Physical mapping and identification of novel genes in human chromosome 21q11

A thesis submitted for the degree of
Doctor of Philosophy.
Faculty of Medicine, University of London.

By

Jürgen Groet

November 1999

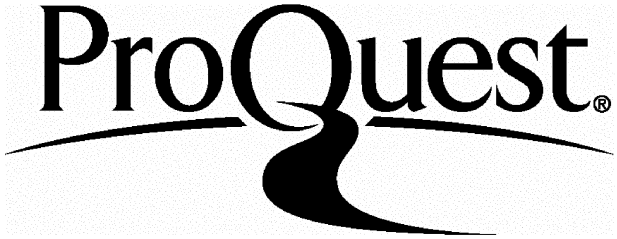
The School of Pharmacy
University of London
29/39 Brunswick Square
London WC1N 1AX

ProQuest Number: 10104723

All rights reserved

INFORMATION TO ALL USERS
The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript
and there are missing pages, these will be noted. Also, if material had to be removed,
a note will indicate the deletion.



ProQuest 10104723

Published by ProQuest LLC(2016). Copyright of the Dissertation is held by the Author.

All rights reserved.
This work is protected against unauthorized copying under Title 17, United States Code.
Microform Edition © ProQuest LLC.

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

For Barbara and Jake

Abstract

A physical map, of approximately 3.2 Mbp, consisting of overlapping genomic inserts contained within bacterial recombinant clones was constructed in the q11 region of human chromosome 21. This region has been associated with Alzheimer's disease, non-small cell lung carcinoma (NSCLC) and a specific form of acute non-lymphocytic leukaemia. Exon trapping was performed, for the identification of novel genes, on twenty-two overlapping clones from the physical map. Ninety-three different species of putative exons were analysed and compared with non-redundant, cDNA and protein databases. The DNA sequence of nineteen exon trapped clones was identical to human cDNA clones. The most significant similarities of translations of the exon trapped products were found with a rat lipase, a human multi-drug resistance protein and a ubiquitin carboxyl-terminal hydrolase from *C. elegans*. Analysis of overlapping cDNA clones, exon trapped material and products from rapid amplification of cDNA ends, led to the identification of two novel genes. The first gene (3803 bp, 1055 amino acids), which was located in a smallest region of overlapping deletion in tumour DNA from patients with NSCLC, showed very high similarity to ubiquitin-specific proteases. This function was studied by an *in vitro* assay, which showed de-ubiquitination of a model fusion protein by the protein product of the novel gene. Sequencing of exons from this gene using genomic tumour DNA from patients with NSCLC and late onset Alzheimer's disease did not show any mutations. The second gene (1888 bp, 373 amino acids) showed similarity to the protein products of novel putative genes located on chromosomes Xq25-26.1 and 6q22.3. The function of this protein is unknown. Analysis of the coding sequence of the gene using genomic DNA from patients with NSCLC, familial early onset Alzheimer's disease, late onset Alzheimer's disease and from Down syndrome patients with acute non-lymphocytic leukaemia failed to show any mutations.

Acknowledgements

A number of people have been very important with their contributions to the work described in this thesis.

Many thanks to Dr. Dean Nizetic, my supervisor, for giving me the chance to carry out a PhD, for his encouragement, vast experimental know how and brilliant theories.

I would like to thank Dr. Jane Ives, the first partner in crime, whose enthusiasm and knowledge of the subject area have helped enormously in the realisation of this work. Also my greatest appreciation for all the work you did concerning corrections and suggestions for the manuscript.

Thanks to Dr. Rachel Flomen for her excellent suggestions concerning some of the techniques.

Also many thanks to the boys, Dr. Andrew South and Mr. Pedro Baptista, for their help in the initial part of the work, quick pints at lunch time and long pints after working hours.

All thanks to our collaborators, who are mentioned in the text, for their contributions to the work.

Special thanks to my parents and Nicole and Jorge for their support and interest in the research.

Lastly, and most importantly, all thanks to my wife Barbara, without whom I would never have started a PhD. Your continual support, especially during moments of self doubt, have been of vital importance during this difficult period. The happiest event in recent time has been the birth of our son Jake.

Table of Contents

Abstract 3

Acknowledgements 4

Table of Contents 5

List of Figures 12

List of Tables 15

Glossary of Abbreviations 17

Chapter 1 Introduction 20

1.1 Towards the identification of all human genes 20

1.2 Methods to identify novel genes 20

1.2.1 Linkage studies and polymorphic markers 20

1.2.2 Physical mapping using genomic libraries 22

1.2.3 Exon trapping and cDNA selection 24

1.2.4 Global, whole transcriptome projects of sequencing and mapping of ESTs 26

1.2.5 The Human Genome Project 27

1.3 Major types of genetic diseases 30

1.3.1 Single gene disorders 30

1.3.1.1 Autosomal dominant 30

1.3.1.2 Autosomal recessive 31

1.3.1.3 X-linked 31

1.3.2 Disorders caused by chromosomal aberrations 32

1.3.3 Somatic cell genetic diseases 33

1.3.4 Complex traits 34

1.4 Genetic diseases mapping to chromosome 21 35

1.4.1 Monogenic/somatic diseases 35

1.4.1.1 SOD1 in motor neuron disease 35

1.4.1.2 APP in familial early onset Alzheimer's disease 37

1.4.1.3 AML1 (CBFA2) in leukaemia 38

1.4.1.4 The EPM1 gene in progressive myoclonus epilepsy 39

1.4.1.5 AIRE in autoimmune-polyendocrinopathy-candidiasis-ectodermal dystrophy 39

1.5 Down syndrome: a trisomy of chromosome 21 40
1.5.1 Phenotypic features of Down syndrome 40
1.5.2 Analysis and models for nondisjunction of chromosome 21 41
1.6 Gene density and resources for identification/mapping of genes on chromosome 21 43
1.6.1 Gene density on chromosome 21 43
1.6.2 Genetic and physical maps 44
1.6.3 Sequencing of chromosome 21 45
1.7 Genotype/phenotype mapping of Down syndrome phenotypes 46
1.7.1 Mouse models 46
1.7.2 Partial trisomy 21 in humans 47
1.8 Human chromosome 21q11 49
1.8.1 Phenotypes associated with chromosome 21q11 49
1.8.1.1 Lung cancer 49
1.8.1.2 Alzheimer's disease 50
1.8.1.3 Leukaemia in Down Syndrome 51
1.8.2 Similarity between 21q11 and other chromosomal regions 53
1.8.2.1 Centromere 53
1.8.2.2 The duplicated region 53
1.8.2.3 Proximal 21q11 54
1.8.3 Genes, cDNA clones, exon trapped material and physical maps of chromosome 21q11 54
1.8.3.1 Gene density of chromosome 21q11 54
1.8.3.2 Physical maps of 21q11 55
1.8.3.3 Characterised genes, cDNA clones and exon trapped material mapping to chromosome 21q11 56
1.9 Aims of the thesis 57

Chapter 2 Materials and Methods 58

2.1 Chemical reagents 58
2.2 Enzymes 59
2.3 Nucleotides and oligonucleotides 59
2.4 Cloning vectors 59
2.5 DNA size markers 60

2.6 Hybridisation membranes and X-ray film 60
2.7 Northern blots 60
2.8 Solutions 60
2.9 PCR Primers 62
2.9.1 IPCR/vectorette primers 62
2.9.2 Marker primers 62
2.9.3 Exon trapping and vector primers 63
2.9.4 Unigene primers 63
2.9.5 STCH primers 64
2.9.6 USP21 primers 64
2.9.7 G31B5 primers 65
2.9.8 APOE primers 66
2.10 Markers used for hybridisation 66
2.11 Somatic cell hybrids 66
2.12 PAC, BAC, Cosmid libraries 66
2.13 Cell lines 66
2.14 *E. coli* strains and constructs 67
2.15 cDNA selected clones, exon trapped material, mapped ESTs (Unigenes) 67
2.16 Materials obtained from the Human Genome Mapping Project Resource Centre 67
2.17 Clinical Material 67
2.18 General Protocols 68
2.18.1 Picking, replicating and processing 68
2.18.2 DNA isolation 69
2.18.3 Standard PCR 69
2.18.4 Electro-elution 70
2.18.5 Southern Blotting 70
2.18.6 Dot blots 70
2.18.7 Labelling of DNA probes 71
2.18.8 Hybridisation 71
2.18.9 PAC/BAC insert isolation 72
2.18.10 Preparation of competent cells 72
2.18.11 Tissue Culture 73

2.18.12 RACE, cloning of RACE products and analysis 73
2.18.13 DNA isolation from lymphoblast cell lines from familial early onset Alzheimer's disease patients 75
2.19 Protocols specific for Chapter 3 75
2.19.1 Inverse PCR (IPCR) 75
2.19.2 Vectorette PCR 76
2.19.3 Generation of end fragments via the degenerate oligonucleotide primer method 76
2.19.4 Isolation of *Sau* 3A I fragments from cosmid inserts 77
2.19.5 Restriction mapping 77
2.19.6 Isolation of genomic DNA for PFGE 77
2.20 Protocols specific for Chapter 4 78
2.20.1 Exon trapping 78
2.21 Protocols specific for Chapter 5 81
2.21.1 Cloning of USP21 81
2.21.2 Cleavage analysis of ubiquitin-Met-β-gal fusion product by USP21 83
2.21.3 Genescan using marker D21S366 84
2.21.4 Hybridisation of 11-mer primers to determine loss of heterozygosity of USP21 85
2.21.5 Sequencing of the USP21 gene using patient DNA 86
2.22 Protocols specific for Chapter 6 87
2.22.1 Direct sequencing on DNA from a BAC clone 87
2.22.2 Sequencing of the G31B5 gene using patient DNA 87

Chapter 3 Bacterial contig map of the 21q11 region 88
3.1 Summary 88
3.2 Introduction 88
3.2.1 Mapping of human chromosome 21 88
3.2.2 Possible involvement of the 21q11 region in a number of disorders 89
3.2.3 Structural features of the pericentromeric DNA of chromosome 21 90
3.3 Results 91
3.3.1 Identification of bacterial clones containing genomic inserts from the 21q11 region 91

3.3.2 Detection of overlapping 21q11 PACs and construction of contigs anchored to molecular markers 94
3.3.2.1 The interval between the markers D21S190 and D21S308 (proximal quarter of the 21q11 region) 94
3.3.2.2 The interval between the markers D21S149 and D21S236 (second quarter of the 21q11 region) 100
3.3.2.3 The interval between the markers D21S371 and D21S13 (third quarter of the 21q11 region) 105
3.3.2.4 The interval between the markers D21S46 and D21S189 (distal quarter of the 21q11 region) 110
3.3.3 Physical measurement of the two gaps in the physical map and closure of gap 2 114
3.3.4 Restriction mapping of the genomic DNA contained in bacterial clones, verification of overlaps, and measurement of the degree of overlap between the clones 118
3.3.5 Localisation of published genes and transcribed DNA fragments on the 21q11 contig map 126
3.3.6 Integrated map of 21q11 127
3.4 Discussion 133
3.4.1 Comparison of the bacterial contig map with maps of uncloned genomic DNA and YAC based maps 133
3.4.2 DNA sequence similarity of proximal 21q11 to sequences on different chromosomes 134
3.4.3 Comparison of the integrated map of the 21q11 region with long range sequencing data 136

Chapter 4 Exon Trapping using genomic clones in the 21q11.2 region 140
4.1 Summary 140
4.2 Introduction 140
4.2.1 Identification of novel genes in human chromosome 21q11.2 140
4.2.2 Exon trapping: a method to identify novel genes 141
4.2.2.1 Exon trapping vectors using one splice site 143
4.2.2.2 Exon trapping vectors using two splice sites 143

4.2.3 Exon trapping using bacterial clones located in the 21q11.2 region 144
4.3 Results 145
4.3.1 Exon trapping using bacterial clones containing overlapping genomic inserts from chromosome 21q11 145
4.3.2 Analysis of the exon trapped clones 148
4.3.2.1 Picking of putative exon trapped clones into 384-well microtitre dishes 148
4.3.2.2 Analysis of random exon trapped clones by sequencing 148
4.3.2.3 Identification of exon trapped clones for further analysis 151
4.3.2.4 Analysis of exon trapped clones via sampling without replacement 152
4.3.2.5 Sequencing and database searches of the exon trapped clones 156
4.3.3 Similarities between exon trapped clones and human cDNA clones 164
4.3.3.1 Exon trapped clone 90B5S-E6 164
4.3.3.2 Exon trapped clone 98L15BB-N7 164
4.3.3.3 Exon trapped clones 98L15S-G9 and 90B5BB-A22 164
4.3.3.4 Exon trapped clones 98L15S-A9 and 90B5BB-C21 166
4.3.3.5 Exon trapped clones 31B5-E21 and 31B5-A3/B39I12-F7 166
4.3.3.6 Exon trapped clones 30P13-F11 and 270M7-A3 169
4.3.3.7 Exon trapped clones 135E14-D6, 73M5-A2, 73M5-I13/111K10-H5, 73M5-H7, 73M5-K13 and 111K10-K12 169
4.3.3.8 Exon trapped clones 135E14-B7 and 135E14-C6 169
4.3.3.9 cDNA clusters A004O28 and A002B43 and exon trapped clone 135E14-B4 173
4.4 Discussion 173

Chapter 5 Identification of a novel gene mapping to proximal 21q encoding a ubiquitin-specific protease 178
5.1 Summary 178
5.2 Introduction 178
5.2.1 Genetic defects of chromosome 21 178
5.2.2 Disorders associated with chromosome 21q11 179
5.3 Results 180
5.3.1 Identification of a novel gene encoding a possible ubiquitin-specific protease 180
5.3.2 *In vitro* characterisation of the novel gene as a ubiquitin-specific 191

protease
5.3.3 Identification of the smallest 21q11 sub-region of overlapping deletion, containing USP21, in non-small cell lung carcinoma patients 194
5.3.4 Sequencing of the USP21 gene in genomic DNA from patients with non-small cell lung carcinoma and sporadic Alzheimer's disease 197
5.4 Discussion 201

Chapter 6 Identification and cloning of a novel gene, G31B5, located in Human chromosome 21q11 207
6.1 Summary 207
6.2 Introduction 207
6.2.1 Examples of genetic defects involving chromosome 21 207
6.2.2 Genetic defects associated with chromosome 21q11 208
6.2.3 Physical maps that include the 21q11 region 209
6.2.4 Prediction of the gene density of chromosome 21q11 209
6.2.5 Genes which map to 21q11 210
6.3 Results 210
6.3.1 Identification of a novel gene mapping to bacterial clones 31B5 and B39I12 210
6.3.2 Direct sequencing using bacterial clone B39I12, to characterise three additional exons of G31B5 214
6.3.3 Identification of a mouse homologue of G31B5 and comparison of the proteins between the two species 217
6.3.4 Sequencing of the gene using genomic DNA from clinical material 219
6.4 Discussion 221

Chapter 7 General discussion 226
Summary of conclusions 232
References 234
Appendix I Tables 265
Appendix II Publications 292

List of Figures

Chapter 1

Fig. 1.1	Cytogenetic map of chromosome 21 and positions of disease-causing genes and the "Down syndrome chromosome region"	36

Chapter 3

Fig. 3.1	Cytogenetic map of chromosome 21 and YAC map of 21q11	92
Fig. 3.2	Two methods for generating end fragments from bacterial clones	95
Fig. 3.3	Overlapping clones in the proximal quarter of 21q11	98
Fig. 3.4	Overlapping clones in the second quarter of 21q11	102
Fig. 3.5	Identification by PCR, of PAC clones from the RPCI 1 library, which contain the markers D21S120 and D21S258	104
Fig. 3.6	Hybridisation of the genomic insert of the PAC clone 31B5 to the total human PAC library	106
Fig. 3.7	Overlapping clones in the third quarter of 21q11	109
Fig. 3.8	Overlapping clones in the distal quarter of 21q11	112
Fig. 3.9	Measurement of the size of the two gaps in the physical map	115
Fig. 3.10	Confirmation that two cosmid clones close gap 2	117
Fig. 3.11	Restriction mapping and verification of the degree of overlap	119
Fig. 3.12	Restriction mapping and determining of the degree of overlap of the minimal tiling path clones	121
Fig. 3.13	Integrated map of 21q11	128
Fig. 3.14	Fully sequenced bacterial clones of the 21q11 region	138

Chapter 4

Fig. 4.1	Strategy of exon trapping	142
Fig. 4.2	Generation of fragments from genomic clones for cloning into pSPL3B	146
Fig. 4.3	Examples of PCR amplified fragments after exon trapping and negative controls	146
Fig. 4.4	Strategy for the analysis of the exon trapped clones	150
Fig. 4.5	Hybridisations for the identification of exon trapped clones for	153

	further analysis	
Fig. 4.6	Analysis of exon trapped clones	155
Fig. 4.7	Localisation of the exon trapped clones	157
Fig. 4.8	cDNA cluster with putative exon 98L15BB-N7	165
Fig. 4.9	cDNA cluster with putative exons 90B5BB-A22 and 98L15S-G9	165
Fig. 4.10	cDNA cluster with putative exons 90B5BB-C21 and 98L15S-A9	167
Fig. 4.11	cDNA cluster with putative exons 31B5-E21 and 31B5-A3/B39I12-F7	167
Fig. 4.12	cDNA cluster with putative exon 30P13-F11	168
Fig. 4.13	cDNA cluster with putative exon 270M7-A3	168
Fig. 4.14	cDNA cluster with putative exons 135E14-D6, 73M5-H7, 73M5-A2, 73M5-K13, 73M5-I13/111K10-H5 and 111K10-K12	170
Fig. 4.15	cDNA cluster with putative exons 135E14-B7 and 135E14-C6	171
Fig. 4.16	A004O28 Unigene cDNA cluster	171
Fig. 4.17	A002B43 Unigene cDNA cluster and the putative exon 135E14-B4	172

Chapter 5

Fig. 5.1	Hybridisation of the Image clone 824710 and the cDNA cluster A002B43 to multiple tissue Northern blots	181
Fig. 5.2	The novel gene USP21; assembly and localisation	183
Fig. 5.3	Nucleotide and amino acid sequence of USP21	186
Fig. 5.4	Comparison of amino acids between the functional domains of USP21 and of other ubiquitin-specific proteases	190
Fig. 5.5	Cloning of USP21 in a prokaryotic expression vector	192
Fig. 5.6	*In vitro* cleavage of ubiquitin-M-β-galactosidase by *USP21*	193
Fig. 5.7	Identification of a smallest region of overlapping deletion in non-small cell lung carcinoma samples	195

Chapter 6

Fig. 6.1	Hybridisation of the exon trapped product 31B5-E21 to a multiple tissue Northern blot	211

Fig. 6.2	Identification of 3' and 5' RACE clones from the exon trapped clone 31B5-E21	211
Fig. 6.3	The novel gene G31B5; assembly and localisation	213
Fig. 6.4	Nucleotide and amino acid sequence of G31B5	215
Fig. 6.5	Nucleotide and amino acid sequence of M31B5, a mouse homologue of the human gene G31B5	218
Fig. 6.6	Comparison of the proteins of G31B5 and M31B5	218
Fig. 6.7	Sequencing of G31B5 from clinical material	220

List of Tables

Chapter 3

Table 3.1	Number of PAC clones detected by YAC hybridisation of 21q11 YACs	93
Table 3.2	Results of hybridisations and PCRs from the proximal quarter of 21q11	97
Table 3.3	Results of hybridisations and PCRs from the second quarter of 21q11	101
Table 3.4	Results of hybridisations and PCRs from the third quarter of 21q11	108
Table 3.5	Results of hybridisations and PCRs from the distal quarter of 21q11	111
Table 3.6	Localisation of B62L20, 29H4 and 82L10 to chromosome 21	135
Table 3.7	FISH analysis of bacterial clones from the 21q11 contig	137

Chapter 4

Table 4.1	Distribution of exon trapped clones in 384-well microtitre dishes	149
Table 4.2	Characteristics of the exon trapped clones	160

Chapter 5

Table 5.1	Sizing of exons and introns, analysis of splice junctions and determination of the phase of the introns of USP21	188
Table 5.2	LOH determined via a polymorphism of USP21	198
Table 5.3	Exon predictions in the genomic interval of the USP21 gene	203

Chapter 6

Table 6.1	Sizing of exons and introns, analysis of splice junctions and determination of the phase of the introns of G31B5	216
Table 6.2	Exon predictions in the genomic interval of the G31B5 gene	223

Appendix I

Table 1	Location of the potential 21q11 PAC clones in 96-well microtitre dishes	265
Table 2	Hybridisation results of the membranes containing the putative exons	269
Table 3	Sequences of the exon trapped clones	280
Table 4	Sequences generated for the analysis of USP21	289
Table 5	Sequences generated for the analysis of G31B5	290

Glossary of Abbreviations

1^0	first round of PCR amplification
2^0	second round of PCR amplification
3^0	third round of PCR amplification
AD	Alzheimer's disease
BAC	bacterial artificial chromosome
bp	base pair
^0C	degrees Celsius
CEPH	Centre d'Etude de Polymorphisme Humain
cDNA	complementary DNA
cM	centi Morgan
cm	centimeter
contig	continuum of DNA, defined by overlapping recombinant DNA clones
CpG	5' CG 3' dinucleotide
cpm	counts per minute
dbEST	EST database
DDBJ	DNA data bank Japan
DMSO	dimethylsulphoxide
DNA	deoxyribonucleic acid
dNTP	3' deoxynucleoside 5'-triphosphate
dATP	3' deoxyadenosine 5'-triphosphate
dCTP	3' deoxycytidine 5'-triphosphate
dGTP	3' deoxyguanosine 5'-triphosphate
dTTP	3' deoxythymidine 5'-triphosphate
DS	Down syndrome
DTT	dithiothreitol
EDTA	ethylenediaminetetraacetic acid
E. coli	*Escherichia Coli*
EMBL	European Molecular Biology Laboratory
ES cells	embryonic stem cells
EST	expressed sequence tag
FEOAD	familial early onset Alzheimer's disease
FISH	fluorescent *in situ* hybridisation
g	acceleration due to gravity
h	hour
HEPES	N-2-hydroxyethylpiperazine-N'-2 ethanesulphonic acid
HGMP	human genome mapping project
HGP	human genome project
HIV *tat* intron	intron of the human immunodeficiency virus transactivator gene

HMFM	Hogness modified freezing medium
IAA	isoamylalcohol
IMAGE	integrated molecular analysis of genomes and their expression
IPTG	isopropylthiogalactoside
Kbp	kilobase pair
kDa	kilo Dalton
kg	kilogram
kV	kilo Volts
LB	Luria-Bertani
LOAD	late onset Alzheimer's disease
lod score	$\log_{10}$ of the odds ratio
LOH	loss of heterozygosity
M	molar
mA	milli Amperes
Mbp	megabase pair
mCi	milliCurie
M-MLV	molony murine leukaemia virus
MOPS	3-[N-morpholino] propane-sulphonic acid
µF	micro Farad
mg	milligram
µg	microgram
mJ	milli Joule
ml	millilitre
µl	microlitre
mM	millimolar
µm	micromolar
mΩ	milli Ohm
mRNA	messenger RNA
MTP	minimal tiling path
ng	nanogram
NSCLC	non-small cell lung carcinoma
NUNC plates	23 by 23 cm square plates
OD	optical density
ORF	open reading frame
PAC	P1 artificial chromosome
PBS	phosphate-buffered-saline
PCR	polymerase chain reaction
PDB	protein data bank
PFGE	pulsed field gel electrophoresis
pmol	picomole

RACE	rapid amplification of cDNA ends
RFLP	restriction fragment length polymorphism
RNA	ribonucleic acid
s	second
SNP	single nucleotide polymorphism
SP	senile plaque
SRO	smallest region of overlapping deletion
SSCP	single strand conformation polymorphism
STC	sequence tagged connector
STS	sequence tagged site
$T°_{an}$	annealing temperature
TAM	transient abnormal myelopoiesis
TEMED	NN'N''N'''-tetramethylethylenediamine
TIGR	the Institute for Genomic Research
tRNA	transfer-RNA
U	unit
3' UTR	3' untranslated
5' UTR	5' untranslated
V	Volt
v/v	volume per volume
w/v	weight per volume
X-gal	5-bromo-4-chloro-3-indolyl β-D-galactopyranoside
YAC	yeast artificial chromosome
2xYT	yeast/tryptone medium

Chapter 1 Introduction

1.1 Towards the identification of all human genes

Enormous progress has been made in the field of human genetics in the last twenty years. The identification of polymorphic markers (Kan and Dozy, 1978) has made it possible to position genes genomically that are involved in inherited genetic diseases (Botstein *et al.*, 1980). This discovery made reverse genetics feasible (Monaco *et al.*, 1986): identification of a gene via its genomic location without prior knowledge of the biochemical properties of the protein it codes for. The cloning of large genomic fragments in bacterial vectors and yeast artificial chromosomes (Collins and Hohn, 1978; Burke *et al.*, 1987; Shizuya *et al.*, 1992; Ioannou *et al.*, 1994) has made it possible to create physical maps that contribute to the positional cloning of genes. These bacterial clone-based physical maps are also very important for sequencing the entire human genome, a project of which approximately 31% (including draft sequences as of 22/10/99) has been completed. The human genomic sequence will ultimately lead to the discovery of all genes, which have been predicted to number 60,000-70,000 (Fields *et al.*, 1994). Other examples of significant contributions of this project will be in determining sequences that initiate, enhance and regulate transcription and also in evolutionary studies. Another outcome of the human genome project is that a very dense genetic map can be created which, in conjunction with the knowledge of the location of all human genes, will make substantial contributions towards the identification of genes involved in complex traits.

1.2 Methods to identify novel genes

The following sections describe some of the approaches and methods that are available to contribute to the identification of novel genes.

1.2.1 Linkage studies and polymorphic markers

Linkage analysis can be used to assign a chromosomal location to an unknown gene which is involved in an inherited disease. This is performed by using polymorphic markers that distinguish between homologous chromosomes. A polymorphism that is in close physical proximity to the disease gene will be inherited in a larger proportion of affected individuals compared with a polymorphism that is either located at greater

distance on the same chromosome (due to recombination), or is present on a different chromosome. In linkage analysis, a likelihood ratio is calculated for different polymorphic markers with respect to the disease locus. This ratio is calculated as the chance that the inheritance pattern of a polymorphic marker and a disease locus could have arisen due to linkage, divided by the chance that it would have been generated if they were not linked. The $\log_{10}$ of this likelihood ratio is also called the lod score ($\log_{10}$ of the odds ratio). A lod score of 3 (linkage 1000x more likely than non-linkage) is generally accepted as a threshold of significance for linkage.

The markers that are used for linkage analysis must be polymorphic in order to follow the inheritance of different chromosomal segments. The first polymorphic markers that were detected in human DNA were restriction fragment length polymorphisms (RFLPs) at the β-globin locus (Kan and Dozy, 1978). The use of RFLPs for linkage analysis was subsequently suggested (Botstein *et al.*, 1980). This was followed by the identification of markers that consisted of a variable number of tandem repeats (Wyman and Whyte, 1980; Nakamura *et al.*, 1987).

These minisatellite markers contain tandem repeated units of 11-60 bp and they proved to be highly polymorphic. However, the majority of the minisatellite markers cluster in subtelomeric ends of chromosomes and can show high mutation rates in germline cells (reviewed by Debrauere *et al.*, 1997) and are therefore not an ideal tool for linkage analysis. The minisatellites were followed by the discovery of polymorphisms containing variable mono-, di-, tri- or tetra-nucleotide repeats (microsatellites), which are also highly polymorphic (Litt and Luty, 1989; Weber and May, 1989). Microsatellites are spaced throughout the genome (Weissenbach *et al.*, 1992) and do not show high mutation rates in meiosis (Debrauere *et al.*, 1997). Large numbers of microsatellite markers were developed from the human genome, and used to generate linkage maps. In this process, carefully selected reference families were used to order the polymorphic markers according to genetic distance (Weissenbach *et al.*, 1992; Dib *et al.*, 1996). These linkage maps can be used as a reference for the localisation of genes that are involved in mono- and polygenic diseases. Data concerning the reference families have been made available to the scientific community at the Centre d'Etude de Polymorphisme Humain (CEPH, http://www.cephb.fr/ceph_presentation.html).

Single nucleotide polymorphisms (SNPs) are present on average in the genome at approximately 1 per 1 Kbp (Wang *et al.*, 1998). Although these markers are not as

polymorphic as the microsatellites, they are abundant and have great potential for automation by using DNA chips. This technology has been successfully used to correctly analyse the majority of 512 SNPs (Wang *et al.*, 1998).

When linkage analysis has sufficiently narrowed down the region, in which a putative candidate gene is located, then a number of techniques can be used to identify the gene. A physical map can be generated using overlapping segments of genomic DNA that have been cloned in bacterial vectors. Eventually, genes can be identified from these clones using, for example, cDNA selection, exon trapping, or sequencing of the genomic clones in combination with gene prediction programmes.

1.2.2 Physical mapping using genomic libraries

Physical maps of overlapping genomic inserts, which are cloned in bacterial vectors or yeast artificial chromosomes (YACs), can be constructed in order to identify genes in a particular region of the genome.

The clones that have been used most often in the generation of physical maps are cosmids, YACs, P1 artificial chromosomes (PACs) and bacterial artificial chromosomes (BACs). Cosmids have been generated by *in vitro* packaging into bacteriophage λ, making use of the λ sequence cos (cohesive end site). The cosmids were subsequently transduced to an *E. coli* strain (Collins and Hohn, 1978; Hohn and Collins, 1980). However, the cosmid insert size of approximately 40 Kbp is relatively small. Another disadvantage is the instability of cosmids, resulting in deletions of the genomic insert in cultured *E. coli* (Fernandez, 1986; Yokobata *et al.*, 1991). Genomic libraries have been constructed in cosmids, for example after enrichment for individual chromosomes via flow sorting (Deaven *et al.*, 1986; de Jong *et al.*, 1989; Nizetic *et al.*, 1991a). A large number of very high resolution physical maps based on cosmid clones have been constructed (for example Osoegawa *et al.*, 1996)

The YAC cloning system is based on the ligation of linear DNA molecules into chromosome arms, which contain the basic functional units of yeast chromosomes (centromere, telomeres and an autonomously replicating sequence) (Burke *et al.*, 1987). The YACs are then transformed to an appropriate yeast strain and selected for an auxotrophic marker. The insert size can be up to 2 Mbp (Chumakov *et al.*, 1992), which is advantageous for covering large pieces of genomic DNA. A number of genomic YAC libraries have been constructed either by using total human genomes or individual

chromosomes (Brownstein *et al.*, 1989; Albertsen *et al.*, 1990; Larin *et al.*, 1991). YACs enriched for a particular chromosome can also be identified within a library constructed from total human genomic DNA, by means of chromosome-specific complex probe hybridisation (Ross *et al.*, 1992). However, YACs have often been found to be chimaeric or unstable (see for instance Korenberg *et al.*, 1995). A further disadvantage of this system is that individual YACs have to be isolated by pulsed field gel electrophoresis (PFGE) to separate them from the yeast chromosomes, resulting in limited yields. Due to these disadvantages, YACs are unsuitable for large scale sequencing projects. Physical maps of overlapping YACs have been constructed for chromosome 21q and for the entire chromosome 21 (Chumakov *et al.*, 1992; Nizetic *et al.*, 1994).

BAC vectors are based on the *E. Coli* F-factor plasmid. They have been shown to maintain stably 300 Kbp of human genomic DNA (Shizuya *et al.*, 1992). PAC vectors are bacteriophage P1 based, and genomic inserts up to 150 Kbp have been shown to be stable in extended culturing (Ioannou *et al.*, 1994). Both vector systems are maintained in *E. Coli*, and large quantities of DNA can therefore be readily isolated. PACs and BACs are ideal substrates for large scale genomic sequencing. A large number of physical maps based on PAC and BAC clones have been constructed (see for example Ohira *et al.*, 1996; Hubert *et al.*, 1997).

In order to generate contigs (continuum of DNA, defined by overlapping recombinant DNA clones), clones have to be identified that are located in specific areas. This can be performed by using specific DNA fragments from the area of interest, as probes. In the construction of a YAC contig that spans chromosome 21q, a large number of sequence tagged sites (STSs) were used to identify chromosome 21 YAC clones by PCR (Chumakov *et al.*, 1992). An STS is a short segment of DNA that can be detected by specific PCR amplification (Olsen *et al.*, 1989), which, ideally, is unique in the human genome. The STSs also identify overlapping genomic clones. The overlaps between clones can also be verified by alu PCR (Nelson *et al.*, 1989), which can generate multiple identical fragments between overlapping genomic clones, or by fingerprinting (Coulson *et al.*, 1986). Gaps between smaller contigs can be closed by using chromosome walking techniques. Vectorette PCR (Riley *et al.*, 1990) and inverse PCR (Ochman *et al.*, 1988) can be used to amplify fragments from the ends of the genomic inserts, which can be used as radio-active probes to identify overlapping clones.

The development of a robotic system contributed greatly to successful physical mapping using genomic clones. This system made it possible to deposit large numbers of clones in microtitre dishes in order to generate genomic libraries consisting of multiple genome equivalents coverage. It assisted in duplication of these libraries and the generation of high density filters (Nizetic *et al.*, 1991a; Nizetic and Lehrach, 1995), by spotting of the clones from multiple microtitre dishes onto nylon membranes. The membranes were incubated on broth containing agar, during which the bacteria from the wells grow as individual spots. Membranes are then processed (Nizetic *et al.*, 1991b) and the DNA crosslinked. The PAC library RPCI 1 (Ioannou *et al.*, 1994), used in this study, was spotted on six membranes, each containing over 18,000 clones, resulting in an approximately 3.3 genome equivalent coverage.

Physical maps of overlapping clones can be used for gene hunting strategies such as exon trapping and cDNA selection. Furthermore, large scale genomic sequencing can be performed using cosmids, PACs or BACs as starting materials.

1.2.3 Exon trapping and cDNA selection

Exon trapping is based on the fact that human genes can consist of multiple exons, which are interrupted by introns. The introns are removed by splicing of the primary transcript in the nucleus. The resulting mRNA, which consists of contiguous exons, is then transported to the cytoplasm. The splicing machinery recognises consensus sequences at the intron/exon borders. A splice donor is present at the 5' site, a splice acceptor at the 3'splice site.

Initially, exon trapping was described using a retroviral shuttle vector which contained one splice donor site (Duyk *et al.*, 1990). Cloning of a splice acceptor site present in genomic DNA would ultimately result in removal, via splicing, of the β-galactosidase gene. After selection, which included X-gal, white colonies were shown to contain exons. This method was subsequently improved by using an exon trapping vector (pSPL1) that contained a splice donor and a splice acceptor site (Buckler *et al.*, 1991). The strategy can be seen in Figure 4.1. Fragments of genomic DNA were introduced into the HIV *tat* intron between the two splice sites. The intron is flanked by two β-globin exons, and the system is under the control of a eukaryotic SV40 promoter. The constructs containing the genomic fragments were subsequently introduced into eukaryotic COS 7 cells. Cloning, in the correct orientation, of an exon in the HIV *tat*

intron will result in splicing between the splice sites of pSPL1 and the exon; the exon will remain part of the mRNA. Total RNA was isolated and RT-PCR was performed using primers specific for the two β-globin exons. Finally, the PCR products were cloned and analysed by sequencing. A number of improvements were made to pSPL1, to allow large scale analysis of genomic DNA (Church *et al.*, 1994). A further improvement, resulting in the exon trapping vector pSPL3B, was made after removal of a cryptic splice site from the HIV *tat* intron (Genzyme, personal communication). Exon trapping has contributed to the isolation of novel genes, including α-adducin and Huntingtin (Taylor *et al.*, 1992; The Huntington's disease collaborative research group, 1993). Also, a systematic approach has been carried out, using a large number of random chromosome 21 cosmids, to identify genes present on this chromosome (Chen *et al.*, 1996).

The main advantage of exon trapping is that low expression levels or tissue specific expression of genes will not be a problem for their identification. Some disadvantages of this technique are: genes with less than three exons will not be identified, cryptic splice sites within genomic DNA will result in background, and exon trapped products can be small, which may result in a lot of additional work to identify full length genes.

The cDNA selection method for the identification of novel genes uses clones containing human genomic DNA that are immobilised on a solid support (Lovett *et al.*, 1991; Parimoo *et al.*, 1991; Korn *et al.*, 1992). Repetitive sequences are subsequently blocked and a population of cDNA fragments containing linkers is hybridised to the immobilised DNA. The specific fragments are eluted from the target DNA after removal of non-specific cDNAs. Amplification of the cDNA clones from the final elution takes place via primers that are complementary to sequences in the linkers. Eventually, these fragments can be cloned in bacterial vectors and analysed by sequencing. cDNA selection has been used extensively as a tool for the identification of novel genes in specific regions (e.g. Peterson *et al.*, 1994; Pennacchio *et al.*, 1996; Ohira *et al.*, 1997). Strategies have also been employed that use pools of random chromosome 21 cosmids for cDNA selection in order to identify novel genes located throughout the chromosome (Cheng *et al.*, 1994; Yaspo *et al.*, 1995).

Some advantages of this technique are: the selected cDNA clones can already represent large parts of the novel gene, all genes can potentially be identified (including genes smaller than three exons), and there is no problem with cryptic splice sites.

Disadvantages are: successful identification of a novel gene depends on using a cDNA library in which the particular cDNA is present in sufficient quantities to be detected, and pseudogenes or members of gene families could select cDNA clones of genes that are not present in the target DNA by cross-hybridisation.

Apart from the specific identification of cDNA clones and exon trapped products using bacterial clones or YACs containing genomic inserts, a global genome-wide mapping of random cDNA clones from a variety of libraries has also been performed. This has been undertaken as one of the goals of TIGR (The Institute for Genomic Research, Rockville, MD, USA) and also of the I.M.A.G.E. consortium (Integrated Molecular Analysis of Genomes and their Expression).

1.2.4 Global, whole transcriptome projects of sequencing and mapping of ESTs

TIGR was founded in 1992. One of its contributions has been to sequence a large number of cDNA clones generated from different human tissues and developmental stages, and to position these into tentative human consensus sequences. This effort identified novel genes and provided information on their expression (Adams *et al.*, 1995).

The I.M.A.G.E consortium was originally initiated in 1993. One of the objectives of the consortium is to sequence clones from high quality cDNA libraries, constructed from different tissues and developmental stages, and to make this information publicly available (Lennon *et al.*, 1996). The cDNA libraries have been distributed to centres in various countries (e.g. HGMP, Hinxton Hall, Cambridge, UK) to make the resources available to the scientific community. An initiative, started by the University of Washington and the pharmaceutical company Merck, has generated a large number of single-pass sequences from human cDNA clones. These have been deposited in dbEST. In 1997, 500,000 human expressed sequence tags (EST, see next section) had been submitted (http://genome. wustl.edu/est/est_general/est_ project_ intro.html).

The 3' ends of human cDNA clones have been sequenced by the collaborators of the I.M.A.G.E. consortium. The reason for this was the observation that the 3' UTRs of genes are unique, even between gene family members, and that 3' UTR's do not normally contain introns (Wilcox *et al.*, 1991). These characteristics, in addition to the poly A tail, which can be used as a starting point for the construction of cDNA libraries, make 3' UTR's ideal for the generation of primers that can amplify a specific product

by PCR. These can then be used for the localisation of the cDNA clones to distinct areas of the genome. The amplified product is known as an EST. A gene map of the genome could then be produced by assigning a genomic location for a large number of cDNA clones. The cDNA clones from the Genbank human EST database were clustered into groups representing distinct genes (Schuler *et al.*, 1996); this resulted in the so-called "Unigene collection". This reduced the large redundancy of multiple cDNA clones that exist in Genbank which were derived from the same gene. ESTs were designed from the different Unigene cDNA clusters and these were physically localised to radiation hybrid panels (these consist of different human chromosomal DNA fragments contained in rodent cell lines, which can be used to establish the chromosomal location of an amplimer) and YAC clones. More than 16,000 human genes were mapped in this way (Schuler *et al.*, 1996). This map was subsequently improved and extended to 30,000 human genes (Deloukas *et al.*, 1998).

1.2.5 The Human Genome Project

The Human Genome Project (HGP) commenced a fifteen year project in 1990, in which the main aims were to sequence the 3 billion base pairs of the human genome and to determine the 60,000-70,000 human genes (http://www.ornl.gov/hgmis/about.html). The project was originally planned for fifteen years, but is now expected to be completed in thirteen, due to technological advances.

Some of the initial goals were to produce a dense genetic map, an STS map with 100 Kbp resolution, the development of software to collect, organise and interpret the genome sequence, and to built up an initial sequence capacity of 50 Mbp per year (Collins and Galas, 1993). These goals were mainly aimed at laying foundations to enable the large scale sequencing projects to be successful. A shift in focus towards sequencing was proposed in 1995 (Gibbs, 1995). Approximately 12 Mbp of human sequence was available at that time. Two different strategies were proposed for sequencing the human genome:

- A whole genome "shotgun" library (cloning of random fragments from the entire human genome) can be produced. This consists of long (5-20 Kbp) and short (0.4-1.2 Kbp) inserts generated from total human genomic DNA, which are cloned into bacterial vectors that are maintained in *E. coli* (Weber and Myers, 1997). Random sequencing of these clones would build up many small contigs. These shorter

contigs could be linked by sequences from the ends of the clones from the longer inserts, if for example progress of a contig is hampered by a repeat region. The intervening sequences can then be resolved specifically. A 10-fold coverage is estimated to suffice for this approach. Advantages in speed, cost and detection of DNA polymorphisms have been suggested (Weber and Myers, 1997). Arguments against this type of strategy have also been published (Green, 1997). The author of this last article questions in particular the successfulness of the finishing stage, for example with respect to the repeats which are frequent and can be large. Questions concerning the reduction in cost are raised, and this method may prove to be more expensive.

- The second approach is a clone-by-clone approach; contigs of bacterial clones containing human genomic DNA are generated, which are subsequently sequenced. This is the method currently used in the HGP. The strategy used by the largest contributors of human genomic sequence (The Sanger Centre and the Washington University Genome Sequencing Center, 1998) is as follows: Bacterial clones are identified using high density STSs. The bacterial clones are assembled into smaller contigs using fingerprinting techniques, and these are joined by the identification of novel bacterial clones using STSs generated from the ends of the smaller contigs. The sequencing phase consists of shotgun cloning of the bacterial clones into M13 vectors, followed by sequencing of the clones. The sequences are assembled in a finishing phase followed by resolving of ambiguities and closure of the remaining gaps. This approach is preferred because it makes use of the mapping information, which helps to verify the order and clone fidelity, and helps co-ordinating efforts of different groups to minimise duplication of work (The Sanger Centre and the Washington University Genome Sequencing Center, 1998). Furthermore, bacterial contigs constructed by individual investigators can be easily incorporated. The mapping information of the clones, together with the immediate release of the data after the sequence of a clone is finalised, make it an ideal substrate for targeted *in silico* identification of novel sequences. Some sequencing centres release the data already fully annotated. The group in Jena (Institute for Molecular Biotechnology, Jena, Germany) use an integrated package (RUMMAGE) of exon and gene prediction, repeat identification, and Blast searches (http://genome.imb-jena.de/). A resource that can also aid in the development of high resolution bacterial contigs is

the development of sequence-tagged connectors (STC). In this approach, ends of BAC clones are sequenced and fingerprints are generated. Sequencing of the genomic inserts of bacterial clones will identify numerous matching STCs. Using the fingerprints will validate the overlapping clones, after which clones can be chosen for sequencing that overlap minimally with the previous ones. A total of 314,000 STC have been deposited into Genbank (Mahairas *et al.*, 1999).

To date (22/10/1999), 448 Mbp (13.9%) of the human genome is ready as finished sequence (99.99% accurate), and 538 Mbp (16.8%) is ready in the form of draft sequence (http://www2.ncbi.nlm.nih.gov/genome/seq/). The final high quality sequence of the human genome is predicted to be completed in 2003, with a working draft version of 90% of the genome estimated to be available by the summer of 2000 (http:/www.ornl.gov/hgmis/project/ progress.html).

The genomic sequence which is produced can be used for the identification of novel genes by using gene prediction programmes. A number of different tools can be used by the prediction programmes (reviewed by Fickett, 1996):

- Repetitive sequences are masked after identification with for example BlastX.
- The non-repetitive sequences can then be compared with peptide sequence databases (after translation in all six reading frames).
- Exons can be predicted by using a neural network approach that analyses different parameters. An example of this is GRAIL (Uberbacher and Mural, 1991) which analyses various different parameters to predict exons.
- Comparisons with known consensus sequences can be performed, e.g. Kozak sequence, splice junctions, TATA box.

Integration of these different tools has been shown to enhance greatly the accuracy of the prediction programmes (Fickett, 1996). A large number of gene prediction programmes have been developed, and a number of these have been compared by testing their performance on a large set of genomic DNA sequences which were known to contain genes (Burset and Guigo, 1996). The correctly identified exons ranged only between 33-51%; a substantial number of false positive and false negative exons were predicted. Combining the output of the different programmes has been shown to improve the prediction of exons in genomic DNA (Burset and Guigo, 1996). Combinations of these programmes are being used to annotate genomic sequences, as was mentioned for the RUMMAGE package used at the Genomic Sequencing Centre in

Jena (Institute for Molecular Biotechnology, Jena, Germany, website-http://genome.imb-jena.de).

The techniques described in this section will lead to the identification of a number of genes which are involved in the traits that are mentioned in Chapter 1.3. Examples of this will be given in the section for human chromosome 21, which is the subject of this thesis.

1.3 Major types of genetic diseases

1.3.1 Single gene disorders

Mutations in single genes that cause genetic defects are typically inherited in an autosomal dominant, an autosomal recessive or an X-linked manner. A large number of single gene disorders follow Mendelian inheritance; some of these are described in the Online Mendelian Inheritance in Man website (OMIM, http://www.ncbi.nlm.nih.gov/omim/). However, a number of genetic disorders involving single genes that do not follow classical Mendelian inheritance have also been described, including triplet repeat expansions and genomic imprinting.

1.3.1.1 Autosomal dominant

This type of genetic defect is expressed in heterozygotes containing one copy of the mutant allele and one copy of the wild type allele. The characteristic pattern of inheritance is vertical, with the disorder passed from one generation to the next. Both males and females can be affected and both can transmit the trait with equal probability. The chance of inheriting the mutant allele is 50% in a mating of an affected heterozygote and a homozygous normal individual.

A number of characteristics have been reported that appear to be exceptions to dominant inheritance. Examples of these are (Schneider and Dawson, 1999):

- Some patients do not have affected parents, this can be due to a *de novo* mutation in the egg or sperm.
- The age of onset may be delayed; affected individuals may die of other causes before the symptoms of the genetic disease appear.
- A reduced penetrance means that only a percentage of the individuals that have inherited the mutant gene express the trait.

- Variable expressivity, in which the inheritance of the mutant gene can lead to different clinical features.

Examples of autosomal dominant disorders include familial hypercholesterolaemia, familial polyposis coli and familial early onset Alzheimer's disease (FEOAD).

1.3.1.2 Autosomal recessive

Individuals suffering from this type of genetic disease are homozygous for the mutant gene. The characteristic pattern of inheritance is horizontal: the disorder is mainly found in one generation. Males and females can be affected by the trait with equal probability. The chance of an affected offspring is 25% in the mating of two individuals that are heterozygous for the mutant allele. Additionally, 50% of the siblings would be phenotypically normal carriers of the mutated gene (heterozygote).

A number of characteristics of autosomal recessive disorders are (Schneider and Dawson, 1999):

- New mutations do occur but may pass through several generations before a descendant mates with another heterozygote, to produce affected offspring.
- The age of onset is often early in life.
- Penetrance is often 100% and expressivity is more uniform than with autosomal dominant disorders.

Examples of autosomal recessive disorders include cystic fibrosis, sickle cell anaemia and progressive myoclonus epilepsy.

1.3.1.3 X-linked

X-linked diseases are caused by mutant genes that are located on the X chromosome. Most of these traits are recessive and therefore predominantly affect males as they have a single X-chromosome. Half of the male offspring from a heterozygote mother will express the defect. All the female descendants of an affected male will be carriers but non of the male descendants will show the trait.

Examples of X-linked diseases are Duchenne muscular dystrophy and haemophilia A and B.

1.3.2 Disorders caused by chromosomal aberrations

These disorders are the result of disruptions, additions or deletions of parts of chromosomes or of entire chromosomes. Improved cytogenetic analysis of metaphase chromosomes in the late 1950's identified three disorders to be of chromosomal origin: Turner syndrome (45,X), Klinefelter syndrome (47,XXY) and Down syndrome (47, +21) (Ford *et al.*, 1959; Jacobs and Strong, 1959; Lejeune *et al.*, 1959).

Down syndrome is caused by trisomy of chromosome 21 and is characterised by a large number of phenotypic features (see Chapter 1.5). Other trisomies involving the autosomes have been observed less frequently in live births. Trisomy 18 (Edwards syndrome) has an incidence of 1 in 8000 births. The phenotypic features include mental retardation, short neck, overlapping fingers, congenital heart defects and renal malformations (Schneider and Dawson, 1999). Trisomy 13 (Patau syndrome) shows an occurrence rate of 1 in 15,000 births. Phenotypic features include polydactyly, cleft lip and palate, renal defects, cardiac defects and mental retardation (Schneider and Dawson, 1999). In contrast to trisomy 21, trisomy of chromosome 13 or 18 usually results in the death of the infant in the first year of life.

The occurrence of an extra chromosome is most often caused by nondisjunction of chromosomes in meiosis I or II (frequencies for chromosome 21 can be found in Chapter 1.5). Mitotic nondisjunction, Robertsonian translocations, isochromosomes, ring chromosomes and translocations can also result in trisomy of the whole or part of the affected chromosomes.

A number of chromosomal disorders are caused by deletions resulting in monosomy of the deleted region. DiGeorge syndrome is characterised by cardiac defects, abnormal facies, thymic hypoplasia, cleft palate and hypocalcaemia. This syndrome is caused by a deletion in chromosome 22q11 (Dallapiccola *et al.*, 1996). Cri-du-Chat syndrome is caused by a deletion of part of 5p12, with an incidence of approximately 1:45,000 (Niebuhr, 1978). The phenotypic features can include severe mental retardation, decreased muscle tone in infancy, low-set ears, renal malformations, small hands and a distinctive cry similar to the mewing of a cat (Niebuhr, 1978).

These chromosomal disorders illustrate that a large number of phenotypes can be identified in the affected individuals. The extent of these chromosomal deletions and duplications indicate that multiple genes could be involved in the formation of the different phenotypes. Furthermore, the genetic background could be important in

determining the phenotypic outcome. For example, duodenal stenosis is observed in only 4-7% of the Down syndrome cases that inherited a complete extra chromosome 21 (Korenberg, 1992).

Linkage mapping, to identify the genes underlying the different phenotypes, is not directly applicable to chromosomal disorders. However, polymorphic markers can be used to identify regions in which rare allelic variants of susceptibility genes are located in trisomies. This is described for acute non-lymphocytic leukaemia in DS in Chapter 1.8.1.3. Patients who carry partial trisomies of a chromosome can be used to link phenotypes to the inherited regions. This is described for Down syndrome in Chapter 1.7.2. Mice can be used as models for different chromosomal disorders by generating aneuploid conditions for single or multiple genes. Examples of this can be found in Chapter 1.7.1.

1.3.3 Somatic cell genetic diseases

These are diseases where changes in the DNA of somatic cells contribute to the disease phenotype. The paradigm for these diseases is cancer. Genes that are important in the progression of cancers are oncogenes and tumour suppressor genes.

Oncogenes (cancer-causing genes), are derived from proto-oncogenes, which promote normal cell growth and cell differentiation. A group of retroviruses was identified that contain additional cellular genes that could transform cells (Stehelin *et al.*, 1976; Spector *et al.*, 1978). It was shown that a human gene from bladder carcinoma that could transform mouse NIH 3T3 cell lines was similar to a gene present in one of the retroviruses; the Harvey sarcoma virus ras gene (Der *et al.*, 1982; Parada *et al.*, 1982). A number of these genes have since been implicated in a variety of human tumours and have been shown to code for important cellular proteins. Their functions include growth factors, growth factor receptors, transcription factors and proteins involved in signal transduction (Cotran *et al.*, 1999). Oncogenes have also been identified which were not detected via retroviruses. An example of this is AML1, which is located on chromosome 21 and encodes a transcription factor (Miyoshi *et al.*, 1991). A detailed description of this oncogene can be found in Chapter 1.4.1.3.

The tumour suppressor theory was originally described by Knudson after observations in sporadic and familial retinoblastoma (Knudson, 1971). He hypothesised that this type of cancer was caused by two mutational events. In the inherited form of the cancer, one

mutation is already present in the germline, the other occurs somatically. In non-hereditary retinoblastoma, two somatic mutations are required (Knudson, 1971). This hypothesis was proven by the identification of mutations in both copies of the Rb gene, which is present on chromosome 13q14 (Lee *et al.*, 1987). The Rb gene has also been shown to be involved in small-cell lung carcinoma, non-small cell lung carcinoma (NSCLC) and breast carcinoma (reviewed by Yokota and Sugimora, 1993). Another tumour suppressor gene that has been shown to be re-arranged in a variety of human cancers is p53. This gene is located on chromosome 17p13.1 and codes for a nuclear protein. Normally, DNA damage results in the rapid induction of the gene. This results in G1 cell-cycle arrest and transcriptional activation of genes that are involved in DNA repair. The p53 protein is involved in induction of apoptosis if DNA repair fails (reviewed in Cotran *et al.*, 1999). Because of the importance of this protein, p53 has been called "guardian of the genome" (Lane, 1992).

1.3.4 Complex traits

Complex traits are disorders whose phenotypes do not follow the classical Mendelian recessive or dominant inheritance attributable to a single gene locus (Lander and Schork, 1994). These disorders can be the result of additive effects of multiple genes and environmental components. Even simple Mendelian inherited gene defects could contain a complex component, for example in combination with variable expression (Lander and Schork, 1994).

Linkage analysis can be used for the identification of areas of the genome in which genes may be present that are involved in complex traits. A number of factors that can make this type of analysis difficult have been identified:

- Incomplete penetrance can lead to individuals, who have inherited a disease-causing allele, not to exhibit the phenotype. Phenocopies can give rise to individuals who do show the phenotype (for example due to environmental causes) but who do not carry the allele that causes the trait. Inheritance of a causative mutation in the BRCA1 gene, that predisposes to breast cancer, incurs a risk of 37%, 66% and 85% in women by the age of 50, 55 and 80 respectively. These risks are 0.4%, 3% and 8% for non-carriers (Easton *et al.*, 1993; Ford *et al.*, 1994).
- Polygenic inheritance, in which mutations in genes at different locations may contribute to the phenotype. Some forms of Hirschsprung disease may require the

simultaneous inheritance of mutations on chromosome 13, 21 and possibly elsewhere (Puffenberger *et al.*, 1994).

- A high frequency in the general population of alleles that cause disease, can severely influence linkage. This has been shown for the type 4 allele of the apolipoprotein E gene, which is associated with an increased risk of early and late onset Alzheimer's disease (AD) (Pericak-Vance *et al.*, 1991, Corder *et al.*, 1993).

Allele sharing methods represent a different method for identifying locations of genes in complex traits. Sib pair analysis uses affected or affected and unaffected sibling pairs to determine the inheritance of chromosomal regions. It identifies areas, which are inherited more often than expected by chance, in which genes that are involved in a phenotype might be located. This type of analysis has identified a number of potential locations for genes involved in late onset Alzheimer's disease (LOAD) (Kehoe *et al.*, 1999).

A number of techniques are available to localise and identify genes that are involved in the aforementioned diseases. Disease genes that are inherited in a Mendelian fashion can be localised to chromosomal regions by linkage analysis. The causative genes can then be identified from overlapping genomic clones that are maintained in bacteria or yeast. Sequencing of the human genome via these overlapping clones will eventually reveal all genes. The identification of genes involved in complex traits is likely to increase due to the identification of possible locations of susceptibility genes, combined with the detection of large numbers of genes via the genomic sequencing project.

1.4 Genetic diseases mapping to chromosome 21

1.4.1 Monogenic/somatic diseases

Causative genes identified for a number of monogenic and somatic diseases, are located on chromosome 21. Some examples of these genes are described in this section. The approximate positions of these genes can be seen in Figure 1.1.

1.4.1.1 SOD1 in motor neuron disease

Amyotrophic lateral sclerosis (ALS)/motor neuron disease is a relatively common neurodegenerative disorder, with an incidence of approximately 1 per 20000 people (De Belleroche *et al.*, 1995). The disease is characterised by the loss of motor neurons in the spinal cord, brain stem and motor cortex. The mean age of onset is 50 years, with death

Cytogenetic map of chromosome 21 and positions of disease-causing genes and the "Down syndrome chromosome region"

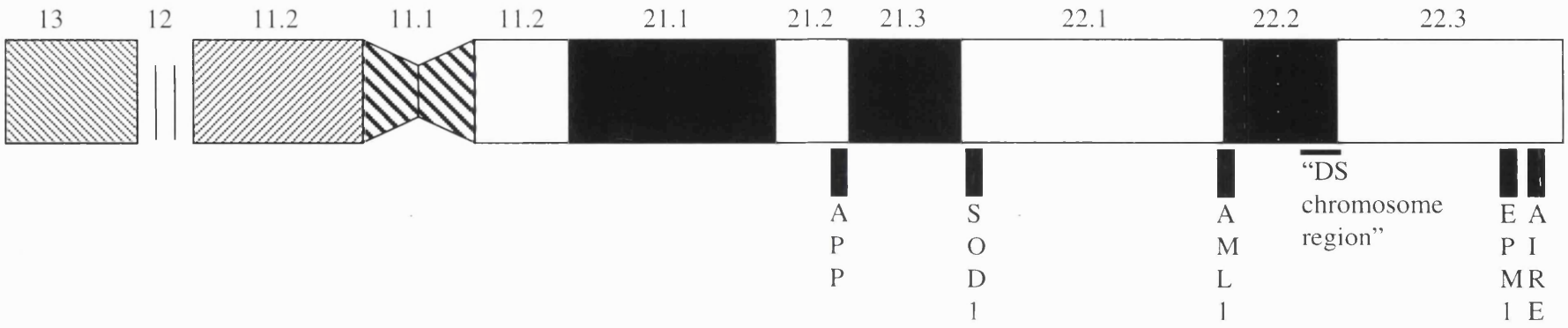


Figure 1.1: The cytogenetic map is shown of Chromosome 21. The approximate positions are indicated of the five genes described in sections 1.4.1.1-1.4.1.5; APP (familial early onset Alzheimer's disease), SOD1 (motor neuron disease), AML1 (acute myeloid leukaemia), EPM1 (epilepsy) and AIRE (APECED). The region which used to be known as the "Down syndrome chromosome region", described in Chapter 1.7.2, is also shown.

occurring in 50% of the cases within the first three years (reviewed by De Belleroche *et al.*, 1995). Death often occurs through respiratory failure due to muscle weakness. Approximately 90% of the cases are sporadic, 10% are familial (Siddique and Deng, 1996). In approximately 20% of the familial cases, mutations in the superoxide dismutase gene (SOD1) have been found (De Belleroche *et al.*, 1995). Initially linkage was established between familial ALS and chromosome 21q22.1 (Siddique *et al.*, 1991). Subsequently, different mutations were detected by single strand conformation polymorphism in the already characterised SOD1 (Rosen *et al.*, 1993; Deng *et al.*, 1993). The enzyme *SOD1* detoxifies the highly reactive superoxide (O_2^-) into O_2 and H_2O_2. How SOD1 mutations cause ALS is not known, but may be due to a decrease of *SOD1* activity and a subsequent increase in O_2^-, or the gain of a toxic function (Siddique and Deng, 1996).

1.4.1.2 APP in familial early onset Alzheimer's disease

One of the characteristics of AD is depositions of extracellular βA4 amyloid peptide (42-43 amino acids) in the brain, known as amyloid plaques. These can also be found in the brains of 30-40 year old patients with Down syndrome (Glenner and Wong, 1984a). The βA4 amyloid, a proteolytic product of the amyloid precursor protein (APP), was initially identified from cerebral blood vessels from AD and Down syndrome patients (Glenner and Wong, 1984 a, b). The full length APP gene was subsequently identified and mapped to chromosome 21 (Kang *et al.*, 1987). The gene was further localised to the cytogenetic band 21q21.2. These authors predicted a protein with one transmembrane domain. A number of functions have been proposed for APP; it may possess cell growth promoting activities, or neuritic outgrowth enhancing capability, it might protect neuronal cells by lowering intraneuronal Ca^{2+} levels or the protein might be involved in wound repair (reviewed by Hendriks and Van Broeckhoven, 1996).

The APP protein is processed by putative β- and γ-secretase, in order to generate βA4 amyloid peptide (Seubert *et al.*, 1993). The β-secretase site is located in the extracellular portion and γ-secretase is present in the transmembrane domain. The β- and γ-secretase enzymes have so far not been identified.

Goate and co-workers showed linkage of a family with familial AD and APP and discovered a causative mutation within the βA4 amyloid part of the APP gene (Goate *et al.*, 1991). Additional mutations have since been detected (e.g. Murrell *et al.*, 1991;

Mullan *et al.*, 1992). The APP mutations account for approximately 5% of the FEOAD cases. Transfections were performed, to neuronal and non-neuronal cells, of wild type APP and cDNAs of the APP gene containing different mutations. These showed either an overproduction of βA4 amyloid or an increased ratio of the more amylogenic 42/43 amino acid βA4 amyloid compared with the 40 amino acid fragment (reviewed by Hendriks and Van Broeckhoven, 1996).

Transgenic mice that overexpress the APP gene at a low rate have failed to show the AD phenotype (Lamb *et al.*, 1993; Pearson and Choi, 1993). A number of mutants have been reported that do exhibit some of the AD characteristics; these animals show a 5-10 fold overexpression of the APP gene in combination with mutations that are causative for FEOAD cases in humans (Games *et al.*, 1995; Sturchler-Pierrat *et al.*, 1997).

1.4.1.3 AML1 (CBFA2) in leukaemia

AML1 (Core binding factor 2, CBFA2) was identified at chromosome 21q22, at a translocation breakpoint which is frequently observed in acute myeloid leukaemia (Miyoshi *et al.*, 1991). The translocation that was used for identification of the gene was the t(8;21)(q22;q22) (involving the AML1 and ETO genes), which is involved in acute myeloid leukaemia of the M2 immunotype. AML1 is also involved in a number of translocations involved in different types of leukaemia. These are t(12;21)(p13;q22) and t(3;21)(q26;q22) (reviewed by Speck *et al.*, 1999). AML1 is a transcription factor, which forms a complex with the non-DNA binding CBFβ, which increases the binding affinity of AML1 for DNA (Wang *et al.*, 1993). Hiebert and co-workers have shown that although the AML1-TEL (t(12;21)(p13;q22)) is still able to bind both the CBFβ subunit and DNA, transcriptional activation via a model enhancer is significantly reduced (Hiebert *et al.*, 1996). The authors believe that the fusion protein acts in a transdominant manner. A homozygous knockout mouse was generated for the AML1 homologue. The mouse embryos showed haemorrhages and all died before embryonic day 12.5. Furthermore, progenitor cells that form the different haematopoietic cells were not observed (Wang *et al.*, 1996). A mouse knock-in, which was heterozygous for the AML1-ETO fusion gene was subsequently constructed. These embryos died before the embryonic age of 13.5 days, from haemorrhages associated with the central nervous system and also showed severe impairment in foetal liver haematopoiesis (Yergeau *et al.*, 1997). The authors conclude that the fusion protein disrupts the function of the

normal copy in a transdominant manner as the phenotypes of the -/- mice from the study by Wang and co-workers (Wang *et al.*, 1996) were very similar. This indicated that AML1 could be required for the development of specific haematopoietic stem cells in the developing embryo (Speck *et al.*, 1999). There is a clear difference between the involvement of AML1 in early haematopoiesis in mice and the leukaemias in a later stage of haematopoiesis in humans. The leukaemias that arise from translocations in a specific haematopoietic lineage could be due to these alterations and other mutations that take place in cells that are already committed to a specific stage in the haematopoietic pathway (Speck *et al.*, 1999).

1.4.1.4 The EPM1 gene in progressive myoclonus epilepsy

Progressive myoclonus epilepsy type 1 (Unverricht-Lundborg disease) is an autosomal recessive epilepsy, which is characterised by an age of onset of between 6 and 13 years (Berkovic *et al.*, 1986). EPM1 is a member of the heterogeneous group of epilepsys whose phenotypes include myoclonic seizures, progressive neurological degeneration and dementia (Berkovic *et al.*, 1986). Linkage was established between the EPM1 locus in Finnish families and chromosome 21q22.3 (Leheshjoki *et al.*, 1991), near the markers D21S25 and D21S154 (Leheshjoki *et al.*, 1993). A bacterial clone contig and a detailed restriction map were developed (Stone *et al.*, 1996). The gene that causes EPM1 was subsequently identified via cDNA selection (Pennacchio *et al.*, 1996). The protein, cystatin B is a member of the cysteine protease inhibitors. Lalioti and co-workers identified a dodecamer repeat expansion 70 nucleotides upstream of the transcription start site to be the common mutation mechanism (Lalioti *et al.*, 1997). Patients were identified who carried two alleles with an expanded repeat (>60 copies) or who carried one allele with an expanded repeat and one allele that showed a normal repeat (2-3 copies) but which had a mutation in the cystatin B gene (Lalioti *et al.*, 1997).

1.4.1.5 AIRE in autoimmune-polyendocrinopathy-candidiasis-ectodermal dystrophy

APECED (autoimmune-polyendocrinopathy-candidiasis-ectodermal dystrophy) is an autoimmune disorder which is inherited in an autosomal recessive manner. It leads to hypoparathyroidism, adrenocortical failure and chronic mucocutaneous candidiasis and a variety of other phenotypes, and is prevalent in the Finnish population (1:25,000)

(Ahonen *et al.*, 1990). The APECED locus was mapped to chromosome 21q22.3 by linkage analysis (Aaltonen *et al.*, 1994). This was further refined by subsequent linkage studies to a region of less than 500 Kbp (Björses *et al.*, 1996). A high-resolution physical map consisting of PAC/BAC and cosmid clones and haplotype analysis narrowed the region down to 350 Kbp, around the PFKL gene (Aaltonen *et al.*, 1997). The gene was eventually identified by sequencing of genomic clones and mutations were detected in patients (Nagamine *et al.*, 1997; The Finnish-German APECED Consortium, 1997). The gene (AIRE: autoimmune regulator) consists of a 545 amino acid open reading frame and contains two PHD zinc-finger motifs (DNA binding) and a proline-rich region (protein-protein interactions/DNA binding). The protein could therefore be involved in regulation of transcription (Nagamine *et al.*, 1997; The Finnish-German APECED Consortium, 1997).

1.5 Down syndrome: a trisomy of chromosome 21

The clinical definition of Down syndrome (DS) was established by John Langdon Down (Down, 1866). It was determined in the late 1950's to be caused by trisomy of chromosome 21 (Lejeune *et al.*, 1959; Jacobs *et al.*, 1959).

1.5.1 Phenotypic features of Down syndrome

DS is the most common genetic birth defect in man, the incidence being around 1:700 live births (Epstein, 1995). In the majority of cases (90%), an extra chromosome 21 is generated by nondisjunction (chromosomes do not separate in meiosis I or meiosis II) in egg or sperm (Antonarakis, 1998). In approximately 5% of cases, an early mitotic nondisjunction in the foetus causes a mosaic individual (Antonarakis *et al.*, 1993). The remaining 5% of cases are caused by isodisomy of chromosome 21, Robertsonian translocations (mainly t(21;21) and t(14;21)) and partial trisomies (very rare). Section 1.5.2 describes current models that could explain the mechanisms of nondisjunction.

A large number of phenotypic features can be seen in people with DS, which are not all expressed in each individual with DS. Mental retardation and neonatal hypotonia are seen in nearly all cases of DS (Korenberg *et al.*, 1994). A number of physical features of DS also show a high incidence, for example: upslanting eyes (82%), loose skin on the nape of the neck (81%), narrow palate (76%), flat nasal bridge (68%), abnormal teeth (61%), open mouth (58%) and protruding tongue (47%) (Epstein, 1986). An

Alzheimer's disease-like neurohistopathology can be seen in virtually all people with DS in the third and fourth decades (Kesslak *et al.*, 1994). Congenital heart diseases are prevalent in 40-50% of DS individuals, which account for approximately 70% of all endocardial cushion defects (Korenberg *et al.*, 1992). Duodenal stenosis is observed in 4-7% of DS patients, which translates to 30-50% of all duodenal stenosis cases in man. DS children are 10-30 times more likely to develop acute leukaemia (1-2% of DS individuals) than non-DS individuals (reviewed in Zipurski *et al.*, 1992). The acute non-lymphocytic leukaemia of the M7 immunotype (ANLL-M7, or megakaryoblastic leukaemia) has a particularly high incidence in DS individuals compared with karyotypically normal children (Zipurski *et al.*, 1987). The challenges of molecular and cellular biology are to identify all the genes of chromosome 21 and to relate the DS phenotypes to the areas of chromosome 21 that are likely to contain the genes responsible for the above mentioned specific phenotypes. Eventually this will lead to the detection of the genes and their roles in the pathogenesis of the different phenotypes of DS. This may lead to a better understanding of mechanisms underlying disorders also in the euploid (non DS) population. Examples of these disorders are: mental retardation, AD, heart defects and childhood leukaemia. Progress towards the identification of the chromosome 21 genes can be found in section 1.6. Genotype/phenotype mapping of different traits of DS are described in section 1.7.

1.5.2 Analysis and models for nondisjunction of chromosome 21

Polymorphic markers are used to establish the parental origin and the meiotic stage of nondisjunction (Antonarakis *et al.*, 1992, Antonarakis *et al.*, 1993, Sherman *et al.*, 1994, Lamb *et al.*, 1996, Yoon *et al.*, 1996). Nondisjunction at meiosis I will result in a trisomic individual with two different copies of chromosome 21 (proximal to a cross-over) from one parent, and one chromosome 21 from the other parent. This is known as a meiosis I error (MI-Maternal meiosis I error, PI-Paternal meiosis I error) and can be detected via heterozygosity of the markers of the nondisjoined chromosomes, close to the centromere. A meiosis II error (MII-Maternal meiosis II error, PII-Paternal meiosis II error) will result in an individual with trisomy 21 which contains two identical copies of chromosome 21 from one parent (proximal to a cross-over), and one chromosome 21 from the other parent. This can be detected by homozygosity of the markers of the nondisjoined chromosomes close to the centromere. Cross-overs can lead to a reduction

to homozygosity distal to the cross-over in a MI/PI nondisjunction, whereas heterozygosity is observed distal to a cross-over in MII/PII nondisjunction. Mitotic errors can be identified by an absence of recombination, resulting in homozygosity between all the markers along the nondisjoined chromosomes. The above mentioned studies (Antonarakis *et al.*, 1992, Antonarakis *et al.*, 1993, Sherman *et al.*, 1994, Lamb *et al.*, 1996, Yoon *et al.*, 1996) showed the origin of nondisjunction to be maternal in approximately 88% of the cases, 8% were paternal in origin and 5% mitotic. Of the maternally derived nondisjunctions, approximately 75% of cases were MI in origin, and 25% MII. The paternal cases showed approximately 40% PI errors and 60% PII nondisjunction (Antonarakis, 1998). So far only increasing maternal age has been associated with increased nondisjunction (Hassold *et al.*, 1984), and is associated with both MI and MII nondisjunctions (Antonarakis *et al.*, 1992, Yoon *et al.*, 1996).

While studying the mechanisms of nondisjunction, Sherman and co-workers showed the genetic map to be decreased in length in MI nondisjunction (Sherman *et al.*, 1991). Subsequently it was shown that this decrease was approximately 50% and that this was due to a decrease in proximal recombination (Sherman *et al.*, 1994). This study also showed a difference between the genetic map of older and younger women, the oldest women showing the greatest reduction. In a similar study, MII nondisjunction showed an increase in recombination, especially in the proximal half of the long arm of chromosome 21. No difference in length of the genetic map was observed in older compared with younger women (Lamb *et al.*, 1996). From these observations the following hypotheses have been suggested:

- In MI attachment of chromosome 21 to the spindle may be lost because distal exchanges increase the likelihood of premature separation or malalignment of the bivalents on the metaphase plate. This would cause the bivalents to segregate randomly to the poles. If they appear at the same pole, then the outcome would be a meiosis I nondisjunction (Hawley *et al.*, 1994).
- The increase of exchanges in MII has led to the hypothesis of chromosome entanglement, in which proximal chiasmata cause the bivalents not to separate at MI, and therefore to travel to the same pole. Separation of the bivalents would then result in MII errors (Koehler *et al.*, 1996; Lamb *et al.*, 1996). Another explanation would be that anaphase I disrupts sister chromatid cohesion, which causes them to move randomly in the second meiotic division (Koehler *et al.*, 1996).

The hypotheses predict a two step model; the first step being the positioning of the exchanges which occur prenatally, the second step being the biological functions that act to separate the chromosomes. This second step would be age related (Sherman *et al.*, 1994; Lamb *et al.*, 1996).

1.6 Gene density and resources for identification/mapping of genes on chromosome 21

It has been estimated that chromosome 21 is likely to contain 560-800 genes (Antonarakis, 1998). Eighty-six full length genes have so far been identified, which can be found on the following website: http://www.expasy.ch/cgi-bin/lists?humchr21.txt. The majority of the genes have been located in the distal part of the long arm (Gardiner, 1996).

1.6.1 Gene density on chromosome 21

A compositional map has been constructed for the long arm of chromosome 21, in which the GC content of a landmark (for instance an STS or a gene), was assessed by hybridisation to a Southern blot containing different fractions from a $CsSO_4$ gradient centrifugation. The gradient separates the DNA fragments according to their GC content. The landmarks, which are linked to a physical map, show which regions of the chromosome are GC rich and therefore gene rich. This showed that the cytogenetic bands 21q11.2 and 21q21 were lowest in GC content, whereas the telomeric R band (21q22.3) had the highest GC percentage (Gardiner *et al.*, 1990).

Further insight into the gene content of different regions of chromosome 21 is provided by analysis of mapping positions of cDNA selected clones, exon trapped material and genes (Reviewed by Gardiner, 1996). The cDNA and exon trapped clones were identified by using random, un-mapped cosmids from a flow-sorted chromosome 21 cosmid library (Cheng *et al.*, 1994; Yaspo *et al.*, 1995). This analysis showed that only 14% of the cDNA selected and exon trapped clones mapped back to 21q11.2 and 21q21, whereas these regions contain approximately 50% of the long arm of chromosome 21. Approximately 7% of the genes which mapped to chromosome 21 were localised to the above mentioned region, which was mirrored by biochemically characterised genes (these were mapped after identification rather than identified in a specific region). These different lines of evidence point to an uneven distribution of

genes along the long arm of chromosome 21, with 21q11-21q21 being potentially gene-poor.

1.6.2 Genetic and physical maps

Genetic and physical maps play an important role in the discovery of genes and the subsequent correlation of the location of these genes with the position of DS phenotypes. A number of resources have been generated to aid in the discovery and mapping of genes on chromosome 21:

- An early, detailed genetic map has been produced (McInnes *et al.*, 1993) and a number of genes that are involved in genetic diseases have been mapped to chromosome 21 (Leheshjoki *et al.*, 1991; Leheshjoki *et al.*, 1993; Aaltonen *et al.*, 1994). The genetic map is also extremely important in determination of the parental origin and state of nondisjunction (section 1.5).
- A panel of somatic cell hybrids is available which contains the entire or partial fragments of chromosome 21 maintained on a rodent background (reviewed by Graw *et al.*, 1995). The panel of partial chromosome 21 fragments divides the chromosome into segments. Hybridisation of radio-active probes to Southern blots containing restriction digested DNA from these cell lines provides information on the mapping position of the probe. The somatic cell hybrids have been used to map STSs (Van Camp *et al.*, 1990), cDNA clones (Yu *et al.*, 1997) and genes (Brodsky *et al.*, 1995).
- A *Not* I map of the long arm of chromosome 21 has been constructed (Ichikawa *et al.*, 1993). This physical map gave an estimate of the length of the long arm of chromosome 21 (38 Mbp), contributed a great number of STSs around the *Not* I sites and provided a scaffold of STSs for the mapping of bacterial clones. It also aids in the identification of novel genes; *Not* I sites contain CpG nucleotides within their recognition site. These nucleotides are clustered in CpG islands, which can often be found at the 5' end of housekeeping genes (Lindsay and Bird, 1987).
- A physical map was constructed of overlapping YAC clones, which encompassed the whole long arm of chromosome 21 (Chumakov *et al.*, 1992). This map was subsequently improved (Korenberg *et al.*, 1995; Gardiner *et al.*, 1995). Another study showed an additional YAC map of overlapping clones of both arms of the chromosome (Nizetic *et al.*, 1994). A cosmid pocket map consisting of groups of

cosmids that were assigned to specific regions was constructed by hybridisation of a subset of these YACs to membranes containing DNA of a flow-sorted chromosome 21 cosmid library (Nizetic *et al.*, 1994). The YAC maps have been used for the localisation of cDNA clones (Cheng *et al.*, 1994) and for the identification of novel genes (Tassone *et al.*, 1995; Xu *et al.*, 1995). Furthermore, they can provide an important stepping stone in the generation of high resolution bacterial contigs.

- A number of bacterial contigs consisting of PAC, BAC and cosmid clones have been constructed in selected regions of chromosome 21 (Patil *et al.*, 1994; Lafreniere *et al.*, 1995; Eki *et al.*, 1996; Osoegawa *et al.*, 1996; Ohira *et al.*, 1996; Stone *et al.*, 1996; Hubert *et al.*, 1997; Hildmann *et al.*, 1999). These contigs have been used to identify novel genes (Ohira *et al.*, 1997) and are being used in large scale sequencing projects of chromosome 21.

1.6.3 Sequencing of chromosome 21

A consortium of different research groups is sequencing chromosome 21, which is estimated to be finished by the end of 1999. The main contributors to the sequence are: Stanford Human Genome Center (Palo Alto, California, USA), Keio University (Tokyo, Japan), University of Tokyo (Tokyo, Japan), Max Planck Institute (Berlin, Germany), Genome Analysis Centre (Braunschweig, Germany), Institute for Molecular Biotechnology (Jena, Germany). The sequences are being collated at the Eleanor Roosevelt Institute (Denver, Colorado, USA) (see http://www-eri.uchsc.edu/chr21/c21index.html), in June 1999, 168 clones and a total of 22.968 Mbp had been deposited. A map of the chromosomal regions that are being sequenced by the different groups can be found at http://chr21.rz-berlin.mpg.de/map1.html. A great number of clones have already been placed in publicly available databases.

This work will greatly enhance the number of chromosome 21 genes that will be discovered in the near future. For example, the sequencing group in Jena subject their sequences to a great number of exon and gene prediction programmes before release on their website. These are then displayed in graphical form (these can be found by following the chromosome 21 link from http://genome.imb-jena.de and selecting individual projects).

1.7 Genotype/phenotype mapping of Down syndrome phenotypes

1.7.1 Mouse models

Mouse chromosome 16 contains a region that is syntenic to part of human chromosome 21, spanning from the gene STCH (21q11) to MX1 (21q22.3) (Reeves *et al.*, 1998). Syntenic regions distal to MX1 of human chromosome 21 have been found on mouse chromosomes 10 and 17 (Carver and Stubbs, 1997). In addition to the synteny to human chromosome 21, mouse chromosome 16 also shows regions that contain similar genes to those on human chromosomes 16, 8, 22 and 3 (Carver and Stubbs, 1997). Mice could theoretically be used as a model for DS if an extra copy is present of the part of chromosome 16 that is syntenic to human chromosome 21. Although it is not believed that the mouse model could exactly mimic DS, it provides a potential way of identifying the genes that are involved in the DS pathologies (Hernandez and Fisher, 1999). A number of different approaches have been described for generating mice that are trisomic for some of the human chromosome 21 genes.

Three partial trisomies used as models for DS are Ts16 (Epstein *et al.*, 1985), Ts65Dn (Davisson *et al.*, 1993) and Ts1Cje (Sago *et al.*, 1998). Ts16 mice are trisomic for mouse chromosome 16 and are thus not trisomic for all the genes located on human chromosome 21 (see previous paragraph). Additionally, they are also in dosage imbalance with genes which are not located on human chromosome 21. Ts16 mice die *in utero*, and many types of analysis cannot be carried out. Ts65Dn mice contain a marker chromosome consisting of the centromere of mouse chromosome 17 and a large portion of mouse chromosome 16, which shows synteny to human chromosome 21. This mouse is trisomic for the segment from the APP (21q21) to the MX1 gene (21q22.3)(Reeves *et al.*, 1995). These mice exhibit a number of symptoms including mild developmental delay, hyperactivity in early development and slightly impaired learning in adult mice (Reeves *et al.*, 1995; Holzman *et al.*, 1996). Although the APP gene showed a 2.2-fold increase in expression in these mice, no senile plaques or neurofibrillar tangles were observed (Reeves *et al.*, 1995). The Ts1Cje mice contain a translocation of distal mouse chromosome 16 (the position where part of human chromosome 21 is located) onto mouse chromosome 12. The trisomy consists of a smaller region than in Ts65Dn, from SOD1 (21q22.1) to MX1 (21q22.3). These mice show impaired learning and decreased locomotor activity (Sago *et al.*, 1998).

Much smaller trisomic regions can be created with the introduction of YACs and PACs. Smith and co-workers introduced human chromosome 21 YACs and PACs from the region around D21S55, by zygote micro-injection. Analysis of the expression of genes present on the YACs in different tissues, showed comparable results to the expression in human tissues (Smith *et al.*, 1995). Analysis of the transgenic animals showed that mice containing the distal part of YAC 152F7 were impaired in learning tests and exhibited decreased locomotor activity (Smith *et al.*, 1997). The distal part of this YAC contains the human homologue of the *Drosophila melanogaster* minibrain gene, which is a transcription factor involved in neurogenesis.

Recently, mice have been generated which contain the whole or part of human chromosome 21 (Hernandez *et al.*, 1999). This was performed by microsome-mediated chromosome transfer. Homologous recombination was carried out whereby the neomycin resistance gene was introduced in a locus near D21S55. Microcells were isolated, irradiated to introduce chromosome breaks which would create deletions in chromosome 21, and fused with mouse embryonic stem (ES) cells. Selection for neomycin resistance was carried out, after which the different clones could be analysed for the human DNA that was retained in the ES cells. Eventually, chimeric mice were created from four ES cell lines, which contained different segments of human chromosome 21, and which were subsequently used to generate trisomic transgenic mice (Hernandez *et al.*, 1999). The mice showed expression of most of the genes that were tested in the appropriate tissues. One of the mouse chimeras showed morphological differences. These experiments showed that it is possible to create mice that are trisomic for different regions of human chromosome 21 and which can be analysed for the different DS phenotypes. This approach could potentially lead to the identification of chromosomal regions in which genes are located that contribute to certain DS phenotypes. Also a genome-wide search can be performed for genes that are involved in the different phenotypes, by placing the partial trisomies on different mouse genetic backgrounds (Hernandez *et al.*, 1999).

1.7.2 Partial trisomy 21 in humans

Genotype/phenotype mapping uses gathered information from patients with rare partial trisomies of chromosome 21 and the phenotypes they exhibit, in order to assign specific areas of chromosome 21 to the gene(s) that are involved in the particular disorders.

These studies were initially performed by studying the karyotype of partial trisomic individuals. They indicated that the 21q22 region, which showed a subset of the Down's phenotypes, including facial features and mental retardation (Niebuhr, 1974; Williams *et al.*, 1975), was important. This region was therefore known as "the Down syndrome chromosome region" (see Figure 1.1 for the approximate location on chromosome 21).

With the identification of chromosome 21 specific markers that could be used for the generation of radio-actively labelled probes, the extent of the partial trisomies could be determined at the molecular level. A careful set of protocols was designed to establish accurately the copy number of the different markers in combination with a protocol for recording the phenotypic features of the patients with partial trisomy of chromosome 21 (Epstein *et al.*, 1991).

The studies that were performed using molecular analysis mapped a substantial number of DS phenotypes to distal 21q, in particular to the region around the marker D21S55 ("Down syndrome chromosome region") (McCormick *et al.*, 1989; Rahmani *et al.*, 1989; Delabar *et al.*, 1993). These studies also indicated that very few of the DS phenotypes would be found in proximal 21q (21q11-21q21). However, Korenberg and co-workers showed that a contiguous gene syndrome, in which genes located in different areas of the chromosome may be responsible for a specific phenotype (Korenberg *et al.*, 1994). This may be a better model for DS. A gene or genes could be present in an individual region if the percentage of cases in partially trisomic DS patients that exclude the candidate region, is the same or lower as the general population. This appears to be true for congenital heart defects (Korenberg *et al.*, 1994). The contiguous gene syndrome takes into account the penetrance and expressivity of the genes involved. The involvement of a single gene or gene cluster could be determined by the overlap between the different partial trisomies showing the phenotype. However, if more genes are involved in a particular trait, they might be located in the non-overlapping regions (Korenberg *et al.*, 1994). The term "Down syndrome chromosome region" was subsequently dropped (Korenberg *et al.*, 1996).

The whole of chromosome 21 would therefore be relevant to the study of the involvement of the different genes in DS. A region on the long arm of chromosome 21 (21q11) has been associated with a number of DS and non-DS related phenotypes and will be the subject of this study.

1.8 Human chromosome 21q11

The chromosomal region 21q11 is associated with transient abnormal myelopoiesis (TAM), acute non-lymphocytic leukaemia in DS (Shen *et al.*, 1995, Cavani *et al.*, 1998), FEOAD (St. George-Hyslop *et al.*, 1987, Goate *et al.*, 1989) and squamous NSCLC (Kohno *et al.*, 1998). The region is estimated to be gene-poor (Gardiner, 1996) and to date only two genes have been identified in this 3.5 Mbp region.

1.8.1 Phenotypes associated with chromosome 21q11

1.8.1.1 Lung cancer

Lung cancer is the most common neoplastic cause of death among men and women in the USA. A study in mortality rates due to neoplasias in the USA, conducted from 1987-1991, showed that of the 172.8 people that die from cancers per 100,000 of the total population, 49.3 were due to lung cancer (SEER Cancer Statistics Review 1973-1991, http://www.meds.com/lung/seer.html#quest_14).

Induced expression of different oncogenes and loss of tumour suppressor genes have been reported for known genes in lung cancer. These include the oncogenes RAS, MYC, c-erbB-2 and BCL-2, and the tumour suppressor genes Rb and P53 (reviewed in Salgia and Skarin, 1998). Additional tumour suppressor genes are likely to be involved in the development and progression of lung cancer, as indicated by loss of heterozygosity (LOH) and homozygous deletions in tumour DNA. Testa and co-workers performed a cytogenetic analysis of 63 NSCLCs and found frequent loss (at least 60%) of the following chromosomal arms: 3p (94% in squamous cell carcinomas), 6q, 8p, 9q, 13q, 17p, 18q, 19p, 21q and 22q (Testa *et al.*, 1994). A different study found the most frequent deletions in the following chromosomal arms in squamous cell carcinoma: 3p (81%), 9p (55%), 9q (67%), 13q (60%), 17p (88%) and 21q (50%) (Sato *et al.*, 1994).

A very frequent LOH has been detected on chromosome 3p, with the smallest region of overlapping deletion (SRO) being detected at the cytogenetic band 3p21 (Testa *et al.*, 1994). Additional studies have reported hemizygous and homozygous deletions in 3p21 (Kok *et al.*, 1987, Rabbitts *et al.*, 1990). Kohno and co-workers analysed the frequent deletions of the long arm of chromosome 21 in NSCLC and found 62.5% LOH of the 21q11 marker D21S120, and 42% LOH with two 21q21 markers (Kohno *et al.*, 1998). A NSCLC cell line, Ma17, contained a homozygous deletion of the 21q11-21q21

region, including the markers D21S318 (21q11) and LL136 (21q21.1). The size of this deletion was estimated to be larger than 3 Mbp (Kohno *et al.*, 1998). ANA (Abundant in Neuroepithelial Area, a member of the *tob/BTG* family of antiproliferative genes), a gene mapping to this interval, was discounted as a tumour suppressor gene in squamous NSCLC, after single strand conformation polymorphism analysis, with tumour samples that showed LOH of the area (Kohno *et al.*, 1998).

1.8.1.2 Alzheimer's disease

AD is a neurodegenerative disorder of the central nervous system, which was first described by Aloïs Alzheimer in 1907 (Alzheimer, 1907). The symptoms of the disease are progressive memory impairment and disorientation, which leads to dementia and death. Although the disease can be diagnosed clinically, a definite diagnosis can only be obtained after a post-mortem examination of the brain. The two brain lesions that are diagnostic for AD, when found together, are neurofibrillary tangles (NFTs) and senile plaques (SPs). The NFTs, which are intra-neuronal inclusions of paired helical fragments, are composed of the microtubule-associated protein tau (Goedert *et al.*, 1988) and ubiquitin (Mori *et al.*, 1987; Perry *et al.*, 1987; Morishima-Kawashima *et al.*, 1993). The SPs are extracellular deposits which consist mainly of βA4 amyloid peptide (Kang *et al.*, 1987), which is a proteolytic product (40-43 amino acids) of the much larger βA4 amyloid precursor protein (APP).

A distinction is made between patients with onset of the disease under and over the age of 65 (early onset [EOAD] and late onset [LOAD] Alzheimer's disease, respectively). The disease can either be inherited in an autosomal dominant fashion, or can be sporadic. The largest group of patients are the sporadic LOAD patients (45%), followed by the familial LOAD patients (30%), the sporadic EOAD patients (15%) and the familial EOAD patients (10%) (see Van Broeckhoven 1995 for a review).

Linkage analysis of polymorphic markers have shown linkage of AD with the marker D21S16 from the 21q11 region in some families with EOAD (St. George-Hyslop *et al.*, 1987, Goate *et al.*, 1989). Subsequent linkage analyses have shown much stronger linkage to markers on chromosome 14q24 (Van Broeckhoven *et al.*, 1992, St. George-Hyslop *et al.*, 1992) in the same families. This resulted eventually in the identification of the gene presenilin 1 (PS-1), located at chromosome 14q24.3, which showed mutations that segregated with the disease (Sherrington *et al.*, 1995). The PS-1 protein

resembles a transmembrane protein with eight transmembrane domains. Additional linkage analysis in some families which excluded 14q24, showed linkage with chromosome 1q31-q42 and EOAD (Levy-Lahad *et al.*, 1995a). PS-2, a gene with homology to PS-1, was mapped within the chromosome 1 candidate region and mutations were detected (Levy-Lahad *et al.*, 1995b, Rogaev *et al.*, 1995).

It has been hypothesised that linkage found in some families with EOAD to the 21q11 region and mutations in PS-1 could be explained by a modifier gene being located at 21q11 that generates or modifies the AD phenotype in conjunction with PS-1 (St George-Hyslop *et al.*, 1992, Van Broeckhoven 1995).

Apart from the presenilin genes, only the APP and apolipoprotein E (APOE) genes have so far been directly implicated in AD. The presenilins account for most of the FEOAD cases (Hutton and Hardy, 1997), with APP accounting for less than 5%. The ε4 allele of the APOE gene, which is a plasma protein involved in cholesterol transport, is associated with an increased risk of early and late onset AD (Corder *et al.*, 1993). In order to identify additional loci that are implicated in AD, a genome wide linkage analysis of affected sib pairs with LOAD was carried out. Positive lod scores were detected in a number of loci, these included a large region of the proximal long arm of chromosome 21 (which contained the 21q11 region). The positive lod scores on proximal chromosome 21q were obtained in the APOE-ε4-negative subgroup and reached a maximum level around the APP gene (Kehoe *et al.*, 1999). A proposed second stage study, in which more sibling pairs and additional markers will be used, should help determine whether a gene involved in LOAD might be present in 21q11.

1.8.1.3 Leukaemia in Down Syndrome

DS children are 10-30 times more likely to develop acute leukaemia than non-DS individuals (reviewed by Zipurski *et al.*, 1992). Although the relative proportions of acute lymphocytic leukaemia (ALL) and acute non-lymphocytic leukaemia (ANLL) have been reported to be similar between older children with DS and karyotypically normal children (Stiller and Kinnier Wilson, 1981), cases of ANLL in DS children under 4 years of age are predominant (Zipurski *et al.*, 1992). Iselius and co-workers have shown that approximately 50% of the ANLL cases in DS individuals are of the M7 subtype (ANLL-M7 or acute megakaryoblastic leukaemia) (Iselius *et al.*, 1990). It has

been estimated that this type of leukaemia could be 400 times as common in DS compared with karyotypically normal children. (Zipurski *et al.*, 1987).

A phenomenon that has been linked with ANLL-M7 is transient abnormal myelopoiesis (TAM), a self regressing leukemoid reaction (Zipurski *et al.*, 1987, 1992). The malignant cells have the properties of megakaryoblastic cells (Coulombel *et al.*, 1987) and approximately 20% of patients with TAM will develop ANLL-M7 (Zipurski *et al.*, 1992). In some cases TAM has been observed in karyotypically normal infants; in these instances trisomy 21 was observed in the leukaemic clone (Abe *et al.*, 1989). These observations suggest an increased risk of childhood leukaemia, especially of ANLL-M7 and TAM, when an extra copy of chromosome 21 is present.

Twenty DS patients with TAM were followed in two studies that assessed which chromosomes were inherited from the parents (Abe *et al.*, 1989; Niikawa *et al.*, 1991). All twenty patients in these studies showed an aab or aab/ab pattern of inheritance of chromosome 21. The aa chromosomes are homozygous, these findings therefore led to the hypothesis of disomic homozygosity (Abe *et al.*, 1989; Niikawa *et al.*, 1991); inheritance of a susceptibility gene in two copies would result in either TAM or ANLL-M7. An explanation for the fact that this type of leukaemia is infrequently observed in karyotypically normal individuals could be that inheritance of two malfunctioning alleles of a susceptibility gene is lethal. Disomic homozygosity can occur after a meiosis I nondisjunction, distal to a crossover, or after a meiosis II nondisjunction, proximal to a crossover, or in mitotic nondisjunction. Both Iselius (Iselius *et al.*, 1990) and Shen (Shen *et al.*, 1995) report an increase of atypical constitutional karyotypes in DS individuals with ANLL-M7 or TAM. The cases include mosaics of chromosome 21, i(21q), ring chromosome 21 and tetrasomy 21. These karyotypes would result in disomic homozygosity.

Niikawa and co-workers (Niikawa *et al.*, 1991) reported a crossover in one of the duplicated chromosome 21, in a patient with TAM and a maternal second meiotic nondisjunction. The crossover took place between 21q11.2 and q21.3, placing the area of disomic homozygosity and the gene that could cause ANLL-M7 or TAM proximal to this crossover. In the same study, they showed an individual with a 47, -21, XY, +inv(21)(q11.2q22.13), +inv(21)(q11.2q22.13) with TAM. The authors argue that the gene involved in ANLL-M7 and TAM might be disrupted in this individual and if so, might be located near the proximal breakpoint. Shen and co-workers (Shen *et al.*, 1995)

performed genetic mapping studies of chromosome 21 in DS individuals. They observed an increase in disomic homozygosity for all markers tested when comparing the non-leukaemic with the ANLL-M7/TAM DS population. This increase was by far the highest in the proximal region of chromosome 21. The markers with the highest proportions of disomic homozygosity were D21S215, S258, S120 and S192, which are all located in 21q11. A more recent study on a different population shows similar results (Cavani *et al.*, 1998).

