

**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

Table of contents	5
List of tables	10
List of figures	11
List of abbreviations	14
Chapter I Introduction	16
1.1 Introduction	17
1.2 Mapping the human genome	19
1.2.1 Broad Features of the genome	19
1.2.2 Genome maps	20
1.3 Large-scale features of the genome sequence	25
1.3.1 Distribution of GC content	25
1.3.2 CpG islands	26
1.4 Coding and non-coding sequence	27
1.4.1 Non-coding features	27
1.4.2 Coding genome features	30
1.5 Gene Identification	32
1.5.1 Traditional approaches	33
1.5.2 Post-genomic era	64
1.5.3 Comparative studies	37
1.6 Functional genomics	40
1.6.1 Expression studies	41
1.6.2 Control of gene expression	42
1.6.3 Proteomics	45
1.7 Model organisms	47
1.7.1 Model organism genome projects	47
1.7.2 Functional studies in model organisms	48
1.8 Bioinformatics	50
1.9 Chromosome 22	52
1.10 This thesis	54
Chapter II Materials and Methods	57
2.1 DNA manipulation methods	58
2.1.1 Polymerase Chain Reaction	58
2.1.2 Gel electrophoresis	59
2.1.3 Restriction enzyme digests	59
2.1.4 DNA purification	60
2.2 Clone resources	61
2.2.1 Libraries used	61

2.2.2 cDNA clone synthesis	63
2.2.3 Vectorette Library Synthesis	68
2.3 Screening	70
2.3.1 Probe labelling	70
2.3.2 Library screening	71
2.3.3 Vectorette PCR	73
2.4 Landmark production	75
2.4.1 Primer design	75
2.4.2 Primer synthesis	75
2.4.3 Fingerprinting	75
2.4.4 SNP verification	77
2.5 RNA manipulation	77
2.5.1 Steps taken to limit contamination with RNase	77
2.5.2 RNA resources	78
2.5.3 RNA isolation	78
2.5.4 Ethanol precipitation	79
2.5.5 Reverse Transcription PCR (RT-PCR)	79
2.5.6 Northern blotting	81
2.6 Cell Culture and Protein Manipulation	81
2.6.1 SDS-PAGE	81
2.6.2 Western blotting	83
2.6.3 Cell culture and transfection	83
2.6.4 Immunofluorescence	84
2.7 Computational analysis	85
2.7.1 ACeDB	85
2.7.2 Sequence analysis	86
2.7.3 Gene annotation	86
2.7.4 BLAST	86
2.7.5 Perl scripts	87
2.7.6 Calculations of specificity and sensitivity of sequence data	87
2.7.7 Phylogenetic analysis	90
2.8 Materials	91
2.8.1 Buffers	91
2.8.2 Cell culture	93
2.8.3 Size markers	94
2.8.4 Primer sequences	95
2.8.5 URLs and ftp sites	96
Chapter III Transcript map of human chromosome 22q13.31	97
3.1 Introduction	98
3.1.1 Gene identification	98

3.1.2 Ab initio prediction packages	99
3.1.3 Sequence similarity	100
3.1.4 Combination	101
3.1.5 Summary	103
3.2 Gene identification on 22q13.31	105
3.3 Genomic landscape of human chromosome 22q13.31	108
3.3.1 Repeat content	108
3.3.2 GC content	109
3.4 Transcript map of a 3.4Mb region of human chromosome 22	112
3.4.1 Sequence analysis	112
3.4.2 Experimental approaches	114
3.4.3 Transcript mapping results	117
3.5 Investigation of expression	118
3.5.1 Northern hybridisation	119
3.5.2 Construction and screening of expression panel	126
3.5.3 EST tissue origin	129
3.5.4 Overall expression results	130
3.6 Experimental testing of ab initio gene predictions	131
3.6.1 cDNA library screens	131
3.7 Final Transcription map results	134
3.8 Analysis of annotated genes	137
3.8.1 General features of annotated genes	137
3.8.2 Splice sites	140
3.8.3 Investigation of full gene translational start sites	141
3.8.4 Polyadenylation signals	145
3.8.5 Promoter Regions	147
3.8.6 Alternative Splices	153
3.8.7 Paralogues	155
3.9 Correlation of expression evidence with annotated gene features	161
3.9.1 Calculation of specificity and sensitivity	163
3.9.2 Further analysis of Genscan and Fgenesh predictions	166
3.10 Discussion	169
Chapter IV Comparative mapping, sequencing and analysis	176
4.1 Introduction	177
4.1.1 Benefits of comparative sequence analysis	177
4.1.2 The Mouse Genome Projects	178
4.1.3 Comparative Analysis	181
4.1.4 This chapter	185
4.2 Production of regional mouse BAC maps	186
4.2.1 Bacterial clone contig construction	186
4.2.2 Fingerprinting	188

