1 |
| WWW | World Wide Web |
| Y | Pyrimidine |
| YAC | Yeast Artificial Chromosome |

## Publication arising from this work

Dunham I., Shimizu N., Roe B. A., Chissoe S., Hunt A. R., Collins J. E., Bruskiewich R., Beare D. M., Clamp M., Smink L. J., Ainscough R., Almeida J. P., Babbage A., Bagguley C., Bailey J., Barlow K., Bates K. N., Beasley O., Bird C. P., Blakey S., Bridgeman A. M., Buck D., Burgess J., Burrill W. D., O'Brien K. P., and *et al*. (1999). The DNA sequence of human chromosome 22. *Nature* **402:** 489-95.