The above mentioned studies show an involvement of chromosome 21 in leukaemia, in particular in ANLL-M7 and TAM, in children with DS. The strongest association with the location of a gene or genes on chromosome 21 that is implicated in these phenotypes is 21q11.

1.8.2 Similarity between 21q11 and other chromosomal regions

1.8.2.1 Centromere

The centromeric region of chromosome 21 contains alphoid sequences that show similarity to the alphoid sequences on chromosomes 13 and 14 (Choo *et al.*, 1991; Trowell *et al.*, 1993)

1.8.2.2 The duplicated region

In an attempt to identify polymorphic markers for linkage analysis at the AD locus in the pericentromeric region of chromosome 21, a phage library was constructed from a mouse-human somatic cell hybrid which contained chromosome 21 as the only human material (Van Camp *et al.*, 1989). One of these phage clones (pVC1.21b-D21S190) hybridised equally strongly to two locations on chromosome 21; in 21qcen-q21.2 and 21q22.1-q22.2 (Van Camp *et al.*, 1990). Subsequent mapping on additional somatic cell hybrids and on long range restriction maps showed that the proximal copy of this marker was located at 21q11.1, at a maximum distance from D21Z1 of 350 Kbp (Van Camp *et al.*, 1992, Van Hul *et al.*, 1993). The size of this duplicated region was estimated, by screening chromosome 21 YAC libraries with the D21S190 probe. Isolation of probes from the extremities of the D21S190 positive YACs and subsequent mapping onto somatic cell hybrids and positioning on restriction maps of the YACs resulted in a size estimation of between 135 and 500 Kbp (Dutriaux *et al.*, 1994). The size of the duplicated region was determined more accurately by hybridisation of

extremities of fragmentation YACs. These were generated from a YAC that spanned the duplicated region and were subsequently positioned onto a restriction map and a somatic cell hybrid panel. This showed the duplicated region to be between 220 and 375 Kbp (Potier *et al.*, 1996). A physical map of PACs and cosmids was eventually constructed spanning both copies of the duplicated region. STSs located just outside, and polymorphisms inside the region were used to distinguish between both copies (Potier *et al.*, 1998).

1.8.2.3 Proximal 21q11

A number of studies have shown similarity between 21q11 and other chromosomes. The markers D21S190 and D21S149 were shown to map to 21q11.1. Both markers cross-hybridised to chromosome 13 but not to 14, 15 or 22 (Van Camp *et al.*, 1990; Van Camp *et al.*, 1992). Long range restriction mapping showed the following order of the 21q11 markers: D21Z1 (centromere)-S190-S215-S149-S253-S258-S13 (21q11) (Van Hul *et al.*, 1993). The markers D21S215, S253 and S258 were analysed for cross-hybridisation to chromosome 13; S215 and S253 hybridised to chromosome 13 whereas S258 did not (Van Hul *et al.*, 1993). The chromosomal location was determined from a number of markers from 21q11. They were localised by PCR analyses using a monochromosomal hybrid panel. This showed that a large number of markers were positioned on additional chromosomes, mainly chromosomes 13, 15 and 18 (Graw *et al.*, 1995).

1.8.3 Genes, cDNA clones, exon trapped material and physical maps of chromosome 21q11

1.8.3.1 Gene density of chromosome 21q11

A compositional map was constructed for the long arm of chromosome 21, in which the GC content of STSs and genes was assessed by hybridisation to a Southern blot containing different fractions from a $CsSO_4$ gradient. The gradient separates the DNA fragments according to their GC content. The STSs and genes which are linked to a physical map show which regions of the chromosome are GC rich and therefore gene rich. These data showed that the cytogenetic bands 21q11.2 and 21q21 were lowest in GC content and the telomeric R band the highest (Gardiner *et al.*, 1990), although only approximately 25% of the 21q11 region could be assessed with the probes that were

used. A study that determined the GC content of YACs from the 21q11.1 and 21q11.2 region, by measuring the buoyant density of the YACs in CsCl gradients, found that the whole region was equally GC poor. The whole region exhibited a GC content of 39% (De Sario *et al.*, 1997). These results predicted that the 21q11 region would be a gene-poor region.

The "gene-poorness" of the 21q11 region, forecasted by the compositional mapping, is compatible with a number of other observations:

- Only two *Not* I sites have been found in genomic DNA in the 3.5 Mbp 21q11 region (Ichikawa *et al.*, 1993). The *Not* I sites can be found in CpG islands, which are often associated with 5' ends of housekeeping genes (Lindsay and Bird, 1987). The distal half of the chromosome shows a much higher density of *Not* I sites (Ichikawa *et al.*, 1993).
- Only two genes were localised in the 21q11 region, after identification via biochemical properties. A much higher density of these genes localised to the distal half of the chromosome (Gardiner, 1996).
- Different studies that used random chromosome 21 cosmids found that the majority of exon trapped and cDNA selected clones mapped distally (reviewed by Gardiner, 1996).

1.8.3.2 Physical maps of 21q11

A number of physical maps have been described encompassing the 21q11 region. Since the association of the region with FEOAD, physical maps have been described (e.g. Van Hul and Van Broeckhoven, 1993) which used long range restriction maps generated from rodent cell lines containing human chromosome 21 as the only human material. A physical map, consisting of overlapping YACs of the long arm of chromosome 21 was constructed (Chumakov *et al.*, 1992). This map was subsequently improved (Korenberg *et al.*, 1995, Gardiner *et al.*, 1995). A physical map consisting of overlapping YACs of both arms of chromosome 21 has also been generated (Nizetic *et al.*, 1994). Cosmids were localised to segments of 21q11 by using YACs that had been mapped in this area as probes for hybridisation. These were hybridised to membranes that contained cosmids which were enriched for chromosome 21 inserts. This was part of the work which resulted in a cosmid pocket map of the entire chromosome (Nizetic *et al.*, 1994). In an earlier study, shorter overlapping YACs were also identified between the markers

D21S13 and S16 in the 21q11 region (Butler *et al.*, 1992). A large number of STSs were generated from these clones, by sequencing the ends of the overlapping YACs (Butler *et al.*, 1992).

1.8.3.3 Characterised genes, cDNA clones and exon trapped material mapping to chromosome 21q11

Two genes have so far been shown to map to chromosome 21q11, STCH and RIP140. STCH is a member of the stress70 chaperone family that participate in protein processing by binding to denatured or misfolded proteins and releasing them by an ATP-dependent mechanism. The STCH gene was localised to chromosome 21q11.1 by PCR using somatic cell hybrids and YACs, which positioned the gene between the markers D21S378 and D21S120 (Brodsky *et al.*, 1995). RIP140 is a nuclear protein that interacts *in vitro* with the AF2 domain of the oestrogen receptor. This domain is involved in activation of transcription. The gene was mapped to 21q11.2 by fluorescent *in situ* hybridisation (Cavailles *et al.*, 1995) and subsequently localised near the marker D21S13 (Katsanis *et al.*, 1998)

A number of studies have mapped fragments of genes to 21q11. These fragments can be helpful in establishing full length genes:

- Sequencing of clones from a microdisected library of chromosome 21q11-21q21, subsequent BlastN searches and mapping of the clones to a somatic cell hybrid panel showed that three clones were identical to human cDNA clones and were localised to 21q11 (Yu *et al.*, 1997).
- Xu and co-workers cDNA selected eleven cDNA clones, using YACs which are localised in 21q11.1 (Xu *et al.*, 1995).
- A parallel study showed cDNA selection of three further cDNA clones that were positioned to 21q11.1 (Tassone *et al.*, 1995).
- An exon trapped clone was shown to map to chromosome 21q11 (Yaspo *et al.*, 1995).
- cDNA clones were selected from a foetal brain cDNA library, using random cosmids. Nine clones were mapped back to YACs that were localised in 21q11 (Cheng *et al.*, 1994).
- Six cDNA clusters that are part of the Unigene clusters, were localised in the proximal 14 cM of chromosome 21 (Schuler *et al.*, 1996).

1.9 Aims of the thesis

The primary aim of this thesis was to make progress towards the identification of all the genes that are present in the 21q11 region. This area of chromosome 21 has been associated with non-small cell lung carcinoma (Kohno *et al.*, 1998), familial early onset Alzheimer's disease (St. George-Hyslop *et al.*, 1987, Goate *et al.*, 1989), and acute non-lymphocytic leukaemia in Down syndrome infants (Abe *et al.*, 1989; Niikawa *et al.*, 1991; Shen *et al.*, 1995; Cavani *et al.*, 1998). The following objectives were set out:

- To construct a physical map of the 21q11 region in bacterial clones, which contain overlapping segments of human genomic DNA. At the start of this project, lower resolution YAC maps were available of this region (Butler *et al.*, 1992; Chumakov *et al.*, 1992; Nizetic *et al.*, 1994). Progress towards the generation of the physical map is described in Chapter 3.
- To use the higher resolution bacterial clones to identify exons of novel genes by utilising exon trapping techniques and to compare the putative exons to cDNA and protein databases. Advances towards this goal can be found in Chapter 4.
- To assess the outcome of the exon trapping experiments and to attempt to generate full length novel genes that are located in the 21q11 region. In addition, to make an initial assessment of the involvement of these genes in the phenotypes associated with the region. Progress in these areas can be found in Chapters 5 and 6.

Chapter 2 Materials and Methods

2.1 Chemical reagents

Agarose	Life Technologies
Low melting point agarose	Life Technologies
Rapid agarose	Life Technologies
Yeast extract	Life Technologies
Bacto-peptone 140	Life Technologies
Agar	Life Technologies
pAMP 1 cloning system	Life Technologies
Foetal bovine serum	Life Technologies
EDTA (0.02% w/v) in saline	ICN
Trypsin (0.05% w/v) EDTA (0.02% w/v) solution	Sigma
Dulbecco's Modified Eagles Medium (DMEM)	Sigma
RPMI 1640	Sigma
Foetal bovine serum, heat inactivated	Sigma
Penicillin/streptomycin (10,000 U/ml/ 10 mg/ml)	Life Technologies
Glutamine (200 mM)	Life Technologies
Prep-A-Gene matrix	Bio-Rad
RNAsol B	Biogenesis
Qiagen Tip 100 columns	Qiagen
RNAse inhibitor	Promega
Wizard resin	Promega
Total human placental DNA (10.6 mg/ml)	Sigma
Bovine serum albumine	Sigma
Yeast tRNA	Sigma
Performance optimised Polymer 6 (POP 6)	Perkin Elmer
Performance optimised Polymer 4 (POP 4)	Perkin Elmer
Ethidium bromide (10 mg/ml)	Fluka
COT I DNA (1 mg/ml)	Life Technologies
Prestained protein molecular weight standards	Life Technologies
Enhanced chemiluminescence assay kit	Amersham
Anti rabbit IgG-horseradish peroxidase	Amersham
Original TA cloning kit	Invitrogen
Marathon-Ready cDNA from bone marrow	Clontech
Fixer	Sigma
Developer	Sigma

pAMP 1. Prep-A-Gene, Qiagen Tip 100 columns, Wizard DNA purification resin, Enhanced chemiluminescence assay kit, the Original TA cloning kit and the Marathon ready cDNA were used according to manufacturers' recommendations.

All other common chemicals were supplied by: Merck BDH, Life Technologies, Fluka, Fisons and Sigma.

2.2 Enzymes

Restriction endonucleases (4-20 U/µl)	Life Technologies/NEB
Alkaline phosphatase from calf intestine (1 U/µl)	Boehringer Mannheim
Proteinase K	Merck BDH
Lysozyme (20.1 U/µg)	Pharmacia
Klenow (6.6 U/µl)	Life Technologies
M-MLV reverse transcriptase (200 U/µl)	Promega
RNAse inhibitor (40 U/µl)	Promega
Taq polymerase (5 U/µl)	Promega
Taq polymerase (5 U/µl)	HT Biotechnology
Pfu Taq polymerase (5 U/µl)	Stratagene
The advantage 2 polymerase mix	Clontech
T4 DNA ligase (1 U/µl)	Amersham
β-Agarase I (1 U/µl)	Calbiochem
T4 Polynucleotide kinase (10 U/µl)	NEB
Big dye cycle sequencing reaction mix	Perkin Elmer
Pronase	Boehringer Mannheim
Sequenase 2.0	Amersham

Restriction endonucleases and modifying enzymes were used with the supplied reaction buffers, unless otherwise stated.

2.3 Nucleotides and oligonucleotides

Hexanucleotides (3 µg/µl)	Life Technologies
Oligonucleotides	Life Technologies, OSWEL
3' deoxynucleoside 5'-triphosphate (dNTPs) (dATP, dGTP, dCTP, dTTP) 100 mM	Promega
α-^{32}P-dCTP (10 mCi/ml)	ICN/Amersham
γ-^{32}P-dATP (10 mCi/ml)	ICN/Amersham

2.4 Cloning vectors

PBluescript^{r} II phagemid vector	Stratagene
PCR2.1 vector	InVitrogen

pSPL3B splicing vector	Life Technologies/Genzyme

2.5 DNA size markers

100 bp DNA ladder (100 bp-2.1 Kbp)	Life Technologies
1 Kb ladder (75 bp-12.2 Kbp)	Life Technologies
High molecular weight DNA markers (8.2-48.5 Kbp)	Life Technologies
λ Ladder (48.5-1000 Kbp)	Bio-Rad
Saccharomyces cerevisiae marker (225-2200 Kbp)	Bio-Rad
Genescan 350 size standards	Perkin Elmer

2.6 Hybridisation membranes and X-ray film

Hybond-N+ membrane	Amersham
Nitrocellulose	Schleicher and Schuell
Blue sensitive X-ray film AX	GRI
Polaroid 667 *professional* film	KJP
Hyperfilm™ECL film	Amersham

2.7 Northern blots

Blots were purchased from Clontech and contained 2 μg of poly A+ RNA per lane. Each lane contained RNA from different tissues:

Human I: heart, brain, placenta, lung, liver, skeletal muscle, kidney, pancreas

Human II: spleen, thymus, prostate, testis, ovary, small intestine, colon, peripheral blood leukocyte

Blots were hybridised using express hyb solution (Clontech) and washed according to manufacturer's recommendations.

2.8 Solutions

Denaturing solution:	0.5 M NaOH, 1.5 M NaCl.
Neutralising solution:	1 M Tris-HCl pH 7.4, 1.5 M NaCl.
PROPK:	50 mM Tris-HCl pH 8.5, 50 mM EDTA, 100 mM NaCl, 1% w/v Na-lauroylsarcosine, 0.015% w/v proteinase K.
HMFM (10x):	360 mM K_2HPO_4, 132 mM KH_2PO_4, 17 mM Na_3citrate, 4 mM $MgSO_4$, 68 mM $(NH_4)_2SO_4$, 44% v/v glycerol.
AL1:	50 mM glucose, 10 mM EDTA, 25 mM Tris-HCl pH 8.0, 0.4% w/v lysozyme.
AL2:	0.2 M NaOH, 1% w/v SDS.
AL3:	60 ml 5 M K-acetate, 11.5 ml glacial acetic acid, 28.5 ml H_2O, pH 4.8.
20x SSC:	3 M NaCl, 0.3 M Na_3citrate pH 7.0.
Chloroform/IAA:	Chloroform/isoamyl-alcohol mixture 24:1.
2x YT:	1.6% w/v tryptone, 1.0% w/v yeast extract and 0.5% w/v NaCl.
Superbroth:	3.2% w/v tryptone, 2.0% w/v yeast extract, 0.5% w/v NaCl, 5 mM NaOH.

LB:	1.0% w/v peptone, 0.5% w/v yeast extract, 0.8% w/v NaCl.
Chloramphenicol (1000x):	34 mg/ml in ethanol.
Kanamycin (1000x):	50 mg/ml in H_2O, filter sterilised 0.2 μm filter.
Ampicillin (1000x):	50 mg/ml in H_2O, filter sterilised 0.2 μm filter.
IPTG	24 mg/ml in H_2O, filter sterilised 0.2 μm filter.
Xgal	20 mg/ml in dimethylformamide.
Church solution:	0.5 M NaPi buffer pH 7.2, 7% w/v SDS, 1 mM EDTA and 0.01% w/v yeast tRNA.
NaPi pH 7.2:	720 ml 1 M Na_2HPO_4, 280 ml 1 M NaH_2PO_4.
NaPi pH 6.8:	463 ml 1 M Na_2HPO_4, 537 ml 1 M NaH_2PO_4.
(pre-) hybridisation solution:	6x SSC, 5x Denhardt's, 0.5% w/v SDS and 0.01% w/v herring sperm DNA.
100x Denhardt's:	2% w/v BSA, 2% w/v Ficoll and 2% w/v polyvinylpyrrolidone.
5x Labelling buffer:	0.25 M Tris-HCl, pH 8.0, 50 mM βME, 100 μM dATP/dGTP/dTTP, 1M HEPES pH 6.6, 0.9 μg/μl random hexamer primers.
10x TBE:	0.9 M Tris base pH 8.0, 0.9 M Boric acid, 20 mM EDTA.
TFB I:	30 mM Kacetate, 50 mM MnCl2·4H2O, 100 mM $RbCl_2$, 10 mM $CaCl_2$, 15% v/v glycerol pH 5.8, filter sterilised.
TFB II:	10 mM MOPS pH 7.0, 75 mM $CaCl_2$·$2H_2O$, 10 mM $RbCl_2$, 15% v/v glycerol, filter sterilised.
SOC:	2.0% w/v tryptone, 0.5% w/v yeast extract, 0.05% w/v NaCl, 5 mM $MgSO_4$, 1 mM $MgCl_2$, 0.4% w/v glucose.
NDS:	1% w/v SDS, 10 mM Tris base pH 9.5, 0.5 M EDTA, filter sterilised.
T0.1E:	10 mM Tris HCl pH 8.0, 0.1 mM EDTA.
1x TE:	10 mM Tris HCl pH 8.0, 1 mM EDTA.
T100E:	10 mM Tris HCl pH 8.0, 100 mM EDTA.
SSSarc:	7% w/v sarcosyl, 4x SSC, 1mM EDTA pH 8.0, filter sterilised.
Cycle sequencing buffer (5x):	400 mM Tris HCl pH 9.0, 10 mM $MgCl_2$.
Guys buffer (10x):	100 mM Tris HCl pH 8.3, 500 mM KCl, 15 mM $MgCl_2$.
5x sequencing buffer:	200 mM Tris HCl pH 7.5, 10 mM $MgCl_2$ 250 mM NaCl.
PBS:	137 mM NaCl, 2.7 mM KCl, 10.1 mM Na_2HPO_4, 1.8 mM KH_2PO_4.
Ethidium bromide:	5 μg/ml ethidium bromide in H_2O.
Solution D	100 ml solution A: 1 ml solution B: 1 ml solution C.
Solution A	2% w/v Na_2CO_3, 0.1 M NaOH, 0.5% w/v SDS.
Solution B	2% w/v NaK-tartrate.
Solution C	1% w/v $CuSO_4$.
3x Sample buffer	30 mM NaH_2PO_4, pH 7.0, 30% v/v glycerol, 0.05% w/v bromophenol blue and 7.5% w/v SDS.

Running buffer	1.5 M Tris, 8 mM EDTA, 0.4% w/v SDS, pH 8.8.
Stacking gel buffer	0.5 M Tris, pH 6.8, 8 mM EDTA and 0.4% w/v SDS.
Electrophoresis buffer	50 mM Tris, 384 mM glycine, 1.8 mM EDTA and 1.8% w/v SDS, pH 8.8.
Transfer buffer	25 mM Tris, 192 mM glycine and 20% methanol v/v, pH 8.5.
Blocking buffer	5% w/v milk powder, 0.02% v/v Tween-20 in PBS.
Wash buffer	2.5% w/v milk powder, 0.02% v/v Tween-20 in PBS.
STE	0.1 M NaCl, 10 mM Tris-HCl pH 8.0, 1 mM EDTA.

2.9 PCR Primers

2.9.1 IPCR/vectorette primers

Primer name	PCR method	Sequence (5'-3')	T^{o}_{an}
primer 1 (PAC)	IPCR/vectorette	CATTTAGGTGACACTATAGAGGATC	60^{0}C
primer 2 (PAC)	IPCR	CGGCCAATTAGGCCTACCCAC	60^{0}C
primer 3 (PAC)	IPCR/vectorette	CGACTCACTATAGGGAGAGGATC	60^{0}C
primer 4 (PAC)	IPCR	CAACCCAGTCAGCTCCTTCC	60^{0}C
B1 (BAC)	IPCR/vectorette	GCTATTTAGGTGACACTATAG	55^{0}C
B2 (BAC)	IPCR	CACAATTCCACACAACATACG	55^{0}C
B3 (BAC)	IPCR/vectorette	GTAATACGACTCACTATAGGG	55^{0}C
B4 (BAC)	IPCR	CAGTTGCGCAGCCTGAATGG	55^{0}C
V	vectorette	CGAATCGTAACCGTTCGTACGAGAATCGCT	55^{0}C 60^{0}C
LAW4L (cosmid)	vectorette	CGCCTCGAGGTGGCTTATCG	60^{0}C
LAW4R (cosmid)	vectorette	ATCATACACATACGATTTAGGTGAC	60^{0}C
Pst I top (linker)	vectorette	GAAGGAGAGGACGCTGTCTGTCGAAGGTAA GGAA CGGACGAGAGAAGGGAGAG	
Pst I bottom (linker)	vectorette	CTCTCCCTTCTCGAATCGTAACCGTTCGTAC GAGAATCGCTGTCCTCTCCTTCTGCA	
Hind III top (linker)	vectorette	AGCTGAAGGAGAGGACGCTGTCTGTCGAAG GTAAGGAACGGACGAGAGAAGGGAGAG	
Hind III bottom (linker)	vectorette	CTCTCCCTTCTCGAATCGTAACCGTTCGTAC GAGAATC GCTGTCCTCTCCTTC	
6MW	DOP	CCGACTCGAGNNNNNNATGTGG	58^{0}C

2.9.2 Marker primers (all primer information was derived from Genome Data Base except H41072, the primers which were designed from the cDNA H41072 [J.H. Ives])

Primer name	Description	Sequence (5'-3')	T^{o}_{an}
D21Z1-A D21Z1-B	Unknown poly-morphism 640-820 bp	CTTCTGTCTAGATTTTAGA CATAGAGATGAACATGG	45^{0}C
D21S369-A D21S369-B	Dinucleotide repeat 173-201 bp	ATGGCCTTGGCTAAATGCTG CTAAGCTGATATGGTAAGTACA	52^{0}C
D21S215-A D21S215-B	Dinucleotide repeat 168-180 bp	TCTAAAACAGTGTGTCTAGC GCTGACGTGACAGTTGTGAG	52^{0}C
D21S297-A D21S297-B	261 bp	GAAGCAGATGCTTATCTATTTTGTC TAGAGTTTTGAGTTCATGTGCATT	56^{0}C
D21S308-A D21S308-B	312 bp	CCAAGCAATCACAAGGGTCT AAACACCACAAGTCTGCTATCTGA	56^{0}C
H41072-F2 H41072-R2	241 bp	AACTCCTGTCACCTCCTCACTG ACACTGTGAATCGTCGCTATG	53^{0}C
D21S149-A D21S149-B	170 bp	CAGAAACCGCTCAGCTTTGC CATTGTCTCACCTCAGCTGC	57^{0}C

D21S318-A D21S318-B	207 bp	ACAATGGGGACCCTTAGC GTGAGTCACAGAGCCTGAACTT	56^{0}C
D21S258-A D21S258-B	Dinucleotide repeat 313-335 bp	AGGCATCTCCAAGATTTCTAGG CCACGGGACAGTGTTCACTA	55^{0}C
D21S378-A D21S378-B	110 bp	GTACACATGATTTTATTGTGTCTA AATTCAGTCAAGGATGACGATTGAC	60^{0}C
D21S120-A D21S120-B	Dinucleotide repeat 318-330 bp	GTGTGTCTGCCATTTCTGGGTGTAG GATCCTGGGACAAAGTAGTCTCTAA	55^{0}C
D21S374-A D21S374-B	100bp	TTGGTTTCCTTGAACATCTTTGTG GCAGAAGGAGAGAAAGACCACTGG	60^{0}C
D21S16-A D21S16-B	Dinucleotide repeat 124-238 bp	TTCCATGCCCAGACTGAAGTTCTGGACTGA CCTATCACCAATGGAGAAAGCCCATAAGGA	57^{0}C
D21S236-A D21S236-B	Dinucleotide repeat 104-128 bp	CCCAAATAAAAAAGAGAACAG CTAAAGAGGACTTCAGAGTAAGG	48^{0}C
D21S371-A D21S371-B	252 bp	ATACGACTCACTAGGGGGATCTC CTCCTGGAAAACAAAAGGACTCTC	60^{0}C
D21S373-A D21S373-B	380 bp	CCATATCAGGCCCTGAATATCAGC CATGAATGGCCAGATGACAGATCC	60^{0}C
D21S13-A D21S13-B	Dinucleotide repeat 120-130 bp	TTTAGATTAGTTTAAGAAACATG GAAAAGATGATTCTTTCCACC	46^{0}C
D21S372-A D21S372-B	650 bp	CCAAAACAACAGTCTAGAGTGAGC TGACTCAAAGGCTTCCATGTAGTC	60^{0}C
D21S192-A D21S192-B	195 bp	CTTTTGACCTTCCCTTTGAGC GAGAACTAACTCACTATCGGG	57^{0}C
D21S172-A D21S172-B		TACAGTGGCAAATGTCATTG GAATATGTGTTAGGTCCTGC	
D21S382-A D21S382-B	72 bp	GCCAGGCGCGAAGGCGATG CCCTCACAGTCGGCGTTTCG	57^{0}C

2.9.3 Exon trapping and vector primers

Primer name	Sequence (5'-3')	T^0_{an}
SD6 SA2	TCTGAGTCACCTGGACAACC ATCTCAGTGGTATTTGTGAGC	55^{0}C
dUSA4 dUSD2	CUACUACUACUACACCTGAGGAGTGAATTGGTCG CAUCAUCAUCAUGTGAACTGCACTGTGACAAGC	60^{0}C
EX-1 EX-2	CTGCACGTGCTCTAGAGTCG TCCCCTCGGGAGATCTCCAG	59^{0}C
M13-R M13-F	AGCGGATAACAATTTCACACAGG CCCAGTCACGACGTTGTAAAACG	55^{0}C
M13R+ M13F+	CACACAGGAAACAGCTATGACC ACGTTGTAAAACGACGGCCAGTG	55^{0}C

2.9.4 Unigene primers

Primer name	Size of product	Sequence (5'-3')	T^0_{an}
Bdac-A Bdac-B	84 bp	GAGCAGCATCTCAGGACAC CTCATCTCCAGCACAGCAC	57^{0}C
4O28-A 4O28-B	178 bp	TACAAGCTTATAGAGAAAGTCAA TATGTATAAAATAGGCAACTAGG	57^{0}C
11705-A 11705-B	110 bp	AATCCAAAGCTCAGAGTTTAATGG CCACAGTTCCTGGTTGGAGT	57^{0}C
13495-A 13495-B	105 bp	TGAATGAACACAAACGTCTATGC GAGCTAGTGTAGAAGCTGCCTACC	60^{0}C
12603-A 12603-B	343 bp	CCTATGTCAATGAAATGCTAGGG CTGCTAATAGCAGCAAACTAAACTG	60^{0}C
2B43-A 2B43-B	260 bp	AATATTATGTATCGTACACAGTG AGACAAAATGACTATACAGACTT	57^{0}C

2.9.5 STCH primers

Primer name	Size of product	Sequence (5'-3')	T^{o}_{an}
STCH-1A STCH-1B	918 bp	AGTCATCACAAGCCTGTTC TTTGACCATTTCCACAGC	50^{0}C

2.9.6 USP21 primers

Primer name	Sequence (5'-3')	T^{o}_{an}
USP-A	TACCAAGGCATCACATGAGC	55^{0}C
USP-B	CGATAATCGGTTTCTGGTGG	55^{0}C
N92952-A	AGAGTTTCAGCTTGCATGTG	55^{0}C
N92952-B	TTAAAGAACCCTGCTTCAGG	55^{0}C
N92952-R2	GAGTGTATATGACAGGTGGC	50^{0}C
USP 3' RACE	CTTCATATTCCACGCATGAACTCTG	72^{0}C, 70^{0}C, 68^{0}C
USPEXP1	CTTGAGCTCACCGTGGAGCAGAACGTGCTG	57^{0}C
USPEXP2	CTTGTCGACTTATCTTCCATCAGCAGGAGTTC	57^{0}C
USP-ER	CCTGAATTCTCTGAATGTAAAG	57^{0}C
USP-EF	AGAGAATTCAGGAAAATCAGGCC	57^{0}C
USP-XR	TGCTCTAGATATTCTGGGTCTCC	57^{0}C
USP-XF	ATATCTAGAGCAGCCATCAAGAAG	57^{0}C
USP-1R	TTATTCTCTGCTATGCTGGC	56^{0}C
USP-2F	CAGGGAGACTGGAATAACTG	56^{0}C
USP-2R	TTCTCTTCATCCGTCTCCTC	56^{0}C
USP-3F	GATGTGAGTGAGTTTACACAC	56^{0}C
USP-3R	TTTGGGACCGGAACCATAGC	56^{0}C
USP-4F	GAGGGAAGAGATCAAGAGAC	56^{0}C
USP-4R	GGCATGTAATCGATAAGGAAC	56^{0}C
USP-5F	AGGAAAGCATATCCAGAATCC	56^{0}C
USP-6R	GATGAGGAACAGCAAGGAATC	56^{0}C
USP-7F	CAGGGAAACAACTATGTATCTC	56^{0}C
His 1	ACAGGTTGTTTGTGGGTTTC	55^{0}C
His 2	CCATGATGATTTTGTCACAGC	55^{0}C
QQD 1	GTTATCTTTGTAAGGAGTTGG	55^{0}C
QQD 2	TCAGCTTTCATTTGGAAGGC	55^{0}C
CGA830-F	CTAATGGAATTTTCCCTCAGG	55^{0}C
CGA830-R	CAGTTTCAGATTTGATGAGCC	55^{0}C
CGA1532-F	GTTTCCTGTCTTTGGGGAAC	55^{0}C
CGA1532-R	CAACTGTCTCTTGCATCAGG	55^{0}C
CGA1922-F	CCATAGATTTTCTCCCATAAGAC	55^{0}C
CGA2282-F	GCAGTCTATTTTGAGCTTTGC	55^{0}C
CGA2282-R	GGACAAACCTGTTGTCACAG	55^{0}C
D21S366-F	TGATACTGTACTGTATACTTACA	56^{0}C
D21S366-R	6-FAM-TGTCCCTTCAGAGACCCCT	56^{0}C
2029-A	TGGGCATATAT	54^{0}C
2029-T	TGGGCTTATAT	54^{0}C
USPE-2F	GGTTCTCCACAGCATGATTC	54^{0}C
USPE-2R	CTAGAGATGAGCATATACCAC	54^{0}C
USPE-3F	TTTGTTCTAGAGAGTTATGGAG	54^{0}C
USPE-3R	GTTCTTTACCTCTCCTTAAGC	54^{0}C
USPE-4F	GGTACTGACTGGTTTAAAGAG	54^{0}C
USPE-4R	ATCCCTCATAACACCATCCC	54^{0}C
USPE-5F	GAATGTGTGATATGCCTGGG	54^{0}C
USPE-5R	GATGGGAGGATACTTCAATCC	54^{0}C
USPE-7F	GGCACCAACACTCTACTTTC	54^{0}C
USPE-7R	TCATATTCTTGGGAAAGGTGC	54^{0}C

USPE-8R	CCCATTGCTAAACTTCACGG	54^{0}C
USPE-9F	GCGTGATAGCCACATAAGTG	54^{0}C
USPE-9R	AACTATCGAGACGAATGAATG	54^{0}C
USPE-10F	AATTTTAGTAGCCAGAAGGCC	54^{0}C
USPE-10R	TGTAAGAAAGACCTAGGGAAG	54^{0}C
USPE-11F	TCCTGTGCCTAAGATTGCAG	54^{0}C
USPE-11R	TTCATCTACCTTGACAAATGG	54^{0}C
USPE-12F	GCATCTGGCCATATTGCATC	54^{0}C
USPE-12R	GTAGGAACATTACCTGTTGAC	54^{0}C
USPE-14F	AATACTGTGGGTTTGGCACG	54^{0}C
USPE-14R	TACACATATCCATCTGAGGTG	54^{0}C
USPE-15F	TTCCTCTAACATGTTTGTAGG	54^{0}C
USPE-16R	ATCACACCACATACCAGAGC	54^{0}C
USPE-17R	TTCCTCACACTTTAAGATAGC	54^{0}C
USPE-18F	ATCTTAGTATAGACTCTGTGG	54^{0}C
USPE-18R	ACTGGAGGAAATGTCGTGAC	54^{0}C
USPE-19F	AGTTGTACTGCCTTGTTTGTG	54^{0}C
USPE-19R	TTATCCTGGGAATCTGCAGG	54^{0}C
USPE-20F	TACTGCTGTTTTTGCTCAGTG	54^{0}C
USPE-20R	GGCTGAAACGAAAGGAATATC	54^{0}C
USPE-21F	AGTAGCGAAACAGTGCATTAC	54^{0}C
USPE-21R	GATGAGGTCACACCTGAATAG	54^{0}C
USPE-22F	CTGACTTTGTTTCTACAAAGG	54^{0}C
USPE-22R	TGCCCAATGTCAGACAAAATG	54^{0}C
USPE-23F	CCGTTTGTTCTAATACTGTTTG	54^{0}C
USPE-23R	ACATCTCAGGTGCCTACTAC	54^{0}C
USPE-24F	ATCATGTTGACAATGATCGTG	54^{0}C
USPE-24R	CTATCCAATCTTAATCCCCAG	54^{0}C
USPE-24R2	GTGATTATACATCAGTTCCGG	54^{0}C
USPE-24A	CCGGAACTGATGTATAATCAC	50^{0}C

2.9.7 G31B5 primers

Primer name	**Description**	**Sequence (5'-3')**	**T^o_{an}**
AP1	linker primer	CCATCCTAATACGACTCACTATAGGGC	72^{0}C, 70^{0}C, 68^{0}C
AP2	linker primer	ACTCACTATAGGGCTCGAGCGGC	72^{0}C, 70^{0}C. 68^{0}C
31B5-E21F	RACE primer	GAGGTGTATTCCTGCAGATGAATCC	72^{0}C, 70^{0}C, 68^{0}C
31B5-E21R	RACE primer	GGATGTCATCTCAGAAGAGGAAGC	72^{0}C, 70^{0}C, 68^{0}C
UG1-F1		TTGATGAGCTCTGTCCACTG	56^{0}C
UG1-R1		AAGTGACTGAACACGCATTCC	56^{0}C
UG1-R2		CTTTGGTATGTGTGACATATGC	56^{0}C
UG1-R3		GTGTAGTCACTTCTGATTTAGTG	56^{0}C
G31B5-A		TTAGACAGCACACTGCTGAC	50^{0}C
G31B5-B		ATTGGATGGCTTTCTCTTGAG	50^{0}C
G31B5-C		GAGGCAGTTTTGGGAATTTCG	50^{0}C
G31B5-D		CAGTTGAATCATCTGGTTTTG	50^{0}C
G31B5-E		GAGAGCTATTTCATGGACAATG	50^{0}C
G31B5-F		TTTTCCTCAGAAAGGGCTTTG	50^{0}C
G31B5-1F		CATATCCTGACTGAAGGAGG	54^{0}C
G31B5-1R		TGCAAAAGTGACATTGTCTTC	54^{0}C
G31B5-2F		TACTTATCAGTTTGGACAGTG	54^{0}C
G31B5-2R		TGCAAAAGTGACATTGTCTTC	54^{0}C
G31B5-3F		CCACAAACTCAGAATGTTCTC	54^{0}C
G31B5-3R		TCATAAAGTACCAAGCACTTC	54^{0}C
G31B5-4F		ATCTTCTCTCTGTACCTAACC	54^{0}C
G31B5-4R		ACCTCACACCCAGGATCTTG	54^{0}C
G31B5-5F		ATGACAACTGTGAACTTCTGC	54^{0}C

G31B5-5R		TTTGACAGACCATTCAGAAGG	54^{0}C
G31B5-6F		CTCTGAAATCTAATGTGTATGG	54^{0}C
G31B5-6R		TCGGTGTTTCCATTTACATGC	54^{0}C
G31B5-7F		CCATGTGCATCTCTATACTTTC	54^{0}C
G31B5-7R		TTATCAGCTGGGGAGAAGAG	54^{0}C
G31B5-8F		CTTGAAAGTGAATGAGTACATG	54^{0}C
G31B5-8R		TGTCTGGGAGGAGATATTTTC	54^{0}C

2.9.8 APOE primers

Primer name	Sequence (5'-3')	T^0_{an}
APOE-F	TGATGGACGAGACCATGAAG	54^{0}C
APOE-R	TTACGCAGCTTGCGCAGGTG	54^{0}C

2.10 Markers used for hybridisation

These markers were previously obtained from C. Van Broeckhoven. They were originally described in the following reports: D21S149, D21S189, D21S190 and D21S192 (Van Camp *et al.*, 1990), D21S253 (Stuyver *et al.*, 1991), D21S13 (Stewart *et al.*, 1985), D21S46 (Korenberg *et al.*, 1987).

2.11 Somatic cell hybrids (for review see, Graw et al., 1995)

WA17: intact normal chromosome 21 (background cell line mouse A9).

2Fur: intact normal chromosome 21 t(21;CHO)(21q11.1>21qter) (background cell line Chinese hamster ovary (CHO)).

6;21: der(6)t(6;21)(6qter>6p25::21q11>21pter)mat (background cell line CHO).

R210: ring (21) (background cell line CHO).

JC6: del(21)(pter>21.15::q22.1>qter) (background cell line CHO).

6918: del(pter>q11.2::q22.1>q22.3) (background cell line CHO).

2.12 PAC, BAC, Cosmid libraries

Cosmid library c102/c103: flow-sorted chromosome 21 libraries from lymphoblastoid GM1416 cells, partially digested with a mixture of *Mbo* I and Dam methylase and ligated into the *Bam* H I site of the Lawrist 4 vector, approximately 10-fold chromosome coverage (Nizetic *et al.*, 1991a).

Total human PAC Library RPCI 1: was constructed from human male lymphocyte DNA with the use of the pCYPAC2N vector, approximately 3.3 genome equivalents (Ioannou *et al.*, 1994).

Total Human BAC library (Research Genetics).

2.13 Cell lines

COS 7B - African green monkey kidney cells.

WA17 - human chromosome 21 contained on a mouse A9 background.

A9 - mouse A9 cell line.

2.14 *E. coli* strains and constructs

XL1 blue : SupE44hsdR17recA1endA1gyrAugthirelA1Lac$^-$F'[proAB$^+$lacI^{q}lacZΔM15Tn10(tet^r)]

TOP10F' :F'{lacI^{q}Tn10(Tet^R)} mcr AΔ(mrr-hsdRMS-mrcBC) Φ80lacZΔM15ΔlacX74recA1 araD139Δ(ara-leu)7697galUgalK rpsL(Str^R)endA1 nupG

PRB105 :contains UBP2 a *S. cerevisiae* ubiquitin-specific protease gene in a pBR amplicon (Amp^r) (Baker *et al.*, 1992).

Ub-M-βgal :contains a ubiquitin-β-galactosidase fusion gene in pACYC 184 (Cm^r) (Baker *et al.*, 1992).

2.15 cDNA selected clones, exon trapped material, mapped ESTs (Unigenes)

TU18: exon trapped clone, as described in Yaspo *et al.*, 1995.

ES0033, ES0040, ES0067, ES0093, ES0136: cDNA selected clones from foetal brain, (Cheng *et al.*, 1994).

A004O28, A002B43, U09860, Bdac5g11, SHGC-11705 and WI-13495: cDNA clones represented by the Unigene amplimers (Schuler *et al.*, 1996).

2.16 Materials obtained from the Human Genome Mapping Project Resource Centre (HGMP-Hinxton Hall, Cambridge)

The following items were obtained:

- Image clones (Lennon *et al.*, 1996) using their IMAGE accession numbers.
- Twenty-one primary PAC pools containing *E. coli* clones from fifteen 384-well microtitre dishes each, from the total PAC library RPCI 1.
- Secondary PCR pools containing *E. coli* clones from individual 384-well microtitre dishes from RPCI 1.
- Individual microtitre dishes, from RPCI 1, for the PCR reactions to determine the final clone address.
- Mono-chromosomal somatic hybrid DNA panel for PCR, containing GM07299 (chromosome 1+X), GM10826B (2), GM10253 (3), HHW416 (4), GM10114 (5), MCP6BRA (6), CLONE21E (7), C4A (8), GM10611 (9), 762-8a (10), JICL4 (11), laA9602+ (12), 289 (13), GM10479 (14), HORLI (15), 2806H7 (16), PCTBA1.8 (17), DL18TS (18), GM10612 (19), GM10478 (20), THYB1.3 (21), PgME25NU (22), HORL9X (X), 853 (Y) and human, mouse and hamster as controls.
- RPCI 1 library replicated at high density on hybond N+ membranes, processed by HGMP.

2.17 Clinical Material

-Alzheimer's Disease cell lines were ordered from the Coriell Insitute for Medical Research (401 Haddon Avenue Camden, New Jersey). These were lymphoblast cell lines, generated from patients with familial early onset Alzheimer's disease (FEOAD). Cells were grown as indicated in 2.18.11 and DNA was isolated as shown in 2.18.13. The two samples used were AG08562, a cell line generated from a 39 year old female with FEOAD from FAD2, a family with inherited EOAD (St. George Hyslop *et al.*, 1987) and 07877A a 53 year old male with FEOAD from FAD4 (St. George Hyslop *et al.*, 1987).

-Genomic DNA from a panel of non-small cell lung carcinoma tumours was obtained from the Croatian Tumour Bank (Spaventi *et al.*, 1994) (Division of Molecular Medicine, Rugjer Boskovic Institute, Bijenicka 54, HR-10000 Zagreb, Croatia). Tumour and normal lung material was surplus clinical material obtained through normal surgical procedures which was directly frozen after surgery. The histological type of the tumour was confirmed by a pathologist. DNA was isolated, using standard molecular biological techniques.

-DNA from late onset Alzheimer's disease patients was isolated from peripheral blood lymphocytes by standard molecular biological techniques. The thirteen patients were (sample nr., gender, age): LOAD 1, F, 78, LOAD 2, F, 83, LOAD 3, M, 70, LOAD 4, F, 75, LOAD 5, M, 86, LOAD 6, F, 84, LOAD 7, M, 71, LOAD 8, F, 81, LOAD 9, F, 84, LOAD 10, F, 82, LOAD 11, M, 72, LOAD 12, F, 77, LOAD 13, M, 84. Samples were kindly provided by Dr. Svetlana Kalanj-Bognar, Department of Chemistry, University of Zagreb Medical School.

-The two ANLL-M7 samples (1559 and 2075) were surplus bone marrow samples from routine clinical biopsies performed at Great Ormond Street Hospital (F. Cotter). Sample 1559 was isolated from a remission sample from a 16 month old Down syndrome boy. Sample 2075 was a presentation sample, obtained from a 22 month old Down syndrome girl. Genomic DNA was previously isolated via standard molecular techniques.

2.18 General Protocols

2.18.1 Picking, replicating and processing

Generally a set of protocols described in Nizetic *et al.*, 1991a and Nizetic and Lehrach, 1995 was followed, with some modifications.

1. Individual clones were picked with sterile picking devices from microtitre dishes, or sterile cocktail sticks from petridishes and deposited in individual wells of microtitre dishes.
2. Clones were grown overnight at 37^0C, in 2x YT, containing 1x HMFM.
3. Replication of clones was performed using disposable 384 pin gadgets and clones were replicated on 22cm/22cm Hybond N+ filters. The membranes were placed on square NUNC plates containing 250 ml 2x YT with agar and 50 µg/ml Kanamycin (PACs) or 50 µg/ml Ampicillin (pAMP 1).
4. Bacteria were grown for 14-16h at 37^0C, after which individual colonies could be identified on the membrane.
5. Lysis of the cells was performed by incubating the membranes on pre-wetted Whatman 3MM paper as follows:
   - 4 minutes with denaturing solution.
   - 4 minutes with denaturing solution on a glass plate, which was situated on a plastic box in a boiling water bath.
   - 4 minutes with neutralising solution.
   - 1 minute on a dry piece of Whatman.
6. Membranes were subsequently incubated in PROPK solution at 37^0C for 30-60 minutes.

7. The DNA was crosslinked to the membrane in a crosslinker at 120 mJ/cm^2, after drying the filters overnight under an inverted NUNC plate to cover from dust.

2.18.2 DNA isolation

1. Single colonies of PACs and BACs were grown for 12-14h, in superbroth, cosmids and pBR 322 derivatives were grown in LB with antibiotics. PACs and cosmids were supplemented with 30 µg/ml Kanamycin, BACs with 12.5 µg/ml Chloramphenicol and pBR 322 derivatives with 50 µg/ml Ampicillin.
2. 2 ml of an overnight culture was centrifuged at 2000g for 5 minutes and the supernatant was removed. Another 2 ml of the culture was added to the same tube and the cells were precipitated at 2000g for 5 minutes.
3. The bacterial pellet was resuspended, by vortexing, in 125 µl AL1 and incubated at room temperature for 5 minutes.
4. The samples were incubated for 5 minutes on ice after addition of 250 µl AL2 and gentle inversions.
5. Samples were then incubated for a further 5 minutes on ice after addition of 187.5 µl AL3 and gentle inversions.
6. The samples were centrifuged for 10 minutes at 13800g and the supernatant transferred to a new Eppendorf tube.
7. 1 ml 100% ethanol was added to the samples, which were then incubated at room temperature for 2 minutes and centrifuged (13800g) for 10 minutes.
8. The resulting pellets were dried and resuspended in 100 µl H_2O.
9. Phenol and chloroform/IAA (24:1 v/v) extractions were performed. 100 µl phenol was added and the samples were either vortexed for 1 minute or inverted for several minutes (for vectors with large inserts to minimise shearing of the DNA). The samples were then centrifuged for 3 minutes after which the aqueous phase was transferred to a new tube, and a chloroform/IAA extraction was carried out in a similar manner.
10. 0.1 volume of 3 M NaAc pH 5.2 and 2 volumes of 100% ethanol were added to the aqueous phase from step 9, and the samples were mixed, incubated on dry ice for 15 minutes and centrifuged for 10 minutes at 13800g.
11. The final DNA pellet was washed with 0.5 ml 70% ethanol, centrifuged for 5 minutes at 13800g, dried at 37^{0}C and resuspended in 30 µl H_2O.

2.18.3 Standard PCR

1. PCR was carried out in 50 µl total volume, with 5 µl 10x reaction buffer, 25 pmol of each primer, 10 nmol each of dNTP and 0.1U Taq Polymerase (HT Biotechnology). A positive control was included when available; a negative control (H_2O) was always performed.

2. Standard reactions were: 1 minute 95^{0}C, (1 minute 95^{0}C, 1 minute annealing temperature, 2 minutes 72^{0}C) 35x, 8 minutes 72^{0}C or 1.5 minutes 95^{0}C, (30s 95^{0}C, 30s annealing temperature, 45s 72^{0}C) 35x, 10 minutes 72^{0}C.

2.18.4 Electro-elution

1. Dialysis membrane was treated by incubation in boiling H_2O for 2 minutes and rinsing in cold H_2O.
2. A gel slice was inserted into a 10 cm length of membrane, 400 µl 1x TE was added and clips were used to seal both ends.
3. Electrophoresis was carried out after placing the dialysis membrane containing the gel slice in an electrophoresis tank with 1x TBE, at 40 mA for 45 minutes.
4. The 1x TE solution containing the DNA was transferred to an Eppendorf tube. DNA was precipitated using 0.1 volume 3.0 M NaAc pH 5.2 and 2 volumes 100% ethanol.
5. The supernatant was removed and the pellet washed by addition of 0.5 ml 70% ethanol, after which the sample was centrifuged at 13800g for 5 minutes.
6. The supernatant was removed and the dried pellet resuspended in 20 µl H_2O.

2.18.5 Southern Blotting

1. Agarose gels were de-purinated by submerging the gel in 0.25 M HCl for 20 minutes. The gel was rinsed in H_2O and incubated in 0.4 M NaOH for 15 minutes.
2. Alkaline capillary blotting was carried out overnight using a tray in which a platform was built using a glass plate. A layer of Whatman 3MM paper was placed on the glass plate with the edges of the paper placed into a 0.4 M NaOH solution. The gel from step 1 was placed on top of this, followed by a piece of Hybond N+ on which the date, the wells and the side of the membrane that would contain the DNA were indicated in pencil. Two pieces of pre-wetted Whatman 3MM paper in 0.25 M NaOH and 1 piece of dry paper were placed on top of the filter, followed by 10 cm of tissues, a glass plate and a 0.5 kg weight. After addition of each layer a pipette was used to eliminate trapped air bubbles, Parafilm was used to avoid tissues coming into contact with the Whatman paper on the glass plate and the NaOH solution was covered with clingfilm to avoid evaporation.
3. After blotting, the blots were rinsed in 6x SSC, and stored in clingfilm at 4^{0}C or used immediately for hybridisation.

2.18.6 Dot blots

1. 6 µl DNA (Chapter 2.18.2) was incubated at room temperature for 15 minutes, after addition of 6 µl 2 M NaOH and 12 µl H_2O.
2. 5 µl of this solution was spotted onto Hybond N+ using a multi-channel pipette and dried by placing the membrane on a piece of Whatman 3MM paper.

3. The membranes were subsequently incubated on pre-wetted Whatman 3MM paper:
   -5 minutes with 0.4 M NaOH.
   -5 minutes with neutralising solution.
4. This was followed by submersion for 15 minutes in 50 mM NaPi pH 7.2.
5. After drying the filters on Whatman 3MM paper, the DNA was crosslinked to the membrane at 120 mJ/cm^2 in a crosslinker.

2.18.7 Labelling of DNA probes

1. 50 ng DNA was mixed with 5 µl 5x labelling buffer in a total volume of (24-X) µl.
2. Samples were incubated at 100^0C for 5 minutes and immediately placed on ice.
3. After a short centrifugation, X µl α-^{32}P dCTP and 1 µl (2U) of Klenow polymerase were added.
4. Samples were mixed and incubated at 37^0C for 2-4 hours.
5. Probes were precipitated by addition of 15 µl H_2O, 10 µl 0.2 M EDTA, 3 µl 10 mg/ml Yeast tRNA, 8 µl 1.6 M NaCl and 154 µl 100% ethanol, a 15 minute incubation on dry ice and centrifugation at 13800g for 10 minutes. The precipitate was resuspended in 110 µl H_2O, 1 µl was used to determine the incorporation of ^{32}P in counts per minute (cpm) in a β-counter.
6. The probes were incubated at 100^0C for 5 minutes and used for hybridisation.
7. Pre-association was performed when probes were likely to contain repeats. Following step 5, 17 µl of 1 M NaPi pH 6.8, and 10 µl 10.8 mg/ml total human placental DNA were added. This was followed by an incubation at 100^0C for 7 minutes and incubation at 65^0C for 2.5-5 hours after which they could be used for hybridisation. The probes were in the range of $5 \cdot 10^5$-$3 \cdot 10^6$ cpm per ml of hybridisation solution.

2.18.8 Hybridisation

1. Southern blots were pre-hybridised at 65^0C in 8 or 25 ml of (pre) hybridisation solution. Dot blots or total human BAC and PAC library filters were pre-hybridised in 8 or 25 ml Church solution.
2. Hybridisation was carried out overnight at 65^0C using fresh (pre-) hybridisation or Church solution (same amount as in the pre-hybridisation). Hybridisation was carried out overnight at 30^0C for the radio-actively labelled oligonucleotides.
3. Blots were washed in 2x SSC, 0.1% w/v SDS for 15 minutes at room temperature and subsequently for 30 minutes at 65^0C in 0.2x SSC, 0.1% SDS. Oligonucleotide hybridised blots were washed at room temperature for 15 minutes in 6x SSC and for 20 minutes in Church solution at 42^0C.
4. Autoradiography was performed using X-ray film, or results were visualised using phosphor-imager screens.

2.18.9 PAC/BAC insert isolation

1. Three individual digestions, each containing 3 µl DNA (Chapter 2.18.2) were restriction digested for 2-4h using 5 U *Not* I per digestion in a total volume of 15 µl.
2. Pulsed field gel electrophoresis (PFGE) was carried out using 1% w/v rapid agarose and 0.5x TBE electrophoresis buffer, at 5.2 V/cm for 14h with a switch time of 3-15 seconds.
3. Inserts were cut out of the gel using a sterile razor, after ethidium bromide staining and UV trans-illumination.
4. To 250 mg of gel slice 150 µl H_2O, 40 µl 1.0 M NaCl and 8 µl 0.5 M EDTA were added, after which the gel slice was dissolved at 68^{0}C for 15 minutes.
5. After 15 minutes incubation at 37^{0}C 2 µl 1 U/µl β-Agarase I was added, followed by overnight incubation at 37^{0}C.
6. Phenol and chloroform/IAA extractions were carried out. 500 µl phenol was added and the samples were vortexed for 1 minute. Samples were then centrifuged for 3 minutes at 13800g after which the aqueous phase was transferred to a new tube, and a chloroform/IAA extraction was carried out in a similar manner.
7. Insert DNA (in ~360 µl) was incubated overnight at -20^{0}C after addition of 5.4 µl 10 mg/ml dextran T40, 95 µl 1 M NaCl and 1062 µl 100% Ethanol.
8. DNA was precipitated by centrifugation at 13800g for 10 minutes. After removal of the supernatant, the pellet was washed with 0.5 ml 70% ethanol. The samples were then centrifuged for 5 minutes at 13800g.
9. The dried pellet was finally dissolved in 20 µl H_2O; 5 µl was used for the generation of a radio-actively labeled probe.

2.18.10 Preparation of competent cells

1. An individual colony from a freshly grown *E. coli* XL-1 blue plate was grown in 5 ml SOC medium for 14 hours at 37^{0}C.
2. 200 ml SOC medium was inoculated with 2 ml of the overnight grown culture and grown until the optical density (OD_{550nm}) reached 0.4-0.48.
3. The culture was transferred to 50 ml Falcon tubes and incubated on ice for 15 minutes.
4. Cells were subsequently precipitated by centrifugation, 10 minutes 1800g at 4^{0}C.
5. The cells from two Falcon tubes were resuspended in 15 ml TFB I and incubated on ice for 2 hours.
6. This was followed by centrifugation of the cells, 5 minutes 1800g at 4^{0}C.
7. Cells were resuspended in 4 ml TFB II, by gently pipetting through a tip from which the end was cut off.
8. 200 µl cells were added to eppendorf tubes, and immediately transferred to liquid nitrogen. Cells were stored at -80^{0}C.

2.18.11 Tissue Culture

1. A vial of frozen cells was quickly defrosted, and transferred to a 15 ml Falcon tube.
2. The cells were centrifuged for 5 minutes at 800g.
3. The pelleted cells were resuspended in 1 ml of appropriate medium; DMEM supplemented with 10% v/v foetal calf serum, 2% v/v pen/strep and 2% w/v glutamine for COS 7B, WA17 and A9.
4. The cells were transferred to a 75 cm^2 tissue culture flask containing 9 ml of pre-warmed medium.
5. They were incubated for 2 days at 37^0C at 5% or 10% CO_2.
6. After this period of time, the cells had formed a confluent monolayer, and were split as follows:
   -The old medium was removed and the monolayer was washed with 2 ml EDTA in saline.
   –2 ml 0.05% w/v trypsin in EDTA/saline was added, and removed.
   –1 ml 0.05% w/v trypsin in EDTA/saline was added and incubated for 3 minutes at 37^0C.
   -Cells were loosened by gentle tapping and checked under the microscope.
   -The cells were washed off the surface using 1 ml medium, and divided between 2 new flasks each containing 10 ml fresh pre-warmed medium.
7. Freezing of cells was carried out by addition of 1 ml medium supplemented with 10% (v/v) DMSO, to the trypsinated cells. Cells were transferred to sterile cryo-tubes, frozen overnight at -20^0C and then transferred to liquid nitrogen.
8. The lymphoblast cell lines, generated from patients with FEOAD grew in suspension and were diluted according to recommendations from the Coriell cell repository (AG08562A 1:2, AG07877A 1:4) in RPMI 1640 which was supplemented with 15% v/v heat inactivated Foetal calf serum and 2% v/v pen/strep. Cells were grown for 3-4 days at 37^0C and 5% CO_2. The cells were subsequently diluted in fresh medium.
9. The cells contained in 10 ml of a confluent culture were used for freezing and subsequent storage. They were first precipitated as indicated in step 2 and subsequently resuspended in RPMI 1640, supplemented with 30% v/v foetal calf serum, 2% v/v pen/strep and 5% v/v DMSO. The cells were frozen as indicated in step 7.

2.18.12 RACE, cloning of RACE products and analysis

1. Marathon-ready cDNA (Clontech), which was generated from bone marrow poly A+ RNA and to which linkers were ligated, was used in a PCR reaction with a gene-specific primer and a linker primer. The following samples were prepared: 2.5 µl Marathon-Ready cDNA (0.1 ng/µl), 0.5 µl (10 µM) AP1 primer (linker primer), 0.5 µl (10 µM) gene-specific primer (31B5-E21F or –E21R or USP 3' RACE), 0.5 µl 10 mM dNTP, 2.5 µl PCR buffer (Clontech), 0.5 µl Advantage 2 Polymerase mix (Clontech), 18 µl H_2O. Negative controls were included (Marathon-Ready cDNA substituted by H_2O) and positive controls were performed with G3PDH gene-specific primers. The PCR reactions were: 94^0C 1 minute, (94^0C 30s, 72^0C 4 minutes) 5x, (94^0C 30s, 70^0C 4 minutes) 5x, (94^0C 20s, 68^0C 4 minutes) 25x, 4^0C.

2. A PCR reaction was carried out with the nested linker primer AP2 and the gene-specific primers (31B5-E21F or –E21R or USP 3' RACE). Reactions and conditions were identical to step 1, except that the Marathon-Ready cDNA was substituted by 1 µl 1:100 diluted PCR product from step 1.
3. 5 µl of the gene-specific and control PCR reactions were loaded in wells of a 1.2% w/v agarose gel and electrophoresis took place for 1h at 100V after which the gel was incubated in ethidium bromide solution. A Polaroid photograph was taken and the PCR fragments were transferred to a Hybond N+ membrane via Southern blotting (Chapter 2.18.5).
4. A radio-active probe was prepared from the PCR product of exon trapped clone 31B5-E21 (generated with the primers EX1 and EX2, see Chapter 2.20.1 step 16). For the RACE products generated with the primer USP 3' RACE, a probe was radio-actively labelled that was prepared from the cDNA insert from the Image clone 307735. These probes were hybridised to the filters prepared in the previous step. Fragments that hybridised specifically to the 31B5-E21 exon trapped clone or to Image clone 307735 were isolated via Prep-a-Gene (31B5-E21F RACE and 31B5-E21R RACE, originated from AP2/31B5-E21F and AP2/31B5-E21R respectively) and cloned using the original TA cloning kit (InVitrogen) as follows: 3 µl of the PCR product (10 ng) was mixed with 0.5 µl 10x ligation buffer, 1 µl (12.5 ng) PCR 2.1 vector and 0.5 µl T4 DNA ligase (ligase, buffer and PCR 2.1 vector were present in the kit). Ligations were incubated overnight at 14^0C.
5. Transformations were carried out as described (Chapter 2.20.1 step 2). A blue/white screening was performed using X-Gal and IPTG.
6. White colonies were analysed via PCR by using 0.5 µl of boiled broth (Chapter 2.20.1 step 16) 1.5 pmol 31B5-E21F or –E21R or USP 3' RACE and 1.5 pmol of the M13F+ or M13R+ primers together with 1 µl 5 mM dNTPs, 0.5 U Taq Polymerase (Promega) and 2.5 µl Guys buffer in 25 µl total volume. The reactions were identical to Chapter 2.20.1 (step 17). PCR products were analysed by agarose gel electrophoresis.
7. Inserts of the expected size (comparable to the RACE products) were sequenced with the appropriate primers. A PCR reaction was performed using 0.5 µl of boiled broth, with 31B5-E21F or UG1-F1 and M13R+ (for 31B5-E21F RACE) and 31B5-E21R or UG1-R1 or UG1-R2 or UG1-R3 and M13R+. PCR reactions were identical to the ones described (Chapter 2.20.1 step 17). Reactions were as follows: 95^0C 2 minutes (95^0C 30s, 56^0C 30s, 72^0C 45s) 35x, 72^0C 9 minutes, 4^0C. Sequencing was performed as described (Chapter 2.20.1 step 17), with primers 31B5-E21F and UG1-F1 (for 31B5-E21F RACE), 31B5-E21R, UG1-R1, UG1-R2 and UG1-R3. The clones containing the RACE products from the 3' end of USP 21 were analysed by a PCR reaction using the primers USP 3' RACE and M13F+. The clones that amplified a specific product were sequenced using the primers USP 3' RACE and USPE-24A. Reactions were identical to those for the 31B5 RACE products.

2.18.13 DNA isolation from lymphoblast cell lines from familial early onset Alzheimer's disease patients

1. The lymphoblast cell lines were grown as described in Chapter 2.18.11.
2. The cells of a 25 ml culture were precipitated by centrifugation at 800g for 5 minutes, resuspended in 20 ml PBS, reprecipitated by centrifugation and resuspended in 0.9 ml STE.
3. This suspension was transferred to a 2 ml tube after which 10 µl H_2O, 10 µl Proteinase K (20 mg/ml) and 50 µl 10% SDS were added. Samples were incubated at 37^{0}C overnight.
4. The samples were incubated for a further 15 minutes at 37^{0}C, after addition of 5 µl (100 mg/ml) RNAse (Qiagen).
5. 1 ml of phenol was added and samples were incubated in a slow wheel for 10 minutes. This was followed by centrifugation for 5 minutes at 13800g, after which the aqueous phase was transferred to a new tube.
6. A total of two additional extractions were carried out by addition of 1 ml phenol/chloroform/isoamylalcohol (25:24:1) to the samples. The incubation and subsequent centrifugation were identical to step 5.
7. To the aqueous phase 0.1 volume of 3 M NaAC pH 5.2 and 2.5 volumes of 100% ethanol were added, after which the samples were incubated in a slow wheel to precipitate the DNA.
8. The DNA was spooled on a glass rod, washed in 70% ethanol and dissolved in 500 µl H_2O.

2.19 Protocols specific for Chapter 3

2.19.1 Inverse PCR (IPCR)

1. 1 µl DNA (~300ng) (Chapter 2.18.2) was digested with either *Bfa* I or *Nla* III (PACs), *Hae* II or *Msp* I (BACs) for 2 h.
2. Phenol and chloroform/IAA extractions and subsequent precipitation of the DNA were carried out (Chapter 2.18.2 steps 9/11).
3. Overnight ligation was performed at 14^{0}C in 100 µl total volume with 10 µl 10x ligation buffer and 1 U T4 DNA ligase.
4. PCR was carried out using standard quantities of dNTPs, Taq Polymerase (HT Biotechnology), primers (Chapter 2.18.3) and 10 µl of ligation mixture, with the primers IPCR 1, 2, 3 and 4 for PAC clones and IPCR B1, B2, B3 and B4 for BAC clones. Primers IPCR 1 and 2 were used for *Bfa* I, IPCR 1 and 2 or 3 and 4 for the *Nla* III, IPCR B1 and B2 for the *Msp* I, and B3 and B4 for the *Hae* II digested samples. PCR programme used: 1 minute 95^{0}C, (1 minute 95^{0}C, 1 minute 60^{0}C, 2 minutes 72^{0}C) 35x, 8 minutes 72^{0}C for IPCR with primers 1,2 and 3,4. For the PCRs with B1, B2 and B3, B4 55^{0}C was used as the annealing temperature. The PCR reactions were identical to the PAC IPCRs.
5. PCR samples were resolved on a 1 % w/v agarose gel.
6. Products were isolated using Prep-A-Gene or Electro-elution (Chapter 2.18.4).

2.19.2 Vectorette PCR

1. 1 µl DNA (~300 ng) (Chapter 2.18.2) was digested with either *Hind* III or *Pst* I for 2h.
2. Phenol and chloroform/IAA extractions and subsequent precipitation of the DNA were carried out (Chapter 2.18.2 steps 9/11).
3. Overnight ligation was carried out at 14^{0}C in 100 µl total volume using 10 µl 10x ligation buffer, 1 U T4 DNA ligase and 100 ng vectorette linker; *Pst* I linker for the *Pst* I digested samples and *Hind* III linker for the *Hind* III digested samples.
4. PCR was carried out using standard quantities of dNTPs, Taq Polymerase (HT Biotechnology), primers (Chapter 2.18.3), and 10 µl of the ligation mixture, with the primers V1 and either IPCR 1 or 3 (PACs), V1 and either IPCR B1 or B3 (BACs) or V1 and either LAW4L or LAW4R (cosmids). PCR programme used: 1 minute 95^{0}C, (1 minute 95^{0}C, 1 minute 60^{0}C, 2 minutes 72^{0}C) 35x, 8 minutes 72^{0}C. For the PCRs including B1 and B2, an annealing temperature was used of 55^{0}C, the PCR reaction was otherwise identical.
5. PCR samples were resolved on a 1% (w/v) agarose gel.
6. Products were isolated using Prep-A-Gene or Electro-elution (Chapter 2.18.4).
7. The linkers for the vectorette PCR were generated in the following way: 10 µl (1 µg/µl) of the top strand oligonucleotide was mixed with 2 µl 10x kinase buffer, 1.5 µl 8 mM ATP, 1 µl (1 U/µl) T4 polynucleotide kinase in a 20 µl total volume. This sample was incubated for 1 hour at 37^{0}C, followed by an incubation at 90^{0}C for 10 minutes. The sample was quickly cooled on ice, 2 µl 1M NaCl was added together with 10.7 µl (1 µg/µl) of the bottom strand oligonucleotide (the quantities of top and bottom strand were calculated for the *Pst* I linker, these amounts result in equimolar ratios). H_2O was added to a final volume of 40 µl. This mixture was incubated at 65^{0}C for 5 minutes after wich the sample was slowly cooled to 37^{0}C over a two hour period. Finally, the linker was incubated for 10 minutes at room temperature and 160 µl 50 mM NaCl in TE was added.

2.19.3 Generation of end fragments via the degenerate oligonucleotide primer method (Wu *et al.*, 1996)

1. To 1 µl of 1:100 diluted PAC DNA (Chapter 2.18.2), 7.5 pmol of 6MW primer, 0.5 µl 2.5 mM dNTPs, 1 µl 5x sequencing buffer and 0.5 µl H_2O were added.
2. The samples were incubated for 3 minutes at 96^{0}C followed by 5 minutes at 30^{0}C, during which 1 µl 0.8 U sequenase 2.0 was added, followed by incubation for 10 seconds (s) at 31^{0}C, 10s at 32^{0}C, 10s at 33^{0}C, 10s at 34^{0}C, 10s at 35^{0}C, 10s at 36^{0}C, 3 minutes at 37^{0}C and 10 minutes at 72^{0}C.
3. During the incubation at 72^{0}C, 45 µl of pre-warmed PCR solution was added, consisting of: 5 µl 10x reaction buffer, 12.5 pmol 6MW primer, 25 pmol primer 1 or 3 (IPCR primers), 4 µl 2.5 mM dNTPs and 1 U Taq Polymerase (HT Biotechnology). PCR reactions were as follows: (95^{0}C 1 minute, 58^{0}C 1 minute, 72^{0}C 2 minutes) 35x, 72^{0}C 5 minutes, 4^{0}C.
4. Samples were analysed on 1% w/v agarose gels and fragments were isolated via Prep-a-Gene or electro-elution.

2.19.4 Isolation of *Sau* 3A I fragments from cosmid inserts

1. 3 µl cosmid DNA (Chapter 2.18.2) was digested with 1 µl *Sau* 3A I in a total reaction volume of 20 µl for 2h.
2. The *Sau* 3A I fragments were separated by 1% w/v agarose gel electrophoresis. The gel was incubated in an ethidium bromide solution and fragments bigger than 1.1 Kbp were isolated from the gel using Prep-A-Gene. The biggest vector derived *Sau* 3A I fragment was 952 bp.

2.19.5 Restriction mapping

1. 1.5 µl DNA (Chapter 2.18.2) was restriction digested individually with 5 U *Not* I, *Sal* I, *Xho* I, *Sac* II, *Mlu* I, *BssH* II or *Eag* I, or with a combination of these enzymes. Digestions were carried out for 4-14 hours in 15 µl total volume. Digestions including *BssH* II were carried out first at 50^0C after which the other enzymes were added followed by incubation at 37^0C.
2. PFGE was carried out, using a 1% w/v rapid agarose gel, and 0.5x TBE electrophoresis buffer for 14h, at 5.2 V/cm with a switchtime of 3-15 seconds.
3. Polaroid photos were taken after ethidium bromide staining.
4. Subsequent Southern blotting was carried out for the appropriate gels (Chapter 2.18.5).

2.19.6 Isolation of genomic DNA for PFGE

1. WA17 and A9 cells were grown as described (Chapter 2.18.11). Trypsinated cells were centrifuged for 5 minutes at 800g, washed in 2 ml PBS and reprecipitated by centrifugation at 800g for 5 minutes. The cells were subsequently resuspended in 1.0 ml PBS.
2. An aliquot was diluted and the cells were counted using a haemocytometer. The cells were then diluted to a concentration of $6x10^7$/ml in PBS.
3. The cells were equilibrated at 37^0C and mixed with an equal volume of 2% w/v low melting point agarose also equilibrated at 37^0C. This solution was used to fill plugmoulds which were subsequently placed on ice for 15 minutes.
4. Agarose blocks containing the cells were placed in a 50 ml Falcon tube containing 20 ml NDS supplemented with 1 mg/ml pronase and incubated at 50^0C for 24 hours.
5. Two further 1 hour incubations were carried out using fresh NDS and blocks were stored at 4^0C or restriction enzyme digestion was carried out.
6. Blocks were incubated in 20 ml T0.1E at 50^0C for 3x30 minutes, and for 30 minutes at 37^0C in the appropriate 1x restriction buffer.
7. A block was digested in 250 µl solution, containing 1x reaction buffer, 1 mM spermidine, 1 mM DTT, 100 µg/ml BSA and 40 U enzyme (*Mlu* I, *Bam* H I, *Sma* I, *Xho* I or *Bst* E II). Digestions were carried out overnight at the appropriate temperatures.

8. Approximately 25% of a block (approximately $7.5x10^6$ cells) was used per lane for PFGE, using a 1% w/v rapid agarose gel in 0.5x TBE electrophoresis buffer. Electrophoresis conditions were: (5-250 Kbp) 14h, 5.2 V/cm and switch time 3-15 seconds, (100-1600 Kbp) 22h, 6.0 V/cm, switch time 40-80 seconds.

2.20 Protocols specific for Chapter 4

2.20.1 Exon trapping

1. -Insert preparation, *Sau* 3A I partial digestion:

   five individual digestions of 3 µl (~ 1 µg) genomic insert containing bacterial clone (Chapter 2.18.2) were digested with 0.05 U *Sau* 3A I, in 15 µl total reaction volume for 1 hour. Digestions were pooled and the DNA precipitated (Chapter 2.18.2 step 10/11). The DNA was resuspended in 20 µl H_2O and loaded in one well of a 0.7% w/v agarose gel. Electrophoresis was performed for 1 hour at 100 V, 50 mA. The lane containing the DNA marker was incubated in an ethidium bromide solution and a Polaroid photo was taken of the gel slice with a ruler for size estimation. A UV source was used to illuminate the gel slice. Subsequently, agarose slices (containing the *Sau* 3A I digested samples) were cut out of the remaining part of the gel. These slices contained DNA in the range of (approximately) 1-5 Kbp. DNA was isolated using Prep-A-Gene and 5 µl was used to estimate concentration by agarose gel electrophoresis.

   -Insert preparation, *Bam* H I/*Bgl* II digestion:

   10 µl DNA (~3 µg) (Chapter 2.18.2) was digested with 2 µl of each enzyme, in a total reaction volume of 50 µl. for 2 hours. Phenol and chloroform/IAA extraction followed by precipitation of the DNA was carried out as described (Chapter 2.18.2 step 9/11). The DNA was resuspended in 20 µl H_2O and 1 µl was used to estimate concentration by agarose gel electrophoresis.

   -pSPL3B preparation (eukaryotic expression vector):

   10 µl (~3 µg) pSPL3B, isolated using Wizard DNA purification resin, was digested with 2 µl *Bam* H I in 50 µl total reaction volume, for 2 hours. Phenol and chloroform/IAA extraction followed by precipitation of the DNA was carried out as described (Chapter 2.18.2 step 9/11). DNA was incubated for 30 minutes with 2.5 U alkaline phosphatase, in 50 µl total reaction volume, at 37^0C. Two phenol extractions, followed by a chloroform/IAA extraction and precipitation of the DNA were carried out (Chapter 2.18.2 step 9/11). The DNA was finally resuspended in 50 µl H_2O and 1 µl was used to estimate concentration by agarose gel electrophoresis.

2. 1 µl of pSPL3B vector (~20 ng), previously digested and treated with alkaline phosphatase, was mixed with 1 µl (~100 ng) *Bam* H I/*Bgl* II digested samples or 3-6.5 µl (~200 ng) of the *Sau* 3A I partially digested samples, with 0.5 µl T4 DNA ligase in 10 µl total volume for 5 hours at room temperature. Ligation reactions were mixed with 100 µl competent *E. coli* XL-1 blue cells (Chapter 2.18.10), and placed on ice for 30 minutes. Negative and positive controls were performed, in which no ligase, ligase or ligase and polynucleotide kinase were added to the previously treated pSPL3B vector. The samples were incubated at 42^0C for 1.5 minutes then immediately transferred to ice. 1 ml

LB was added, and the cells were incubated at 37^{0}C for 1 hour with occasional mixing. Cells were plated on LB plates containing 50 µg/ml ampicillin.

3. Colonies containing the pSPL3B ligated samples were resuspended by adding 2 ml LB containing 50 µg/ml ampicillin to the plates. A sterile cell scraper was used to resuspend the cells. The LB was pipetted from the plates into a flask containing 30 ml LB and 50 µg/ml ampicillin, and were grown overnight at 37^{0}C, in a horizontal shaker at 200 rpm.
4. DNA was isolated using Qiagen tip 100 columns. An additional DNA precipitation was carried out (Chapter 2.18.2 steps 10/11) and DNA was resuspended in 30 µl H_2O, concentration being determined by spectrophotometric measurement (λ=260nm).
5. COS 7B cells were grown as described (Chapter 2.18.11); one flask of cells was used (per transfection). Trypsinated cells were centrifuged for 5 minutes at 800g, washed in 2 ml PBS and reprecipitated by centrifugation at 800g for 5 minutes. The cells were resuspended in 0.7 ml cold PBS and added to 15 µg DNA in 100 µl PBS in a pre-cooled cuvette. After a 10 minute incubation on ice, the samples were electroporated using a Bio-Rad electroporator. Electroporation conditions were: 1.2 kV, 25 µF, 100 mΩ for 0.5 seconds. The cells were added to a tissue culture flask containing 10 ml fresh DMEM and placed in a CO_2 incubator for 48 hours.
6. Fresh medium was added to the cells after 24 hours.
7. Cells were trypsinated and centrifuged at 800g for 5 minutes, and were resuspended in 100 µl RNAzol B, for RNA isolation. 10 µl chloroform was added, samples were incubated on ice for 5 minutes, then centrifuged at 17860g for 15 minutes at 4^{0}C. The aqueous layer was transferred to a new tube and an equal volume of isopropanol was added, mixed, incubated on ice for 15 minutes and centrifuged at 17860g for 15 minutes at 4^{0}C. Precipitated RNA was washed with 0.5 ml 80% v/v ethanol and centrifuged for 8 minutes at 14000g at 4^{0}C. The dried pellet was resuspended in 10 µl H_2O.
8. 5 µl RNA was mixed on ice with 5 µl 5 mM dNTPs, 1 µl RNAse inhibitor, 4 µl 5x reverse transcriptase (RT) buffer and 0.5 µl SA2 primer (7.2 pmol), in a 20 µl reaction volume. Samples were incubated at 65^{0}C for 10 minutes, transferred to ice for 2 minutes, and 1 µl M-MLV RT was added, after which the samples were incubated at 42^{0}C for 1.5 hours.
9. A primary PCR was carried out, using 10 µl of the cDNA from step 8, 5 µl of Guys buffer, 4 µl 2.5 mM dNTPs, 18 pmol of SD6 and SA2 primers and 5 U Taq Polymerase (Promega), in a 50 µl reaction volume. The PCR programme used was: (1 minute 95^{0}C, 1 minute 55^{0}C, 5 minutes 72^{0}C) 6x. 2 µl (20 U) *Bst* X I was added to the samples and incubated overnight at 55^{0}C.
10. 0.5 µl *Bst* X I was added and the samples were incubated for a further 2 hours at 55^{0}C. A secondary PCR was carried out in which 5 µl of the *Bst* X I digested sample was mixed with 4.5 µl Guys buffer, 2 µl 5 mM dNTPs and 20 pmol of each of the primers dUSA 4 and dUSD 2 in 50 µl reaction volume. The PCR conditions were: 5 minutes 95^{0}C, 3 minutes 80^{0}C during which 2.5 U of Taq Polymerase (Promega) was added, followed by (1 minute 94^{0}C, 1 minute 60^{0}C, 3 minutes 72^{0}C) 30x, 4^{0}C.
11. 5 µl was analysed on a 2% w/v agarose gel. 20 µl was mixed with 80 µl H_2O and DNA was isolated using Prep-A-Gene, which was performed to remove the primers from the preparation.

12. The concentration of DNA was estimated on a 2% w/v agarose gel, and approximately 100 ng was used to transform to *E. coli* XL1-blue cells using pAMP 1.
13. Equal amounts of colonies were picked into 384 well plates per transfection (both for *Sau* 3A I and *Bam* H I/*Bgl* II cloning).
14. Spotting and processing of membranes was carried out as described (Chapter 2.18.1).
15. Hybridisations to these filters, and to dot blots of the bacterial clones the exon trapped clones originated from, were carried out as described (Chapter 2.18.8).
16. Probes that were hybridised to these filters were derived in the following way (Chapter 2.18.7):

    -Genomic insert from the bacterial clones (Chapter 2.18.9).

    -COT I DNA (Life Technologies) was used, 50 ng per labelling.

    -HIV *tat* intron, pSPL3B was digested with *Sal* I/*Hind* III for 2 hours at 37^{0}C, after which agarose gel electrophoresis was carried out. The 1.9 Kbp fragment, containing ~90% of the intron, was isolated from the agarose gel using Prep-A-Gene. Approximately 50 ng was used per labelling.

    -3^0 BB EX1/EX2 was generated via PCR using ~100 pg of RT PCR product from step 10, after transfection with the *Bam* H I/*Bgl* II digested samples. The PCR reaction was carried out as in step 10, with 25 pmol of primer EX1 and EX2. PCR programme: 1.5 minutes 95 ^{0}C (30s 95^{0}C, 30s 59^{0}C, 1 minute 72^{0}C) 33x, 9 minutes 72^{0}C. The PCR product was isolated as described in step 11.

    -3^0 S EX1/EX2 was generated as above, after transfection of genomic clones that had been restriction digested with *Sau* 3A I.

    -Probes from individual exon trapped products were generated as follows. A clone from an individual well from a microtitre dish was grown overnight on an LB plate containing 50 µg/ml ampicillin. A single colony from this plate was inoculated in 2 ml LB broth supplemented with 50 µg/ml ampicillin. 10 µl from this culture was mixed with 10 µl H_2O and boiled for 5 minutes. Subsequently 0.5 µl of this mixture was used in a PCR reaction containing 12.5 pmol EX1 and EX2 primers, 1 µl 5 mM dNTPs, 0.5 U Taq Polymerase (Promega), 2.5 µl Guys buffer in 25 µl total volume. The PCR programme was identical to the one described for the generation of PCR products for 3^0 BB EX1/EX2. The PCR products were isolated as described in step 10, and their concentration was checked via agarose gel electrophoresis.

    -cDNA clones from database searches were requested from the Human Genome Mapping Project Resource Centre and were used as probes. They were grown overnight on LB plates supplemented with 50 µg/ml ampicillin. An individual colony was inoculated in 5 ml LB containing 50 µg/ml ampicillin and grown overnight at 37^{0}C. DNA was isolated as described (Chapter 2.18.2). Insert was separated from the vector using the appropriate restriction endonucleases as instructed from the information on cDNA library construction. The cDNA insert was isolated using agarose gel electrophoresis and the Prep-A-Gene isolation method.
17. Exon trapped clones were sequenced using an ABI 310 automated sequencer. 0.5 µl boiled broth from step 16 was used in a PCR reaction with 1.5 pmol of primer M13F+ and M13R+, 1 µl 5 mM dNTPs, 0.5 U Taq Polymerase (Promega), 2.5 µl Guys buffer in 25 µl total volume. Reactions were as follows: 95^{0}C 1 minute 15s (95^{0}C 45s, 55^{0}C 45s, 72^{0}C 1 minute 30s) 34x, 72^{0}C 5 minutes, 4^{0}C.

PCR products were precipitated by addition of 55 µl H_2O, 8 µl 3 M NaAc pH 5.2 and 160 µl ethanol. After a 20 minutes incubation at -20^0C, DNA was precipitated by centrifugation for 20 minutes at 13800g, the supernatant was removed and pellets were washed twice with 70% v/v ethanol, the samples being centrifuged each time for 5 minutes at 13800g. Finally, pellets were dried in a heating block and resuspended in 20 µl H_2O. Concentrations of the PCR products were determined by agarose gel electrophoresis. Approximately 50 ng of the PCR product was added to 3.2 pmol primer (M13F+ or M13R+), 4 µl Big Dye cycle sequencing mix and 4 µl cycle sequencing buffer, in a total volume of 20 µl. Cycle sequencing was carried out in a Perkin Elmer 2400 PCR machine using the following programme (96 ^{0}C 10s, 50 ^{0}C 5s, 60 ^{0}C 4 minutes) 25x. The ABI sequencer was set up according to manufacturer's recommendations, using POP 6 polymer and a 47 cm capillary. Analysis of the sequence was automatic, although a manual check was also performed.

18. Database searches were performed at The National Center for Biotechnology Information, http://ncbi.nlm.nih.gov/. Open reading frames (ORFs) were analysed by using the 'ORF finder' option at this website. The sequence ATGGATGGATGG was added to the forward frame (5'-3') and TACCTACCTACC to the reverse frame (5'-3'). ORFs were directly compared to the SwissProt protein database. From the http://ncbi.nlm.nih.gov/ webpage the option 'Blast Sequence Similarity Searching' (Altschul *et al.*, 1997) was chosen (http://ncbi.nlm.nih.gov/BLAST/). The Blast searches were performed in basic Blast search 2.0, using the following databases: (nucleotide), nr- All non-redundant Genbank+EMBL+DDBJ+PDB sequences (but no EST, STS, GSS), dbEST- non-redundant Database of Genbank+EMBL+DDBJ EST division, the mouse EST database and the other EST databases (EST that are not human or mouse).

2.21 Protocols specific for Chapter 5

2.21.1 Cloning of USP21

1. Cloning of USP21 RT-PCR was performed using bone marrow Poly A+ RNA (Clontech). The RT step was performed using 4 µl RNA (200 ng), 5 µl 5 mM dNTPs, 1 µl RNAse inhibitor, 7 µl 5x RT buffer and 1 µl hexanucleotides (3 µg/µl) in 33 µl total volume. Samples were incubated at 65^0C for 10 minutes, transferred to ice for 2 minutes, and 2 µl M-MLV RT was added, after which the samples were incubated at 42^0C for 1.5 hours. Negative controls were performed to which H_2O instead of RT was added.
2. A subsequent PCR was carried out using 2 µl from the reverse transcribed bone marrow sample or from the negative control. PCR reactions were performed as described in Chapter 2.21.5 step 1. The PCR programme was: 95^0C 1 minute (95^0C 1 minute, 57^0C 1 minute, 72^0C 2 minutes) 33x, 72^0C 10 minutes, 4^0C. The following primer pairs were used, USPEXP1 and USP-ER, USP-EF and USP-XR, USP-XF and USPEXP2. The primers USPEXP1 and USPEXP2 contained a *Sac* I and a *Sal* I restriction endonuclease site respectively. The primers USP-ER, USP-EF and USP-XR, USP-XF contained an *Eco* R I and an *Xba* I site respectively which are present in the USP21 gene at nucleotide position 1255 (*Eco* R I) and 2419 (*Xba* I).

3. Agarose gel electrophoresis was carried out and the PCR products were isolated via Prep-A-Gene. The RT-PCR products were subsequently digested in 20 µl total volume using 10 U per enzyme in the appropriate 1x reaction buffer, with *Sac* I/*Eco* R I (USPEXP1/USP-ER product), *Eco* R I/*Xba* I (USP-EF/USP-XR product) and *Xba* I/*Sal* I (USP-XF/USPEXP2 product), to create sticky ends for ligation. pBluescript (~ 1 µg), isolated using Wizard DNA purification resin, was digested with *Sac* I/*Eco* R I, *Eco* R I/*Xba* I or *Xba* I/*Sal* I (10 U of each restriction enzyme) in 20 µl total volume, for 2 hours. Phenol and chloroform/IAA extraction followed by precipitation of the DNA was carried out as described (Chapter 2.18.2 steps 9/11).
4. Ligations were carried out overnight at 14^0C in 10 µl total volume with 1 µl (20 ng) digested pBluescript, 1 µl 10x T4 DNA ligase buffer, 1 µl (20 ng) digested RT-PCR product and 1 µl T4 DNA Ligase (1 U/µl).
5. Transformations were carried out as described in Chapter 2.20.1 step 2. Transformants were analysed by blue/white screening using 24 µg/ml IPTG and 20 µg/ml Xgal in the LB plates. DNA was isolated from white colonies (Chapter 2.18.2), and the constructs were analysed by restricyion digestion with the endonuclease sites used for cloning. This resulted in the constructs USP-A (*Sac* I/*Eco* R I fragment), USP-B (*Eco* R I/*Xba* I fragment) and USP-C (*Xba* I/*Sal* I fragment).
6. The three different constructs were analysed by sequencing for point mutations caused by PCR amplification. The primer combinations for analysing the different constructs were M13R+/USP-1R, USP-2F/USP-2R and USP-3F/M13F+ for USP-A, M13F+/USP-5F, USP-4R/USP-4F and USP-3R/M13F+ for USP-B, M13R+/USP-6R and USP-7F/M13F+ for USP-C. PCR reactions were identical to the ones described in Chapter 2.20.1 step 17. The PCR programme was: 95^0C 1 minute (95^0C 1 minute. 56^0C 1 minute, 72^0C 2 minutes) 35x, 72^0C 10 minutes, 4^0C. PCR products were precipitated as described in Chapter 2.20.1 step 17. Subsequent cycle sequencing reactions were carried out with the above mentioned primers for the respective PCR products. PCR conditions for cycle sequencing, precipitation of the products and sequencing were also carried out as described in Chapter 2.20.1 step 17.
7. The *Eco* R I/*Xba* I fragment of USP-B and the *Xba* I/*Sal* I fragment of USP-C were isolated from agarose gel and co-ligated (20 ng of each insert) in the *Sal* I/*Eco* R I sites of pBluescript. Ligation, transformation and analysis of clones was carried out as indicated in steps 4/5 of this section. This resulted in the construct USP-BC.
8. The *Sac* I/*Eco* R I fragment of USP-A was isolated from agarose gel via Prep-a-Gene and ligated in the *Sac* I/*Eco* R I sites of USP-BC, resulting in the construct USP-ABC; the USP21 gene from nucleotide 203-3367 was cloned into pBluescript.
9. The *Sac* I/*Sal* I insert from USP-ABC was eventually cloned into the *Sac* I/*Sal* I restriction sites of the prokaryotic expression vector pQE30 (Qiagen) (pQE-USP21), which places the construct behind an in-frame ATG, under the control of the IPTG inducable T5 promoter.

2.21.2 Cleavage analysis of ubiquitin-Met-β-gal fusion product by USP21

1. *E. coli* XL1-blue cells were transformed (Chapter 2.20.1 step 2) with 50 pg of the following constructs: pACYC-Ub-Met-β-gal (contains ubiquitin-β-galactosidase fusion gene, Baker *et al.*, 1992) co-transformed with pQE30, pQE-USP21 or pRB105 (positive control, containing the UBP2 ubiquitin-specific protease from *S. Cerevisiae*, Baker *et al.*, 1992). Transformants were selected with the appropriate antibiotics (pACYC-Ub-Met-β-gal 42 μg/ml chloramphenicol, the co-transformants 42 μg/ml chloramphenicol and 75 μg/ml carbenicillin) and were analysed by DNA isolation and restriction endonuclease analysis.
2. Overnight cultures were grown in LB with the antibiotics as described in step 1, inoculated from a single colony from XL1-blue, containing pACYC-Ub-Met-β-gal/pQE30, pACYC-Ub-Met-β-gal/pQE-USP21 or pACYC-Ub-Met-β-gal/pRB105. 10 ml fresh LB with antibiotics were inoculated with the overnight cultures to an OD_{600} of 0.1, after which the cultures were grown to an OD_{600} of 1. The cultures were split and to one half of the culture IPTG was added to a final concentration of 0.2 mM. All cultures were incubated at 37^0C for 90 minutes with continuous shaking.
3. Protein fractions were prepared from the induced/uninduced cultures by disruption of the cell membrane of the *E. coli* cells by using a glass potter. Protein fractions were frozen and stored at -20^0C.
4. A Folin-Lowry assay was performed in duplicate to quantify the different protein fractions. The protein in 10 μl of the homogenate was precipitated by addition of 500 μl 10% trichloroacetic acid, mixing and centrifugation for 5 minutes at 13800g. The pellet was air dried at room temperature, after which it was resuspended in 500 μl solution D. The samples were incubated for 10 minutes at room temperature, after which 50 μl Folin/H_2O solution (1:1) was added. The samples were incubated for 30 minutes at room temperature followed by addition of 500 μl H_2O and the OD_{750} was measured from the different protein fractions. Standard BSA solutions of known concentration were measured to construct a calibration curve for quantification.
5. Total protein from the different homogenates was extracted with methanol/chloroform to concentrate and purify the sample before SDS gel electrophoresis and western blotting. 10 μg total protein in a volume of 100 μl was mixed with 400 μl methanol and centrifuged briefly up to 10000g 100 μl chloroform was added, vortexed and again centrifuged up to 10000g, 300 μl H_2O was added, mixed by vortexing and centrifuged for 1 minute at 13000g. Most of the supernatant was removed and 100 μl methanol was added, after which the proteins were precipitated by centrifugation for 4 minutes at 13000g. The supernatant was removed and the pellet was dried in a vacuum desiccator. The protein pellets were resuspended in 4 μl 3x sample buffer, 2 μl 100 mM DTT and 6 μl H_2O, and boiled for 5 minutes in a water bath.
6. Electrophoresis of the protein fractions was performed using a pre-made 7% resolving acrylamide gel, which was made by mixing 31 ml H_2O with 15 ml gel running buffer, 14 ml 30% w/v acrylamide solution and 30 μl TEMED. The solution was degassed for 30 minutes in a vacuum, 300 μl 10% w/v ammonium persulphate was added and the solution was poured in a gel tray containing 11 gel plates. The gel plates were removed after polymerisation for an hour and stored in 1:4 diluted gel running

buffer at 4^0C. A stacking gel was prepared prior to electrophoresis by mixing 2.3 ml H_2O, 1 ml stacking gel buffer, 0.65 ml 30% w/v acrylamide solution and 5 µl TEMED. This solution was degassed for 15 minutes in a vacuum and 80 µl 10% w/v ammonium persulphate was added, after which the solution was loaded onto the top of the resolving gel. A comb was inserted and the gel was left for 10 minutes to polymerise.

7. Protein samples from step 5 were loaded using a Hamilton syringe into individual wells of the gel described in step 6. Prestained protein molecular weight standards were used for size markers. Electrophoresis was carried out at 10 mA for 10 minutes and at 15 mA for approximately 1.5 hours in electrophoresis buffer until the bromophenol blue band had migrated to the bottom of the gel.
8. The protein was transferred to nitrocellulose as follows: The nitrocellulose was soaked in transfer buffer and a sandwich consisting of a layer of foam, 2 sheets of blotting paper, the pre-soaked nitrocellulose, the acrylamide gel, 2 more sheets of blotting paper and a further layer of foam was assembled in a transfer cassette. This cassette was submerged in transfer buffer and proteins were transferred to nitrocellulose using a Hoefer Transfer Electrophoresis unit at 150-200 mA for 2.5 hours.
9. The transfer cassette was disassembled and the nitrocellulose membrane was incubated overnight at 4^0C in blocking buffer, to block non-specific binding sites for the antibodies. The membrane was washed in PBS and incubated at room temperature for 1 hour in 2.5% w/v milk powder in PBS to which the anti-β-galactosidase rabbit serum (r12741, R.D. Everett) was added at a final concentration of 1/20.000. The membrane was subsequently washed (four times 10 minutes) in washing buffer. The nitrocellulose membrane was then incubated with the secondary antibody, anti-rabbit IgG-horseradish peroxidase (HRP) diluted 1:1000 in 2.5% w/v milk powder in PBS. The membrane was again washed (four times 10 minutes) in washing buffer.
10. For detection of specific antibody binding, an enhanced-chemiluminescence assay kit was used (Amersham). The kit contains a chemiluminescence agent that is activated in the presence of antibody linked HRP and H_2O_2. The membrane was removed from the PBS and was incubated for 5 minutes in a 40:1 v/v mixture of solution A and B of the kit, as recommended by the manufacturer. The membrane was then exposed for different times (1-5 minutes) to Hyperfilm^{TM}ECL film, in a developing cassette.
11. Films were developed by submerging in Kodak developer for approximately 2 minutes, rinsed in H_2O, and submerged in Kodak fixer. The film was rinsed in H_2O and left to air dry.

2.21.3 Genescan using marker D21S366

1. A PCR reaction was performed with the primers D21S366-F and D21S366-R. The primer D21S366-R contained a 5' fluorescent label, 6-FAM. PCR reactions were set up as described in Chapter 2.21.5 step 1, the PCR programme was: 95^0C 1 minute 30s (95^0C 30s, 56^0C 30s, 72^0C 45s) 35x, 72^0C 9 minutes, 4^0C.

2. PCR products were analysed by agarose gel electrophoresis. 12 µl deionised formamide and 0.5 µl Genescan 350 size standards (Perkin Elmer) were added (to 1 µl 1:20 diluted PCR reaction), mixed and heated at 95^{0}C for 3 minutes.
3. Electrophoresis of the samples was performed using an ABI 310 Genetic Analyser. The capillary size was 47 cm, the polymer POP 4 and the programme module was GS STR POP4 C. Individual runs lasted approximately 30 minutes; electrophoresis of the samples was performed in triplicate. Data was analysed using Genescan software (Perkin Elmer).

2.21.4 Hybridisation of 11-mer primers to determine loss of heterozygosity of USP21

1. PCR was performed on:

   -genomic DNA, isolated from normal and tumour tissue from patient #61. Primers used were His 1 and His 2, amplifying the T/A polymorphism at nucleotide position 2029 of the USP21 gene.