4.2.3 Landmark content mapping	188
4.2.4 Tile Path Clones	190
4.2.5 Features of the sequence-ready bacterial clone map	191
4.2.6 Sequencing	192
4.3 Comparative sequence analysis	196
4.3.1 Dot plot analysis	196
4.3.2 PIP analysis - investigation of exonic conserved sequences	199
4.3.3 Integration of mouse genomic data into 22ace	200
4.4 Correlation of comparative genomic data with 22q13.31 transcript map	203
4.5 Investigation of intronic and intergenic conserved sequences	207
4.5.1 Correlation of Genscan predictions with human-mouse conserved sequences	208
4.5.2 Test for expression	208
4.6 Finished mouse sequence analysis	209
4.6.1 Mouse gene annotation	209
4.6.2 Human-mouse finished sequence alignment	212
4.6.3 GC content	218
4.6.4 Repeat content	220
4.6.5 Comparison of coding regions	223
4.6.6 Splice site comparison	227
4.6.7 Regulatory regions	229
4.7 Chromosome 22 sequence gap	232
4.8 Localisation of synteny breakpoint	235
4.8.1 Definition of the junction region	235
4.8.2 The junction region	239
4.9 Discussion	241
Chapter V Functional characterisation of protein coding genes from 22q13.31	249
5.1 Introduction	250
5.1.1 In silico methods	250
5.1.2 Experimental approaches to determining protein function	254
5.1.3 Summary	256
5.2 Previously published functional data	257
5.3 In silico analysis	258
5.3.1 Intrinsic feature analysis	259
5.3.2 Domain Analysis	262
5.3.3 Orthologues	267
5.3.4 In silico prediction of subcellular localisation	277
5.4 Experimental analysis of subcellular localisation	279
5.4.1 Overall strategy	279
5.4.2 Selection and generation of full-length cDNA clones	280
5.4.3 Addition of T7.Tag	285
5.4.4 Expression in COS-7 cells	287
5.4.5 Analysis of T7.Tag protein subcellular location	291
5.5 Data integration	298
5.6 Discussion	299

Chapter VI Discussion	306
6.1 Summary	307
6.2 Genomic sequence	307
6.3 Gene annotation	308
6.4 Mouse genomics	310
6.5 Functional studies	312
6.6 Conclusion	317
Chapter VII References	316
Appendices	
Appendix 1	342
Appendix 2	357
Appendix 3	CD
Appendix 4	360
Appendix 5	363
Appendix 6	364
Appendix 7	CD
Appendix 8	CD

List of tables

Table 1.1	Properties of chromosome bands seen with standard Giemsa staining	20
Table 1.2	The model organisms initially proposed for genome sequencing	48
Table 1.3	Syndromes linked to chromosome 22 genes	53
Table 2.1	Details of the mouse genomic library used	61
Table 2.2	cDNA resources used during the course of this project	62
Table 2.3	RNA resources used during the course of this project	78
Table 2.4	Perl scripts used during the course of this project	87
Table 2.5	1 kb ladder (GibcoBRL)	94
Table 2.6	Benchmark™ Prestained Protein Ladder (GibcoBRL)	95
Table 2.7	Useful URL and ftp sites	96
Table 3.1	% repeat coverage and density	109
Table 3.2	GC content, amount of DNA and isochore correspondence	110
Table 3.3	Initial feature identification in 22q13.1	112
Table 3.4	Distribution of generated cDNA sequences	117
Table 3.5	Expected and obtained transcript sizes from Northern blot hybridisations	124
Table 3.6	Key to tissue identity	129
Table 3.7	Sequence reads from PCR products amplified from Genscan predictions	132
Table 3.8	Number and type of annotated gene features	135
Table 3.9a	Pseudogenes annotated within 22q13.31	135
Table 3.9b	Genes annotated within 22q13.31	136
Table 3.10	Mean and median values for a range of protein coding gene properties	138
Table 3.11	Possible downstream ATG translation initiation mechanisms	144
Table 3.12	The presence/absence of polyadenylation signals and cleavage sites	147
Table 3.13	Correlation of predicted promoter regions and CpG islands with gene annotation	152
Table 3.14	Potential alternative splices from 22q13.31	154
Table 3.15	Genes putatively paralogous to full genes from 22q13.31	
Table 3.16	Correlation analysis of the evidence used to annotate genes	164
Table 3.17	Genscan and Fgenesh predictions	167
Table 3.18	Correlation Genscan and Fgenesh predictions with annotated genes	168
Table 4.1	Numbers of pools, markers and isolated clones in the initial library screens	188
Table 4.2	Numbers of pools, end STSs and isolated clones in gap closure screens	190
Table 4.3	Clone contig data	190
Table 4.4	Incorporation of marker information into mouse contigs A, B and C	191
Table 4.5	Overview of PIP results	200
Table 4.6	Mouse clones and orthologous regions of HSA22q13.31 selected for percentage identity analysis	201
Table 4.7	Correlation analysis of the evidence available from different organism genome or gene identification projects	205
Table 4.8	Genscan predictions that do not overlap annotated true exons, but overlap human-mouse conserved regions	208
Table 4.9	The annotated mouse genes	210
Table 4.10	Percentage identities of mouse and human gene sequences	225