   -genomic DNA, isolated from normal tissue, from patient #9, using the primers His 1 and His 2, this product contained an A at nucleotide position 2029 of the USP21 gene.

   -vector DNA, from construct USP-B, with the primers M13R+ and USP-5F. This product contained a T at position 2029 of the USP21 gene.

   PCRs were performed as described in Chapter 2.21.5 for the His 1 and His 2 primers, and as in Chapter 2.21.1 for the M13R+ and UBH5F primers.
2. Agarose gel electrophoresis was carried out with these PCR products, using a 2% w/v agarose gel. Two identical panels on the same gel were created by loading similar quantities of the four products twice. The gel was incubated in ethidium bromide, a Polaroid picture was taken and a Southern blot was prepared as described in Chapter 2.18.5.
3. The blotting set-up was dismantled, the blot rinsed in 6x SSC and the DNA was cross-linked to the filter at 120 mJ/cm^2. The two parts of the blot were hybridised individually to the oligonucleotides 2029A and 2029T, which were radio-actively labelled via γ-^{32}P-dATP and T4 polynucleotide kinase. The blots were pre-hybridised at room temperature in 8 ml SSSarc for 30 minutes with continuous shaking. Blots were hybridised overnight to the radio-actively labelled oligonucleotides in cold SSSarc (4^{0}C) in the coldroom, under continuous shaking. A radio-active phosphate group was added to the oligonucleotides by incubation of 2 µl of 10x T4 polynucleotide kinase buffer, 1 µl (30 ng) primer, 3 µl γ-^{32}P-dATP (10 mCi/ml) and 1 µl T4 polynucleotide kinase (10 U/µl) in 20 µl total volume at 37^{0}C for 30 minutes. 20 µl T100E was then added, after which the radio-actively labelled oligonucleotides were added to the hybridisation solution containing the different blots.
4. Blots were washed for 2 hours in SSSarc in the cold room and then overnight at room temperature, after which the blots were exposed to a phosphor-imager screen. Blots were analysed using "spotfinder" software (Molecular Dynamics).
5. Hybridised oligonucleotides were stripped from the blots by incubation at 65^{0}C in 0.1x SSSarc. The blots were subsequently hybridised to radio-actively oligonucleotide His 1, which hybridises to the four PCR products, to normalise the two blots. Hybridisation was carried out overnight in 8 ml

Church solution. Blots were then washed in 6x SSC at room temperature for 15 minutes and in Church solution at 42^{0}C for 30 minutes. Finally, blots were exposed to a phosphor-imager screen and analysed using "spotfinder".

2.21.5 Sequencing of the USP21 gene using patient DNA

1. PCR for sequencing of the His, Cys and QQD domain and the 'CGA' residues at nucleotide position 830, 1532, 1922 and 2282 was carried out with the primers USPE-5F and USPE-5R (Cys box)(619-754), QQD 1 and QQD 2 (980-1027), His 1 and His 2 (1964-2077), CGA830-F and CGA830-R (755-841), CGA1532-F and CGA1532-R (1504-1665), CGA1922-R and UBH5F (1922-1963), CGA2282-F and CGA2282-R (2193-2349) respectively. The numbers in brackets refer to the USP21 sequence (Figure 5.3) that was sequenced using these primers. 50 ng of genomic DNA was used in the PCR reaction, which also contained 12.5 pmol of the appropriate primers, 1 µl 5 mM dNTPs, 0.5 U Taq Polymerase (HT Biotechnology) and 2.5 µl 10x reaction buffer in 25 µl total volume. Reactions were as follows: 95^{0}C 1 minute 15s (95^{0}C 45s, 55^{0}C 45s, 72^{0}C 1 minute) 35x, 72^{0}C 9 minutes, 4^{0}C.
2. Amplification, by PCR, of the exons 2-24 and adjacent splice sites of USP21 was carried out with the following primers: USPE-2F/2R (245-322), USPE-3F/3R (323-467), USPE-4F/4R (468-591), USPE-5F/R (592-754), CGA830-F/R (755-841), USPE-7F/7R (842-979), QQD1/USPE-8R (980-1056), USPE-9F/9R (1057-1130), USPE-10F/10R (1131-1279), USPE-11F/11R (1280-1407), USPE-12F/12R (1408-1503), CGA1532-F/R (1504-1665), USPE-14F/14R (1666-1892), USPE-15F/CGA1922-R (1893-1963), His 1/USPE-16R (1964-2192), CGA2282F/USPE-17R (2193-2393), USPE-18F/18R (2394-2536), USPE-19F/19R (2537-2715), USPE-20F/20R (2716-2794), USPE-21F/21R (2795-2874), USPE-22F/22R (2875-2998), USPE-23F/23R (2999-3194), USPE-24F/24R (3195-3803), USPE-24F/24R2 (3195-3568). The numbers in brackets refer to the USP21 sequences (Figure 5.3) that were sequenced using these primers. The PCR reactions were identical to those described above (Chapter 2.21.5 step 1). The PCR programme used was: 95^{0}C 1 minute 30s (95^{0}C 30s, 54^{0}C 30s, 72^{0}C 1 minute) 35x, 72^{0}C 9 minutes, 4^{0}C.
3. The primers APOE-F and APOE-R were used to distinguish between the APOE-ε4 and the ε2/3 alleles. These primers amplify the nucleotides 3786-4035 of the APOE gene. The PCR reactions were identical to those described above. The PCR programme used was: 95^{0}C 1 minute 30s (95^{0}C 30s, 54^{0}C 30s, 72^{0}C 1 minute) 35x, 72^{0}C 9 minutes, 4^{0}C.
4. PCR products were precipitated as described in Chapter 2.20.1 step 17. Subsequent cycle sequencing reactions were carried out with the primers USPE-5F/QQD 1/His 1/CGA830-R/CGA1532-R/CGA1922-R or CGA2282-R for the PCR reactions performed in step 1 of this section. The primers USPE-2R/3F/4F/5F/CGA830-R/USPE-7R/8R/9R/10F/11F/12R/CGA1532-R/USPE-14F/15F/16R/CGA2282-F/USPE-18R/19R/20R/21F/22F/23F/24F/24R or USPE-24R2 were used to sequence the appropriate PCR products described in step 2. The primers APOE-F and –R were used to

sequence the PCR products from step 3. PCR conditions for cycle sequencing, precipitation of the products and sequencing were carried out as in Chapter 2.20.1 step 17.

2.22 Protocols specific for Chapter 6

2.22.1 Direct sequencing on DNA from a BAC clone

1. BAC DNA was isolated as described in Chapter 2.18.2. 10 µl DNA (~ 3µg) was digested overnight with 50 U of the restriction enzyme *Hind* III, in a total volume of 100 µl.
2. The restriction digested DNA was isolated using Wizard purification resin, after which an additional DNA precipitation was carried out as described in Chapter 2.18.2 steps 10/11.
3. Cycle sequencing, subsequent purification of the PCR products and sequencing were carried out as described in Chapter 2.20.1 step 17.

2.22.2 Sequencing of the G31B5 gene using patient DNA

1. The coding DNA from the 8 exons present in G31B5 was amplified by PCR using the following primers: G31B5-1F/R (1-139), G31B5-2F/2R (140-211), G31B5-3F/3R (212-361), G31B5-4F/4R (362-491), G31B5-5F/5R (492-643), G31B5-6F/6R (644-850), G31B5-7F/7R (851-1001), G31B5-8F/8R (1002-1204). The numbers in brackets refer to the G31B5 sequences (Figure 6.4) that were sequenced using these primers. The PCR reactions were identical to the ones described in Chapter 2.21.5 step 1. The PCR programme used was: 95^0C 1 minute 30s (95^0C 30s, 54^0C 30s, 72^0C 1 minute) 35x, 72^0C 9 minutes, 4^0C.
2. PCR products were precipitated as described in Chapter 2.20.1 step 17. Subsequent cycle sequencing reactions were carried out with the primers G31B5-1F/2R/3R/4F/5F/6F/7F/8F. PCR conditions for cycle sequencing, precipitation of the products and sequencing were carried out as in Chapter 2.20.1 step 17.

Chapter 3 Bacterial contig map of the 21q11 region

3.1 Summary

A physical map of genomic DNA is presented in the 21q11 region, consisting of overlapping recombinant P1 Artificial Chromosomes (PACs), Bacterial Artificial Chromosomes (BACs) and cosmids. PACs were identified by initial hybridisation of recombinant Yeast Artificial Chromosomes (YACs) from the 21q11 region and these were ordered into initial contigs. PAC/BAC and cosmid clones were identified to extend and link these initial contigs. The total genomic DNA covered in overlapping bacterial clones was approximately 3.4 Mbp in length and contained one 200 Kbp gap. A restriction map is presented of the minimal tiling path (MTP) of clones, which was used to estimate the size of the contig and to reconfirm the overlap between adjacent clones. The positions of the two already characterised genes are indicated together with 24 markers, five cDNA selected clones (Cheng *et al.*, 1994), one exon trapped clone (Yaspo *et al.*, 1995) and two unique clusters of expressed human sequences that are present in the Unigene collection (Schuler *et al.*, 1996).

3.2 Introduction

3.2.1 Mapping of human chromosome 21

Chromosome 21 is the smallest of the human autosomes and is 40-50 Mbp long. Down's syndrome (DS), which is trisomy of chromosome 21, is the most common genetic birth defect in man, the incidence being around 1:700 live births (Epstein, 1995).

Chromosome 21 has one of the most advanced mapping statuses in the human genome project, which is probably due to its small size and large genetic interest (Shimizu *et al.*, 1995). This has been accomplished by means of detailed genetic and physical maps, including:

- Genetic maps (Antonarakis *et al.*, 1989; McInnes *et al.*, 1993).
- A physical *Not* I map which spans the long arm of chromosome 21 (Ichikawa *et al.*, 1993).
- Overlapping YAC maps of the long arm, or of both arms of chromosome 21 (Chumakov *et al.*, 1992; Nizetic *et al.*, 1994; Korenberg *et al.*, 1995; Gardiner *et al.*, 1995).

- A cosmid pocket map covering the entire chromosome (Nizetic *et al.*, 1994).
- Megabase size bacterial contigs of selected regions (Patil *et al.*, 1994; Lafraniere *et al.*, 1995; Eki *et al.*, 1996; Osoegawa *et al.*, 1996; Ohira *et al.*, 1996; Stone *et al.*, 1996; Hubert *et al.*, 1997; Hildmann *et al.*, 1999).

Most of the physical maps are located in the distal half of the long arm of chromosome 21, which harbours more genes than the proximal half (Gardiner, 1996) and contains a region that used to be known as the "DS chromosome region".

3.2.2 Possible involvement of the 21q11 region in a number of disorders

Studies of partial trisomies and correlation to phenotypical features of DS attributed most of these dismorphic features to the trisomy of a region around D21S55 (Delabar *et al.*, 1993), which was therefore named the "DS chromosome region". Another study indicated that a more complex model might be more appropriate to explain DS, which makes increased dosage of genes throughout the entire long arm interesting in the study of the molecular basis of various features of DS (Korenberg *et al.*, 1994).

An example of a disorder linked to DS with a possible gene (or genes) mapping to the proximal half of chromosome 21 is acute non-lymphocytic leukaemia of the M7 immunotype (ANLL-M7). DS infants are 10-30 times more likely to develop one of the different forms of acute leukaemia, compared to non-DS individuals (Zipurski *et al.*, 1992). Approximately 50% of the ANLL cases in DS children are of the M7 immunotype and this type is approximately 400 times more frequent in DS than in karyotypically normal infants (Zipurski *et al.*, 1987). Transient abnormal myelopoiesis (TAM), a self regressing leukemoid reaction, is also observed in DS children (Zipurski *et al.*, 1987; 1992). Increased disomic homozygosity was found in the pericentromeric region of chromosome 21 both in DS individuals with TAM and ANLL-M7 (Abe *et al.*, 1989; Niikawa *et al.*, 1991; Shen *et al.*, 1995; Cavani *et al.*, 1998), in particular around the markers D21S215, D21S258, D21S120 and D21S192. These markers are all located in 21q11 and it has been proposed that this region might contain a gene with tumour suppressor or growth control function (Niikawa *et al.*, 1991; Ohta *et al.*, 1996).

Linkage was shown with markers in the 21q11 region and some families with early onset Alzheimer's disease (FEOAD) (St. George-Hyslop *et al.*, 1987; Goate *et al.*, 1989). Mutations were eventually found in these families in presenilin 1 which is located on 14q23.3 (Sherrington *et al.*, 1995). It has been hypothesised that the 21q11

region might contain a modifier gene, which could modify the impact of presenilin 1 (St George-Hyslop *et al.*, 1992; Van Broeckhoven, 1995).

Hemizygous and homozygous deletions of the 21q11-q21 region have been reported in tumour DNA from patients with non-small cell lung carcinoma (NSCLC). Two different studies reported that at least 60% and 50% of the tumour DNA showed loss of 21q in squamous NSCLC (Testa *et al.*, 1994; Sato *et al.*, 1994). Kohno and co-workers analysed the frequency of deletions of the long arm of chromosome 21 in NSCLC and found the highest frequency (62.5%) loss of heterozygosity (LOH) of the 21q11 marker D21S120 (Kohno *et al.*, 1998). A NSCLC cell line, Ma17, contained a homozygous deletion of the 21q11-21q21 region including the markers D21S318 (21q11) and LL136 (21q21.1) (Kohno *et al.*, 1998). This study demonstrated that a tumour suppressor that is involved in NSCLC may be present in the region 21q11-q21.

3.2.3 Structural features of the pericentromeric DNA of chromosome 21

One difficulty in generating a physical map of genomic DNA of the 21q11 region in bacterial clones is likely to be in the areas that show sequence similarities to different locations in the genome. This can either be sequence similarity between 21q11 and a different region located on chromosome 21 or elsewhere in the genome.

The centromeric region of chromosome 21 shows alphoid sequences that are very similar to the alphoid sequences on chromosome 13 and 14 (Choo *et al.*, 1991; Trowell *et al.*, 1993)

The region adjacent to the centromere on the long arm of chromosome 21 (proximal 21q11) shows similarity to both a different region on chromosome 21 and to different chromosomes:

- An area of 21q11, known as the duplicated region, was originally identified with the marker pVC1.21b-D21S190, which showed two different mapping positions on chromosome 21 (Van Camp *et al.*, 1990). The proximal copy is located at 21q11, the distal at 21q22.1. The size of the duplicated region is approximately 200 Kbp and a contig of overlapping genomic DNA in bacterial clones has been constructed for both proximal and distal copies (Potier *et al.*, 1998).
- The markers D21S190, D21S215, D21S149 and D21S253, which were localised to proximal 21q11, cross-hybridised to chromosome 13, whereas a more distal marker, D21S258, was chromosome 21 specific (Van Camp *et al.*, 1992: Van Hul *et al.*, 1993). This was confirmed by Graw and co-workers who mapped different 21q11

markers onto a single chromosome hybrid panel. Of the nine markers tested between D21Z1 (centromere) and D21S378, eight showed similarity to different chromosomes, particularly to chromosomes 13 and 18 (Graw *et al.*, 1995).

We have generated a physical map of overlapping PAC/BAC and cosmid clones of a 3.4 Mbp genomic region, which covers 3.2 Mbp and contains one 200 Kbp gap, and contains most of the 21q11 region. It contains the markers that showed an increase in disomic homozygosity in DS individuals with TAM/ANLL-M7 and the markers that showed linkage with some of the families with FEOAD. It also contains approximately half of the region homozygously deleted in the NSCLC cell line, Ma17. This work will greatly enhance the identification of novel genes and is an ideal starting material for large-scale sequencing projects.

3.3 Results

3.3.1 Identification of bacterial clones containing genomic inserts from the 21q11 region (in collaboration with J. H. Ives and D. Nizetic).

YACs that were localised to the 21q11 region (see Figure 3.1) were isolated previously (D. Nizetic). These were 1A2y21, 7A1y21, CEPH-255d1, 6A6y21, 2G2y21 and 5A11y21 (Nizetic *et al.*, 1994). They were radio-actively labelled, pre-associated with an excess of sheared total human genomic DNA and individually hybridised to the total human PAC library (RPCI 1, Ioannou *et al.*, 1994). These hybridisations were carried out by D. Nizetic and J.H. Ives. The RPCI 1 library was previously spotted in duplicate on six high density nylon membranes (one set of six membranes represented one copy of the library). Protocols are detailed in Chapters 2.18.1, 2.18.7 and 2.18.8. Membranes were exposed to X-ray films, which were analysed, resulting in 1775 positively hybridising clones for the six YAC hybridisations. PACs that hybridised to more than three YACs were eliminated from further analysis, which left 1441 clones. Sixty-eight of these referred to clones resulting from hybridisation to two or three YACs that were not overlapping according the YAC map (Nizetic *et al.*, 1994). These were also eliminated. This resulted in 1373 PAC clones with an average insert of 120 Kbp (average insert of RPCI 1) from YACs covering approximately 3.6 Mbp (see Figure 3.1). The genome equivalent coverage would therefore be 50, which is much higher than the three to four genome equivalents this PAC library has been estimated to contain (HGMP, personal communication). The majority of this excess is likely to be

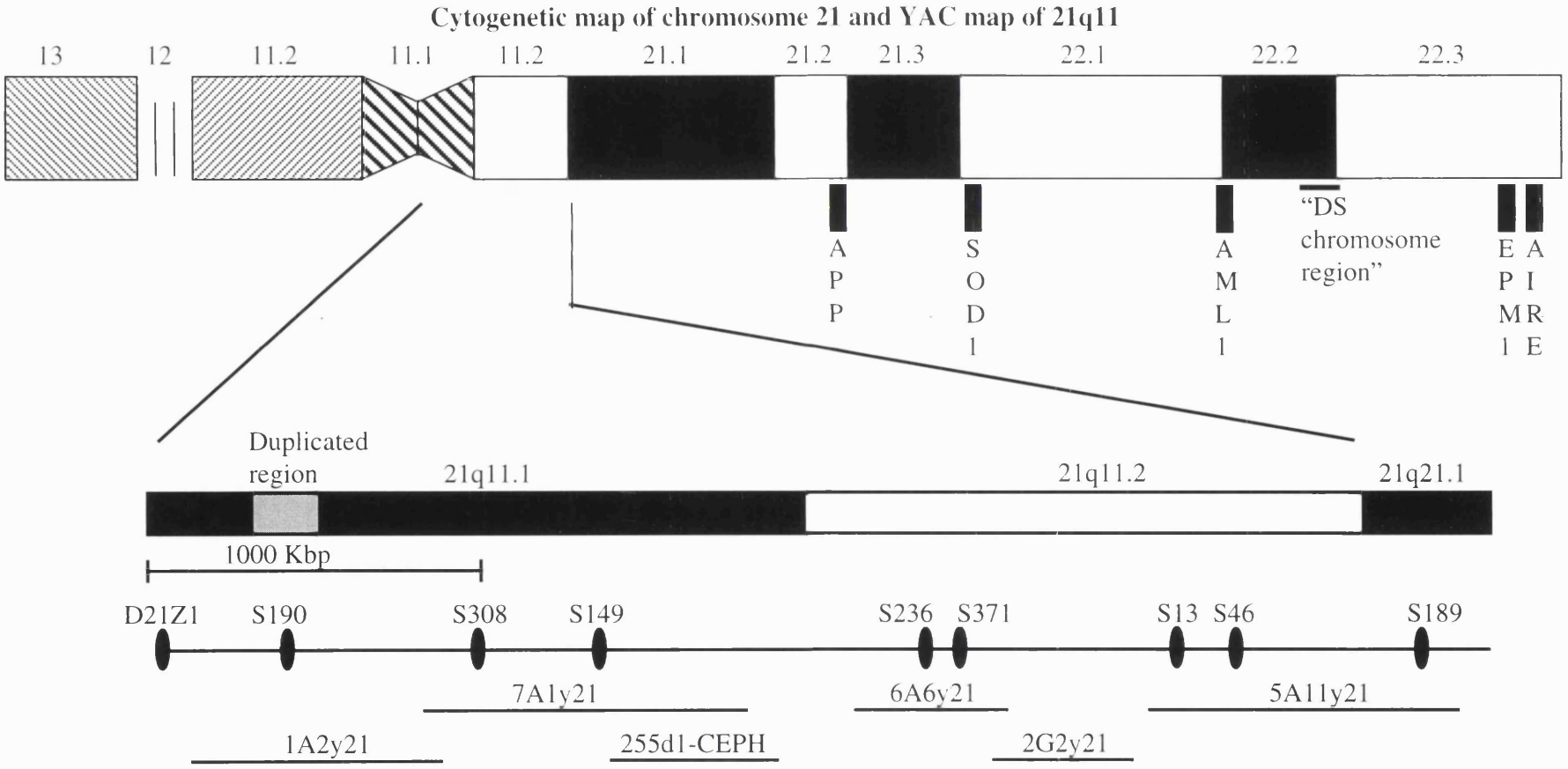


Figure 3.1: The cytogenetic map is shown of Chromosome 21 at the top. Underneath, five genes are Indicated; APP (familial early onset Alzheimer's disease), SOD1 (motor neuron disease), AML1 (acute myeloid leukaemia), EPM1 (epilepsy) and AIRE (APECED). The region which used to be known as the Down syndrome chromosome region is indicated. The bottom half of the figure shows YACs mapping to the 21q11 region (Nizetic *et al.*, 1994) that were used in hybridisation experiments to the total human PAC library RPCI 1 (Ioannou *et al.*, 1994) and the markers that border the segments described in sections 3.3.2.1-3.3.2.4 (D21 is omitted).

Number of PAC clones detected by YAC hybridisation of 21q11 YACs

YAC clones used for hybridisation	Number of PAC clones detected
1A2y21	191
7A1y21	136
7A1y21/255d1-CEPH	17
255d1-CEPH /6A6y21	38
255d1-CEPH /2G2y21	30
6A6y21/2G2y21	2
2G2y21	123
5A11y21	23

Table 3.1: YACs used for hybridisation to the PAC library RPCI 1 and numbers of PACs hybridising that were chosen for further analysis.

caused by high numbers of weakly positive clones, which were included in the analysis. A further 813 clones hybridised only to either one of the following YACs, CEPH-255d1 (528 clones) or 6A6y21 (285 clones). Due to the high number, these clones were also not included in the initial analysis of bacterial clones containing genomic inserts from the 21q11 region. The 560 remaining clones were chosen for further analysis. Table 3.1 shows to which YACs these 560 clones hybridised.

The 560 clones were inoculated into individual wells of six 96-well microtitre dishes containing LB broth and grown overnight (Chapter 2.18.1). Table 1 of Appendix I shows the different clones from RPCI 1 and the position in the 96-well plates. The clones from these six 96-well plates were inoculated in two 384-well microtitre dishes, and grown overnight. From these two 384-well plates, 48 pairs of nylon filters were produced for hybridisation experiments, by spotting the clones on Hybond N+ membranes, growing them overnight and processing the filters (Chapter 2.18.1). These filters were called the 21q11 filters. An example of these filters can be seen in Figure 3.2C

3.3.2 Detection of overlapping 21q11 PACs and construction of contigs anchored to molecular markers

3.3.2.1 The interval between the markers D21S190 and D21S308 (proximal quarter of the 21q11 region)

Initially the marker D21S190 and a cDNA selected clone, ES0033 (Cheng *et al.*, 1994), were used as probes for hybridisation to the 21q11 filters. The insert from ES0033 was amplified by PCR from phage stock (J.H. Ives), using the vector primers M13 F and R, after which the fragment was isolated from agarose via Prep-A-Gene. As can be seen from Table 3.2, eight clones hybridised to D21S190 and six to the cDNA clone. The marker D21S294 (Chumakov *et al.*, 1992, Potier *et al.*, 1998) was used as a hybridisation probe to the 21q11 filters to distinguish clones from the distal duplicated region (D21S294 is located just centromeric of the distal duplicated region). Five clones hybridised to the D21S294 probe, two of which were included in the D21S190 hybridisation, 22E2 and 50G11 (Table 3.2). These five clones were omitted from further analysis. End fragments from five genomic inserts, which hybridised to D21S190, were isolated using inverse PCR (IPCR) and vectorette PCR, and these were hybridised to the 21q11 filters. The strategy of both methods can be seen in Figure 3.2A and 3.2B. An

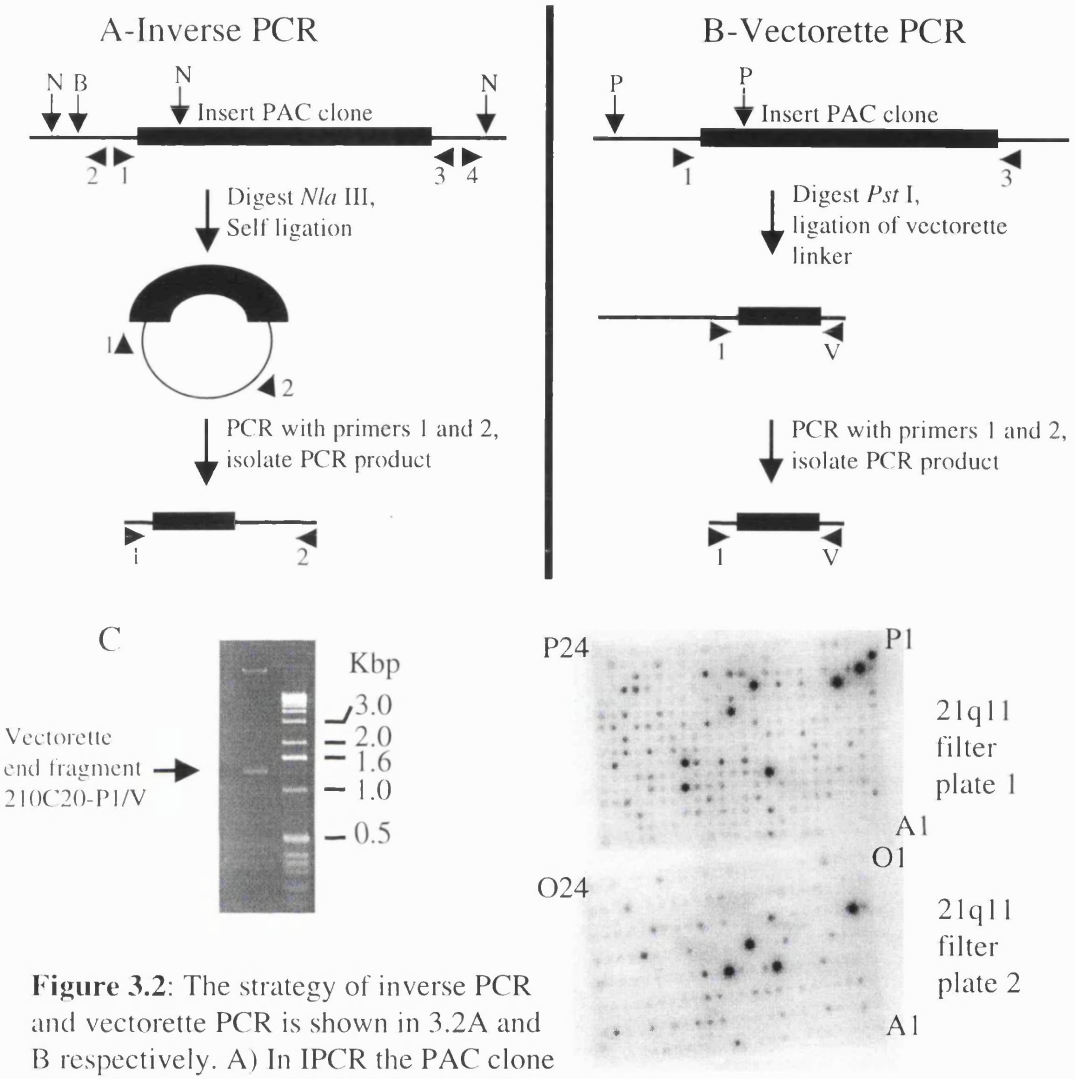


Figure 3.2: The strategy of inverse PCR and vectorette PCR is shown in 3.2A and B respectively. A) In IPCR the PAC clone containing the genomic insert is digested with *Nla* III (N) or *Bfa* I (B), after which T4 DNA ligase is added to the fragments which are present in a large reaction volume to enhance recircularisation. A PCR reaction is performed with primers located near the vector/insert border. The gel purified PCR products can be used as probes in hybridisations. B) In vectorette PCR, the clone is restriction digested with, for example *Pst* I (P). A vectorette linker, compatible with the *Pst* I site is ligated and a PCR is performed. The linker and the primer that anneals to this linker have been designed to contain mismatches. For elongation from primer V to occur, first the complementary strand has to be generated from the vector primer 1 or 3. C). An example of a hybridisation of radio-labelled vectorette probe 210C20-P1/V (clone 210C20 digested with *Pst* I, amplification with primers 1 and V). The photograph of the gel on the left shows the 1.4 Kbp fragment, which is purified from the gel and hybridised to the 21q11 filters on the right. The positive clones: 1E17, 1F10, 1G17, 1K13, 1M4, 1M11, 1N2, 1O1, 2G9, 2G13, 2I11 and 2K2 from the 384-well plate, translate into 1C9, 4C5, 1D9, 1F7, 2G2, 1G6, 4G1, 1H1, 5D5, 5D7, 5E6 and 6F1 respectively in the 96-well plates (see section 3.3.1). The final addresses of these clones in RPCI 1 can be found in Table 1 of Appendix I.

example of a hybridisation of radio-actively labelled vectorette fragment 210C20-P1/V (the PAC clone 210C20 digested with *Pst* I after which a *Pst* I cassette was ligated and PCR amplification performed with IPCR primer 1 and vectorette primer V) can be seen in Figure 3.2C, using the 21q11 filters as the target. Both methods have been described previously (IPCR: Ochman *et al.*, 1988, vectorette PCR: Riley *et al.*, 1990) but have been adapted to be used with the vectors in this study. The following vectorette and IPCR end fragments were generated, 156L9-N1/2 (the PAC clone 156L9, digested with *Nla* III and amplified with IPCR primers 1 & 2), 156L9-N3/4, 210C20-P1/V, 11P10-P1/V and 216O16-P1/V. 156L9-N1/2 hybridised to the following clones on the 21q11 filters: 22E2, 38J19, 48P20, 50G11, 77P15, 133G21, 156L9, 194E16, 194P20 and 281D20. The clones 22E2, 38J19, 48P20, 50G11 also hybridised positively to D21S294, and were thus located in the distal duplicated region. The clones 77P15, 133G21 194P20 and 281D20 hybridised to 156L9-N1/2 and to the cDNA clone ES0033. Clones from the 21q11 filters that positively hybridised to 156L9-N1/2, 156L9-N3/4, 210C20-P1/V, 11P10-P1/V and 216O16-P1/V can be seen in Table 3.2, and were used to position the clones as shown in Figure 3.3. From the PACs 32M22 and 100M18, that hybridised to ES0033, vectorette end fragments were generated (32M22-P1/V and 100M18-P1/V). From the thirty-five and twenty-eight PACs that hybridised to these fragments respectively, a subset is indicated in Table 3.2 and used in Figure 3.3. Four clones were common between these hybridisations and the hybridisation from the IPCR fragment from the D21S190-containing clone 156L9-N1/2 (77P15, 133G21, 194P20 and 281D20).

DNA from the cosmid clone Q50H6, that contained the marker D21S215 (Ohta *et al.*, 1996) was isolated, and reconfirmed by PCR to contain D21S215. To generate fragments from the genomic insert of this cosmid, it was digested with *Sau* 3A I, after which fragments were isolated as described in Chapter 2.19.4. Hybridisation of *Sau* 3A I fragments from cosmid Q50H6 to the 21q11 filters failed to identify positively hybridising PAC clones. Hybridisation of this composite probe to the total human PAC library identified one clone, 210P24, which was shown, by PCR, to contain D21S215. Hybridisation of the genomic insert of clone 210P24 to the total PAC library showed one additional clone, 22301 which also contained D21S215. Genomic inserts from PAC clones were generated by using the endonuclease *Not* I, which has a recognition site on either side of the pCYPAC2N vector, thus releasing the insert from the vector. These were resolved by pulsed field gel electrophoresis (PFGE), an example of which can be

Results of hybridisations and PCRs from the proximal quarter of 21q11

Labelled probe or amplimer	Hybrisation or PCR	Positive clones										
D21S190	Hybr.	10H17	11P10	22E2	50G11	156L9	194E16	210C20	216O16			
ES0033	Hybr.	32M22	77P15	100M18	133G21	194P20	281D20					
156L9-N1/2	Hybr.	22E2	38J19	48P20	50G11	77P15	133G21	156L9	194E16	194P20	281D20	
156L9-N3/4	Hybr.	6C17	10H17	127M18	156L9	187K17	188B2	193D6	210C20			
210C20-P1/V	Hybr.	11P10	22E2	38J19	48P20	50G11	77P15	133G21	156L9	194E16	194P20	210C20
		281D20										
11P10-P1/V	Hybr.	11P10	22E2	38J19	48P20	50G11	77P15	133G21	156L9	194E16	194P20	210C20
		281D20										
216O16-P1/V	Hybr.	11P10	22E2	38J19	48P20	50G11	77P15	156L9	194E16	194P20	210C20	216O16
		281D20										
D21S294	Hybr.	22E2	38J19	48P20	50G11	209F16						
100M18-P1/V	Hybr.	32M22	77P15	100M18	133G21	194P20	281D20					
32M22-P1/V	Hybr.	32M22	74G19	103L15	110L15	112E7	281D20					
127M18 insert	Hybr.	6C17	10H17	11P10	16C2	127M18	156L9	187K17	188B2	193D6	210C20	
Q50H6 *Sau* 3A I	Hybr	210P24										
210P24 insert	Hybr.	210P24	223O1	B62L20	B174B13	B203I11						
B62L20 insert	Hybr.	1C18	16C2	29H4	30P22	50P15	94F2	98M24	210P24	223O1		
29H4 insert	Hybr.	1C18	29H4	30P22	32C15	50P15	94F2	98M24	248P15			
D21S369	PCR	16C2	B62L20									
D21S215	PCR	210P24	223O1	Q50H6	B62L20	B174B13	B203I11					
D21S297	PCR	1C18	29H4	30P22	32C15	50P15	94F2	98M24	B62L20			
D21S308	PCR	29H4	30P22	32C15								
H41072	PCR	29H4	30P22	32C15	248P15							

Table 3.2: Hybridisation and PCR results are shown of the region D21S190-D21S308 on chromosome 21q11. The first column shows the probes and the amplimers that were used, the second column shows if this was a hybridisation (Hybr.) or a PCR. The other columns show the positively hybridising or PCR positive clones. In all hybridisations the 21q11 filters were used, except for the genomic inserts of 127M18, 210P24, B62L20, 29H4 and *Sau* 3A I fragments from Q50H6; these were hybridised to the total human PAC library. The clone 210P24 was also hybridised to the total human BAC library.

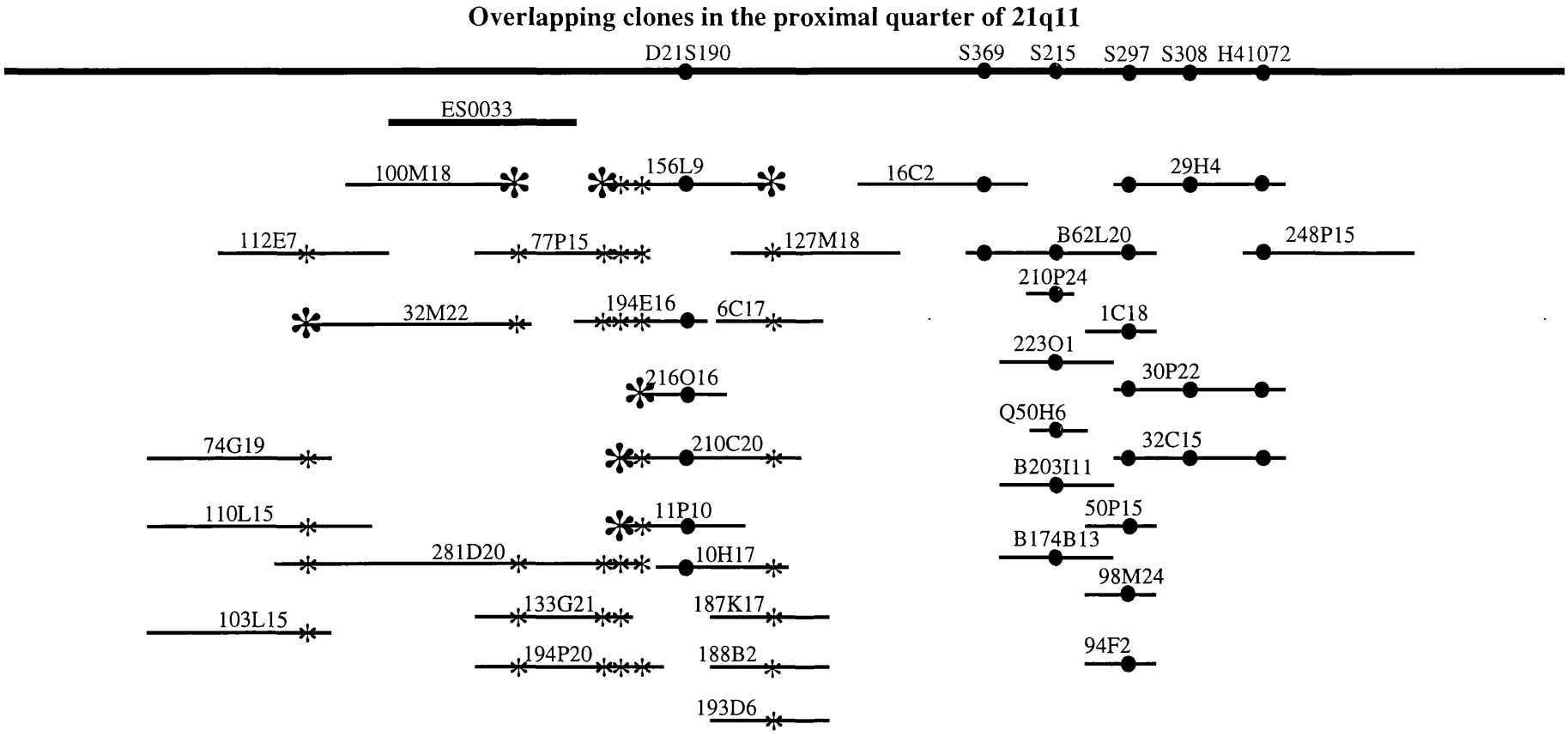


Figure 3.3: A contig is shown of genomic inserts of bacterial clones in the 21q11 region, including the markers D21S190-D21S308. The contig was generated by hybridisation of the marker D21S190 and cDNA selected clone ES0033. Subsequent hybridisations of end probes of the different clones generated by IPCR and vectorette PCR, hybridisation of genomic inserts from the bacterial clones and PCR amplification of markers, as indicated in Table 3.2 were used to position the clones. ✻ -endprobe generated, ✳ -clones hybridising to the end probe on the 21q11 filters (as indicated in Table 3.2). ● -clones either positive by hybridisation (for D21S190) or by PCR for the other markers.

seen in Figure 3.10A lane 6. The genomic insert was subsequently isolated from the agarose gel (see Chapter 2.18.9). An example of a hybridisation of a genomic insert, from a bacterial clone onto membranes containing DNA from the total human PAC library, can be seen in figure 3.8. The genomic insert from 210P24 was subsequently hybridised to the total human BAC library (Shizuya *et al.*, 1992). This hybridisation revealed the following clones, B62L20, B174B13 and B203I11 which contained D21S215. B62L20 was shown, by PCR, to contain additionally the markers D21S369 and D21S297. The genomic insert of the BAC clone B62L20 hybridised to fifteen clones of the total human PAC library, the most important ones are shown in Table 3.2. The genomic insert from the clone 127M18, which was present on the 21q11 filters and hybridised to the IPCR fragment 156L9-N3/4, was shown to hybridise to twenty clones of the total human PAC library (see Table 3.2 for a subset of these clones). The PAC 16C2 was one of the positively hybridising clones, which also hybridised to the genomic insert of B62L20 (see Table 3.2). B62L20 and 16C2 also shared the marker D21S369. These results link the contig from the D21S190 region with the D21S369, D21S215 and D21S297 region and also orients the contig, since D21S190 is located centromeric from D21S215 (Van Hul *et al.*, 1993).

The cDNA selected clone ES0099 (Cheng *et al.*, 1994) was shown to be very similar to the human EST H41072 (J.H. Ives). The insert from this EST, which was generated by PCR using the M13 forward and reverse primers (J.H. Ives), was hybridised to the total human PAC library. The positively hybridising clones were screened by PCR with two primers specific for H41072, H41072 F2 and R2. From one clone a product of the correct size was amplified: 29H4. This clone was shown by PCR to contain the markers D21S297 and D21S308. The clone 29H4 was also identified by B62L20, linking this clone to the others in the contig. The insert of 29H4 was hybridised to the total human PAC library, showing fifteen positively hybridising clones; the most important ones are shown in Table 3.2. Five additional PAC clones were shown to be common between hybridisations of 29H4 and B62L20 to the total human PAC library; these were 1C18, 30P22, 50P15, 94F2 and 98M24. Six additional clones were shown to contain D21S297; 1C18, 30P22, 32C15, 50P15, 94F2 and 98M24. The clones 30P22 and 32C15 showed PCR amplification of marker D21S308, and 30P22, 32C15 and 248P15 were shown to contain the H41072 amplimer.

These results are shown in Table 3.2 and the positioning into a contig, using the data from this table, is shown in Figure 3.3. The genomic insert from the PAC 248P15,

which extends furthest telomerically, was isolated and used as a radio-actively labelled probe in hybridisation experiments to the total human PAC library (Ioannou *et al.*, 1994). This resulted in approximately 250 positively hybridising clones of the total human PAC library including the clones 29H4 and 30P22. The PAC clone 32C15 did not hybridise to this insert which may be explained by the fact that the overlap between the two clones could be very small. This high number of hybridising clones could be due to low to medium copy repeats present in the insert which are not competed out by hybridisation of excess total human placental DNA. This increases the difficulty of the identification of additional overlapping clones by hybridisation. The presence of low and middle-copy repetitive sequences have been reported to be frequent in this area (Yu *et al.*, 1997).

3.3.2.2 The interval between the markers D21S149 and D21S236 (second quarter of the 21q11 region)

The markers D21S149 and D21S253 were hybridised to the 21q11 filters, and six clones were identified in total, three of which were shared between the two markers (48P5, 72H5 and 214G14). When the insert of the cDNA selected clone ES0093 (Cheng *et al.*, 1994) was used as a radio-actively labelled probe and hybridised to the 21q11 filters, eleven PAC clones were identified, two of which also contained D21S253 (142I11 and 256M13). Hybridisation of ES0040 to the 21q11 filters revealed four clones, which were all in common with ES0093 (63K18, 82L10, 120E20 and 168N11). The clones that hybridised to D21S149, D21S253, ES0093 and ES0040 can be seen in Table 3.3. PCR was performed with primers that amplified the marker D21S318 using DNA from the identified PAC clones, the following clones showed amplification of a fragment of the correct size; 63K18, 76I22, 82L10, 120E20, 142I11, 168N11 and 256M13. Fragments from the ends of the genomic inserts of the bacterial clones were generated from a number of clones by IPCR and vectorette PCR. These fragments were used for hybridisation to the 21q11 filters. They were 72H5-B1/2 (the PAC clone 72H5, which was digested with *Bfa* I after which amplification was performed with the IPCR primers 1 & 2), 207O15-N3/4, 258B10-B1/2, 214G14-B1/2, 72H5-N3/4, 82L10-P1/V and 168N11-H3/V (the PAC clone 168N11, digested with *Hind* III, after which a *Hind* III cassette was ligated and a specific PCR fragment was amplified with IPCR primer 3 and the vectorette primer V). The fragment 72H5-B1/2 hybridised only to the PAC clones 73H5 and 214G14. Since no other fragments from the inserts were isolated from

Results of hybridisations and PCRs from the second quarter of 21q11

Labelled probe or amplimer	Hybridisation or PCR	Positive clones									
D21S149	Hybr.	48P5	72H5	214G14							
D21S253	Hybr.	48P5	72H5	142I11	207O15	214G14	256M13				
ES0093	Hybr.	63K18	76I22	82L10	120E20	124J6	132O8	134I9	142I11	168N11	256M13
		258B10									
ES0040	Hybr.	63K18	82L10	120E20	168N11						
72H5-B1/2	Hybr.	72H5	214G14								
207O15-N3/4	Hybr.	72H5	207O15	214G14							
258B10-B1/2	Hybr.	63K18	72H5	76I22	82L10	134I9	142I11	168N11	256M13	258B10	
214G14-B1/2	Hybr.	63K18	72H5	76I22	82L10	124J6	132O8	134I9	142I11	207O15	214G14
		256M13	258B10								
72H5-N3/4	Hybr.	63K18	72H5	76I22	82L10	142I11	168N11	207O15	256M13	258B10	
82L10-P1/V	Hybr.	63K18	82L10	168N11							
168N11-H3/V	Hybr.	63K18	120E20	168N11							
D21S258	PCR	90B5	98L15	183N4	256I9						
D21S120	PCR	90B5	126N20	184J15	B330G15						
98L15-H1/V	Hybr.	63K18	76I22	82L10	120E20	124J6	132O8	134I9	142I11	256M13	258B10
D21S48	Hybr.	31B5	B39I12								
31B5 insert	Hybr.	31B5	126N20	184J15	B10K6	B39I12	B246N7	B330G15			
184J15 insert	Hybr.	31B5	90B5	126N20	184J15						
D21S318	PCR	63K18	76I22	82L10	120E20	142I11	168N11	256M13			
D21S378	PCR	90B5	183N4	256I9							
D21S374	PCR	31B5	184J15	B246N7	B330G15						
D21S16	PCR	31B5	B10K6	B246N7	B330G15						
D21S236	PCR	31B5	B10K6	B39I12	B246N7	B330G15					

Table 3.3: Hybridisation and PCR results are shown of the region D21S149-D21S236 on chromosome 21q11. The first column shows the probes and amplimers that were used, the second column shows if this was a hybridisation (Hybr.) or a PCR. The other columns show the positively hybridising or PCR positive clones. The 21q11 filters were used in all the hybridisations, except D21S48 and the genomic inserts from the clones 31B5 and 184J15; these were hybridised to the total human PAC library. The insert of 31B5 was also hybridised to the total human BAC library.

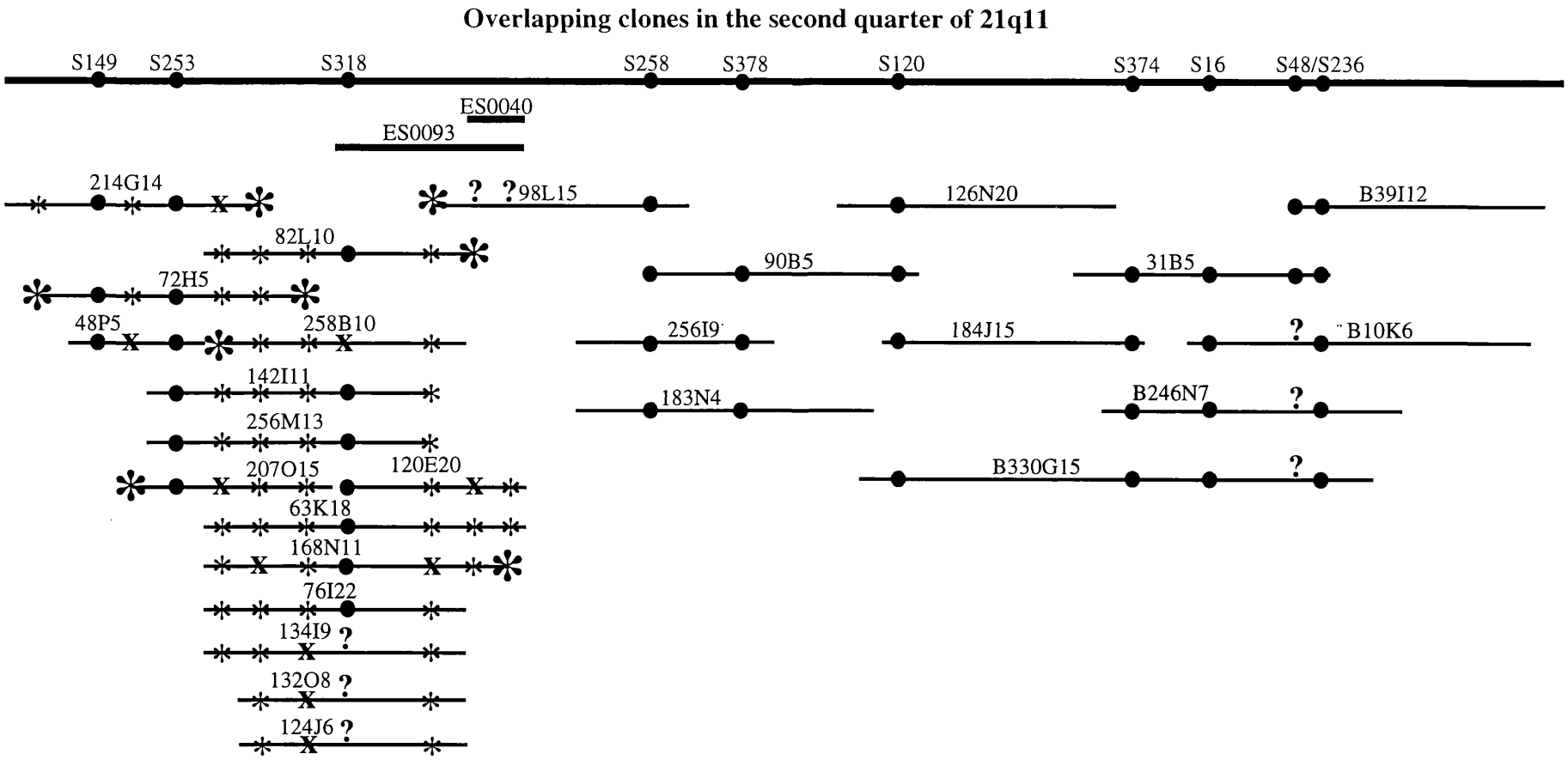


Figure 3.4: A contig is shown of genomic inserts of bacterial clones in the 21q11 region including the markers D21S149-D21S236. The contig was generated by hybridisation of the markers D21S149 and D21S253 and cDNA selected clones ES0093 and ES0040. Subsequent hybridisations of end probes of the different clones generated by IPCR and vectorette PCR, hybridisation of genomic inserts and analysis of marker content, as indicated in Table 3.3, were used to construct this contig. ✻-endprobe generated, ⁎-clones that hybridised positively on the 21q11 filters to the endprobe (as indicated in Table 3.3), **X**-clone did not hybridise to the endprobe or was negative for a marker. **?**-clones not tested. ● -clones either positive by hybridisation (for D21S149 and D21S253) or by PCR for the other markers.

this end of the contig, this would make 214G14 the "gap-facing" clone. The fragment from the other end of 72H5, 72H5-N3/4, hybridised to a large number of clones including the clones that hybridised to ES0040, ES0093 and D21S318 (63K18, 82L10 and 168N11). Fragment 168N11-H3/V hybridised to clones 63K18, 120E20 and 168N11. No additional clones were identified on the 21q11 filters using the fragments from the inserts compared with D21S149, D21S253, ES0040 and ES0093. The positively hybridising clones from the hybridisation experiments described above, can be seen in Table 3.3. Their position in a contig deduced from this table is shown in Figure 3.4. Four IPCR end fragments, 72H5-N3/4, 82L10-P1/V, 207O15-N3/4 and 258B10-B1/2, did not hybridise to all the expected clones on the 21q11 filters. As shown in Figure 3.4 and Table 3.3, 72H5-N3/4 did not hybridise to 124J6, 132O8 and 134I9, 82L10-P1/V did not hybridise to 120E20, 207O15-N3/4 did not hybridise to the PAC 48P5 and 258B10-B1/2 did not hybridise to 207O15 and 214G14. This is likely to be due to regions of similarity of DNA sequences between chromosome 21 and other chromosomes (Van Camp *et al.*, 1992; Van Hul *et al.*, 1993; Graw *et al.*, 1995).

To extend the contig additional clones had to be identified in the total human PAC/BAC or to cosmid libraries. The genomic insert of clone 214G14 was isolated and used as a radio-actively labelled probe and hybridised to the total human PAC library (Ioannou *et al.*, 1994). This insert hybridised to approximately 300 PAC clones in the total human PAC library. Like 248P15, the clone from the opposite end of this gap (see Chapter 3.3.2.1), repeats within the insert are probably not competed against and result in the high number of clones hybridising to this insert. This could again be explained by low and middle-copy repetitive elements that have been reported to be present at high frequency in the area (Yu *et al.*, 1997). This results in a gap in the physical map between the clones 248P15 and 214G14.

The clones containing the markers D21S258 and D21S120 were identified in three rounds of PCR amplification using the primers D21S258-A and –B and D21S120-A and –B, as can be seen in Figure 3.5. The first round included 21 pools, each pool containing the DNA from fifteen 384-well microtitre dishes from RPCI 1. DNA from individual 384-well microtitre dishes, from the pools that showed specific amplification from the first round, were PCR amplified in a second round of screening. This resulted in identification of the microtitre dish in which the positive clone(s) was/were located. The final well addresses were established in a final round of screening, in which DNA

Identification by PCR, of PAC clones from the RPCI 1 library, which contain the markers D21S120 and D21S258

A: First round of PCR with D21S120 and D21S258 amplimers

PCR pools	A	B	C	D	E	F	G	H	I	J	K
plates in pools	1-15	16-30	31-45	46-60	61-75	76-90	91-105	106-120	121-135	136-150	151-165
D21S120						+	+				
D21S258						+			+		

PCR pools	L	M	N	O	P	Q	R	S	T	U
plates in pools	166-183	184-198	199-213	214-228	229-243	244-258	259-273	274-290	291-305	306-321
D21S120	+					+				
D21S258		+								

B: Second round of PCR with D21S120 and D21S258 amplimers

Plate	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90
D21S120															
D21S258															+
Plate	91	92	93	94	95	96	97	98	99	100	101	102	103	104	105
D21S120								+							
Plate	166	167	168	169	170	171	172	173	177	178	179	180	181	182	183
D21S120															+
Plate	244	245	246	247	248	249	250	251	252	253	254	255	256	257	258
D21S120													+		
Plate	121	122	123	124	125	126	127	128	129	130	131	132	133	134	135
D21S258						+									
Plate	184	185	186	187	188	189	190	191	192	193	194	195	196	197	198
D21S258	+														

C: Third round of PCR with D21S120 and D21S258 amplimers



D: Determining the final well address of clone 98L15

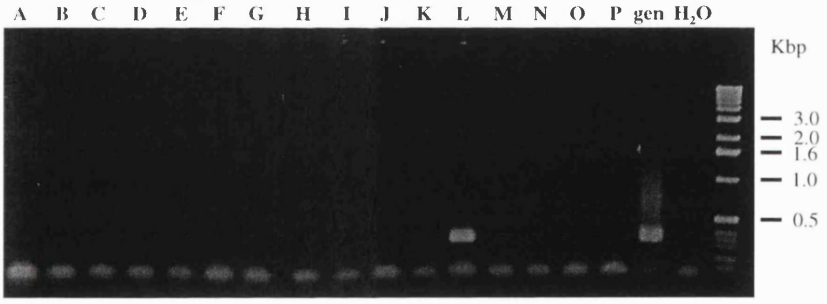


Figure 3.5: The PCR analysis was carried out in three rounds. **A:** In the first round, PCR was performed on 21 pools, each containing DNA from fifteen 384-well microtitre dishes from RPCI 1. **B:** A second round of PCR was used to determine the microtitre dish in which the positive clones were located. The third round established the well addresses of the positive clones, by using DNA from the clones which make up individual columns and, eventually, DNA from single clones. + indicates a fragment of the right size in the respective pool, microtitre dish and row. Empty squares indicate that no fragment was amplified. H_2O and genomic DNA were included and were always negative and positive respectively. The D21S120-containing clones are: 90B5, 126N20 and 184J15; 90B5, 98L15, 183N4 and 256I9 contain D21S258. **D:** An example of determining the final address of the clone 98L15 is indicated (photo). This clone was detected via pool G, plate 98 and column A15-P15. Clones that were present in this column were analysed separately (see photo of the PCRs on a 1.5% w/v agarose gel), this resulted in the identification of the PAC 98L15. Gen- genomic DNA. Size markers are indicated on the right hand side of the photo, in Kbp.

of rows and columns from the microtitre dishes was used in PCR amplification of the markers. These were 90B5, 98L15, 183N4 and 256I9 for D21S258 and 90B5, 126N20 and 184J15 for D21S120. Hybridisation of the vectorette end probe from the PAC clone 98L15 (98L15-H1/V) hybridised to ten clones (see Table 3.3), that were described earlier in this section, linking the markers D21S258 and D21S120 to D21S149 and D21S253. This orients the contig, as D21S149 and D21S253 are located centromeric to D21S258 (Van Hul *et al.*, 1993).

D21S48 was hybridised to the total human PAC library, which resulted in eleven positively hybridising clones. One of these PAC clones was 31B5, a clone that originally hybridised only to 6A6y21, and for this reason was initially omitted from the 21q11 filters (see Chapter 3.3.1). Hybridisation of the genomic insert generated from the clone 31B5 to the total human PAC library additionally identified 126N20 and 184J15. The result of part of this hybridisation can be seen in Figure 3.6, in which 31B5 hybridised to 126N20. The previous hybridisation was confirmed by using the insert of 184J15, which hybridised to 31B5, 90B5, 126N20 and 184J15. Hybridisation of the insert of 31B5 to the total human BAC library identified four clones, B10K6, B39I12, B246N7 and B330G15. These, together with the other clones from this part of the contig, were assessed for their marker content using primers that PCR amplify the markers D21S378, D21S374, D21S16 and D21S236. D21S378 was present on the clones 90B5, 183N4 and 256I9, and D21S374 was present on 31B5, 184J15, B246N7 and B330G15. The clones 31B5, B10K6, B246N7 and B330G15 contained the marker D21S16, and D21S236 was present on 31B5, B10K6, B39I12, B246N7 and B330G15. These results can be seen in Table 3.3. The hybridisation data and the PCR results are assembled and shown in Figure 3.4.

In order to identify additional clones telomeric to B39I12, the genomic insert from this clone was hybridised to the total human PAC library. The only positive clone was 31B5 which was positioned centromerically to B39I12. This leaves a gap telomerically of B39I12.

3.3.2.3 The interval between the markers D21S371 and D21S13 (third quarter of the 21q11 region)

D21S13 was hybridised to the 21q11 filters which resulted in identification of clones 71F7, 91E10 and 270M7. Hybridisation of PCR amplified insert of the cDNA selected clone ES00136 (Cheng *et al.*, 1994) to the same filters showed positive hybridisation of

Hybridisation of the genomic insert of the PAC clone 31B5 to the total human PAC library

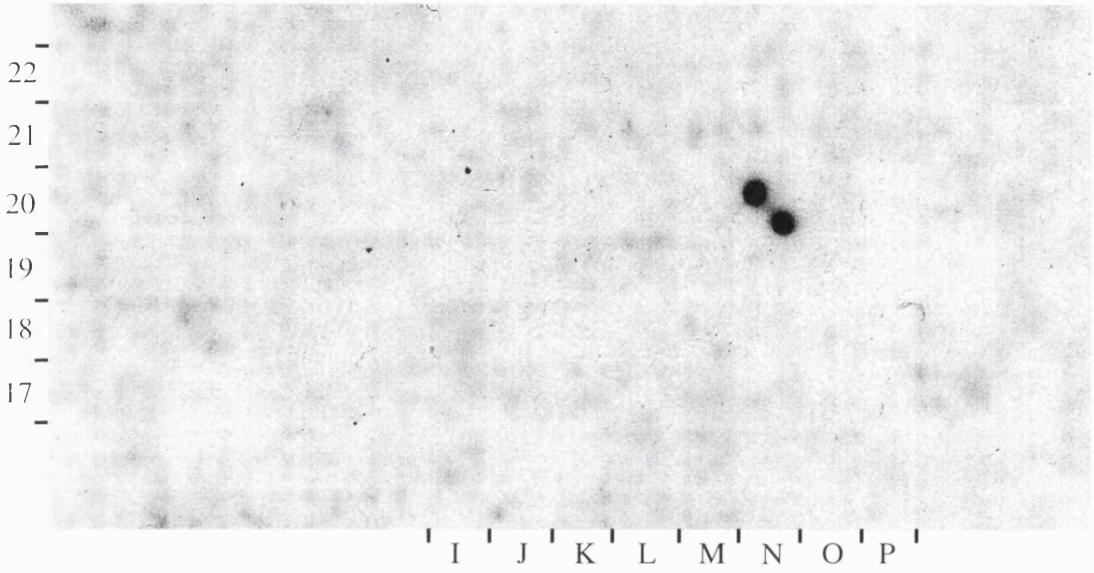


Figure 3.6: The genomic insert of the PAC clone 31B5 which was isolated after *Not* I digestion and pulsed field gel electrophoresis (see Chapter 2.18.9), was radio-actively labelled (see Chapter 2.18.7). This probe was hybridised to the total human PAC library RPCI 1, which was represented on six high density membranes. The membranes were subsequently exposed to X-ray film and a representation of part of one membrane can be seen above. The final co-ordinate of the positively hybridising clone can be calculated from its position and the pattern of the two signals (the library is spotted in duplicate). The positively hybridising clone in the above picture is 126N20. Additional clones that hybridised to 31B5 were: 31B5 and 184J15 (not shown).

24B1, 70G18, 70H17, 247N12 and 270M7; ES0040 hybridised to 70G18, 70H17, 71F7, 91E10 and 270M7. From these clones the following end fragments of the genomic insert were obtained by IPCR and vectorette PCR: 270M7-B1/2, 70G18-P1/V, 71F7-N3/4, 91E10-N3/4 and 24B1-B1/2, which were hybridised to the 21q11 filters. 71F7-N3/4 and 91E10-N3/4 hybridised to the same clones; 70G18, 70H17, 71F7, 91E10, 247N12 and 270M7. 270M7-B1/2 identified eight clones, two were in common with hybridisations described above (24B1 and 270M7), six were clones not identified previously on the 21q11 filters: 30P13, 61O14, 119P15, 260H6, 270D8 and 287N14. From these clones two vectorette end fragment were generated, 270D8-P1/V and 30P13-P1/V, and hybridised to the 21q11 filters. 270D8-P1/V hybridised to six clones that were also identified by 270M7-B1/2 (24B1, 30P13, 260H6, 270D8, 270M7 and 287N14). 30P13-P1/V revealed the clones 30P13, 260H6 and 270D8; no new clones were identified. The results of all the above mentioned hybridisations are shown in Table 3.4, and the results were used to order the clones into a contig, which is shown in Figure 3.7. PCR analysis of primers amplifying D21S373 showed amplification with 270D8, and 30P13, which orients the contig on the genomic map (the marker D21S373 is located centromerically to D21S13 (Butler *et al.*, 1992)).

Two PAC clones, 190N10 and 153I22, were identified by the YACs 6A6y21 and 2G2y21. The PAC 153I22 was shown, by PCR, to contain the marker D21S373 and therefore overlaps with 270D8 and 30P13. This was confirmed by hybridisation of the genomic insert from clone 153I22 to the total human PAC library, which hybridised to 8C10, 30P13, 37O12, 153I22, 190N10, 216O5, 260H6 and 270D8. Additionally, the vectorette end 153I22 H1/V hybridised to the clones 30P13, 153I22, 260H6 and 270D8 on the 21q11 filters. Insert of the clone 190N10 hybridised to 8C10, 153I22 and 190N10, and was the only clone that contained marker D21S371. Hybridisation of the insert from 190N10 to the total human BAC library identified the following clones B41F2, B145A11 and B146I7. None of these clones contain the markers D21S371 or D21S373. Inserts of 8C10 and 153I22 both hybridised to B41F2, B145A11 and B146I7, which places these clones between D21S371 and D21S373, overlapping with 190N10 and 153I22. These results leave a gap in the contig map distal to B39I12 (Chapter 3.3.2.2) and proximal to 190N10.

On the telomeric side of this contig (D21S13-containing clones), the genomic insert of 71F7 was hybridised to the total human PAC library. This insert hybridised to the following clones: 70G18, 70H17, 71F7, 91E10, 247N12 and 270M7. All these clones

Results of hybridisations and PCRs from the third quarter of 21q11

Labelled probe or amplimer	Hybridisation or PCR	Positive clones								
D21S13	Hybr.	71F7	91E10	270M7						
ES00136	Hybr.	24B1	70G18	70H17	247N12	270M7				
ES0040	Hybr.	70G18	70H17	71F7	91E10	270M7				
270M7-B1/2	Hybr.	24B1	30P13	61O14	119P15	260H6	270D8	270M7	287N14	
70G18-P1/V	Hybr.	24B1	30P13	70G18	70H17	247N12	270M7			
71F7-N3/4	Hybr.	70G18	70H17	71F7	91E10	247N12	270M7			
91E10-N3/4	Hybr.	70G18	70H17	71F7	91E10	247N12	270M7			
24B1B1/2	Hybr.	24B1	30P13	61O14	119P15	194P24	260H6	270D8	287N14	
270D8-P1/V	Hybr.	24B1	30P13	260H6	270D8	270M7	287N14			
30P13-P1/V	Hybr.	30P13	260H6	270D8						
D21S373	PCR	30P13	37O12	153I22	270D8					
153I22 insert	Hybr.	8C10	30P13	37O12	153I22	190N10	216O5	260H6	270D8	
153I22 H1/V	Hybr.	30P13	153I22	260H6	270D8					
190N10 insert	Hybr.	8C10	153I22	190N10	B41F2	B145A11	B146I7			
D21S371	PCR	190N10								
71F7 insert	Hybr.	70G18	70H17	71F7	91E10	247N12	270M7	B194B7	B293H17	B328G21
B194B7 insert	Hybr.	71F7	91E10	103K19	270M7					

Table 3.4: Hybridisation and PCR results are shown of the region D21S371-D21S13 on chromosome 21q11. The first column shows the probes and amplimers that were used, the second column shows if this was a hybridisation (Hybr.) or a PCR. The other columns show the positively hybridising and PCR positive clones. The 21q11 filters were used in all hybridisations except for the genomic inserts of the clones 153I22, 190N10, 71F7 and B194B7; these were hybridised to the total human PAC library. The inserts from 190N10 and 71F7 were also hybridised to the total human BAC library.

Overlapping clones in the third quarter of 21q11

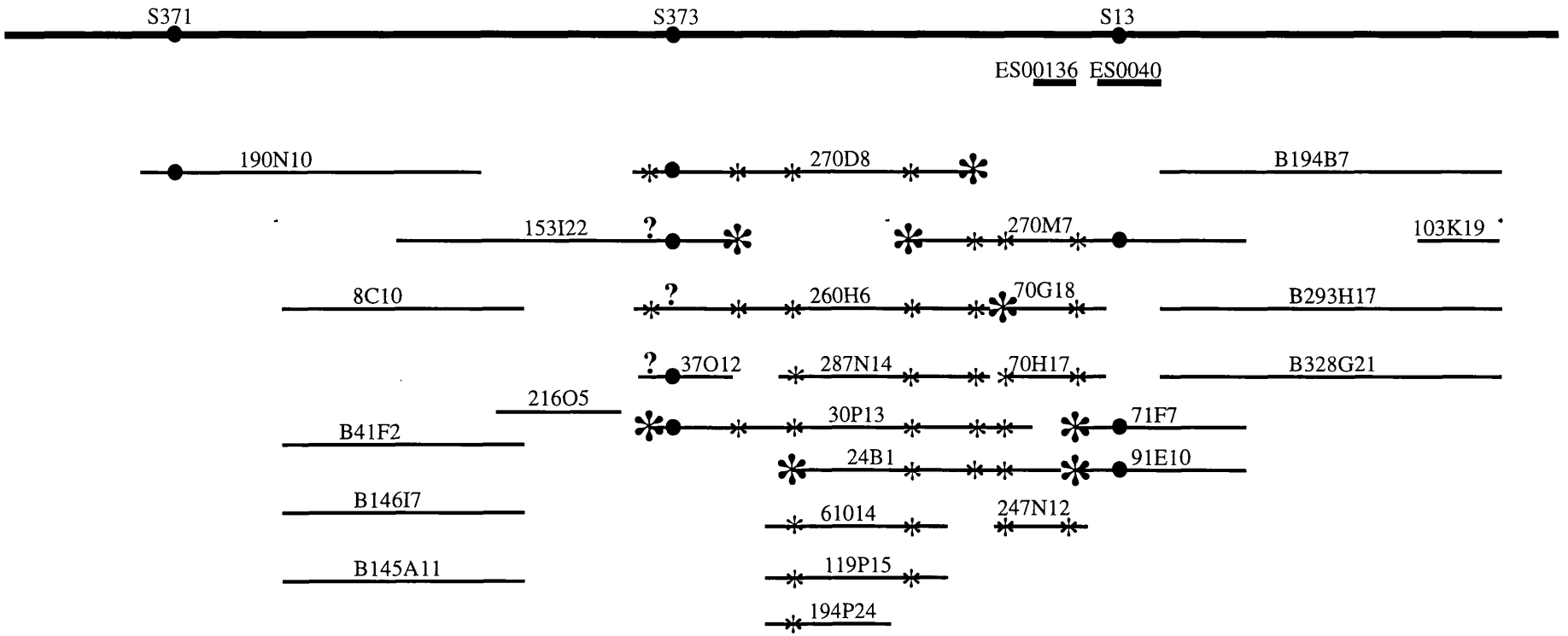


Figure 3.7: A contig is shown of genomic inserts of bacterial clones in the 21q11 region including the markers D21S371-D21S13. The contig was generated by hybridisation of the marker D21S13 and cDNA selected clones ES0040 and ES00136. Subsequent hybridisations of end probes of the different clones generated by IPCR and vectorette PCR, hybridisation of genomic inserts and PCR analysis of the marker content of different clones, indicated in Table 3.4, contributed to this contig. ✻ -endprobe generated, ⁎ -clones hybridised by the endprobe on the 21q11 filters (as indicated in Table 3.4), ?-clones not tested. ● -clones either positive by hybridisation (for D21S13) or by PCR for the other markers.

were already identified on the 21q11 filters. The same insert was also hybridised to the total human BAC library, which identified the clones B194B7, B293H17 and B328G21. These clones were shown, by PCR, not to contain the marker D21S13. An insert of the BAC clone B194B7 was hybridised to the total human PAC library and although this experiment showed a large number of positive clones, it hybridised to 71F7, 91E10 and 270M7 (D21S371-S13 contig) and to 103K19 (D21S46-S189 contig, see Chapter 3.3.2.4). All the hybridisation results can be seen in Table 3.4 and the contig which was established using these results in Figure 3.7.

3.3.2.4 The interval between the markers D21S46 and D21S189 (distal quarter of the 21q11 region)

The markers D21S46, D21S192, D21S172 and D21S189 were hybridised to the 21q11 filters by labelling using random primers (Chapter 2.18.7), the primers D21S172-A and –B were hybridised radio-actively (Chapter 2.21.4). The clones that hybridised positively were 10J13 for D21S46, 75G13 for D21S192, 101E10, 190A9 and 265A22 for D21S172 and 135E14 for D21S189. By using IPCR and vectorette PCR a number of end fragments were generated from bacterial clones from this region. These were either clones that were positive for one of the above mentioned markers, or PACs that were initially identified by the YAC 5A11y21. As can be seen from Table 3.1, there were only twenty-three PAC clones identified by this YAC. The vectorette end fragment 265A22-H1/V hybridised to 75G13 and 265A22, linking the markers D21S192 and D21S172. 190A9-P1/V hybridised to 190A9 and 265A22; the opposite end of this clone. 190A9-N3/4, identified 111K10 and 190A9. Both insert ends were generated from 73M5 which was one of the twenty-three clones that hybridised to the YAC 5A11y21. 73M5-H3/V hybridised to 73M5, 111K10, 128A18 and 147H3; 73M5-P1/V identified 73M5 and 135E14. These experiments link the markers D21S192-D21S172-D21S189. The hybridisation results can be seen in Table 3.5, which also contains the additional hybridisations of 111K10-N3/4 and 147H3-P1/V. The results were used to position the clones into a contig which can be seen in Figure 3.8.

The genomic inserts from the clones 10J13 and 71F7 were hybridised to high density filters containing DNA of chromosome 21 cosmids from the libraries c102/103 (Nizetic *et al.*, 1991a), in order to close the gap between these clones. 71F7 is a D21S13 positive clone (see Chapter 3.3.2.3); D21S13 is located 220-300 Kbp from D21S46 (Van Hul and Van Broeckhoven, 1993). The insert from clone 71F7 hybridised to cosmids from

Results of hybridisations and PCRs from the distal quarter of 21q11

Labelled probe or amplimer	Hybridisation or PCR	Positive clones			
D21S46	Hybr.	10J13			
D21S192	Hybr.	75G13	B318B17		
D21S172	Hybr.	101E10	190A9	265A22	
D21S189	Hybr.	135E14	B40D24		
265A22-H1/V	Hybr.	75G13	265A22	B6L13	
190A9-P1/V	Hybr.	190A9	265A22	B162O10	
190A9-N3/4	Hybr.	111K10	190A9		
73M5-H3/V	Hybr.	73M5	111K10	128A18	147H3
73M5-P1/V	Hybr.	73M5	135E14	B40D24	
111K10-N3/4	Hybr.	111K10	190A9		
147H3-P1/V	Hybr.	111K10	128A18	147H3	
c102 65F03 T7	Hybr.	c102 65F3	10J13		
c102 44E10 T7	Hybr.	c102 44E10	c102 65F3		
c102 44E10 *Sau* 3A I	Hybr.	75G13			
10J13 insert	Hybr.	10J13	103K19	B318B17	
103K19-DOP primer	Hybr.	103K19	B194B7	B293H17	B328G21
B194B7-M B1/B2	Hybr.	103K19	B194B7	B293H17	B328G21
265A22 insert	Hybr.	B6L13	B162O10	B193O15	
135E14 insert	Hybr.	B40D24			
D21S372	PCR	10J13			
D21S382	PCR	73M5	111K10	128A18	147H3

Table 3.5: Hybridisations and PCR results are shown of the region D21S46-D21S189 on chromosome 21q11. The first column shows the probes that were used, the second column shows if this was a hybridisation (Hybr.) or a PCR. The other columns show the positively hybridising or PCR positive clones. All hybridisations were carried out using the 21q11 filters, except for the genomic inserts of the clones 10J13, 265A22 and 135E14; these were hybridised to the total human BAC library. The *Sau* 3A I fragments from c102 44E10 and the insert from 10J13 were hybridised to the total human PAC library.

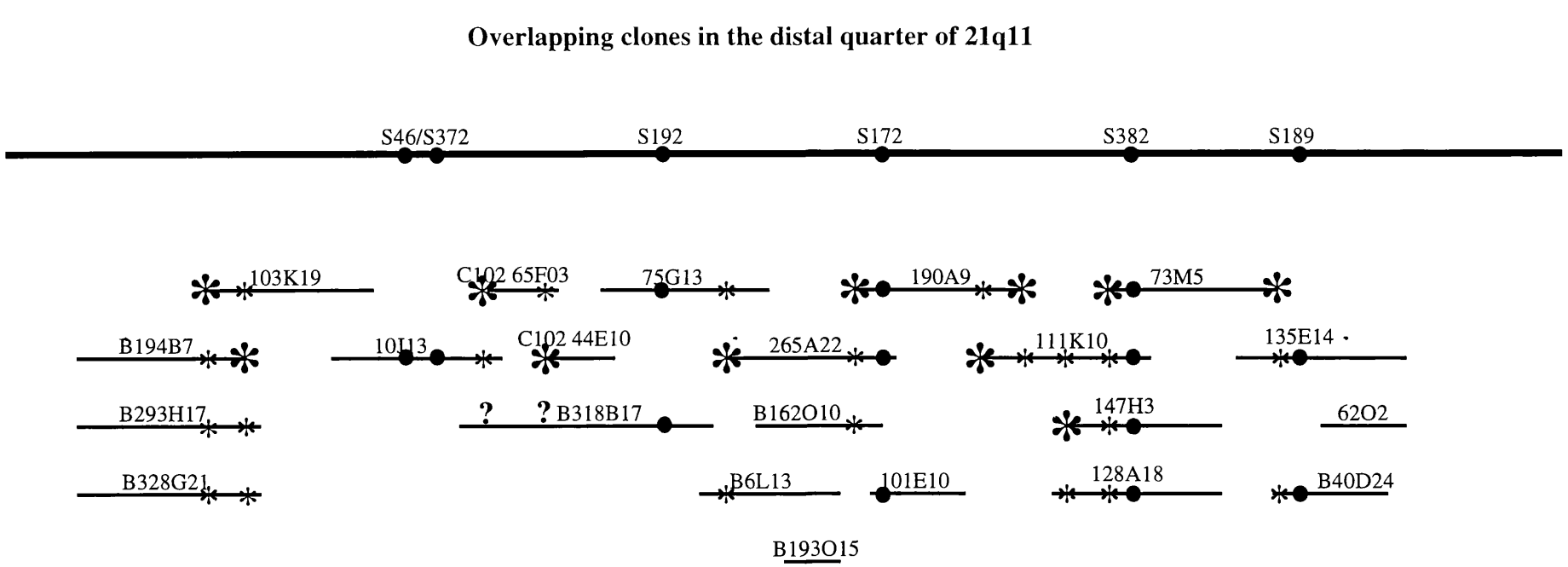


Figure 3.8: A contig is shown of genomic inserts of bacterial clones in the 21q11 region including the markers D21S46-D21S189. The contig was generated by hybridisation of the markers D21S46, D21S192, D21S172 and D21S189. Subsequent hybridisations of end probes of the different clones generated by IPCR and vectorette PCR, hybridisation of genomic inserts from clones and PCR amplification of markers, as indicated in Table 3.5, contributed to the generation of this contig. ✻ -endprobe generated, ⁎ -clones hybridised on the 21q11 filters by the endprobe (as indicated in Table 3.5), ?-clones not tested. ● -clones either positive by hybridisation (for D21S46, D21S192, D21S172 and D21S189) or by PCR for the other markers.

the pockets 22, 23 and 24 (the cosmid pocket map is described in Nizetic *et al.*, 1994) and PAC clone 10J13 to the pockets 24, 25 and 26. No cosmids were common between the two hybridisations. DNA from the 24 cosmids from pocket 24 was isolated, as these pockets could contain clones that close the gap, and dot blots were made which included DNA from both PACs. An RNA probe generated from the T7 end of the cosmid c102 65F03 (A.P. South), which was the only pocket 24 cosmid hybridising to 10J13, identified c102 65F03 and 10J13. The T7 hybridisation of another cosmid clone, c102 44E10 from pocket 24, hybridised to c102 44E10 and c102 65F03 (A.P. South). When *Sau* 3A I fragments from the clone c102 44E10 were hybridised to the total human PAC library, they identified one clone, 75G13. Instead of closing the gap between 10J13 and 71F7, these experiments closed the gap between 10J13 and 75G13. This can be explained, for example, by deletions in YAC clones which would place cosmids in the wrong pockets. Closing this gap, linked the marker D21S46 to the markers D21S192-D21S172-D21S189 and oriented the contig with D21S46 centromeric to D21S192 (Van Hul and Van Broeckhoven, 1993) The insert hybridisation of 10J13 to the total human PAC library identified two clones 10J13 and 103K19. One end fragment was generated from this clone, 103K19-DOP primer 1 (Wu *et al.*, 1996; Chapter 2.19.3), which hybridised to 103K19, B194B7, B293H17 and B328G21 (see Chapter 3.3.2.3). This result was confirmed by the hybridisation of the IPCR fragment generated from the end of the genomic insert from B194B7, which was digested with *Msp* I and amplified with primers B1 and B2 (B194B7-M B1/B2).

Additional clones in this region were identified by hybridising inserts of 10J13, 265A22 and 135E14 to the total human BAC library. B318B17 was identified by 10J13 and was shown to contain the marker D21S192. The BAC clones B6L13, B162O10 and B193O15 hybridised to the PAC clone 265A22. These clones were positioned by hybridisation of some of the insert end clones mentioned earlier in this section; 265A22-H1/V hybridised to B6L13, and 190A9-P1/V identified B162O10. The genomic insert from the clone 75G13 hybridised to B6L13 and B162O10 (not shown). 111K10-N3/4 failed to hybridise to these clones and none of the BAC clones contained D21S192 (except B318B17). The BAC clone B40D24 was identified by hybridisation to 135E14.

Two additional markers were placed on the contig: primers for D21S372 showed amplification of a specific PCR product with PAC clone 10J13, and the clones 73M5, 111K10, 128A18 and 147H3 were shown to contain marker D21S382. The PCR and

hybridisation results can be seen in Table 3.5, and positioning of the genomic inserts using these results can be seen in Figure 3.8.

3.3.3 Physical measurement of the two gaps in the physical map and closure of gap 2

Estimation of the size of gap 1, which is located distally from the clone 248P15 and proximal from 214G14 (sections 3.3.2.1 and 3.3.2.2) was performed using DNA of WA17 (a hybrid cell line containing only human chromosome 21 DNA on a mouse background (Graw *et al.*, 1995)) and A9 (the mouse background cell line). Agarose plugs, containing genomic DNA of WA17 and A9, were prepared (Chapter 2.19.6), and the DNA from each plug was digested individually with *Mlu* I. The agarose plugs, which contained the DNA of approximately $7.5x10^6$ cells, were loaded on a pulsed field gel (PFG) and electrophoresis was performed, separating fragments in the region of 100-2000 Kbp. The gels were Southern-blotted onto Hybond N+ membranes which were used in hybridisation experiments, using genomic inserts from the PAC clones 29H4 and 98L15 as probes. In both hybridisations, a fragment of approximately 1060 Kbp was observed, which can be seen in Figure 3.9A. The size of this fragment corresponds well with the *Mlu* I fragment of 1100 Kbp in WA17, which was described as containing the markers D21S215, D21S149, D21S253 and D21S258 (Van Hul *et al.*, 1993). The most proximal of these two *Mlu* I restriction sites is located between the markers D21S190 and D21S215 (Van Hul *et al.*, 1993). In the restriction map generated from the bacterial clones which were mapped in this study, three *Mlu* I fragments were present between these two markers (located within the clones 127M18, 16C2 and B62L20, see Figure 3.13). The more distal of the two *Mlu* I sites is located at the position of the marker D21S258. D21S258 is located on the PAC clones 98L15 and 90B5 (Chapter 3.3.2.2) and these clones show the presence of an internal *Mlu* I site (see Figure 3.12 and 3.13). The gap size in panel A of Figure 3.9 is calculated using the *Mlu* I site present on PAC clone 16C2. Calculation of the distance from this restriction site to and including 248P15 is between 398 and 460 Kbp (size estimates are used from the restriction maps shown in Figure 3.12). The distance on the other side of the gap from 214G14 to the *Mlu* I site on 98L15 is calculated as between 426 and 451 Kbp. From these distances an estimate of the size of gap 1 of between 149 and 236 Kbp can be calculated. If, for the estimation of the gap size, the proximal *Mlu* I sites located on the

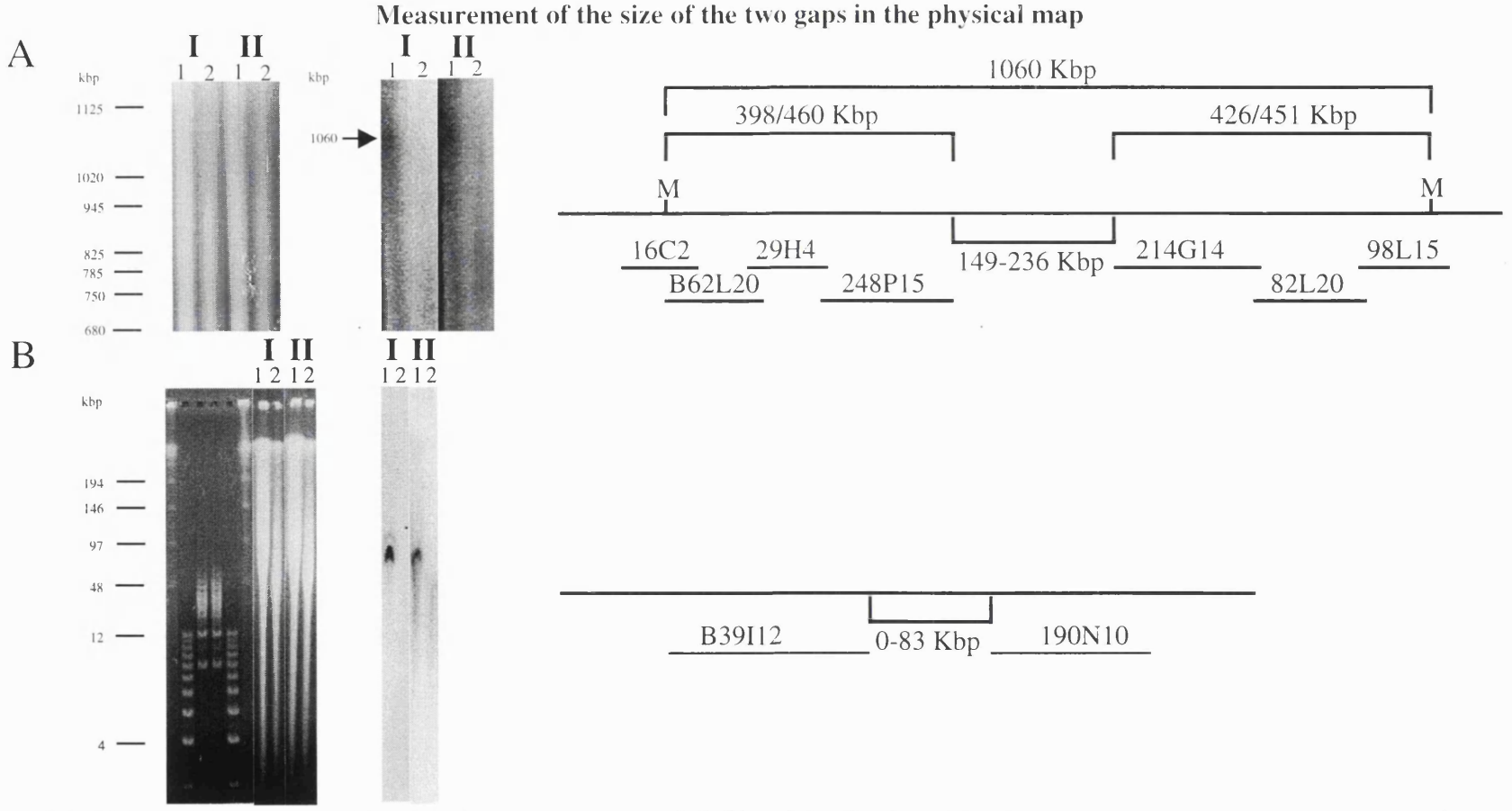


Figure 3.9: Sizing of the two gaps. Panel A shows an agarose gel (left) and the Southern blot that was generated using the agarose gel, after hybridisation (right). Lane 1 WA17, lane 2 A9 both digested with *Mlu* I. Blot I was hybridised with genomic insert from 29H4, blot II with genomic insert from 98L15. The size of the gap is indicated on the right. Panel B shows an agarose gel (left) of WA17 (lane 1) and A9 (lane 2), both digested with *Xho* I. A blot of this gel was hybridised using the genomic insert from B39I12 (I) and 190N10 (II). An identical 83 Kbp fragment was observed, indicating the gap to be smaller than 83 Kbp

clones 127M18 or B62L20 were used, then the estimated size would be 73 Kbp smaller or 37 Kbp larger respectively.

Estimation of the size of gap 2, which is located between the clones B39I12 and 190N10, was performed using DNA of WA17 and A9. Agarose plugs, containing genomic DNA of WA17 and A9, were prepared as described (see Chapter 2.19.6), and the DNA from each plug was digested individually with *Bam* H I, *Sma* I, *Xho* I and *Bst* E II. A PFG was prepared containing two identical sections in which the agarose plugs from the different restriction digests were applied in the same order (DNA from approximately $7.5x10^6$ cells per well). DNA markers were included for size estimation. Electrophoresis conditions were such that DNA fragments in the size range of approximately 5-250 Kbp were separated. The gel was stained with ethidium bromide and a Polaroid photo was generated after which the gel was Southern blotted. The two halves of the blot were separated and were used as targets for hybridisation experiments, with the probes B39I12-PB1/V and 190N10-H3/V. These fragments, when used as probes, were shown previously in separate hybridisation experiments only to hybridise to themselves and not to the neighbouring clone (the dot blots contained DNA of 31B5/B39I12 and 190N10/153I22 respectively), and it was concluded that they were generated from the ends of the genomic inserts nearest to the gap. The results of the hybridisation of B39I12-PB1/V and 190N10-H3/V to the PFG of the WA17 and A9 restriction digested DNA can be seen in Figure 3.9B. A specific band of approximately 83 Kbp can be seen in the *Xho* I digested WA17 both for B39I12-PB1/V and the 190N10-H3/V hybridisation. This indicated the maximum size of the gap to be 83 Kbp (see Figure 3.9B).

Genomic inserts of the bacterial clones B39I12 and 190N10 were hybridised to the chromosome 21 cosmid libraries ICRF c102/103 (Nizetic *et al.*, 1991a). The clones B39I12 and 190N10 hybridised to the cosmid pockets (Nizetic *et al.*, 1994) 19/20 and 20/21/22 respectively, indicating that cosmids from pocket 20 might be the most likely to bridge this gap. DNA was isolated from the 51 cosmids that were present in this pocket and dot blots were made on Hybond N+ (A.P. South). Hybridisation of the gap-facing end of clone B39I12 (B39I12-PB1/V) showed hybridisation to c102 68C10. This clone was previously shown to hybridise to the SP6 RNA probe of c102 132H4 (A.P. South), a 190N10 positive cosmid. These possible "gap-filling" clones were further analysed using additional hybridisation experiments. New dot blots were made containing DNA from the clones 190N10, c102 132H4, c102 68C10, and B39I12, and

Confirmation that two cosmid clones close gap 2

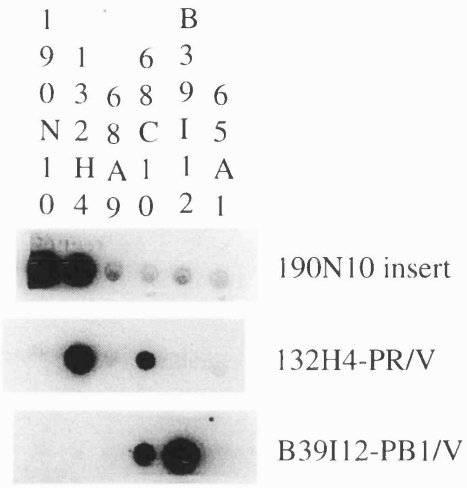


Figure 3.10: Three identical dot blots are shown. The position of the dots and the DNA they contain is indicated with the clone names above the blots. The probes that were used for the three dot blots are indicated on the right. The genomic insert of the PAC 190N10 hybridised to c102 132H4 and a vectorette end probe from this clone hybridised to c102 68C10. This latter clone is also identified by the vectorette end probe B39I12-B1/V from genomic clone B39I12. This closes the gap between 190N10 and B39I12 via c102 132H4 and c102 68C10. The cosmids c102 65A1 and c102 68A9 are negative controls.

two clones that were used as negative controls (c102 65A1, c102 68A9). As can be seen from Figure 3.10, the genomic insert from the PAC clone 190N10 hybridised to 190N10 and c102 132H4. A vectorette-derived fragment from the end of cosmid c102 132H4 (132H4-PR/V) hybridised to c102 132H4 and c102 68C10. B39I12 PB1/V, a vectorette end probe from BAC B39I12, hybridised to B39I12 and c102 68C10. These experiments confirmed that clones c102 68C10 and c102 132H4 were overlapping clones and closed the gap between B39I12 and 190N10.

3.3.4 Restriction mapping of the genomic DNA contained in bacterial clones, verification of overlaps, and measurement of the degree of overlap between the clones

From the bacterial clones that were identified in Chapter 3.3.2 and 3.3.3, a minimal tiling path (MTP) of overlapping clones was chosen for further analysis. These clones were 77P15, 156L9, 127M18, 16C2, B62L20, 29H4, 248P15, 214G14, 82L20, 98L15, 90B5, 126N20, 31B5, B39I12, c102 68C10, c102 132H4, 190N10, 153I22, 270D8, 30P13, 270M7, B194B7, 103K19, 10J13, c102 65F3, c102 44E10, 75G13, 265A22, 190A9, 111K10, 73M5 and 135E14.

From these clones a restriction map, using the enzymes *Not* I, *Xho* I and *Sal* I was constructed using single, double and triple digestions and PFGE. An example of this can be seen in Figure 3.11A, which shows a photograph of a PFG with fragments generated by the enzymes mentioned above, of the PAC clone 31B5. As can be seen from Figure 3.11D, the PAC vector is 17 Kbp long, contains two *Not* I sites on either side of the vector, a *Sal* I site at the SP6 end of the vector and two *Xho* I sites, one approximately 7 Kbp from the SP6 end, and one 5 Kbp from the T7 end. Lanes 2 and 6 in Figure 3.11A show that there are no extra *Sal* I and *Not* I sites present in the insert; the 17 Kbp fragment in lanes 6 and 7 is the vector fragment between the two *Not* I sites. Lane 1, in which the clone is digested with *Xho* I, shows fragments of approximately 50, 40, 30, 15 and 5 Kbp. The 5 Kbp fragment is an internal PAC vector fragment. When lane 1 is compared with lane 3, in which clone 31B5 is restriction digested with *Xho* I and *Xho* I/*Sal* I respectively, the 30 Kbp fragment is reduced to two fragments of 23 and 7 Kbp, indicating that a 23 Kbp fragment with an *Xho* I restriction site borders the SP6 end of the vector. The 15 Kbp fragment is reduced to 10 and 5 Kbp by digestion with *Xho* I and *Xho* I/*Not* I respectively (compare lanes 1 and 5). This places a *Xho* I site 10 Kbp from the T7 end of the PAC vector. Consequently, the 40 and 50 Kbp *Xho* I fragments

Restriction mapping and verification of the degree of overlap

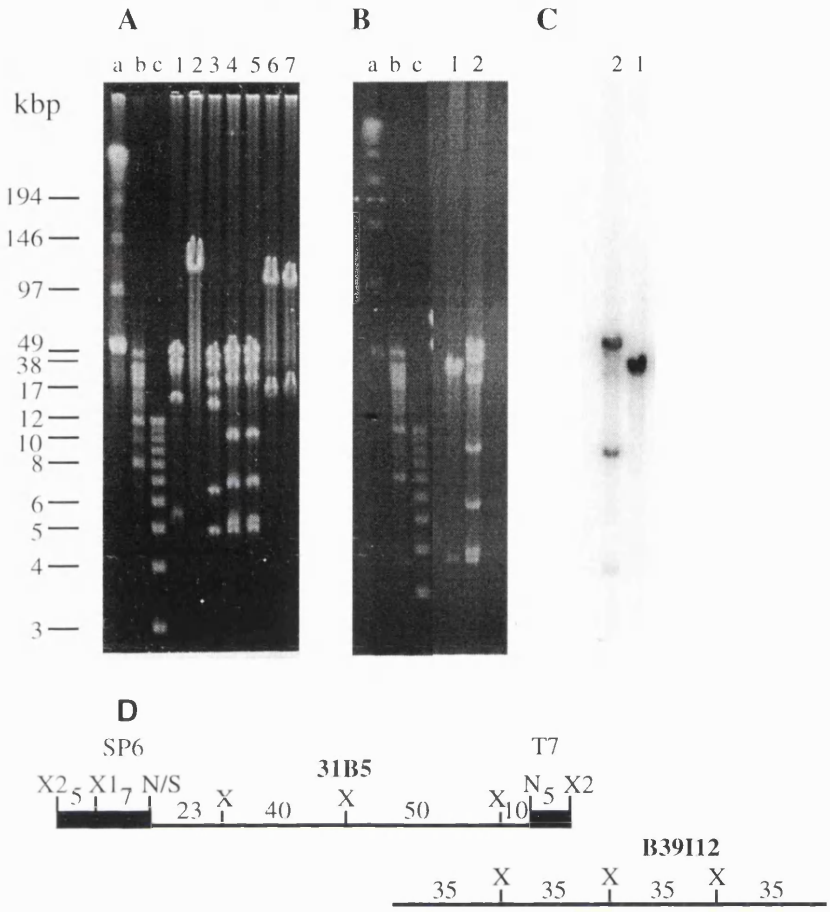


Figure 3.11: Panel A-restriction mapping of PAC 31B5 using PFGE. Lane 1-*Xho* I; lane 2-*Sal* I; lane 3-*Xho* I/*Sal* I; lane 4-*Not* I/*Xho* I/*Sal* I; lane 5-*Not* I/*Xho* I; lane 6-*Not* I; lane 7-*Not* I/*Sal* I. Panel B-PFGE of two overlapping clones from the MTP. Lane 1-BAC B39I12 digested with *Not* I/*Xho* I/*Sal* I, lane 2-PAC 31B5 restriction digested with *Not* I/*Xho* I. The lanes numbered a, b, c contain molecular weight markers, a-λ ladder, b-high molecular weight markers, c-1 kb ladder. Panel C-Verification and sizing of the overlapping segment; the Southern blot from B was hybridised to the genomic insert from B39I12, the 10 and 50 Kbp fragments hybridised positively. The cross-hybridising 5 Kbp fragment is a vector-derived fragment. Panel D-Restriction map of 31B5 and the overlap of the two clones as deduced from A, B and C. N-*Not* I; S-*Sal* I; X1, X2-two *Xho* I sites in the vector. SP6 and T7 are the RNA polymerase promoters flanking the *Bam* H I cloning site of the PAC vector.

must therefore be internal. The clones from the MTP were additionally restriction digested with the enzymes *Bss*H II, *Sac* II, *Mlu* I and *Eag* I. Single digestions using these enzymes individually or in combination with *Not* I, *Sal* I and *Xho* I were performed to position these restriction enzyme sites to the clones of the MTP. The enzymes *Bss*H II, *Sac* II, *Mlu* I and *Eag* I, like *Not* I, contain CpG dinucleotides in their recognition sites. Clusters of CpG dinucleotides, known as CpG islands, are known to be often associated with 5' ends of housekeeping genes (Lindsay and Bird, 1987). Three CpG islands have been identified previously (Van Hul and Van Broeckhoven, 1993), two of which correspond to the *Not* I restriction sites from the *Not* I physical map (Ichikawa *et al.*, 1993). These can be found at D21S13 and D21S382. Data presented here show that the same restriction enzyme recognition sites were found at these two genomic *Not* I sites compared with those reported (Van Hul and Van Broeckhoven, 1993); these are present on the clones 270M7 (D21S13) and 111K10 (D21S382). The third CpG island is present approximately 90 Kbp centromeric from D21S16 (Van Hul and Van Broeckhoven, 1993). These authors used a number of restriction enzymes to identify the CpG island, however the work described here utilised two of these enzymes; *Sac* II and *Eag* I. These two enzyme sites are present on the clones 90B5 and 126N20, approximately 90 Kbp centromeric from D21S16 (see Figures 3.12 and 3.13). The complete results can be seen in Figure 3.12, which shows all the restriction mapped clones.

Hybridisation experiments were carried out between adjacent clones to reconfirm overlaps. Neighbouring clones were restriction digested with *Not* I, *Sal* I and *Xho* I (or with *Not* I/*Xho* I, when *Sal* I sites were not present in the genomic insert) and PFGE was performed. The PFG was blotted onto Hybond N+ and the resulting membrane was used in a hybridisation experiment in which the radio-actively labelled genomic insert of one of the clones, also present on the filter, was used as a probe. An example of this can be seen in Figure 3.11B and C. Figure 3.11B shows a PFG of the *Not* I/*Sal* I/*Xho* I digested BAC B39I12 (lane 1) and *Not* I/*Xho* I digested PAC 31B5. The BAC B39I12 consists of four *Xho* I fragments of approximately 35 Kbp (Figure 3.11D). This gel was blotted onto a Hybond N+ filter, which was hybridised to radio-actively labelled genomic insert of B39I12. The result of this hybridisation can be seen in Figure 3.11C. The insert of B39I12 hybridises to the 10 and 50 Kbp fragments of 31B5 resulting in an overlap, as indicated in Figure 3.11D. This was performed with the clones throughout the MTP. These experiments were performed to reconfirm overlaps and to estimate the

Restriction mapping and determining of the degree of overlap of the minimal tiling path clones

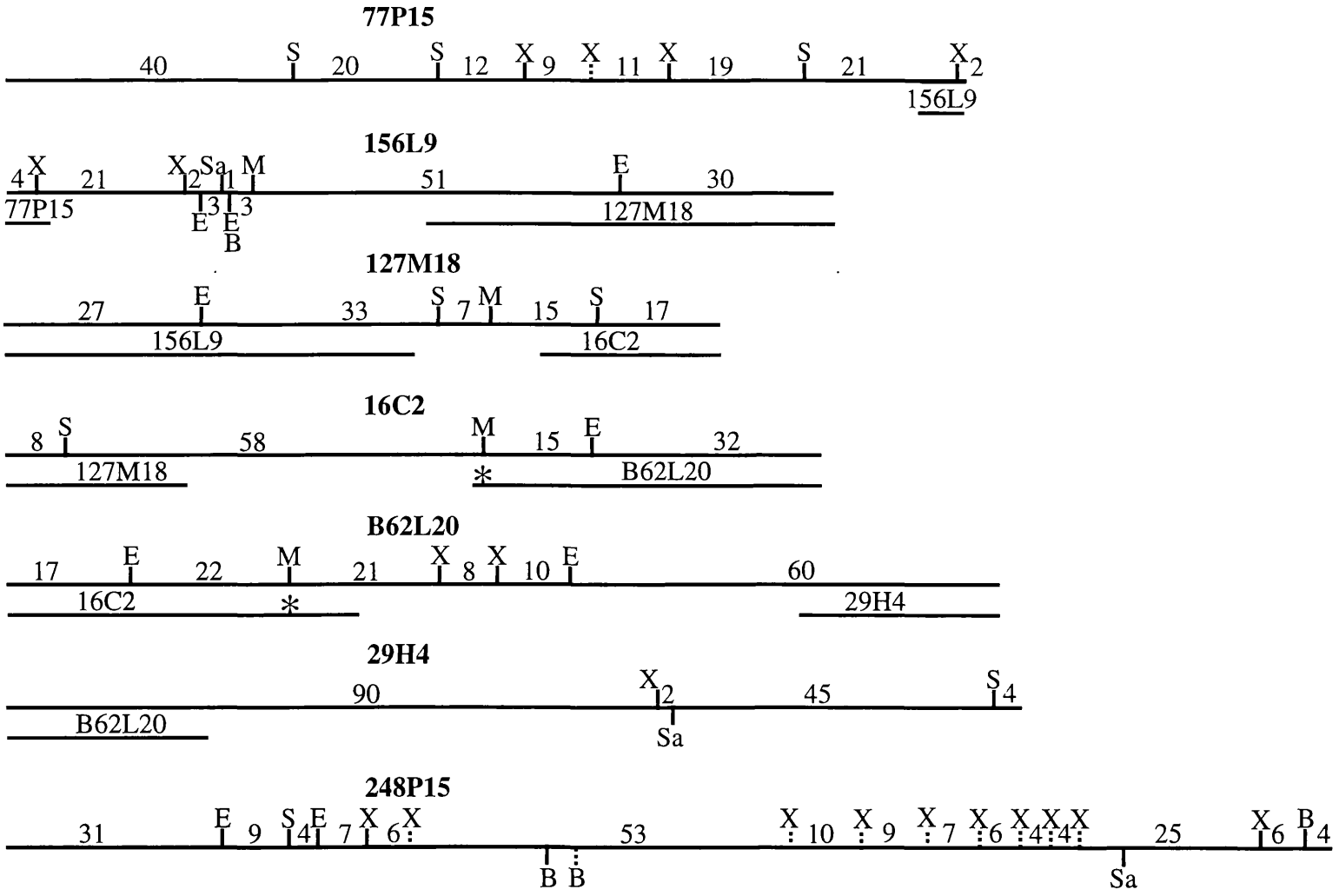


Figure 3.12

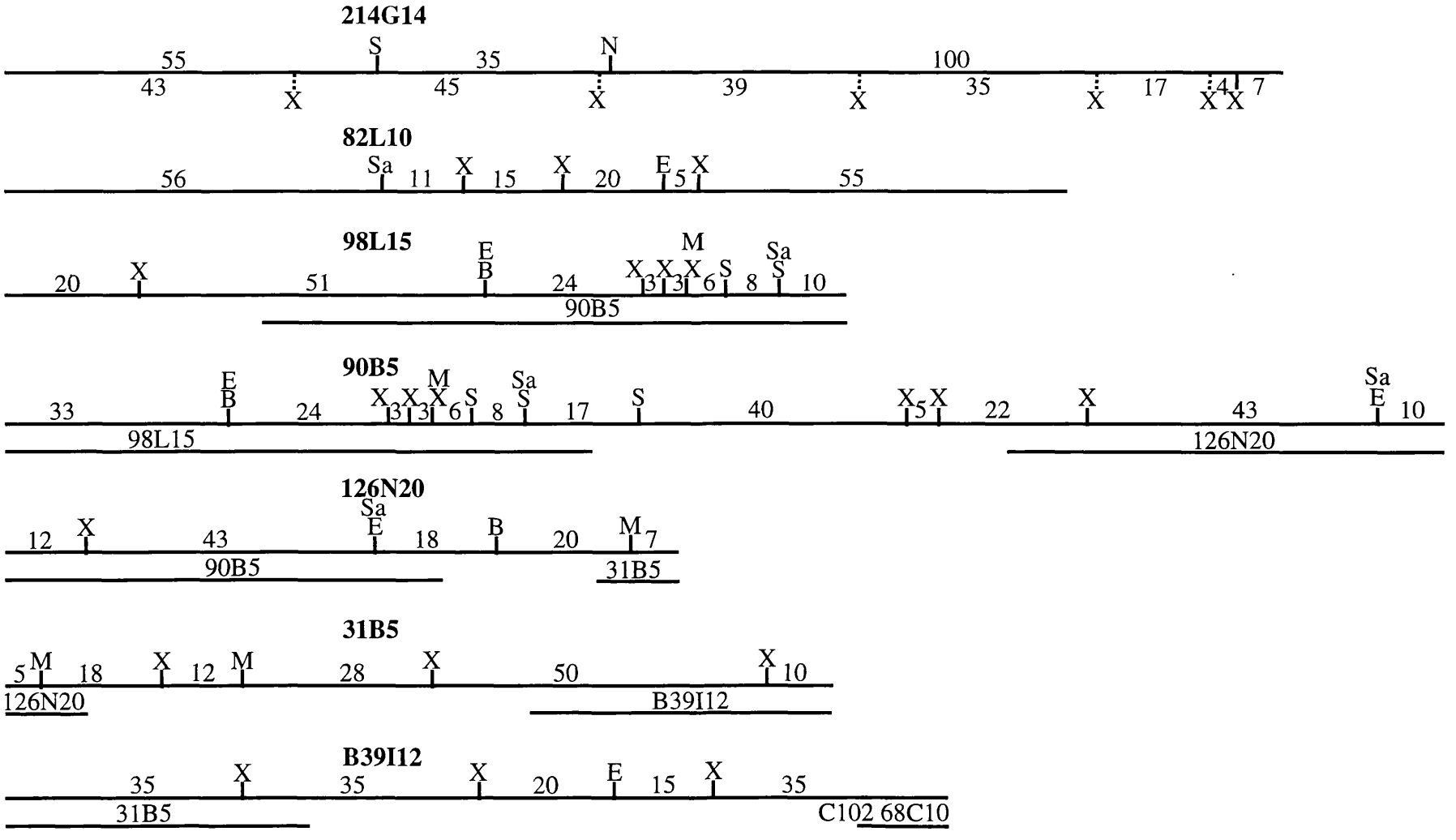

214G14
55
S
35
N
100
43
X
45
X
39
X
35
X
17
4
7
X
X
82L10
56
Sa
11
X
15
X
20
E
5
X
55
98L15
20
X
51
E
B
24
M
X
3
X
3
X
6
S
8
Sa
S
10
90B5
90B5
33
E
B
24
X
3
X
3
X
6
M
S
8
Sa
S
17
S
40
X
5
X
22
X
43
Sa
E
10
98L15
126N20
126N20
12
X
43
Sa
E
18
B
20
M
7
90B5
31B5
31B5
5
M
18
X
12
M
28
X
50
X
10
126N20
B39I12
B39I12
35
X
35
X
20
E
15
X
35
31B5
C102
68C10

Figure 3.12

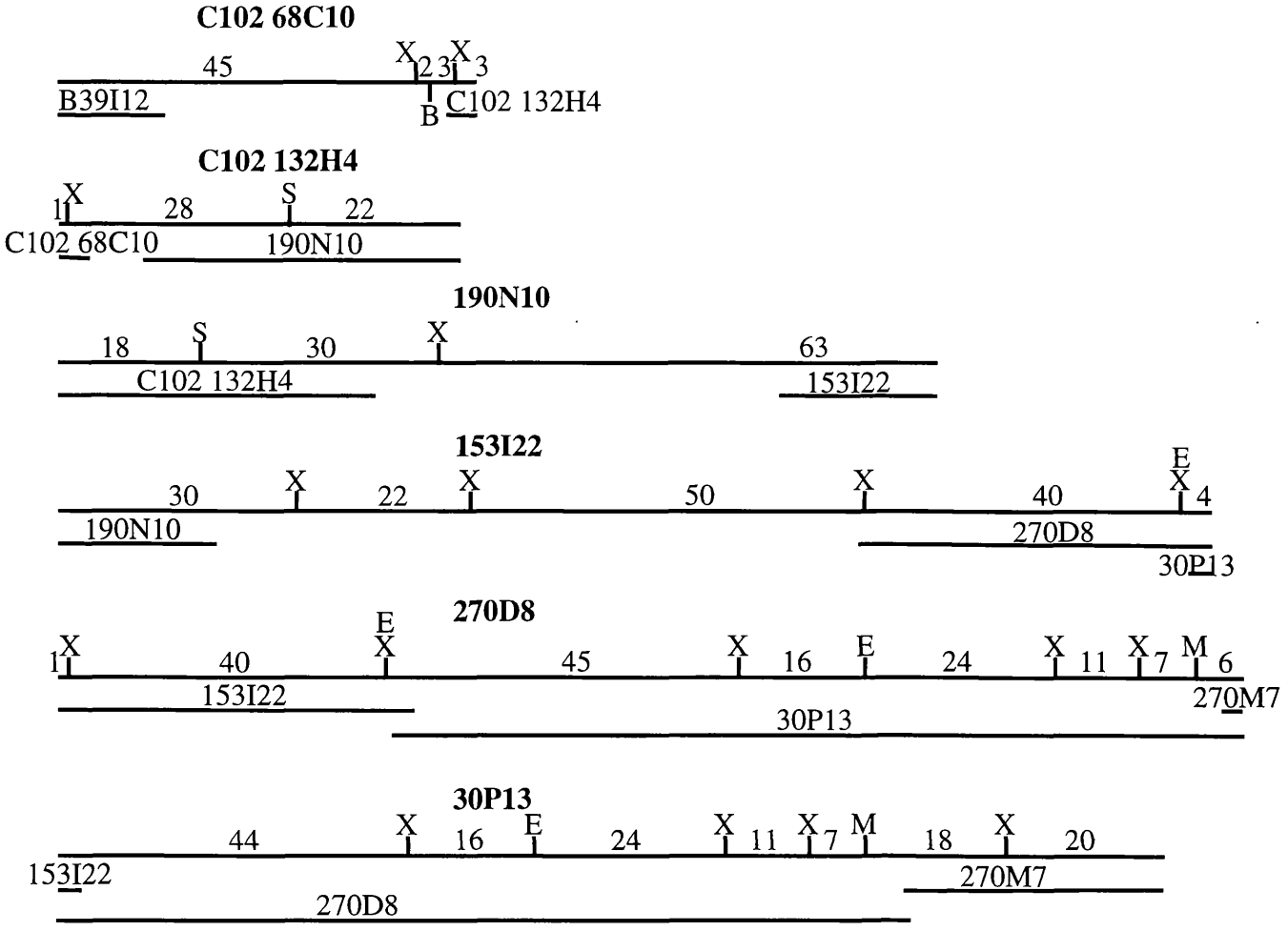

C102 68C10
45
X
2
3
X
3
B39I12
B
C102 132H4
C102 132H4
X
1
28
S
22
C102 68C10
190N10
190N10
18
S
30
X
63
C102 132H4
153I22
153I22
30
X
22
X
50
X
40
E
X
4
190N10
270D8
30P13
270D8
X
1
40
E
X
45
X
16
E
24
X
11
X
7
M
6
153I22
270M7
30P13
30P13
44
X
16
E
24
X
11
X
7
M
18
X
20
153I22
270M7
270D8

Figure 3.12

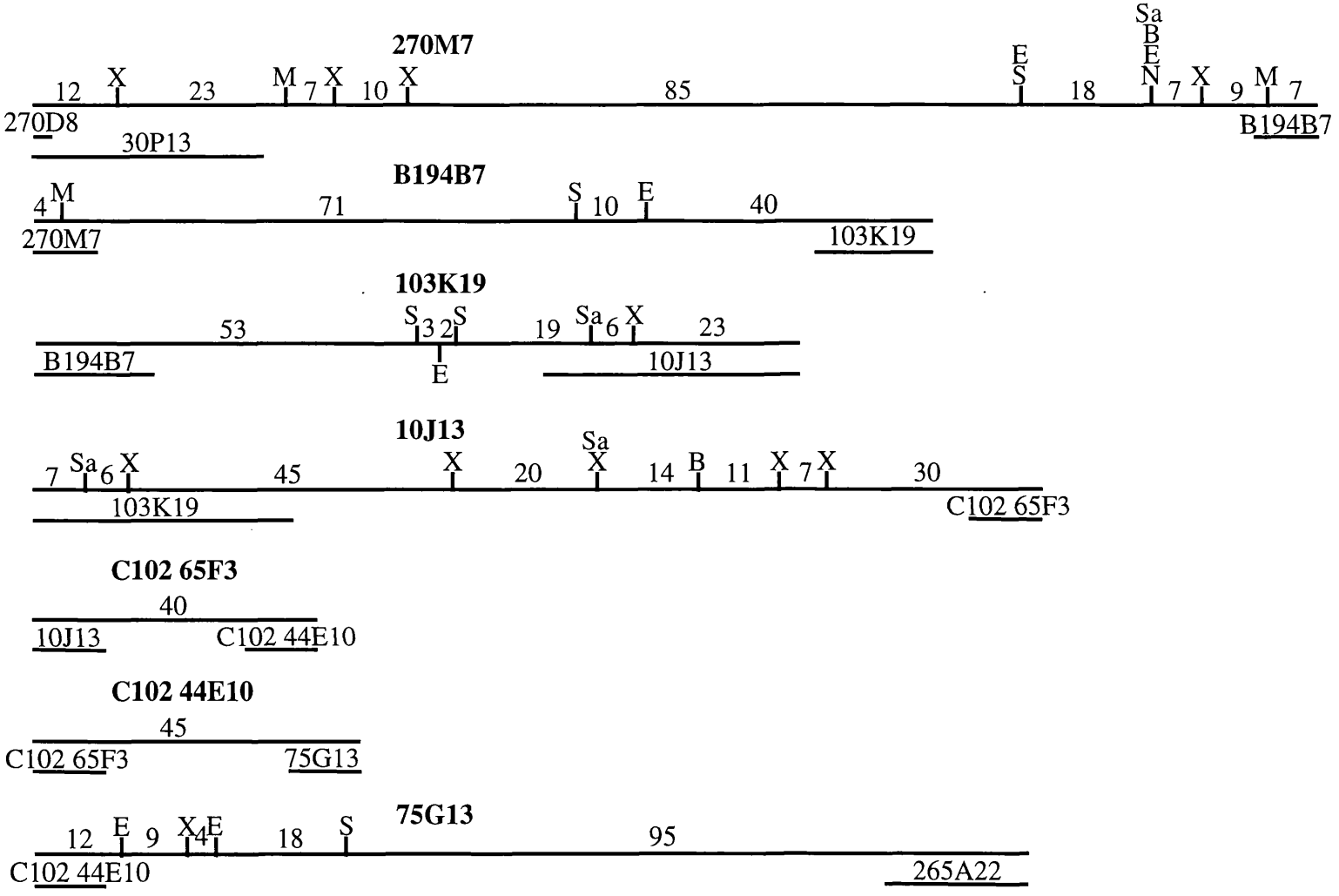

270M7
12 X 23 M 7 X 10 X 85 E S 18 Sa B E N 7 X 9 M 7
270D8
30P13
B194B7
B194B7
4 M 71 S 10 E 40
270M7
103K19
103K19
53 S 3 2 S 19 Sa 6 X 23
B194B7
E
10J13
10J13
7 Sa 6 X 45 X 20 Sa X 14 B 11 X 7 X 30
103K19
C102 65F3
C102 65F3
40
10J13
C102 44E10
C102 44E10
45
C102 65F3
75G13
75G13
12 E 9 X 4 E 18 S 95
C102 44E10
265A22

Figure 3.12

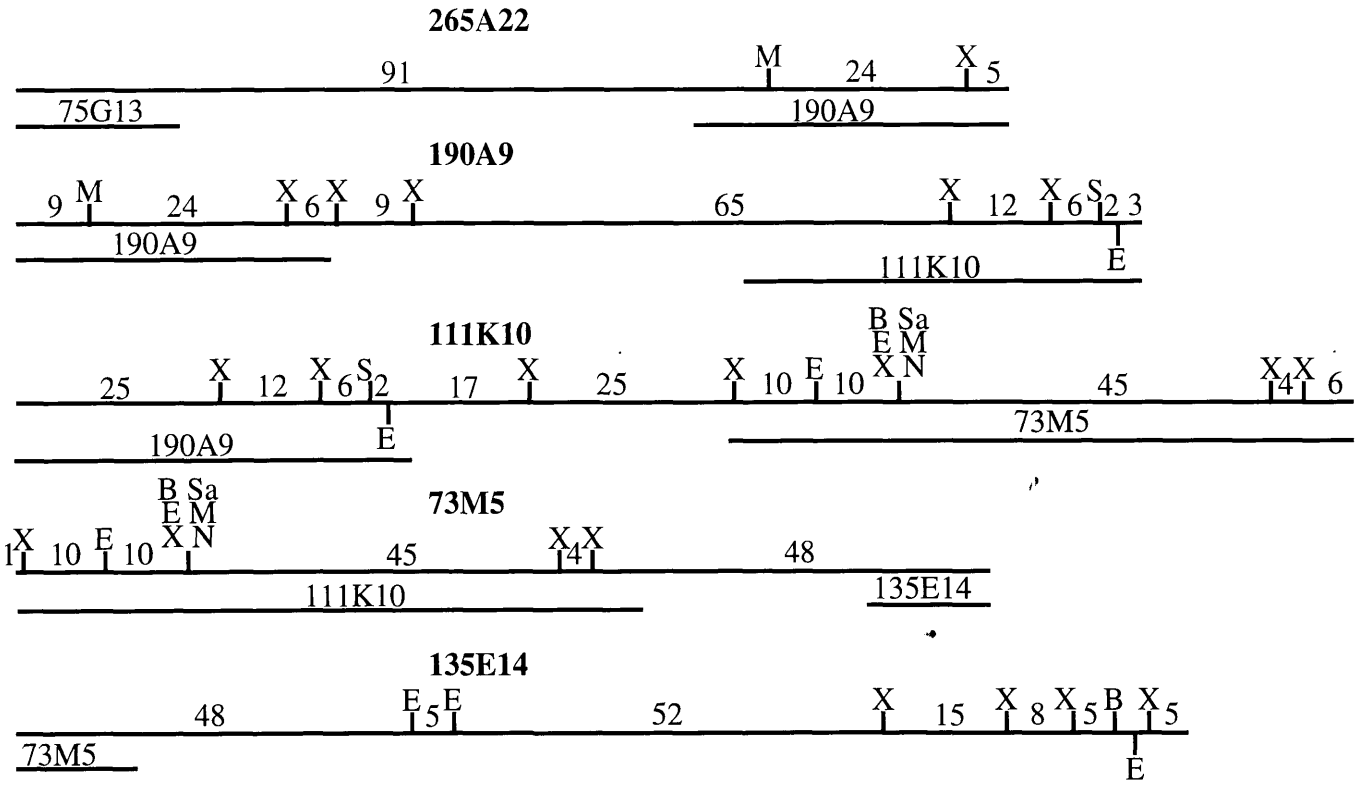


Figure 3.12: Restriction mapping of the Minimal Tiling Path clones. Restriction maps were constructed using digestions of single and combinations of the following enzymes; X-*Xho* I, S-*Sal* I, N-*Not* I, B-*Bss* H II, M-*Mlu* I, E-*Eag* I, Sa-*Sac* II. Subsequently, neighbouring clones were restriction digested with N, X and S or with N, X and a PFGE was carried out, followed by Southern blotting. The genomic insert from one of the clones present on the filter was used in a hybridisation experiment, to identify the fragments in common with overlapping clones. These experiments were performed to reconfirm overlaps and to determine the extent of the overlap between clones. The lines on the left and right hand side underneath the clones indicate the overlap. Restriction endonuclease recognition sites were identical in the overlapping regions unless indicated by ∗. An example of these experiments can be seen in Figure 3.9. Dotted lines indicate that the order of the fragments could not be determined. Numbers indicate distances in Kbp.

degree of overlap between the clones. This was important to establish the total length of the contig and gave an indication of the amount of redundancy for the genomic DNA sequencing centres. All clones showed clear overlaps when inserts of neighbouring clones were used as probes in hybridisation, except for the overlaps between 29H4 and 248P15, 214G14 and 82L10, and 82L10 and 98L15. This could be due to the fact that these clones only have a minimal overlap. The insert of 248P15 hybridised weakly to 29H4 on the total human PAC library (Chapter 3.3.2.1). From the clones proximal to 77P15, no restriction map is indicated in Figure 3.12 as no consensus in the restriction map in the overlap between 77P15 and the more proximal clones could be found. Similarity in DNA sequence between different chromosomes, as has been shown for this region (Van Camp *et al.*, 1992; Van Hul *et al.*, 1993), could be one of the reasons for this problem.

3.3.5 Localisation of published genes and transcribed DNA fragments on the 21q11 contig map

Two genes have been reported to map in the area; STCH and RIP140. STCH is a member of the stress70 chaperone family (Otterson *et al.*, 1994). RIP140 is a nuclear protein that interacts *in vitro* with the AF2 domain of the oestrogen receptor. STCH has been localised to the PAC clones 90B5 and 126N20 (A.P. South). This was performed by hybridisation of a radio-active probe, generated from an RT-PCR product which was specific for part of the STCH gene (A.P. South), to nylon membranes containing *Eco* R I digested MTP clones. RIP140 is localised in the overlap between the clones 30P13 and 270M7 (Katsanis *et al.*, 1998).

A radio-active probe of TU18, an exon trapped clone which was mapped to the 21q11 YACs 5A11y21 and 2C12y21 (Yaspo *et al.*, 1995), was hybridised to the 21q11 filters. Two clones hybridised to this exon trapped clone, 190A9 and 111K10 indicating the mapping position of TU18 (J.H. Ives).

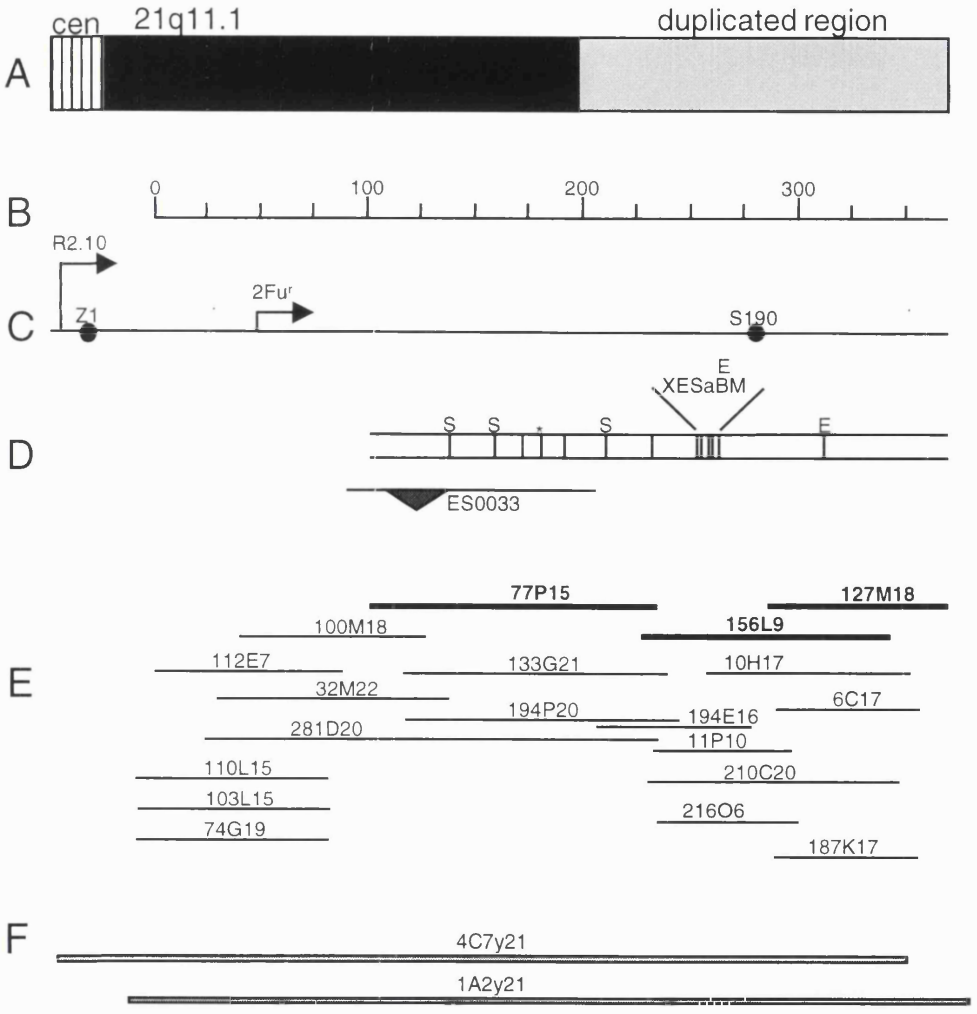
Six unique clusters of expressed human sequences (part of the Unigene collection, Schuler *et al.*, 1996 and http://www.ncbi.nlm.nih.gov/Unigene/Hs.Home.html/) which were mapped between the markers D21S1904 and D21S1905 (these markers include human chromosome 21p, 21q11 and part of 21q21), were analysed for possible positioning on the contigs generated in this study. These Unigene cDNA clusters were: the enteropeptidase precursor gene (accession number U09860) and the unidentified transcripts A002B43, Bdac5g11, A004O28, SHGC-11705 and WI-13495. These

Unigene cDNA clusters were subsequently mapped to the 21q11 contig by amplification of specific PCR fragments from the Unigenes. The primers that were used were identical to the ones used to map the Unigene cDNA clusters onto the radiation hybrid map (see http://www.ncbi.nlm.nih.gov/Unigene/Hs.Home.html/). The following primers were used (Chapter 2.9.4): 12603-A and –B (enteropeptidase precursor gene), 2B43-A and –B (A002B43), Bdac-A and -B (Bdac5g11), 4O28-A and -B (A004O28), 11705-A and –B (SHGC-11705) and 13495-A and –B (WI-13495). Pools were made in which DNA from four or five MTP clones (see section 3.3.4) were mixed, after which PCR analysis was performed with the above mentioned primer combinations. A positive control containing genomic DNA amplified a specific product and a H_2O control was negative in all cases. The Unigene cluster A004O28 amplified a specific product in the pool containing DNA from the clones B194B7, 103K19, 10J13, 75G13 and 265A22, PCR on DNA from individual clones showed mapping of this Unigene to the PAC clone 265A22. A specific product was amplified for A002B43 in the pool containing the clones 190A9, 111K10, 73M5 and 135E14. A PCR using these clones individually, localised this Unigene cDNA cluster to 135E14. The other Unigene clusters failed to amplify a PCR product from the MTP clones. The Unigene cDNA cluster Bdac5g11 was subsequent localised to chromosome 16, the enteropeptidase precursor gene, SHGC-11705 and WI-13495 are probably located distal from the 21q11 region (as described in the website http://www.ncbi.nlm.nih.gov/Unigene/Hs.Home.html/). The position of the genes, cDNA selected clones, the exon trapped clone and the Unigene clusters are indicated in Figure 3.13.

3.3.6 Integrated map of 21q11

When all the data of the sections 3.3.2-3.3.5 are combined, an integrated map of the 21q11 region can be established which is shown in Figure 3.13. This map integrates the information from the markers located on the different bacterial clones, the positioning of the genomic inserts from the bacterial clones and mapping of cDNA selected material (Cheng *et al.*, 1994) as described in section 3.3.2. It also includes information on the size of the gap (section 3.3.3) and a restriction map of the MTP clones from which the degree of overlap between the clones was determined, as shown in section 3.3.4. The somatic cell hybrid lines containing partial chromosome 21 material were placed according to published mapping information (Graw *et al.*, 1995). Amplimers from the markers D21S369, D21S215, D21S297, D21S308, D21S318, D21S258, D21S236 and

Figure 3.13:The following four pages show the PAC/ BAC/ cosmid contig map of the 21q11 region. This page shows an example of the different sections. The figure consists of: **A**-cytogenetic banding. **B**-size scale in 100 Kilobasepairs (Kbp). **C**-schematic representation of the chromosome with markers (●), genes (▬) and intervals of different somatic cell hybrid cell lines containing breakpoints (┌►). **D**- Restriction map, generated from the minimal tiling path clones; no letter or X=*Xho* I, S=*Sal* I, N=*Not* I, E=*Eag* I, B=*Bss* H II, M=*Mlu* I, Sa=*Sac* II, * indicates that the order of *Xho* I fragments could not be determined. **E**-PAC/BAC/cosmid contig and putative transcripts (▼). The clones that form the minimal tiling path are indicated in bold. These clones are positioned according to the restriction map and the overlaps between adjacent clones. **F**- YACs in the 21q11 region. The PAC/ BAC and cosmid clones were identified and positioned using the hybridisations and PCRs described in section 3.3.2, not all clones are shown. The restriction map in section D was shown in 3.3.4, in which the extent of the overlaps was also determined. The mapping positions of the putative transcripts were described in sections 3.3.2 and 3.3.5.



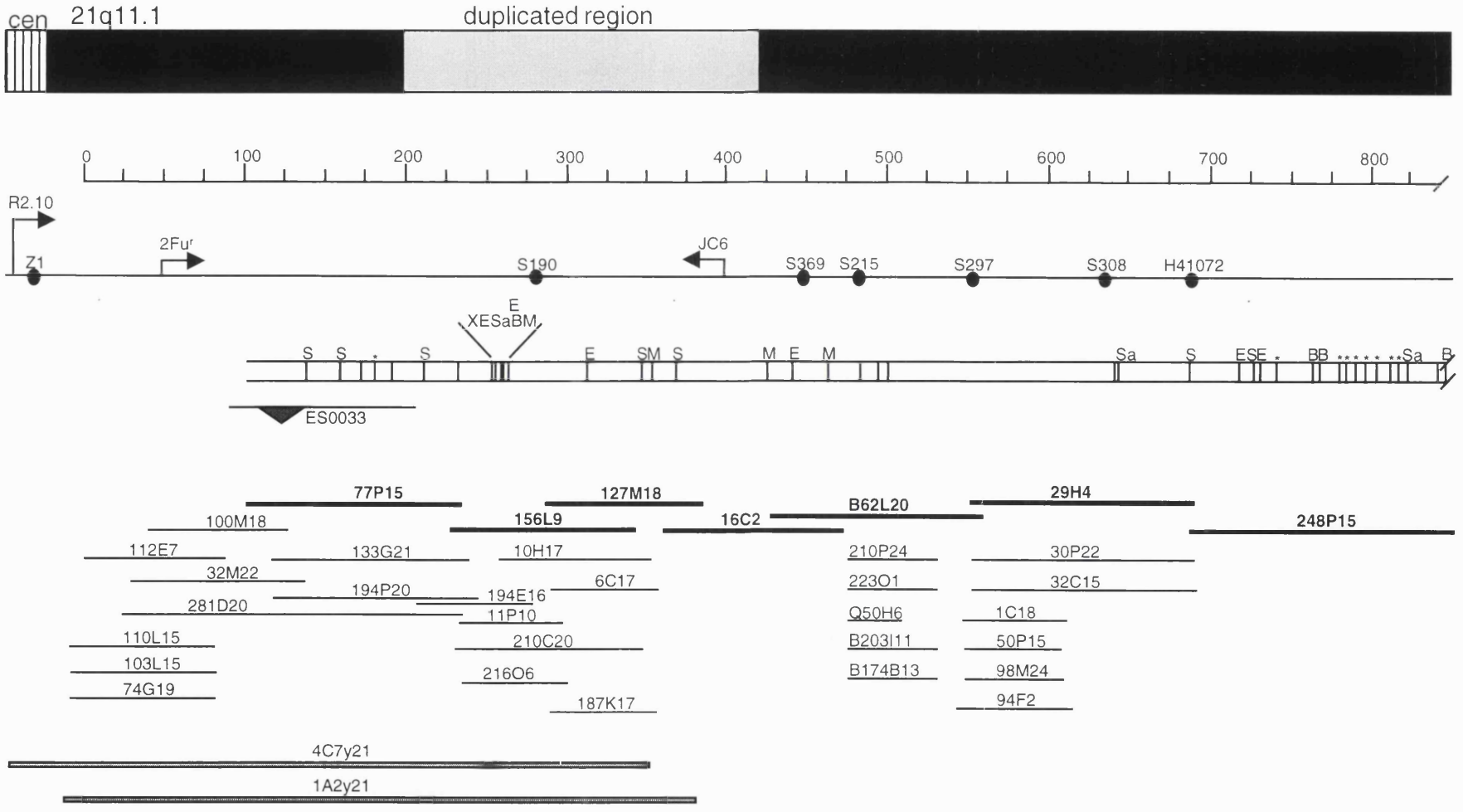

cen
21q11.1
duplicated region
0
100
200
300
400
500
600
700
800
R2.10
Z1
2Fur
S190
JC6
S369
S215
S297
S308
H41072
E
XESaBM
S S * S E SM S M E M Sa S ESE * BB * * * * * * Sa B
ES0033
77P15
127M18
B62L20
29H4
100M18
156L9
16C2
248P15
112E7
133G21
10H17
210P24
30P22
32M22
194P20
194E16
6C17
223O1
32C15
281D20
11P10
Q50H6
1C18
110L15
210C20
B203I11
50P15
103L15
216O6
B174B13
98M24
74G19
187K17
94F2
4C7y21
1A2y21

Figure 3.13

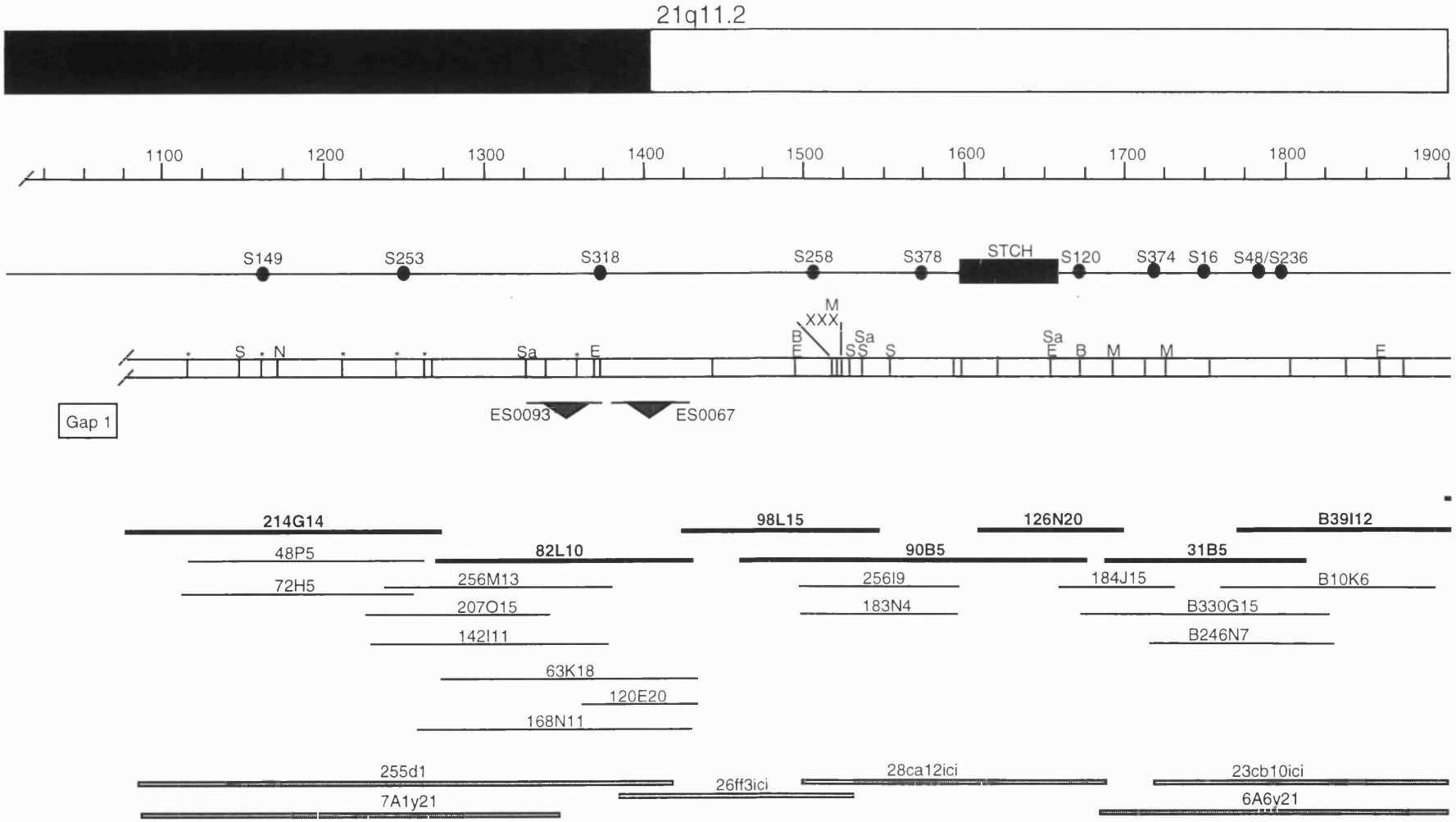

21q11.2
1100
1200
1300
1400
1500
1600
1700
1800
1900
S149
S253
S318
S258
S378
STCH
S120
S374
S16
S48/S236
M
XXX
B
F
Sa
S
N
Sa
E
S
S
Sa
F
B
M
M
F
Gap 1
ES0093
ES0067
214G14
98L15
126N20
B39I12
48P5
82L10
90B5
31B5
72H5
256M13
256I9
184J15
B10K6
207O15
183N4
B330G15
142I11
B246N7
63K18
120E20
168N11
255d1
26ff3ici
28ca12ici
23cb10ici
7A1y21
6A6y21

Figure 3.13

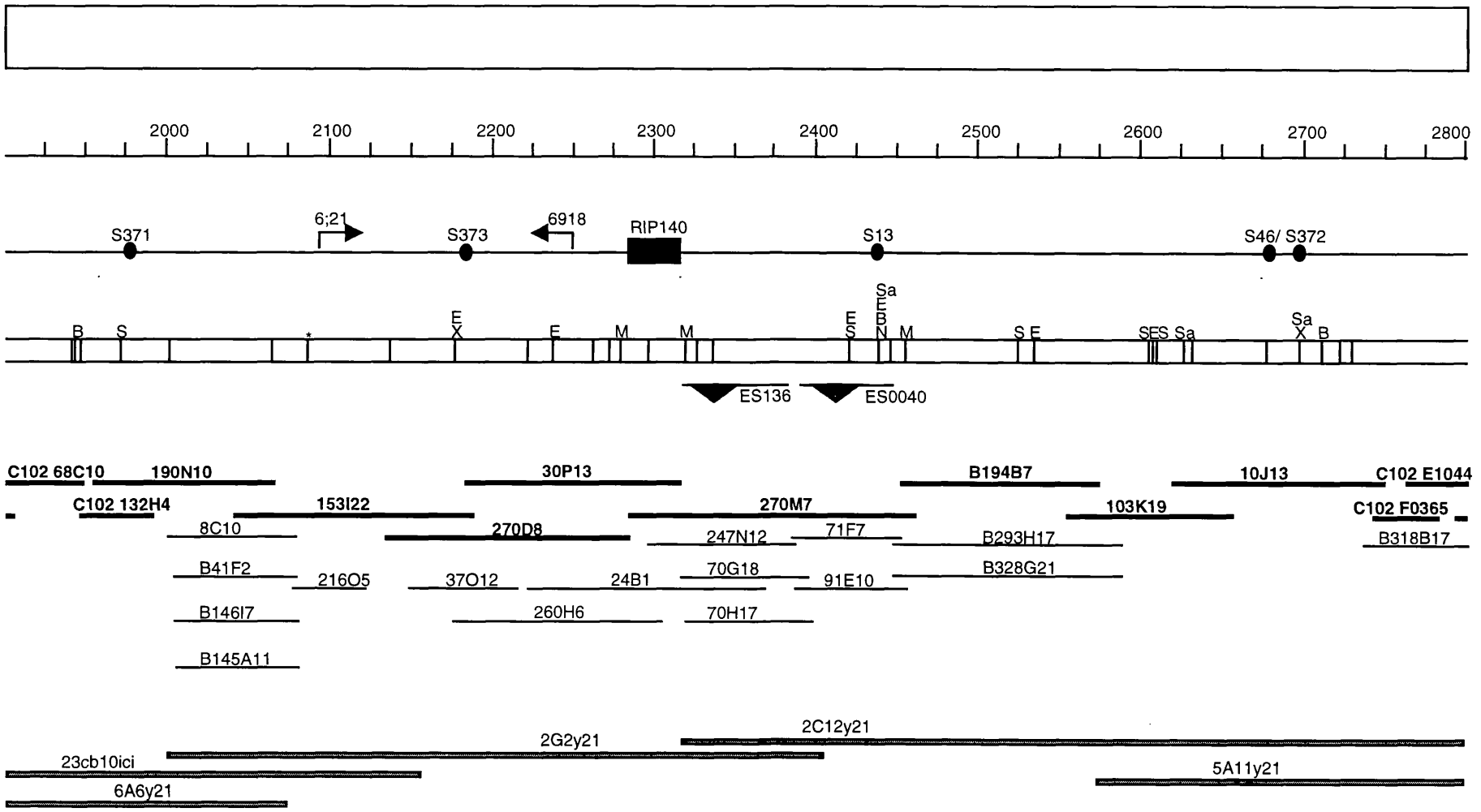

2000
2100
2200
2300
2400
2500
2600
2700
2800
S371
6;21
S373
6918
RIP140
S13
S46/ S372
ES136
ES0040
C102 68C10
190N10
30P13
B194B7
10J13
C102 E1044
C102 132H4
153I22
270M7
103K19
C102 F0365
8C10
270D8
247N12
71F7
B293H17
B318B17
B41F2
216O5
37O12
24B1
70G18
91E10
B328G21
B146I7
260H6
70H17
B145A11
2C12y21
2G2y21
23cb10ici
5A11y21
6A6y21

Figure 3.13

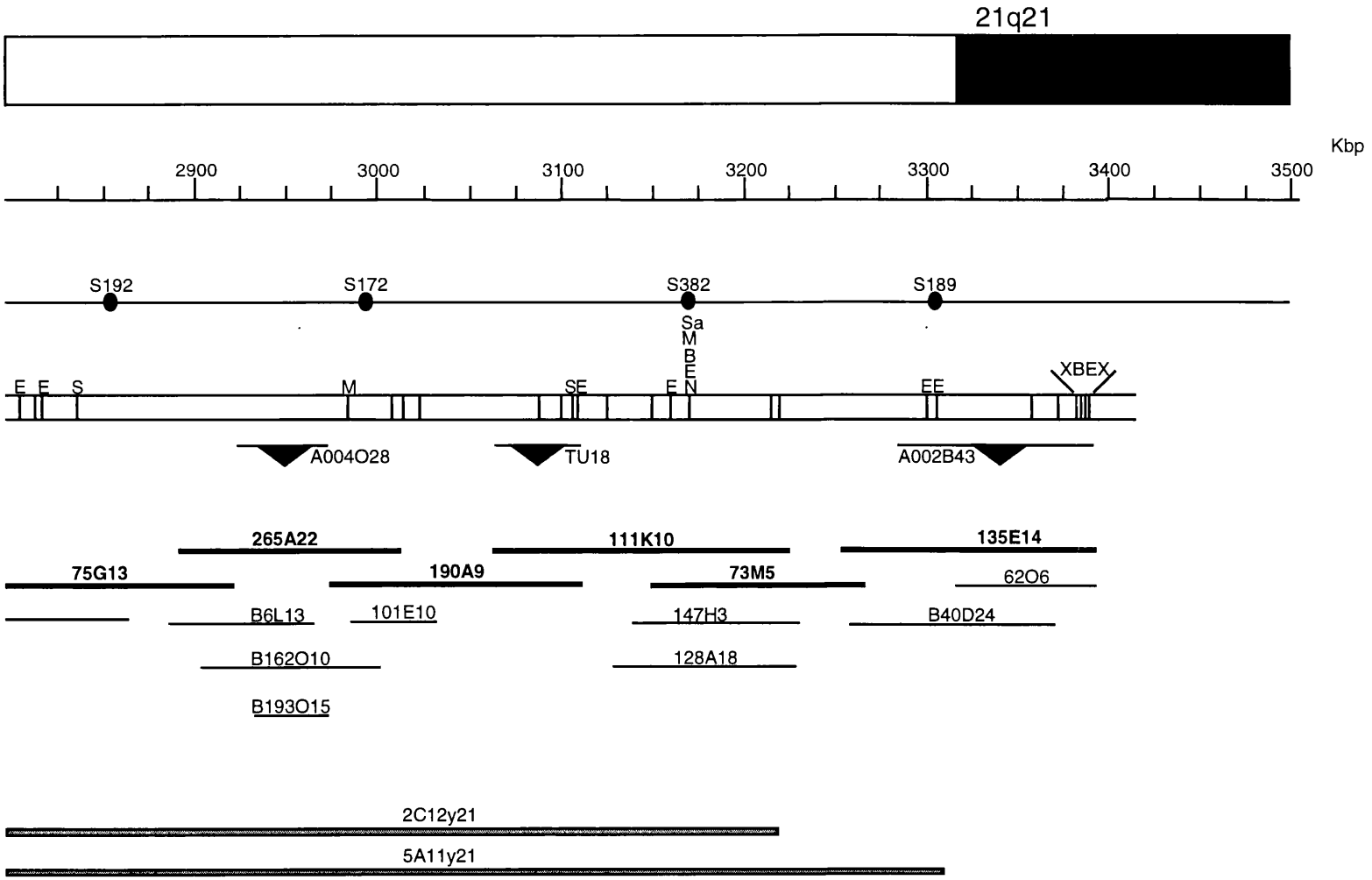

21q21
Kbp
2900
3000
3100
3200
3300
3400
3500
S192
S172
S382
S189
Sa
M
B
E
N
E E S M SE E N EE XBEX
A004O28
TU18
A002B43
265A22
111K10
135E14
75G13
190A9
73M5
62O6
B6L13
101E10
147H3
B40D24
B162O10
128A18
B193O15
2C12y21
5A11y21

Figure 3.13

D21S373 were used to assess if these markers were localised on the somatic cell hybrid lines indicated in Figure 3.13. The results were identical to those previously published (Graw *et al.*, 1995). The markers D21S258, D21S236 and D21S373 were not analysed by Graw and co-workers. Lastly the map contains information on the position of the only two characterised genes thus far in the area, STCH and RIP140, the exon trapped clone TU18, and the two unigene clusters A004O28 and A002B43.

3.4 Discussion

3.4.1 Comparison of the bacterial contig map with maps of uncloned genomic DNA and YAC based maps

The total genomic distance that is spanned by the contigs is 3.4 Mbp of which 3.2 Mbp is contained in the two contigs (94%); one gap of approximately 200 Kbp remains. Approximately 59% of the contigs consists of three- to seven-fold coverage; 13% is covered by a single clone. In the part of the map where the YAC contig has been reliable (80% of the total distance) PAC, BAC or cosmid clones have been identified covering this whole distance. A total of 97 clones are shown in Figure 3.13, eleven additional clones are not shown in this figure (compared with Figures 3.3, 3.4, 3.7, 3.8). A large number of clones were detected by hybridisations of the vectorette fragments 32M22-P1/V and 100M18-P1/V (35 and 28), of which ten are indicated in Figures 3.3 and 3.13.

A comparison of the order of markers from four different studies (Van Hul and Van Broeckhoven, 1993; Butler *et al.*, 1992; Gardiner *et al.*, 1995; Korenberg *et al.*, 1995) with data presented here shows the order to be almost identical. A long range restriction map of genomic DNA, in which markers were mapped to restriction fragments, showed the order to be D21S190-S215-S149-S253-S258-S16-S48-S13-S192-S382(Eag 101) (Van Hul and Van Broeckhoven, 1993), which is identical to that obtained here (Figure 3.13). An overlapping YAC map in the D21S16-S13 region showed the order of markers to be D21S378-S374-S16-S373-S13-S372 (Butler *et al.*, 1992), being identical to the order of markers in Figure 3.13. Two studies (Gardiner *et al.*, 1995; Korenberg *et al.*, 1995) verify the YAC contig established by Chumakov and co-workers (Chumakov *et al.*, 1992). The order of the markers in these studies is: D21S369/S215/S297/S308-S318-S378/S120-S374/S236-S16-S371-S373-S13-S372-S192-S172-S382 (Gardiner *et al.*, 1995) and D21S215/S369/S297-S318-S120-S374-S16-S48/S236-S371-S373-S13-S372-S192-S172-S382 (Korenberg *et al.*, 1995). These orders are also identical to those

described here, except for D21S236 which was found to be centromeric to D21S16 in the study by Gardiner and co-workers. Here D21S236 was found to be telomeric of D21S16, which is likely to be correct as this was the same order detected by Korenberg and co-workers. Two *Not* I sites have been reported in this area, these are D21S13 and D21S382 (LL52), which are separated by 0.72 Mbp of genomic DNA (Ichikawa *et al.*, 1993). In this study, *Not* I sites were found at these two markers and a comparable distance between the markers (see Figure 3.13). Van Hul and Van Broeckhoven (Van Hul and Van Broeckhoven, 1993) found the same restriction endonuclease sites at D21S13 and D21S382 (*Eag* 101) as described here, and they found a distance of 700 Kbp between the two markers. In addition, the study by Van Hul and Van Broeckhoven shows a distance of 1320 Kbp between D21S16 and D21S382 (1420 Kbp for the present study, see Figure 3.13) and 2800-3120 Kbp between D21S190 and D21S382 (2895 Kbp, Figure 3.13). These comparisons show that both the order of the markers and the genomic distances are in close agreement with previously published studies. In particular, concordance of distances across the whole region with those obtained by direct sizing of uncloned genomic DNA fragments (Van Hul and Van Broeckhoven, 1993) diminishes the possibility that many sizeable deletions are present in the bacterial clones that form the contig.

3.4.2 DNA sequence similarity of proximal 21q11 to sequences on different chromosomes

The high number of clones hybridising to the vectorette fragments 32M22-P1/V and 100M18-P1/V on the 21q11 filters, which are located close to the centromere, could indicate that these fragments show sequence similarities to different chromosomal locations. Although no clones were found on the 21q11 filters that hybridised to the α-satellite probe D21Z1 (not shown), this marker cannot be far from the proximal end of the contig as it is located at a maximum distance of 350 Kbp from D21S190 (Van Hul *et al.*, 1993). The clones that span the proximal duplicated region (133G21, 156L9, 127M18 and 16C2) were confirmed to originate from this copy by mapping of unique polymorphisms (Potier *et al.*, 1998) (see Figure 3.13). The restriction map, which was generated before the data by Potier and co-workers was available, was constructed using 77P15 instead of 133G21. When the markers from the clones B62L20 and 29H4 are compared with the localisation of these markers on different chromosomes (see Graw *et al.*, 1995) then chromosome 21 is the only chromosome which all the markers have in

Localisation of B62L20, 29H4 and 82L10 to chromosome 21

	B62L20	29H4	82L10	Chromosome location
D21S369	+			21, 18
D21S215	+			21, 10, 14, 15, 18
D21S297	+	+		21, 17
D21S308		+		21 and all except 5, 16, 17, X, Y
D21S318			+	21, 2, 9, 9/13, 18, 20
ES0067			+	21, 22

Table 3.6: deduction of the chromosomal location of three clones B62L20, 29H4 and 82L10 by comparison of their marker content with the chromosomal location of the D21 markers (from Graw *et al.*, 1995) and the EST ES0067 (from Chiang *et al.*, 1995). + indicates the marker or EST was present on the clone. The only chromosome in common between the markers and the EST on the respective clones is chromosome 21.

common, suggesting these two clones do contain genomic inserts from chromosome 21 (see Table 3.6). B62L20 showed hybridisation by fluorescent *in situ* hybridisation (FISH) to chromosomes 21 and 18p11.2, confirming this region to have sequence similarity between the two chromosomes (T.A. Jones). The PAC clone 82L10 is also likely to be located on chromosome 21 as this is the only chromosomal location in common between D21S318 (Graw *et al.*, 1995) and ES0067 (Chiang *et al.*, 1995). The region of 21q11 specific to chromosome 21 begins just centromeric to the clone 98L15, between the markers D21S318 and D21S258 (see Figure 3.13). The marker D21S258 was found not to cross-hybridise to GB3, a cell line containing human chromosome 13 on a hamster background (Van Hul *et al.*, 1993). Using a human mono-chromosomal hybrid panel (HGMP) it was found that an amplimer of the marker D21S258 would amplify a specific fragment only in the cell line containing human chromosome 21. Furthermore, the marker D21S378 was found to be the most proximal of eight chromosome 21 specific markers (Graw *et al.*, 1995). In this study, it was found that D21S378 and D21S258 were localised to the same PAC clone, 90B5. Furthermore, the clones 98L15 and 90B5 hybridised only to chromosome 21q11.2 by FISH (see Table 3.7, T.A. Jones, J. Crolla, T. Hildmann personal communications), The clone 63K18 just centromeric to 98L15 hybridised by FISH to the following chromosomes: 21, 2, 9, 13, 14, 15, 18, 19, 22, again indicating that the region specific for chromosome 21 starts between the markers D21S318 and D21S258. The previously mentioned clones, and other clones that were localised by FISH on metaphase spreads of human chromosomes, are indicated in Table 3.7.

3.4.3 Comparison of the integrated map of the 21q11 region with long range sequencing data

The final verification of the physical map will come from genomic sequencing of the clones. The clones are being sequenced by groups in Jena, Berlin and Braunsweich, and progress can be monitored via the following web sites: http://chr21.rz-berlin.mpg.de/, http://genome.imb-jena.de/, http://genome.gbf.de/seqproj.html, http://www.eri.uchsc.edu/chr21/c21index.html . To date (28/10/1999) seventeen clones from the MTP have been fully sequenced and are available from GDB/EMBL. These are: 133G21 (AJ239321), 156L9 (AJ239322), 127M18 (AJ239318), 16C2 (AL049911) 29H4 (AL078475), 98L15 (AL178615), 126N20 (AF130249), 31B5 (AF130247), B39I12 (AF165138), 190N10 (AF130351), 10J13 (AF130248), 75G13 (AJ006998), 265A22

FISH analysis of bacterial clones from the 21q11 contig

Clone	Chromosomal location	Source		Clone	Chromosomal location	Source
32M22	21, 18, 2	T.H.		B39I12	21	T.H.
94H11	21, 16, 10	T.H.		190N10	21	T.H.
77P15	21, 18, 2	T.H.		153I22	21	T.H.
156L9	21, 2	T.H.		30P13	21	T.H.
127M18	21 2x	T.H.		270M7	21	T.H.
16C2	21 2x	T.H.		B194B7	21 2x	T.H.
B62L20	21, 18	T.J.		103K19	21	T.H.
223O1	21, 18	T.J.		10J13	21	T.H.
214G14	21, 18	T.H.		c102 65F03	21 2x	T.H.
256M13	21, 18, 13, 2	T.H.		75G13	21	T.H.
82L10	21	T.H.		265A22	21	T.H.
63K18	21,2,9,13,14,15,18,19,22	T.H		190A9	21	T.H.
98L15	21	T.H./J.C		111K10	21	T.H.
90B5	21	T.H./T.J		73M5	21	T.H.
126N20	21	T.H.		135E14	21	T.H.
31B5	21	T.H.				

Table 3.7: Clones from the contig indicated in Figure 3.11 that were used in FISH analysis of metaphase spreads of human chromosomes, and the chromosomes the clones hybridise to. The clones can be found in Figure 3.11 and are shown in the table from centromere (top left) to telomere (bottom right). The information was derived from T. Hildmann (T.H.), T.A. Jones (T.J.) and J. Crolla (J.C.)

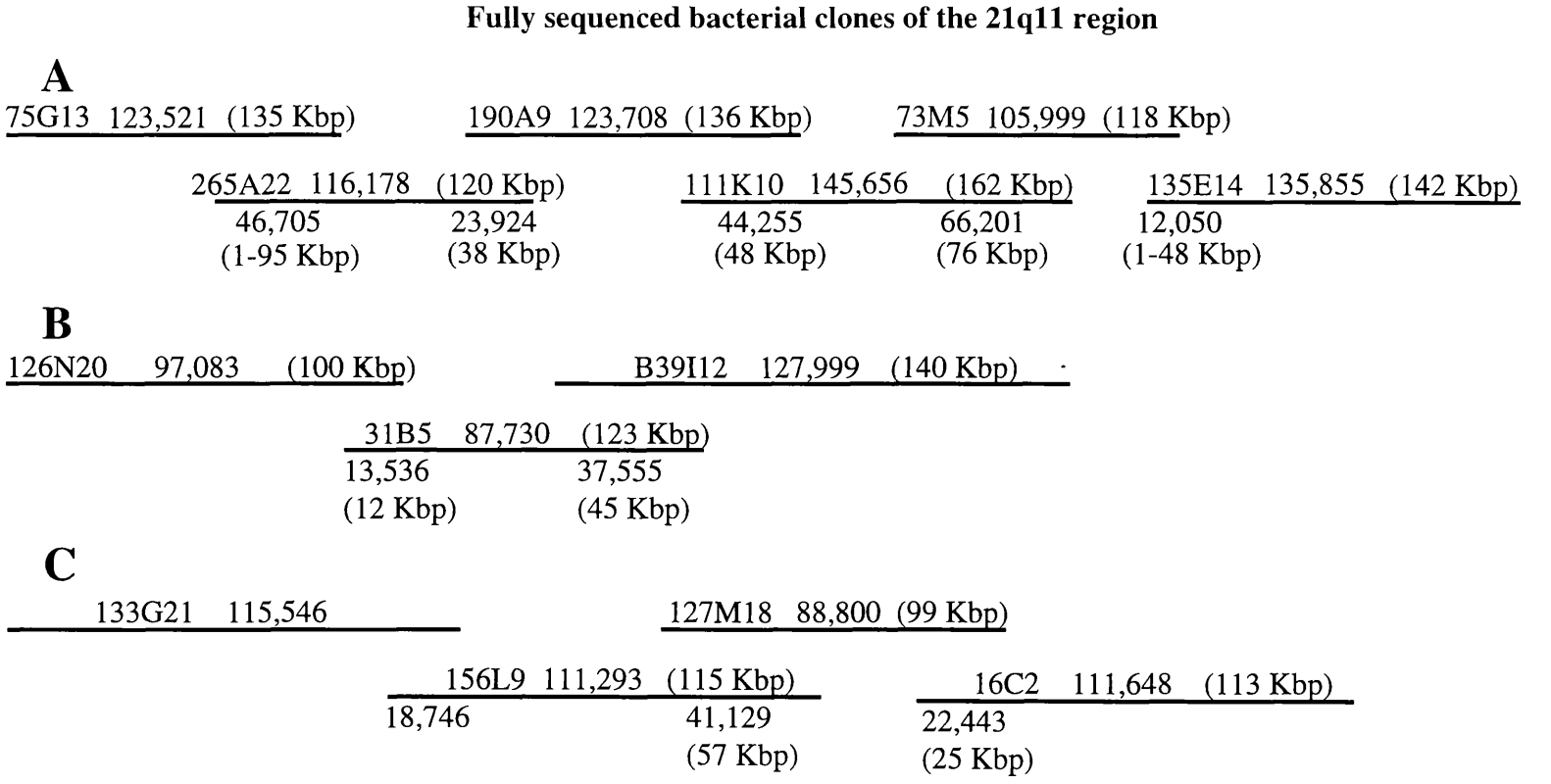


Figure 3.14: Fully sequenced genomic clones that show overlaps. Three contigs are shown. The horizontal lines represent the different clones, which are drawn to scale. Above the clones their names are indicated followed by the total length in bp (from GDB) and the estimated length from the restriction maps (in brackets in Kbp). Underneath each overlap the total size in bp is shown, with the estimated overlap from the restriction maps in brackets in Kbp (estimates derived from Figure 3.10). A -The total length of the clones is 750,917 bp, 193,085 is present in the overlaps. The total size of the contig is 557,832 bp (estimated in Figure 3.11 as 640 Kbp). B -The length of the clones is 312,812 bp, the overlap is 51,091, the total length of this contig is 261,721 bp (the estimation in Figure 3.11 is 300 Kbp). C -Length of the clones 427,287 bp, overlap 82,317, length of the contig is 344,970 (estimation from Figure 3.11 is 355). The total length of the three contigs is 1.16 Mbp.

(AJ009632), 190A9 (AJ006997), 111K10 (AL034449), 73M5 (AJ010597) and 135E14 (AJ010598). The clones 133G21/156L9/127M18/16C2, 126N20/31B5/B39I12 and 75G13/265A22/190A9/111K10/73M5/135E14 are expected to overlap from the hybridisation data (Chapter 3.3.2.1, 3.3.2.2, 3.3.2.4 and 3.3.4), which is shown in Figure 3.13. The length of the individual clones and the extent of the overlap, as determined by genomic sequencing of these clones, can be seen in Figure 3.14. The clones show the expected overlaps, compared with Figure 3.13, which result in total continuous genomic sequences of 344,970, 261,721 and 557,832 bp for the three contigs. Size estimations of the clones and overlaps made from PFGs are, on average, 10% larger than the genomic sequence (see Figure 3.14). Additional clones from the contig are being sequenced and are nearing completion. For the clones that are located in the areas with sequence similarities to different chromosomes, additional sequence tagged sites could be designed from the genomic sequence to verify the location of the clones.

In conclusion, this chapter shows a physical map of overlapping bacterial clones containing genomic inserts from the 21q11 region. This map spans a 3.4 Mbp region composed of two contigs of a total of 3.2 Mbp genomic DNA, and one gap of approximately 200 Kbp. A number of exon trapped, cDNA selected and Unigene cDNA clusters have been positioned on the map. With the aid of gene identification techniques such as exon trapping and cDNA selection, together with genomic sequencing of the bacterial clones, additional genes are likely to be found which can be analysed for expression, and function and assessed for their involvement in different phenotypes that are associated with the region.

Chapter 4 Exon Trapping using genomic clones in the 21q11.2 region

4.1 Summary

Exon trapping was performed using the minimal tiling path clones represented in the physical map from Chapter 3 (Figure 3.13). Twenty-two overlapping clones from the 21q11 region were used in forty exon trapping experiments. More than 7000 clones were included in the analysis and over 3000 were analysed. Ninety-three putative exons were described, and their location is indicated on the physical map from Chapter 3. Sixty-three putative exons were found to contain an open reading frame consisting of the entire sequence. Nineteen exons were identical to human cDNA clones and one cDNA clone was identical to six different exon trapped clones. Amino acid sequences generated from the exon trapped clones were compared with known and putative proteins; the best matches were found between putative exons and a rat lipase, a multidrug resistance protein and a ubiquitin carboxyl-terminal hydrolase.

4.2 Introduction

Chromosome 21 is the smallest of the human autosomes and is 40-50 Mbp long. Down syndrome (DS), which is caused by trisomy of chromosome 21, is the most common genetic birth defect in man, the incidence being around 1:700 live births (Epstein, 1995). More than twenty phenotypical features are associated with DS, although these are not all present in each individual with DS. A study that compared phenotypes of DS with the genotype of partial trisomies mapped most of the DS characteristics in the region 21q22.2 in the distal half of the long arm (Delabar *et al.*, 1993). A different study suggested that other regions of chromosome 21 might also be important in studying genes that are involved in DS phenotypes (Korenberg *et al.*, 1994).

4.2.1 Identification of novel genes in human chromosome 21q11.2

Three different disorders have been associated with the 21q11 region. These are acute non-lymphocytic leukaemia of the M7 immunotype (ANLL-M7)/ transient abnormal myelopoiesis (TAM), Familial early onset Alzheimer's disease (FEOAD) and non-small cell lung carcinoma (NSCLC) (see Chapter 3.2.2). We are therefore interested in

studying genes that are located in the 21q11 region, and their possible implication in disease phenotypes.

A number of different methods can be used to identify genes. Some examples of these are:

- cDNA selection (Lovett *et al.*, 1991; Parimoo *et al.*, 1991; Korn *et al.*, 1992). This technique is based on hybridisation of cDNA clones to human genomic DNA. The specific hybridisation of cDNA species to the clones which contain genomic DNA can lead to the identification of genes that are present within the genomic fragments. A more detailed description can be found in Chapter 1.2.3.
- Exon trapping (Duyk *et al.*, 1990; Buckler *et al.*, 1991). This method utilises the splice sites that are present at intron/exon boundaries of exons. Exons of genes can be identified from clones containing genomic DNA. Exon trapping is discussed in more detail in Chapter 1.2.3 and in this Chapter.
- Exon/gene prediction programmes (Uberacher and Mural, 1991; Burge and Karlin, 1997). Genomic sequence is subjected to different exon and gene prediction programmes, which can lead to the identification of novel genes in the area (see also Chapter 1.2.5).

There are a number of advantages and disadvantages to these methods (see sections 1.2.3., 1.2.5). A combination of the three methods will probably be necessary to identify the majority of genes in a genomic region. No genomic sequence for the 21q11 region was available for gene prediction programmes at the start of this study. In this study we chose the exon trapping method, because of the suggestion that the 21q11 region may contain genes that are only transcribed in certain tissues (Gardiner, 1996). The cDNA selection technique has been performed in parallel (R. Flomen).

4.2.2 Exon trapping: a method to identify novel genes

Exon trapping is a technique which can be used to identify potential exons from genomic DNA (Duyk *et al.*, 1990; Buckler *et al.*, 1991; Church *et al.*, 1994). It makes use of the fact that exons of many human genes are interrupted by introns. Introns are removed from the primary transcript by splicing, and the resulting mRNA consists only of exons. The splicing machinery recognises consensus sequences at the intron/exon boundaries to remove the introns. Exon trapping makes use of splice sites that are present in the genomic DNA, to capture exons.

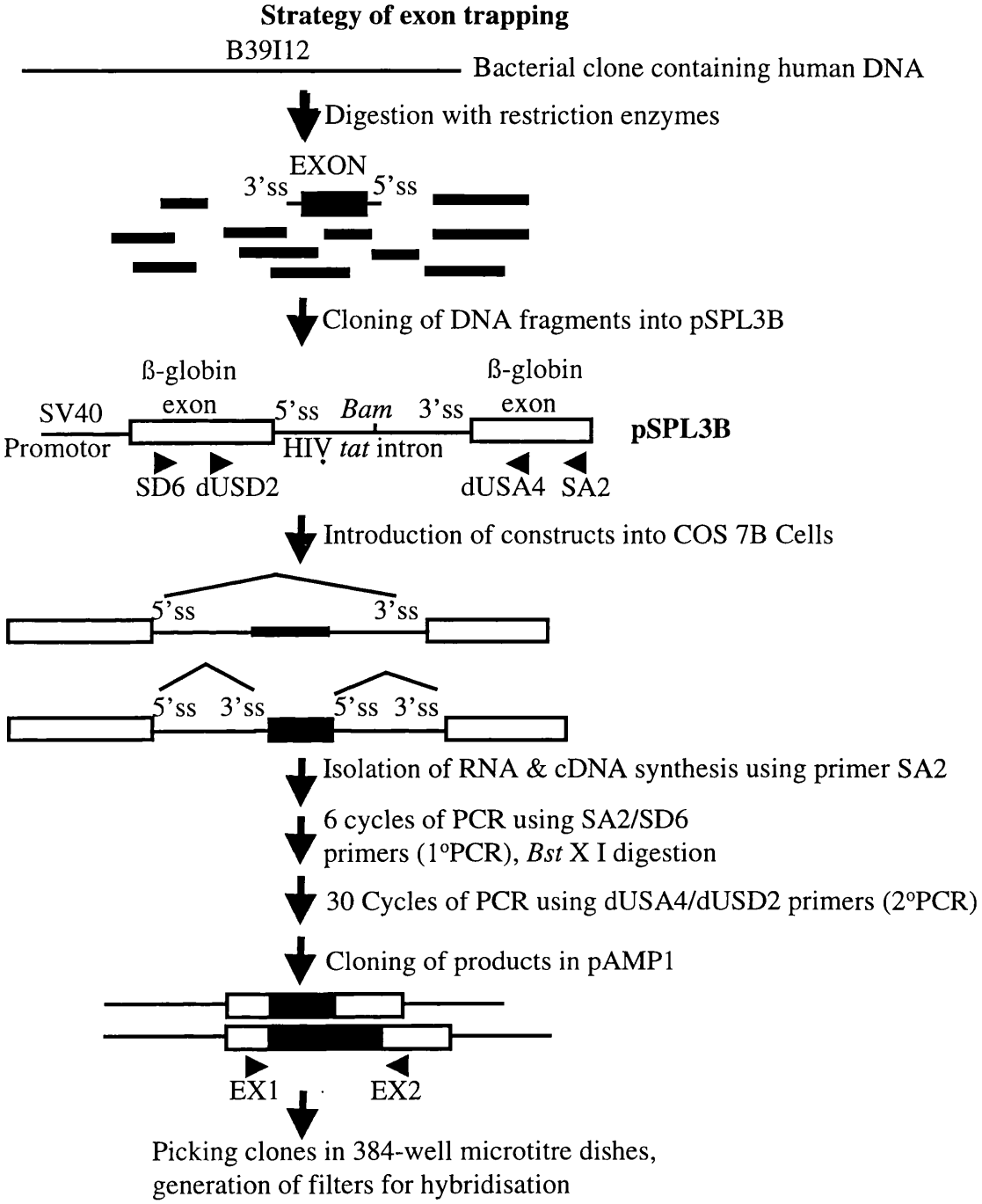


Figure 4.1: The exon trapping protocol is shown from bacterial clone to exon trapped clones. The bacterial clones were digested partially with *Sau* 3A I or completely with *Bam* H I/*Bgl* II and these fragments were cloned into the *Bam* H I site of pSPL3B. DNA was isolated from these constructs and they were introduced into COS 7B Cells. After a 48 hour incubation, RNA was isolated from the COS 7B cells, which was reverse transcribed and specifically amplified via PCR, using the primers SA2/SD6. A *Bst* X I digestion was carried out to eliminate species which still contained the intron (*Bst* X I sites are present in the intron) or which did not contain a putative exon (*Bst* X I half sites are present adjacent to the HIV *tat* intron, which form a functional site when the intron is spliced out without the presence of a putative exon). A nested PCR was carried out with the primers dUSA4/dUSD2, after which the fragments were cloned using the pAMP1 system. The EX1/EX2 primers are indicated, which were used for the generation of specific probes for further analysis

4.2.2.1 Exon trapping vectors using one splice site

A number of systems have been described that use one splice site in the vector, which is used for exon trapping. Duyk and co-workers designed a retroviral shuttle vector, which contains a splice donor site. Introduction of a splice acceptor site, present within genomic DNA would ultimately result in the removal of the β-galactosidase gene, and introduction of a putative exon. The DNA is eventually cloned into *E coli*. The clones that have lost the β-galactosidase gene are subsequently identified and analysed (Duyk *et al.*, 1990).

An extension of the use of the exon trapping vector by Duyk and co-workers was subsequently reported (Datson *et al.*, 1994). Genomic DNA is cloned into the vector, after which it is directly introduced into COS cells. The additional use of a primer that amplifies from the poly A tail means that both internal and 3'-terminal exons can be trapped.

A similar approach to the isolation of 3'-terminal exons had already been reported for a different exon trapping vector (Krizman and Berget, 1993). These authors reported a vector which contains two exons, the introduction of a 3'-terminal exon results in the production of stable mRNA species. Specific primers are used to analyse the putative exons.

4.2.2.2 Exon trapping vectors using two splice sites

Initially, the pSPL1 vector was designed that uses two splice sites for exon trapping (Buckler *et al.*, 1991). An improved vector, pSPL3, was constructed by the addition of *Bst* X I sites in order to significantly reduce the number of clones that were generated without the occurrence of splicing events (Church *et al.*, 1994). Additionally, cryptic splice sites were removed from pSPL3, resulting in the vector pSPL3B. This abolished the exon trapped products that arose via splicing using these cryptic splice sites (Genzyme, personal communication).

The exon trapping vector pSPL3B contains two splice sites, at either end of an HIV *tat* intron. This intron is flanked by two β-globin exons from rabbit, and the construct is under the control of a eukaryotic SV40 promoter (Figure 4.1). Genomic fragments can be cloned in the multi-cloning site which is present in the HIV *tat* intron, after which these constructs are introduced into eukaryotic cells. The primary transcript from the pSPL3B vector consists of the two β-globin exons, the HIV *tat* intron and a genomic

fragment. If an exon is cloned with its splice sites in the correct orientation, then it would remain part of the mRNA after splicing of the intronic sequences (see Figure 4.1). RNA is isolated from the eukaryotic cells and reverse transcriptase PCR is then applied using the primers SA2 and SD6, which are complementary to DNA sequence that is present in the two β-globin exons. This is followed by a *Bst* X I digestion. *Bst* X I sites are present within the multi-cloning site and this restriction digestion decreases the number of species that still contain part of the HIV *tat* intron. A further two *Bst* X I half sites are present adjacent to either end of the HIV *tat* intron. These form a functional site when splicing occurs between the two HIV *tat* intron splice sites. A reduction in clones without putatively trapped exons is the result of the *Bst* X I digestion. The digestion is followed by a second PCR (for example with the primers dUSD2 and dUSA4, see Figure 4.1). The PCR products are subsequently cloned into a bacterial vector and can be analysed by sequencing of the inserts.

4.2.3 Exon trapping using bacterial clones located in the 21q11.2 region

Exon trapping was performed in order to identify novel genes in the 21q11 region. Twenty-two genomic clones from the region were used. The exon trapping system using a vector (pSPL3B) with two splice sites, was chosen because of:

- The availability of the vector (Life Technology, Genzyme).
- The availability of a detailed and accurate set of protocols (Life Technology).
- The reliability of the system, which has been shown by the numerous papers from different groups reporting the successful isolation of putative exons (see for example Taylor *et al.*, 1992; The Huntington's disease collaborative research group, 1993; Yaspo *et al.*, 1995; Chen *et al.*, 1996).

In order to identify novel genes in the 21q11 region, exon trapping was performed using twenty-two genomic clones with overlapping genomic inserts from this region (see Figure 3.13). These genomic clones have been described in Chapter 3 and in Groet *et al.*, 1998. Ninety-three putative exons were analysed and nineteen were shown to be identical to partially sequenced cDNA clones (human expressed sequence tag [EST] database). This work is likely to lead to identification of novel genes which can be tested for their involvement in the above mentioned phenotypes.

4.3 Results

4.3.1 Exon trapping using genomic clones containing overlapping genomic inserts from chromosome 21q11.

Twenty-two overlapping genomic clones containing genomic DNA from chromosome 21q11 (see Figure 3.13) were selected for exon trapping experiments, these were: 98L15, 90B5, 126N20, 31B5, B39I12, c102 68C10, c102 132H4, 190N10, 153I22, 30P13, 270M7, B194B7, 103K19, 10J13, c102 65F3, c102 44E10, 75G13, 265A22, 190A9, 111K10, 73M5 and 135E14. DNA was isolated from these clones, as described in Chapter 2.18.2. The size and the restriction endonuclease recognition sites were confirmed by *Not* I/*Xho* I digestion and pulsed field gel electrophoresis (PFGE), and then compared with the restriction maps of the individual clones (see Figure 3.12).

The twenty-two clones were digested with the restriction endonucleases *Bam* H I and *Bgl* II or *Sau* 3A I. The *Bam* H I/*Bgl* II samples were then phenol/chloroform extracted and precipitated, and the *Sau* 3A I digested samples were gel purified. *Sau* 3A I digestion was carried out under conditions that favoured partial digestion, and fragments in the size range of 1-5 Kbp were isolated by agarose gel electrophoresis. Examples of this can be seen in Figure 4.2. Fragments from the digested clones were ligated into the *Bam* H I site of the eukaryotic expression vector pSPL3B, which contains an SV40 early promoter for transcription in mammalian cells (see Figure 4.1), and were then transformed into *E. coli* XL1-blue cells. Colonies were counted and the average number obtained for the *Sau* 3A I preparations was 251, with the range being 40 (when cloning cosmid clone fragments) to 603. The average number of colonies for the *Bam* H I/*Bgl* II ligations was 198, with the range being 73 (when cloning cosmid clone fragments) to 384. The control ligations showed an average of 12 colonies for digested pSPL3B vector (no ligase), 8 for self ligation of pSPL3B (ligase added) and 622 for the positive control (ligase and T4 polynucleotide kinase added). The different transformants at this stage of the exon trapping protocol represented the original genomic DNA-containing bacterial clones, cloned in small fragments in the expression vector.

The colonies from individual transformation experiments were scraped from the plates and grown overnight as one culture, after which DNA was isolated from these pools for transfection into mammalian COS 7B cells. DNA from *Sau* 3A I and *Bam* H I/*Bgl* II

Generation of fragments from genomic clones for cloning into pSPL3B

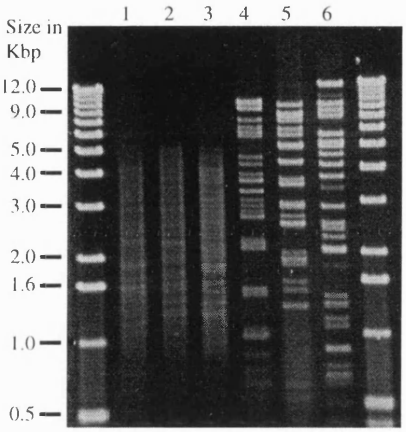


Figure 4.2: 0.7% w/v agarose gel of bacterial clones containing genomic DNA after digestion with *Sau* 3A I or *Bam* H I/*Bgl* II. Lanes 1&4-B39I12, lanes 2&5-190N10, lanes 3&6-103K19. Lanes 1-3 contain the *Sau* 3A I partially digested clones, after isolation of the ~ 1-5 Kbp fragments. Lanes 4-6 show the *Bam* H I /*Bgl* II digested samples after precipitation of DNA. The lanes on the outsides of the gel contain a 1 Kb ladder; sizes are indicated in Kbp.

Examples of PCR amplified fragments after exon trapping and negative controls

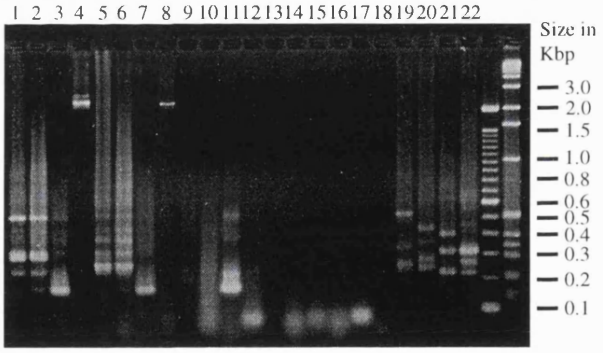


Figure 4.3: 2% w/v agarose gel showing the final stage after RNA isolation and RT-PCR, after PCR amplification with dUSA4 and dUSD2. Lane 1- B39I12BB PCR products (these were generated after digestion of the bacterial clone B39I12 with the restriction endonucleases *Bam* H I and *Bgl* II) before cloning in pAMP1. Lane 2- B39I12BB PCR products before removal of primers. Lane 3- B39I12BB final PCR products in a preparation in which *Bst* X I digestion had been omitted. Lane 4- B39I12BB PCR products of a sample in which RT and *Bst* X I were excluded. Lanes 5-8 similar to lanes 1-4 for B39I12S. Lanes 9-12 after pSPL3B transfection. Lanes 13-16 as lanes 1-4 after transfection with H_2O. Lane 17-H_2O control of the first PCR which was used in the second PCR. Lane 18- H_2O control for the second PCR. Lanes 19-22 examples of PCR products before UDG cloning, 19-98L15S, 20-90B5S, 21-c102 68C10/132H4BB, 22-c102 68C10/132H4S. The two lanes on the right contain the DNA markers, the lane on the left is the 100 bp ladder, the other the 1 Kb ladder, sizes are indicated in Kbp.

preparations of eighteen different PACs and BACs was transfected individually. DNA from *Sau* 3A I and *Bam* H I/*Bgl* II preparations of the four cosmids was pooled in pairs; c102 132H4 with c102 68C10 and c102 65F3 with c102 44E10. The vector pSPL3B and H_2O were transfected as negative controls. A positive control was not available, but as can be seen from Table 4.2 and Appendix I Table 2 (Table 90B5S, nrs. 6,7), exon trapped clone 90B5S-E6 was shown to be part of the previously published gene STCH, which was localised to clone 90B5 (see Figure 3.13).

The COS 7B cells that were transfected with the expression vector were grown for 48 hours after which total RNA was isolated. This was reverse transcribed with primer SA2, which anneals specifically to pSPL3B, after which six cycles of PCR were carried out with the SA2 and SD6 primers. The PCR samples were restriction digested with *Bst* X I, and a nested PCR was carried out with the primers dUSA4 and dUSD2 (Figure 4.1). A *Bst* X I digestion was carried out to eliminate species which result from splicing of the complete HIV *tat* intron, including genomic DNA (*Bst* X I half sites are present adjacent to the HIV *tat* intron, which form a functional site when the intron is spliced out without the presence of a potential exon). Final preparation of the RT-PCR products and negative controls of transfection of DNA from clone B39I12 can be seen in Figure 4.3: lanes 1-4 contain PCR products after transfection with DNA from B39I12 *Bam* H I/*Bgl* II preparation (B39I12BB). Lane 1 shows the PCR products before cloning into pAMP1, lane 2 before removal of primer dimers (Chapter 2.20.1 step 11). Lane 3 shows predominantly the 174 bp fragment which was the result of a splicing event in which all the intervening sequences were removed; this resulted in the β-globin exons being spliced together. Omission of *Bst* X I in this preparation is the cause of the abundance of this product in lane 3. Lane 4 shows the result when RT and *Bst* X I were excluded; it shows the amplification of an approximately 2.3 Kbp fragment, which is probably the result from the co-isolation of DNA with the RNA. The amplified product in this lane probably contains the complete HIV *tat* intron, which separates the β-globin exons in pSPL3B. This is not likely to be a problem as this product is not visible in lane 1, where a *Bst* X I digestion was carried out. Furthermore, the analysis of filters containing exon trapped clones includes hybridisation with the HIV *tat* intron probe, to identify these events which can then be omitted from further analysis. Lanes 5-8, 9-12, 13-16 show similar preparations for transfection of DNA from the BAC clone B39I12, which was digested with *Sau* 3A I (B39I12S), pSPL3B and H_2O, respectively. The final preparation of pSPL3B before cloning into pAMP1 showed a limited amount of the 174

bp product which is probably due to an incomplete digestion with *Bst* X I. The RT-PCR preparation and the other H_2O controls do not show any products (lanes 13-16), indicating that amplification of PCR products with the primers SA2, SD6, dUSA4 and dUSD2 is specific to pSPL3B. The PCR products were cloned into pAMP1 by UDG cloning. Uracil DNA glycosylase removes uracil residues that are present in the primers dUSA4 and dUSD2; the sticky ends that were thus generated anneal to pAMP1 (Life Technologies), after which the exon trapped clones were transformed into XL1-blue cells (Figure 4.1).

4.3.2 Analysis of the exon trapped clones

4.3.2.1 Picking of putative exon trapped clones into 384-well microtitre dishes

Individual exon trapped clones were transferred into wells of a 384-well microtitre dish and cultured. The different microtitre dishes and the exon trapped clones they contained can be seen in Table 4.1. For most genomic clones one 384-well microtitre dish was prepared (except for 90B5 and 98L15, see Table 4.1). Equal amounts of clones from the *Bam* H I/*Bgl* II and *Sau* 3A I preparations were picked, except for B39I12 (due to insufficient *Bam* H I/*Bgl* II clones, *Sau* 3A I clones were picked in the remaining wells). In addition, one row of the microtitre dish was filled with clones from the pSPL3B transfection, and one row with clones from the H_2O transfection (except 126N20, 31B5 and 135E14). In every plate a number of wells containing broth were left empty. This served as a control for contamination and was an extra labelling code for easier identification of membranes derived from these plates, after hybridisation.

4.3.2.2 Analysis of random exon trapped clones by sequencing

Initially 17 random clones were sequenced, as part of the analysis of the exon trapping experiments. These clones were derived from the microtitre dish containing the exon trapped clones from 90B5BB and 98L15BB. This generated the following results (J.H. Ives): five clones showed no insert or an insert of just a few bp and the 2 β-globin exons were spliced together, showing that the *Bst* X I digestion did not completely eliminate this event. Three identical clones showed a 100% match to an *E. coli* open reading frame (ORF). Three clones contained an insert but were identical. Three clones were unique and showed inserts of different sizes. A further three clones were unique and identical to clones that were independently analysed, as described in the next section.

Distribution of exon trapped clones in 384-well microtitre dishes

	Rows of the 384-well microtitre dishes in which exon trapped clones from the following experiments were grown.				
Clone	*Bam* H I /*Bgl* II digested	*Sau* 3A I digested	pSPL3B control	H_2O control	Empty wells
90B5BB 98L15BB	A-G H-N		O	P1-16	P17-24
98L15S		A-N	O	P1-16	P17-24
90B5S		A-N	O	P1-16	P17-24
126N20	ACEGIKMO	BDFHJLNP			P16,18,20,22,24
31B5	ACEGIKMO	BDFHJLNP			P16,18,20,22,24
B39I12	A-E13	E14-N	O	P	A19-24
68C10/132H4	A-G	H-N	O	P	B19-24
190N10	A-G	H-N	O	P	I19-24
153I22	A-G	H-N	O	P	D19-24
30P13	A-G	H-N	O	P	H19-24
270M7	A-G	H-N	O	P	P19-24
B194B7	A-G	H-N	O	P	A21-24, B21-24
103K19	A-G	H-N	O	P	E19-24
10J13	A-G	H-N	O	P	C19-24
65F3/44E10	A-G	H-N	O	P	M19-24
75G13	A-G	H-N	O	P	K19-24
265A22	A-G	H-N	O	P	L19-24
190A9	A-G	H-N	O	P	N19-24
111K10	A-G	H-N	O	P	J19-24
73M5	A-G	H-N	O	P	O19-24
135E14	ACEGIKMO	BDFHJLNP			P16,18,20,22,24

Table 4.1: Indicated in this table are the different 384-well microtitre dishes and the exon trapped clones they contain. Column 1- bacterial clones containing DNA from chromosome 21q11, which was the starting material in the exon trapping experiments. Individual rows symbolise separate 384-well microtitre dishes containing exon trapped clones derived from the clones indicated in column 1. Column 2- letters indicate rows in the microtitre dish that were filled with exon trapped clones after *Bam* H I/*Bgl* II digested samples. Column 3, as 2, but with *Sau* 3A I digested samples. Column 4, as 2, with clones after pSPL3B transfection. Column 5, as 2, containing clones after transfection with H_2O. Column 6 contains wells of the microtitre dish that were filled with broth but not inoculated.

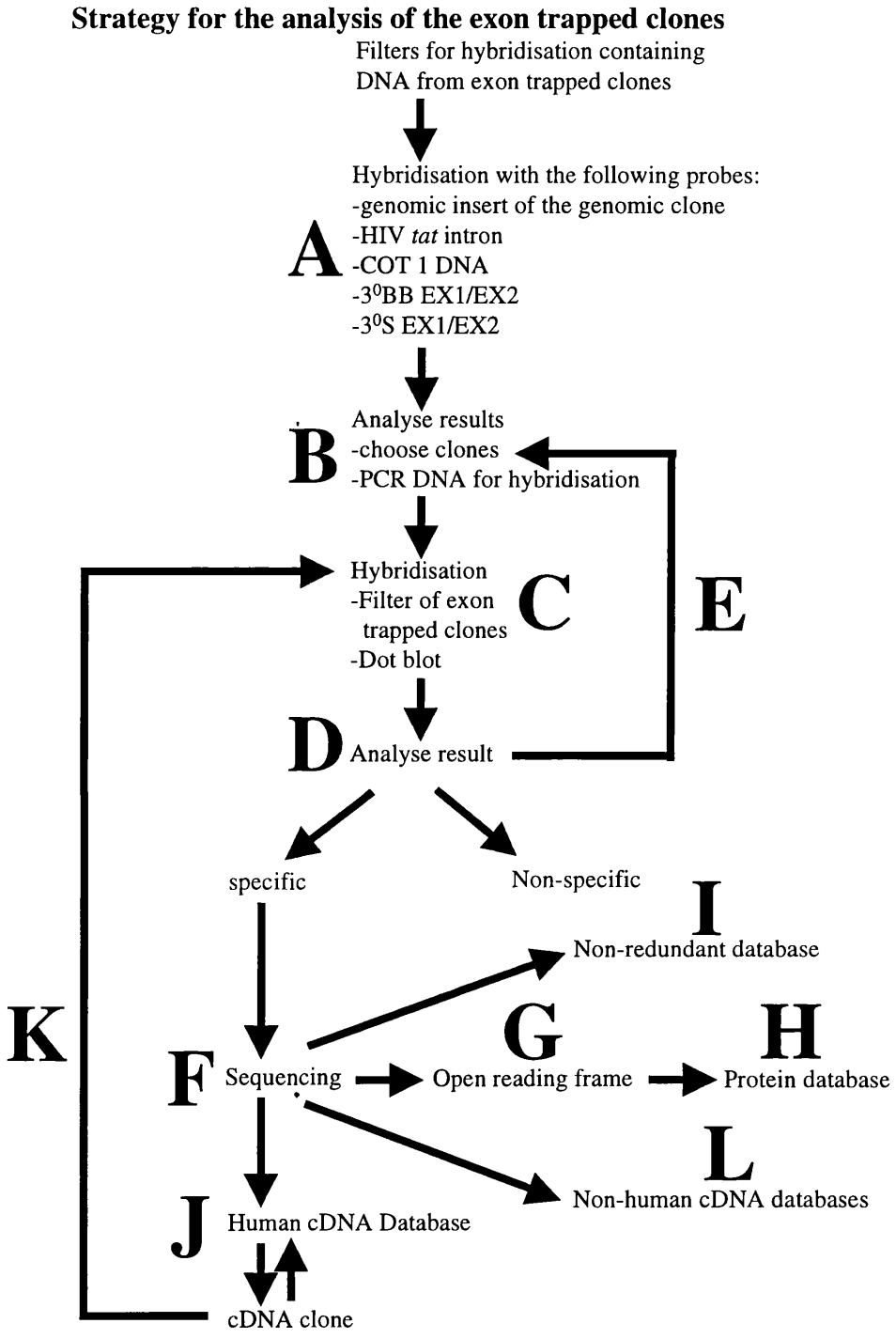


Figure 4.4: Strategy for analysis of exon trapping experiments, from spotted clones on nylon filters to putative exons. Section A: Initial hybridisations. Section B analysis of these results, selection of clones and PCR amplification from these clones. Section C shows hybridisation of this DNA to the filters containing DNA from the exon trapped clone, with analysis of these results in section D. Section E: additional clones for analysis. Section F: sequencing. Database searches are indicated in sections G-J and L. K: hybridisation of cDNA clones to the filters containing the exon trapped clones.