Table 4.11	TRANSFAC screen results	231
Table 4.12	Sequence clones adjacent to and spanning the syntenic breakpoint.	236
Table 5.1	Available functional information for 12 mRNAs and/or proteins encoded within human chromosome 22q13.31	258
Table 5.2	Domain-containing proteins	267
Table 5.3	Key to figures 5.3 and 5.4	270
Table 5.4	Potential orthologues of proteins from 22q13.31	273
Table 5.5	Cloned cDNAs from 22q13.31	281
Table 5.6	Discrepancies discovered between cDNA clone and genomic sequences	282
Table 5.7	Generation of N- and C-terminally T7 tagged cDNA inserts	287
Table 5.8	SDS-PAGE expected and obtained protein sizes	290
Table 5.9	Subcellular localisation of 16 proteins encoded within 22q13.31	296
Table 5.10	Overall functional characteristics of 27 protein coding genes encoded within human chromosome 22q13.31	298

List of figures

Figure 1.1	Protein-coding gene features	32
Figure 2.3	Measures of sequence correlation with annotated gene structures	88
Figure 3.1	Automated analysis strategy	105
Figure 3.2	Chromosome 22 additional analysis	106
Figure 3.3	An example of the ACeDB display	107
Figure 3.4	Repetitive and non-repetitive DNA coverage (%) for region of interest	109
Figure 3.5	Transcript map of 22q13.31	111
Figure 3.6	Example of vectorette PCR	115
Figure 3.7	Vectorette cDNA library screens	116
Figure 3.8	Results from 41 Northern blots	121
Figure 3.9	Example of an RT-PCR experiment	127
Figure 3.10	Transcription profiles for 41 genes annotated in 22q13.31	128
Figure 3.11	e-profile results from dJ22E13.C22.3a (Em:AL160111)	130
Figure 3.12	Vectorette library screens of Genscan predictions	132
Figure 3.13	Splice donor and acceptor consensus sequences of 379 introns in 22q13.31	141
Figure 3.14	Translational start site consensus	142
Figure 3.15	Analysis of the sequence contexts surrounding 27 initiator codons from 22q13.31	143
Figure 3.16	An example of Blixem output from ACeDB	146
Figure 3.17	Correlation of predicted promoter and transcription start site regions with 27 annotated full genes	150
Figure 3.18	Venn diagram showing the number of full gene structures and their correlation with different kinds of promoter prediction algorithms	151
Figure 3.19	Approximate positions of genes putatively paralogous to full genes on 22q13.31	
Figure 3.20	Schematic showing a region of interchromosomal duplication on chromosome 22	157
Figure 3.21	Annotated dot plot from identifying an intrachromosomal duplication	159