These clones were: 90B5BB-C21, 98L15BB-L11 and 98L15BB-N7. These experiments show that random sequencing will identify large numbers of identical, as well as empty, clones.

4.3.2.3 Identification of exon trapped clones for further analysis

In order to reduce the workload involved in characterisation of empty and identical exon trapped clones, a strategy was followed which is shown in Figure 4.4. Filters representing the exon trapped clones in the 384-well plates were produced for hybridisation experiments (Chapter 2.18.1). Hybridisations were carried out to identify clones for further analysis and were performed for all exon trapped clones (Figure 4.4A), except for those using radio-actively labelled genomic insert from the cosmids. Genomic inserts from the cosmid clones, can not be easily isolated by endonuclease digestion, since no appropriate endonuclease sites are present near the vector/insert boundaries. A representative example of the hybridisations can be seen in Figure 4.5.1-5 using the membranes containing the exon trapped clones from the BAC clone B39I12 as a starting point (see strategy in Figure 4.4A). The radio-active probes that were used in these hybridisations were:

- The genomic fragment of the bacterial clone, isolated as described in Chapter 2.18.9, hybridised to the corresponding membrane containing the exon trapped clones. Figure 4.5.1 shows an example of the hybridisation of the insert of clone B39I12 to the exon trapped products generated from the same clone. As can be seen from Figure 4.5.1, this type of hybridisation was difficult to interpret, because of a poor signal to noise ratio (see also Chapter 4.4). The clones that showed the strongest hybridisation signal are depicted in Figure 4.5.6 as number 1.
- HIV *tat* intron, isolated as described in Chapter 2.20.1 step 16. The hybridisation result of this probe can be seen in Figure 4.5.2. Positively hybridising clones were identified and are shown in Figure 4.5.6 as number 2. These result from non-splicing events of the HIV *tat* intron, or from co-isolation of pSPL3B DNA with RNA, followed by incomplete *Bst* X I digestion.
- COT 1 DNA, which hybridises to clones containing repetitive sequences. The positively hybridising clones H20 and K5, from the filter containing DNA from the exon trapped products from B39I12, can be seen in Figure 4.5.3. They are shown in Figure 4.5.6 as number 3.

- 3^0 BB EX1/EX2, which is a composite probe that was generated by a third PCR amplification with primers EX1/EX2 (see Figure 4.1 and Chapter 2.20.1 step 16) after the RT-PCR reaction (after PCR with dUSA4 and dUSD2) in the exon trapping experiment of B39I12BB. This composite probe is most likely to hybridise to the most abundant or larger species of exon trapped clones present on the membranes. It is therefore likely to help in distinguishing clones that contain putative exons from clones that do not. The hybridisation result can be seen in Figure 4.5.4 and the strongest hybridising clones are indicated in Figure 4.5.6 as number 4.
- 3^0 S EX1/EX2, as above for the exon trapping experiment of clone B39I12 with *Sau* 3A I. This hybridisation can be seen in Figure 4.5.4 and the strongest hybridising clones are shown as number 5 in Figure 4.5.6.

The analysis of these five hybridisations can be seen in Figure 4.5.6. Clones were selected that showed strongest hybridisation to either the genomic insert from the bacterial clones, 3^0 BB EX1/EX2 or 3^0 S EX1/EX2, or a combination of these and that did not hybridise to the HIV *tat* intron and Cot I DNA (see Figure 4.4B).

4.3.2.4 Analysis of exon trapped clones via sampling without replacement

The clones that were selected for further analysis from exon trapping experiment B39I12 were A1, D2, E18 and F7. DNA from these clones was amplified using the EX1/EX2 primers to generate a PCR product specific for the clone. This DNA was used to generate a radio-active probe which was hybridised to the filter containing the exon trapped clones, and to dot blots containing DNA of the parental genomic clone (B39I12) and its neighbours (strategy in Figure 4.4C). Hybridisation of the exon trapped clone to the dot blots was performed to show hybridisation to the parental genomic clone and to position putative exons onto the physical map. Hybridisation to the filters containing the exon trapped clones was performed to establish the clones that contained identical or overlapping inserts, which could then be omitted from further analysis. This is called "sampling without replacement". An example of the hybridisation of clone B39I12-D2 can be seen in Figure 4.6.1. The positively hybridising clones were analysed and their locations in the 384-well microtitre dish were added to Figure 4.5.6, as number 7, which resulted in Figure 4.6.2 (strategy in Figure 4.4D). Figure 4.6.2 contains the additional information from hybridisation of B39I12-E18 (nr. 6), B39I12-F7 (nr. 8) and B39I12-A1 (nr. 9). Additional clones could then be chosen for further rounds of hybridisation and analysis (strategy in Figure 4.4E), for example B39I12-I1 or B39I12-

Hybridisations for the identification of exon trapped clones for further analysis

Figure 4.5.1

Insert probe, number 1 in Fig. 4.5.6

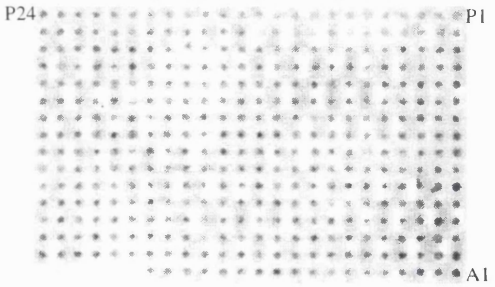


Figure 4.5.2

HIV intron probe, number 2 in Fig. 4.5.6

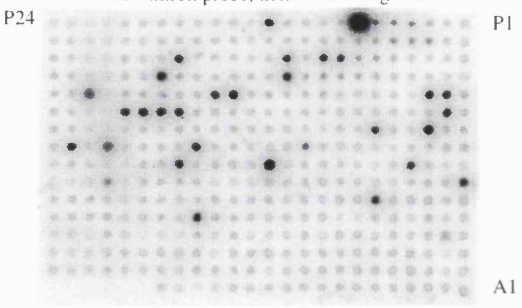


Figure 4.5.3

Cot 1 probe, number 3 in Fig. 4.5.6

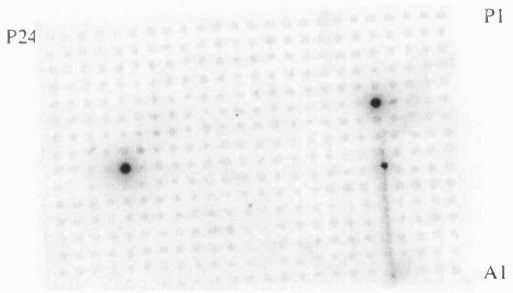


Figure 4.5.4

3⁰BB EX1/EX2, number 4 in Fig. 4.5.6

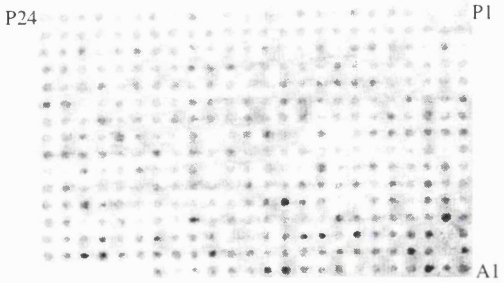


Figure 4.5.5

3⁰S EX1/EX2, number 5 in Fig. 4.5.6

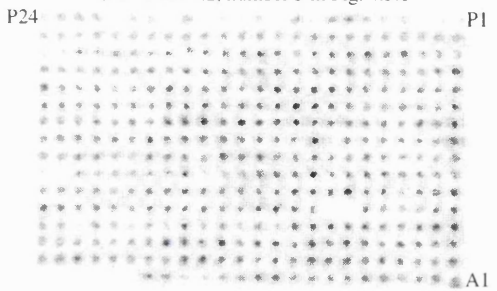


Figure 4.5.6

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
A	1		1,4								1,4	4												
B	1,4		4	1,4											4						4	4		
C			1,4				4		1,4	4				5							4			
D	5	1,4,5						4		1						4								
E		4									1,4					2		1						
F	1					2	1,5																	
G	2								5															
H				2								2					2	4		3				
I	1,4								5	2		1	1	1		2					2		2	
J	1,5	1	1,2			2				5														
K	1,4,5	2	1		3				5	5							2	2	2	2				
L	1	1,2	1,2											2	2							2		
M	1,5	1									2							2						
N	1	1		1				2	2		2	5					2							
O																								
P							2					2												

Figure 4.5: Hybridisation results of filters containing exon trapped clones from clone B39I12. Hybridisation of genomic insert from clone B39I12 (4.5.1), HIV *tat* intron (4.5.2), COT 1 DNA (4.5.3), 3^0BB EX1/EX2 (4.5.4), 3^0S EX1/EX2 (4.5.5), and the analysis of these results (4.5.6). The clones present in the wells indicated in grey were chosen for further analysis.

Figure 4.6.1

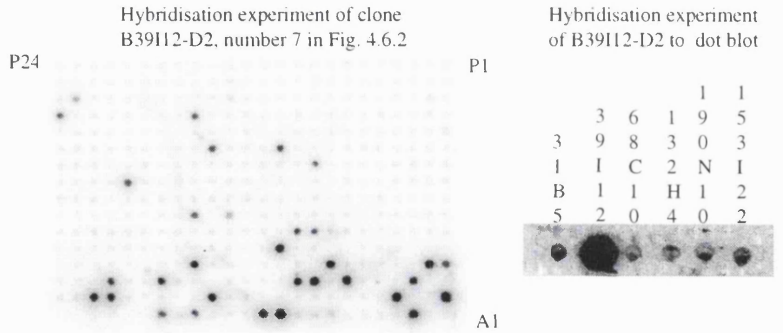


Figure 4.6.2

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
A	1,9		1,4,7				8	8	9	9	1,4,7	4,7		9	9	7	8	7						
B	1,4,7		4,9	1,4,7		9	9			8			9		4,7	9	9				4,7	4,7	9	8
C			1,4,7		9	9	4,7		1,4,7	4,7				5,8	9	9	8	7	9		4,7	8	9	
D	5,7	1,4,5,7	8	9			9	4,7		1	8	8		9	8	4,7		8			8		9	
E	8	4	9	9	9				9		1,4,7		9			2	8	1,6		8		9		8
F	1,6					2	1,5,8	8	7	7	8		9		8	8	8							8
G	2	8	8		8		8	8	5,8					7		7	8		9			8		
H			9	2						8		2				9	2	4		3			9	9
I	1,4	9							5,8	2		9				2	8	8		7	2		2	
J	1,5,8	1	1,2			2			7	5,6		1	1,8	1	8			8						
K	1,4,5,8	2	1		3			8	5,8	5,8	7		8		7	8	2	2	2	2			8	
L	1	1,2	1,2											2	2		8	8	8			2		
M	1,5,8	1									2					7		2	6					7
N	1	1	8	1,6		8		2	2		2	5,8		8			2,8	8	9		8	8	7	
O																								
P							2					2												

Figure 4.6: 4.6.1 Hybridisation experiment of clone B39I12-D2 to a 384-well filter of exon trapped clones from B39I12, and to dot blots containing DNA from B39I12 and neighbouring clones (indicated above the dots). 4.6.2 384-well grid from Fig 4.5.6 with additional hybridisation data from B39I12-A1 (number 9), B39I12-D2 (number 7), B39I12-E18 (number 6), and B39I12-F7 (number 8).

M2 (Figure 4.6.2). Appendix I Table 2 shows all the data for hybridisation to the filters containing the exon trapped clones. It contains the initial hybridisations of genomic insert from the bacterial clones, 3^0 BB EX1/EX2, 3^0 S EX1/EX2 (Chapter 2.20.1 step 16), HIV *tat* intron and Cot I DNA. Appendix I Table 2 also shows the subsequent hybridisations of all the exon trapped clones that were analysed, together with information on subsequent cDNA clones, which were identified via BlastN analysis of sequenced exon trapped clones. Individual grids that represent the different membranes containing DNA of the exon trapped clones in Appendix I Table 2 is identical to Figure 4.6.2. Figure 4.7 shows the position on the physical map of the exon trapped clones that mapped back specifically to the genomic clones. The analysis of the dot blots which were responsible for these data is identical to the one described for Figure 4.6.1.

4.3.2.5 Sequencing and database searches of the exon trapped clones

All the clones that mapped back specifically to the dot blots were sequenced (strategy Figure 4.4F). Five clones were analysed that did not map back specifically to the genomic clones they were derived from. These were: 98L15S-D12, 90B5S-A1, 90B5S-G5, 73M5-H10 and 73M5-K16. Two of these were also sequenced: 90B5S-G5 was very similar to an ORF from *E. coli*, and 73M5-H10 was derived from the PAC vector (see Table 4.2). PCR products for sequencing were amplified using M13 primers (M13F+ and R+, see Chapter 2.20.1 step 17). These products were cycle-sequenced with the same M13 primers and analysis was carried out on an ABI 310 automated sequencer. Ninety-three clones were sequenced that mapped back to the genomic clones. These clones are listed in Table 4.2. The sequences of the ninety-three exon trapped clones can be seen in Table 3 of Appendix I. The sequences were analysed against a number of databases:

- It was determined if the putative exon sequence contained ORFs (strategy Figure 4.4G). This was performed at http://www.ncbi.nih.gov/gorf. The sequence ATGGATGGATGG (5'-3') was added before the 5' end of the putative exon, and CCATCCATCCAT (5'-3') was added behind the 3' end. This was performed to provide start codons in every frame, as the programme that finds ORFs uses ATG codons to start open reading frames; ORFs can be missed if a start codon is not present. The results of this analysis for the sequenced exon trapped products can be seen in Table 4.2, which shows the largest ORFs.

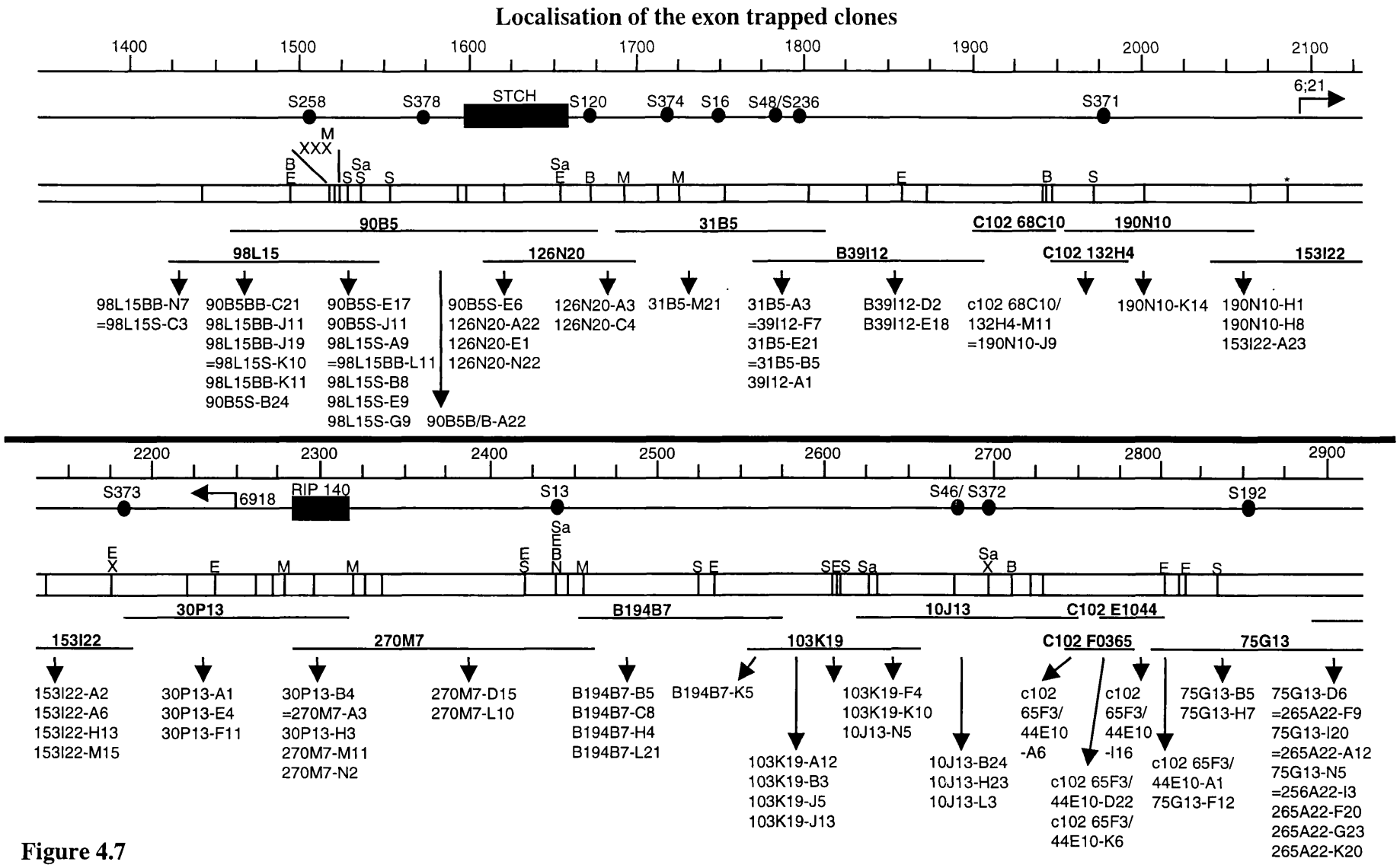

Localisation of the exon trapped clones
1400
1500
1600
1700
1800
1900
2000
2100
S258
S378
STCH
S120
S374
S16
S48/S236
S371
6;21
M
XXX
B
E
Sa
S
S
S
Sa
E
B
M
M
E
B
S
*
90B5
31B5
C102 68C10
190N10
98L15
126N20
B39I12
C102 132H4
153I22
98L15BB-N7
=98L15S-C3
90B5BB-C21
98L15BB-J11
98L15BB-J19
=98L15S-K10
98L15BB-K11
90B5S-B24
90B5S-E17
90B5S-J11
98L15S-A9
=98L15BB-L11
98L15S-B8
98L15S-E9
98L15S-G9
90B5S-E6
126N20-A22
126N20-E1
126N20-N22
90B5B/B-A22
126N20-A3
126N20-C4
31B5-M21
31B5-A3
=39I12-F7
31B5-E21
=31B5-B5
39I12-A1
B39I12-D2
B39I12-E18
c102 68C10/
132H4-M11
=190N10-J9
190N10-K14
190N10-H1
190N10-H8
153I22-A23
2200
2300
2400
2500
2600
2700
2800
2900
S373
6918
RIP 140
S13
S46/ S372
S192
Sa
E
X
E
M
M
S
Sa
B
N
M
S
E
S
E
S
Sa
Sa
X
B
E
E
S
30P13
B194B7
10J13
C102 E1044
153I22
270M7
103K19
C102 F0365
75G13
153I22-A2
153I22-A6
153I22-H13
153I22-M15
30P13-A1
30P13-E4
30P13-F11
30P13-B4
=270M7-A3
30P13-H3
270M7-M11
270M7-N2
270M7-D15
270M7-L10
B194B7-B5
B194B7-C8
B194B7-H4
B194B7-L21
B194B7-K5
103K19-A12
103K19-B3
103K19-J5
103K19-J13
103K19-F4
103K19-K10
10J13-N5
10J13-B24
10J13-H23
10J13-L3
c102
65F3/
44E10
-A6
c102 65F3/
44E10-D22
c102 65F3/
44E10-K6
c102
65F3/
44E10
-I16
c102 65F3/
44E10-A1
75G13-F12
75G13-B5
75G13-H7
75G13-D6
=265A22-F9
75G13-I20
=265A22-A12
75G13-N5
=256A22-I3
265A22-F20
265A22-G23
265A22-K20

Figure 4.7

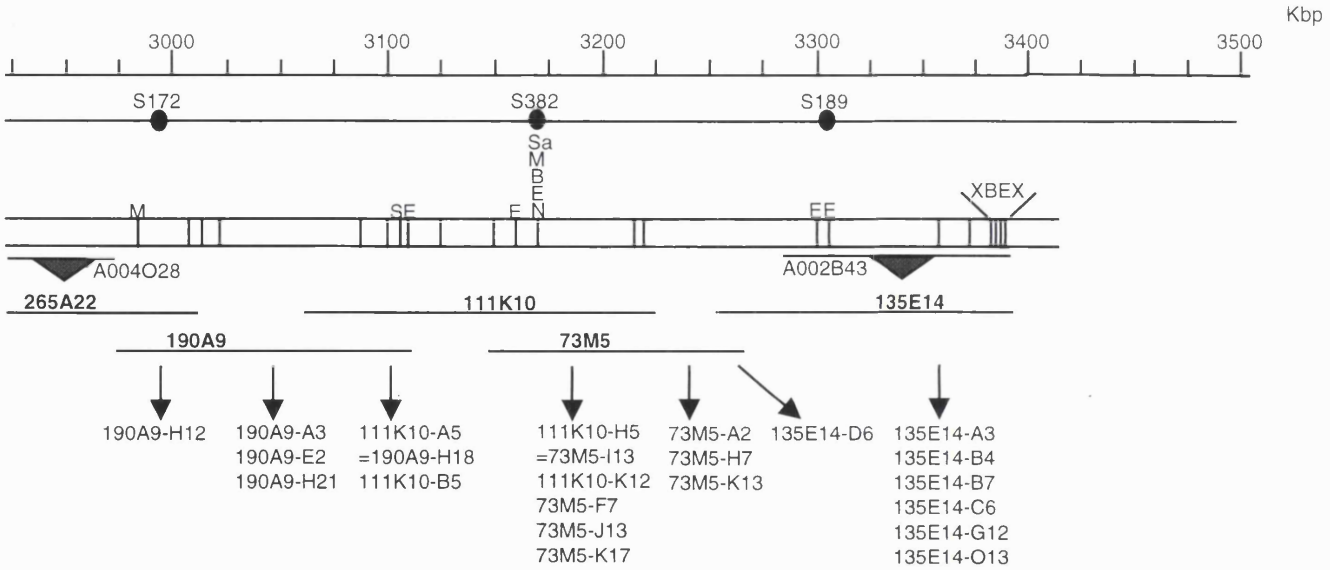


Figure 4.7: shows in three horizontal sections the mapping position of the ninety-three different exon trapped clones that were sequenced. This figure was adapted from Figure 3.11. It shows from top to bottom:
-scale in Kbp.
-representation of the chromosome with STSs (●), fully characterised genes (▬), segments of the chromosome contained within somatic cell hybrids on a rodent background (⌐→ ←⌐).
-Restriction map, vertical lines indication positions of restriction endonucleases. B=*BssH* II, E=*Eag* I, M=*Mlu* I, Sa=*Sac* II, N=*Not* I, S=*Sal* I, X=*Xho* I, if no enzyme is indicated = an *Xho* I site.
-cDNA clusters A004O28, AA002B43 (Schuler *et al.*, 1996).
-Minimal tiling path of genomic DNA in bacterial clones used in exon trapping experiments.
-Exon trapped clones and their position on the physical map. This was determined from hybridisation of these clones to dot blots of the relevant genomic clones. [=: two clones contain identical sequences].

- The ORFs were then compared to the non-redundant SwissProt database (strategy Figure 4.4H). This facility is directly available from the ORF-finding webpage. This compares the translated amino acid sequence of the exon trapped clones to protein sequences in the database. The best matches are depicted in Table 4.2. These include similarities to a lipase from rat (exon trapped clone 98L15BB-N7), a multidrug resistance protein (90B5BB-A22, 98L15S-G9, 126N20-A22 and 126N20-N22) and a ubiquitin carboxyl-terminal hydrolase from *Caenorhabditis elegans* (135E14-C6).
- The sequence was compared with sequences in the non-redundant Genbank+EMBL+DDBJ+PDB databases (strategy Figure 4.4I). This was carried out at: http://www.ncbi.nih.gov/Blast/. These searches show, for example, similarities to alu and L1 repetitive elements, *E. coli* DNA or human genomic sequence. Matches to these sequences can be found in Table 4.2. So far, all exon trapped clones that mapped back to the parental genomic clones have been shown to be identical to the sequenced genomic DNA. The exact mapping position on the fully sequenced genomic clones (Chapter 3.4.3) has not been shown in Table 4.2, but this can be determined by performing a BlastN search with the sequences of the exon trapped products (Table 3 of Appendix I) to the Genbank non-redundant database.
- Searches were performed against the non-redundant database of human expressed sequence tag (EST) entries in Genbank (Figure 4.4J), this was performed at the website http://www.ncbi.nih.gov/Blast/. This shows resemblance to ESTs present in the database, which can be a first stepping stone in establishing full length genes. Table 4.2 shows if putative exons show similarity to ESTs. A number of I.M.A.G.E. (Image) cDNA clones (Lennon *et al.*, 1996) which can be obtained from resource centres (HGMP, Hinxton) were used as probes in hybridisations to filters containing the exon trapped clones (Chapter 2.20.1 step 16, Figure 4.4K). This can identify new species of putative exons mapping to the same cDNA (see for example Image clone 824710 in Figure 4.14).
- Searches were performed against the non-redundant Genbank databases of mouse ESTs and the non-mouse and non-human ESTs (Figure 4.4L). This was performed at the website http://www.ncbi.nih.gov/Blast/. The only exon trapped clone that showed similarity only to a non-human EST was 270M7-L10. This clone resembled the rabbit EST AI052988, from an adult aorta cDNA library. The other exon trapped

Characteristics of the exon trapped clones

Exon trapped clone	size	ORF, protein similarities	ESTs	Comments
90B5BB-A22	297bp	98AA, R, multidrug resistance-associated protein-2 (*Rattus norvegicus*, 7e-4)	AI076090 e-134, H47627 e-107, N54854 e-100, N58518 e-7	
90B5BB-C21	244bp	80AA, R, putative RNA binding Protein 7e-4	AA639269 e-125	
98L15BB-J11	160bp	52AA-stop, R		
98L15BB-J19 =98L15S-K10	94bp	31AA, R		
98L15BB-K11	177bp	Stop-58AA, F		
98L15BB-N7 =98L15S-C3	494bp	164AA, R, PS-PLA1 (*Rattus norvegicus*) 1e-24	AA483957 e-120	
90B5S-B24	85bp	28AA, F		
90B5S-E6	341bp	113AA, R, STCH 2e-59	AA307875 0.0, T06636 e-165, AF000596 e-165, AA332837 e-108	
90B5S-E17	150bp	49AA, F, reverse transcriptase related protein (*H. sapiens*) 7e-20	101 blast hits	L1 repetitive element, maps back
90B5S-G5	204bp	67AA, F, ORF of *E. coli* 5e-30	H14022 e-36, H13987 2e-29	*E coli* genomic DNA (non-redundant database), does not map back
90B5S-J11	115bp	3x38AA, 37AA, F/R, human homologue of yeast mismatch repair gene 2e-6	100 blast hits	MER repetitive element, maps back
98L15S-A9 =98L15BB-L11	126bp	42AA, F	D61628 3e-65, AA737090 3e-62, AA742713 9e-50, AA639269 2e-38, AA648334 9e-16	
98L15S-B8	37bp	2x11AA, 12AA, F/R		
98L15S-E9	58bp	2x19AA, F/R		
98L15S-G9	62bp	2x20AA, F/R, canicular multidrug resistance-associated protein 2 E=8.3	H47929 4e-25, AI076090 e-16, H47627 9e-8	
126N20-A3	213bp	44AA-stop-25AA, F		
126N20-A22	142bp	47AA, R, multidrug resistance protein (*H. sapiens*) 2e-4		Exon 31 of MRP1
126N20-C4	159bp	53AA, F		

Table 4.2

126N20-E1	47bp	4x15AA, F/R		
126N20-N22	354bp	97AA, R, multidrug resistance protein (*H. sapiens*) e-10		Exon 29 of MRP1
31B5A3=B39I12-F7(218bp,52AA)	152bp	50AA, R	AA471280 3e-16	
31B5-E21 =31B5-B5	358bp	119AA, R	AA035703 8e-19	
31B5-M21	75bp	2x25AA, F/R		
B39I12-A1	132bp	20AA-stop-21AA, F		
B39I12-D2	354bp	118AA, R		
B39I12-E18	392bp	71AA-stop-58AA, F		
68C10/132H4-M11 =190N10-J9	132bp	2x43AA, 44AA, F/R		
190N10-H1	87bp	29AA, R		
190N10-H8	99bp	33AA, R		
190N10-K14	106bp	35AA, R		
153I22-A2	227bp	75AA, R		
153I22-A6	126bp	2x42AA, 41AA, F/R	120 blast hits	OFR repetitive element, maps back
153I22-A23	192bp	43AA, 39AA, 37AA, R		
153I22-H13	155bp	Stop-50AA, R		
153I22-M15	146bp	48AA, F		
30P13-A1	264bp	78AA, 62AA, 60AA, F/R	44 blast hits	Repetitive element, maps back
30P13-B4 =270M7-A3	103bp	34AA, R	AA604555 8e-19, AA632790 7e-16	
30P13-E4	55bp	18AA, 17AA, R		
30P13-F11	158bp	40AA, 35AA, 32AA, F/R	AI192191 7e-82, AA043647 9e-60, H58539 3e-53, H83025 e-21	Contains MER 10 repeat, maps back
30P13-H3	96bp	31AA, F		
270M7-D15	113bp	3x37AA, F/R, ORF1; MER 37 putative transposase, similar to pogo element, *H. sapiens* 4e-7	99 blast hits	MER 37 transposable element
270M7-L10	147bp	48AA, R		
270M7-M11	100bp	32AA, R		
270M7-N2	381bp	43AA, F		
B194B7-B5	125bp	41AA, R		

Table 4.2

B194B7-C8	115bp	38AA, R		
B194B7-H4	208bp	52AA, 36AA, 32AA, F/R		
B194B7-K5	78bp	3x25AA, F/R		
B194B7-L21	125bp	41AA, F		
103K19-A12	166bp	2x36AA, F		
103K19-B3	316bp	39AA, F		
103K19-F4	208bp	31AA-stop-37AA, R, ALU1 subfamily J (*H. sapiens*) 9e-5	104 blast hits	ALU repetitive element, maps back
103K19-J5	145bp	14AA-stop-32AA, R		
103K19-J13	95bp	31AA, F		
103K19-K10	82bp	2x27AA, R		
10J13-B24	253bp	6xORF, 29-44AA, F/R		
10J13-H23	236bp	78AA, F		
10J13-L3	97bp	32AA, R		
10J13-N5	175bp	58AA, R		
65F3/44E10-A1	106bp	2x35AA, 34AA, F/R	91 blast hits	MER10 repetitive element, maps back
65F3/44E10-A6	139bp	45AA, F		
65F3/44E10-D22	199bp	Stop-64AA, F		
65F3/44E10-I16	145bp	48AA, F		
65F3/44E10-K6	172bp	57AA, 56AA, R		
75G13-B5	97bp	7AA-stop-24AA, F		
75G13-D6 =265A22-F9	121bp	39AA, 40AA, F		
75G13-F12	199bp	Stop-64AA, F		
75G13-H7	126bp	2x41AA, F		
75G13-I20 =265A22-A12	69bp	3x22AA, 23AA, F/R, human homologue of yeast mismatch repair gene 4e-3	100 blast hits	OFR.t1 repetitive element, maps back
75G13-N5 =265A22-I3	134bp	2x22AA, 13AA, F/R	102 blast hits	MER 18 repetitive element, maps back
265A22-F20	458bp	39AA, 37AA, 35AA, 33AA, 31AA, F/R		
265A22-G23	388bp	35AA, 33AA, F		
265A22-K20	324bp	46AA, R		
190A9-A3	241bp	25AA-stop-53AA, R		

Table 4.2

190A9-E2	349bp	54AA-stop-61AA, F		
190A9-H12	165bp	54AA, 50AA, R, ORF2 (*C. familiaris*) 2e-6, reverse transcriptase (*H.sapiens*) 5e-4	100 blast hits	L1 repetitive element, maps back
190A9-H21	174bp	58AA, R		
111K10-A5 (first 105 bp=190A9-H18)	275bp	48AA, F		
111K10-B5	253bp	67AA, 46AA, 43AA, F/R		
111K10-H5 =73M5-I13	145bp	48AA, R	AI092963 2e-76, AA482291 2e-76	
111K10-K12	232bp	36AA-stop-40AA, R	AI092963 6e-37, AA482291 6e-37	
73M5-A2	163bp	54AA, R	AA482201 e-80	
73M5-F7	138bp	46AA, 45AA, F/R		
73M5-H7	87bp	29AA, R	AA482201 3e-42	
73M5-H10	205bp	68AA, F, repA protein phage P7 2e-33		Does not map back specifically
73M5-J13	154bp	50AA, R		
73M5-K13	124bp	40AA, R	AI092963 e-24, AA482291 4e-24	
73M5-K17	128bp	42AA, F		
135E14-A3	145bp	48AA, 47AA, F		
135E14-B4	79bp	2x26AA, 25AA, F/R	Z45010 8e-9, AA644617 8e-9	A002B43 Unigene cluster
135E14-B7	143bp	47AA, R	W52737 e-55, AA24980 7e-11	
135E14-C6	689bp	229AA, R, probable ubiquitin carboxyl-terminal hydrolyse K02C4.3 (*C. elegans*) 7e-13	AA081200 0.0, AA385363 e-152, AA382386 e-132, W05786 7e-98, AA307805 7e-98, AA525465 e-90, W52737 e-65	
135E14-D6	77bp	2x25AA, F/R	AA482201 5e-13	
135E14-G12	113bp	2x37AA, R	93 blast hits	Contains repeat, maps back
135E14-O13	133bp	40AA, 39AA, F/R		

Table 4.2: The names of the exon trapped clones that were sequenced are shown (column 1), together with their size (column 2). Column 3 shows sizes of the open reading frames (ORFs) and the frame in which the ORFs were found (F/R, Forward/Reverse): reverse is the expected frame compared with the β-globin exons in pSPL3B and it shows similarity to proteins in the SwissProt database. Column 4: Human ESTs that were partly identical to the analysed clones. Column 5: Additional comments.

clones with resemblance to non-human ESTs also showed similarity to human ESTs. These are described in Chapter 4.3.3.

4.3.3 Similarities between exon trapped clones and human cDNA clones

4.3.3.1 Exon trapped clone 90B5S-E6

The exon trapped clone 90B5S-E6 is part of the previously published gene STCH (Otterson, 1994), which is a member of the stress70 chaperone family. This exon trapped clone was identified by hybridisation of a radio-actively labelled probe to the membrane containing the exon trapped clones from the genomic clone 90B5 (*Sau* 3A I preparation), which contains the gene (see Figure 3.13). The labelled fragment was generated by RT-PCR with the primers STCH 1A and 1B from bone marrow RNA (J.H. Ives), the fragment was isolated from agarose gel by the Prep-A-Gene method. Hybridisation results can be seen in Table 2 of Appendix I (384-well grid of 90B5S, nrs. 6 and 7). Two mouse ESTs (AI226564 and AI226600) and one rat EST (H32090) showed similarity to the exon trapped clone 90B5S-E6.

4.3.3.2 Exon trapped clone 98L15BB-N7

As shown in Figure 4.8, clone N7 from the exon trapping experiment 98L15BB was shown to be very similar to an EST from Ewings sarcoma; Image clone 910220. This clone was sequenced, the cDNA insert being 323 bp. Table 2 in Appendix I shows that hybridisation of this EST to the filter containing the 98L15BB exon trapped clones resulted in positive clone N7 (nr. 6 in Appendix I Table 2). This clone shows similarity to a rat lipase gene (e-24, Table 4.2), as determined from the SwissProt comparison of the amino acid sequence.

4.3.3.3 Exon trapped clones 98L15S-G9 and 90B5BB-A22

Figure 4.9 shows that 98L15S-G9 is identical to part of two human ESTs, Image clones 193617 and 1673263. Subsequent BlastN searches with these ESTs revealed the other clones in this cluster. Inserts of Image clones 193617, 248395 and 248318 were hybridised to the filters containing the exon trapped products (nrs. 16, 15 and 18 in 98L15/90B5 BB, nrs. 20 and 19 in 98L15S, and nrs. 20, 19 and 22 in 90B5S respectively, see Appendix I Table 2). This revealed an additional exon trapped clone mapping back to the ESTs; clone 90B5BB-A22 was sequenced and mapped to the ESTs as shown in Figure 4.9. Identical clones on the exon trap filters were revealed by

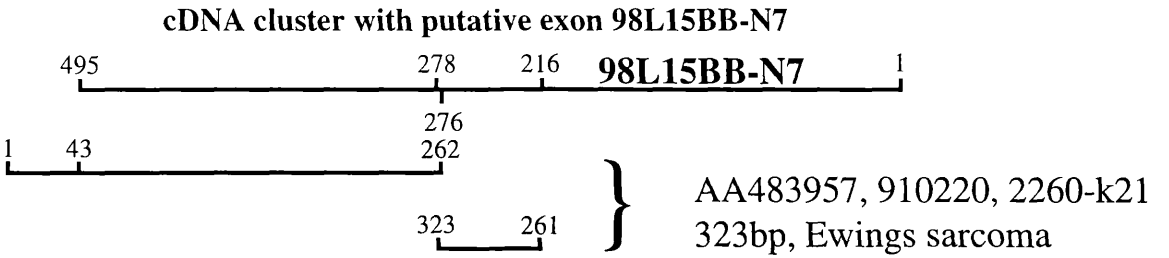


Figure 4.8: Identical sequence between human EST 910220 and exon trapped clone 98L15BB-N7. The EST is identified using the following information: AA483957 is the accession nr. for the sequence; 910220 is the Image accession nr.; 2260-k21 is the HGMP accession nr.; 323bp shows the insert size of the cDNA clone; Ewings sarcoma is the tissue the cDNA library was constructed from.

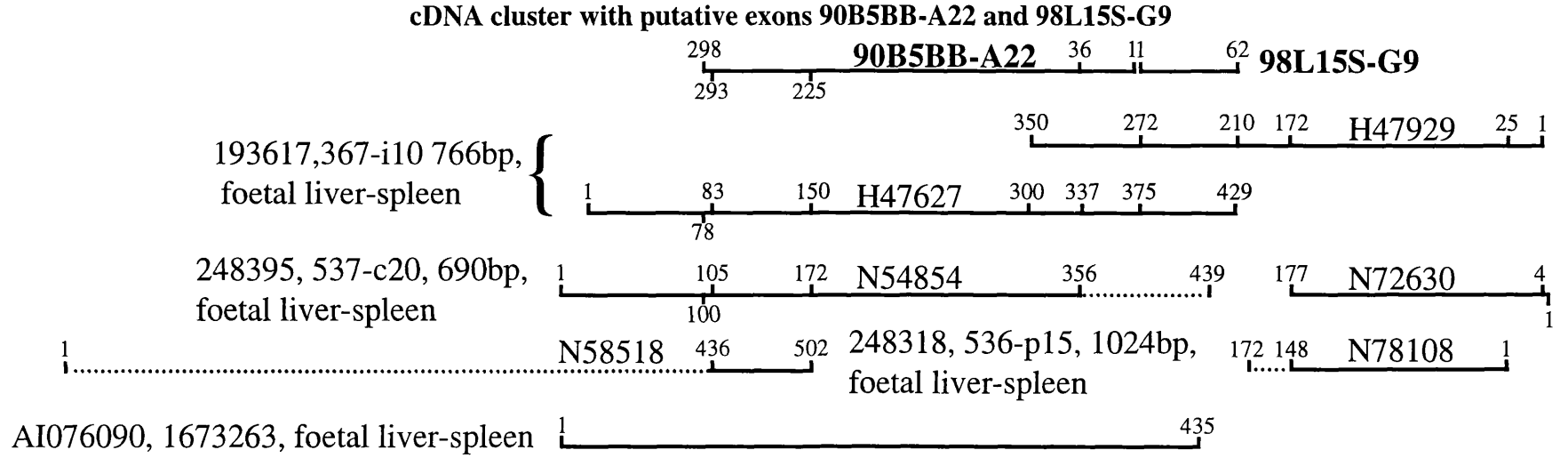


Figure 4.9: Human EST cluster, which is partly identical to 90B5BB-A22 and 98L15S-G9. Additional information with EST clones is as described in Figure 4.8. A dotted line indicates no similarity.

hybridisation (nr. 19 in 98L15/90B5BB, and nr. 5 in 90B5S, Appendix I Table 2). 90B5BB-A22 showed similarity to rat multidrug resistance associated protein 2 and 98L15S-G9 showed homology with a similar protein of *Arabidobsis thaliana.*

4.3.3.4 Exon trapped clones 98L15S-A9 and 90B5BB-C21

Figure 4.10 shows the similarity of 98L15S-A9 to a human EST cluster. The ESTs contain a head-to-tail repeat of 119bp, separated by 9bp. One of these repeats is also present in 98L15S-A9 although its mapping position is the right hand one in Figure 4.10 (the sequence of 98L15S-A9 starts with the nine unique nucleotides, followed by the repeat). Inserts of the cDNA clones of 124053, 1182683 and 1183050 were used as probes for hybridisation to the filters containing the exon trapped clones (nrs. 12, 10, 11 in 98L15/90B5BB, nrs. 7, 8, 9 in 98L15S, and nrs. 8, 9, 10 in 90B5S, see Appendix I Table 2). Hybridisation of 1182683 revealed additional exon trapped clones other than the clones represented by 98L15S-A9. These were all one species and are represented by 90B5BB-C21 (nr. 13 in 98L15/90B5BB, 12 in 98L15S and 13 in 90B5S, see Appendix I Table 2). SwissProt analysis shows the highest similarity of 90B5BB-C21 with a putative RNA-binding protein from *Homo sapiens.*

4.3.3.5 Exon trapped clones 31B5-E21 and 31B5-A3/B39I12-F7

Part of the exon trapped clone 31B5-E21 was shown to be identical to the EST 360025, which identified the ESTs 1286563 and 136169; this can be seen in Figure 4.11. Hybridisation of the insert of the Image clone 1286563 identified clones represented by 31B5-E21 and 31B5-A3 (Appendix I Table 2, Table 31B5 nrs. 6,7,11). 31B5-A3, which is contained within B39I12-F7, is identical to part of AA471280, which resembles T71075 (Figure 4.11). The clones 31B5-E21 and 31B5-A3/B39I12-F7 also show similarity to mouse ESTs. The ESTs with the accession numbers AA560016, AA560511 and AA798427 resemble both exon trapped clones very closely and show them to be neighbouring exons. The mouse cDNA clones AA596448 and AA288461 show similarity to 31B5-E21, whereas AA756594 and AA921588 show similarity to 31B5-A3/B39I12-F7.

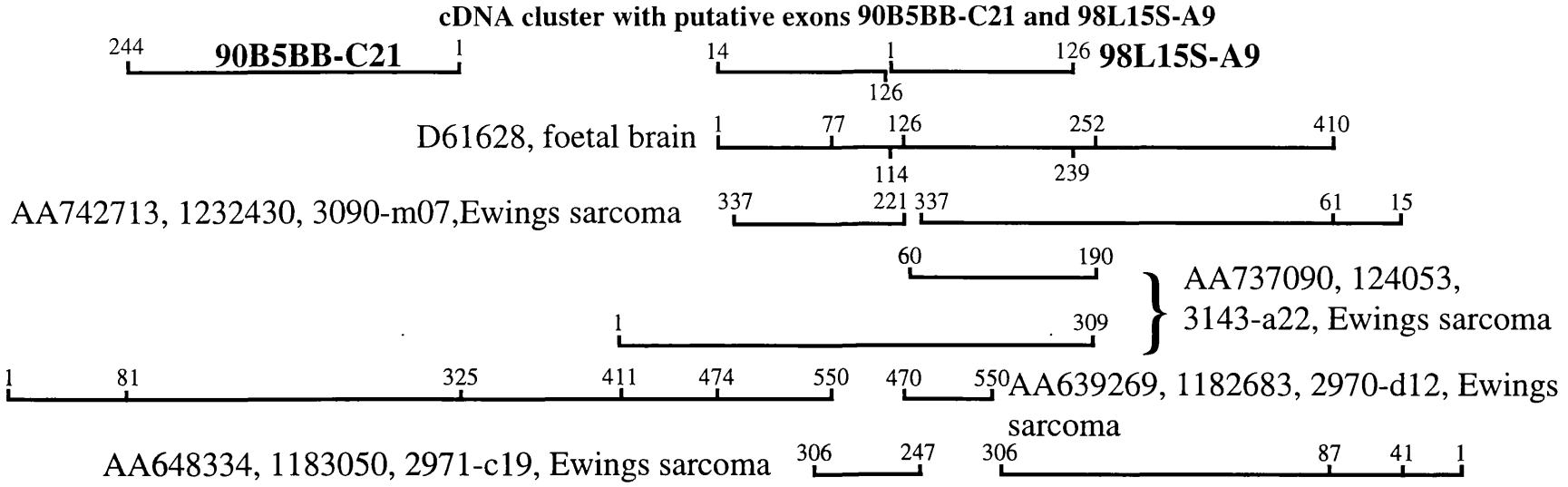


Figure 4.10: A human EST cluster which are partly identical to exon trap clones 90B5BB-C21 and 98L15S-A9 is shown. Additional information with EST clones is as described in Figure 4.8.

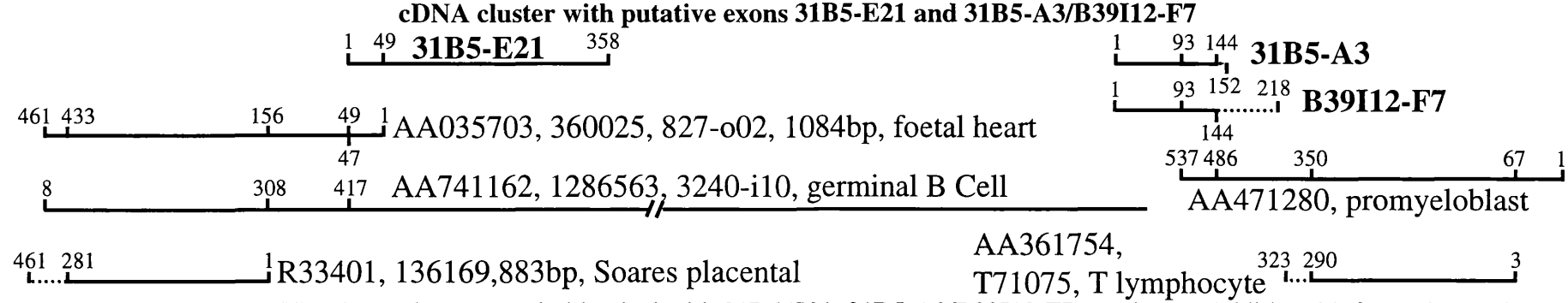


Figure 4.11: Human EST clones that are partly identical with 31B5-E21, 31B5-A3/B39I12-F7 are shown. Additional information with EST clones is as described in Figure 4.8. A dotted line indicates no similarity. Only the first 417 bp of the Image clone 1286563 is present in the human EST database.

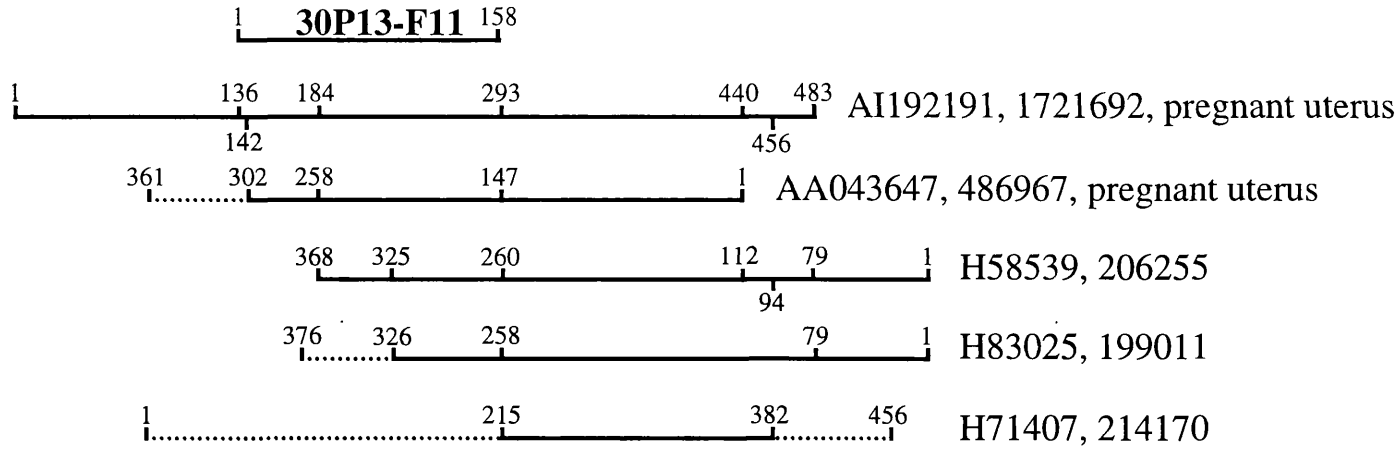


Figure 4.12: Exon trapped clone 30P13-F11 and the similarity to human ESTs is shown. Additional information with EST clones is as described in Figure 4.8. A dotted line indicates no similarity.

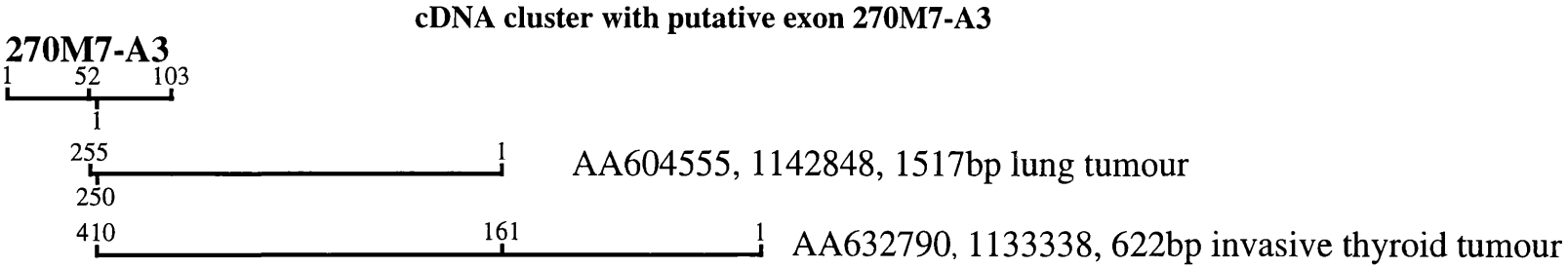


Figure 4.13: Sequence similarity is shown between the putative exon 270M7-A3 and the human EST clones 1142848 and 1133338. Additional information with EST clones is as described in Figure 4.8.

4.3.3.6 Exon trapped clones 30P13-F11 and 270M7-A3 (J.H. Ives)

The exon trapped clone 30P13-F11 showed similarity to four human cDNA clones as shown in Figure 4.12. 270M7-A3 was very similar to two cDNA clones, which can be seen in Figure 4.13.

4.3.3.7 Exon trapped clones 135E14-D6, 73M5-A2, 73M5-I13/111K10-H5, 73M5-H7, 73M5-K13 and 111K10-K12

Figure 4.14 shows that 135E14-D6, 73M5-A2 and 73M5-I13/111K10-H5 were identical to part of the ESTs AA482201 or AA482291. These sequences are derived from the same cDNA clone, 824710. This clone was sequenced and hybridised to the exon trap filters of 111K10 and 73M5 (hybridisation results of clone 824710 to a dot blot showed positive signals for the PAC clones 111K10 and 73M5 but did not hybridise to 135E14). These experiments revealed three extra species of exon trapped clones, represented by 73M5-H7, 73M5-K13 and 111K10-K12 (see Appendix I Table 2). These clones were hybridised to the filters containing the exon trapped clones, sequenced, and subsequently positioned on the ESTs as indicated in Figure 4.14. A number of non-human ESTs showed similarity to these exon trapped clones. The mouse cDNA clones AA217497, AI390583, AI465477 and the rat EST AI231533 resembled the exons 73M5-I13/111K10-H5 and 111K10-K12, and the rat clones AI501530 and AI030865 were similar to 111K10-K12. Additionally, the mouse EST AA930443 resembled 73M5-A2.

4.3.3.8 Exon trapped clones 135E14-B7 and 135E14-C6

Sequencing of clone 135E14-B7 and a BlastN search to the human EST database showed similarity to Image clone 338628 (Figure 4.15). A subsequent Blast search of this EST identified the cDNA clones with the accession numbers 549350, 299864 and AA248980. Hybridisation of the insert of cDNA clone 299864 (nr. 13 in 135E14 exon trap products in Appendix I Table 2) identified clones represented by 135E14-C6 (nr. 14 in the same Table). Sequencing of this clone revealed the mapping position and additional ESTs in this cluster (Figure 4.15). Clone 135E14-C6 showed similarity (7e-13) to a putative ubiquitin carboxyl-terminal hydrolase from *Caenorhabditis elegans* (Table 4.2). Three clones from the mouse EST database (AI464870, AA267185 and AA267185) show similarity to the clone 135E14-C6.

cDNA cluster with putative exons 135E14-D6, 73M5-H7, 73M5-A2, 73M5-K13, 73M5-I13/111K10-H5 and 111K10-K12

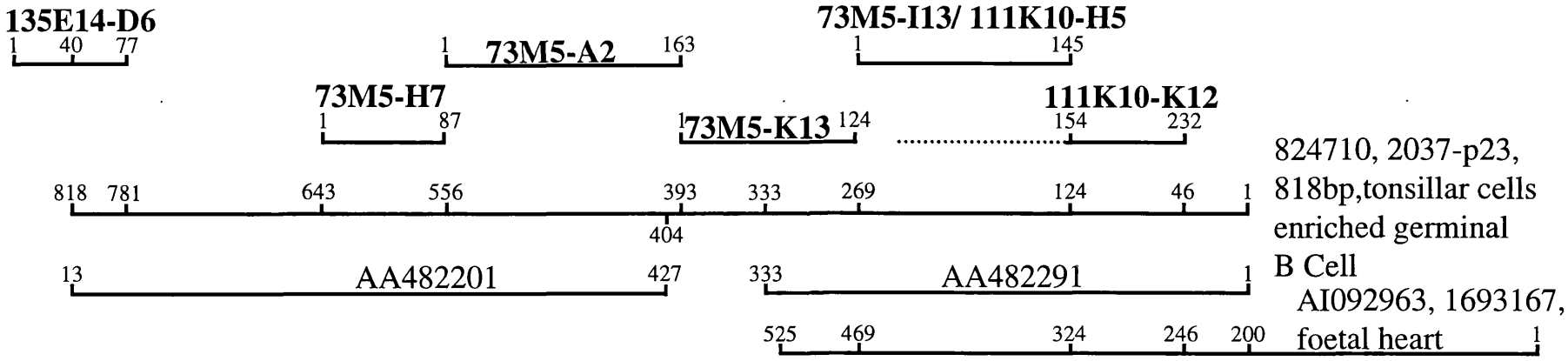


Figure 4.14: Sequence similarity is shown between EST clones 824710, 1693167 and exon trapped clones mapping to these ESTs. Additional information with EST clones is as described in Figure 4.8. A dotted line indicates no similarity.

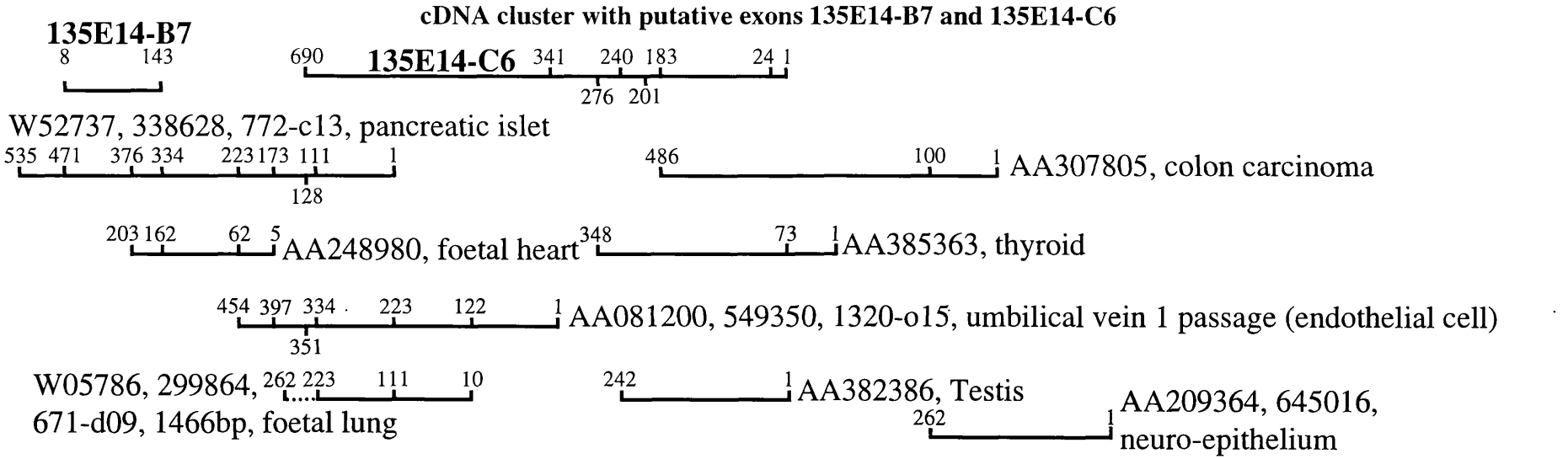


Figure 4.15: Sequence similarity is shown between exon trap clones 135E14-B7, 135E14-C6 and EST clones. Information with the EST clones is as described in Figure 4.8. A dotted line indicates no similarity.

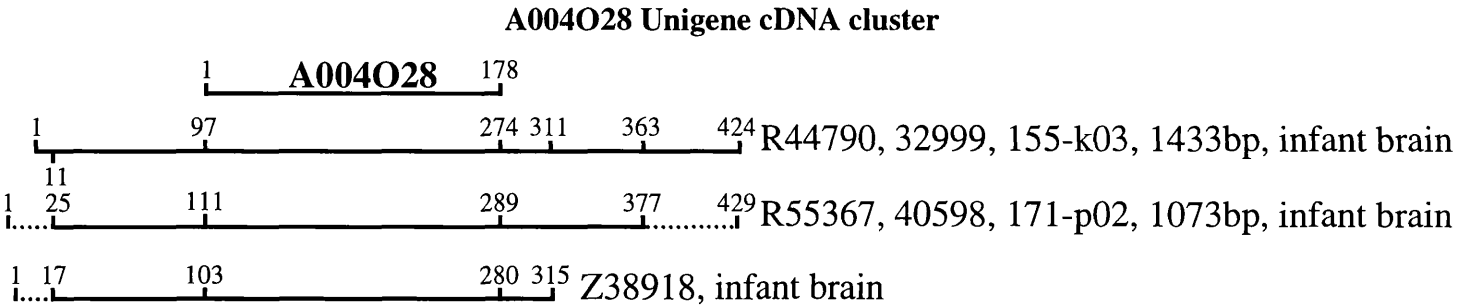


Figure 4.16: This figure shows the Unigene cluster A004O28. No exon trapped clones were detected that hybridised to these cDNA clones. Additional information with EST clones is as described in Figure 4.8 A dotted line indicates no similarity.

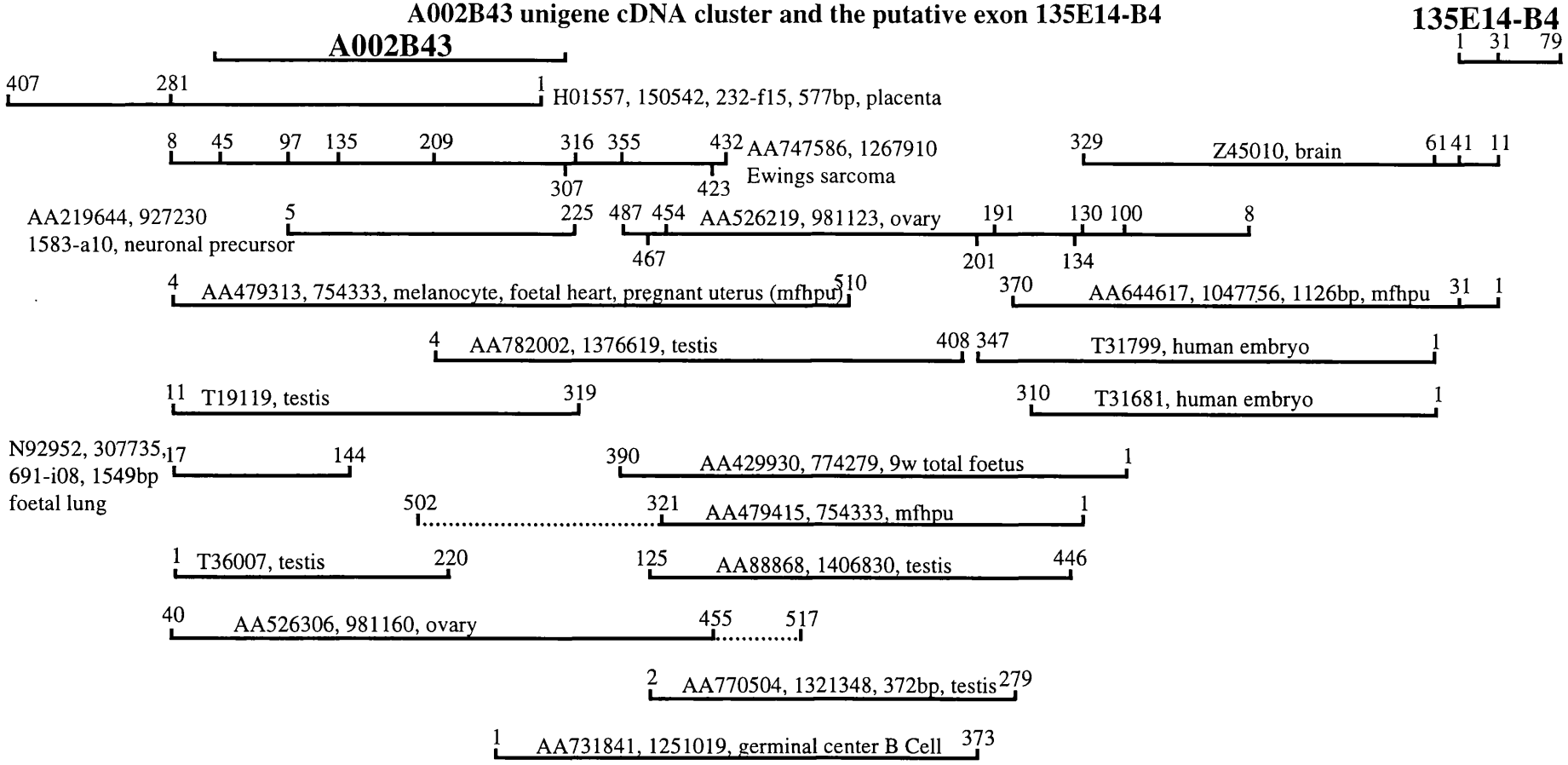


Figure 4.17: This figure depicts the human ESTs forming the Unigene cluster A002B43. Exon trap clone 135E14-B4 was shown to be identical to part of ESTs Z45010, AA644617 and hybridised to insert from the cDNA clone 307735. Additional information with EST clones is as described in Figure 4.8. A dotted line indicates no similarity.

4.3.3.9 cDNA clusters A004O28 and A002B43 and exon trapped clone 135E14-B4

cDNA clones from two cDNA clusters that were localised to the area (Chapter 3.3.5) were also used as probes in hybridisation experiments to the exon trapped products of the appropriate clones. Radio-active probes made from inserts of Image clones 32999 and 40598 of the A004O28 cluster failed to hybridise to any of the exon trapped clones from the PAC clone 265A22 (Figure 4.16), but did map back to this clone on dot blots. Exon trapped clone 135E14-B4 was shown to be identical to two clones of the cDNA cluster A002B43 (Z45010 and 1047756, see Figure 4.17). This was subsequently confirmed by hybridisation of the insert of clone 307735 to the exon trapped products of 135E14 (see nr. 9 for 135E14-B4 and nr. 6 for 307735 in Appendix I Table 2 for clone 135E14). The clone 135E14-B4 also appeared to be very similar to two cDNA clones from the mouse EST database, AA590785 and AA555695.

4.4 Discussion

A total of twenty-two overlapping bacterial clones containing genomic inserts from the 21q11 region were used in forty exon trapping experiments. These clones are the MTP clones telomeric to and including 98L15 (see Figure 3.13). Clones from this region of the contig were chosen for exon trapping for the following reasons:

- The region contains the marker D21S120 which showed the highest frequency of loss of heterozygosity in non-small cell lung carcinoma (Kohno *et al.*, 1998).
- Three of the four markers that showed an increase in disomic homozygosity in DS individuals with ANLL-M7 are present in this region (D21S258, D21S120 and D21S192 (Abe *et al.*, 1989; Niikawa *et al.*, 1991; Ohta *et al.*, 1996; Shen *et al.*, 1995; Cavani *et al.*, 1998)) (Figure 3.13).
- Linkage was found with the marker D21S16 in some families with EOAD (St. George-Hyslop *et al.*, 1987, Goate *et al.*, 1989), which is present on the PAC clone 31B5 (Figure 3.13).
- This part of the contig is specific for chromosome 21 (rather than the region proximal to the clone 98L15, see Chapter 3.4.2), which will simplify subsequent identification of full length genes.

In total, forty exon trapping experiments were performed (excluding the negative controls), which consisted of *Bam* H I/*Bgl* II and *Sau* 3A I digested genomic clones. From these exon trapping experiments, twenty-one 384-well microtitre dishes were

generated containing putative exon trapped clones (Table 4.1) and from these, membranes were prepared for further analysis. A total of 7105 exon trapped clones were present in the gridded collection.

The filters containing DNA from the potential exons were hybridised with the HIV *tat* intron and COT 1 DNA, to identify clones that contained the intron or repetitive DNA respectively. The clones containing the HIV *tat* intron probably resulted from co-isolation of DNA with RNA (see Figure 4.3 lanes 4 and 8) or from RNA molecules still containing the complete or partial intron (non-splicing events or splicing using a cryptic splice site within the intron), which had not been removed by *Bst* X I digestion. A total of 279 (3.9%) of the exon trapped clones still contained HIV *tat* intron and 224 (3.2%) clones contained repetitive DNA, though not all clones containing repetitive DNA were removed by COT 1 hybridisation. Table 4.2 shows that fourteen of the ninety-five exon trapped clones still contain repetitive DNA. A possible explanation for this could be that the repeats present in these exon trapped clones are not similar enough to the repeats which are most abundant in COT 1 DNA to be detected by high stringency hybridisation. Nevertheless, twelve of these fourteen clones mapped back specifically to the genomic clone they were derived from.

Three probes were hybridised to the exon trapped filters that assisted in the identification of clones containing an insert; no probe could be generated that distinguished exon trapped clones that contain an insert from empty ones.

- The genomic insert from the PAC and BAC clones was hybridised to the membrane containing the exon trapped products generated from the respective genomic clones. This probe consisted of a large number of radio-actively labelled fragments which represent the genomic insert. The average genomic insert was approximately 120 Kbp. The average insert from the putative exons was 180 bp (from Table 4.2) resulting in a poor signal-to-noise ratio (see Figure 4.5.1). The best signal-to-noise ratio is probably achieved with the largest inserts from the exon trapped clones.
- 3^0 BB EX1/EX2 and 3^0 S EX1/EX2 fragments were generated, via a nested PCR, from the PCR products, which were amplified with the primers dUSA4 and dUSD2 in the exon trapping protocol. The radio-active probe derived from these PCR products consisted of specific fragments (EX1 and EX2 are located just adjacent to the putative exon inserts, see Figure 4.1) representing the exon trapped products present on the membranes. This composite probe is likely to hybridise mostly to the larger and more abundant species of exon trapped products, which are present on the

membranes. An example can be seen in Table 2 in the Appendix I; four exon trapped clones derived from the genomic clone 190N10 (190N10-K14, -H8, -J9 and H1) hybridised to 235 of the 330 exon trapped clones on this filter (72%).

The strategy which was followed to reduce the sequence analysis of identical and empty clones was highly successful; only two identical exon trapped clones from one parental genomic clone were sequenced (31B5-E21 and 31B5-B5), and no empty clones were sequenced. An even greater reduction in the unnecessary sequencing of identical clones can be achieved when filters, which contain exon trapped clones from overlapping genomic clones, are included in the hybridisation. This has not been carried out for these experiments. Eleven of the ninety-five sequenced clones are completely or partially identical to exon trapped clones that were derived from neighbouring genomic clones (Table 4.2).

The sequences of ninety-five different exon trapped clones were determined, of which ninety-three mapped back specifically to the parental genomic clones. Eighty-seven of these were hybridised back to filters containing exon trapped clones derived after *Bam* H I/*Bgl* II and *Sau* 3A I digestion (approximately half of the eighty-seven clones originated from either method of digestion). Forty-three clones hybridised to identical exon trapped clones, which were generated from both *Bam* H I/*Bgl* II and *Sau* 3A I digestions, twenty-two hybridised only to clones from the *Sau* 3A I digestion and another twenty-two hybridised only to clones generated after *Bam* H I/*Bgl* II digestion. This showed that the two different methods for subcloning the genomic clones in pSPL3B both generated significant amounts of different putative exons. The ninety-three clones hybridised positively to 2598 clones (36.6%) on the twenty-one exon trapped filters. Sixty-three clones contained an ORF of the whole sequence, forty-eight of these showed the ORF in the expected reading frame compared with the rabbit β-globin exons (see Table 4.2). Nineteen of the exon trapped clones, that mapped back were identical to cDNA clones in the human EST database and eight closely resembled non-human ESTs (mouse, rat and rabbit). The exon trapped clones that were identical to cDNA clones can be seen in Figures 4.8-4.17; the clone 90B5S-E6 is part of the characterised gene STCH (not shown). This gene was shown to be located between D21S378 and D21S120 (Brodsky *et al.*, 1995) on the genomic PAC clones 90B5 and 126N20 (Figure 3.13). The fact that an exon of this gene was identified from the genomic clone 90B5, indicates that the exon trapping experiments might also lead to the identification of novel genes in the area. From Figures 4.8-4.17 it can be seen that one

cDNA clone was identical to six exon trapped clones, four cDNA clones were identical to two exons and a further four were identical to one exon trapped clone. Eighteen of the nineteen exon trapped clones with similarity to cDNA clones had an ORF of the entire clone (the exception was 111K10-K12) and eighteen also showed the ORF in the expected reading frame (except 98L15S-A9). Part of the clone 111K10-K12 did not show similarity with the Image cDNA clone 824710 (indicated with a dotted line in Figure 4.14). This may have been caused by cryptic splice sites present in the genomic DNA. The clone 111K10-K12 was found to be part of the novel gene USP21 (Chapter 5). This gene did not contain the part of 111K10-K12 which showed no similarity to the Image clone 824710. Another example of this can be seen in Figure 4.11 where part of the clone B39I12-F7 did not resemble the cDNA clone AA471280. The sequence indicated by the dotted line in Figure 4.11 was not part of the novel gene G31B5 (Chapter 6). This shows that an exon trapped clone does not have to contain a perfect ORF and can be chimaeric, consisting of an exon and a piece of genomic DNA not represented in the mRNA. Additionally, exon trapped clones that do not show an ORF consisting of the entire clone could be present partially or totally in the 5' or 3' UTRs.
A number of interesting similarities between translations of ORFs of the exon trapped clones and protein sequences in the SwissProt database are shown (see Table 4.2).

- The clone 98L15BB-N7 shows similarity to PS-PLA1 (e-24), a lipase from rat.
- A number of clones show similarities (up to e-10) to the multidrug resistance or associated proteins, these are 90B5BB-A22, 98L15S-G9, 126N20-A22 and 126N20-N22.
- The translation of the ORF from 135E14-C6 shows resemblance to a putative ubiquitin carboxyl-terminal hydrolase from *Caenorhabditis elegans* (7e-13).

In conclusion, a total of twenty-two bacterial clones containing overlapping genomic inserts, which total approximately 2 Mbp of sequence, were used in exon trapping experiments. More than 7000 exon trapped clones were included in the analysis, which generated ninety-three putative exons that mapped back specifically to the genomic clones. The analysis of the putative exons included mapping of their position on the genomic contig, determination of ORFs, comparison of ORF translation with the SwissProt protein database and identification of similarities to human cDNA clones present in databases. The gene STCH, which is present within the genomic clones that

were used for exon trapping, functioned as a positive control. An exon of this gene was successfully trapped showing that the method worked. The closest similarities of amino acid translations were found between putative exons and a rat lipase, a multidrug resistance protein and a ubiquitin carboxyl-terminal hydrolase. This work is likely to result in the identification of novel full length genes that can be tested for their involvement in the phenotypes that are associated with the 21q11 region.

Chapter 5 Identification of a novel gene mapping to proximal 21q encoding a ubiquitin-specific protease

5.1 Summary

The exon trapping experiments described in Chapter 4 identified a cDNA clone (Image clone 824710) that showed sequence similarity to six putative exons (see Figure 4.14). Hybridisation of this cDNA clone to a Northern blot identified an mRNA molecule of approximately 4.1 Kbp in size, in all tissues tested. A gene was subsequently identified using genomic sequence from the inserts of two bacterial clones, overlapping cDNAs and additional exon trapped clones. The gene is 3803 bp long, contains twenty-four exons and occupies approximately 149 Kbp of genomic DNA. The gene contains a 1055 amino acid open reading frame (ORF), which shows similarities to ubiquitin-specific proteases (USPs). This function has been proven after cloning of the gene in a prokaryotic expression vector and de-ubiquitination of a ubiquitin-β-galactosidase fusion protein in *E. coli*. Sequencing of the gene using genomic DNA from patients suffering from non-small cell lung carcinoma (NSCLC) and loss of heterozygosity (LOH) for polymorphic markers in the area, does not show mutations. Also, no mutations were detected in a limited number of patients with late onset Alzheimer's disease (LOAD).

5.2 Introduction

5.2.1 Genetic defects of chromosome 21

Down syndrome (DS) is caused by trisomy of chromosome 21 (Lejeune *et al.*, 1959; Jacobs *et al.*, 1959). More than twenty phenotypes can be identified in DS individuals, although not every individual with DS will show all the phenotypes. Some of these phenotypes are: mental retardation, an increased risk of childhood leukaemia, congenital heart disease, Alzheimer's disease and hypotonia (Epstein *et al.*, 1991; Zipurski *et al.*, 1992, Korenberg *et al.*, 1992; 1994; Kesslak *et al.*, 1994). A number of genetic defects in genes on chromosome 21 have also been identified associated with diseases in karyotypically normal individuals. Some examples are: SOD 1 in amyotrophic lateral sclerosis (Rosen *et al.*, 1993), the EPM1 gene in progressive myoclonus epilepsy (Pennacchio *et al.*, 1996; Lalioti *et al.*, 1997), APP in patients with familial early onset Alzheimer's (FEOAD) (Goate *et al.*, 1991) and autoimmune polyendocrinopathy-

candidiasis-ectodermal dystrophy (The Finnish-German APECED Consortium, 1997; Nagamine *et al.*, 1997).

5.2.2 Disorders associated with chromosome 21q11

We are interested in the genes that map to the pericentromeric region of chromosome 21, which might contain genes involved in the following phenotypes:

- acute non-lymphocytic leukaemia of the immunotype M7 (ANLL-M7) (megakaryoblastic leukaemia). This is approximately 400 times more frequent in DS children compared with karyotypically normal infants (Zipurski *et al.*, 1987). A self regressing transient abnormal myelopoieisis is also thought to be related to ANLL-M7 and is almost exclusively observed in DS children (Zipurski *et al.*, 1987; 1992). In DS individuals with either ANLL-M7 or TAM, an increase in disomic homozygosity of polymorphic markers, has been observed in the 21q11 region (Abe *et al.*, 1989; Niikawa *et al.*, 1991; Shen *et al.*, 1995; Ohta *et al.*, 1996; Cavani *et al.*, 1998). It has therefore been proposed that this region might contain a gene with tumour suppressor or growth control function that is involved in the pathogenesis of ANLL-M7 and TAM. (Niikawa *et al.*, 1991; Ohta *et al.*, 1996).
- Analysis of polymorphic markers have shown linkage of AD with the marker D21S16 from the 21q11 region in some FEOAD patients (St. George-Hyslop *et al.*, 1987; Goate *et al.*, 1989). Although mutations were subsequently found in the same families in the presenilin-1 gene on chromosome 14, the linkage found between these families and 21q11 might be explained by the presence of a putative modifier gene in this region (St George-Hyslop *et al.*, 1992; Van Broeckhoven, 1995). Genome wide linkage analysis of affected sib pairs with LOAD was carried out in which a distinction was made between those that contained an APOE-ε4 allele compared to those that did not. Positive lod scores were detected in the APOE-ε4 negative subgroup in a large region of the proximal long arm of chromosome 21 (including the 21q11 region), the highest score was found just proximal of the APP gene (Kehoe *et al.*, 1999). A proposed second stage study, in which more sibling pairs and additional markers will be used, might reveal whether a gene involved in late onset Alzheimer's disease (LOAD) is present in 21q11.
- Studies showed that the long arm of chromosome 21 was often deleted in squamous NSCLC (Sato *et al.*, 1994; Testa *et al.*, 1994). LOH of polymorphic markers of the

21q11-q21 region varied between 62 and 42% in squamous NSCLC, the highest loss being found with the marker D21S120, which is located in 21q11.2 (Kohno *et al.*, 1998). The same authors observed a homozygous deletion of the 21q11-q21 region in a lung cancer cell line, and estimated it to be larger than 3 Mbp.

A novel full length gene is described in this Chapter, located in the cytogenetic band 21q11. Exons from this gene were identified via exon trapping from the PAC clones 73M5 and 135E14, which were described in Chapter 4. Overlapping cDNAs and the exon trapped clones were used in conjunction with the genomic sequence from two PAC clones 73M5 and 135E14, which were identified in Chapter 3 (Figure 3.13) The amino acid sequence derived from this gene showed similarities to ubiquitin-specific proteases (USPs). It was determined if this gene could function as a USP, by cloning of the coding region of the gene into a prokaryotic expression vector and co-expression with a ubiquitin-β-galactosidase fusion gene. Mutation detection in DNA samples from patients suffering from NSCLC and LOAD were also performed.

5.3 Results

5.3.1 Identification of a novel gene encoding a possible ubiquitin-specific protease.

Exon trapping experiments showed six different exon trapped clones (135E14-D6, 73M5-H7, 73M5-A2, 73M5-K13, 73M5-I13/111K10-H5 and 111K10-K12) to be identical to part of the Image cDNA clone 824710, which can be seen in Figure 4.14. The cDNA insert from this Image clone was isolated from an agarose gel via Prep-A-Gene, after restriction digestion with *Not I/Eco* R I. This fragment was radio-actively labelled and hybridised to the Human I Northern blot (Clontech). The result of this experiment can be seen in Figure 5.1A; a 4.1 Kbp fragment was visible in all tissues present on the blot. The signal was most intense in skeletal muscle, with a less intensely hybridising product of 4.9 Kbp also being visible in all lanes. The PCR product that was used to map the cDNA clones (represented by the Unigene cluster A002B43) to proximal 21q (Schuler *et al.*, 1996) and to the PAC clone 135E14 (Chapter 3.3.5; Groet *et al.*, 1998), was also used as a probe for hybridisation. It was radio-actively labelled and hybridised to the Human II multiple tissue Northern blot (Clontech). A fragment of 4.1 Kbp hybridised in all the tissues present on the Northern blot, but was most intense in testis. This tissue showed an additional band of approximately 1.4 Kbp (Figure 5.1B). Hybridisation to similarly sized products on the Northern blot and their presence on

Hybridisation of the Image clone 824710 and the cDNA cluster A002B43 to multiple tissue Northern blots

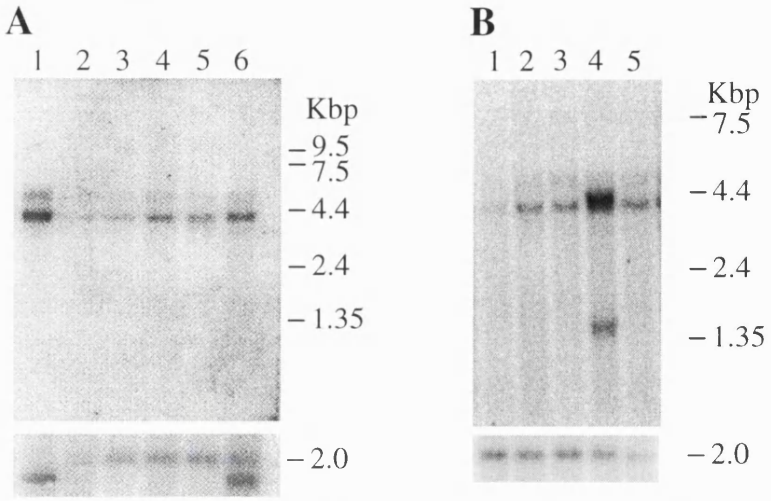


Figure 5.1 : Northern blots (Clontech) containing 2 μg poly A+ RNA per lane. The top picture in panel A shows hybridisation of the insert of Image cDNA clone 824710 to the Human I Northern blot: lane 1 -skeletal muscle, lane 2 -liver, lane 3 -lung, lane 4 -placenta, lane 5 -brain, lane 6 -heart. Panel B shows hybridisation of a PCR product representing the cDNA cluster A002B43 to the Human II Northern blot, lane 1 -colon (mucosal lining), lane 2 -small intestine, lane 3 -ovary, lane 4 -testis, lane 5 - prostate. The bottom picture in both panels shows the same blots hybridised to the β-actin probe (Clontech). A product of 4.1 Kbp is observed with both probes in all tissues. A minor product of approximately 4.9 Kbp can also be seen. The highest level of expression for cDNA 824710 is in skeletal muscle. Expression is highest in testis for A002B43, in this tissue an extra 1.4 Kbp fragment can also be seen.

overlapping genomic clones indicated that the Image cDNA clone 824710 and the cDNA cluster A002B43 may be part of the same gene.

The majority of this novel gene was identified by overlapping sequence information from partially sequenced cDNA clones (BlastN searches using the indicated accession numbers at http://ncbi.nlm.nih.gov/BLAST/) and sequences of exon trapped clones (see Table 3 of Appendix I). The sequence of the two most distally mapping clones of the contig from Chapter 3, 73M5 and 135E14 was fully determined and publicly released (Chromosome 21 Sequencing Consortium, Braunsweich, Germany) thus becoming accessible from Genbank (73M5-AJ010597, 135E14-AJ010598) during the identification of this gene. The mapping position of the different exons in relation to the PAC clones 73M5 and 135E14 is shown in Figure 5.2. This figure also shows the different cDNA and exon trapped clones that were used to identify the gene. A BlastN search was performed with the Image clone 824710 to the non-redundant Genbank+EMBL+DDBJ+PDB databases, and sequences were found which were identical to DNA sequences at eight different positions on clone 73M5. Six of these were also identical to the six exon trapped clones, which were previously sequenced (Figure 4.14). Exons 7 and 8 were present on both PAC clones 73M5 and 135E14. One additional cDNA clone (AI092963) was identified when a BlastN search was performed to the human EST database, using the sequence from Image clone 824710. This cDNA clone did not reveal additional exons, but did show additional sequence from the first exon of the gene. A BlastN search of exon 135E14-D6 to the human EST database detected the cDNA clone AA209364. This cDNA clone revealed two additional exons, exon 9 mapped to PAC clones 73M5 and 135E14, exon 10 only to 135E14. AA209364 was identical to part of the cDNA clone AA307805, which was determined via a BlastN search to the human EST database. AA307805 identified the additional exons 11-13 (see Figure 5.2) and had already been shown to overlap with a previously sequenced exon trapped clone 135E14-C6 (see Figure 4.15). A BlastN search to the non-redundant database showed that this exon trapped clone consisted of four different exons (exons 13-16 in Figure 5.2). From Figure 4.15 it had previously been determined that cDNA clone W52737 overlapped with exon trapped clone 135E14-C6. W52737 was shown to be identical to exons 16-19, exon trapped clone 135E14-B7 being exon 18. In order to confirm that exons 18 and 19 were neighbouring exons in the mRNA, an RT-PCR experiment was carried out. This was performed for two reasons: the quality of this part of the sequence of cDNA clone W52737 in the database was poor, and sequence from a

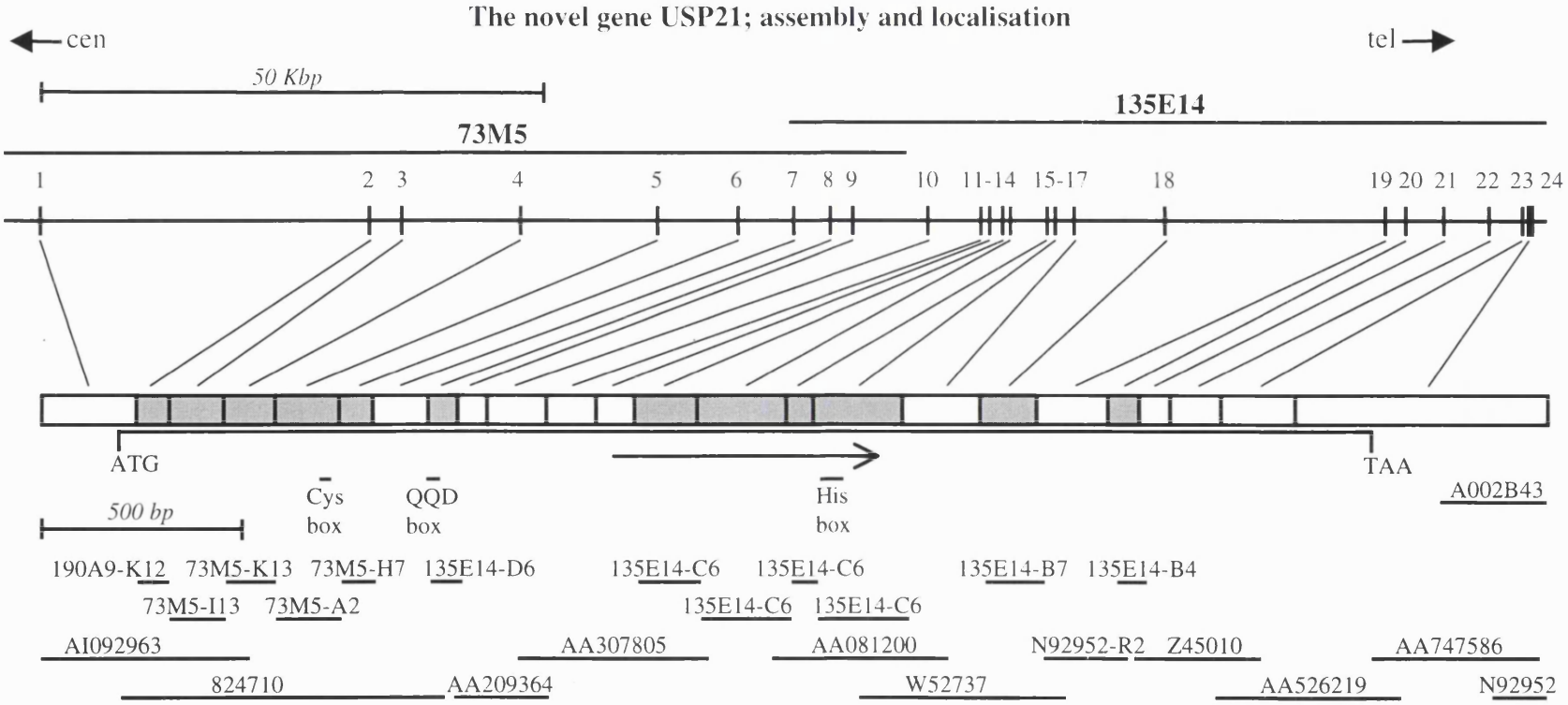


Figure 5.2: Composite picture of the novel gene USP21. The genomic location of the 24 exons is shown on the PAC clones 73M5 and 135E14 (accession numbers AJ010597 and AJ010598 resp.). The largest ORF, 3165 bp 1055 amino acids, and the direction of transcription (cen-tel) is shown below the gene. The exon trapped clones that were isolated in Chapter 4 which are part of this gene are indicated in grey. The start and stop codons are shown, as are the functionally important domains, the Cys, His and QQD boxes. The cDNA cluster represented by the Unigene clone A002B43 and the Image cDNA clone 824710, which were used as probes for the Northern blot, are indicated.The exon trapped clones and the cDNA clones which were used for identification of the gene are indicated underneath. A scale marker is indicated for the genomic DNA (50 Kbp) and for USP21 (500 bp).

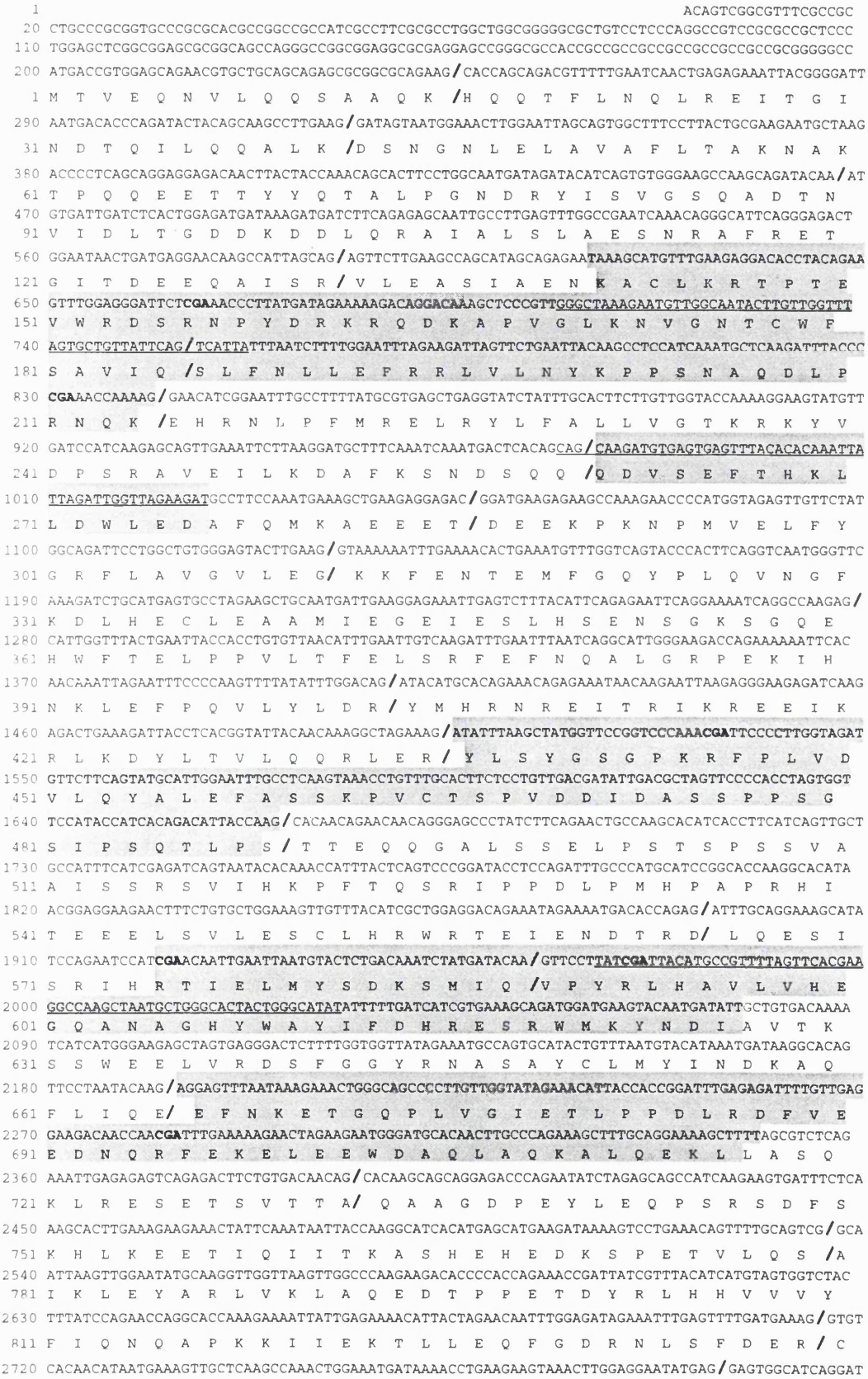
cDNA clone (Image 307735, N92952-R2 see next section) mapped between exons 18 and 19 in the genomic DNA sequence from clone 135E14. RT-PCR was performed using bone marrow mRNA (Clontech). This was reverse transcribed with random primers after which a PCR was performed with primers USP A and B, which are present in exons 18 and 19 of the novel gene, respectively. This PCR product was sequenced using the primer USP A. The sequence showed exons 18 and 19 to be neighbours in the gene (the sequence is shown in Table 4 of Appendix I); the sequence was identical to the genomic sequence from 135E14 (AJ010598).

Image cDNA clone 307735/N92952, which is part of the A002B43 Unigene cluster (Figure 4.17), was requested from HGMP. DNA was isolated from this clone, after which a PCR reaction was performed using primers M13F+ and M13R+ amplifying the cDNA insert. Both ends of this PCR product were sequenced using the M13F+ and M13R+ primers. The sequence generated by the M13F+ primer was identical to the 3' end of the gene (identical to N92952). The other sequence (generated with the primer M13R+) mapped between exons 18 and 19 (the sequence is shown in Table 4 of Appendix I; N92952-M13R+). From this last sequence, primer N92952-R2 was designed, which was used to extend the known sequence at this end of the cDNA clone 307735. This sequence bridged between W52737 and 135E14-B4, and was identical to exons 19 and 20. The mapping position of the sequence generated with the primer N92952-R2 can be seen in Figure 5.2 and the sequence is shown in Table 4 of Appendix I (N92952-R2). As can be seen from Figure 4.17, 135E14-B4 (exon 20) is identical to cDNA clones that are part of a large group of overlapping cDNA clones. The Unigene amplimer A002B43 (Schuler *et al.*, 1996) is also part of this cluster (Figure 4.17). BlastN searches to the non-redundant database of the overlapping cDNA clones Z45010, AA526219, AA747586 and N92952 revealed the exons 21 to 24. Figure 4.17 shows that six cDNA clones all start with approximately the same sequence (see Figure 4.17, cDNA clones AA747586, AA479313, T19119, N92952, T36007 and AA526306), suggesting that this is the 3' end of the gene. In addition, 3' rapid amplification of cDNA ends (RACE) was performed to confirm these findings (the protocol is detailed in Chapter 2.18.12). PCR products were amplified using the linker primer AP1 and the gene-specific primer USP 3' RACE. A subsequent PCR was performed using a dilution from the first PCR reaction, the primers being AP2 (nested linker primer) and USP 3' RACE. Fragments generated by these PCR reactions were separated by agarose gel electrophoresis and subsequently Southern blotted. A hybridisation using the radio-

actively labelled insert from Image clone 307735 (which contains N92952) was used to identify positively hybridising fragments from the 3' end of USP21. A 200 bp fragment hybridised positively to the probe, which was then isolated from an agarose gel, cloned into a TA vector (InVitrogen) and transformed into *E. coli.* Colonies were analysed for specific inserts by PCR (described in Chapter 2.20.1 step 16), using the primers USP 3' RACE and M13F+ or M13R+. The products of two RACE clones were sequenced from these PCR products via the primers USP 3' RACE and USPE-24A. The 3' end of the gene (indicated in Figure 5.3), corresponds to nucleotide 74,206 of the genomic sequence from the PAC clone 135E14. The two 3' RACE clones show a poly A tail after nucleotides 74,200 and 74,204 (not shown) confirming this to be the 3' end of the gene.

As previously mentioned, an additional 1.4 Kbp hybridising mRNA fragment was observed in testis when a Northern blot was hybridised to the Unigene probe A002B43 (Figure 5.1). This probe is located in the 3' UTR and is identical to part of the human EST N92952. This EST represents the sequence from one end of the Image clone 307735. The other end of this Image cDNA clone mapped between exons 18 and 19 in the genomic DNA, as discussed above. This sequence was determined by using the M13R+ primer and is homologous to nucleotides 44,992-45,167 of PAC clone 135E14, and contains an ATG start codon, which is in frame with the ORF of exons 19-24. The primers N95952-A and B were designed to establish if this could be the alternative start of transcription, which may give rise to the 1.4 Kbp mRNA fragment in testis (these primers amplify a PCR product from PAC clone 135E14 corresponding to nucleotides 44,996-45,166). The PCR product was isolated via Prep-A-Gene, radio-actively labelled and hybridised to human Northern II blot (Clontech), which contains testis. The probe failed to hybridise to a specific product on the Northern, but did hybridise to the PAC clone 135E14 (not shown). Although this specific PCR product is not part of the 1.4 Kbp testis mRNA species, it could be an alternative start of transcription in foetal lung from which tissue the cDNA clone 307735 was derived. Foetal lung RNA is not present on the human Northern II blot. The PCR product which contained exon 10 of the novel gene that was generated with the primers USPE-10F/R using genomic DNA from NSCLC tumour from patient #13 (see also Chapter 5.3.4), was used as a radio-active probe. This was hybridised to the Clontech human Northern II blot, which contained testis; both the 4.1 and the 1.4 Kbp fragment hybridised positively to this probe. The use

Nucleotide and amino acid sequence of USP21



```
 841 H  N  I  M  K  V  A  Q  A  K  L  E  M  I  K  P  E  E  V  N  L  E  E  Y  E  /E  W  H  Q  D
2810 TATAGGAAATTCAGGGAAACAACTATGTATCTCATAATTGGGCTAGAAAATTTTCAAAGAGAAAG/TTATATAGATTCCTTGCTGTTCCTC
 871 Y  R  K  F  R  E  T  T  M  Y  L  I  I  G  L  E  N  F  Q  R  E  S  /Y  I  D  S  L  L  F  L
2900 ATCTGTGCTTATCAGAATAACAAAGAACTCTTGTCTAAAGGCTTATACAGAGGACATGATGAAGAATTGATATCACATTATAGAAGAGAA
 901 I  C  A  Y  Q  N  N  K  E  L  L  S  K  G  L  Y  R  G  H  D  E  E  L  I  S  H  Y  R  R  E
2990 TGTTTGCTA/AAATTAAATGAGCAAGCCGCAGAACTCTTCGAATCTGGAGAGGATCGAGAAGTAAACAATGGTTTGATTATCATGAATGAG
 931 C  L  L  /K  L  N  E  Q  A  A  E  L  F  E  S  G  E  D  R  E  V  N  N  G  L  I  I  M  N  E
3080 TTTATTGTCCCATTTTTGCCATTATTACTGGTGGATGAAATGGAAGAAAAGGATATACTAGCTGTAGAAGATATGAGAAATCGATGGTGT
 961 F  I  V  P  F  L  P  L  L  L  V  D  E  M  E  E  K  D  I  L  A  V  E  D  M  R  N  R  W  C
3170 TCCTACCTTGGTCAAGAAATGGAAC/CACACCTCCAAGAAAAGCTGACAGATTTTTTGCCAAAACTGCTTGATTGTTCTATGGAGATTAAA
 991 S  Y  L  G  Q  E  M  E  P/  H  L  Q  E  K  L  T  D  F  L  P  K  L  L  D  C  S  M  E  I  K
3260 AGTTTCCATGAGCCACCGAAGTTACCTTCATATTCCACGCATGAACTCTGTGAGCGATTTGCCCGAATCATGTTGTCCCTCAGTCGAACT
1021 S  F  H  E  P  P  K  L  P  S  Y  S  T  H  E  L  C  E  R  F  A  R  I  M  L  S  L  S  R  T
3350 CCTGCTGATGGAAGATAAACTGCACACTTTCCCTGAACACACTGTATAAACTCTTTTTAGTTCTTAACCCTTGCCTTCCTGTCACAGGGT
1051 P  A  D  G  R  * 1055
3440 TTGCTTGTTGCTGCTATAGTTTTTAACTTTTTTTTATTTTAATAACTGCAAAAGACAAAATGACTATACAGACTTTAGTCAGACTGCAGA
3530 CAATAAAGCTGAAAATCGCATGGCGCTCAGACATTTTAACCGGAACTGATGTATAATCACAAATCTAATTGATTTTATTATGGCAAAACT
3620 ATGCTTTTGCCACCTTCCTGTTGCAGTATTACTTTGCTTTTATCTTTTCTTTCTCAACAGCTTTCCATTCAGTCTGGATCCTTCCATGAC
3710 TACAGCCATTTAAGTGTTCAGCACTGTGTACGATACATAATATTTGGTAGCTTGTAAATGAAATAAAGAATAAAGTTTTATTTATGGCTA
3800 CCTA 3803
```

Figure 5.3: Complete nucleotide and translated peptide sequence of USP21. This gene contains a 199 bp 5' UTR, a 3165 bp, 1055 amino acid open reading frame and a 439 bp 3' UTR. USP21 contains 24 exons, which are separated by /. The areas coloured grey represent nucleotide sequences, which were determined in patients with non-small cell lung carcinoma. These areas contain the Cys, QQD, and His boxes (these are underlined), and sequences around CGA codons (CGA codons are indicated in bold). The CGA codons towards the 3' end of the coding sequence have not been sequenced. This sequence was submitted to Genbank (accession number AF134213, which will be released on 03-01-2000) and which will be published (Groet *et al.*, in press)

Sizing of exons and introns, analysis of splice junctions and determination of the phase of the introns of USP21

Exon				Intron		
Nr.	Size (bp)	3' Acceptor	5' Donor	Nr.	Phase	Size (bp)
1	199+45		CAGAAG/gtgagg	1	0	32452
2	78	ctctag/CACCAG	TTGAAG/gtatgg	2	0	3028
3	145	ttgaag/GATAGT	ATACAA/gtaagt	3	1	11762
4	124	tttcag/ATGTGA	TAGCAG/gtaaaa	4	2	13474
5	163	attaag/AGTTCT	ATTCAG/gtaagg	5	0	8092
6	87	tttcag/TCATTA	CAAAAG/gtaaaa	6	0	5277
7	138	caatag/GAACAT	CAGCAG/gtagtt	7	0	3550
8	77	tcacag/CAAGAT	GGAGAC/gtaagt	8	2	2251
9	74	ctttag/GGATGA	TTGAAG/gtagag	9	1	7487
10	149	ttatag/GTAAAA	CAAGAG/gtgagt	10	0	5192
11	128	tttcag/CATTGG	GGACAG/gtatgg	11	2	799
12	96	tgatag/ATACAT	AGAAAG/gtattt	12	2	1142
13	162	taacag/ATATTT	ACCAAG/gtaaaa	13	2	611
14	227	atacag/CACAAC	CCAGAG/gtaaga	14	1	3327
15	71	ccttag/ATTTGC	ATACAA/gtaagt	15	0	799
16	229	atttag/GTTCCT	TACAAG/gtaaga	16	1	1718
17	201	ttgcag/AGGAGT	CAACAG/gtttgt	17	1	8849
18	143	tagaag/CACAAG	CAGTCG/gtaaga	18	0	21727
19	179	ttccag/GCAATT	TGAAAG/gtaatt	19	2	1819
20	79	tttaag/GTGTCA	TATGAG/gtaatg	20	0	3723
21	80	ctacag/GAGTGG	AGAAAG/gtaaag	21	2	4255
22	124	ttcaag/TTATAT	TTGCTA/gtaagt	22	0	3269
23	196	attaag/AAATTA	TGGAAC/gtaagt	23	1	311
24	170+439	ttgcag/CACACC				
	3803	(c,a)ag/(A,G)	(A,C)AG/gt(a,g)agt			**144,914**
		70,100,100/65	78,91,78/100,100,96,74,61,44			

Table 5.1: Columns 1 and 2 show the number and the size of the exons, respectively. Exon 1 contains a 199 bp 5' UTR, exon 24 a 439 bp 3' UTR. The total size of the gene indicated in bold is 3803 bp. Columns 3 and 4 show the splice junctions, column 3 the 3' acceptor, and column 4 the 5' donor. The consensus splice junctions are indicated underneath (c,a)ag/(A,G) for the 3' acceptor and (A,C)AG/gt(a,g)agt for the 5' donor (Grant *et al.*, 1997). The percentages of the individual nucleotides from the USP21 gene compared with the consensus sequence are indicated in the bottom row. Column 5 shows the intron number. Column 6 indicates the phase of the introns and column 7 shows the size of the individual introns, with the total intronic size indicated in bold. The intronic sizes were calculated from the sequence from the genomic clones 73M5 and 135E14 (accession nrs. AJ010597 and AJ010598 respectively). The total genomic distance that this gene occupies is 148,717 bp.

of 5' and 3' RACE from testis cDNA may be used to elucidate the exons that are present in the 1.4 Kbp product.

The total length of the sequence of the novel gene is 3803 bp, which was determined from the overlapping cDNA and exon trapped clones. This number, in combination with a poly A tail, corresponds well with the 4.1 Kbp fragment observed as the main band on the Northern blot. The first in-frame start codon can be found at nucleotide 200, from which a 3165 bp open reading frame starts, coding for a protein of 1055 amino acids. This is followed by a 439 bp 3' UTR. The complete nucleotide and amino acid sequence can be seen in Figure 5.3. The full length sequence was submitted to Genbank (accession number AF134213, to be released on 03-01-2000) and has been accepted for publication (Groet *et al.*, in press) This gene consists of 24 exons (shown in Figure 5.2 and Table 5.1), and spans approximately 149 Kbp of genomic DNA. The sizes of the individual exons can be found in Table 5.1. This table also shows the sequences of the intron/exon boundaries, which correspond well with the consensus, which is also shown. The phase and sizes of the introns are also indicated in Table 5.1.

The amino acid sequence was subsequently compared with protein sequences in the SwissProt database. The novel protein showed the best matches with ubiquitin-specific proteases (USPs), the e values ranging from e^{-6}-e^{-24}. The e value is the probability of hits that are expected by chance, when searching a particular database (http://www.ncbi.nlm.nih/BLAST/blast_FAQs.html). A low score indicates a significant similarity. Similarities in this protein sequence were detected with proteins from different organisms, including *Caenorhabditis elegans*, *Homo sapiens*, *Saccharomyces cerevisiae* and *Drosophila melanogaster.* USPs are a group of proteins that can cleave ubiquitin from precursor proteins, and can regenerate ubiquitin from poly-ubiquitin units, which are formed during targeting of proteins for degradation, via the 26S proteasome pathway (Wilkinson, 1997). The three main, short, protein domains which are conserved in USPs are the Cys, His and QQD boxes. Figure 5.4 shows the similarities in amino acid sequences between different USPs and the novel protein characterised in this study, henceforth referred to as USP21 (ubiquitin-specific protease on chromosome 21). Figure 5.4 shows that USP21 shares the key amino acids necessary for the function of USPs with different members of this family and is likely to be a USP. A less significant similarity is found between the USP21 sequence and the heavy chain of non-muscle myosin ($7e^{-05}$).

Comparison of amino acids between the functional domains of USP21 and of other ubiquitin-specific proteases

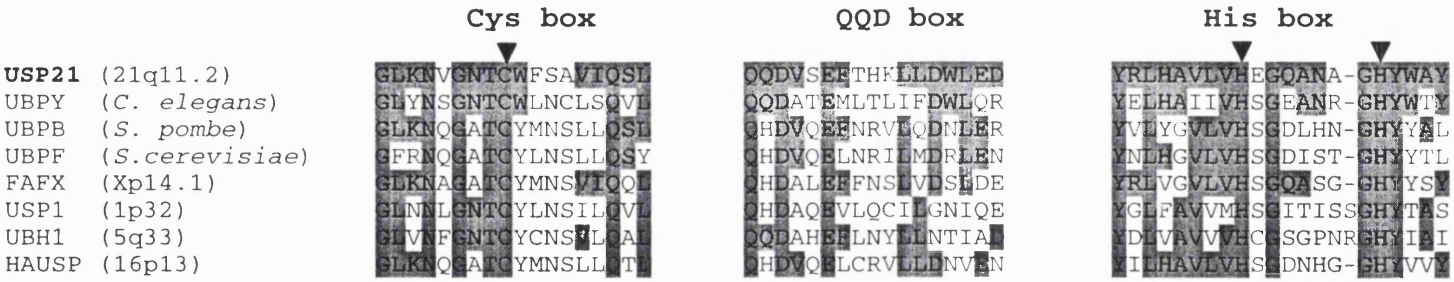


Figure 5.4: The different USP genes are indicated on the left hand side. The chromosomal location is indicated for the human genes. The amino acids that are in common between the family members are indicated in grey. The key functional amino acids for the Cys and the His boxes are indicated by a triangle.

5.3.2 *In vitro* characterisation of the novel gene as a ubiquitin-specific protease

(protein assay and Western blotting in collaboration with P. Chazot)

In order to establish if the USP21 protein can function as a USP, co-expression of a ubiquitin-β-galactosidase fusion gene and USP21 in an *E. coli* host was performed. USP21 was cloned as indicated in Figure 5.5. RT-PCR was performed using bone marrow poly A+ RNA (Clontech). The resulting cDNA was used to amplify (by PCR), the coding sequence of USP21 in three parts. Unique *Eco* R I and *Xba* I restriction endonuclease sites within the coding region of USP21, and *Sac* I and *Sal* I restriction sites which were present in the primers USPEXP 1 and 2 were used. The primer combinations that were used were USPEXP1/USP-ER, USP-EF/USP-XR and USP-XF/USPEXP2 (Figure 5.5). After amplification and isolation of the specific PCR products, the respective fragments were digested with *Sac* I/*Eco* R I, *Eco* R I/*Xba* I and *Xba* I/*Sal* I and cloned in the corresponding restriction sites of pBluescript. This resulted in the constructs USP-A, -B and -C. The three inserts were sequenced using internal and M13 primers (Chapter 2.21.1). The sequence of the three inserts was identical to the USP21 sequence indicated in Figure 5.3, except for one single, silent A to T substitution at nucleotide 2029. This nucleotide was also present in clinical material from some of the patients with NSCLC and can be classed as a natural sequence variation (single nucleotide polymorphism). Inserts of USP-B and -C were co-ligated in the *Eco* R I/*Sal* I sites (USP-BC) and the insert of USP-A was then introduced (USP-ABC). The *Sac* I/*Sal* I fragment was subsequently cloned into the identical sites of the prokaryotic expression vector pQE30 (Qiagen), resulting in pQE-USP21. This construct contained the coding sequence of USP21 from nucleotide 203-3367, in frame with the ATG initiation codon of pQE30. Co-transformants were generated in *E. coli* of a construct containing a ubiquitin-Met-β-galactosidase (Ub-M-βgal) fusion gene with the following vectors: pQE30 (negative control), pRB105 a *S. cerevisiae* USP (UBP2) gene (positive control) or pQE-USP21. Co-transformants containing the different constructs were grown uninduced or induced with IPTG, after which protein lysates were prepared. Approximately 10 μg of total protein from these lysates was loaded on a 7% SDS-PAGE gel and a Western blot was prepared. Detection of products was established using an anti-β-galactosidase polyclonal rabbit antiserum, an anti-rabbit IgG-HRP secondary antibody and an enhanced-chemiluminescence assay kit (Amersham). Figure 5.6 shows the result from the Western blot; odd and even numbered lanes show protein

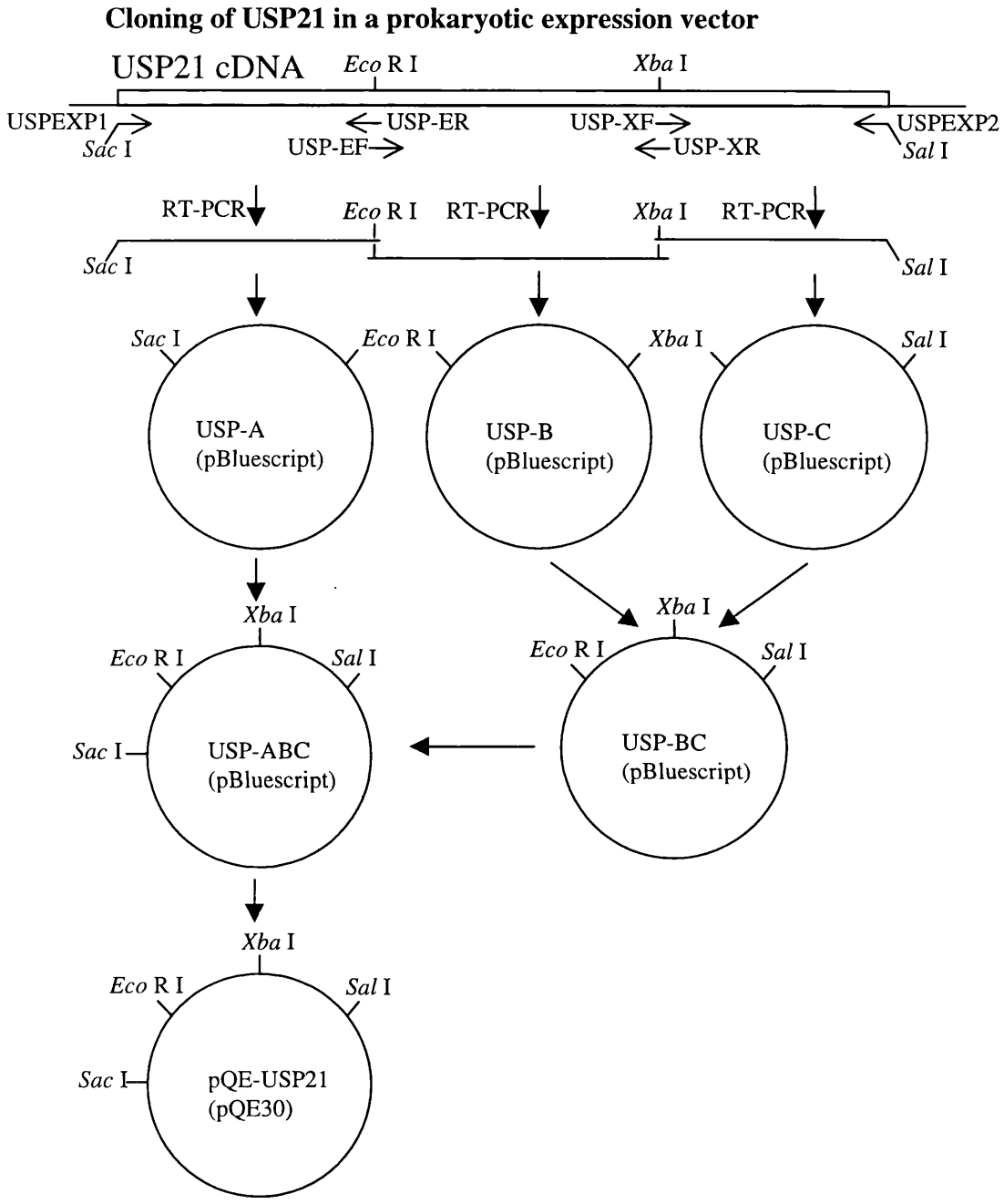


Figure 5.5: The cloning of the USP21 coding DNA in the prokaryotic expression vector pQE30 is shown. PCR was performed from reverse transcribed bone marrow RNA with the indicated primers (horizontal arrows), after which the three separate fragments were cloned individually in pBluescript, resulting in USP-A, -B, -C. These fragments were sequenced to eliminate PCR artefacts. Fragments B and C were cloned in pBluescript (USP-BC), after which fragment A was introduced (USP-ABC). The *Sac* I-*Sal* I fragment, containing the coding sequence of USP21, was subsequently cloned into the same sites in pQE30, resulting in the construct pQE-USP21. The USP coding DNA was cloned in frame with the ATG initiation codon from pQE30.

In vitro* cleavage of ubiquitin-M-β-galactosidase by *USP21

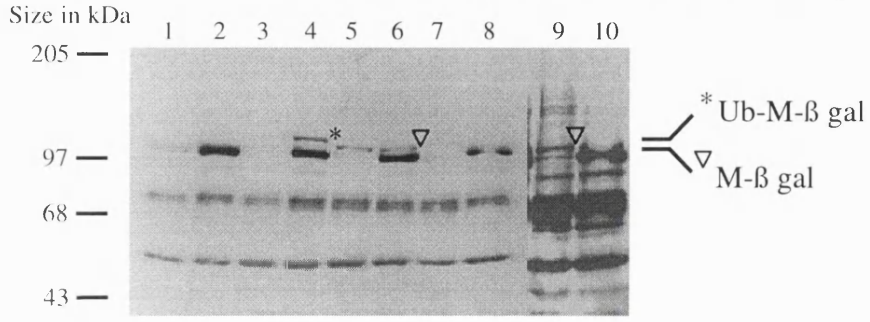


Figure 5.6: Co-expression of a ubiquitin-M-β-galactosidase (Ub-M-βgal) fusion gene with the USP21 gene (construct pQE-USP21), to characterise the USP21 gene as a ubiquitin-specific protease. Clones were generated using XL-1 blue cells which were transformed with the following constructs: Ub-M-βgal + pQE30 (negative control), Ub-M-βgal + pQE-USP21 (USP21 coding DNA cloned in pQE30), Ub-M-βgal + UBP2 (positive control, USP from *S. cerevisiae*). Protein lysates were prepared from cultures which were either induced or uninduced with IPTG, and a Western blot was prepared which is shown. Fragments were detected using anti β-galactosidase polyclonal rabbit antiserum and enhanced chemiluminescence. Lanes 1,2 -XL1-blue, lanes 3, 4 -Ub-M-βgal + pQE30, lanes 5, 6 -Ub-M-βgal + pQE-USP21, lanes 7, 8- Ub-M-βgal + UBP 2. Lanes 9 and 10 show a long exposure of lanes 7 and 8. Lanes 5 and 6 show a specific fragment (marked ∇), which has a higher mobility compared with the fragment in lane 4 (marked ∗) and is comparable to the fragment in lane 9. This is due to *USP21* (lanes 5 and 6) and *UBP2* (lanes 9 and 10) cleaving the fusion protein. The resulting M-βgal product is approximately 8 kDa smaller than the Ub-M-βgal fusion product. This shows that *USP21* can cleave the fusion product *in vitro* and can function as a ubiquitin-specific protease. The other cross-hybridising fragments are non-specific as can be seen from lanes 1 and 2 (XL1-blue). The molecular weight marker is indicated on the left hand side in kDa.

samples uninduced and induced with IPTG, respectively. Lanes 1 and 2, which contain proteins from the *E. coli* host, show cross-hybridising protein products. A specific product can be seen in lane 4, corresponding to the Ub-M-βgal fusion protein (indicated with a star). This lane contains proteins from *E. coli* which co-expressed the Ub-M-βgal fusion gene with pQE30. In lanes 5 and 6, which contain proteins from *E. coli* and which express the Ub-M-βgal and pQE-USP21 genes, a product of slightly greater mobility can be seen (indicated via a triangle). This is due to cleavage of the 8 kDa ubiquitin part from the Ub-M-βgal fusion protein. This same product can be seen in lane 9, which contains proteins from *E. coli* containing Ub-M-βgal and UBP2 (positive control). The Western blot shows that USP21 is able to cleave ubiquitin from a Ub-M-βgal fusion product *in vitro* and that this gene can therefore be classed as a USP.

5.3.3 Identification of the smallest 21q11 sub-region of overlapping deletion, containing USP21, in non-small cell lung carcinoma patients

Figure 5.7 compares work that has been performed by Kohno and co-workers and our group (J. H. Ives, T. A. Jones and D. Nizetic). Kohno and co-workers have shown allelic loss in patients with NSCLC of between 42 and 62.5% with the markers D21S1256/D21S1433 (21q21.1) and D21S120 (21q11.2) respectively (Kohno *et al.*, 1998). They also reported a NSCLC cell line containing a homozygous deletion from D21S318 (21q11.2) to D21S18 (21q21.1). These markers are located centromeric and telomeric respectively, to the indicated region of Figure 5.7. This deletion is approximately 4.5 Mbp. A study of the polymorphic markers D21S120, D21S258, D21S172 and D21S1256 in the region showed loss of heterozygosity (LOH) in 13 out of 34 informative cases (38%) of squamous NSCLC with one or more markers (J.H. Ives). DNA from tumour material from patient #47 showed that three markers showed LOH, but D21S1256 did not. Tumour DNA from patient #61 showed LOH at D21S1256 but not at D21S172. This suggested a smallest region of overlapping deletion (SRO). Fluorescent *in situ* hybridisation (FISH) of the PAC clones 90B5, and 126N20, and BAC clone B39I12 to slides containing paraffin embedded sections from tumour #61 showed two signals in 71% of 301 interphase nuclei and only one signal in 29% (T.A. Jones). This is probably due to the thinness of the slice (4 micrometer); some nuclei may have been cut in half (T.A. Jones) thus enabling the observation of only one signal instead of two. When PAC clones 73M5 and 135E14 were used as FISH probes

Identification of a smallest region of overlapping deletion in non-small cell lung carcinoma samples

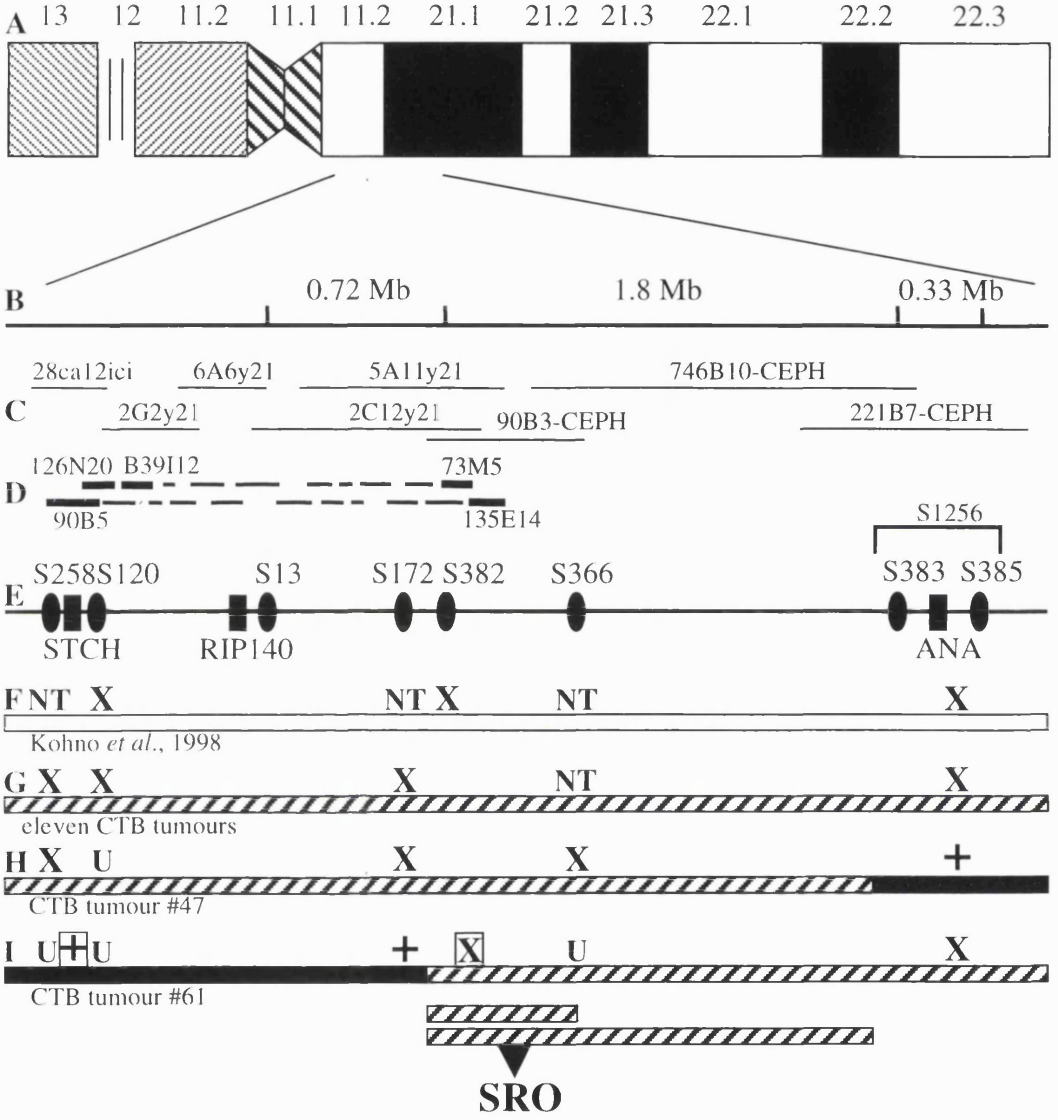


Figure 5.7: A-Cytogenetic map of chromosome 21. B-*Not* I map of the indicated chromosomal region (Ichikawa *et al.*, 1993). C-Overlapping YAC map of this region (Chumakov *et al.*, 1992, Nizetic *et al.*, 1994, Korenberg *et al.*, 1995). D-Overlapping PACs, BACs and cosmids (Chapter 3; Groet *et al.*, 1998); the indicated clones were used for FISH to tumour #61. E-Location of markers (ovals) and genes (squares). F-Data from Kohno *et al.*, 1998; the markers S120 and S1256 showed LOH between 42 and 62.5% in patients with squamous NSCLC, the whole region was homozygously deleted in a NSCLC cell line. G-Eleven CTB tumours that showed LOH with the markers S258, S120, S172 and S1256. H-Tumour sample #47; LOH was detected in S258, S172 and S366, S1256 was present in two copies. I-CTB tumour #61, S1256 was detected in one copy, S172 was present in two copies. FISH of the clones 90B5, 126N20 and B39I12 shows predominantly two signals, in this tumour sample. The PAC clones 73M5 and 135E14 show predominantly single signals in FISH analysis of tumour sample #61. From #47 and #61 can be deduced that the SRO spans from between S172/73M5 to between S366/S1256. **X**-LOH, **+**-no LOH, **NT**- not tested, **U**-uninformative, [X]-FISH, showing predominantly two signals, [+]-FISH, showing predominantly one signal.

to sections from the same tumour (#61), predominantly single signals were observed (61% of 294 counted nuclei, T.A. Jones). The amount of single signals could even be higher, as nuclei that originally contained one chromosome could lose this signal and could therefore not be counted, because the nucleus had been cut in half. Additionally, a small percentage of non-tumour cells will also be present in the sections, which can explain why the cells show two signals.

In order to analyse the NSCLC samples #47 and #61 further, the polymorphic marker D21S366 (Bosch *et al.*, 1996) was used, which maps between D21S172 and D21S1256. This marker is located on the YACs 90B3-CEPH and 746B10-CEPH (Bosch *et al.*, 1996); 90B3-CEPH overlaps with the distal part of the contig, containing marker D21S382 (presented in Chapter 3) (Korenberg *et al.*, 1995). The YAC 746B10-CEPH contains the marker D21S383 (Gardiner *et al.*, 1995., Korenberg *et al.*, 1995), positioning D21S366 between D21S172 and D21S1256, as shown in Figure 5.7. LOH was assessed for this marker using the tumour samples #47 and #61. The primers D21S366-F and fluorescently labelled 6-FAM-D21S366-R were used in PCR amplification, and products were generated from tumour and normal tissue from patients #47 and #61 (Chapter 2.21.3). Data were analysed using Genescan analysis software (PE applied Biosystems). Ratios of the peak heights from both alleles from normal tissue divided by those from the tumour DNA of sample #47 resulted in a value of 1.45. The ratio required to confirm LOH is either less than 0.5 or greater than 1.5 (Microsatellite Analysis 8, PE Applied Biosystems). Due to the fact that these values are close and a percentage of DNA from normal cells was present in the original tumour sample, it may be concluded that the marker D21S366 is heterozygously deleted in DNA from tumour #47. The marker D21S366 was uninformative in sample #61. This result narrows the distal end of the SRO down and places it between D21S366 and D21S1256. The proximal end is located between D21S172 and D21S382 which would allow a size estimation for the SRO of between 250 Kbp and 2.5 Mbp (Figure 5.7).

The hemizygous deletion of the genomic region corresponding to the PAC clones 73M5 and 135E14, as established by FISH, was studied in more detail. The sequence of USP21 was used as a tool, since the gene was found to be contained within these two PACs. The patient corresponding to #61 showed a single silent nucleotide polymorphism at nucleotide 2029 of the USP21 sequence, the same polymorphism that was cloned in pQE-USP21. Cycle sequencing was performed using the primer His 1 of the PCR products from genomic DNA from normal and tumour tissue of patient #61.

These were generated with the primers His 1 and His 2 (these primers amplify part of the USP21 gene: nucleotides 1964-2077). This PCR product contains the A/T polymorphism at nucleotide 2029. Peak heights of the A and T signal in nucleotide position 2029 were compared between normal and tumour tissue. Table 5.2 shows the relative peak height of A in normal tissue to be 1.81 times as high as in the tumour sample, suggesting LOH. This was further analysed by gel electrophoresis and Southern blotting of the PCR products generated by the primers His 1 and His 2 from normal and tumour tissue from patient #61. Two identical blots were prepared which were hybridised separately to radio-actively labelled 11-mer oligonucleotide probes. These probes were identical to the nucleotides 2025-2035, with either the A or the T at position 2029 (primer 2029-A and 2029-T, respectively). Blots were exposed to a phosphor-imager screen and the intensities of the different bands on the blot were analysed using spotfinder software (Image QuaNT^{TM}, Molecular Dynamics). Values for the different intensities of the A and T oligonucleotide probes hybridising to normal and tumour PCR products can be seen in Table 5.2. These values have been normalised by subsequent hybridisation of radio-actively labelled His 1 oligonucleotide to the same blots. The ratio of A/T of normal DNA is 1.67 times higher then the same ratio in tumour DNA, again suggesting LOH. PCR products with either an A or a T at position 2029 were included in both blots (Chapter 2.21.4) to measure cross-reactivity, which was around 17% for both oligonucleotides.

5.3.4 Sequencing of the USP21 gene in genomic DNA from patients with non-small cell lung carcinoma and sporadic Alzheimer's disease.

In order to establish whether the USP21 gene can function as a classical (two hit) tumour supressor gene, mutation detection was carried out using DNA of patients with NSCLC that exhibits heterozygous deletions of the region. This included sequencing of the functionally important domains, to detect missense or nonsense mutations, and sequencing of CGA codons which could be converted into TGA stop codons. additionally 23 out of the 24 exons were sequenced from genomic DNA of two patients. The Cys, QQD and His boxes of USP21 were sequenced using genomic DNA derived from normal and tumour DNA from 13 patients with squamous NSCLC, that showed LOH for the markers D21S120, D21S258, D21S172 or D21S1256 (see Figure 5.7). The Cys, QQD and His boxes show a high degree of similarity between different USP proteins and are important functional domains (Wilkinson 1997; Figure 5.4). They

LOH determined via a polymorphism of USP21

Sequencing	A (cm)	T (cm)	A/T	#61N A/T #61T A/T
#61N	13.4	23.4	0.573	1.81
#61T	7.5	23.7	0.316	
Oligo hybridisation	A (counts)	T (counts)	A/T	#61N A/T #61T A/T
#61N	30719	32818	0.936	1.67
#61T	19802	35283	0.561	

Table 5.2: Results confirming the LOH as indicated by FISH with the PAC clones 73M5 and 135E14 in the lung tumour sample #61. The top half of the table shows peak heights of sequencing of normal (N) and tumour (T) DNA at nucleotide position 2029 of the USP21 gene. The ratio of A/T of normal versus tumour is 1.81:1, suggesting LOH. The bottom half of the table shows normalised values for the intensity of hybridisation of an 11-mer oligonucleotide, hybridised to a blot containing PCR products from normal and tumour DNA from patient #61, generated with the primers His 1 and His 2. The 11 mer oligonucleotides are identical to the nucleotides 2025-2035 of USP21 and contain either an A or a T at position 5, which corresponds to nucleotide position 2029 of USP21. The ratio of A/T of normal versus tumour is 1.67:1, again indicating LOH. The results were normalised by subsequent hybridisation of the His 1 primer.

could therefore be altered in tumour material from patients with squamous NSCLC. The patients that showed LOH of the region were; #5, #13, #16, #47, #61, #99, #136, #166, #184, #225, #227, #231, and #247. Specific PCR fragments were generated using the primers USPE-5F and USPE-5R (amplifying nucleotides 619-754, which contain the Cys box), QQD 1 and QQD 2 (amplifying nucleotides 980-1027, which contain the QQD box except for the first amino acid of this motif, which is located in the previous exon) and His 1 and His 2 (amplifying nucleotides 1964-2077, which contain the His box). Cycle sequencing was performed using these PCR products with the primers USPE-5F, QQD 1 and His 1 respectively. The sequences of the different boxes were identical between normal and tumour DNA and corresponded to the nucleotide sequences indicated in Figure 5.3. The only difference compared with the original USP21 sequence was detected in patients #16, #61 and #231, where a heterozygous A and T were found at nucleotide position 2029 of the USP21 gene (see also section 5.3.2 and 5.3.3). This natural variant is silent, GCA and GCT codons both code for arginine. The other samples all show a homozygous A at this position.

Sequences around CGA codons in USP21 were also determined. These codons are known to be the most frequent sites of causative point mutations in tumour suppressors due to de-amination of CpG cytosines, converting them into thymines and creating a TGA stop codon (Mancini *et al.*, 1997). The CGA codons that were sequenced are located at nucleotides 830, 1532, 1922, 2282 of the USP21 sequence (Figure 5.3). Specific PCR products were generated from genomic tumour DNA from samples #5, #13, #16, #47, #61, #99, #136, #166, #184, #225, #227, #231, and #247 using the primers CGA830-F and –R (contains CGA830), CGA1532-F and –R (contains CGA1532), CGA1922-R and USP-5F (contains CGA1922), and CGA2282-F and –R (contains CGA2282). Cycle sequencing was carried out using the –R primer of the four corresponding CGA-containing PCR products. The following nucleotides were sequenced, referring to the position in the sequence as indicated in Figure 5.3: 755-841 (CGA830), 1504-1665 (CGA1532), 1922-1963 (CGA1922), and 2193-2349 (CGA2283). No differences were detected compared with the sequence in Figure 5.3. Other CGA codons are present at nucleotide 665 and 1973 of USP21. These were sequenced with the Cys and His box respectively, and no mutations were detected. The CGA codons at positions 3044, 3161, 3214, 3223 and 3344, towards the 3' end of the coding region, were not sequenced.

In two squamous NSCLC tumour samples that showed LOH in the region, the whole gene was sequenced from genomic DNA, except for exon 1 (Chapter 2.21.5). The two tumour samples were from patients #5 and #13. The primers that were used to generate the PCR products from genomic DNA were USPE-2F/2R, 3F/R, 4F/R, 5F/R, CGA830F/R, 7F/R, QQD1/8R, 9F/9R, 10F/R, 11F/R, 12F/R, CGA1532F/R, 14F/R, 15F/CGA1922R, His 1/16R, CGA2282F/17R, 18F/R, 19F/R, 20F/R, 21F/R, 22F/R, 23F/R, 24F/R, and 24F/R2. The numbers of the primers refer to the different exons as indicated in Figure 5.2 and are located in the intronic sequence adjacent to the exons. The primers USPE-5F/R, CGA830F/R and CGA1532F/R amplify the exons 5, 6 and 13 respectively and these sequences were already analysed (see previous paragraph). The following primers were used for the cycle sequencing of their respective PCR products: USPE-2R, 3F, 4F, 5F, CGA830R, 7R, 8R, 9R, 10F, 11F, 12R, CGA1532R, 14F, 15F, 16R, CGA2282F, 18R, 19R, 20R, 21F, 22F, 23F, 24F, 24R and 24R2. The nucleotide sequence of USP21, which was determined via sequencing using the different primers, can be found in Chapter 2.21.5. Exons and splice sites adjacent to the exons (see Table 5.1) were compared with the genomic sequence. No mutations were found in any of the exons and splice junctions that were sequenced. Two polymorphisms were found, one was a G/A present at nucleotide position 3455 of the USP21 gene, in the tumour sample of patient #13. This variant was also present in the normal tissue from the same patient and was located in the 3'UTR. An A is present at this position in the gene sequence which was derived from the genomic clone 135E14 (Figure 5.3). The chromosome containing the G residue at position 3455 is likely to have been lost from tumour material, from patient #13, this was established by comparing peak heights of the A and G residues between normal and tumour DNA from this patient. The other polymorphism was found just adjacent to exon 14 in tumour DNA from patient #5. This sequence is **g/a**atacagCACAAC, where the lower case indicates intronic sequence. The sequence from the PAC clone 135E14 (AJ010598) indicates a G at nucleotide position 22,265. The same polymorphism is found in DNA from normal tissue from the same patient. The chromosome containing the G residue is again likely to have been lost from the tumour DNA (again established by comparing peak heights of both residues in the sequence of normal and tumour DNA). No fragments could be amplified from genomic DNA that included exon 1. This is probably due to the high GC content of the 5' UTR (84%) and the sequence upstream from the gene. This region contains the CpG island around the marker D21S382.

Ubiquitin has been found in the senile plaques and neurofibrillary tangles of patients with AD (Mori *et al.*, 1987; Perry *et al.*, 1987; Morishima-Kawashima *et al.*, 1993). Furthermore, positive lod scores were found in the 21q11 region after linkage analysis of affected sib pairs with LOAD (Kehoe *et al.*, 1999). These findings could mean that USP21 might be involved in AD. For this reason, mutation detection was carried out in LOAD patients, by sequencing of the Cys, QQD and His boxes using thirteen late onset Alzheimer's disease samples (Chapter 2.17). The primers that were used for the analysis were identical to those described for sequencing of the NSCLC samples. No mutations were detected; an A/T heterozygote at nucleotide position 2029 was detected in patients 12 and 14. The APOE alleles that these samples contain are unknown, except for patients 6 and 13 which are APOE-ε4 negative (see next section).

Genomic DNA from two LOAD samples (LOAD 6 and 13, Chapter 2.17) was used to sequence exons 2-24 of USP21. The primers that were used and the method of analysis was carried out was identical to the NSCLC samples #5 and #13 (previous section). The LOAD samples were analysed for APOE-ε4 allele content. This was established by PCR amplification using genomic DNA from the two samples and the primers APOE-F and APOE-R (Chapter 2.21.5). These primers amplify the nucleotides 3786-4035 of the APOE-ε4 gene (accession nr. M10065), which contain nucleotide 3932 of this allele. This nucleotide is a C residue in APOE-ε4, and a T residue in APOE-ε3 and ε2, which causes a Cys112>Arg substitution in the ε4 allele. Cycle sequencing was carried out with both primers individually. Both samples showed a homozygous T residue at position 3932 and therefore did not contain the ε4 allele. No mutations were detected in any of the 23 exons and adjacent splice sites in the two LOAD samples. An example of the quality of the sequence can be seen in Figure 6.7.

5.4 Discussion

The full length sequence of a novel gene located on the PAC clones 73M5 and 135E14 was determined after initial identification via exon trapping (Figure 4.14, 4.15, 4.17). This was achieved by comparison of overlapping cDNA clones and exon trapped material with the genomic sequence, which was contained in the bacterial clones 73M5 and 135E14. The order and the presence of the twenty-four exons in the mRNA was re-confirmed by RT-PCR of the coding sequence from bone marrow RNA (Chapter 5.3.2). Northern blots revealed a 4.1 Kbp fragment in all tissues tested. A less intensely

hybridising fragment of 4.9 Kbp was also observed in all tissues tested, which could be explained by alternative splicing, or an alternative 3' end. A cDNA clone that extends significantly at the 3' end of USP21 has been identified in the human EST database (Image clone 150542, Figure 4.17).

The total length of the gene is 3803 bp, it consists of twenty-four exons, occupies a genomic distance of approximately 149 Kbp and contains a 3165 bp, 1055 amino acid ORF. The 5' end of the gene is located at the CpG island D21S382/LL56 (Ichikawa *et al.*, 1993). The first in frame ATG start codon is located at nucleotide position 200 (see Figure 5.3). Although the start codon does not show the most optimal context for initiation of translation, which is GCCACCatgG (Kozak, 1996), it does contain a purine at nucleotide –3 (Figure 5.3), which makes it an adequate initiation codon (Kozak, 1996). Furthermore, no downstream ATG codons are present. At the 3' end AATAAA polyadenylation signals are present 33 and 26 bp before the 3' end (see Figure 5.3). The 3' end was originally determined from cDNA clones from the A002B43 cluster (see Figure 4.17, AA526306, AA479313, T19119, AA747586, T36007 and N92952). The sequence similarity between these and three cDNA clones (AI093898, AI276958 and AI308246) that are not shown in Figure 4.17 all stop within 15 bp of their 3' end. The analysis of two 3' RACE clones confirms this (Chapter 5.3.1).

5' RACE was also performed, as described for the 3' RACE (section 5.3.1). Four 5' RACE clones were sequenced and all four were identical to the USP21 sequence (not shown). No USP21 sequence was found 5' from nucleotide 197, which is probably due to the high GC content of the 5' region (84%). Comparison of the USP21 protein with amino acid translations of mouse ESTs shows that the thirty-six amino acids of USP21 closest to the amino-terminal end are identical in mouse. Amino acid translations of the nucleotide sequence upstream of the start codon do not show a significant similarity, suggesting that the ATG at nucleotide 200 is the start codon in USP21 (not shown).

Although the sequence of USP21 was readily determined via overlapping cDNA clones and exon trapped material, the comparison of the exon prediction programmes with the twenty-four exons can give some indication about the efficiency of the prediction of other genes in the area. The Genbank report of the PACs 73M5 and 135E14 (accession nrs. AJ010597, AJ010598 resp.) was analysed for exon predictions and compared with the location of the USP21 exons. For both PACs three exon prediction programmes were used; Grail, MZEF and Xpound for 73M5 and Grail, MZEF and Genscan for 135E14 (Uberbacher and Mural, 1991; Kamb *et al.*, 1995; Burge and Karlin, 1997;

Exon predictions in the genomic interval of the USP21 gene

	Number of exons (total nr. of exons in USP21)	Exons predicted by number of exon prediction programs		
		3	2	1
USP21 exons rightly predicted	20 (24)	4	10	6
USP21 exons wrongly predicted	62	1	2	59

Table 5.3: The number of USP21 exons that are rightly predicted are shown (20 out of 24). The first 9 exons were predicted from the PAC 73M5 (AJ010597) using Grail, MZEF and Xpound; the other 15 from the PAC 135E14 (AJ010598) with Grail, MZEF and Genscan. Out of the 20 rightly predicted exons, four were predicted by three exon prediction programs, ten by two programmes and six by one programme. In the genomic interval where USP21 is located, 62 exons were predicted wrongly. Only one of these was predicted by all three prediction programmes, two by two programmes and 59 by one programme.

Zhang, 1997). Twenty of the twenty-four exons were correctly predicted (the exons 1, 5, 6 and 24 were not predicted). Sixty-two exons were wrongly predicted in the interval where USP21 is located, showing that both false positives and false negatives were anticipated. Table 5.3 shows that the majority of the exons predicted by two or more prediction programmes to be part of USP21; fourteen of the seventeen exons (82%). This shows that, if a putative exon in this region is predicted by more exon prediction programmes, then there is a greater chance that the exon is part of the gene. This analysis did not take into account the occurrence of genes within genes, or the fact that wrongly predicted exons could be alternatively spliced exons.

The coding region of USP21 was cloned into a prokaryotic expression vector and co-expressed with a ubiquitin-β-galactosidase fusion gene. This showed that *USP21* can cleave the ubiquitin from β-galactosidase in the fusion protein (section 5.3.2). From this study in *E. coli*, and the fact that this novel gene shows similarity especially in the Cys and His boxes (Wilkinson, 1997), it can be concluded that the novel gene codes for a *USP*. These proteins are involved in the ubiquitin proteasome pathway, in which proteins are being degraded physiologically (see Schwartz and Ciechanover, 1999 for a recent review). In this pathway, ubiquitin, a 76 amino acid protein, is activated by the ubiquitin activating enzyme. The activated ubiquitin is then transferred to a ubiquitin-conjugating protein, whose function is to associate the activated ubiquitin with a ubiquitin-protein ligase. The role of these ligases is to bind directly, or via a protein complex, to a protein that needs to be degraded and to promote the transfer of ubiquitin. The ubiquitin molecule is attached via the glycine at amino acid position 76 via an ε-amino bond to a lysine residue of the target protein. Additional ubiquitin molecules can then be attached to the lysine at position 48 of the previous ubiquitin moiety. The protein which contains the poly-ubiquitin chain is the appropriate substrate for degradation by the 26S proteasome complex. Ubiquitin carboxyl-terminal hydrolases are a group of enzymes that can remove small peptide fragments from the proximal end of the poly-ubiquitin chain, after proteasome degradation. The USPs can then remove one ubiquitin molecule at a time from the poly-ubiquitin chain, starting at the proximal end, to regenerate ubiquitin molecules and also de-ubiquitinate the substrate, preventing its degradation (reviewed by Wilkinson, 1997). Another function of the USPs is to generate ubiquitin from precursor proteins, which are the protein products transcribed and translated as ubiquitin-fusion proteins. Also, a USP linked to the 26S proteasome

has been described (Lam *et al.*, 1997), which can remove ubiquitin molecules from the distal end. This could perhaps function as a proof-reading enzyme by removing poly-ubiquitin chains that are too short. Proteins with long poly-ubiquitin chains would be hydrolysed before the chain could be removed. Only proteins that are completely ubiquitinated would be degraded (Wilkinson, 1997). It has been shown that USPs are involved in different biological processes, for example development, growth and oncogenesis, regulation of DNA and chromosome structure (reviewed by Wilkinson, 1997). Examples of these include the fat facets gene involved in eye development in *Drosophila melanogaster* (Huang *et al.*, 1995), the human tre-2 gene which has been shown to function as a dominant oncogene (Nakamura *et al.*, 1992; Papa and Hochstrasser, 1993) and human Ubp-M which is thought to play a role in the regulation of mitotic chromatin (Cai *et al.*, 1999).

In order to determine if the USP21 gene can act as a classical tumour suppressor gene (both copies of the gene disrupted) in NSCLC, mutation detection via sequencing was carried out using tumour material from patients with squamous NSCLC. In these patients hemizygous deletions were detected in the region in which USP21 is located. In this study, no mutations were found in twenty-three out of twenty-four exons sequenced from two patients. Causative mutations could be present in the first exon, which was not sequenced due to a high G/C content, or in non-coding DNA. Although the USPs contain a core catalytic domain of approximately 450 amino acids, the amino and carboxyl terminal extensions could be involved for example in substrate specificity (Wilkinson, 1997). It would be interesting to determine if the protein product of the USP21 gene could interact with other proteins by performing yeast two hybrid screens.

Markers, located in a large region of the long arm of chromosome 21, showed positive lod scores in patients with LOAD who were APOE-ε4 negative (Kehoe *et al.*, 1999). Also, ubiquitin was found in the senile plaques and neurofibrillary tangles in AD (Mori *et al.*, 1987; Perry *et al.*, 1987; Morishima-kawashima *et al.*, 1993). The possible involvement of USP21 was therefore tested by mutation detection of patients with LOAD. Sequencing of USP21 using genomic DNA from two patients with LOAD showed no mutations. Again, it cannot be ruled out that a mutation might be present in the first exon or the non-coding DNA from USP21. Mutation detection using more patients with LOAD that do not contain the APOE-ε4 allele should be carried out to draw firm conclusions about a possible involvement of USP21 in this phenotype.

We have cloned and described a novel gene, mapping just on the telomeric side of the marker D21S382, which is located in a smallest region of overlapping deletion in squamous NSCLC. Similarity with known proteins predicted this gene to code for a ubiquitin-specific protease. This function was shown by an assay in which the protein product of the novel gene cleaved the ubiquitin moiety from a ubiquitin-β-galactosidase fusion protein in *E. coli.* Studies that determined the minimal region of LOH were described (J.H. Ives, T.A. Jones) and showed that this gene was located within this region. Sequencing of the conserved domains and the 'CGA' codons of the gene in thirteen patients with squamous NSCLC, and 94% of the whole gene in two of those patients failed to identify mutations. No mutations were detected in sequencing important functional domains using DNA from thirteen patients with LOAD. Additionally, 94% of the whole gene was sequenced in two of the thirteen patients with LOAD and again, no mutations were detected.

Chapter 6 Identification and cloning of a novel gene, G31B5, located in Human chromosome 21q11.

6.1 Summary

The exon trapping experiments from Chapter 4 identified a putative exon (31B5-E21) which was shown to be identical to part of the Image cDNA clone 1286563. This cDNA clone hybridised additionally to the putative exon 31B5-A3/39I12-F7, resulting in a cluster of two putative exons and five cDNA clones (Figure 4.11). Rapid amplification of cDNA ends (RACE) was performed from the exon trapped clone 31B5-E21, to generate clones which were specific for the 3' and 5' ends of the novel gene. The sequenced RACE products, cDNA's, and the two exon trapped clones were assembled into one contiguous sequence and compared with the genomic sequence, which was determined from the PAC clone 31B5 (accession nr. AF130247). Part of the novel gene was located on this PAC clone. The total length of the gene is 1888 bp (373 amino acid open reading frame (ORF)), it contains eight exons and is known as G31B5. The full length gene is in good agreement with the main 2.1 Kbp fragment detected by the hybridisation of a radio-actively labelled exon of this gene to a multiple tissue Northern blot. A mouse homologue is described from overlapping cDNA clones (1335 bp 364 amino acids), which shows a high degree of similarity with the human gene at the amino acid level. The coding region, present in eight exons, and the adjacent splice sites were sequenced in genomic DNA from patients with acute non-lymphocytic leukaemia M7 (ANLL-M7), familial early onset Alzheimer's disease (FEOAD), late onset Alzheimer's disease (LOAD) and non-small cell lung carcinoma (NSCLC). No mutations were found, however two different polymorphisms were discovered.

6.2 Introduction

6.2.1 Examples of genetic defects involving chromosome 21

Chromosome 21 is the smallest of the human autosomes and is approximately 40-50 Mbp in length. When present in three copies it causes Down syndrome (DS), which can be characterised by more than twenty phenotypic features, which may not all be present in each individual with DS. The most common phenotypes in DS are hypotonia, mental retardation and an Alzheimer's disease-like phenotype in the third and fourth decade of life (Epstein *et al.*, 1991; Korenberg *et al.*, 1994; Korenberg, 1995; Kesslak *et al.*,

1994). A number of genetic defects have been identified in karyotypically normal people, in genes that map to chromosome 21. Some examples are: SOD 1 in amyotrophic lateral sclerosis (Rosen *et al.*, 1993), APP in familial early onset Alzheimer's disease (Goate *et al.*, 1991), the EPM1 gene in progressive myoclonus epilepsy (Pennacchio *et al.*, 1996; Lalioti *et al.*, 1997), and autoimmune-polyendocrinopathy-candidiasis-ectodermal dystrophy (The Finnish-German APECED Consortium, 1997; Nagamine *et al.*, 1997).

6.2.2 Genetic defects associated with chromosome 21q11

The cytogenetic band q11 of human chromosome 21, which flanks the centromere, may contain genes that are involved in a number of phenotypes.

An increase in leukaemia, in particular of ANLL-M7 has been reported in DS infants (Zipurski *et al.*, 1992). Also a transient abnormal myelopoiesis (TAM, a self-regressing leukaemic reaction), which might be linked to ANLL-M7, has been observed in neonates and small children with DS (Zipurski *et al.*, 1987; 1992). It has been hypothesised that the 21q11 region may contain a gene or genes involved in ANLL-M7 and TAM, as disomic homozygosity of the polymorphic markers in the area was observed in individuals with these disorders (Abe *et al.*, 1989; Niikawa *et al.*, 1991; Shen *et al.*, 1995; Cavani *et al.*, 1998).

Linkage has been shown in some families with FEOAD and the 21q11 region, in particular with the marker D21S16 (St. George-Hyslop *et al.*, 1987; Goate *et al.*, 1989). Causative mutations have since been found in the presenilin-1 gene in the same families (Sherrington *et al.*, 1995). However it has been hypothesised that a modifier gene might be present in 21q11 involved in the generation of the phenotype in these patients (St George-Hyslop *et al.*, 1992; Van Broeckhoven, 1995). Linkage analysis of affected sib pairs with LOAD, detected positive lod scores in a large region of the proximal long arm of chromosome 21 (including the 21q11 region) in the APOE-ε4 negative subgroup. The lod score peaked around the APP gene (Kehoe *et al.*, 1999). The authors have proposed a second stage study, in which more sibling pairs and additional markers will be used to confirm lod scores and to narrow down regions in which causative genes may be located.

Frequent deletions have been detected in the long arm of chromosome 21 in squamous NSCLC (Testa *et al.*, 1994; Sato *et al.*, 1994). A study into the loss of heterozygosity

(LOH) of the long arm of chromosome 21 in this disease, found the 21q11 marker D21S120 to be the most frequently deleted (Kohno *et al.*, 1998).

6.2.3 Physical maps that include the 21q11 region

A number of physical maps have been reported in the 21q11 region. These include long range restriction maps (Van Hul and Van Broeckhoven, 1993) and a *Not* I map of the long arm of chromosome 21, which showed two *Not* I sites in this region (Ichikawa *et al.*, 1993). Physical maps have also been constructed which consisted of overlapping 21q11 genomic inserts as Yeast Artificial Chromosomes (Butler *et al.*, 1992; Chumakov *et al.*, 1992; Nizetic *et al.*, 1994). Work illustrated in this thesis has described how a physical map was constructed containing overlapping segments of this region in bacterial clones (Chapter 3; Groet *et al.*, 1998). Sequencing of these overlapping clones will determine the entire sequence of this segment of chromosome 21 (see for instance http://www-eri.uchsc.edu/chr21/c21index.html for progress).

6.2.4 Prediction of the gene density of chromosome 21q11

The proximal half of the long arm of chromosome 21, which includes 21q11, is predicted to be very gene-poor (reviewed by Gardiner, 1996). This was evident from:

- The lack of *Not* I sites (Ichikawa *et al.*, 1993), which cluster in CpG islands that are often associated with the 5' ends of housekeeping genes (Lindsay and Bird, 1987).
- The number of cDNA selected clones that mapped back to the proximal region of chromosome 21q after cDNA selection from random cosmids was much lower, compared with the distal region (reviewed by Gardiner, 1996).
- The number of exon trapped clones that mapped back to this region was also small, as was determined from preliminary, global exon trapping approaches (Yaspo *et al.*, 1995; Chen *et al.*, 1996).
- The region was shown to be rich in A and T residues (Gardiner *et al.*, 1990; De Sario *et al.*, 1997), which is associated with gene-poor areas of the genome (Bernardi, 1995).
- Genes have been identified via their biological function mapped predominantly to the distal half of the chromosome (Gardiner, 1996).

All these observations indicate that the cytogenetic band 21q11 is likely to be a gene-poor region.

6.2.5 Genes which map to 21q11

Prior to this study, only two genes were reported to be located in 21q11, STCH and RIP140 (Brodsky *et al.*, 1995; Cavailles *et al.*, 1995; Katsanis *et al.*, 1998). STCH is a member of the stress70 chaperone family that participate in protein processing by binding to denatured or misfolded proteins and releasing them by an ATP-dependent mechanism (Otterson *et al.*, 1994). RIP140 is a nuclear protein that interacts *in vitro* with the AF2 domain of the oestrogen receptor, which is involved in activation of transcription (Cavailles *et al.*, 1995). In order to identify novel genes in the 21q11 region, that could be tested for the different phenotypes that are linked to this region, exon trapping was performed on twenty-two overlapping genomic clones (Chapter 4). This resulted in the identification of a novel full length gene USP21 located near the marker D21S382; this novel gene is a member of the ubiquitin-specific proteases (Chapter 5).

In this Chapter a second full length novel gene is described. The gene, which mapped back to the bacterial clones 31B5 and B39I12, was identified from a cluster of exon trapped clones and cDNAs. RACE was performed to generate 3' and 5' products. The exon trapped clones, cDNAs and RACE products were assembled and the sequence was compared with the two genomic clones 31B5 and B39I12, which became available during this identification of this novel gene. The gene was shown to consist of eight exons. Mutation detection was performed by sequencing of the coding region of the gene using genomic DNA from DS infants with ANLL-M7, genomic tumour DNA from squamous NSCLC with LOH of the region, DNA from lymphoblast cell lines, which were generated from patients with FEOAD and from genomic DNA from LOAD patients.

6.3 Results

6.3.1 Identification of a novel gene mapping to bacterial clones 31B5 and B39I12.

The two exon trapped clones 31B5-E21 and 31B5-A3/B39I12-F7 overlapped the Image cDNA clone 1286563, as shown in Appendix I Table 2 and Figure 4.11. These exon trapped and cDNA clones form a cluster with the cDNA clones 136169, 360025, AA471280 and AA361754 (Figure 4.11). The insert from exon trapped clone 31B5-E21 was specifically amplified by PCR (using the primers EX1 and EX2, see Chapter

Hybridisation of the exon trapped product 31B5-E21 to a multiple tissue Northern blot

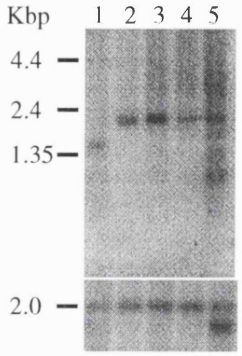


Figure 6.1: Hybridisation of radio-actively labelled PCR product of the exon trapped product 31B5-E21 to a multiple tissue Northern blot (Human I- Clontech). Each lane contains approximately 2 µg Poly A+ RNA. Lane 1 -liver, lane 2 -lung, lane 3 -placenta, lane 4- brain, lane 5- heart. A 2.1 Kbp fragment can be seen in lanes 2, 3, 4 and 5. Lane 1 contains a 1.5 Kbp fragment and lane 5 a fragment of 700 bp. The lower panel shows hybridisation of the same membrane with a β-actin probe (Clontech).

Identification of 3' and 5' RACE clones from the exon trapped clone 31B5-E21

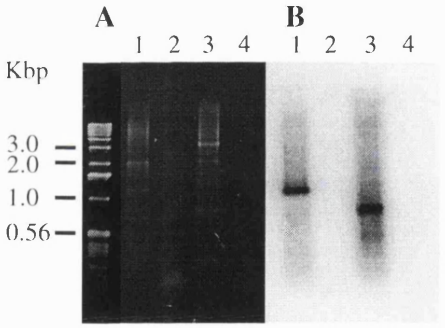


Figure 6.2: A-shows a photograph of a 1.2% w/v agarose gel, which contains RACE products generated via sequence information from the exon trapped product 31B5 -E21. The initial PCRs were performed using Marathon-Ready cDNA (Clontech) and the primers AP1 (linker primer) and 31B5-E21R or -E21F (gene-specific primer). This was followed by a PCR with AP2 primers (nested) and the same gene-specific primers. These last PCR products are shown in this figure. Lanes 1 & 2- PCRs performed with the primers AP2 and 31B5-E21R, lanes 3&4- as lanes 1 & 2 with the primers AP2 and 31B5-E21F, lanes 1&3- 1 µl of a 1:100 dilution of the initial PCR was added to the reaction, lanes 2&4- negative control. B-shows hybridisation of the Southern blot generated from the gel depicted in A, using a radio-actively labelled PCR product from the exon trapped clone 31B5-E21.

2.20.1), and was subsequently isolated from agarose gel via Prep-A-Gene. This PCR product was radio-actively labelled and hybridised to a multiple tissue Northern blot (Human I, Clontech). The result can be seen in Figure 6.1. This hybridisation shows an mRNA fragment of approximately 2.1 Kbp in lung, placenta, brain and heart, a fragment of 1.5 Kbp in liver and an additional 700 bp fragment in heart.

The primers 31B5-E21F and 31B5-E21R were designed for RACE, from the exon trapped clone 31B5-E21 (see Chapter 2.18.12). RACE was performed using Marathon-Ready cDNA, which was generated from bone marrow poly A+ to which linkers were ligated (Clontech). An initial PCR reaction was carried out using the linker primer AP1 and either the gene-specific primer 31B5-E21R or 31B5-E21F. This was followed by a subsequent PCR using the nested linker primer AP2 and the same gene-specific primers. A total of 5 µl out of 25 µl from these PCR reactions was used in electrophoresis on a 1.2% w/v agarose gel. A photograph of the gel can be seen in Figure 6.2A. The DNA from this gel was transferred to Hybond N+ via Southern blotting. This membrane was used for hybridisation of the radio-actively labelled PCR product from the exon trapped clone 31B5-E21 (amplified via PCR, using the primers EX1 and EX2). Figure 6.2B shows the specific fragments that hybridised to this probe; an approximately 1.2 Kbp fragment, which was generated with the primers AP2 and 31B5-E21R, and an 0.8 Kbp fragment for AP2 and 31B5-E21F. These fragments were isolated from the remainder of the PCR reaction via an agarose gel and Prep-A-Gene, the fragments were cloned using the TA cloning Kit (InVitrogen). The transformants were grown on LB plates which contained IPTG/Xgal for blue/white screening. White colonies were analysed by PCR for amplification of a specific product using the appropriate gene-specific primer (31B5-E21F or 31B5-E21R) and either M13F+ or M13R+. In both cases, a specific fragment was amplified of a comparable size to the initial RACE products (two out of five white colonies for 31B5-E21R, one was used for further analysis and one out of twenty clones for 31B5-E21F). The products were named 31B5-E21R RACE and 31B5-E21F RACE. Cycle sequencing was performed for both RACE products, and PCR products were generated with the primers 31B5-E21R/M13R+, UG1-R1/M13R+, UG1-R2/M13R+ and UG1-R3/M13R+ from the RACE product 31B5-E21R RACE. The primers 31B5-E21R, UG1-R1, UG1-R2 and UG1-R3 were subsequently used to sequence the appropriate PCR products by cycle-sequencing. The primers 31B5-E21F/M13R+ and UG1-F1/M13R+ were used to amplify specific products from 31B5-E21F RACE, the primers 31B5-E21F and UG1-F1 were used for cycle sequencing.

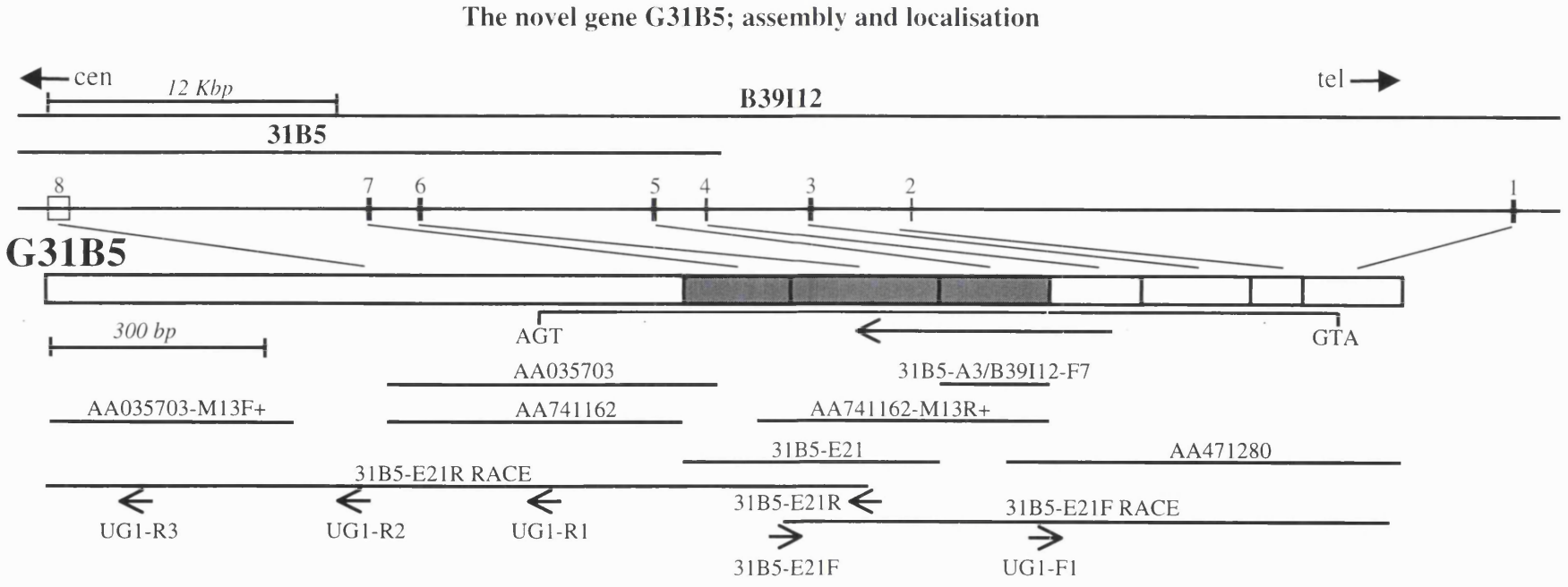


Figure 6.3: Composite picture of the novel gene G31B5. The genomic location of the eight exons is shown on the PAC clones 31B5 (AF130247) and B39I12 (AF165138). The largest ORF (1119 bp, 373 amino acids) and the direction of transcription (tel-cen) is shown below the gene. The exons shown in grey were those identified via exon trapping (31B5-E21 and 31B5-A3/B39I12-F7). The cDNA clones AA035703, AA035703-M13F+, AA741162, AA741162-M13R+, AA471280, the exon trapped clones 31B5-E21 and 31B5-A3/B39I12-F7 and the two RACE products 31B5-E21R Race and 31B5-E21F RACE are shown with their position relative to the gene. The position of primers used for sequencing the RACE products are shown below these products (arrowheads). Scale markers are indicated for the genomic distance (12 Kbp) and for the gene G31B5 (300 bp).

Sequencing of 31B5-E21R RACE revealed a poly A sequence. The positions of the gene-specific primers can be seen in Figure 6.3 and the sequences are shown in Table 5 of Appendix I. The cDNA clones AA035703 and AA741162 were ordered from the HGMP resource centre (Image clone nrs. 360025 and 1286563 respectively) and both ends of the cDNA insert were sequenced using the M13F+ and M13R+ primers (Chapter 2.20.1 step 17). The sequence of AA035703 generated with the primer M13R+ was identical to Genbank AA035703. The other end of this cDNA clone, generated by sequencing using the primer M13F+, is shown in Table 5 of Appendix I. Similarly, the Genbank sequence AA741162 is generated using the primer M13F+, the sequence AA741162-M13R+ is shown in Table 5 of Appendix I. BlastN searches were performed of the non-redundant Genbank+EMBL+DDBJ+PDB databases, with the sequences from the different cDNAs, exon trapped clones and sequenced RACE products. All these products mapped back to the PAC clone 31B5, of which the complete sequence became available in Genbank (accession number AF130247). The overlapping cDNAs, exon trapped clones and RACE products can be seen in Figure 6.3. Five exons of this gene (exons 4-8), which will be called G31B5, are present on the genomic DNA contained within 31B5 and can be seen in Figure 6.3. The ORF, and the direction of transcription (tel-cen) can also be seen in Figure 6.3. The first 361 bp of G31B5 map telomerically to 31B5 and consist of an unknown number of exons. The sequence of the genomic clone B39I12 (AF165138) only became available after work, to establish how many exons the 361 bp at the 5' end of the gene contained, was completed (see Chapter 6.3.2).

6.3.2 Direct sequencing using bacterial clone B39I12, to characterise three additional exons of G31B5.

In order to determine the exact number of exons contained within the first 361 bp of the novel gene, direct sequencing was performed using primers derived from the gene sequence. The BAC clone B39I12 (the sequence of which only became available after this work was completed), which overlaps with 31B5 (see Figure 3.13) and was likely to contain the additional exons, was digested with *Hind* III. DNA was then isolated using Wizard purification resin (Chapter 2.22.1) and Cycle sequencing was performed with the primers G31B5-A, B, C, D, E and F (see Figure 6.7 for the location of the primers). Initially, the primers G31B5-A and F were used for sequencing, and were eventually shown to map to exons 1 and 3 respectively. Both primers were used to

Nucleotide and amino acid sequence of G31B5

```
   1          ATTAGACAGCACACTGCTGACTGTTTTCAGTTGTTTCTGTAACAGCAGAAAGTGCACTCACTAGGAGTAGTCAGAATTCAAA
  83 ATGCTCAAGAGAAAGCCATCCAATGTTTCAGAGAAGGAGAAACATCAAAAACCAAAG/CGAAGCAGCAGTTTTGGGAATTTCGATCGTTTT
   1 M  L  K  R  K  P  S  N  V  S  E  K  E  K  H  Q  K  P  K   R  S  S  S  F  G  N  F  D  R  F
 173 CGGAATAATTCTTTATCAAAACCAGATGATTCAACTGAG/GCACATGAAGGAGATCCCACAAATGGAAGTGGAGAACAAAGTAAAACTTCA
  31 R  N  N  S  L  S  K  P  D  D  S  T  E   A  H  E  G  D  P  T  N  G  S  G  E  Q  S  K  T  S
 263 AATAATGGAGGCGGTTTGGGTAAAAAAATGAGAGCTATTTCATGGACAATGAAGAAAAAAGTGGGTAAAAAGTACATCAAAGCCCTTTCT
  61 N  N  G  G  G  L  G  K  K  M  R  A  I  S  W  T  M  K  K  K  V  G  K  K  Y  I  K  A  L  S
 353 GAGGAAAAG/GATGAGGAAGATGGAGAGAATGCCCACCCATATAGAAACAGTGACCCTGTGATTGGGACCCACACAGAGAAGGTGTCCCTC
  91 E  E  K   D  E  E  D  G  E  N  A  H  P  Y  R  N  S  D  P  V  I  G  T  H  T  E  K  V  S  L
 443 AAAGCCAGTGACTCCATGGATAGTCTCTACAGTGGACAGAGCTCATCAA/GTGGCATAACAAGCTGTTCAGATGGTACAAGTAACCGGGAC
 121 K  A  S  D  S  M  D  S  L  Y  S  G  Q  S  S  S  S   G  I  T  S  C  S  D  G  T  S  N  R  D
 533 AGCTTTCGACTGGATGACGATGGCCCCTATTCAGGACCATTCTGTGGCCGTGCCAGAGTGCATACGGATTTCACGCCAAGTCCCTATGAC
 151 S  F  R  L  D  D  D  G  P  Y  S  G  P  F  C  G  R  A  R  V  H  T  D  F  T  P  S  P  Y  D
 623 ACTGACTCCCTCAAAATCAAG/AAAGGAGACATCATAGACATTATTTGCAAAACACCAATGGGGATGTGGACAGGAATGTTGAACAATAAA
 181 T  D  S  L  K  I  K   K  G  D  I  I  D  I  I  C  K  T  P  M  G  M  W  T  G  M  L  N  N  K
 713 GTGGGAAACTTCAAATTCATTTATGTGGATGTCATCTCAGAAGAGGAAGCAGCCCCCAAGAAAATAAAGGCAAACCGAAGGAGTAACAGC
 211 V  G  N  F  K  F  I  Y  V  D  V  I  S  E  E  E  A  A  P  K  K  I  K  A  N  R  R  S  N  S
 803 AAAAAATCCAAGACTCTGCAGGAGTTCCTAGAGAGGATTCATCTGCAG/GAATACACCTCAACACTTTTGCTCAATGGTTATGAGACTCTA
 241 K  K  S  K  T  L  Q  E  F  L  E  R  I  H  L  Q   E  Y  T  S  T  L  L  L  N  G  Y  E  T  L
 893 GAAGATTTAAAAGATATAAAAGAGAGTCACCTCATTGAATTAAATATTGAAAACCCAGATGACAGAAGAAGGTTACTATCAGCTGCTGAA
 271 E  D  L  K  D  I  K  E  S  H  L  I  E  L  N  I  E  N  P  D  D  R  R  R  L  L  S  A  A  E
 983 AACTTCCTTGAAGAAGAAA/TTATTCAAGAGCAAGAAAATGAACCTGAGCCCCTATCCTTGAGCTCAGACATCTCCTTAAATAAGTCACAG
 301 N  F  L  E  E  E  I   I  Q  E  Q  E  N  E  P  E  P  L  S  L  S  S  D  I  S  L  N  K  S  Q
1073 TTAGATGACTGCCCAAGGGACTCTGGTTGCTATATCTCATCAGGAAATTCAGATAATGGCAAAGAGGATCTGGAGTCTGAAAATCTGTCT
 331 L  D  D  C  P  R  D  S  G  C  Y  I  S  S  G  N  S  D  N  G  K  E  D  L  E  S  E  N  L  S
1163 GACATGGTACATAAGATTATTATCACAGAGCCAAGTGACTGAACACGCATTCCCAACTATATATCTACAGATGCATTCCATTTTAACTCT
 361 D  M  V  H  K  I  I  I  T  E  P  S  D  * 373
1253 TCTTGAGCTAAAACGTCAAATAGGAGAGGAAGATAAGATAAATATTTGTAAATAAAACCTAAAGTTTAAATGTTTTAATCTGAATAATTG
1343 TACATAAAATTTTGTATCTCTAACATTCCAAATTACTGTCAATAAAATATATATTTATTATTTTAAATGCTATGTGTTAATATTTCACTT
1433 GCTTGTATTAGAAAGGCAAAATGTAAGACTTTGGTATGTGTGACATATGCTTTATTTGGCTTTATTTTACAAGTACAGTATCTGCAAAAA
1523 ACAAAGTAACCTTTTTTCATACCTGCCAGTTTTGAATTTATATATGTTATTGAACAAATAGTAATAGAGGATTCGCTGTTGAAACAAGTT
1613 GTCCAAGCAATGTTATATTCATTTTTATACTTATTGGGAAAGTGTGAGTTAATATTGGACACATTTTATCCTGATCCACAGTGGAGTTTT
1703 AGTAATTATATTTTGTTGATTTCTTCATTTTGTTTTCTGGTATAAAAGTAGAGATAATGTGTAGTCACTTCTGATTTAGTGAAACCAATT
1793 GTAATAATTGTGGAAATGTTTTGTCTTTAAGTGTAAATATTTTAAAATTTGACATACCCTAATGTTAATAATAAAAAGAACTATTTGCAT
1883 ATTGCA 1888
```

Figure 6.4: G31B5 consists of an 82 bp 5' UTR, an 1119 bp 373 amino acid ORF and a 687 bp 3' UTR. The 8 exons are separated by /.

Sizing of exons and introns, analysis of splice junctions and determination of the phase of the introns of G31B5

Exon				Intron		
No.	Size (bp)	3' Acceptor	5' Donor	No.	Phase	Size (bp)
1	82+57		CCAAAG/gtaagg	1	0	24983
2	72	ttatag/CGAAGC	ACTGAG/gttcgt	2	0	4108
3	150	aactag/GCACAT	GAAAAG/gtaaat	3	0	4318
4	130	ctgaag/GATGAG	CATCAA/gtaagg	4	1	1982
5	152	catcag/GTGGCA	ATCAAG/gtgaga	5	0	9574
6	207	ccttag/AAAGGA	CTGCAG/gtgagc	6	0	1936
7	151	ttacag/GAATAC	AAGAAA/gtaagt	7	1	12327
8	200+687	ttacag/TTATTC				
	1888	(c,a)ag/(A,G)	(A,C)AG/gt(a,g)agt			**59,228**
		57,100,100/71	86,100,71/100,100,86,86,86,43			

Table 6.1:. Columns 1 and 2 show the number and size of the G31B5 exons, respectively. Exon 1 contains an 82 bp 5' UTR, exon 8 a 687 bp 3' UTR. The total size of the gene is indicated in bold. Columns 3 and 4 show the splice junctions, column 3 the 3' acceptor and column 4 the 5' donor. The consensus splice junctions are indicated underneath (c,a)ag/(A,G) for the 3' acceptor and (A,C)AG/gt(a,g)agt for the 5' donor (Grant *et al.*, 1997). The percentages of the individual nucleotides of the G31B5 gene compared with the consensus sequence are indicated in the bottom row. Column 5 shows the intron number. Column 6 indicates the phase of the introns and column 7 shows the size of the individual introns. The intronic sizes are calculated from the sequence from the genomic clone 31B5 (accession nr AF130247) and B39I12 (AF165138). The total genomic distance that the gene occupies is 61,116 bp.

sequence the exon/intron boundary. This showed that 72 bp of the G31B5 sequence located between the two primers was still unaccounted for, which may be explained by a separate exon. This was confirmed by direct sequencing with the primers G31B5-C and –D, which were generated from the intervening sequence. Sequencing from the BAC B39I12 with the primers G31B5-B and E completed the sequencing across the exon/intron boundaries. The three additional exons can be seen in Figures 6.3 and 6.7. Figure 6.7 shows additionally the location of the different primers that were used to determine the exon/intron boundaries. The sequences of the exon/intron boundaries, which were generated with the primers G31B5-A to –F, can be seen in Appendix I Table 5. The sequences were identical to the respective sequences of the genomic clone B39I12.

The total length of the G31B5 is 1888 bp, which in combination with a poly A tail corresponds well with the 2.1 Kbp fragment from the Northern blot. The gene contains an 82 bp 5' UTR, a 1119 bp, 373 amino acid ORF and a 687 bp 3' UTR. The nucleotide and amino acid sequences can be seen in Figure 6.4. The size of the eight exons, the intron/exon borders, and the phase and size of the introns is shown in Table 6.1. A BlastP search to the SwissProt database and translations of coding databases revealed significant similarity to the KIAA0790 (AB018333) protein (3e-69, 44% identity and 61% positives in 346 amino acids), which is located on chromosome 6q22.3 (determined via the website http://www.ncbi.nlm.nih.gov/UniGene/Hs.Home.html). Similarity was also detected with dj753P9.2 (3676220), a novel putative protein from the PAC clone 753P9, which is located on human chromosome Xq25-26.1 (2e-51, 48% identity and 63% positives in 247 amino acids). The functions of these two proteins are not known.

6.3.3 Identification of a mouse homologue of G31B5 and comparison of the proteins between the two species

A BlastN search of the mouse EST database with the G31B5 sequence revealed 11 overlapping cDNA clones; AA560016, AA560511, AA798427, AA756594, AA921588, AA273615, AA763828, AA596448, AA288461, W70804 and AA032413. The cDNA clones AA273615, AA756594, AA560016, AA596448 and AA032413 form a minimal tiling path. BlastN was performed with these cDNA clones to the mouse EST database and these results were used to generate a consensus sequence, by comparing sequences from the overlapping cDNA clones. This sequence is shown in Figure 6.5

Nucleotide and amino acid sequence of M31B5, a mouse homologue of the human gene G31B5

```
   1                                GCAGCAACTGTGCTCACTAGAGGTGGCCAGAATTCAAAATGCTAAAGAGGAAGCCATCCA
  61 ATGCCTCAGACAAGGAGAAACACCAAAAACCGAAGCGCAAGCAGTTTTGGGAATTTTGACCGTTTTCGGAATAATTCCGTATCAAAATCT
   1 M  P  Q  T  R  R  N  T  K  N  R  S  A  S  S  F  G  N  F  D  R  F  R  N  N  S  V  S  K  S
 151 GATGACTCAATTGAGGTCCATGACAGGGAACTGACAAATGGAAGTGAAGAACAAAGCAAGACTTCAAGCAGTGGAGGAAGTCTAGGTAAA
  31 D  D  S  I  E  V  H  D  R  E  L  T  N  G  S  E  E  Q  S  K  T  S  S  S  G  G  S  L  G  K
 241 AAAGTGAGGGCTATTTCATGGACAATGAAGAAAAAAGTAGGAAAAAAATACATCAAAGCTCTTTCTGAGGAAAAAGAGGAAGAAAGTGGA
  61 K  V  R  A  I  S  W  T  M  K  K  K  V  G  K  K  Y  I  K  A  L  S  E  E  K  E  E  E  S  G
 331 GAGGAAGCCCTCCCATATCGGAACAGTGATCCCATGATTGGAACACATACAGAGAAGATCTCACTCAAAGCCAGCGACTCTATGGACAGT
  91 E  E  A  L  P  Y  R  N  S  D  P  M  I  G  T  H  T  E  K  I  S  L  K  A  S  D  S  M  D  S
 421 CTTTACAGTGGGCAGAGTTCATCAAGTGGAATAACAAGCTGTTCAGATGGAACAAGTAATCGGGACAGCTTCCGACTGGATGATGACAGC
 121 L  Y  S  G  Q  S  S  S  S  S  G  I  T  S  C  S  D  G  T  S  N  R  D  S  F  R  L  D  D  D  S
 511 CCCTACTCAGGGCCATTCTGTGGCCGGGCCAAAGTGCACACAGACTTCACGCCAAGTCCCTATGACACCGACTCCCTTAAAATCAAGAAA
 151 P  Y  S  G  P  F  C  G  R  A  K  V  H  T  D  F  T  P  S  P  Y  D  T  D  S  L  K  I  K  K
 601 GGAGACATCATAGACATTATCTGCAAAACACCAATGGGAATGTGGACAGGGATGCTGAACAATAAAGTGGGAAACTTCAAATTTATTTAT
 181 G  D  I  I  D  I  I  C  K  T  P  M  G  M  W  T  G  M  L  N  N  K  V  G  N  F  K  F  I  Y
 691 GTGGATGTTATCTCAGAAGAGGAAGCAGCTCCTAAGAAGATAAAGGTCCCCAGAAGCAGGAGAAGAGAGAACCACCAGACTATCCAGGAG
 211 V  D  V  I  S  E  E  E  A  A  P  K  K  I  K  V  P  R  S  R  R  R  E  N  H  Q  T  I  Q  E
 781 TTCTTAGAGAGGATTCACCTTCAGGAATACACTTCAACACTTCTGCTTAATGGCTATGAGACCCTGGATGACTTGAAGGATATCAAAGAA
 241 F  L  E  R  I  H  L  Q  E  Y  T  S  T  L  L  L  N  G  Y  E  T  L  D  D  L  K  D  I  K  E
 871 AGTCATCTCATTGAATTAAACATTGCAGATCCAGAAGACAGGGCGAGGCTGCTGTCTGCTGCCGAGAGTCTCCTGGATGAAGAAACCACT
 271 S  H  L  I  E  L  N  I  A  D  P  E  D  R  A  R  L  L  S  A  A  E  S  L  L  D  E  E  T  T
 961 GTAGAACATGAAAAGGAATCCGTGCCTCTATCCTCAAACCCAGACATCTTGAGTGCATCACAGCTAGAGGACTGCCCGAGGGACTCTGGC
 301 V  E  H  E  K  E  S  V  P  L  S  S  N  P  D  I  L  S  A  S  Q  L  E  D  C  P  R  D  S  G
1051 TGTTATATCTCATCGGAAAACTCAGATAATGGCAAGGAAGATCTGGAGTCAGAAAATCTGTCTGACATGGTGCAGAAGATTGCCATCACA
 331 C  Y  I  S  S  E  N  S  D  N  G  K  E  D  L  E  S  E  N  L  S  D  M  V  Q  K  I  A  I  T
1141 GAATCCAGTGACTGAACACACACTCTCATGTCCACACCCGCAGCCGAATTCCATTTGAACTTCTTGTATGCTACAGGATCAAATGGAAGG
 361 E  S  S  D  * 364
1231 TTTCCAAATGTACCCCAAAGTGCTAAGTGTTTCAGTCTGTACAATTGTTCATGTAAATTTATATTTCTCACTTACCAATGAGTGTGAATA
1321 AAATGTATATTTATT 1335
```

Figure 6.5: This gene consists of a 60 bp 5' UTR, a 1092 bp, 364 amino acid ORF and a 183 bp 3' UTR. The gene sequence was assembled using sequences of overlapping mouse cDNA clones from the database.

Comparison of the proteins of G31B5 and M31B5

```
Human:  11 EKEKHQKPKRSSSFGNFDRFRNNSLSKPDDSTEAHEGDPTNGSGEQSKTSNNGGGLGKKMRAISWTMKKKVGKKYIKALSEEKDEEDGEN 100
           +  ++ K + +SSFGNFDRFRNNS+SK DDS E H+ + TNGS EQSKTS++GG LGKK+RAISWTMKKKVGKKYIKALSEEK+EE GE
Mouse:   3 QTRRNTKNRSASSFGNFDRFRNNSVSKSDDSIEVHDRELTNGSEEQSKTSSSGGSLGKKVRAISWTMKKKVGKKYIKALSEEKEEESGEE  92

Human: 101 AHPYRNSDPVIGTHTEKVSLKASDSMDSLYSGQSSSSGITSCSDGTSNRDSFRLDDDGPYSGPFCGRARVHTDFTPSPYDTDSLKIKKGD 190
           A PYRNSDP+IGTHTEK+SLKASDSMDSLYSGQSSSSGITSCSDGTSNRDSFRLDDD PYSGPFCGRA+VHTDFTPSPYDTDSLKIKKGD
Mouse:  93 ALPYRNSDPMIGTHTEKISLKASDSMDSLYSGQSSSSGITSCSDGTSNRDSFRLDDDSPYSGPFCGRAKVHTDFTPSPYDTDSLKIKKGD 182

Human: 191 IIDIICKTPMGMWTGMLNNKVGNFKFIYVDVISEEEAAPKKIKANRRSNSKKSKTLQEFLERIHLQEYTSTLLLNGYETLEDLKDIKESH 280
           IIDIICKTPMGMWTGMLNNKVGNFKFIYVDVISEEEAAPKKIK  R    +   T+QEFLERIHLQEYTSTLLLNGYETL+DLKDIKESH
Mouse: 183 IIDIICKTPMGMWTGMLNNKVGNFKFIYVDVISEEEAAPKKIKVPRSRRRENHQTIQEFLERIHLQEYTSTLLLNGYETLDDLKDIKESH 272

Human: 281 LIELNIENPDDRRRLLSAAENFLEEEIIQEQENEPEPLSLSSDISLNKSQLDDCPRDSGCYISSGNSDNGKEDLESENLSDMVHKIIITE 370
           LIELNI +P+DR RLLSAAE+ L+EE   E E E  PLS + DI L+ SQL+DCPRDSGCYISS NSDNGKEDLESENLSDMV KI ITE
Mouse: 273 LIELNIADPEDRARLLSAAESLLDEETTVEHEKESVPLSSNPDI-LSASQLEDCPRDSGCYISSENSDNGKEDLESENLSDMVQKIAITE 361

Human: 371 PSD 373
            SD
Mouse: 362 SSD 364
```

Figure 6.6: Comparison at the amino acid level of the human gene G31B5 and a mouse homologue M31B5 (e-169).

and is 1335 bp long. A single large ORF was identified, consisting of 364 amino acids. This mouse gene contains a 60 bp 5' UTR, a 1092 bp, 364 amino acid ORF and a 183 bp 3' UTR. A BlastP database search showed similarity to the same proteins as G31B5, KIAA0790 (7e-70) and dj753P9.2 (2e-56). The protein sequences from 31B5 and the mouse homologue were compared by using the BlastP option of the Blast 2 sequences, from the webpage http://www.ncbi.nlm.nih.gov/gorf/bl2.cgi. Figure 6.6 shows the result of this comparison: the two proteins have a high degree of similarity, e-169, 81% identity and 88% positives in 363 amino acids. It is important to note that this gene was assembled entirely *in silico* via cDNA sequences from the mouse EST database, and could therefore contain some sequencing errors.

6.3.4 Sequencing of the gene using genomic DNA from clinical material

The location of this gene indicated that it could be implicated in a number of phenotypes that were associated with the region (this is explained in more detail in section 6.4). Mutation detection was therefore performed by sequencing using genomic DNA from the following patients:

- Two DS patients with ANLL-M7 (1559 and 2075, Chapter 2.17).
- Two lymphoblast cell lines generated from two patients with FEOAD (AG08562 and AG07877A, Chapter 2.17).
- Two LOAD patients, (LOAD 6 and 13, Chapter 2.17), who were negative for APOE-ε4 (Chapter 5.3.4).
- DNA from tumour material from two patients with NSCLC (#5 and #13, Chapter 5.3.3) that showed LOH of the region.