	within chromosome 22	
Figure 3.22	Alignment of the amino acid sequences of bK126B4.C22.2 and dJ222E13.C22.1	160
Figure 3.23	Alignment of the nucleotide sequences of bK126B4.C22.3 and dJ222E13.C22.2	160
Figure 3.24	Specificity and sensitivity of sequence evidence alignment with the 22q13.31 transcript map	165
Figure 3.25	Specificity and sensitivity of the alignment of ab initio gene prediction programs with a variety of annotated human sequences	168
Figure 4.1	Contig construction strategy combining both landmark-content mapping and restriction enzyme fingerprinting	179
Figure 4.2	Screening strategy	187
Figure 4.3	Example of landmark-content mapping	189
Figure 4.4	Bacterial clone contigs containing mouse genomic sequence spanning regions of conserved synteny with a) human chromosome 22q13.31 and b) human chromosome 22q13.1	193
Figure 4.5	Mouse BAC clone contigs spanning orthologous regions of human chromosome 22	195
Figure 4.6a	Annotated dot plot of the human sequence of 22q13.31 (X-axis) and orthologous mouse (Y-axis) sequences from MMU 15	197
Figure 4.6b	Annotated dot plot of the human sequence of a 1.96 Mb region of 22q13.1 (X-axis) and orthologous mouse (Y-axis) sequences from MMU15 and MMU8	198
Figure 4.7	Sensitivity and specificity of MatchReport BLAST results from three mouse clone sequences against the equivalent human genomic sequence	202
Figure 4.8	22ace display showing the region surrounding the gene dJ526I14.C22.2	204
Figure 4.9	Specificity and sensitivity of different comparative sequence data with the 22q13.31 transcript map	209
Figure 4.10	Alignment of the human and mouse annotated genes	211
Figure 4.11	Annotated dot plot of the mouse (x-axis) and human (y-axis) sequences.	213
Figure 4.12	Percentage identity plot calculated by PipMaker for the human interval TLL1 to dJ345P10.C22.4, compared with sequence from the region of conserved synteny on mouse chromosome 15	215
Figure 4.13	Human and mouse GC distribution	219
Figure 4.14	Comparison of human and mouse CpG island GC content (A) and length (B)	221
Figure 4.15	Repeat density (A) and genomic coverage by repeats (B) for human and mouse	222
Figure 4.16	Scatter plots depicting (A) exon sizes and (B) intron sizes between human and mouse gene structures. (C) A more detailed view of the 500 bp exon interval is also shown	224
Figure 4.17	A) Alignment of Biklk and BIK	227
Figure 4.18	The splice acceptor and donor sites for human (A) and mouse (B)	228
Figure 4.19	Sequence alignment of mouse and human sequence upstream of TLL1 and bM121M7.1	230
Figure 4.20	Diagram showing GC content, gene content and repeat content (mouse	233

	sequence only) of sequence spanning an ‘unclonable’ sequence gap in human chromosome 22	
Figure 4.21	Repetitive and non-repetitive DNA distribution of 30216bp of mouse sequence, spanning an equivalent ‘unclonable’ sequence gap in human chromosome 22	234
Figure 4.22	Annotated dot plot of regions of mouse chromosome 8 and 15 available sequences (Y-axis) against the syntenic region of human chromosome 22 sequence (X-axis)	237
Figure 4.23	Comparative maps define the MMU8.	238
Figure 4.24	Comparative sequence analysis defines the MMU8:15 chromosome junction region on human chromosome 22	239
Figure 5.1	PIX display out put showing analysis of the translated coding sequence of dJ222E13.C22.1 (isoform a)	262
Figure 5.2	Results from both the secondary structure and domain analysis	264
Figure 5.3	Clustalw alignment of the amino acid sequence of dJ222E13.C22.1 against five homologous protein sequences identified from a BLASTP search of the NCBI nonredundant protein sequence database	269
Figure 5.4	Phylogenetic tree derived from the above alignment using the Phylowin package	270
Figure 5.5	Phylogenetic trees derived using NJ methodology from clustalw protein alignments	272
Figure 5.6	The predicted domain and secondary structures of both proteins from 22q13.31 and functionally characterised potential orthologues	275
Figure 5.7	Predicted subcellular localisation	278
Figure 5.8	Blixem alignment of dJ549K18.C22.1 cDNA clone sequencing reads	281
Figure 5.9	Visual display from Gap4 database	283
Figure 5.10	Inspection of the forward and reverse sequence traces from two (of 24) individuals	284
Figure 5.11	Schematic of the mammalian cell expression vector pCDNA3-T7-C	286
Figure 5.12	Schematic showing strategy used to generate N- and C- terminally T7-tagged clones	288
Figure 5.13	Western blot analysis of transiently transfected COS-7 cells. N- and C-terminally tagged constructs are shown	289
Figure 4.14	Examples of immunofluorescence experiments of COS-7 cells, transiently transfected with N- and C- terminally T7 tagged constructs	292
Figure 5.15	An example of possible aggresome formation	297
Figure 5.16	Schematic representation of the regulation of ADAM13 by X-PACSIN2	302

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 Laboratories
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.