The primers were designed from intronic sequences adjacent to the individual exons and sequence upstream and downstream of the gene (see Figure 6.7 for the position of the primers). They were constructed from the sequence of the genomic clone 31B5 (AF130247) and from direct sequencing of the first three exons of G31B5 (see Appendix I Table 5). The primers used for amplification of the different exons were G31B5-1F/R, 2F/R, 3F/R, 4F/R, 5F/R, 6F/R, 7F/R, and 8F/R, and these products were sequenced using the primers G31B5-1F, 2R, 3R, 4F, 5F, 6F, 7F, and 8F respectively. The numbers of the primers refer to the different exons, which are indicated in Figure 6.7. The nucleotide sequence of G31B5 which was determined using the different primer combinations can be found in Chapter 2.22.2. No mutations were found in the

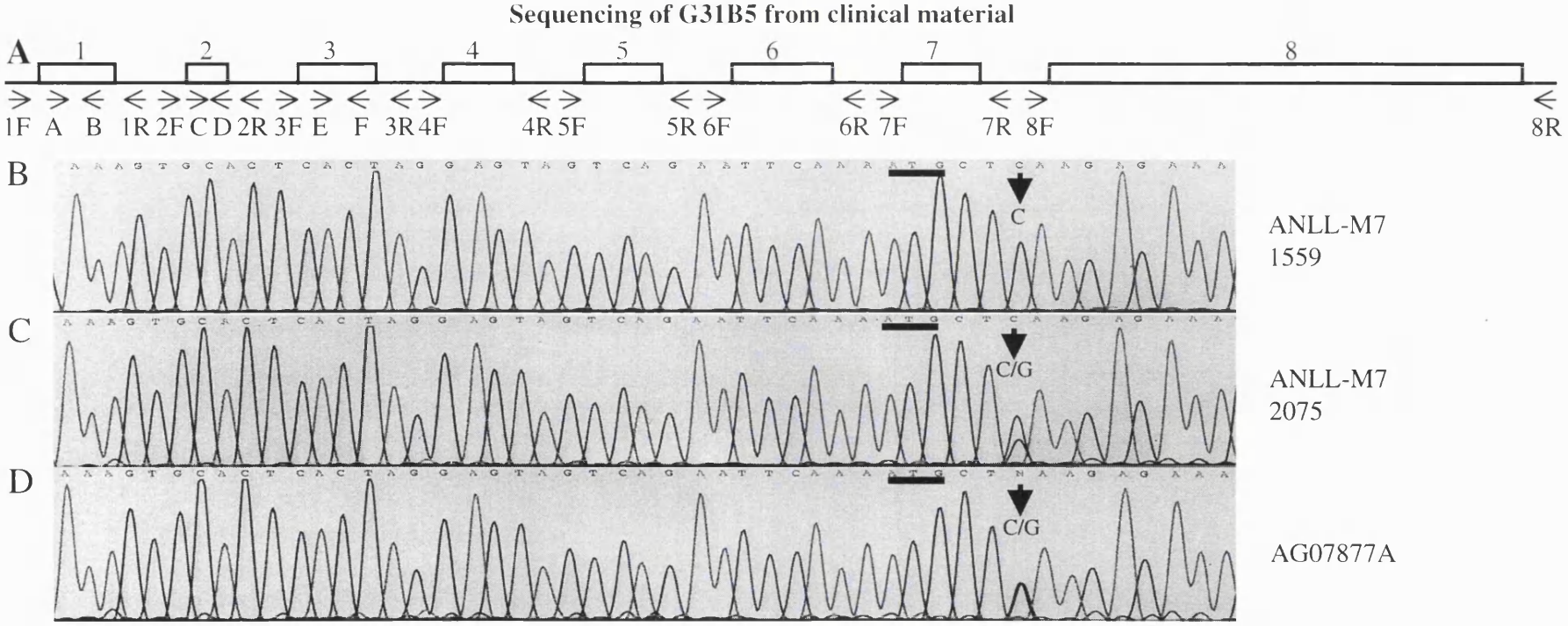


Figure 6.7: Sequencing of G31B5 from genomic DNA. A-The 8 exons of G31B5 and the primers that were used for sequencing are depicted. The primers G31B5-A, -B, -C, -D, -E and -F were used for direct sequencing to determine the intron/exon structure of the first three exons. The other primers were used for sequencing of their respective exons from genomic DNA for mutation detection. B/C/D-shows the quality of the sequence and a silent polymorphism in the third nucleotide of the second codon of G31B5. The polymorphism is indicated via an arrow and the ATG start codon with a horizontal bar. B-sample of Down syndrome patient 1559 with ANLL-M7 showing a homozygous C at nucleotide position 88 of G31B5. C-identical to B for patient 2075; a heterozygous C/G is present at position 88. D-identical to B for a patient with Alzheimer's disease; a heterozygous C/G is present at position 88. The polymorphism in C shows the trisomic state of this DS sample as the G signal is approximately half the strength of the C signal, compared to the equal strength signals for the normal karyotype in patient AG07877A.

genomic DNA from the different patients in the eight exons or the 3' and 5' splice sites. Exons 1-7 were fully sequenced, but only the coding region was sequenced from exon 8 (up to and including nucleotide 1204, see Figure 6.4). At nucleotide position 921 (exon 7), a homozygous A residue was found in all sequenced samples compared with a C residue in the sequence from clone 31B5 (AF130247). This results in a histidine instead of a proline at position 280 in the protein. In the mouse sequence, a histidine is also present at this position (amino acid 272, Figure 6.5), suggesting this to be a sequencing error in the genomic sequence of clone 31B5. This was confirmed after cycle sequencing of exon 7 from the PAC clone 31B5, which revealed an A residue at nucleotide position 921 of G31B5.

A silent single nucleotide polymorphism was found in the two FEOAD cell lines and in one Down's patient with ANNL-M7, at nucleotide position 88. A C/G heterozygote was found in these patients (see Figure 6.4 for the gene sequence and Figure 6.7 for the polymorphism). This is the third nucleotide in the leucine codon which remains unchanged in the protein. The sequence of G31B5 and the genomic clone B39I12 show a C residue at this nucleotide. The polymorphism shown in Figure 6.7 suggests that it is possible to detect disomic homozygosity via cycle sequencing. This may help in the identification of the susceptibility gene that might be involved in ANLL-M7 and TAM in DS (Niikawa *et al.*, 1991). A G/A heterozygote was additionally found in a LOAD patient (13) at nucleotide 64 in the 5' UTR. An example of the quality of the sequence, which is important for the accuracy of mutation detection, can be seen in Figure 6.7.

6.4 Discussion

The novel gene G31B5 was initially discovered via exon trapping. This technique identified two exon trapped clones, 31B5-E21 and 31B5-A3/B39I12-F7, that mapped back to the same cDNA clone (Image accession nr. 1286563). The full length sequence of the gene was determined via overlapping cDNA clones, exon trapped material and 3' and 5' RACE products, which were generated from primers designed from the sequence of the exon trapped clone 31B5-E21. Five exons were found to map back to the fully sequenced bacterial clone 31B5, and three further exons were determined via direct sequencing using the clone B39I12. The bacterial clones 31B5 and B39I12 were overlapping and mapped to a region between the markers D21S120 and D21S371 (see Figure 3.13).

The total length of the gene is 1888 bp, which, together with the poly A tail, corresponds well with the 2.1 Kbp fragment that hybridised to the exon 31B5-E21 on a Northern blot in lung, placenta, brain and heart. The mRNA fragments of 1.5 Kbp in liver and 700 bp in heart, that hybridise to the same probe, could be explained by alternative splicing. Hybridisations to additional human Northern blots, using the same probe, showed a 2.1 Kbp fragment in stomach, thyroid, spinal cord, bone marrow and spleen. No expression was detected in lymph node, trachea, testis and adrenal gland (not shown).

G31B5 contains a 1119 bp, 373 amino acid ORF, from the first ATG start codon. The 5' end of the gene was determined by comparing the 31B5-E21F RACE product with the cDNA clone AA471280. The sequence similarity of these two clones with the genomic clone 31B5 starts within 13 bp of each other, suggesting that this is the 5' end. Two in-frame termination codons can be found twenty-one and fifteen base pairs upstream from the start codon (Figure 6.4). The AAAatgC sequence surrounding the start codon is an adequate context for the start of translation (Kozak, 1996). An AATAAA poly-adenylation signal can be found, twenty-seven bp before the 3' end. The 3' end of the gene was determined from 31B5-E21R RACE and eight cDNA clones (AA831672, AA994323, AI223999, AA747998, AI371399, N67227, AA063573 and C01451, these clones are not shown in Figure 6.3). The 3' end, as indicated in Figure 6.4, is similar to the 3' end of two cDNA clones (AI371399 and AA994323). The RACE clone and one cDNA clone contain 1 bp less, two cDNA clones contain 2 bp less, two have 5 bp less and one has 7 bp less than the sequence that is shown in Figure 6.4.

The best evidence for the validity of the protein sequence is the comparison with a mouse homologue. The mouse protein is very similar to G31B5; 364 amino acids, 81% identity, 88% positives in 363 amino acids (Figure 6.6). There could however be some errors in the mouse sequence, as this gene was assembled via cDNA sequences from the mouse EST database.

Three exon prediction programmes (Grail, Genscan and MZEF, Uberbacher and Mural, 1991; Burge and Karlin, 1997; Zhang, 1997) have been used in the Genbank submitted sequence of the PAC clones 31B5 (AF130247) and B39I12 (AF165138). Six of the eight exons (exons 2-7) of G31B5 that map back to the PAC clones, were predicted precisely. Exons 1 and 8 were only partly anticipated. Eighteen exons were wrongly predicted in the interval where G31B5 is located. Fifteen of these were only predicted by one of the programmes, and three by two prediction programmes. Exons 2-7 of

Exon predictions in the genomic interval of the G31B5 gene

	Number of exons (total nr. of G31B5 exons)	Exons predicted by number of exon prediction programs		
		3	2	1
G31B5 exons correctly predicted	6 (8)	6	0	0
G31B5 exons wrongly predicted	18	0	3	15

Table 6.2: The number of G31B5 exons that were correctly predicted are shown (6 out of 8). The first three exons were predicted from the BAC B39I12 (AF165138), the other five from the PAC 31B5 (AF130247) with Grail, MZEF and Genscan. All six correctly predicted exons were predicted by all three exon prediction programs. Eighteen exons were predicted wrongly in the genomic interval where G31B5 is located. Three of these were predicted by two programmes and fifteen by one programme.

G31B5 were predicted by all three programmes, showing again (see also Chapter 5.4) that multiple predictions increase the chance that a putative exon is part of a gene (Table 6.2). This analysis does not take genes within genes into account, it also does not consider the fact that some of the wrongly predicted exons could be alternatively spliced exons in certain tissues.

Similarity was found between *G31B5* and two other proteins in the SwissProt database; dj753P9.2 and KIAA0790. No function has been predicted for either of these proteins. The amino acid sequence of G31B5 was subjected to PROSITE (PC/Gene version 11.0) and a number of phosphorylation and glycosylation sites were predicted. No membrane associated domains were predicted and also no secretory signal was found. However, two bipartite nuclear localisation signals near the amino terminal end were found, which suggest that this may be a nuclear protein. Further analysis will be necessary to determine this.

The mapping position of the gene and the fact that the region is gene-poor (Gardiner, 1996) justified a preliminary investigation of this gene's involvement in the following phenotypes:

- The gene maps to the clones 31B5 and B39I12 between the markers D21S120 and D21S192 (see Figure 3.13), markers that were associated with ANLL-M7 (Abe *et al.*, 1989; Niikawa *et al.*, 1991; Shen *et al.*, 1995; Cavani *et al.*, 1998).
- The marker D21S16 that showed linkage in some families with FEOAD is located on the PAC 31B5.
- The marker most often deleted in squamous NSCLC was D21S120 (Kohno *et al.*, 1998), which is located just centromeric to this gene (see Figure 3.13). This latter marker was one of the markers on 21q that showed positive lod scores in sib pair analysis of LOAD samples that were APOE-ε4 negative (Kehoe *et al.*, 1999), although it was not the highest lod score on chromosome 21.

To test the possibility that this gene was involved in these phenotypes, mutation detection was carried out by direct sequencing. Exons 1-7 and the coding region from exon 8 of G31B5 were sequenced using genomic DNA from two Down syndrome patients with ANLL-M7, genomic DNA from two lymphoblast cell lines generated from two FEOAD patients, genomic DNA from tumour material from two patients with squamous NSCLC, and genomic DNA from two APOE-ε4 negative patients with LOAD. No mutations were found in any of the patient material tested. However, the

sample size was small and additional patient material will have to be sequenced to draw any firm conclusions about the involvement of G31B5 in ANLL-M7, FEOAD and LOAD. Sequencing of a silent polymorphism present in the coding region of two trisomic patients with ANLL-M7 and a disomic patient with FEOAD, suggested that it might be possible to detect a causative mutation which could show disomic homozygosity (see Figure 6.7). Sequencing of this gene in squamous NSCLC tumour DNA was performed before the smallest region of overlapping deletion (SRO) was determined, this subsequently ruled G31B5 out, as it was not located within the SRO.

We have identified a second novel full length gene (G31B5). Initially, a cluster of cDNA and exon trapped clones were identified via exon trapping and subsequent database searches (Figure 4.11). Amplification and sequencing of 5' and 3' RACE products and the assembly of these RACE fragments, exons and cDNA clones, determined the full length gene (Figure 6.3). The gene consists of eight exons, is 1888 bp in size and contains a 373 amino acid ORF. Similarity was detected between this and two other proteins in the SwissProt database. No putative function can be predicted for this novel protein, although it may be localised in the nucleus. A mouse homologue was identified via overlapping cDNA clones. This mouse protein showed a high degree of similarity to *G31B5*. Initial mutation detection in a limited number of samples showed no mutations in material from patients that suffered from ANLL-M7, squamous NSCLC, FEOAD and LOAD.

Chapter 7 General discussion

In order to be able to identify novel genes in the 21q11 region a contig was constructed consisting of overlapping genomic inserts, which were present in bacterial vectors. The overlapping clones consisted of approximately 3.2 Mbp of genomic DNA, one approximately 200 Kbp gap remained (Chapter 3). The clones were made available to the human genome project for sequencing of the region. The German Sequencing Consortium analysed the clones via fluorescent *in situ* hybridisation (FISH), fingerprinting and hybridisation (specific fragments were generated from T3, T7 and SP6 promoters) (T. Hildmann, personal communication). The clones were subsequently distributed and sequencing has commenced (see Chapter 3.4.3). The clones that showed hybridisation to multiple human chromosomes by FISH should be analysed further to be certain as to their chromosomal location. Multiple clones could be sequenced in order to detect differences from which STSs could be made that would distinguish between different chromosomal locations. The analysis of the duplicated region present in 21q11 and 21q22.1 is a good example: RFLP analysis was necessary to establish the location of the bacterial clones. Determining which of these clones are located in chromosome 21q11 will also identify the chromosomal location of the clones that are not present in 21q11. This information will assist the centres involved in sequencing those areas, and will avoid duplication of the work.

The potentially very complicated structure of the proximal part of 21q11 with respect to sequence similarities between different chromosomal regions and over large distances, could also indicate that clone-by-clone sequencing is preferred over whole genome shotgun sequencing. It is likely that the finishing phase in whole genome shotgun sequencing could not resolve these areas. This would result in a clone-by-clone approach being undertaken for these regions, resulting in duplication of work.

Twenty-two clones that form the minimal tiling path of approximately 2 Mbp of the bacterial contig were used in exon trapping experiments. More than 7000 clones were included in the analysis and ninety-three individual exon trapped products mapped back specifically to the parental bacterial clones. Nineteen of these were identical to human cDNA clones (Chapter 4, Table 4.2). A number of open reading frames generated from the exon trapped clones showed resemblance to proteins present in the SwissProt database.

Experiments could be conducted in order to isolate full-length genes. Hybridisation of human 21q11 exons to Southern blots, which contain restriction digested genomic DNA from mouse, could identify sequences which are conserved between human and mouse. These exon trapped clones could then be used in hybridisation experiments onto dotblots that contain cDNA species from different tissues, and multi-tissue Northern blots. This would identify expression patterns, and show the size of the full-length transcripts of the exons that are part of genes. Alternatively, expression patterns and transcript sizes could be detected by hybridisation of all exon trapped clones. The identification of the DNA sequence of the full-length genes could then, for example, be performed by RACE or by screening of cDNA libraries.

The clone 98L15BB-N7 showed similarity to rat lipase PS-PLA1. This rat protein is secreted from activated platelets and hydrolyses the fatty acyl residue at the sn1 position of serine phospholipids (Sato *et al.*, 1997), which are components of cell membranes. Phospholipase A2, a related enzyme which hydrolyses fatty acyl residues at the sn2 position of phospholipids, has been shown to be activated by different colony stimulating factors (Dellapuca and Gallicchio, 1996) and could therefore be involved in the differentiation of haematopoietic precursors. Interestingly, the rat PS-PLA1 gene was identified via the purification of the PS-PLA1 protein from activated platelets; megakaryoblastic leukaemia (acute non-lympocytic leukaemia-M7 (ANLL-M7)) originates from the precursors that form platelets. The human homologue of PS-PLA1 has recently been identified, but this gene does not show expression in platelets (Nagai *et al.*, 1999). It would therefore be interesting to identify the full length novel gene, to show where it is expressed and to determine if it can function as a lipase and elucidate the substrate it hydrolyses. Furthermore, experiments can be conducted similar to the ones described by Dellapuca and Gallicchio (1996) to show if the protein is involved in haematopoietic differentiation.

A number of exon trapped products (90B5BB-A22, 98L15S-G9, 126N20-A22 and 126N20-N22) showed similarity to multidrug resistance- or multidrug resistance-associated proteins. The similarities in amino acid sequences between these putative exons and the thirty-one exons containing human multidrug resistance-associated protein (MRP1) are 98L15S-G9 with exon 6, 90B5BB-A22 with exons 7 and 8, 126N20-N22 with exons 28 and 29 and 126N20-A22 with exon 31. These proteins that are part of the ATP-binding cassette transporter superfamily, can cause multidrug resistance in, for example, cancer cells that are treated with a variety of cytostatic drugs.

The definition of multidrug resistance is the simultaneous acquisition of resistance to multiple drugs that are dissimilar in structure and function, and which does not result from the accumulation of independent mutations, each conferring resistance to a single drug (Gottesman *et al.*, 1995). These proteins contain multiple membrane spanning regions as well as two nucleotide binding domains (reviewed by Cole and Deeley, 1998). Experiments to elucidate the full length novel gene present on chromosome 21q11 have shown that multiple stop codons are present within the exons that show similarity to the multidrug resistance genes (R.H. Flomen, J. Groet, unpublished results). For example, parts of the amino acid sequence of the putative exon 126N20-N22 show similarity to MRP1 in three different reading frames. It is unlikely that this has been caused by sequencing errors in our sequence of 126N20-N22, as this is identical to the genomic sequence of the PAC clone 126N20 (AF130249). However, different results indicate that part of this novel putative gene is transcribed, which could therefore still have a specific function. The fact that, so far, all human cDNA clones that were found to be identical to part of this gene were identified from a foetal liver-spleen cDNA library (see Figure 4.9), might indicate a specific function in either or both of these tissues. As this gene is expressed in bone marrow and possibly in foetal liver, two tissues in which haematopoiesis takes place, this warrants mutation detection in DS individuals with ANLL-M7/transient abnormal myelopoiesis (TAM). This should ideally include promoter, enhancer and intronic sequences. Studies could be performed that compare levels of transcript of the gene between affected and unaffected DS infants. A specific antibody could be raised against the protein to show cellular localisation. Additionally, this antibody could be used in immunoprecipitation experiments to detect putative interacting proteins. In parallel, yeast two hybrid screens would also discover potential protein-protein interactions.

An initial comparison between the exon trapping experiments (Chapter 4) and preliminary cDNA selection (R.H. Flomen) can be made. The genomic clones that were used for cDNA selection were 90B5, 126N20, 30P13 and 270M7, and the source of mRNA was bone marrow. Species of cDNA clones were detected that were identical to the full-length genes STCH (present on the PAC clones 90B5 and 126N20) and RIP140 (located on 30P13 and 270M7).

Hybridisation was performed (R.H Flomen) using exon trapped clones, that mapped back to 90B5, 126N20, 30P13 or 270M7 (Figure 4.7), as probes to filters containing

DNA from the clones resulting from the cDNA selection approach. The following exon trapped clones hybridised to clones that were derived via cDNA selection:

- 90B5BB-A22, 98L15S-G9. These exons are part of the DNA sequence that shows similarity to multidrug resistance genes (Figure 4.9 and Chapter 4.3.3.3)
- 98L15S-A9. This exon is part of a human EST cluster (see Figure 4.10).
- 126N20-E1.

A number of cDNA selected clones were detected that were partly identical to two independent clusters of human ESTs, which have, so far, not been detected by exon trapping. However, corresponding exon trapped species may still be present as the cDNA selected clones have not been hybridised to the respective filters containing DNA from the exon trapped clones. One further cDNA selected clone showed a large open reading frame (136 amino acids), but did not detect any human ESTs.

Expansion of the cDNA selection procedure to the other clones that were used for exon trapping and extensive analysis of both methods will be necessary to detect the majority of genes in the region and for the comparison of both methods. The gene prediction programmes that use the sequence of the genomic clones in the 21q11 region will also have to be carefully analysed.

Full length genes have so far been determined from two transcriptional units that were identified via exon trapping. G31B5 is a novel gene with an unknown function (see Chapter 6). The total length of the gene, which consists of eight exons, is 1888 bp and it codes for a 373 amino acid protein, which shows a high degree of similarity to a mouse homologue. Human proteins that show similarity to the novel protein are located on chromosome 6q22.3 and Xq25-26.1. No mutations have been detected in the coding DNA from this gene in material from patients with diseases that have been associated with this region. These were DNA from two patients with ANLL-M7, two patients with familial early onset Alzheimer's disease (FEOAD), two patients with late onset Alzheimer's disease (LOAD) and two patients with non-small cell lung carcinoma (NSCLC). Additional experiments could be performed, similar to the ones described for the gene which is partly similar to the multidrug resistance gene.

USP21 is the second full length novel gene identified in this study (Chapter 5). This gene consists of twenty-four exons, is 3803 bp long and codes for a 1055 amino acid protein. The gene was shown to be located in a smallest region of overlapping deletion, which is part of a region that is frequently deleted in non-small cell lung carcinoma. This novel protein showed strong similarities with ubiquitin-specific proteases,

especially in the functional domains. This function was proven *in vitro*, via cloning of the coding region into a prokaryotic expression vector and cleavage of a ubiquitin-β-galactosidase fusion protein. This family of proteins can remove ubiquitin moieties from the proximal end, after degradation of the target protein and removal of the proximal peptide fragment molecules (reviewed by Wilkinson, 1997). Interestingly, ubiquitin has been detected in senile plaques and neurofibrillary tangles in the brains of patients with Alzheimer's disease (AD) (Mori *et al.*, 1987; Perry *et al.*, 1987; Morishima-Kawashima *et al.*, 1993). This could indicate a possible involvement of proteins of the USP family in the development of AD. The fact that almost all people with Down syndrome will generate AD-like neurohistopathological changes in the brain (Kesslak *et al.*, 1994), indicates that genes on chromosome 21 could be important in AD. An exceptional DS individual of 78 years that did not show the neurohistopathological changes in post-mortem brain analysis, and was partially trisomic for chromosome 21, had the proximal part of the chromosome present only in two copies (Prasher et al, 1998). The region that remained disomic contained the APP gene and USP21.

Causative mutations have been shown in the APP gene in FEOAD, and it may therefore be involved in DS-related AD. In Ts65Dn mice, no senile plaques or neurofibrillary tangles were observed, although the mice were trisomic for the APP gene and a 2.2-fold increase in mRNA level for this gene was detected (Reeves *et al.*, 1995). The segment proximal to APP on mouse chromosome 16 (including the region where the human USP21 gene is located) is only present in two copies in Ts65Dn mice. Although these observations are far from conclusive, they indicate that USP21 might potentially be involved in AD. Further analysis of the specific function of USP21 could potentially determine the involvement of this gene in AD.

No mutations were detected in the gene in two patients with LOAD. However, the sample size was too limited to draw firm conclusions. Moreover, the entire USP21 sequence was not determined in these two patients and therefore mutations could be present in these areas.

One of the loci most frequently deleted in non-small cell lung carcinoma is 3p21 (Testa *et al.*, 1994). A USP (Unph or USP4) was identified at 3p21.3 which showed elevated levels of expression in human lung tumours (Gray *et al.*, 1995). Subsequently, reduced expression of this gene was detected in small cell lung carcinoma cell lines, which led to the suggestion that USP4 may be a candidate tumour suppressor gene (Frederick *et*

al., 1998). However, sequencing of USP21 using genomic DNA from patients with non-small cell lung carcinoma and loss of heterozygosity of the area, did not show mutations. Although we cannot exclude that mutations are present in the first exon, which could not be sequenced due to a high content in G and C residues.

Similar experiments could be performed to the ones described for the gene which is partly similar to the multidrug resistance gene. Changes in methylation patterns, and potential changes in expression levels, can be detected by analysis of the CpG island at the 5' end of the gene. Over-expression of the USP21 protein, in for example Ts65Dn mice (Chapter 1.7.1), could be performed to determine if this protein can assist in the generation of the AD-like phenotype. Knock-out mice would also aid in suggesting a possible specific function for the protein.

Summary of conclusions

This study has resulted in the generation of a physical map of chromosome 21q11 in bacterial clones containing overlapping human genomic inserts. This will contribute, via sequencing of the genomic DNA, to the eventual identification of all the genes located in chromosome 21q11. These can then be tested for involvement in the diseases that have been associated with the region. So far, exon trapping has led to the identification of two novel full length genes as well as putative exons and overlapping cDNA clones of potentially very interesting genes.
The specific conclusions from the experimental chapters are:

Chapter 3- A physical map is presented of the 21q11 region, consisting of overlapping PACs, BACs and cosmids. The map covers a region of approximately 3.4 Mbp in length and contains one gap of approximately 200 Kbp. The positions of two already characterised genes are indicated, together with twenty-four markers, five cDNA selected clones (Cheng *et al.*, 1994), one exon trapped clone (Yaspo *et al.*, 1995) and two cDNA clusters, which are represented in the Unigene collection (Schuler *et al.*, 1996).

Chapter 4- Describes exon trapping from bacterial clones represented in the physical map from Chapter 3. Twenty-two overlapping clones, covering a region of approximately 2 Mbp, were used in forty exon trapping experiments. More than 7000 clones were included in the analysis, over 3000 were analysed. Ninety-three putative exons were sequenced and their mapping positions were determined. These sequences were compared to non-redundant, cDNA and protein databases. Sixty-three contained a perfect open reading frame, nineteen putative exons were very similar to cDNA clones.

Chapter 5- The exon trapping experiments from Chapter 4 identified a cDNA clone that was identical to six different exon trapped clones. Hybridisation of this cDNA clone to a Northern blot showed a 4.1 Kbp mRNA fragment. The full length sequence of the gene was subsequently determined using genomic sequence from the bacterial clones 73M5

and 135E14, and from overlapping cDNAs and nine exon trapped clones. The gene is 3803 bp long, contains 24 exons and occupies approximately 149 Kbp of genomic DNA. The ORF is 1055 amino acids long. Protein prediction indicated similarity to ubiquitin-specific proteases. This function was proven *in vitro*, after cloning of the gene into a prokaryotic expression vector and cleavage of a ubiquitin-β-galactosidase fusion protein. The gene was shown to reside in a smallest region of overlapping deletion in squamous NSCLC. Mutation detection, by sequencing, using genomic DNA from a limited number of patients with squamous NSCLC or LOAD did not show any mutations.

Chapter 6- The exon trapping experiments from Chapter 4 identified a putative exon (31B5-E21) that was overlapping with the Image cDNA clone 1286563. This cDNA clone hybridised to the putative exon 31B5-A3/39I12-F7, resulting in a cluster of two putative exons and five cDNA clones (Figure 4.11). Overlapping cDNA clones were sequenced and assembled and 3' and 5' rapid amplification of cDNA end (RACE) products were generated, which were subsequently cloned and sequenced. After RACE products and cDNA clone sequences had been assembled and compared with the sequenced genomic DNA from PAC clone 31B5 (accession nr. AF130247), a new gene was discovered. The total length of the gene is 1888 bp (containing a 373 amino acid ORF). A mouse homologue was identified using overlapping cDNA clones (1335 bp, containing a 364 amino acid ORF). Mutation detection by sequencing, using genomic DNA from a limited number of patients with NSCLC, LOAD, FEOAD and ANLL-M7, did not show any mutations.

References

Aaltonen J., P. Björses, L. Sandkuijl, J. Perheentupa and L. Peltonen. 1994. An autosomal locus causing autoimmune disease: Autoimmune polyglandular disease type 1 assigned to chromosome 21. *Nature Genetics* **8:** 83-87.

Aaltonen J., N. Horelli-Kuitunen, J.-B. Fan, P. Björses, J. Perheentupa, R. Myers, A. Palotie and L. Peltonen. 1997. High-resolution physical and transcriptional mapping of the polyendocrinopathy-candidiasis-ectodermal dystrophy locus on chromosome 21q22.3 by FISH. *Genome Research* **7:** 820-829.

Abe K., K. Tadashi and N. Niikawa. 1989. Disomic homozygosity in 21-trisomic cells: a mechanism responsible for transient myeloproliferative syndrome. *Human Genetics* **82:** 313-316.

Adams A.D., A.R. Kerlavage, R.D. Fleischmann, R.A. Fuldner, C.J. Bult, N.H. Lee, E.F. Kirkness, K.G. Weinstock, J.D. Gocayne, O. White *et al.* 1995. Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature* **377:** 3-17.

Ahonen P., S. Myllärniemi, I. Sipilä and J. Perheentupa. 1990. Clinical variation of autoimmune-polyendocrinopathy-candidiasis-ectodermal dystrophy (APECED) in series of 68 patients. *N. Eng. J. Med.* **322:** 1829-1836.

Albertsen H.M., H. Abderrahim, H.M. Cann, J. Dausset, D. Le Paslier and D. Cohen. 1990. Construction and characterization of a yeast artificial chromosome library containing seven haploid human genome equivalents. *Proc. Natl. Acad. Sci. USA* **87:** 4256-4260.

Altschul S.F., T.L. Madden, A.A. Schäffer, J. Zhang, W. Miller and D.J. Lipman. 1997. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucl. Acids Res.* **25:** 3389-3402.

Alzheimer, A. 1907. Über eine eigenartige Erkrankung der Hirnrinde. *Centralblatt für Nervenheilkunde und Psychiatrie* **30:** 177-179.

Antonarakis S.E., A.C. Warren, M.K. McKormick, J.G. Lewis, P.A. Hieter and A. Chakravarti. 1989. Molecular mapping of chromosome 21 and the region responsible for Down syndrome. *Prog. Clin. Biol. Res.* **311:** 29-43.

Antonarakis S.E., M.B. Petersen, M.G. McInnes, P.A. Adelsberger, A.A. Schinzel, F. Binkert, C. Pangalos, O. Raoul, S.A. Slaugenhaupt, M. Hafez *et al.* 1992. The meiotic stage of nondisjunction in trisomy 21: determination by using DNA polymorphisms. *Am. J. of Hum. Genet.* **50:** 544-550.

Antonarakis S.E., D. Avramopoulos, J-L. Blouin, C. Conover Talbot and Schinzel A.A. 1993. Mitotic errors in somatic cells cause trisomy 21 in about 4.5% of cases and are not associated with advanced maternal age. *Nature Genetics* **3:** 146-150.

Antonarakis S.E. 1998. 10 Years of *Genomics*, chromosome 21, and Down syndrome. *Genomics* **51:** 1-16.

Baker R.T., J.W. Tobias and A. Varshavsky. 1992. Ubiquitin-specific proteases of *Saccharomyces cerevisiae. J. Biol. Chem.* **267:** 23364-23375.

Berkovic S.F., F. Andermann, S. Carpenter and L.S. Wolfe. 1986. Progressive myoclonus epilepsies: specific causes and diagnosis. *N. Engl. J. Med.* **315:** 296-305.

Bernardi G. 1995. The human genome: organisation and evolutionary history. *Annu. Rev. Genetics* **29:** 445-476.

Björses P., J. Aaltonen, A. Vikman, J. Perheentupe, G. Ben-Zion, G. Chiumello, N. Dhal, P. Heideman, J.J.G Hoorweg-Nijman, L. Mativon *et al.* 1996. Genetic homogeneity of autoimmune polyglandular disease type 1. *Am. J. Hum. Genet.* **59:** 879-886.

Bosch A., J. Guimera, S. Graw, K. Gardiner, I. Chumakov, D. Patterson and X. Estivill. 1996. Integration of 30 CA-repeat markers into the cytogenetic, genetic and YAC maps of human chromosome 21. *Eur. J. Hum. Genet.* **4:** 135-142.

Botstein D., R.L. White, M. Skolnick and R.W. Davis. 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* **32:** 314-331.

Brodsky G., G.A. Otterson, B.B. Parry, I. Hart, D. Patterson and F.J. Kaye. 1995. Localization of STCH to human chromosome 21q11.1. *Genomics* **30:** 627-628.

Brownstein B.H., G.A. Silverman, R.D. Little, D.T. Burke, S.J. Korsmeyer, D. Schlessinger and M.V. Olson. 1989. Isolation of single-copy human genes from a library of yeast artificial chromosome clones. *Science* **244:** 1348-1351.

Buckler A.J., D.D. Chang, S.L. Graw, J.D. Brook, D.A. Haber, P.A. Sharp and D.E. Housman. 1991. Exon amplification: A strategy to isolate mammalian genes based on RNA splicing. *Proc. Natl. Acad. Sci. USA* **88:** 4005-4009.

Burge C. and S. Karlin. 1997. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268:** 78-94.

Burke D.T., G.F. Carle and M.V. Olson. 1987. Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. *Science* **236:** 806-812.

Burset M. and R. Guigo. 1996. Evaluation of gene structure prediction programs. *Genomics* **34:** 353-367.

Butler R., D.J. Ogilvie, P.Elvin, J.H. Riley, R.S. Finniear, G. Slynn, J.E.N. Morten, A.F. Markham and R. Anand. 1992. Walking, cloning and mapping with yeast artificial chromosomes: a contig encompassing D21S13 and D21S16. *Genomics* **12:** 42-51.

Cai S.-Y., R.W. Babbit and V.T. Marchesi. 1999. A mutant deubiquitinating enzyme (Ubp-M) associates with mitotic chromosomes and blocks cell devision. *Proc. Natl. Acad. Sci. USA* **96:** 2828-2833.

Carver E.A. and L. Stubbs. 1997. Zooming in on the human-mouse comparitive map: genome conservation re-examined on a high-resolution scale. *Genome Research* **7:** 1123-1137.

Cavailles V., S. Dauvois, F. L'Horset, G. Lopez, S. Hoare, P.J. Kushner and M.G. Parker. 1995. Nuclear factor RIP140 modulates transcriptional activation by the estrogen receptor. *The EMBO J.* **14:** 3741-3751.

Cavani S., F., C. Perfumo, A. Argusti, M. Pierluigi, L. Perroni, K. Schmiegelow, M.B. Petersen, F. Cotter, P. Strigini, F. Dagna-Bricarelli and D. Nizetic. 1998. Cytogenetic and molecular study of 32 Down syndrome families: potential leukaemia predisposing role of the most proximal segment of chromosome 21q. *Brit. J. of Haem.* **103:** 213-216.

Chen H., R. Chrast, C. Rossier, M.A. Morris, M.D. Lalioti and S.E. Antonarakis. 1996. Cloning of 559 potential exons of genes of chromosome 21 by exon trapping. *Genome Research* **6:** 747-760.

Cheng J.-F., V. Boyartchuk and Y. Zhu. 1994. Isolation and mapping of a human chromosome 21 cDNA: Progress in constructing a chromosome 21 expression map. *Genomics* **23:** 75-84.

Chiang P.-W., G. Dzida, J. Grumet, J.-F. Cheng, W.-J. Song, E. Crombez, M.L. Van Keuren and D.M. Kurnit. 1995. Expressed sequence tags from the long arm of human chromosome 21. *Genomics* **29:** 383-389.

Choo K.H., B. Vissel, A. Nagy and P. Kalitsis. 1991. A survey of the genomic distribution of alpha satellite DNA on all the human chromosomes, and derivation of a new consensus sequence. *Nucl. Acids Res.* **19:** 1179-1182.

Chumakov I., P. Rigault, S. Guillou, P. Ougen, A. Billaut, G. Guasconi, P. Gervy, I. LeGall, P. Soularue, L. Grinas *et al.* 1992. Continuum of overlapping clones spanning the entire human chromosome 21q. *Nature 359:* 380-387.

Church D.M., C.J. Stotler, J.L. Rutter, J.R. Murrell, J.A. Trofatter and A.J. Buckler. 1994. Isolation of genes from complex sources of mammalian genomic DNA using exon amplification. *Nature Genetics* **6:** 98-105.

Cole S.P.C. and R.G. Deeley. 1998. Multidrug resistance mediated by the ATP-binding cassette transporter protein MRP. *Bioessays* **20:** 931-940.

Collins J. and B. Hohn. 1978. Cosmids: a type of plasmid gene-cloning vector that is packageable *in vitro* in bacteriophage λ heads. *Proc. Natl. Acad. Sci. USA* **75:** 4242-4246.

Collins F. and D. Galas. 1993. A new five-year plan for the U.S. Human Genome Project. *Science* **262:** 43-46.

Corder E.H., A.M. Saunders, W.J. Strittmatter, D.E. Schmechel, P.C. Gaskell, G.W. Small, A.D. Roses, J.L. Haines and M.A. Pericak-Vance. 1993. Gene dose of Apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science* **261:** 921-923.

Cotran R.S., V. Kumar and T. Collins. 1999. Neoplasia. In "Robbins pathologic basis of disease". (ed. R.S. Cotran, V. Kumar and T. Collins) pp. 260-312. W.B. Saunders company, Philadelphia.

Coulombel L., M. Derycke, J.L. Villeval, C. Leonard, J. Breton-Gorius, M. Vial, P. Bourgeois and G. Tchernia. 1987. Characterization of the blast cell population in two neonates with Down's syndrome and transient myeloproliferative disorder. *Brit. J. of Haem.* **66:** 69-76.

Coulson A., J. Sulston, S. Brenner and J. Karn. 1986. Towards a physical map of the genome of the nematode *Caenorhabditis elegans*. *Proc. Natl. Acad. Sci. USA* **83:** 7821-7825.

Dallapiccola B., A. Pizutti and G. Novelli. 1996. How many breaks do we need to catch on 22q11. *Am. J. Hum. Genet.* **59:** 7-11.

Datson N.A., G.M. Duyk, G.-J. B. Van Ommen and J.T. Den Dunnen. 1994. Specific isolation of 3'-terminal exons of human genes by exon trapping. *Nucl. Acids Res.* **22:** 4148-4153.

Davisson M.T., C. Schmidt, R.H. Reeves, N.G. Irving, E.C. Akeson, B.S. Harris and R.T. Bronson. 1993. Segmental trisomy as a mouse model for studying Down syndrome. *Prog. Clin. Biol. Res.* **384:** 117-133.

Deaven L.L., M.A Van Dilla, M.F. Bartholdi, A.V. Carrano, L.S. Cram, J.C. Fuscoe, J.W. Gray, C.E. Hildebrand, R.K. Moyzis and J. Perlman. 1986. Construction of human chromosome-specific DNA libraries from flow-sorted chromosomes. *Cold Spring Harb. Symp. Quant. Biol.* **51:** 159-167.

De Belleroche J., R. Orrell and A. King. 1995. Familial amyotrophic lateral sclerosis/motor neuron disease (FALS): a review of current developments. *J. Med. Genet.* **32:** 841-847.

Debrauere H., C.G. Gendrel, S. Lechat and M. Dutreix. 1997. Differences and similarities between various tandem repeat sequences: minisatellites and microsatellites. *Biochimie* **79:** 577-586.

De Jong P.J., K. Yokobata, C. Chen, F. Lohman, L. Pederson, J. McNinch and M. Van Dilla. 1989. Human chromosome specific partial digest libraries in lambda and cosmid vectors. *Cytogenet. Cell Genet.* 51: 985.

Delabar J.M., D. Theophile, Z. Rahmani, Z. Chettouh, J.L. Blouin, M. Prieur, B. Noel and P.M. Sinet. 1993. Molecular mapping of twenty-four features of Down syndrome on chromosome 21. *Eur. J. Hum. Genet.* **1:** 114-124.

Dellapuca R. and V.S. Gallicchio. 1996. The regulation of phospholipase-A2 (PLA-2) by cytokines expressing hematopoietic growth-stimulating properties. *Proc. Soc. Exp. Biol. Med.* **212:** 174-184.

Deloukas P., G.D. Schuler, G. Gyapay, E.M. Beasley, C. Soderlund, P. Rodriguez-Tomé, L. Hui, T.C. Matise, K.B. McKusick, J.S. Beckmann *et al.* 1998. A physical map of 30,000 human genes. *Science* **282:** 744-746.

Deng H.X., A. Hentati, J.A. Tainer, Z. Iqbal, A. Cayabyab, W.Y. Hung, E.D. Getzoff, H. Ping, B. Herzfeldt, R.P. Roos *et al.* 1993. Amyotrophic lateral sclerosis and structural defects in Cu, Zn superoxide dismutase. *Science* **261:** 1047-1051.

Der C.J., T.G. Krontiris and G.M. Cooper. 1982. Transforming genes of the human bladder and lung carcinoma cell lines are homologous to the ras genes of Harvey and Kirsten sarcoma viruses. *Proc. Natl. Acad. Sci. USA* **79:** 3637-3640.

De Sario A., G. Roizes, N. Allegre and G. Bernardi. 1997. A compositional map of the cen-q21 region of human chromosome 21. *Gene* **194:** 107-113.

Dib C., S. Fauré, C. Fizames, D. Samson, N. Drouot, A. Vignal, P. Millasseau, S. Marc, J. Hazan, E. Seboun *et al.* 1996. A comprehensive genetic map of the human genome based on 5,264 microsatellite markers. *Nature* **380:** 152-154.

Down J.L.H. 1866. Observation on an ethnic classification of idiots. *Lond. Hosp. Rep.* **3:** 259-262.

Dutriaux A., J. Rossier, W. Van Hul, D. Nizetic, D. Theophille, J.M. Delabar, C. Van Broeckhoven and M.-C. Potier. 1994. Cloning and characterisation of a 135- to 500-kb region of homology on the long arm of human chromosome 21. *Genomics* **22:** 472-477.

Duyk G.M., S. Kim, R.M. Myers and D.R. Cox. 1990. Exon trapping: a genetic screen to identify candidate transcribed sequences in cloned mammalian genomic DNA. *Proc. Natl. Acad. Sci. USA* **87:** 8995-8999.

Easton D., D. Ford and J. Peto. 1993. Inherited susceptibility to breast cancer. *Cancer surveys* **18:** 95-113.

Eki T., M. Abe, K. Furuya, N. Fujishima, H. Kishida, A. Shiratori, K. Yokoyama, D. Lepaslier, D. Cohen and Y. Murakami. 1996. 1.8 Megabases fine physical map encompassing IFNAR and AML1 loci on human-chromosome 21q22.1 *DNA Sequence* **6:** 95-108.

Epstein C.J., D.R. Cox and L.B. Epstein. 1985. Mouse trisomy 16: an animal model for human trisomy 21 (Down's syndrome). *Ann. Proc. Natl. Acad. Sci. USA* **450:** 157-177.

Epstein C.J. 1986. In "The consequences of chromosome imbalance. Principles, mechanisms, and models" (eds. P.W. Barlow, P.B. Green and C.C. Wylie), pp. 256. New York, Cambridge University Press.

Epstein, C.J., J.R. Korenberg, G. Anneren, S.E. Antonarakis, S. Ayme, E. Courchesne, L.B. Epstein, A. Fowler, Y. Groner, J.L. Huret *et al.* 1991. Protocols to establish genotype-phenotype correlations in Down's syndrome. *Am. J. Hum. Genet.* **49:** 207-235.

Epstein C.J. 1995. Down syndrome (trisomy 21). In "The metabolic and molecular bases of inherited disease" (eds. C.R. Scriver, A.L. Beaudet, W.S. Sly and D. Valle), pp. 749-794. McGraw-Hill, New-York.

Fernandez C., D. Larhammar, B. Servenius, L. Rask and P.A. Peterson. 1986. Spontaneous insertions into cosmid vector: a warning. *Gene* **42:** 215-219.

Fickett J.W. 1996. Finding genes by computer: the state of the art. *Trends in Genetics* **12:** 316-320.

Fields C., M.D. Adams, O. White and J.C. Venter. 1994. How many genes in the human genome? *Nature Genetics* **7:** 345-346.

Ford C.E., K.W. Jones, P.E. Polani, J.C. de Almeira and J.H. Briggs. 1959. A sex-chromosome anomaly in a case of gonadal dysgenesis (Turner's syndrome). *Lancet* 711-713.

Ford D., D.F. Easton, D.T. Bishop, S.A. Narod, D.E. Goldgar and the breast cancer linkage consortium. 1994. Risks of cancer in BRCA1-mutation carriers. *Lancet* **343:** 692-695.

Frederick A., M. Rolfe and M.I. Chiu. 1998. The human UNP locus at 3p21.31 encodes two tissue-selective, cytoplasmic isoforms with deubiquitinating activity that have reduced expression in small cell lung carcinoma cell lines. *Oncogene* **16:** 153-165.

Games D., D. Adams, R. Alessandrini, R. Barbour, P. Berthelette, C. Blackwell, T. Carr, J. Clemens, T. Donaldson, F. Gillespie. 1995. Alzheimer-type neuropathology in transgenic mice overexpressing V717F β-amyloid precursor protein. *Nature* **373:** 523-527.

Gardiner K., B. Aissani and G. Bernardi. 1990. A compositional map of human chromosome 21. The *EMBO J.* **9:** 1853-1858.

Gardiner K., S. Graw, H. Ichikawa, M. Ohki, A. Joetham, P. Gervy, I. Chumakov and D. Patterson. 1995. YAC analysis and minimal tiling path construction for chromosome 21q. *Som. Cell and Mol. Genet.* **21:** 399-414.

Gardiner K. 1996. Base composition and gene distribution: critical patterns in mammalian genome organisation. *Trends in Genetics* **12:** 519-524.

Gibbs R.A. 1995. Pressing ahead with human genome sequencing. *Nature Genetics* **11:** 121-125.

Glenner G.G. and C.W. Wong. 1984a. Alzheimer's disease: initial report of the purification and characterisation of a novel cerebrovascular amyloid protein. *Biochem. Biophys. Res. Commun.* **122:** 885-890.

Glenner G.G. and C.W. Wong. 1984b. Alzheimer's disease and Down's syndrome: sharing of a unique cerebrovascular amyloid fibril protein. *Biochem. Biophys. Res. Commun.* **122:** 1131-1135.

Goate A.M., A.R. Haines, M.J. Owen, M. Farrell, L.A. James, L.A.C. Lai, M.J. Mullan, P. Roques, M.N. Rossor, R. Williamson *et al.* 1989. Predisposing locus for Alzheimer's Disease on chromosome 21. *Lancet* **8634:** 352-355.

Goate A., M.-C. Chartier-Harlin, M. Mullan, J. Brown, F. Crawford, L. Fidani, L. Giuffra, A. Haynes, N. Irving, L. James *et al.* 1991. Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer's disease. *Nature* **349:** 704-706.

Goedert M., C.M. Wischek, R.A. Crowther, J.E. Walker and A. Klug. 1988. Cloning and sequencing of the cDNA encoding a core protein of the paired helical filament of Alzheimer disease: identification as the microtubule-associated protein tau. *Proc. Natl. Acad. Sci. USA* **85:** 4051-4055.

Gottesman M.M., C.A. Hrycyna, P.V. Schoenlein, U.A. Germann and I. Pastan. 1995. Genetic analysis of the multidrug transporter. *Annu. Rev. Genet.* **29:** 607-649.

Grant C.E., E.U. Kurz, S.P.C. Cole and R.G. Deeley. 1997. Analysis of the intron-exon organization of the human multidrug-resistance protein gene (MRP) and alternative splicing of its mRNA. *Genomics* **45:** 368-378.

Graw S.L., K. Gardiner, K. Hall-Johnson, I. Hart, A. Joetham, K. Walton, D. Donaldson and D. Patterson. 1995. Molecular analysis and breakpoint defenition of a set of human chromosome 21 somatic cell hybrids. *Som. Cell and Mol. Genet.* **21:** 415-428.

Gray D.A., J. Inazawa, K. Gupta, A. Wong, R. Ueda and T. Takahashi. 1995. Elevated expression of Unph, a proto-oncogene at 3p21.3, in human lung tumors. *Oncogene* **10:** 2179-2183.

Green P. 1997. Against a whole-genome shotgun. *Genome Research* **7:** 410-417.

Groet J., J.H. Ives, A.P. South, P.R. Baptista, T.A. Jones, M-L Yaspo, H. Lehrach, M.-C. Potier, C. Van Broeckhoven and D. Nizetic. Bacterial contig map of the 21q11 region associated with Alzheimer's disease and abnormal myelopoiesis in Down syndrome. 1998. *Genome Research* **8:** 385-398.

Groet J., J.H. Ives, T.A. Jones, M. Danton, R.H. Flomen, D. Sheer, R. Hrascan, K. Pavelic and D. Nizetic. Narrowing of the region of allelic loss in squamous non-small cell lung carcinoma and cloning of a novel ubiquitin-specific protease gene from the deleted segment. *Genes, Chromosomes and Cancer* in press.

Hassold T.J., D. Chiu and J.A. Yamane. 1984. Parental origin of autosomal trisomies. *Ann. Hum. Genet.* 48: 129-144.

Hawley R.S., J.A. Frazier and R. Rasooly. 1994. Commentary: Separation anxiety: the etiology of nondisjunction in flies and people. *Hum. Mol. Genet.* **3:** 1521-1528.

Hendriks L. and C. Van Broeckhoven. 1996. The βA4 amyloid precursor protein gene and Alzheimer's disease. *Eur. J. Biochem.* **237:** 6-15.

Hernandez D. and Fisher E.M.C. 1999. Mouse autosomal trisomy: two's company, three's a crowd. *Trends in Genetics* **15:** 241-247.

Hernandez D., P.J. Mee, J.E. Martin, V.L.J. Tybulewicz and E.M.C. Fisher. 1999. Transchromosomal mouse embryonic stem cell lines and chimeric mice that contain freely segregating segments of human chromosome 21. *Hum. Mol. Genet.* **8:** 923-933.

Hiebert S.W., W. Sun, J.N. Davis, T. Golub, S. Shurtleff, A. Buijs, J.R. Downing, G. Grosveld, M.F. Roussel, D.G. Gilliland *et al.* 1996. The t(12;21) translocation converts AML-1B from an activator to a repressor of transcription. *Mol. Cell. Biol.* **16:** 1349-1355.

Hildmann T., X. Kong, J. O'Brien, L. Riesselman, H.M. Christensen, E. Dagand, H. Lehrach and M.L. Yaspo. 1999. A contiguous 3-Mb sequence-ready map in the S3-MX region on 21q22.2 based on high-throughput nonisotopic library screenings. *Genome Research* **9:** 360-372.

Hohn B. and J. Collins. 1980. A small cosmid for efficient cloning of large DNA fragments. *Gene* **11:** 291-298.

Holzman D.M., D. Santucci, J. Kilbridge, J. Chua-Couzens, D.J. Fontana, S.E. Daniels, R.M. Johnson, K. Chen, Y. Sun, E. Carlson *et al.* 1996. Developmental abnormalities and age-related neurodegeneration in a mouse model of Down syndrome. *Proc. Natl. Acad. Sci. USA* **93:** 13333-13338.

Huang Y., R.T. Baker and J.A. Fischer-Vize. 1995. Control of cell fate by a deubiquitinating enzyme encoded by the *fat facets* gene. *Science* **270:** 1828-1831.

Hubert R.S., S. Mitchell, X.N. Chen, K. Ekmekji, C. Gadomski, Z.G. Zun, D. Noya, U.J. Kim, C.R. Chen, H. Shizuya, M. Simon, P.J. de Jong and J.R. Korenberg. 1997. BAC and PAC contigs covering 3.5 Mb of the Down syndrome congenital heart disease region between D21S55 and MX1 on chromosome 21. *Genomics* **41:** 218-226.

Hutton M. and J. Hardy. 1997. The presenilins and Alzheimer's disease. *Hum. Mol. Genet.* **6:** 1639-1646.

Ichikawa H., F. Hosoda, Y. Arai, K. Schimizu, M. Ohira and M. Ohki. 1993. A *Not* I restriction map of the entire long arm of human chromosome 21. *Nature Genetics* **4:** 361-365.

Ioannou P.A., C.T. Amemiya, J. Garnes, P.M. Kroisel, H. Shizuya, C. Chen, M.A. Batzer and P.J. de Jong. 1994. A new bacteriophage P1-derived vector for the propagation of large human DNA fragments. *Nature Genetics* **6:** 84-89.

Iselius L., P. Jacobs and N. Morton. 1990. Leukaemia and transient leukaemia in Down syndrome. *Human Genetics* **85:** 477-485.

Jacobs P., A. Baikie, W. Court-Brown and J. Strong. 1959. The somatic chromosomes in mongolism. *Lancet* 710-711.

Jacobs P., and J. Strong. 1959. A case of human intersexuality having a possible XXY sex-determining mechanism. *Nature* **183:** 302-303.

Kamb A., C. Wang, A. Thomas, B.S. De Hoff, F.H. Norris, K. Richardson, J. Rine, M.H. Skolnick and P.R. Posteck Jr. 1995. Software trapping: a strategy for finding genes in large genomic regions. *Comput. Biomed. Res.* **2:** 140-153.

Kan Y.W. and A.M. Dozy. 1978. Polymorphism of DNA sequence adjacent to human β-globin structural gene: relationship to sickle mutation. *Proc. Natl. Acad. Sci USA* **75:** 5631-5635.

Kang J., H. Lemaire, A. Unterbeck, J.M. Salbaum, C.L. Masters, K. Grzeschik, G. Multhaup, K. Beyreuter and B. Müller-hill. 1987. The precursor of Alzheimer's disease amyloid A4 protein resembles a cell-surface receptor. *Nature* **325:** 733-736.

Katsanis, N., J.H. Ives, J. Groet, D. Nizetic and E.M.C. Fisher (1998). Localization of receptor interacting protein 140 (RIP 140) within 100kb of D21S13 on 21q11, a gene-poor region of the human genome. *Human Genetics* **102:** 221-223.

Kehoe P., F. Wavrant-De Vrieze, R. Crook, W.S. Wu, P. Holmans, I. Fenton, G. Spurlock, N. Norton, H. Williams, N. Williams *et al.* 1999. A full genome scan for late onset Alzheimer's disease. *Hum. Mol. Genet.* **8:** 237-245.

Kesslak J.P., S.F. Nagata, I. Lott and O. Nalcioglu. 1994. Magnetic resonance imaging analysis of age related changes in the brains of individuals with Down's syndrome. *Neurology* **44:** 1039-1045.

Knudson A.G. 1971. Mutation and cancer: statistical study of retinoblastoma. *Proc. Natl. Acad. Sci USA* **68:** 820-823.

Koehler K.E., C.L. Boulton, H.E. Collins, R.F. French, K.C. Herman, S.M. Lacefield, L.D. Madden, C.D. Schuetz and R.S. Hawley. 1996. Spontaneous X chromosome MI and MII nondisjunction events in Drosophila Melanogaster oocytes have different recombinational histories. *Nature Genetics* **14:** 406-414.

Kohno T., M. Kawanishi, S. Matsuda, H. Ichikawa, M. Takada, M. Ohki, T. Yamamoto and J. Yokota. 1998. Homozygous deletion and frequent allelic loss of the 21q11.1-q21.1 region including the ANA gene in human lung carcinoma. *Genes, Chromosomes and Cancer* **21:** 236-243.

Kok K, J. Osinga, B. Carritt, M.B. Bavis, A.H. van der Hout, A.Y. van der Veen, R.M. Landsvater, L.F.M.H. de Leij, H.H. Berendsen, P.E. Postmus *et al.* 1987. Deletion of a DNA sequence at the chromosomal region 3p21 in all major types of lung cancer. *Nature* **330:** 578-581.

Korenberg J.R., M.L. Croyle and D.R. Cox. 1987. Isolation and regional mapping of DNA sequences unique to human chromosome 21. *Am. J. Hum. Genet.* **41:** 963-978.

Korenberg J.R., C. Bradley and C.M. Disteche. 1992. Down syndrome: molecular mapping of the congenital heart disease and duodenal stenosis. *Am. J. Hum. Genet.* **50:** 294-302.

Korenberg J.R., X.-N. Chen, R. Schipper, Z. Zun, R. Gonsky, S. Gerwehr, N. Carpenter, C. Daumer, P. Dignan, C. Disteche *et al.* 1994. Down syndrome phenotypes: The consequences of chromosomal imbalance. *Proc. Natl. Acad. Sci. USA* **91:** 4997-5001.

Korenberg J.R., X.-N. Chen, S. Mitchell, S. Fannin, S. Gerwehr, D. Cohen and I. Chumakov. 1995. A high fidelity physical map of human chromosome 21q in yeast artificial chromosomes. *Genome Research* **5:** 427-443.

Korenberg J.R. 1995. Mental modelling. *Nature Genetics* **11:** 109-111.

Korenberg J.R., J. Aaltonen, C. Brahe, D. Cabin, N. Creau, J.M. Delabar, J. Doering, K. Gardiner, R.S. Hubert, J. Ives, *et al.* 1996. Report and abstracts of the sixth international workshop on human chromosome 21 mapping 1996. Cold Spring Harbour, New York, USA. May 6-8, 1996. *Cytogenet. Cell Genet.* **79:** 21-52.

Korn B., Z. Sedlacek, A. Manca, P. Kioschis, D. Konecki, H. Lehrach and A.M. Poustka. 1992. A strategy for the selection of transcribed sequences in the Xq28 region. *Hum. Mol. Genet.* **1:** 235-242.

Kozak M. 1996. Interpreting cDNA sequences: some insights from studies on translation. *Mammalian Genome* **7:** 563-574.

Krizman D.B. and S.M. Berget. 1993. Efficient selection of 3'-terminal exons from vertebrate DNA. *Nucl. Acids Res.* **21:** 5198-5202.

Lafraniere R.G., P.J. de Jong and G.A. Rouleau. 1995. A 405-kb cosmid contig and HindIII restriction map of the progressive myoclonus epilypsy type 1 (EPM1) candidate region in 21q22.3. *Genomics* **29:** 288-290.

Lalioti M.D., H.S. Scott, C. Buresi, C. Rossier, A. Bottani, M.A. Morris, A. Malafosse and S.E. Antonarakis. 1997. Dodecamer repeat expansion in cystatin B gene in progressive myoclonus epilypsy. *Nature* **386:** 847-851.

Lam Y.A., W. Xu, G.N. DeMartino and R.E. Cohen. 1997. Editing of ubiquitin conjugates by an isopeptidase in the 26S proteasome. *Nature* **385:** 737-740.

Lamb B.T., S.S. Sisodia, A.M. Lawler, H.H. Slunt, C.A. Kitt, W.G. Kearns, P.L. Pearson, D.L. Price and J.D. Gearhart. 1993. Introduction and expression of the 400 kilobase precursor amyloid protein gene in transgenic mice. *Nature Genetics* **5:** 22-29.

Lamb N.E., S.B. Freeman, A. Savage-Austin, D. Pettay, L. Taft, J. Hersey, Y. Gu, J. Shen, D. Saker, K.M. May *et al.* 1996. Susceptible chiasmate configurations of chromosome 21 predispose to non-disjunction in both maternal meiosis I and meiosis II. *Nature Genetics* **14:** 400-405.

Lander E.S. and N.J. Schork. 1994. Genetic dissection of complex traits. *Science* **265:** 2037-2048.

Lane D.P. 1992. P53, guardian of the genome. *Science* **358:** 15-16.

Larin Z., A.P. Monaco and H. Lehrach. 1991. Yeast artificial chromosome libraries containing large inserts from mouse and human DNA. *Proc. Natl. Acad. Sci. USA* **88:** 4123-4127.

Lee W.-H., R. Bookstein, F. Hong, L.-J. Young, J.-Y. Shew and E.Y.-H.P. Lee. 1987. Human retinoblastoma susceptibility gene: cloning, identification, and sequencing. *Science* **235:** 1394-1399.

Leheshjoki A.-E., M. Koskiniemi, P. Sistonen, J. Miao, J. Hästbacka, R. Norio and A. de la Chapelle. 1991. Localization of a gene for progressive myoclonus epilepsy to chromosome 21q22. *Proc. Natl. Acad. Sci. USA* **88:** 3696-3699.

Leheshjoki A.-E., M. Koskiniemi, R. Norio, S. Tirrito, P. Sistonen, E. Lander, and A. de la Chapelle. 1993. Localisation of the EPM1 gene for progressive myoclonus epilepsy on chromosome 21: linkage disequilibrium allows high resolution mapping. *Hum. Mol. Genet.* **2:** 1229-1234.

Lejeune J., M. Gautier and R. Turpin. 1959. Etudes des chromosomes somatique de neuf enfants mongoliens. *C.R. Acad. Sci., Paris* **248:** 1721-1722.

Lennon G., C. Auffray, M. Polymeropoulos and M.B. Soares. 1996. The I.M.A.G.E. Consortium: an integrated molecular analysis of genomes and their expression. *Genomics* **33:** 151-152.

Levy-Lahad E., E.M. Wijsman, E. Nemens, L. Anderson, K.A.B. Goddart, J.L. Weber, T.D. Bird and Schellenberg G.D. 1995a. A familial Alzheimer's disease locus on chromosome 1. *Science* **269:** 970-973.

Levy-Lahad E., W. Wasco, P. Poorkaj, D.M. Romano, J. Oshima, W.H. Pettingell, C. Yu, P.D. Jondro, S.D. Schmidt, K. Wang *et al.* 1995b. Candidate gene for the chromosome 1 familial Alzheimer's disease locus. *Science* **269:** 973-977.

Lindsay S. and A.P. Bird. 1987. Use of restriction enzymes to detect potential gene sequences in mammalian DNA. *Nature* **327:** 336-338.

Litt M. and J.A. Luty. 1989. A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. *Am. J. Hum. Genet.* **44:** 397-401.

Lovett M., J. Kere and L.M. Hinton. 1991. Direct selection: a method for the isolation of cDNAs encoded by large genomic regions. *Proc. Natl. Acad. Sci. USA* **88:** 9628-9632.

Mahairas G.G., J.C. Wallace, K. Smith, S. Swartzell, T. Holzman, A. Keller, R. Shaker, J. Furlong, J. Young, S. Zhao *et al.* 1999. Sequence-tagged connectors: a sequence approach to mapping and scanning the human genome. *Proc. Natl. Acad. Sci. USA* **96:** 9739-9744.

Mancini D., S. Shiva, P. Ainsworth and D. Rodenhiser. 1997. Constitutively methylated CpG dinucleotides as mutation hot spots in the Retinoblastoma gene (RB 1). *Am. J. Hum. Genet.* **61:** 80-87.

McCormick M.K., A. Schinzel, M.B. Petersen, G. Stetten, D.J. Driscoll, E.S. Cantu, L. Tranebjerg, M. Mikkelson, P.C. Watkins and S.E. Antonarakis. 1989. Molecular genetic approach to the characterization of the "Down syndrome region" of chromosome 21. *Genomics* **5:** 325-331.

McInnes M.G., A. Chakravarti, J. Blaschak, M.B. Petersen, V. Sharma, D. Avramopoulis, J.L. Blouin, U. Konig, C. Brahe, T.C. Matise *et al.* 1993. A linkage map of human chromosome 21: 43 PCR markers at average intervals of 2.5 cM. *Genomics* **16:** 562-571.

Miyoshi H., K. Shimizu, N. Maseki, Y. Kanako and M. Ohki. 1991. t(8;21) breakpoints on chromosome 21 in acute myeloid leukaemia are clustered within a limited region of a single gene, AML1. *Proc. Natl. Acad. Sci. USA* **88:** 10431-10434.

Monaco A.P., R.L. Neve, C. Colletti-Feener, C.J. Bertolson, D.M. Kurnit and L.M. Kunkel. 1986. Isolation of candidate cDNAs for portions of the Duchenne muscular dystrophy gene. *Nature* **323:** 646-650.

Mori H., J. Kundo and Y. Ihara. 1987. Ubiquitin is a component of paired helical filaments in Alzheimer's disease. *Science* **235:** 1641-1644.

Morishima-Kawashima M., M. Hasegawa, K. Takio, M. Suzuki, K. Titani and Y. Ihara. 1993. Ubiquitin is conjugated with amino-terminally processed tau in paired helical filaments. *Neuron* **10:** 1151-1160.

Mullan M., F. Crawford, K. Axelman, H. Houlden, L. Lilius, B. Winblad and L. Lannfelt. 1992. A pathogenic mutation for probable Alzheimer's disease in the APP gene at the N-terminus of β-amyloid. *Nature Genetics* **1:** 345-347.

Murrell J., M. Farlow, B. Ghetti and M.D. Benson. 1991. A mutation in the amyloid precursor protein associated with hereditary Alzheimer's disease. *Science* **254:** 97-99.

Nagai Y., J. Aoki, T. Sato, K. Amano, Y. Matsuda, H. Arai and K. Inoue. 1999. An alternative splicing form of phosphatidylserine-specific phospholipase A_1 that exhibits lysophosphatidylserine-specific lysophospholipase activity in humans. *J. Biol. Chem.* **274:** 11053-11059.

Nagamine K., P. Peterson, H.S. Scott, J. Kudoh, S. Minoshima, M. Heino, K. J. E. Krohn, M.D. Lalioti, P.E. Mullis, S.E. Antonarakis *et al.* 1997. Positional cloning of the APECED gene. *Nature Genetics* **17:** 393-398.

Nakamura Y., M. Leppert, P. O'Connell, R. Wolff, T. Holm, M. Culver, C. Martin, E. Fujimoto, M. Hoff, E. Kumlin and R. White. 1987. Variable number of tandem repeat (VNTR) markers for human gene mapping. *Science* **235:** 1616-1622.

Nakamura T., J. Hillova, R. Mariage-Samson, M. Onno, K. Huebner, L.A. Cannizzaro, L. Boghossian-Sell, C.M. Croce and M. Hill. 1992. A novel transcriptional unit of the tre oncogene widely expressed in human cancer cells. *Oncogene* **7:** 733-741.

Nelson D.L., S.A. Ledbetter, L. Corbo, M.F. Victoria, R. Ramirez-solis, T.D. Webster, D.H. Ledbetter and C.T. Caskey. 1989. *Alu* polymerase chain reaction: a method for rapid isolation of human-specific sequences from complex DNA sources. *Proc Natl. Acad. Sci. USA* **86:** 6686-6690.

Niebuhr E. 1974. Down's syndrome, the possibility of a pathogenetic segment on chromosome no. 21. *Humangenetik* **21:** 99-101.

Niebuhr E. 1978. The Cri-du-chat syndrome, epidemiology, cytogenetics and clinical features. *Human Genetics* **44:** 227-275.

Niikawa N., D. Han-Xiang, K. Abe, N. Harada, T. Okada, H. Tsuchiya, I. Akaboshi, I. Matsuda, Y. Fukushima, Y. Kaneko *et al.* 1991. Possible mapping of the gene for transient myeloproliferative syndrome at 21q11.2. *Human Genetics* **87:** 561-566.

Nizetic D., G. Zehetner, A.P. Monaco, L. Gellen, B.D. Young and H. Lehrach. 1991a. Construction, arraying, and high-density screening of large insert libraries of human chromosomes X and 21: their potential use as reference libraries. *Proc. Natl. Acad. Sci. USA* **88:** 3233-3237.

Nizetic D., R. Drmanac and H. Lehrach. 1991b. An improved bacterial colony lysis procedure enables direct DNA hybridisation using short (10, 11 bases) oligonucleotides to cosmids. *Nucl. Acids Res.* **19:** 182.

Nizetic D., L. Gellen, R. Hamvas, R. Mott, A. Grigoriev, R. Vatcheva, G. Zehetner, M.L. Yaspo, A. Dutriaux, C. Lopes *et al.* 1994. An integrated YAC-overlap and 'cosmid pocket' map of the human chromosome 21. *Hum. Mol. Genetics* **3:** 759-770.

Nizetic D. and H. Lehrach 1995. In "DNA Cloning 3, a practical approach" (eds D.M. Glover and B.D. Hames). pp 49-79. Oxford University Press Inc., New York.

Ochman H., A.S. Gerler and D.L. Hartl. 1988. Genetic applications of an inverse polymerase chain reaction. *Genetics* **120:** 621-623.

Ohira M., H. Ichikawa, E. Suzuki, M. Iwaki, K. Suzuki, F. Saitoohara, T. Ikeuchi, I. Chumakov, H. Tanahashi, K. Tashiro, Y. Sakaki and M. Ohki. 1996. A 1.6 Mb P1-based physical map of the Down syndrome region on chromosome 21. *Genomics* **33:** 65-74.

Ohira M., N. Seki, T. Nagase, E. Suzuki, N. Nomura, O. Ohara, M. Hattori, Y. Sakaki , T. Eki, Y. Murakami *et al.* 1997. Gene identification in 1.6-Mb region of the Down syndrome region on chromosome 21. *Genome Research* **7:** 47-58.

Ohta T., M. Nakano, T. Tsujita, K. Abe, K. Osoegawa, T. Yamagata, K.-I. Yoshiura, Y. Jinno, E. Soeda, Y. Nakamura and N. Niikawa. 1996. Isolation of a cosmid clone corresponding to an inv(21) breakpoint of a patient with transient abnormal myelopoiesis. *Am. J. Hum. Genet.* **58:** 544-550.

Olson, M., L. Hood, C. Cantor and D. Botstein. 1989. A common language for physical mapping of the human genome. *Science* **245:** 1434-1435.

Osoegawa K., R. Susukida, S. Okano, S. Minoshima, N. Shimizu, P.J. de Jong, J. Groet, J. Ives, H. Lehrach, D. Nizetic and E. Soeda. 1996. An integrated map with cosmid/PAC contigs of a 4-Mb Down syndrome critical region. *Genomics* **32:** 375-387.

Otterson G.A., G.C. Flynn, R.A. Kratzke, A. Coxon, P.G. Johnston and F.J. Kaye. 1994. Stch encodes the 'ATPase core' of a microsomal stress 70 protein. *The EMBO J.* **13:** 1216-1225.

Papa F.R. and M. Hochstrasser. 1993. The yeast DOA4 gene encodes a deubiquitinating enzyme related to a product of the human tre-2 oncogene. *Nature* **366:** 313-319.

Parada L.F., C.J. Tabin, C. Shih and R.A. Weinberg. 1982. Human EJ bladder carcinoma oncogene is homologue of Harvey sarcoma virus ras. *Nature* **297:** 474-478.

Parimoo S., S.R. Patanjali, H. Shukla, D.D. Chaplin and S.M. Weissman. 1991. cDNA selection: efficient PCR approach for the selection of cDNAs encoded in large chromosomal DNA fragments. *Proc. Natl. Acad. Sci. USA* **88:** 9623-9627.

Patil N., A. Peterson, A. Rothman, P.J. de Jong, R.M. Mayers and D.R. Cox. 1994. A high resolution physical map of 2.5 Mbp of the Down syndrome region on chromosome 21. *Hum. Mol. Genet.* **3:** 1811-1817.

Pearson B.E. and T.K. Choi. 1993. Expression of the human β-amyloid precursor protein gene from a yeast artificial chromosome in transgenic mice. *Proc. Natl. Acad. Sci. USA* **90:** 10578-10582.

Pennacchio L.A., A.-E. Leheshjoki, N.E. Stone, V.L. Willour, K. Virtaneva, J. Miao, E. D'Amato, L. Ramirez, M. Faham, M. Koskiniemi *et al.* 1996. Mutations in the gene encoding cystatine B in progressive myoclonus epilypsy (EPM1). *Science* **271:** 1731-1734.

Pericak-Vance M.A., J.L.Bebout, P.C. Gaskell Jr, L.H. Yamaoka, W.-Y. Hung, M.J. Alberts, A.P. Walker, R.J. Bartlett, C.A. Haynes, K.A. Welsh *et al.* 1991. Linkage studies in familial Alzheimer disease: evidence for chromosome 19 linkage. *Am. J. Hum. Genet.* **48:** 1034-1050.

Perry G., R. Friemann, G. Shaw and V. Chau. 1987. Ubiquitin is detected in neurofibrillary tangles and senile plaque neurites of Alzheimer's disease brains. *Proc. Natl. Acad. Sci. USA* **84:** 3033-3036.

Peterson A., N. Patil, C. Robbins, L. Wang, D.R. Cox and R.M. Myers. 1994. A transcript map of the Down syndrome critical region on chromosome 21. *Hum. Mol. Genet.* **3:** 1735-1742.

Potier M.-C., A. Dutriaux and R. Reeves. 1996. Use of YAC fragmentation to delimit a duplicated region on human chromosome 21. *Mammalian Genome* **7:** 85-88.

Potier M.-C., A. Dutriaux, R. Orti, J. Groet, N. Gibelin, G. Karadima, G. Lutfalla, A. Lynn, C. Van Broeckhoven, A. Chakravarti *et al.* 1998. Two sequence-ready contigs spanning the two copies of a 200-kb duplication on human 21q: partial sequence and polymorphisms. *Genomics* **51:** 417-426.

Prasher V.P., M.J. Farrer, A.M. Kessling, E.M. Fisher, R.J. West, P.C. Barber and A.C. Butler. 1998. Molecular mapping of Alzheimer-type dementia in Down's syndrome. *Ann. Neurol.* **3:** 380-383.

Puffenberger E.G., E.R. Kauffman, S. Bolk, T.C. Matise, S.S. Washington, M. Angrist, J. Weissenbach, K.L. Garver, M. Mascari, R. Ladda *et al.* 1994. Identity-by-descent and association mapping of a recessive gene for Hirschsprung disease on chromosome 13q22. *Hum. Mol. Genet.* **3:** 1217-1225.

Rabbitts P., J. Bergh, J. Douglas, F. Collins and J. Waters. 1990. A submicroscopic homozygous deletion at the D3S3 locus in a cell line isolated from a small cell lung carcinoma. *Genes, Chromosomes & Cancer* **2:** 231-238.

Rahmani Z., J.-L. Blouin, N. Creau-Goldberg, P.C. Watkins, J.-F. Mattei, M. Poissonnier, M. Prieur, Z. Chettouh, A. Nicole, A. Aurias *et al.* 1989. Critical role of the D21S55 region on chromosome 21 in the pathogenesis of Down syndrome. *Proc. Natl. Acad. Sci. USA* **86:** 5958-5962.

Reeves R.H., N.G. Irving, T.H. Moran, A. Wohn, C. Kitt, S.S. Sisodia, C. Schmidt, R.T. Bronson and M.T. Davisson. 1995. A mouse model for Down syndrome exhibits learning and behaviour deficits. *Nature Genetics* **11:** 177-184.

Reeves R.H., E. Rue, J. Yu and F.-T. Kao. 1998. Stch maps to mouse chromosome 16, extending the conserved synteny with human chromosome 21. *Genomics* **49:** 156-157.

Riley J., R. Butler, R. Finniear, D. Jenner, S. Powell, R. Anand, J.C. Smith and A.F. Markham. 1990. A novel, rapid method for the isolation of terminal sequences from Yeast Artificial Chromosome (YAC) clones. *Nucl. Acids Res.* **18:** 2887-2890.

Rogaev E.I., R. Sherrington R., E.A. Rogaeva, G. Levesque, M. Ikeda, Y. Liang, H. Chi, C. Lin, K. Holman, T. Tsuda *et al.* 1995. Familial Alzheimer's disease in kindreds with missense mutations in a gene on chromosome 1 related to the Alzheimer's disease type 3 gene. *Nature* **376:** 775-778.

Rosen D.R, T. Siddique, D. Patterson, D.A. Figlewicz, P. Sapp, A. Hentati, D. Donaldson, J. Goto, J.P. O'Regan, H.-X. Deng *et al.* 1993. Mutations in the Cu/Zn superoxide dismutase gene are associated with familial amyotrophic lateral sclerosis. *Nature* **362:** 59-63.

Ross M.T., D. Nizetic, C. Nguyen, C. Knights, R. Vatcheva, N. Burden, C. Douglas, G. Zehetner, D.C. Ward, A. Baldini and H. Lehrach. 1992. Selection of a human chromosome 21 enriched YAC library using a chromosome-specific composite probe. *Nature Genetics* **1:** 284-290.

Sago H., E.J. Carlson, D.J. Smith, J. Kilbridge, E.M. Rubin, W.C. Mobley, C.J. Epstein and T.-T. Huang. 1998. TsCje, a partial trisomy 16 mouse model for Down syndrome, exhibits learning and behavioral abnormalities. *Proc. Natl. Acad. Sci. USA* **95:** 6256-6261.

Salgia R. and A.T. Skarin. 1998. Molecular abnormalities in lung cancer. *J. of Clin. Oncol.* **16:** 1207-1217.

Sato S., Y. Nakamura and E. Tsuchiya. 1994. Difference of allelotype between squamous cell carcinoma and adenocarcinoma of the lung. *Cancer Research* **54:** 5652-5655.

Sato T., J. Aoki, Y. Nagai, N. Dohmae, K. Takio, T. Doi, H. Arai and K. Inoue. 1997. Serine phospholipid-specific phospholipase A that is secreted from activated platelets. *J. Biol. Chem.* **272:** 2192-2198.

Schneider N. and B. Dawson. 1999. Genetic disorders. In "Robbins pathologic basis of disease" (ed. R.S. Cotran, V. Kumar and T. Collins) pp. 139-187. W.B. Saunders company, Philadelphia.

Schuler G.D., M.S. Boguski, E.A. Stewart, L.D. Stein, G. Gyapay, K. Rice, R.E. White, P. Rodriguez-tomé, A. Aggerwal, E. Bajorek *et al.* 1996. A gene map of the human genome. *Science* **274:** 540-546.

Schwartz A.L. and A. Ciechanover. 1999. The ubiquitin-proteasome pathway and pathogenisis of human disease. *Annu. Rev. Med.* **50:** 57-74.

Seubert P., T. Oltersdorf, M.G. Lee, R. Barbour, C. Blomquist, D.L. Davis, K. Bryant, L.C. Fritz, D. Galasko, L.J. Thal *et al.* 1993. Secretion of β-amyloid precursor protein cleaved at the amino terminus of the β-amyloid peptide. *Nature* **361:** 260-263.

Shen J.J., B.J. Williams, A. Zipurski, J. Doyle, S.L. Shermann, P.A. Jacobs, A.L. Shugar, S.W. Soukop and T.J. Hassold. 1995. Cytogenetic and molecular studies of Down syndrome individuals with leukemia. *Am. J. Hum. Genet.* **56:** 915-925.

Sherman S.L., N. Takaesu, S.B. Freeman, M. Grantham, C. Phillips, R.D. Blackston, P.A. Jacobs, A.E. Cockwell, V. Freeman, I. Uchida *et al.* 1991. Trisomy 21: association between reduced recombination and nondisjunction. *Am. J. of Hum. Genet.* **49:** 608-620.

Sherman S.L., M.B. Petersen, S.B. Freeman, J. Hersey, D. Pettay, L. Taft, M. Frantzen, M. Mikkelsen and T.J. Hassold. 1994. Non-disjunction of chromosome 21 in maternal meiosis I: evidence for a maternal age-dependent mechanism involving reduced recombination. *Hum. Mol. Genet.* **3:** 1529-1535.

Sherrington R., E.I. Rogaev, Y. Liang, E.A. Rogaeva, G. Levesque, M. Ikeda, H. Chi, C. Lin, G. Li, K. Holman *et al.* 1995. Cloning of a gene bearing missense mutations in early-onset familial Alzheimer's disease. *Nature* **375:** 754-760.

Shimizu, N., S.E. Antonarakis, C. Van Broeckhoven, D. Patterson, K. Gardiner, D. Nizetic, N. Creau, J.-M. Delabar, J. Korenberg, R. Reeves *et al.* 1995. Report of the fifth international workshop on human chromosome 21 mapping 1994. *Cytogenet. Cell Genet.* **70:** 157-182.

Shizuya H., B. Birren, U.-J. Kim, V. Mancino, T. Slepak, Y. Tachiiri and M. Simon. 1992. Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. *Proc. Natl. Acad. Sci. USA* **89:** 8794-8797.

Siddique T., D.A. Figlewics, M.A. Pericak-Vance, J.L. Haines, G. Rouleau, A.J. Jeffers, P. Sapp, W.Y. Hung, J. Bebout, D. McKenna-Yasek *et al.* 1991. Lineage of a gene causing familial amyotrophic lateral sclerosis to chromosome 21 and evidence of genetic locus heterogeneity. *N. Eng. J. Med.* **324:** 1381-1384.

Siddique T. and H.-X. Deng. 1996. Genetics of amyotrophic lateral sclerosis. *Hum. Mol. Genet.* **5:** 1465-1470.

Smith D.J., Y. Zhu, J. Zhang, J.-F. Cheng and E.M. Rubin. 1995. Construction of a panel of transgenic mice containing a contiguous 2-Mb set of YAC/P1 clones from human chromosome 21q22.2. *Genomics* **27:** 425-434.

Smith D.J., M.E. Stevens, S.P. Sudanagunta, R.T. Bronson, M. Makhinson, A.M. Watabe, T.J. O'Dell, J. Fung, H.-U.G. Weier, J.-F. Cheng and E.M. Rubin. 1997. Functional screening of 2 Mb of human chromosome 21q22.2 in transgenic mice implicates minibrain in learning defects associated with Down syndrome. *Nature Genetics* **16:** 28-36.

Spaventi R., L. Pecur, K. Pavelic, Z.P. Pavelic, S. Spaventi and P.J. Stambrook. 1994. Human Tumour Bank in Croatia: a possible model for a small bank as part of the future European Tumour Bank network. *Eur. J. Cancer* **30A:** 419.

Speck A.A., T. Stacy, Q. Wang, T. North, T.-L. Gu, J. Miller, M. Binder and M. Marin-Padilla. 1999. Core-binding factor: a central player in hematopoiesis and leukemia. *Cancer Research* **59:** 1789s-1793s.

Spector D.H., H.E. Varmus and J.M.Bishop. 1978. Nucleotide sequences related to the transforming gene of avian sarcoma virus are present in DNA of uninfected vertebrates. *Proc. Natl. Acad. Sci. USA* **75:** 4102-4106.

Stehelin D., H.E. Varmus, J.M. Bishop and P.K. Vogt. 1976. DNA related to the transforming gene(s) of avian sarcoma viruses is present in normal avian DNA. *Nature* **260:** 170-173.

Stewart G.D., P. Harris, J. Galt and M.A. Ferguson-Smith. 1985. Cloned DNA probes regionally mapped to human chromosome 21 and their use in determining the origin of non-disjunction. *Nucl. Acids Res.* **13:** 4125-4132.

St George-Hyslop P.H., R.E. Tanzi, R.J. Polinsky, J.L. Haines, L. Nee, P.C. Watkins, R.H. Myers, R.G. Feldman, D. Pollen, D. Drachman *et al.* 1987. The genetic defect causing familial Alzheimer's Disease maps on chromosome 21. *Science* **235:** 885-890.

St George-Hyslop P., J. Haines, E. Rogaev, M. Mortilla, G. Vaula, M. Pericak-Vance, J-F Foncin, M. Montesi, A. Bruni, S. Sorbi *et al.* 1992. Genetic evidence for a novel familial Alzheimer's disease locus on chromosome 14. *Nature Genetics* **2:** 330-334.

Stiller C. and L.M. Kinnier Wilson. 1981. Down syndrome and leukemia. *Lancet* 1343.

Stone N.E., J.-B. Fan, V. Willour, L.A. Pennacchio, J.A. Warrington, A. Hu, A. de la Chapelle, A.-E. Leheshjoki, D.R. Cox and R.M. Myers. 1996. Construction of a 750-kb bacterial clone contig and restriction map in the region of human chromosome 21 containing the progressive myoclonus epilepsy gene. *Genome Research* **6:** 218-225.

Sturchler-Pierrat C., D. Abramowski, M. Duke, K.-H. Wiederhold, C. Mistl, S. Rothacher, B. Ledermann, K. Burki, P. Frey, P.A. Paganetti. 1997. Two amyloid precursor protein transgenic mouse models with Alzheimer disease-like pathology. *Proc. Natl. Acad. Sci. USA* **94:** 13287-13292.

Stuyver L., G. Van Camp, A. Van De Voorde, C. Van Broeckhoven and H. Van Heuverswyn. 1991. Isolation of chromosome 21 encoded expressed sequences. Human gene mapping 11. 1991. *Cytogenet Cell Genet.* **58:** 2039.

Tassone F., H. Xu, H. Burkin, S. Weissman and K. Gardiner. 1995. cDNA selection from 10Mb of chromosome 21 DNA: efficiency in transcriptional mapping and reflections of genome organization. *Hum. Mol. Genet.* **4:** 1509-1518.

Taylor S.A.M., R.G. Snell, A. Buckler, C. Ambrose, M. Duyao, D. Church, C.S. Lin, M. Altherr, G.P. Bates *et al.* 1992. Cloning of the α-adducin gene from the Huntington's disease candidate region of chromosome 4 by exon amplification. *Nature Genetics* **2:** 223-227.

Testa J.R., J.M. Siegfried, Z. Liu, J.D. Hunt, M.M. Feder, S. Litwin, J.-Y. Zhou, T. Taguchi and S.M. Keller. 1994. Cytogenetic analysis of 63 non-small cell lung carcinomas: recurrant chromosome alterations amid frequent and widespread genomic upheaval. *Genes, chromosomes & Cancer* **11:** 178-194.

The Finnish-German APECED Consortium. 1997. An autoimmune disease, APECED, caused by mutations in a novel gene featuring two PHD-type zinc-finger domains. *Nature Genetics* **17:** 399-403.

The Huntington's disease collaborative research group. 1993. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* **72:** 971-983.

The Sanger Centre and the Washington University Genome Sequencing Center. 1998. Toward a complete human genome sequence. *Genome Research* **8:** 1097-1108.

Trowell H.E., A. Nagy, B. Vissel and K.H.A. Choo. 1993. Long-range analysis of the centromeric regions of human chromosomes 13, 14 and 21: identification of a narrow domain containing two key centromeric DNA elements. *Hum. Mol. Genet.* **2:** 1639-1649.

Uberbacher E.C. and R.J. Mural. 1991. Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. *Proc. Natl. Acad. Sci. USA* **88:** 11261-11265.

Van Broeckhoven C., H. Backhovens, M. Cruts, G. de Winter, M. Bruyland, P. Cras and J-J Martin. 1992. Mapping of a gene predisposing to early-onset Alzheimer's Disease to chromosome 14q24.3. *Nature Genetics* **2:** 335-339.

Van Broeckhoven C.L. 1995. Molecular genetics of Alzheimer Disease: identification of genes and gene mutations. *Eur. Neurol.* **35:** 8-19.

Van Camp G., P. Stinissen, W. Van Hul, H. Backhovens, A. Wehnert, A. Vandenberge and C. Van Broeckhoven. 1989. Selection of human chromosome 21-specific DNA probes for genetic analysis in Alzheimer's dementia and Down syndrome. *Human Genetics* **83:** 58-60.

Van Camp G., W. Van Hul, H. Backhovens, P. Stinissen, A. Wehnert, D. Patterson, A. Vandenberge and C. Van Broeckhoven. 1990. Physical mapping of chromosome 21 DNA markers in Alzheimer's disease region using somatic cell hybrids. *Som. Cell and Mol. Genet.* **16:** 241-249.

Van Camp G., M. Cruts, H. Backhovens, A. Wehnert and C. Van Broeckhoven. 1992. Unique sequence homology in the pericentromeric regions of the long arms of chromosomes 13 and 21. *Genomics* **12:** 158-160.

Van Hul W., G. Van Camp, L. Stuyver, J.M. Delabar, M.G. McInnes, A.C. Warren, S.E. Antonarakis and C. Van Broeckhoven. 1993. A contiguous physical map of the pericentromeric region of chromosome 21q between D21Z1 and D21S13E. *Genomics* **15:** 626-630.

Van Hul W. and C. Van Broeckhoven. 1993. A longe-range restriction map from the APP gene to the centromere of chromosome 21. In Alzheimer's disease: Advances in clinical and basic research (ed. B. Corain, K. Iqbal, M. Nicolini, B. Winblad, H. Wisniewski and P. Zatta), pp.235-242. John Wiley & Sons, New York, NY.

Wang S., Q. Wang, B.E. Crute, E.N. Melnikova, S.R. Keller and N.A. Speck. 1993. Cloning and characterization of subunits of the T-cell receptor and murine leukemia virus enhancer core-binding factor. *Mol. Cell. Biol.* **13:** 3324-3339.

Wang S., T. Stacy, M. Binder, M. Marin-Padilla, A.H. Sharpe and N.A. Speck. 1996. Disruption of the Cbfa2 gene causes necrosis and hemorrhaging in the central nervous system and blocks definitive hematopoiesis. *Proc. Natl. Acad. Sci. USA* **93:** 3444-3449.

Wang D.G., J.-B. Fan, C.-J. Siao, A. Berno, P. Young, R. Sapolsky, G. Ghandour, N. Perkins, E. Winchester, J. Spencer *et al.* 1998. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280:** 1077-1082.

Weber J.L. and P.E. May. 1989. Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *Am. J. Hum. Genet.* **44:** 388-396.

Weber J.L. and E.M. Myers. 1997. Human whole-genome shotgun sequencing. *Genome Research* **7:** 401-409.

Weissenbach J., G. Gyapay, C. Dib, A. Vignal, J. Morissette, P. Millasseau, G. Vaysseix and M. Lathrop. 1992. A second-generation linkage map of the human genome. *Nature* **359:** 794-801.

Wilcox A.S., A.S. Khan, J.A. Hopkins and J.M. Sikela. 1991. Use of 3' untranslated sequences of human cDNAs for rapid chromosome assignment and conversion to STSs: implications for an expression map of the genome. *Nucl. Acids Res.* **19:** 1837-1843.

Wilkinson K.D. 1997. Regulation of ubiquitin-dependent processes by deubiquitinating enzymes. *FASEB J.* **11:** 1245-1256.

Williams J.D., R.L. Summit, P.R. Martens and R.A. Kimbrell. 1975. Familial Down syndrome due to t(10;21)translocation: evidence that the Down phenotype is related to trisomy of a specific segment of chromosome 21. *Am. J. Hum. Genet.* **27:** 478-485.

Wu C., S. Zhu, S. Simpson and P.J. de Jong. 1996. DOP-vector PCR: a method for rapid isolation and sequencing of insert termini from PAC clones. *Nucl. Acids Res.* **24:** 2614-2615.

Wyman A.R. and R. White. 1980. A highly polymorphic locus in human DNA. *Proc. Natl. Acad. Sci. USA* **77:** 6754-6758.

Xu H., H. Wei, F. Tassone, S. Graw, K. Gardiner and S.M. Weissman. 1995. A search for genes from the dark band regions of human chromosome 21. *Genomics* **27:** 1-8.

Yaspo M.-L., L. Gellen, R. Mott, B. Korn, D. Nizetic, A. Poustka and H. Lehrach. 1995. Model for a transcript map of human chromosome 21: isolation of new coding sequences from exon and enriched cDNA libraries. *Hum. Mol. Genet.* **4:** 1291-1304.

Yergeau D.A., C.J. Hetherington, Q. Wang , P. Zhang, A.H. Sharpe, M. Binder, M. Marin-Padilla, D.G. Tenen, N.A. Speck and D.-E. Zhang. 1997. Embryonic lethality and impairment of haematopoiesis in mice heterozygous for an AML1-ETO fusion gene. *Nature Genetics* **15:** 303-306.

Yokobata K., B. Trenchak and P.J. de Jong. 1991. Rescue of unstable cosmids by *in vitro* packaging. *Nucl. Acids Res.* **19:** 403-404.

Yokota J. and T. Sugimura. 1993. Multiple steps in carcinogenesis involving alterations of multiple tumor suppressor genes. *FASEB J.* **7:** 920-925.

Yoon P.W., S.B. Freeman, S.L. Sherman, L.F. Taft, Y. Gu, D. Pettay, W. Dana Flanders, M.J. Khoury and T.J. Hassold. 1996. Advanced maternal age and the risk of Down Syndrome Characterized by the meiotic stage of chromosomal error: a population based study. *Am. J. of Hum. Genet.* **58:** 628-633.

Yu J., S. Tong, Y. Shen and F.-T. Kao. 1997. Gene identification and DNA sequence analysis in the GC-poor 20 megabase region of human chromosome 21. *Proc. Natl. Acad. Sci. USA* **94:** 6862-6867.

Zhang M.Q. 1997. Identification of protein coding regions in the human genome by quadratic discriminant analysis. *Proc. Natl. Acad. Sci. USA* **94:** 565-568.

Zipurski A., M. Peeters and A. Poon. 1987. Megakaryoblasic leukemia and Down's syndrome: a review. *Pediatr. Hematol. Oncol.* **4:** 211-230.

Zipurski A., A. Poon and J. Doyle. 1992. Leukemia in Down syndrome. *Pediatr. Hematol. Oncol.* **9:** 139-149.

Appendix I Tables

Location of the potential 21q11 PAC clones in 96-well microtitre dishes

1A1	1C4	**1C1**	8P3	**1E1**	23M5	**1G1**	46D1
1A2	1E6	**1C2**	9F1	**1E2**	24B1	**1G2**	46K6
1A3	1G17	**1C3**	10G23	**1E3**	25O12	**1G3**	46M24
1A4	1G7	**1C4**	10H17	**1E4**	25P24	**1G4**	46N17
1A5	1H21	**1C5**	10J13	**1E5**	26M10	**1G5**	47N11
1A6	1P24	**1C6**	10J7	**1E6**	28C15	**1G6**	48P20
1A7	2J11	**1C7**	11A24	**1E7**	28E18	**1G7**	48P3
1A8	2M6	**1C8**	11J19	**1E8**	28P11	**1G8**	48P5
1A9	3D1	**1C9**	11P10	**1E9**	29G24	**1G9**	49A20
1A10	3G22	**1C10**	12B10	**1E10**	30P13	**1G10**	49D5
1A11	3J7	**1C11**	12H4	**1E11**	31G17	**1G11**	49L12
1A12	3P15	**1C12**	13B16	**1E12**	32I18	**1G12**	50F14
1B1	4A9	**1D1**	13F23	**1F1**	32M22	**1H1**	50G11
1B2	4D1	**1D2**	13M7	**1F2**	34A17	**1H2**	50G17
1B3	4L23	**1D3**	14L10	**1F3**	34B13	**1H3**	50N18
1B4	4L24	**1D4**	16E19	**1F4**	34H24	**1H4**	50P17
1B5	4M2	**1D5**	17N7	**1F5**	34M7	**1H5**	50P7
1B6	4O17	**1D6**	19D22	**1F6**	36J1	**1H6**	51G16
1B7	5L6	**1D7**	20M4	**1F7**	38J19	**1H7**	51N22
1B8	6B21	**1D8**	20N11	**1F8**	39O3	**1H8**	53J15
1B9	6C17	**1D9**	22E2	**1F9**	42C18	**1H9**	53P6
1B10	7F1	**1D10**	22J3	**1F10**	42E13	**1H10**	54F17
1B11	7G13	**1D11**	22O3	**1F11**	43O5	**1H11**	54G21
1B12	8A22	**1D12**	22P6	**1F12**	44L23	**1H12**	54O23
2A1	55J3	**2C1**	64J11	**2E1**	72A18	**2G1**	77M13
2A2	55O1	**2C2**	64J19	**2E2**	72H5	**2G2**	77P15
2A3	56I19	**2C3**	65M23	**2E3**	72O10	**2G3**	78A24
2A4	56J8	**2C4**	66E20	**2E4**	73K18	**2G4**	78A5
2A5	57N10	**2C5**	66M6	**2E5**	73M5	**2G5**	78E18
2A6	58I13	**2C6**	67A21	**2E6**	73O12	**2G6**	78G20
2A7	59B23	**2C7**	67F13	**2E7**	74G19	**2G7**	78N12
2A8	59H17	**2C8**	67F22	**2E8**	74H9	**2G8**	79N5
2A9	60B14	**2C9**	67K6	**2E9**	74O13	**2G9**	80J24
2A10	60D20	**2C10**	67N22	**2E10**		**2G10**	80K19
2A11	60P21	**2C11**	68E24	**2E11**	75C2	**2G11**	81C2
2A12	61A1	**2C12**	68F21	**2E12**	75D10	**2G12**	81N18
2B1	61E22	**2D1**	68G23	**2F1**	75D4	**2H1**	81O23
2B2	61I16	**2D2**	68M16	**2F2**	75G13	**2H2**	81P23
2B3	61M19	**2D3**	69J5	**2F3**	75H15	**2H3**	82B20
2B4	61O14	**2D4**	70G18	**2F4**	75H21	**2H4**	82B3
2B5	62B10	**2D5**	70G5	**2F5**	75I20	**2H5**	82E14
2B6	62D10	**2D6**	70H17	**2F6**	75J24	**2H6**	82G13
2B7	62L22	**2D7**	70J18	**2F7**	75L12	**2H7**	82L10
2B8	62O6	**2D8**	71E8	**2F8**	75N22	**2H8**	82M24

Table 1

2B9	62P23	**2D9**	71F7	**2F9**	76A24	**2H9**	83L5
2B10	63K18	**2D10**	71I13	**2F10**	76I22	**2H10**	83O23
2B11	63L9	**2D11**	71K24	**2F11**	77G4	**2H11**	84H16
2B12	64G14	**2D12**	71L19	**2F12**	77I10	**2H12**	84J17
3A1	84O4	**3C1**	96F17	**3E1**	104M21	**3G1**	113K21
3A2	85B5	**3C2**	98M20	**3E2**	104O17	**3G2**	114A24
3A3	85C22	**3C3**	100M18	**3E3**	105A12	**3G3**	114A9
3A4	85G21	**3C4**	100P19	**3E4**	105B17	**3G4**	114C16
3A5	85J10	**3C5**	101B21	**3E5**	105C13	**3G5**	114N11
3A6	85M20	**3C6**	101C10	**3E6**	106B12	**3G6**	115I15
3A7	87H19	**3C7**	101D22	**3E7**	106F2	**3G7**	115L18
3A8	89K22	**3C8**	101E10	**3E8**	106K6	**3G8**	116I21
3A9	90A15	**3C9**	101F4	**3E9**	106L7	**3G9**	116J11
3A10	90F20	**3C10**	101N16	**3E10**	107I8	**3G10**	116M13
3A11	90G18	**3C11**	103E4	**3E11**	107P21	**3G11**	117B4
3A12	91E10	**3C12**	103I12	**3E12**	108B10	**3G12**	117L14
3B1	91H9	**3D1**	103I19	**3F1**	108J20	**3H1**	117N22
3B2	91M21	**3D2**	103J19	**3F2**	108M18	**3H2**	117O23
3B3	93H17	**3D3**	103K19	**3F3**	109F6	**3H3**	118E16
3B4	93L18	**3D4**	103L15	**3F4**	109N19	**3H4**	118J16
3B5	94F3	**3D5**	104C11	**3F5**	110L15	**3H5**	118O13
3B6	94H11	**3D6**	104D15	**3F6**	110N4	**3H6**	118P3
3B7	94J18	**3D7**	104D23	**3F7**	111G14	**3H7**	119E12
3B8	94N13	**3D8**	104E17	**3F8**	111K10	**3H8**	119E17
3B9	94P21	**3D9**	104F10	**3F9**	112E7	**3H9**	119F19
3B10	95B16	**3D10**	104K11	**3F10**	112F17	**3H10**	119P15
3B11	95L24	**3D11**	104K14	**3F11**	113F21	**3H11**	120C24
3B12	96B19	**3D12**	104L10	**3F12**	113J11	**3H12**	120E20
4A1	121H20	**4C1**	131H14	**4E1**	146F2	**4G1**	156L9
4A2	122C13	**4C2**	131L13	**4E2**	146K23	**4G2**	158B11
4A3	122E23	**4C3**		**4E3**	147H3	**4G3**	158C10
4A4	123B8	**4C4**	132O8	**4E4**	149N1	**4G4**	158C4
4A5	123D2	**4C5**	133G21	**4E5**	150O6	**4G5**	158D20
4A6	123J6	**4C6**	133J18	**4E6**	151D1	**4G6**	158F12
4A7	123N18	**4C7**	134I9	**4E7**	151M11	**4G7**	158K5
4A8	124B10	**4C8**	134M2	**4E8**	152D4	**4G8**	158M18
4A9	124E17	**4C9**	135D11	**4E9**	152H2	**4G9**	159H11
4A10	124G3	**4C10**	135E14	**4E10**	152L12	**4G10**	160J5
4A11	124H22	**4C11**	136D14	**4E11**	153I21	**4G11**	160P20
4A12	124J6	**4C12**	136O6	**4E12**	153K1	**4G12**	161B4
4B1	124K19	**4D1**	136P6	**4F1**	153L2	**4H1**	161F2
4B2	125O10	**4D2**	137H19	**4F2**	153M4	**4H2**	161I14
4B3	126B20	**4D3**	139B17	**4F3**	153N5	**4H3**	161J6
4B4	126E22	**4D4**	139N13	**4F4**	154M13	**4H4**	161M12
4B5	126M24	**4D5**	140O22	**4F5**	155B11	**4H5**	162A14
4B6	127G16	**4D6**	142B10	**4F6**	155C3	**4H6**	162L2
4B7	127M18	**4D7**	142I11	**4F7**	155I7	**4H7**	162O12
4B8	128A18	**4D8**	143J22	**4F8**	155J13	**4H8**	163B6

Table 1

4B9	128E2	**4D9**	144O20	**4F9**	155M10	**4H9**	162I23
4B10	130I9	**4D10**	145E23	**4F10**	155M13	**4H10**	164H3
4B11	130N2	**4D11**	146A10	**4F11**	156E22	**4H11**	164J22
4B12	131G22	**4D12**	146E22	**4F12**	156K9	**4H12**	164K23
5A1	167E22	**5C1**	181M3	**5E1**	206I10	**5G1**	225A23
5A2	168L24	**5C2**	182D15	**5E2**	207L18	**5G2**	226G9
5A3	168N11	**5C3**	186J13	**5E3**	207O15	**5G3**	226J16
5A4	168P13	**5C4**	186J22	**5E4**	209F16	**5G4**	226M9
5A5	171E24	**5C5**	187K17	**5E5**	210A14	**5G5**	227G3
5A6	171H13	**5C6**	188B2	**5E6**	210C20	**5G6**	227K21
5A7	171M19	**5C7**	188F1	**5E7**	214F15	**5G7**	227O9
5A8	172C5	**5C8**	190A9	**5E8**	214G14	**5G8**	228G2
5A9	172O5	**5C9**	191B2	**5E9**	214G16	**5G9**	228K10
5A10	173A6	**5C10**	191F13	**5E10**	214O7	**5G10**	229K6
5A11	173B1	**5C11**	191L13	**5E11**	215A17	**5G11**	230N6
5A12	173M21	**5C12**	192C21	**5E12**	215L21	**5G12**	231N9
5B1	176L12	**5D1**	192D22	**5F1**	216O16	**5H1**	75C11
5B2	176P5	**5D2**	192P16	**5F2**	218D22	**5H2**	132M11
5B3	177F3	**5D3**	193D6	**5F3**	218E11	**5H3**	
5B4	177L10	**5D4**	193H1	**5F4**	220K12	**5H4**	
5B5	177N5	**5D5**	194E16	**5F5**	220L20	**5H5**	
5B6	178B18	**5D6**	194L2	**5F6**	221E7	**5H6**	
5B7	178L18	**5D7**	194P20	**5F7**	221H12	**5H7**	
5B8	178L5	**5D8**	194P24	**5F8**	221N12	**5H8**	
5B9	180J6	**5D9**	195F16	**5F9**	222D22	**5H9**	
5B10	180L14	**5D10**	199O21	**5F10**	222K1	**5H10**	
5B11	180O17	**5D11**	200I17	**5F11**	223N8	**5H11**	
5B12	181A6	**5D12**	101J19	**5F12**	224L2	**5H12**	
6A1	242B22	**6C1**	258B10	**6E1**	273C24	**6G1**	289O7
6A2	242L8	**6C2**	258E10	**6E2**	273H23	**6G2**	290G22
6A3	244A13	**6C3**	259A1	**6E3**	274I22	**6G3**	
6A4	245A11	**6C4**	259J5	**6E4**	275M23	**6G4**	291H16
6A5	245G17	**6C5**	260H6	**6E5**	277F15	**6G5**	291I9
6A6	245H4	**6C6**	261J24	**6E6**	279F16	**6G6**	291J15
6A7	245J9	**6C7**	262C12	**6E7**	279J6	**6G7**	293N11
6A8	246I20	**6C8**	262E18	**6E8**	279M12	**6G8**	293O10
6A9	247L10	**6C9**	262L10	**6E9**	279N21	**6G9**	291H14
6A10	247M12	**6C10**	262N20	**6E10**	280K24	**6G10**	
6A11	247N12	**6C11**	263F2	**6E11**	280N14	**6G11**	
6A12	247O10	**6C12**	263L5	**6E12**	281C13	**6G12**	
6B1	247O16	**6D1**	263N8	**6F1**	281D20	**6H1**	232E4
6B2	248G6	**6D2**	265A22	**6F2**	282C15	**6H2**	233C14
6B3	248G8	**6D3**	266L2	**6F3**	282L17	**6H3**	233K18
6B4	248H23	**6D4**	267D22	**6F4**	283A7	**6H4**	234E1
6B5	249L2	**6D5**	269L2	**6F5**	283D5	**6H5**	235I18
6B6	250J2	**6D6**	270D8	**6F6**	283L17	**6H6**	236H9
6B7	251I9	**6D7**	270E18	**6F7**	283M21	**6H7**	236O21
6B8	252A1	**6D8**	270M7	**6F8**	286K18	**6H8**	237C14

Table 1

6B9	253O3	**6D9**	272B20	**6F9**	286L23	**6H9**	237G15
6B10	254C5	**6D10**	272L23	**6F10**	287F23	**6H10**	239C1
6B11	256M13	**6D11**	272M2	**6F11**	287N14	**6H11**	239C7
6B12	257P13	**6D12**	272N20	**6F12**	289M21	**6H12**	241B11

Table 1: The 560 PAC clones that were selected for further analysis from the hybridisations of the 21q11 YACs to the total human PAC library, RPCI 1 (Ioannou *et al.*, 1994), and their position in the six 96-well microtitre dishes.

Hybridisation results of the membranes containing the putative exons

98L15/90B5BB

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
A																					15,16,20	15,16,18,19		
B	2,9,10			7						2,9-12,17			5,10,13				15,16,20	15,16,20						
C																					10,13			
D																	15,16,20							
E				16,20				15,16,20					10,11,12								2,9-12			2,9-12,17
F						2,9-12,17		15,16,20							15,16,18,19									
G			2,9-12,17																					
H		5,10,13	7		10,13	9-12,17				10,13			14				5,10,13		7		5,10,13	9-12,17		
I		9-12,17	2,9-12,17	5,10,13					14	14	14			10,13	1a,7		14		5,10,13	14				
J	10,13	1a,4,7			14			1a,7	5,10,13		1a,8	14			5,10,13	14			1a,14	14				
K	9-13,17	7,13	5,10,13	9-12,17	5,10,13		10,13	14	14	7	1a,7							1a,9-12,17	2					9,13
L	9-13,17	10,13	10,13	5,10,13,14			14			7	1a,9-12,17			9-12,17			1a,9-12,17		14					
M	9-14	9-12,17		1a,9-12,17		14			1a,9-12,17	10,13	14		1a,9-12,17			9-12,17		5,10,13	5,10,13	9-12,17	9-12,17	9-12,17		
N	10,14	10,13				9,10	1a,6	5,10,13								14	14		9-12,17		14	10,13		
O																								
P																								

98L15/90B5BB: 1a-insert 90B5, 2-HIV *tat* intron, 3-COT 1 DNA, 4-90B5BB 3^0 EX1/2, 5-98L15BB 3^0 EX1/2, 6-insert image 910220, 7-98L15BB-K11, 8-98L15BB-J11, 9-98L15BB-L11, 10-IMAGE 1182683, 11-IMAGE 1183050, 12-IMAGE 124053, 13-90B5BB-C21, 14-98L15BB-J19, 15-IMAGE 248395, 16-IMAGE 193617, 17-98L15S-A9, 18-IMAGE 248318, 19-90B5BB-A22, 20-98L15S-G9.

98L15S

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
A		17			17	15			5,7,8,9			17			2	17	13					17		
B						17		4,14		13		4		17	17	15			5,8-10		14			15
C	2,4	1b,2,4,17	1b,14	15		13			13	13			3,13	18					13					
D	5,7,8,9	15		17			13,17	5,7-10	4,13,17	15	17	4,18		17		16,19,20			17				14,17	
E					2				4,13,15	15	13			17							18			
F		2,4,15	17		13				13							13,17		15		5,7-10	13	13	5,7,8,9,10	
G		15	15,16,19,20				2,17	18	4,16,17,19,20		17		17	17	17		13					15		17
H		4	1b,15	15	4,15			8,12	14	4,8,11,12	14	6	4,17	4,15	13	17	8		17		8,12,16	8,12,17		2
I	16,19,20		17		16,19,20	4,8,12	16,19,20	4,8,12	8	15	8	17	14	11	8,9	15	16,19,20	11			15	15		14
J		16,19,20		15	4,11	4,16,17,19,20		15	14		13	15	8,12	4,8,12	8,9	8		15	17	17			17	
K	4	4,8,12		8,12	8,9			13	14	4,17	16,19,20	4,15		14	4,17	4,8,12			15		4,8,12	8,12		15
L				4,14			14	4,11	17		4,8,12	4,15		4,8,12	8,12		16,19,20			17				
M	12	14	16,19,20	14	17	11	4,11		15		15		17	2,4			17	15		14	1b,8,12		1b	
N			15	15		16,19,20	8,12	8,12			1b,11	4	8,12	11	8,12	15			16,19,2(11			16,19,20	15	
O		2																						
P																								

98l15S: 1b-insert 90B5, 2-HIV *tat* intron, 3-COT 1 DNA, 4-98L15S 3^0 EX1/2, 5-98L15S-A9, 6-image 910220, 7-image 124053, 8-image 1182683, 9-image 1183050, 10-98L15BB-L11, 11-98L15BB-K11, 12-90B5BB-C21, 13-98L15S-B8, 14-98L15S-C3, 15-98L15S-E9, 16-98L15S-G9, 17-98L15S-K10, 18-98L15S-D12, 19-image 248395, 20-image 193617. No positives: image 248318.

Table 2

90B5S

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
A	4,15		17	6,7		15		17				2		16	15		18	2		15			6,7	
B		16	1a,2	2			2	4,18		17	17		6,7		15	15			6,7	18				4,18
C		17			1a,6,7	2	17	15	2,17	18	22				15		6,7	16	6,7			2	2	
D	2		15			16		11,17				18	15	17		15			15	16	6,7	6,7	2	
E			4,18		2	6,7	17		15		2				2	6,7	1b,16		17	15		6,7		
F		17			15				16				6,7	6,7	15	2,6,7	18		22	16	16		2	6,7
G	2				1a,17			17							18		16	2	2	6,7				2
H	2,4	1a,4,9,13	2	3	4,12	4,9,13	2			2	23	5,19,20,22	5,19,20	2,9		5,19,20,22	6,7		9	2,4,23	2	2	9,13	9,13
I	6	1a,4,9,13	9,13	6,7,13	9,13	22							4,9,13	6,7,13	4,9,13	5,19,20,22		4,9,13		4,9,13				9,13
J	9	2		2	8,9,10,21	2		2,4		2	1b,4,14,23		2		2	1a,6,7	2			4,9		2		
K		5,19,20,22	1b,4,8,9,10,11,21	3		1a								4,9,13	4,9,13	2	9,10	2	6,7		6,7		6,7	1b,4,9,10,14,23
L	1a,4,9,13			9,13	1a,4	5,19,20,22	4,12	19,20,23	1b,8,9,10,11,21	2,9,10			1b,14	4,9,13			2		5,19,22	5,19,20,22	9,13		4,9,13	5,9,13
M	9		6,7	2	14		2,6,7	6,7	4,9,13	1a,4,9,13		6,7		2	9			19,20,23				4,9,13	4,9,13	22
N	9	2	1b,8,9,10,11,21		1a,9	9,10			1a	1b,14			2		9			1a	6,7	2		9	4,9,13	
O																								
P																								

90B5S: 1a-insert 90B5, 1b-insert 98L15, 2-HIV *tat* intron, 3-COT 1 DNA, 4-90B5S 3^0 EX1/2, 5-90B5BB-A22, 6-STCH-A, 7-90B5S -E6, 8-image 124053, 9-image 1182683, 10-image 1183050, 11-98L15BB-L11, 12-98L15BB-K11, 13-90B5BB-C21, 14-90B5S-J11, 15-90B5S-A1, 16-90B5S-E17, 17-90B5S-G5, 18-90B5S-B24, 19-image 248395, 20-image 193617, 21-98L15S-A9, 22-image 248318, 23-98L15S-G9.

126N20

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
A	4,7		1,9	4,7				7					4,7			7		7	7	7		1,4,7,10		
B	10		2,4	4		1,8			2	2	2	2			2		2			2	2			
C			7	1,6	7	7			7						7		7	7			7		7	3,7
D	2						2	2	2						2,3		2		2		2		2	2
E	1,7		7				7				7		4,7		7	7					7			
F	2					2	2		2				2		2	2		1,2,10	2	2				
G		4,7	7	4	7				7		7	7	7	7	7		7		1,6				1	
H	2	1,8			2	1,10	2							1,10	2	2								
I	4,7	4				7		7			1					7		5		7				4,7
J		3		2	1,10		2				2					2					2			
K		7	7		7				7	7	7	7					7	7	7	7				
L							2	2			2	2	2		2		2				2		2	1,5,10
M			7	7	7	7				4,7	1								4		7			
N		2	2					2	2		2				2	2	2		2	5		1,5,8		2
O	7				7								7				7							
P										2	2	2	3	2	2		2		2					

126N20: 1-insert 126N20, 2-HIV *tat* intron, 3-COT 1 DNA, 4-126N20BB 3^0 EX1/2, 5-126N20S 3^0 EX1/2, 6-126N20-C4, 7-126N20 -E1, 8-126N20-N22, 9-126N20-A3, 10-126N20-A22.

Table 2

31B5

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
A			4,5,6				6		5,6	4,5,6,11			6	4,5,6,11			6		5,6,11	9				
B	4,5,6,11	6,11	7,8,10,11		8,10			6	4,5,6,11	4,5,6,11	7,8,10	3	4,5,6,11	5,6,7,11	7,8,10	4,5,6,11		3	6	6	7			
C		6			5,6								4											
D		3					6	6			4,5,6,11	6	6		4,7,8,10,11	4,5,6,11	4,5,6,11	5,6,11	4,5,6,11			5,6	3	
E									3						6		6		6		1,5,7,8,10			
F									6	8	3		2	8,10			3				3,5		6	
G	6,11	6						4,6,11		5,7,8,10					6			6	5,6			6		
H		6,11			6,11		6,11					1	4,5,6,11	6,11	8,10	6	6,11	10			5,6,11			
I	9																6	6						
J		2							6	3				4,5,6,11	3		5		3	6				
K				4,6,11								4,5,6,11			4,5,6,11							4,5,6,11	4,5,6,11	
L	5,6,11			6,11					5		4,5,6,11						3		6					
M			4,5,6,11		4,5,6,11		4,6,11		6,11		4,6,11		6,11		4,6,11	9	4,5,6,11	6,11	9		4,9		6	
N						5				5	3			6,11		2			4,6,11					
O	6		9			4,6,11	9			1,4,6,11	4	4,6,11	4,5,11		9	4,6,11		4,5,6,11	3	5	4,9			5,6,11
P		2							3,5		7,8,10,11								3,4		3		1,7	

31B5: 1-insert 31B5, 2-HIV *tat* intron, 3-COT 1 DNA, 4-31B5BB 3^0 EX1/2, 5-31B5S 3^0 EX1/2, 6-31B5-A3, 7-31B5-E21, 8-image 360025, 9-31B5-M21, 10-31B5-B5, 11-image 1286563.

B39I12

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
A	1,9		1,4,7				8	8	9	9	1,4,7	4,7		9	9	7	8	7						
B	1,4,7		4,9	1,4,7		9	9			8			9		4,7	9	9				4,7	4,7	9	8
C			1,4,7		9	9	4,7		1,4,7	4,7				5,8	9	9	8	7	9		4,7	8	9	
D	5,7	1,4,5,7	8	9			9	4,7		1	8	8		9	8	4,7		8			8		9	
E	8	4	9	9	9				9		1,4,7		9			2	8	1,6		8		9		8
F	1,6					2	1,5,8	8	7	7	8		9		8	8	8							8
G	2	8	8		8		8	8	5,8					7		7	8		9			8		
H			9	2						8		2				9	2			3			9	9
I	1,4	9							5,8	2		9				2	8	8		7	2		2	
J	1,5,8	1	1,2			2			7	5,6		1	1,8	1	8			8						
K	1,4,5,8	2	1		3			8	5,8	5,8	7		8		7	8	2	2	2	2			8	
L	1	1,2	1,2											2	2		8	8	8			2		
M	1,5,8	1									2					7		2	6					7
N	1	1	8	1,6		8		2	2		2	5,8		8			2,8	8	9		8	8	7	
O																								
P							2					2												

B39I12: 1-insert B39I12, 2-HIV *tat* intron, 3-COT 1 DNA, 4-B39I12BB 3^0 EX1/2, 5-B39I12S 3^0 EX1/2, 6-B39I12-E18, 7-B39I12-D2, 8-B39I12-F7, 9-B39I12-A1.

Table 2

c102 68C10/132H4

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
A	6		6		6						6		6	6				6		6		4,6		
B				6		6	6		5,6			6		6										
C	6								5,6					5,6						6				
D	6					6				4,6			6								6			
E				6						4,6								6				6		6
F			6			6					4,6				6							6		
G								4,6		4,6						4,6								5,6
H	6		6			6		4,6		4,6		4,6		4,6				4,6	4,6	4,6	4,6		5,6	5,6
I		6	5,6	6	6	6		6	6		4,6	4,6		4,6	4,6	6				4,6		6	5,6	
J	6		4,6	6		6	6		6			4,6	4,6				6	6		4,6		6		5,6
K			5,6	6		6		6	6	6		6	6	6		6		6	6	6		6	5,6	5,6
L			5,6				6	6				6		6						4,6	4,6	4,5,6	5,6	
M		6			5,6				6		4,6	4,6	6	6	2	6		4,6	4,6				4,5,6	4,5,6
N	6	6					6		6	6				6			6	4,6		4,6	4,6	4,6		4,6
O						2																		
P																2								

c102 68C10/132H4: 2-HIV *tat* intron, 3-COT 1 DNA, 4-c102 68C10/132H4BB 3^0 EX1/2, 5-c102 68C10/132H4S 3^0 EX1/2, 6-c102 68C10/132H4-M11.

190N10

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
A	5,6	6	6	6	6	6	6			6	6	6	6			6	6	6	6	6	6	6	7	
B	6	6	6	6	7	6	6	6	6	7	6		6	6	6	6		6	6	6	7	6	6	6
C	6	3		6		6		6		6	6	6	6	6	6				6	6		6	6	6
D	6	6	6		6	6	6	6	6			6	3	6	6	6	6	6	6	6		6	3	6
E	6	6	6	6	3		6	6		6	6	6	7	6	6	6	6		6	6	6	4,5,6,7	6	
F	6	3	6		6		6			6	6	6		6		6	6			6	3		6	3
G	6		6	1,6		6	3		6		6	4,6	6	9	6	6	6	6	6	6	6	6	6	7
H	1,9	7		1,7	6	2	6	5,7	1,6	7	6	4,5,6	4,6		6,7		6	6		7	7			
I	9	6	7	1		7		6		5,7		6		3	7		6							
J		7	7	6		6	6		1		7		7	7	6					6				
K	6	6		5		3,7			1,3		7			4,5,6	6	6				7	6	7	7	
L	6	6	6	6		7		3		6	6	2	7		7	6	7		7		7		6	7
M	4,6	9		6	7		3	3		4,5,6	6	7			2,7		7	7	6	7	6	7	6	
N	9	3	7	7	6,7	6	4,6	4,6				3,4		7	7		3		3			6	9	
O																								
P																								

190N10: 1-insert 190N10, 2-HIV *tat* intron, 3-COT 1 DNA, 4-190N10BB 3^0 EX1/2, 5-190N10S 3^0 EX1/2, 6-190N10-K14, 7-190N10 -H8, 8-190N10-J9, 9-190N10-H1.

Table 2

153I22

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
A		4,10	3,4	3,4	4,10	4,5,9	3,4,5,9	4	4,5,9		4,10		4,10			3,4	6	3,4	4,5,9	4,9	4,9	3,4	1,4,6	
B			3,4		3,4	6			3,4	3,4			4,10	3,4					3,4,10	10	10			
C		3,4	3,4				3,4												4	4	4			3,4
D			1		3,4			3,4		3,4,5,9		3,4		4	3,4	3,4	3							
E			4		4,10					4	3,4		3,4						3,4			3,4		
F		3,4									3,4	4,10				3,4	4,6		3,4					3,4,5,9
G	4,10		4,9	4,10		4,5,9							3,4	3,4		4,10	3,4		4	4,10				3,4
H	5,7	3,5	5,7		3,5	4,5,9	2	2	2,5,7	3,5	2		5,8	2	5,8	2		3,5	2	2		2	3,5	2
I	2			5,7		2	3,5	5,8	2		2	2	2	5,7	2	5,7	2	2	2	5,7	3,5	2	3	5
J	5,8	2	2,5,7		5	5,7	2	5,7	2	2	5	3,5		2,5,7	3	2	2	2		5,8	7	7		2
K	2	2,8		2		2	2	3	3	2	7		2		3,4,9	3	2	2	5,7		5,7	3	2	
L	3,5		2	5,7	5,7		7	2	2	2	2	2	3,5	3,5	1,5,7	1	5,7	2		5,8		3,5	5,7	2
M		2	5,7	3,5		2	2	2		3,5		2	2		1,5,7	2	5	5,7		3,5	2	3	2	
N		2		7	5		2	2	2	2	1,5,7		2	2,3,5		2	3	1,5,7	2	1,5	2,3	7		3,5
O																								
P																								

153I11: 1-insert 153I22, 2-HIV *tat* intron, 3-COT 1 DNA, 4-153I22BB 3^0 EX1/2, 5-153I22S 3^0 EX1/2, 6-153I22-A23, 7-153I22-M15, 8-153I22-H13, 9-153I22-A6, 10-153I22-A2.

30P13

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
A	4,5,10	1,4	1,5,9	1	1				6	6			6	5,9		5,9	6	5,10	5,9	7	6			
B	4,5,10	1,4,5,7,10	1	1,6	1,10	6	6	7	10			6			6		6		10	9	7		10	
C		6	1,10				1,10		4	4,10	7	6	10	7				6	10	4,7		10		7
D		6	7			9	1	9	10	4,5,10	4,5,10	6	7		10	6	10	10		4,7		6	4	10
E	10	10		1,4,7	4		6			6	10	6	4,10		4	6		10		9	6	5,9	4,5,10	
F			6	6	6	5,6	6	6	6	4,10	9	3	4,7		4	7	6,7		7,10		1,10	1,10	5,10	
G		4	6	4,7			5	4,5,6	9	6	10		4			10		6	10	7		5,9	6	7
H			8	4,10	2	10		1	7	7	2,4		4			2								
I	7	7	7	1,5	1,3	4,7			7	4,7		2	4,7	4,5,10			3	2			2	2		
J	4,7		4	1	1,3		5	10	4,10			4		7	5,7,10		8		3	7			2,7	
K	7			7		3			2		4,5,7,10			4	5,10	7			10	2	7	1,7		
L	2	7	2	5,6			3		7		3	6		7	2	10				7	4	8	4,7	7
M		7	4	2	6	7			7			5,6		2	7	7			2	3	6		2	2,4
N				2	2		7	7	1,5,10	1,4			7	5,10		6	7		4,7	7	4,10	2		
O																								
P																								

30P13: 1-insert 30P13, 2-HIV *tat* intron, 3-COT 1 DNA, 4-30P13BB 3^0 EX1/2, 5-30P13S 3^0 EX1/2, 6-30P13-B4, 7-30P13-E4, 8-30P13-H3, 9-30P13-F11, 10-30P13-A1.

Table 2

270M7

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
A			4,6	4,5,6				4,6	4,6	4,6		4,6	4,5,6								4,6			
B		4,6				4,8							5		4,6				4,6				4,6	4,6
C				4,6		4,6	4,6	4,6	4,6	4,6	4,6		1,5,9		4,5,6	4,5,6		4,6	4,6					
D	4,6			4,6	4,6	4,6		4,6	4,6	4,6			4,6		1,4,8		4,6	4,6		4,6	4,6			
E		4,6					3	4,6				4,6				4,6		4,6				4,6		4,6
F	4,6			4,6	5	4,6	4,6				4,6		5	4,6	5									
G	4,6	4,6									4,6		4,6	4,5,6			4,6						6	
H	5											5,9												
I													5											
J																	2							5,9
K																1,5,9						5	5,10	
L									5	1,5,7														
M	5							2			1,5,9					1,5								
N		5,10						2								5,9						1		
O							2	2	2	2	2	2												
P							2	2	2	2	2	2												

270M7: 1-insert 270M7, 2-HIV *tat* intron, 3-COT 1 DNA, 4-270M7BB 3^0 EX1/2, 5-270M7S 3^0 EX1/2, 6-270M7-A3, 7-270M7-L10, 8-270M7-D15,9-270M7-M11, 270M7-N2.

B194B7

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
A		4,6	4,6			8	4,7	5,8	8		4,5,6		8		8	6	5,8			8				
B	8			4,6	1,4,5,6	5,8		4,7	4,6		5,8	5,8	5,6	5,8	4,56		7							
C			6	4,5,6				1,5,8					8				6	4,5,7	5,6		5,8	5,8	5,8	5,8
D	4,5,6	5,8						6	8	8		4,5,6			3	5,8							8	8
E			1,8	3																8				
F																							6	
G									6	5,6														
H	4,7	9		4,9		4,7						4,6		4,9	4,7		7			7	4,7			4,5,7
I	4,7				4,7		7	9		6			7		7	7	7				7	1		
J	1,7			4,7			10	5,6		5,6			7		7		7	9				10	5,7	
K	7			4,7	1,10					4,7			7			4,6	4,5,7						7	
L		10			4,7		1,4,7	7		3	7			4,6,7					6		5,7			7
M	10				5,6	7	7		9			4,5,7												7
N		7	10			7			6								4,5,7						6	
O																							5	5
P			2																					5

B194B7: 1-insert B194B7, 2-HIV *tat* intron, 3-COT 1 DNA, 4-B194B7BB 3^0 EX1/2, 5-B194B7S 3^0 EX1/2, 6-B194B7-B5, 7-B194B7-L21, 8-B194B7-C8, 9-B194B7-H4, 10-B194B7-K5.

Table 2

103K19

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
A			3,4	9			9	6				1,4,6		3,4		9			4,6		3,4	3,4	6	6
B	3,4	1,10	1,8	3,4		1,3,4,10			3,4			4,10				3,4			3,4		9			3,4
C		3,4		4,9	3,4		3,4	9	3,4	3,4		1,4,6	3,4					9		3,4	4,10		3,4	9
D	3,4			6		6	6				1,6			3,4	3,4	1,6		6					10	10
E				3,4	9						3,4				9		6	3,4						
F	3,4			1,4,10	3,4			1,4,9	1,4,10				1,4,6								3,4		9	10
G						1,4,6	9	3,4			4,10			3,4	6							3,4	9	
H				1,11	1,8		3,4	9		3,4	3,4			9										
I	1,11										9			9						2		9	3,4	
J	9				1,7	1,7			1,7		3,4		4,9	4,9	4,9	3,4	3,4	3,4			8		3,4	
K	2									1,10	7									3,4				
L	3,4		1,7	2	1,8			3,4	3,4	1,9				3,4	3,4	3,4	1,7			8				
M			9	1,7													3,4		3,4			3,4	1,7	
N							9			3,4	3,4			8	2	9		9	1,7					3,4
O																								
P																						2		

103K19: 1-insert 103K19, 2-HIV *tat* intron, 3-COT 1 DNA, 4-103K19BB 3^0 EX1/2, 5-103K19S 3^0 EX1/2, 6-103K19-A12, 7-103K19 -J5, 8-103K19-B3, 9-103K19-J13, 10-103K19-F4, 11-103K19-K10.

10J13

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
A					3	3,4				3,4				2	3,4					3,4				3,4
B									7					7							3,4			1,4,7
C																								
D			7						3,4	7		3,4					1		3,4			3,4		3,4
E									3,4			4,7	3,4				1							
F						4			7	4,7	3,4							7			7			
G	3,4,6	7	3,4,6				1					7				3,4					4,7	7		
H	6	5,6		1,5,6			5,6	8	4,5,6		4,5,6	8	4,5,6					4,5,6		8	1		5,9	
I		6,9		8					8		8	8	8		8				8		8		8	4,5
J	8			4,5,6	4,6	9			5,9	5		8	8			5,9						9			
K			2	1	4,5,6	4,5,6		8		4,5,6	6	6		8		5,6	5,9			8				
L	8	4,5,6	1,8	1,5	8	4,5,6	8			4,5,6	6	8			5		8	4,5,6	4,5,6	4,5,6	8	4,5,6	2	
M		4,5,6	4,5,6				5,6,9		5,9	8	5,6										4,6			
N		8		4,5,6	1,4,5,6	8		5	6			3,4		8					5,9		8			
O																								
P			2				2																	

10J13: 1-insert 10J13, 2-HIV *tat* intron, 3-COT 1 DNA, 4-10J13BB 3^0 EX1/2, 5-10J13S 3^0 EX1/2, 6-10J13-N5, 7-10J13-B24, 8-10J13 -L3, 9-10J13-H23.

Table 2

c102 65F3/44E10

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
A	4,8				4,8	4,9	4,5,7				4,6	4,6	4,8	4,8				4,5,7	4,6	4,6,8	4,7,8	4,8	4	4
B	6	8	6		4,6	4,7		4,8			6	8			4,8	4,6	4,5,7,9		6	6	8	8	9	6
C						4,5,7	4,5,7	4,6	4,6	4,8	4,6	9	6	9			5,7	8			6		5,7	8
D		4,7		4,7		4,8	5,7		5,7		8	8	5,7	8	8				8			4,6	8	6
E	8	6			8			4,8			6	5,7			8			8	8	8	5,7	8		
F	6	8			10			8		4,6		8	4	6		5,7	6	6		6	8	6		
G	8		9			8	6	8	9	6		6	8	4,8	4,6,8		8	6			6	9		6
H	9	5,7		4,5,7					4,5,7	10				9		5,7		10		10			4,5,7	4,5,7
I	10					2		2	4,5,7				4,5,7		4,5,7	5,10	5,10	4,5,7	5		4,5,7	4,5,7		
J	5,7		4,5,7	10				10		4,5,7		5,7			5,7					4,5,7	5,7	5,7	10	
K					10	4,5,7		5,7	4,5,7	5,7	2				4,5,7			4,5,7		10		5,7	5,7	5,10
L			4,5,7	4,5,7		4,5,7			10			4,5,7							4,5,7				5,7	
M			5,7		4,5,7	4,5,7		9	4,5,7	5	5	10	5	7	5	7		5						
N									7	7	7		7		7	2					6	4,6		
O																								
P	2																				2			

c102 65F3/44E10: 2-HIV *tat* intron, 3-COT 1 DNA, 4-c102 65F3/44E10BB 3^0 EX1/2, 5-c102 65F3/44E10S 3^0 EX1/2, 6-c102 65F3/44E10-D22, 7-c102 65F3/44E10-K6, 8-c102 65F3/44E10-A1, 9-c102 65F3/44E10-A6, 10-c102 65F3/44E10-I16.

75G13

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
A			6	9					4,5,6							4,5,6	9	4,5,6			5,6		11	6
B	6	9		6	5,10						4,5,6	4,5,6	10	5,7		4,5,6		5	4,7,10		4,9		1,4,5,6	10
C	7	8	1,5,6															4,5,6	4,9				4,6	
D			1,6			1s,5,6													6	4,5		10		
E																	4,5,6,9		4,5,6					
F	7	1,6							10			1,4,9				4,9						1,5,6	10	
G		10				5,6	4,5,6,9	5,6										4,5,6			4,5,6			
H	11	3	1,8		10	3	1,11				1,4,6				4,6	11	4,5,6	6	8			8	8	
I		8	1,8,11	1,8		8		7			8	1	7		6	6,8				4,5,6,7	3	4,6		
J			8		7	1,4,6					11	8	11	7	7	7		8	1	1			4,6	8
K	7	1,5	1,4,6	11	8		4,5,6			7	1,5			8		4,5,6								
L	4,6	11	1,11	7	1	11	6	8			7	8	7	5,6	5,6	10	8		7		11		6	5,6
M		7	7		7	6		7	11			5,7	1,4,5,6			8			7			8	1,5,7	
N			1		1,4,5,6	4,6	11	11		3	8	8				5,8	5		8	1	1,7	4,6		
O																								
P												2		6										

75G13: 1-insert 75G13, 2-HIV *tat* intron, 3-COT 1 DNA, 4-75G13BB 3^0 EX1/2, 5-75G13S 3^0 EX1/2, 6-75G13-D6, 7-75G13-I20, 8-75G13-N5, 9-75G13-F12, 10-75G13-B5, 11-75G13-H7.

Table 2

265A22

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
A	5,8					4,5,8			5,8			4,10		4,9					4,10			10	10	
B	9		4,5,8	10				4,10			4,5,8	4,10		5,8				4,10	4,10	4,10	5,8			
C	8	4,10	4,10			4,5,8,10	9	4,10	4,10		8	8	9	9			4,5,8	4,9			1,4,5,6,9		1,4,5,6	
D	4,10	9	1	8			4,10	4,10		4,9	10	4,5,10	1,4,5,6	5,8	4,6				10				5,8	10
E		4,10	1			8	4,5,6				1,4,6	4,10	4,9		4,10						10	5,8	5,8	
F	4,5,6		1,9	10		5,8		9	1,4,6	10		1,4,6			4,10	4,10				1,9	5,8		5,8	
G	10		1	9	5,8	5,8	1,4,5,6	5,8					4,10	9						10		5,8	1,4,9	4,9
H									4,5,8		7													
I			1,4,5,8				4,5,8			8			9		1,5,7		4,8							8
J								4,5,8									4,5,8		4,8			9	4,8	8
K					4,5,8		4,5,8					4,5,8	4,8						4,5,8	1,5,7		4,8		
L					9				8			5,8						4,8						
M	1,4,9		1				7		1,4,5,6,8					7		1	5,8		4,8	8			1,7	
N											4,5,8	5,8		4,8	7	4,5,8		8						
O																								
P																2								

265A22: 1-insert 265A22, 2-HIV *tat* intron, 3-COT 1 DNA, 4-265A22BB 3^0 EX1/2, 5-265A22S 3^0 EX1/2, 6-265A22-F9, 7-265A22 -K20, 8-265A22-I3, 9-265A22-G23 & -F20, 10-265A22-A12.

190A9

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
A			1,7	8	1,4,7		5,9,10			8	6		9,10				6		1,4,7			4,7		8
B	8		1,4,7	10	4			4,9,10	1	1	9,10							9,10	1,9,10	1				1,8
C	5,9,10						1,4,7	1,8		1,4,9,10		1,9,10	10	1		8	9,10	9,10	9,10				1,4,7	1
D		9,10			1,10				10	1								8	1					9,10
E		1,4,9,10						9,10					10	10	9			10	9	9,10			9,10	
F		5,9,10	1,4,7	5,6			6	4,7	10			9,10			9,10			1,4,7		9,10		1	9,10	1,7
G			9,10								9,10	9,10			8	9,10	1,4,7	10		4,7	7			
H		4,9,10		6		3	6			6		5,6	5,6	5,6				1,8	9,10		5,9,10		9,10	6
I		5,6				3	3	6		5,6		6		4,8	6	2		5,6	5,6					9,10
J		6	6	4,5,6	5,9,10	9,10	5,6	3		6	9,10	6				5,6	2			5,6			3,6	9,10
K							1,4,8			2	6	9,10	6		5,6		1,6,9,10	6	1,8	6	10			
L	6	5,9,10	6	5,6	6	5,9,10				6	9,10	9,10	3	6		6		6		6				
M					9,10	5,9,10	5,6				6			6	9,10	2		9	8		6		6	
N	6	4,5,9,10		6	5,9,10	6		6	9,10	6		8	6	9,10	10	9,10	9,10							
O																								
P											2			2									2	

190A9: 1-insert 190A9, 2-HIV *tat* intron, 3-COT 1 DNA, 4-190A9BB 3^0 EX1/2, 5-190A9S 3^0 EX1/2, 6-190A9-H12, 7-190A9-A3, 8-190A3-H18, 9-190A3-E2, 10-190A3-H21.

Table 2

111K10

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
A			4,7		1,4,7					3	4,7						4,7			3,4,7			4,7	
B					1,4,8			3					4,7		4,7					3	4,7			
C				1	1,4,7																			
D					4,7				1,4	4,7	1,4				4,7						2,3			3
E		1,8		4,7	4,7		3		4,7		3	4,7		4,7			4,7	4,7						
F			4,7								4,7		3						1,4,7					
G		1,4,7	3							1,4,7			3											
H	5,6,10	5,6,10			5,9,10			2			5,6,10	5,9,10		2	9,10	6,10	2			1,6,10		1		9,10
I		2											1,5,6,10					9,10		1,9,10		1,6,10		
J	5		6,10	5,9,10	5	2				9,10	5	1,4	4											
K		5,6,10										1,5,6,10	5,9,10		1,5,6,10				1,5,6,10		2			
L						5,9,10	6,10	5,6,10								1,5,9,10			1,5,6,10					5,6,10
M			5,6,10			5,6,10					2								1	1,9,10	6,10	6,10		5,6,10
N	5,6,10	1									4		2			1			1,5,6,10		4			1,5
O																								2
P														2										2

111K10: 1-insert 111K10, 2-HIV *tat* intron, 3-COT 1 DNA, 4-111K10BB 3^0 EX1/2, 5-111K10S 3^0 EX1/2, 6-111K10-K12, 7-111K10 -A5, 8-111K10-B5, 9-111K10-H5, 10-image 824710.

73M5

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
A	4,8,14	4,9,13		4,9,13	4,8,14	4,5,8,14		4,9,13	4,9,13	4,8,14	1,3	9,13	4,8,14			9,13	1,3			3		4,5,8,9,14		4,8,14
B	9,13	9,13	3	4,8,14	4,8,14	4,9,13	1,3		9,13	9,13	4,8,14		4,8,14	9,13		4,5,8,14	4,5,8,14	12	12				4,8,14	
C	12	4,8,14	4,8,14	4,8	4,9,13		4,9,13	4,8,14		4,8,14	4,8,14	4,8,14		9	4,5,8,9,13,14	9,13				4,5,8,14	3	4,5,8,14	1	8,14
D	4,8,9,13,14		4,8,14	1,3,4,9,13	1,4,9,12,13	4,8,14		1,3,4,5,8,14	4,5,8,14	4,5,8,9,13,14	9,13	1,8,12		1,3,4,9,13	4,9,13	4,9,13	4,5,8,9,13,14	4,9,13	4,5,8,14	4,5,8,9,13,14	8,14	1,9	9,13	4,8,12
E		12	4,8,14	4,8,14	3	12		9,13	9,13	9,13	4,5,8,14		1,4,9,12,13	4,5,8,14	4,8,14	4,8,14	4,9,13			4,8,14		4,8,14	4,9,13	
F	3	1,3		4,8,14		4,8,14	1,4,12	4,14		3	1,3		4,9,13		9,13	4,5,8,14	4,5,8,14		4,8,14	4,9,13			4,8,14	
G		4,8,14	4,8,14	4,9,13	4,9,13	1	4,12		3	1		4,8,14		4,9,13	4,9,13	4,5,8,12		4,9,13	4,5,8,14				4,8,14	
H		4,9,13	4,8	4,8	4,8		13,15			1,6,10			1,4	14				7,13	13,15			13,15	5,11	
I											4,5,8,14		1,5,7,13		13,15	6,13	16				9,13			7,13
J							1,13,15,16	4,8,14			4,9,13		4,5,8,14			5,11		13,15			11			4,9,13
K													13,16		13,16	1,10,14	5,11	11				13,16		
L					1		13,15					1							8,14	2	13,15	7,13		5,11
M		1	13,15			4,8,14	1			4,8,14				7,13,15	4,9,13	1		13,16	5,11			11		4,9,13
N		11				13,15		4,8,14			11		13,15		7,13								5,11	
O																								
P																								

73M5: 1-insert 73M5, 2-HIV *tat* intron, 3-COT 1 DNA, 4-73M5BB 3^0 EX1/2, 5-73M5S 3^0 EX1/2, 6-73M5-H10, 7-73M5-I13, 8-73M5-J13, 9-73M5-A2, 10-73M5-K16, 11-73M5-K17, 12-73M5-F7, 13-image 824710, 14-111K10-K12, 15-73M5-H7, 16-73M5-K13.

Table 2

135E14

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
A	12	12	1,7	11	12	12		12	12	10,12	7,10,12	10,12	12	11	10,12	12	12	10		11	10,12	12	1,12	1,12
B		7	14	6,9		12	5,6,8		14							6,9	12		6,9				6,9	
C				12	12	13,14	12	12	12	12	12	12	12	12	12	12	12		12	12	12	12		
D				5		1,10		10	12			6,9	6,9	5,8	10	5,8		6,9	10	6,9	12	12,14		
E	11	12	12	12		1	12	12		12	12	12	12	12	12	12	12	12	12	11		11		12
F			12	6,9	12	7	10		5,8	6,9	12	12	6,9	6,9		10	7	6,9		10		10	12	10
G	12	12	12	12	12	1,12	12	12	12	12	7,12	5,12	12	12	12			12	12	12	12	12		12
H		10		7		1	6,9		7,10		12	6,9,12	12	12		14	10	10			8	5,8		
I	12		12	12		12	12	7		12	12	12	12	11	12		12			12		1,12		12
J		1	5,8	7		6,9	12			5	5,8,12	5,8,10,12	12	10,12	7	10			12	10				
K	12	11	12	12	12	11	12	11	12	1,13,14	12	10,12	12	12	12	12	12	12		12			12	12
L	1	12			12		5,8	7	12	6,9		6,9	5,8	12	7	6,9	5,8			12		6	10	
M	12	12	12	12		12	11	13,14	1,12,13,14		12	12	12	12	12		12	12	12	12	12	12	12	12
N	12		8	10	6,9	12	12	7			10		12	6,9	10	5,8		6,9	10	7		10		
O	1,13,14	12	12				12	12	12	7,12			1,11		12	12	12		10		12		11	12
P	8	8	10	8	8			6,9	10	8	6,9	8			6,9,12		10				10		7	

135E14: 1-insert 135E14, 2-HIV *tat* intron, 3-COT 1 DNA, 4-135E14BB 3^0 EX1/2, 5-135E14S 3^0 EX1/2, 6-image 307735, 7-135E14 -A3, 8-135E14-B7, 9-135E14-B4, 10-135E14-D6, 11-135E14-O13, 12-135E14-G12, 13-image 299864, 14-135E14-C6.

Table 2 : Data, showing hybridisation of insert, HIV *tat* intron, COT 1 DNA, 3^0 BB EX1/EX2, 3^0 S EX1/EX2, putative exons, Image cDNA clones that showed homology to putative exons and cDNA clones that were mapped in Unigene clusters, to the filters containing the exon trapped products. Above each table is indicated which genomic clone the products were derived from. The lists below the tables show all the probes that were hybridised to the individual filters; the numbers of these probes correspond to the numbers in the tables and indicate which clones the probes hybridised to.

Sequences of the exon trapped clones

90B5BB-A22
AGTAATTACTTTAGGCTATAAGAGACCTTTGGAAAGAGAGGATCTTTTTGAACTAAAGGAAA
GTGATTCCTTCTGCACTGCGTGTCCCATCTTTGAAAAACAATGGAGAAAGGAAGTTTTAAGG
AATCAAGAGAGGCAAAAAGTAAAGTTATGTAGGTATCTTGGTATAAAGAGGCACATATCAA
GAACCATCTCTACTCTATGCATTGTGGACACCTTTAAATCCATCCTGATTCAAGTTGCCTTTA
TTCAAAGTGTTTGCTGATATTTTGTCCTTCACTAGCCCACTCATAATGAA

90B5BB-C21
CTGTAGTTGTGTGAATTTATCTTAACACAGCTCTCAAAACTTTGGTTAGCTGGTTCAGAAGAG
CGAGAACTCCCAAATCGATACTGCACGTTAATTGGTCTTCCATATAAACGAATTCCATTCAG
CAAAGCTATGGCATAAGACACCGATTCTGGGTGTTTAAAGCAGACAAATCCAAAAGACTTT
GGCTTTCCTTCTCTGTCTTTGCATATAGTCACTTGGTAGTGGCCCCGCCTATAAGAAA

98L15BB-J11
TCATTGAGGTTTCTCTTTGGTAGGATCATCACAGGCNTGGCAGGTTCTTCTGTAAAAGTTCAC
TGATTTTTGCTCCTCCCTCTTGGTTGAAAAGGTGAGGTAGTTTGTTTTCTCCCACTTTAGCTTG
TGGTTGTGGGAAGGATTTTTAAAGTCACAGGCT

98L15BB-J19/98L15S-K10
CTTAGCTGAGTTCGTAGAATGAGATGCCAGCCTTCTGCATAGAACACATACCACACTCAGTA
AGGGGTTCCAGTGAAGCAGATTCCATTTCCAC

98L15BB-K11
TAGTTGCTCCAGACGACTTCTGTTGGGGAAAATAAATGGACTGCAAAACTTATACAGCATCA
ACAGAATGAATCTTTCCAACAAGCCGAAACATTGAGTATTGTGGCACAGAATATGCCCCACC
CATTACTCAATCTAGATATCCTTTTATTCCACCGTCTCATGATTTTCTTTTTC

98L15BB-N7~98L15S-C3
CTGTTATTCTTCCAAGTTGACCATGAAATATCTTTCCAACAAATCCACTGATATGAGCCCCTA
AGCTCACACCTATGAAATGAAAATTGTCAAGAGATGCACCATGCTTCAAAAGATTTTTAATG
TGCACACTCAAACTCACAGCAACTTTTCTGGTGTTTTTAACTGCTCTATTATAAATAAAAGTT
GTAGCACCCCGGCTCCAGTCTACTACAATTACATTCATATCTTCTTCATTCAGCAAAATCCTT
ACGAAGTTCTGAAGCCATAATGGGATGGAGCCTACTGGTCTGTATCCGTGAATAAGCCAGAC
TGTTTTCTTTTGTGTGTTGAAATTAACATTAAGTGAGTTATTTTGTTCAAACAGTGGCTCAGC
ACAGTTTAGGTTGTTCCTTGTATACATCATCAGAATGGTCTCTATTCTCGGAATAAATAAATC
TCTGAAGGAATCCTTTACACTTAGCTGAGAGAATTCAAGGCATGGTCTTTTATTAT

90B5S-B24
CGTGTAGTCAATGTACCTCATCTATCAAGCATATTCCCTGTGAGATTTGTTCTTCCTATTCCTC
GTCTCTTTAAGTATTGTGAAC

90B5S-E6
CTTAAATGGGTATCTGCCAATTTCAGCCTCCAACTCTTCTGCGGTAAAAATCTTGCCTATGAA
TCTTTTGGCATCATATATTGTGTTTTGAGGATTTGAATCTGCCAGCTCTACGCTTTCATATCCC
ACATATACATCATTGTCAGTAAAAGACACCATGCTGGGTATGCTGATATGCCCATTTTCATCT
GGAATCACCTTTACTTTTCCTGTGCCAGGAAAAAACACCCCAACAGAACAATAGGTGGTGCC
AAGATCAATACCAATCACTTTAGGAGTAGGCAATGGTAAATACTGTTGTGCCAAATAGCCGG
CCAACAGGAGAGTCAAAACAGCCGATC

90B5S-E17
CTTAAATGTCAATGGGATAAATGCCCCAATTGAAAGACACAGAGTGGCAAACTGGATAAAG
TGCCAAGACCCATTTGTATGCTGTCTTCAAAAGACCCATCTCACATGCAATGACAAACATGG
GCTCAAAATAAAGGGATGGAGAAGAAT

90B5S-G5
TGGCAATGGTGTTAATGGCGTACTGAAAGGTGCAGCTGTTGGCGCTGGTGTTGGTGCAGTAA
CAGAAAAAGGCCGCGACGGTAAAAATGCGCGTAAAGGTGCAAAAGTGGGTGCCGCCGTTGG
TGCTGTCACTGGCGTTCTGACCGGCAATGGCCTCGAAGGGGCGATTAAAGGTGCCGTGATTG
GCGGTACGGTGGTGCAATT

90B5S-J11
CCGCTCTCTCTCTGTTTCTCCTGCTCCGGCGGTGTAAGAAGTGTCTGCTTCCCCTTCGCCTTCC
TCCATGATTGTAAGTTTCCTGAGGCATCCCCAGCCATGCTTCCTGTACAGG

98L15S-A9~98L15BB-L11
CTTTATCAGATGACAGCTCCACTTCCTAATAGTGCATCCGTGTCTTCCTCACTGAATCATGTT
CCAGATCTTGAGGCTGGACCCAGCTCATATAAATGGACTCACCAACAACCAAGTGACTCTGA
C

98L15S-B8
CAGCAAGTCATGAACATCTCAGGGCTCACTTTCTCAC

98L15S-E9
CTGTTCTTTCTTTGTAGTCTTCTTCTCATATTTAATTAATGTTGAAGCATCTTGTGTC

98L15S-G9
CTGCTAAACCAGGAATATGTCACTTTACTGAAAAATGAGGCACTTTTCTCTGGACTGCATTT

126N20-A3
TTTCATATACACTCTTTATTTACTCTTTTCATTTATGCATTCATTTAATATATTTAAGCATCCA
CAAGGTATTTAGCACTGTTGTAGGAAGTGAATATACAGCAAAGACCAAGACGACAAAGCTC
CGTCTTCCATGATGCCCCATTTCAGAGTCTGTGTTGTACTTCCCTGTTTACTTCTCTTTCTTCC
CTGTGTCTTGGGCACATCCACCAC

126N20-A22
AGCTAAAGTAATTTGTTTTTTTGTCACTGAGTCTTGTGTTATTCCAGCATCCGTTGTCATTTCA
TAGAAAAGTCCTTTCTGACGTATCAAATTCTGAGGCGCTTCGAATTCCACAATGCTTCCTGA
GTCTAGAACATGCACC

126N20-C4
CAGATAGGCCAGGTGGGATCATGCAAAAAGCAGAATCTTTGGAGCCAGACAGACCTGCCTT
CACATGCCAATACACAATCTGAGAGAACATTCTCCAGGTCTTGTGAGCCTCAGGCAACTCAG
CTCAAAAATTGGGATCACAATCCCTTCTTTGCAGGG

126N20-E1
CTTCGTGTGGTGTAGACAAGAACTCTTTGGAACCTCCAGCTCACTTT

126N20-N22
CTGAGGTTCTCGCCACCCTCTGAAATCTCAGTAGCTTCTCAGGAAGGGACTGCACAAATTGT
TTCAAGTGACACAATTCCAGCACTTCCCACAGCTTGCTATCAGAATATTTGTTTAGGGGATCC
AGGTTCATTTGAAGGGTTCCAGAAAATAAAACAGGGTGCCAAAGCAGGAAATAGTCCTGCA
TAATCTTACTGTGGAATAATATTAAGTTTGCCATGAAGGTCATGAAGTCCAATAGTTGATAC
ATCAATTCCGTCAATAATAATTTCGCCTCCTGCTCTCTCCACAACTTCTAAATAAGCAATTTG
ATAGTGTGGATTTGCCTGCTCCAGTCCTGCCCACAATTCCAAT

31B5-A3
CTTGATTTTGAGGGAGTCAGTGTCATAGGGACTTGGCGTGAAATCCGTATGCACTCTGGCAC
GGCCACAGAATGGTCCTGAATAGGGGCCATCGTCATCCAGTCGAAAGCTGTCCCGGTTACTT
GTACCATCTGAACAGCTTGTTATGCCAC

39I12-F7
CTTGATTTTGAGGGAGTCAGTGTCATAGGGACTTGGCGTGAAATCCGTATGCACTCTGGCAC
GGCCACAGAATGGTCCTGAATAGGGGCCATCGTCATCCAGTCGAAAGCTGTCCCGGTTACTT
GTACCATCTGAACAGCTTGTTATGCCACCAGTTCATCAACAGTGGTCTCAACATGTCTTTGAT
AGTGAAATGGACTCCAACATTCAGAGGTCAC

Table 3

31B5E21
TTTCTTCTTCAAGGAAGTTTTCAGCAGCTGATAGTAACCTTCTTCTGTCATCTGGGTTTTCAAT
ATTTAATTCAATGAGGTGACTCTCTTTTATATCTTTTAAATCTTCTAGAGTCTCATAACCATTG
AGCAAAAGTGTTGAGGTGTATTCCTGCAGATGAATCCTCTCTAGGAACTCCTGCAGAGTCTT
GGATTTTTTGCTGTTACTCCTTCGGTTTGCCTTTATTTTCTTGGGGGCTGCTTCCTCTTCTGAG
ATGACATCCACATAAATGAATTTGAAGTTTCCCACTTTATTGTTCAACATTCCTGTCCACATC
CCCATTGGTGTTTTGCAAATAATGTCTATGATGTCTCCTTT

31B5-M21
CTTTGTAATTATCATCAGGCAATGATGAGTCCTGGGGAAAGTATGGATCTTAAATTCGAGGA
TCTCTTCTGCACC

39I12-A1
TTTTTTCCATGTGTTAGGATTTGTATTTCCGAATGAAGCTCTTCCTACTGTAGTCTTCTAGAGC
CTTTTGTCTTGGTTGGCTGCAGAGCTTATCACGACCTATTTGGTCATTTGCTTCCCTCTCCTGT
TATA

39I12-D2
CCATGAATGTAGGAGATCCATATCTGAGCAGGAAGACTTGAGACGTGTGTGGCGAATACAA
GAGGAGCAGTGGTCCCAGGGTCCAACTTGTGCTATTTGGAAATCAAGACATGTCCAGAGAG
GGTTTTCAGGAAAAGGCTTCCAGTGATGCCACATGTAAGCTTCACCTTCATACCAAGCTAAT
TTATCTTCCAAGTTGCAGTTATTTAAAGTGCTGGATCTACCCAGTGATGCAGAAAGAGTATC
CAATCTAATCTCCATTTCTTCCTTCCTCCTCGTGCTAACACTATTCACTTTCTTTTCTCGGGAC
TTGGTTAATTTCAAACAAGAAATCAACTGTGTCATCTTCATCTT

39I12-E18
TTGAGAGCCAGGTACTTGTCCAAGGTGCTCATGTCCATTATTATGTTTAATCTTCTCAAAACC
CCAGGAGGTAGACATTATTTTTATCTCCAATTTACAAAGAAGAAAAAAGCCTGTGGTCCTAT
AGAGCAACAGCACGGGCATTCCAGGAGCTTGTTAGCCATGTGCATTTTCCAGTCTCCCCCAG
GCTTTATCCATTTGGATGTTTATAGGTAACTTACTACATGCTGTCTAAAGATGGCAGACAGA
AAGTTTCCACTGTCCGGTCCCATTCTTGCTTCCAGTTGGTGCTTGCATGCCAAGGAGAAATGG
CCTCATTCCATTGTTTACAGTCACCAGAAATTCACCCCTAAATTGATGTCTCACTTTGCAAAG
GCTTCTTCTTCTTCTTT

68C10/132H4=190N10-J9
CACGAAGAGATTGTTTGGTGTTTGTTCACTCTCTCCAAAGGGTGGGATAGGAGAATCAGTTC
TACTAGACATGTTTTCTTGGTTAGCTTGTTGCTCTGGGATTGGCTCATATAAGCTATCCATAG
AGCCCTC

190N10-H1
CTTGTTGGCAATGTTCACAGAAATAAAGCTTTGGTCAAAGGAACCAGTTTCTCTTTGGCATA
CAGATGAATGACATTCAGAGGCATC

190N10-H8
AGTTCTGGAGAGCTGGAAGTCTGAGATCAGGATGCCGCCTCGTTGAGTTTTCCTAAGGATCT
TCTTCCTTGTCGCAGATGATTGTCTTTTCTTTGCATC

190N10-K14
CTTGGGTTGGAGAGTGGAGATGGTACTGGCTGAGAGGGACACAACGTGGTCTTCTGAAGCA
CTAGAAATTTGTATCTTCATTTGGATGGTGGTTCTATAGATGTGT

153I22-A2
CTCATCTCTCATCTTGGCTAAGCCAATGAACAAACACCAATGATTAAATGAGTCCAGGATGT
GAATCAACAGACCAGGTATCTTCCAGGGACCATTTTTTGGGGCTTGGGATTCCCTAAGTAGT
CCATTTAGGGGCCCTGGGAAGTTTGCCACATCCCATGTGAAAGCAACACGTTCTCTTGGGCC
TGATGTAATAACTGACACATAATCCATTCCAGATTTCTTTT

153I22-A6
AATCATGGTGGAAGAGGAAGCAAACATGTCTTTCTTCACATGGTGGCAGGAATGAGAAGTG
CAGAACAAAAGGGGGAAAAGCCCCTTATAAAACCATCGGATCTCATTGTGCCATTGCACTCC
AGC

153I22-A23
CTTCCGAGTTGTAGTTTGACAGCTACGTTGATAAGCTAAATCCAAATAAAATTAGCCTTCAG
TTAGATGCTCTCACGTTACTGTAATAAACTTCCTGGGACATACAGTGTCCTTTGGAGGACCA
GGAGCCTCGCCCAAACAGACTTGAACAACTGTTGACCAGACTTGACAGCATGCAGTCCACA
GAAGTGC

153I22-H13
CCAGGCCCATTTCAATCTTCTCAATACTAACAGGCAACTTCCACAGTTAACTCTGGGCTTCTT
GTGTAACAAGTTTGTTACTTTGTCGTTGGTCATGAAGCAGGGCTCCTTAAAGATCTTCAGGTA
TGGACACAGGTAGAGACAATTCATTTATC

153I22-M15
GGTTATGGCAATTTCTACAATGAAGATAAACAGGGATTATGCAGTTTGGTTGACATTCAGGT
CTCCAGTATGTGCTCTTTGGCCTTTCGTGGTCACACTATGTGCTGCTCTGCTTCTATTTATATC
ATCTTCATCAAATAACAGGC

30P13-A1
GAGCTCCAAGTTGAACAACTATCCACACAGAAAACCACCCTCATAAGAACCAAAAATCAGC
TTGGTTACCAGCTCAGCCACAGTGGGATAGAGCACCAAGCAGGCACCTGGGGTCCTTGATTC
CAGGCCTTGGATTCTGGATGGCATTTCTGGACAAGTCCTGGGCCAGAGGGGAGTCAACTGCC
CTGAAGAGAAAGACCCAGGCCTGGCAGCATTTACCACAAGCTGACTTCAGAGTCCTTGGGC
CTTGAGTCAACATAGGAG

30P13-B4=270M7A3
CTGAAATCTGGAATCAATATTTATAAATATTTCCAGGTTCATCCATGAACCTGTTGATGAGCT
GTCAGTTTCAGAGCGCCTTGTCCTTTTCAGATATATACTC

30P13-E4
CTGCACCAACATCTACAAAGGAAATAAAGCGCTGAGTTTTGTCTGCCAGTGAGCC

30P13-F11
CTCAGGAATGACTTGGTGTTGTCTTCATGGTAGTGAATGAGTTCTCACTCTACGAGTTCACGT
GATGTACTGGCTCCCTCTTTGCCTTCCACCATGACTGTAAGCTTCCTGAGGCCTCACCACAAG
CCAAACAGATGCACGTGCCATGCTTGCACAAC

30P13-H3
AGGCAATCTTCTCTTTGTACTTGATTTTAAAATATTAATATTTCATTCATGGTGTCTTCTCAAA
TTAAATTTCTCAGGGCTACTGGACCTCAGATA

270M7-D15
CTGCTTCATCGTGCACTTTTATGTTATCCAGGCAGCTTCTATCCTTAAACTTCATGAACAAAC
CTCTGCTAGCTTCCAACTTTTCTTCTGCAGTTTCTTCATCTCCTCTCGGC

270M7-L10
CGCCCTCCTCCTTGTATTGGTCGGGACCTGTGCTTGGCACCACTGCTGTGAGAGAAAGGTAG
GGCCGTAATCTGTTTTATTGTATTGCATATTGTGACACCCCAGGAATATTCAACATGCGCTCT
TTCTAGGGCACGTGCTGTAAGT

270M7-M11
AGTGCTAGGTAGCTACTCACTTCCAGGAGTGATGAGAACTTCCACATTTGAAAGATAAGCTT
CCATTCCAAATGTTTCCCCCATTTGTTGGGGATAGATT

Table 3

270M7-N2
TATCTGAGGTCCAGTAACCCTGAGAAATTTAATTTGAGAAGACACCATGAATGAAATATTAA
TATTTTAAAATCAAGTACAAAGAGAAGATTGCCTGTGAGTGGTTATTCTCTTCAGTTGTCCA
ATAGCTGTTATTTATCAAATTAACACCAAAACGTTCAAACTCTCTAAAAATGAATAGCTTTG
CTTTATAAGGAGAGTTGTGAAAACTTATTAAGAACTCCTGCACATATTATTTAAAGATACCA
TTTACACGTGAGAGTAATTATAAATGAAATATATTAAATGTCCGAAAGTAGCACCATGTGGC
AGAATTTTCAGGAAAAGAAATCGAATTAGTACTGGAAGAAGACATTCAAATAAGAGAACCA
CATTTTCAAG

194B7-B5
CTGAAAGAACGGGTCTTTTAATCAATCGGTAGGTAATATTTTTCCATGGAGCATGTGGTTCTT
GAGGACAAATCTGTCCAGCTTTCGTTAATGTGGGTCTCATGAAACTGCGAGGGTGCAAGTGC

194B7-C8
CTGAGATTCATGGGTCCAGGTTGCAGTAGTAAAAGTCCCTTGGCTCAGGCTTCCTCTGTTGCT
CCCAGTAAGAATGCAGAGAAAAAAAAAAAATCCGAGTTTTAGGCTGTTGTGT

194B7-H4
TTAGTAAGACAAAAATCGGAACAAATGCTGAAAACCCCAGCTGCATGCTTTAGCTGACCAG
ATTCCAACAAGGAAATTCTCTATAGTTTCCATGAGGTGGAAGATACTCCAATGTCTTTCTTCT
GTGTTGAAGAACTGGGAAAACTGGATTCCAATTCTCACTGTTCATTGCAGAAGAGTGGGGTG
GCTCAGCCTTCAATGATCTCAC

194B7-K5
TCTTTATCTGTGTATGGCTGAACAATGATCTCTTCAGAGGCACCACAAAAGCTTCCCACCAA
GATTCTGTTTGGGTAA

194B7-L21
TATAAGCAGAGATTAGGTGTCTCCTTTCACACACTTGACTTGGAGAAATTGGCAAAAACTCT
TCCTCTCCCTCAATTCCATGGCTTGCTCATGACCAAAACCTTCCCCCAGGCTTTTAAATATAC

103K19-A12
CTATATATTTTTGGATGATGCGCACTTCCTACTGAAAGAAGCTCCTCTTCATGTAGCTGACAA
TGAAGATTACAAGGGTAGAGCCTGGCTTACTTGTGTCAAGACGATATGAGTTAACTGAAGGT
CACTTTGAGCACACTTTATTTACCATGCCTCCCAGGCACTG

103K19-B3
CTTTCTCAGGGACGCAATCCTCAGTGATTTAGTCAAGGTGCTATATCCTTTTCAGGAAGGTTC
CCAAGTAATTCCATGGTTCCATTGTCTCTACCTTTCTAAATCAGGCTACTCTCTGTGACTCCA
AAGTGATTTTTCATCCTGGCCTAAGTCATGAACTCCTGGGCATTTTTTTCCCGAGAATTCCAA
GTCTAAGCGTTCTTGTCAGGGCTCACTCTTTCTATGTCTTAGATCCACATGACTTTGTTTCTTT
TTAAGAATTGACTGAAGCTTCTCACCAGGTGCATTGGCAGGAAAGGTTGTCAAAGCTCATAC
C

103K19-F4
ATTTAACAATTATCAAAATTTGGCGCCAGGTATGGTGGCTCACATCTGTAATCCTGCCACTTT
GGAAGGCTGAGGTGTGTGGATTGGTTGAGCCCAGGAGTCTGAGACCAGCCTAGGCAGCACA
GTGAGACCTGGTCTTGTTTCCTTGATTTCCGTTTTCATTGTTGGCATTTTCTTGTATTCTCTTTG
TTTTGTAGCACCAACTTTC

103K19-J5
CTTAATGTAGCTCTGAAATAGAGAGGCATGCTGTGGTTCTGGCTCTTCACTTCAGGGAGGGA
GGGGGGTGTAGGAGGAGTATCCTGTGACATGGACAGTTAGTGGCCTCTCTGCTGCATCTGTC
TTCTTCTAATGGGATGGTCAC

103K19-J13
CTGGTTTCTCCTTGAAAAAAGAATTTGAAGGACTGATTTTGATGTTAGCTTCTTGCAGTTGCT
CCAAGGCACCAAAGGATTTCTTTACTCAGATT

103K19-K10
CTTGAAGCACAATTTTAATGGCAGCATGATTAGCATCTGGAAAACTAGGGGGCAGACAGAC
CAGCCTCCTTTGAGGTGACCT

10J13-B24
CTGGTAATCTCTCCATTTCTTCTTCATCTGTCTTTCCCCTGGAAGCATCCTTCCGGTGGTGGGG
GCAGGGGGTGCTTTTGTCCAATCTGAAATGGCTCTCCTCTGTTTGAAAGCTTCTTCTCAACTA
TAATCTTGGACCCTCTGTACCTCCTGGTCATTTCTTTTGTCTCTCTCGATTTGGATTACCTGCT
TTCTGATCTTTAGTCTTGCTCCTTCTTGGTTTATTCCATTATGGGTAGAGCATCCACTAGTT

10J13-H23
CTGCCTCCTTGCTTGACTGTGTTCTGAAGGAATTCAGGAGCAGAACTGCTGCCATCTGCAGA
GACTTTGAATCCAGGCTTGAAGAAGGGACACTGCAGTTGATTCAGAAGAAGGTTTTTCCAGG
CATTATCCCAGAATCTGTCTCCCTGGAGATTGAGGGCCTCCCAGCTGAGCAAAGGCTGCAGC
AGCACTCCCCTGCTGGATCTTCTGGCAGCCATTTTACTTGGGATTCAGTC

10J13-L3
CTCTCTTTCTCAGTATTAAACATACTATTTGGATTATAATACAGTCTTCATAAACACTAGTTG
AGTGAAGAAAGCACATGGAACCATGATTTTCATC

10J13-N5
CTTGGTCCACTTGACAAACGATGTTCTAAGAGTGAAGGTCAAGTGGGCCTTTTTCTTTACACT
CCACAAGTAGCTCAAAACATTCCACACTGCAGTCTTTGTCAGATTTGGTCCGTGGCTTCCTGA
TACAGGATTTTGCTCTTCTTTGAGTTTCCAGTTTTCTTTTTTAATTTCC

65F3/44E10-A1
AGCTCTGCAGGCTGTACACACATGGTGCCTGCGTCTGCCCGGCTTCTTGGGAGGTCTCAGGG
AGCTTTTACTCATGGTGGAAGGTGAAGCAGGAGCAGGTGTCACA

65F3/44E10-A6
CTAGGAGATGCTTGAGGAATTGGAAGGACCAACCCAGAGTCTATCTCTCTTTGTTGTTGCTT
ACTGAGTGTTCTCCTTGGTGGAAGTGGCTTCTATTCCATTATAGGCAGAGCACTGCAGGGCT
GTTTATTTCCTGTAT

65F3/44E10-D22
CAAATAACATAGATTCAAATTAGTAAAGCAAAAGCAAATGGAAGGAGACGTCTGTCCTTGG
ATAGGATTATGCAGCTTCATAAAGATATCAGTTTTTTTCAGTTTCATGTATAAATTTAGTGCA
ATTCAAATACAAATAAGTGTTCATTTGTTTGCTTGTTTTTCTCACATAATGCATCAGAAAGAA
CTCCAAATGGAT

65F3/44E10-I16
CTGCAATCAACAAAAAATTCCTGGTACTTTTTTTTTATACAGCCAGAAAATGTTATTATCTCT
TCTTCCCATGGAGGTTGGAATGGAGAGAAGACAGCAAGAAGTCTTGACTTTCAGGTAGAAG
ACACAATAGTCCAGGGGAGAC

65F3/44E10-K6
CTCCTAGACTTGTGCTTTCTCCAGCTTTGCATGCTCTCTTGATTACTATTAAACCTTTTCGGTT
TATCTCTAGAATATGACTTGGAAGCCAAGGTGGTTTTGTGAAATGGGTCTGAGTTTTTCCATT
GATGATATTTGAGTTCAAATCCTGGTTCAGTTTCATTGCCTTCAT

75G13-B5
CCGTGTCCCGAGGCCTAGTTGCTAACACATTGTCGTGATGAAAGAGCTCAGATCAAATCTAA
ATTTTGTGTTAGAAGTCCATTGCTGTATTAATCAT

75G13-D6=265A22-F9
CGTGGGTATTCAACTTGTGCCGGCGTCGATACTTCCACAGGGGAGTTTTCTCCAGGTGGTGC
GACAGCACAGGCAGGTTGCGCCGGTGGAACTTGATATTCCCTCACTCATGCCACCCCTA

75G13-F12
CAAATAACATAGATTCAAATTAGTAAAGCAAAAGCAAATGGAAGGAGACGTCTGTCCTTGG
ATAGGATTATGCAGCTTCATAAAGATATCAGTTTTTTTCAGTTTCATGTATAAATTTAGTGCA
ATTCAAATACAAATAAGTGTTCATTTGTTTGCTTGTTTTTCTCACATAATGCATCAGAAAGAA
CTCCAAATGGAT

75G13-H7
CCCTCTGTTGGGGTCCACAGACAACATTGGATGGAGATGGCCACCAAGTGGCCTCAGGACA
AGGAGCTCAGATTACTGCTTCCCAGGGTTTCCAGCTGAACAATGCGTATATTGTGAAGGTCC
CAT

75G13-I20=265A22-A12
AGTTCTGCAGGGCTGGGGAGGCCTCAGGAAACTTACAGTCACGGCAGAAGGGGAAGCAAAC
ACGTCCTT

75G13-N5=265A22-I3
CAGCCTCCAGAACTGAAAGAAGTAAATTTCTGTTCTTCATAAGTTACCCAGTCTCAGGTATTT
TATTATAGCAGCACCAAATGAACTAAGACACTTTCCTTCAGTCTGTTCTCAACATAGAAGCC
AGAATGATG

265A22-F20
CAGATGTGGCCCCTGGATCTTGGACTTTCCAGCCTCCAGAACTGTGAGTCAAATAAGCTTCT
AGCTTTTGTAAGTTACTCTGTCTGTAGTAGTCTGTTGTAGCAGCAAGAAGCCGACTAAGACA
CCTTATTGCGATCCAGTATCAGTTCAGAAGCAGGTATCAGGCTATGAATCACATTTTGAGTA
GCACTGGCTTAGCGGTCTTCATCCAGTGTATTTCTAGAGGTTCACTACTAATTCCTATGGAGA
CCTTTGTTTTCCAGAACAAAAATTAAATCAAGATGCAGATGAAAGTCAATACCTCCCTCTTA
AAAATAGAAATCTTCTTTGTATTCATCATGCCTTAATATTGTATTCTCTTCTTACACCAAATA
GTTTCTCTGGGTGCCTTTGAAGCAATACACAAGAACCTTGGAAGGAACAATAGCTGGTTCCA
CTGAAGGAGGAGAACTGGGCAG

265A22-G23
CAGATGTGGCCCCTGGATCTTGGACTTTCCAGCCTCCAGAACTGTGAGTCAAATAAGCTTCT
AGCTTTTGTAAGTTACTCTGTCTGTAGTGGTCTGTTGTAGCAGCAAGAAGCCGACTAAGACA
CCTTATTGGGATCCAGTATCAGCTCAGAAGCAGGTATCAGGCTATGAATCACATTTTGAGTA
GCACTGGCTTAGCGGTCTTCATCCAGTGTATTTCTAGAGGTTCACTACTAATTCCTATGGAGA
CCTTTGTTTTCCAGAACAAAAATTAAATCAAGATGCAGATGAAAGTCAATACCTCCCTCTTA
AAAATAGAAATCTTCTTTGTATTCATCATGCCTTAATATTGTATTCTCTTCTTACACCAAATA
GTTTCTCTGGGTGC

265A22-K20
CCAAAAGCCAAGACCAGAACCAAAATATATCTAAAAGGTCAAGTCCAGAGAGGTAAAGTTT
GTTAGGACGTTTATCCCGTTTCATCTTGAATGTGAAGAACATAGTATAGTATGTAGCACACC
TAAAGAGAAATAAGCCCTATCACGAGAGAAAACTTTGAGGATAACTCTTTACTTGTTTTCTC
CATTCAGAACATCAGAAACATGAATCTGGTATAGAAGGCTTTCCAAATCCAAGGTTCATTCA
TTCAGCAAATATTCACTGGGTCTGTTGTTCTAGATATCTGTCCATCGAATCCTGACAGCAGAT
TCTTCAGCTCCAGT

190A9-A3
TCAGGTGAGAGCAAATTCTCCTGACAGGCAGCCTTCTCACTGTCATATTGGCTCTTCTTCCCA
ATTAGCCTCTTCTTTTCTAATTGCTTAGCAAAGTTGTGCAGGGGATTGCTCAAAAATATGCAA
CTCTTCAGAAGCCTAGGCTTGGAACCATCATCTTGTCACTCCCGCCACATTTGATTTGCCGAA
ACATGTCCTGAGGCTAGCCCAGTTTCAGAGGAGAGGAGATGGACACCACCTC

190A9-E2
TATATTAATAGAAGGCTTGTCTGAAGCAGCATCTCTTAGCATCTCATCTGGCCCAGATGTGG
GCACTGCTGTTGAGAATGAATTGCTATCGTTCTTGTATGCACCACGAAGTAAAGATTCCAAA
CCATGTGTGGAAAATCCAGTTGAGAGAACTAAATCATCGTAACTGCTTGGCCGCCAGGCTGT
TCCTTATCCTCTTTGCCCTTTGCAGTGGGAAATAAGATACTGCCTTCCAAAAGGAATCATCAC
GGTGGGAAATTACCTTTGTTTACTTGGTTTTGACATTTGGAGAATGAGGTTTTCTCACCACAA
GATAGAATACCAAGTGAATCTTGTTCTTCACACGTTC

Table 3

190A9-H12
CTCTTATGAAAACAAATGACTCTCCAAGCCTTAAAAAAGAAGAAAATCCTGCCTTTTGCAAC
AACCTGGATGAGCCTGGAGGACATTACGCTAAGTGAAATAAGCCAGGCACAGAAAGACAAA
TATTGCATGATCTCACACATGTGGAGTCCAAAAAAGTCAAAC

190A9-H21
TTGTATTTTGAGCTGGCGTGGTGTCCTATTTGCAGGTAATTTCCCACCGTGATGATTCCTTTT
GGAAGGCAGTATCTTATTTCCCACTTCAAAAGCAGATTTCTTTTTCTTGTCTGAGAGTTCTTT
CTCCATTGCTTGAGGATTATTTTGTTGCCGCCTCACCATAGTGATTTC

111K10-A5
CCCATGAGACCAGCTGCCTCCTCAGGACAATTTCTTCCTGACACCTTTGAAACTTACGGATG
AGTTTTTCTGCTACGCACGAACTCTACTTCTCAGCTCCGCCTGCTAGCAGAACACAACATCG
AGCAACCAAGAGATGGTCTCTGGCTCCTTTCTGGATTCCTAAGCTCTGGCATCCACATTGGC
ATCCACAATAAGTTCAAAAAATTGAACTTATTTTAAATACATAATGATAGATGAATGAGAGC
CTGGGTATCGTAGTTTGGTCTTCAAGC

111K10-B5
CTGGGTTGGCTCAATCTTGCTTGTCCTGTGGTTTCAGAAATCTAAATTGATTATTTGGATATC
TTTTTCGTGTATTTTAACGCTGTGTTGTGCCCACCAAATTAGGTTGCTCAAATGTCTGATCCA
AGAGGATTATGGAAGCATTCTCGAGTATTTTCATCTTCCGTGTTCCTCTGGGCTTTCTCTCTT
GAAGTCATTTTATTAGTCATCCTCAAATGGCTTCTACTGATGTTTCTGATAACTTGATGCTGC
C

111K10-H5
TTGTATCTGCTTGGCTTCCCACACTGATGTATCTATCATTGCCAGGAAGTGCTGTTTGGTAGT
AAGTTGTCTCCTCCTGCTGAGGGGTCTTAGCATTCTTCGCAGTAAGGAAAGCCACTGCTAAT
TCCAAGTTTCCATTACTATC

111K10-K12
CTGGGTTGGGTAAATATCCGAAATTTGAAAATCTGAAATCTGAAATAAGCCAGTGAGCATTT
CCTTTCAACATCTTGTTGGTTCTCAGAAAATATGGAGTTTTCAAATTGTGGAGCATTTCTTAT
TTCTGGCTTGGAGTTTTGGGATGTGCAACCTTCAAGGCTTGCTGTAGTATCTGGGTGTCATTA
ATCCCCGTAATTTCTCTCAGTTGATTCAAAAACGTCTGCTGGTG

73M5-A2
CTGAATAACAGCACTAAACCAACAAGTATTGCCAACATTCTTTAGCCCAACGGGAGCTTTGT
CCTGTCTTTTTCTATCATAAGGGTTTCGAGAATCCCTCCAAACTTCTGTAGGTGTCCTCTTCA
AACATGCTTTATTCTCTGCTATGCTGGCTTCAAGAACT

73M5-F7
CCAAAGCCTCCAAAGCAAACATTTCCCTCCATGCCCATAACGTCCCAAAACTGCGTAGGGTC
CCGAACTCCGGCCCCTTTCTGTGTCACGCGTTCTCTCCCCTCGCCTTTCTAGGAATTTCACAA
GCCTTGAACTTGC

73M5-H7
CTTTTGGTTTCGGGGTAAATCTTGAGCATTTGATGGAGGCTTGTAATTCAGAACTAATCTTCT
AAATTCCAAAAGATTAAATAATGA

73M5-H10
CTTGATCTGATAGAGCTGGAGGCTGATTCGAAGTTGTGGGAGCTGTTCCAGCTTGATTATCG
CGTTCTGTTGCAACACCACGCCTTGCGTGCCCTTCCGAAGAAAGAAGCTGCACAAGCCATTT
ACACTTTCATCGAAAGCCTTCCGCAGAACCCGTTGCCGCTATCGTTCGCGCGAATCCGTGAG
CGCCTGGCTTTGCAGTCAG

73M5-J13
CTGGGTTGGGTAAATATCCGAAATTTGAAAATCTGAAATCTGAAATAAGCCAGTGAGCATTT
CCTTTCAACATCTTGTTGGTTCTCAGAAAATATGGAGTTTTCAAATTGTGGAGCATTTCTTAT
TTCTGGCTTGGAGTTTTGGGATGTGCAAC

73M5-K13
CTGCTAATGGCTTGTTCCTCATCAGTTATTCCAGTCTCCCTGAATGCCCTGTTTGATTCGGCC
AAACTCAAGGCAATTGCTCTCTGAAGATCATCTTTATCATCTCCAGTGAGATCAATCACAT

73M5-K17
CTTTTTAATCCTCACTGCTTCATCAGTGTTTATGAATGTTCAGTTCTGGGTTACACTCATCTTC
AGCTGTGGTCTGGGTGGACTTGGATTTTGTAACCAGACTTGTAAAACCTTTGGCCTGCTGTTG
C

135E14-A3
TTTGTTCAGTTACCAGTTTATTGCCTTGCAGCTCCAAATCCATCCTCCATTGCTCTGCCTGTGA
TACTGGAGCTAAACCTTATAAATATTTCTCCTCTGTCAGATGCCCAATGTTCGTGTTGTCAGC
AGATGGCTCTAGGAAACA

135E14-B4
CTCATATTCCTCCAAGTTTACTTCTTCAGGTTTTATCATTTCCAGTTTGGCTTGAGCAACTTTC
ATTATGTTGTGACAC

135E14-B7
CGACTGCAAAACTGTTTCAGGACTTTTATCTTCATGCTCATGTGATGCCTTGGTAATTATTTG
AATAGTTTCTTCTTTCAAGTGCTTTGAGAAATCACTTCTTGATGGCTGCTCTAGATATTCTGG
GTCTCCTGCTGCTTGTG

135E14-C6
ATATTTAAGCTATGGTTCCGGTCCCAAACGATTCCCCTTGGTAGATGTTCTTCAGTATGCATT
GGAATTTGCCTCAAGTAAACCTGTTTGCACTTCTCCTGTTGACGATATTGACGCTAGTTCCCC
ACCTAGTGGTTCCATACCATCACAGACATTACCAAGCACAACAGAACAACAGGGAGCCCTA
TCTTCAGAACTGCCAAGCACATCACCTTCATCAGTTGCTGCCATTTCATCGAGATCAGTAATA
CACAAACCATTTACTCAGTCCCGGATACCTCCAGATTTGCCCATGCATCCGGCACCAAGGCA
CATAACGGAGGAAGAACTTTCTGTGCTGGAAAGTTGTTTACATCGCTGGAGGACAGAAATA
GAAAATGACACCGGAGATTTGCAGGAAAGCATATCCAGAATCCATCGAACAATTGAATTAA
TGTACTCTGACAAATCTATGATACAAGTTCCTTATCGATTACATGCCGTTTTAGTTCACGAAG
GCCAAGCTAATGCTGGGCACTACTGGGCATATATTTTTGATCATCGTGAAAGCAGATGGATG
AAGTACAATGATATTGCTGTGACAAAATCATCATGGGAAGAGCTAGTGAGGGACTCTTTTGG
TGGTTATAGAAATGCCAGTGCATACTGTTTAATGTACATAAATGATAAGGCACAGTTCCTAA
TACAAG

135E14-D6
GTCTCCTCTTCAGCTTTCATTTGGAAGGCATCTTCTAACCAATCTAATAATTTGTGTGTAAAC
TCACTCACATCTTG

135E14-G12
CACCCAAACTGGATGAGATAAGGAATGGTCTTCCAGAGCCTCTGGAGGGAACACAGCCTTG
CCAACACCTTGATTTCAGCCAAGGAAACTGATGTTGAACTTCTGACTCCAGA

135E14-O13
CCTTAGGACTAGTAGCTGCATCCTACAGTTGTGATCCCCGGGTTACTCTCTAATGCCTCTG
TACTCACATCTCCTCTGTTGAAATATGTAGTGTCGGCTTTCCTGATTGGACACATGCTGGT
AGTGGTGCTCT

Table 3: Sequences of the exon trapped clones analysed in Chapter 4.

Sequences generated for the analyses of USP21

USPA
AAGTCCTGAAACAGTTTTGCAGTCG/GCAATTAAGTTGGAATATGCAAGGTTGGTTAAGTTG
GCCCAAGAAGACACCCCACCAGAAACCGATTATCG

N92952-M13R+
CAAAAGAGTTTCAGCTTGCATGTGAATAAATTATTTTAGTTTAAAAACCTTTTTTTAATGTTA
ACTATCTTCTGTGGCAGATCATGATGACACCGAACATGCAAGGTATTATCATGGCGATAGGT
AAATCCAGGAGTGTATATGACAGGTGTGGCCCTGAAGCAGGGTTCTTTAAGGCAATTAAGTT
GGAATATGCAAGGTTGGTTAAGTTGGCCCAAGAAGACACCCCACCAGAAACCGATTATCG

N92952-R2
TTCTTTAAGGCAATTAAGTTGGAATATGCAAGGTTGGTTAAGTTGGCCCAAGAAGACACCCC
ACCAGAAACCGATTATCGTTTACATCATGTAGTGGTCTACTTTATCCAGAACCAGGCACCAA
AGAAAATTATTGAGAAAACATTACTAGAACAATTTGGAGATAGAAATTTGAGTTTTGATGAA
AGGTGTCACAACATAATGAAAGTTGCT

Table 4: The first sequence was generated by RT-PCR from bone marrow RNA. The primers USP A and B were used to specifically amplify the PCR product, which was sequenced with the primer USP A. This confirmed that exons 18 and 19 were neighbouring exons. The border between the two exons is indicated with /. The two following sequences were generated from the Image cDNA clone 307735, which was amplified with the vector primers M13F+ and M13R+. Cycle sequencing was carried out with M13R+, from which the sequence is indicated here. This sequence maps between exons 18 and 19 of USP21 in the genomic DNA. The primer N92952-R2 was generated from this sequence, which was used to extend the sequence at this end of the Image clone. This sequence was identical to exons 19 and 20 of USP.

Sequences generated for the analyses of G31B5

Clone 31B5-E21R RACE, cycle sequencing primer 31B5-E21R
AAATAAAGGCAAACCGAAGGAGTAACAGCAAAAAATCCAAGACTCTGCAGGAGTTCCTAGA
GAGGATTCATCTGCAGGAATACACCTCAACACTTTTGCTCAATGGTTATGAGACTCTAGAAG
ATTTAAAAGATATAAAAGAGAGTCACCTCATTGAATTAAATATTGAAAACCCAGATGACAG
AAGAAGGTTACTATCAGCTGCTGAAAACTTCCTTGAAGAAGAAATTATTCAAGAGCAAGAA
AATGAACCTGAGCCCCTATCCTTGAGCTCAGACATCTCCTTAAATAAGTCACAGTTAGATGA
CTGCCCAAGGGACTCTGGTTGCTATATCTCATCAGGAAATTCAGATAATGGCAAAGAGGATC
TGGAGTCTGAAAATCTGTCTGACATGGTACATAAGATTATTATCACAGAGCCAAGTGACTGA
ACACGCATTCCCAACTATATATCT

Clone 31B5-E21R RACE, cycle sequencing primer UG1-R1
CTATATATCTACAGATGCATTCCTTTTTAACTCTTCTTGAGCTAAAACGTCAAATAGGAGAGG
AAAATAAGATAAATATTTGTAAATAAAACCTAAAGTTTAAATGTTTTAATCTGAATAATTGT
ACATAAAATTTTGTATCTCTAACATTCCAAATTACTGTCAATAAAATATATATTTATTATTTT
AAATGCTATGTGTTAATATTTCACTTGCTTGTATTAGAAAGGCAAAATGTAAGACTTTGGTAT
GTGTGACATATGCTTTATTTGGCTTTATTTTACAAGTACAGTATCTGCAAAA

Clone 31B5-E21R RACE, cycle sequencing primer UG1-R2
TTATTTTACAAGTACAGTATCTGCAAAAACAAAGTAACCTTTTTTCATACCTGCCAGTTTTGA
ATTTATATATGTTATTGAACAAATAGTAATAGAGGATTCGCTGTTGAAACAAGTTGTCCAAG
CAATGTTATATTCATTTTTATACTTATTGGGAAAGTGTGAGTTAATATTGGACACATTTTATC
CTGATCCACAGTGGAGTTTTAGTAATTATATTTTGTTGATTTCCTCATTTTGTTTTCTGGTATA
AAAGTAGAGATAATGTGTAGTCACTTCTGATTTAGTGAAACCAATTGTAATAATTGTGGAAA
TGTTT

Clone 31B5-E21R RACE, cycle sequencing primer UG1-R3
AATTGTAATAATTGTGGAAATGTTTTGTCTTTAAGTGTAAATATTTTAAAATTTGACATACCC
TAATGTTAATAATAAAAAGAACTATTTGCATATTGCTAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAA

Clone 31B5-E21F RACE, cycle sequencing primer 31B5-E21F
GGAACTCCTGCAGAGTCTTGGATTTTTTGCTGTTACTCCTTCGGTTTGCCTTTATTTTCTTGGG
GGCTGCTTCCTCTTCTGAGATGACATCCACATAAATGAATTTGAAGTTTCCCACTTTATTGTT
CAACATTCCTGTCCACATCCCCATTGGTGTTTTGCAAATAATGTCTATGATGTCTCCTTTCTTG
ATTTTGAGGGAGTCAGTGTCATAGGGACTTGGCGTGAAATCCGTATGCACTCTGGCACGGCC
ACAGAATGGTCCTGAATAGGGGCCATCGTCATCCAGTCGAAAGCTGTCCCGGTTACTTGTAC
CATCTGAACAGCTTGTTATGCCACTTGATGAGCTCTGTCCACTGTAGAGACTATCCATGGAG
TCACTGGCTTTGAGGGACACCTTCTCTGTGTGGG

Clone 31B5-E21F RACE, cycle sequencing primer UG1-F1
TGGAGTCACTGGCTTTGAGGGACACCTTCTCTGTGTGGGTCCCAATCACAGGGTCACTGTTTC
TATATGGGTGGGCATTCTCTCCATCTTCCTCATCCTTTTCCTCAGAAAGGGCTTTGATGTACT
TTTTACCCACTTTTTTCTTCATTGTCCATGAAATAGCTCTCATTTTTTTACCCAAACCGCCTCC
ATTATTTGAAGTTTTACTTTGTTCTCCACTTCCATTTGTGGGATCTCCTTCATGTGCCTCAGTT
GAATCATCTGGTTTTGATAAAGAATTATTCCGAAAACGATCGAAATTCCCAAAACTGCTGCT
TCGCTTTGGTTTTTGATGTTTCTCCTTCTCTGAAACATTGGATGGCTTTCTCTTGAGCATTTTG
AATTCTGACTACTCCTAGTGAGTGCACTTTCTGCTGTTACAGAAACAACTGAAAACAGTCAG
CAG

AA035703-M13F+
TATGCAAATAGTTCTTTTTATTATTAACATTAGGGTATGTCAAATTTTAAAATATTTACACTT
AAAGACAAAACATTTCCACAATTATTACAATTGGTTTCACTAAATCAGAAGTGACTACACAT
TATCTCTACTTTTATACCAGAAAACAAAATGAAGAAATCAACAAAATATAATTACTAAAACT
CCACTGTGGATCAGGATAAAATGTGTCCAATATTAACTCACACTTTCCCAATAAGTATAAAA
ATGAATATAACATTGCTTGGACAACTTGTTTCAACAGCGAATCCTCTATTACTATTTGTTCAA
TAACATATATAAATTCAAAACTGGCAGGTATGAAAAAAGGTT

Table 5

AA741162-M13R+
AGCTGTTCAGATGGTACAAGTAACCGGGACAGCTTTCGACTGGATGACGATGGCCCCTATTC
AGGACCATTCTGTGGCCGTGCCAGAGTGCATACGGATTTCACGCCAAGTCCCTATGACACTG
ACTCCCTCAAAATCAAGAAAGGAGACATCATAGACATTATTTGCAAAACACCAATGGGGAT
GTGGACAGGAATGTTGAACAATAAAGTGGGAAACTTCAAATTCATTTATGTGGATGTCATCT
CAGAAGAGGAAGCAGCCCCCAAGAAAATAAAGGCAAACCGAAGGAGTAACAGCAAAAAAT
CCAAGACTCTGCAGGAGTTCCTAGAGAGGATTCATCTGCAGGAATACACCTCA

G31B5-A
TCACTAGGAGTAGTCAGAATTCAAAATGCTCAAGAGAAAGCCATCCAATGTTTCAGAGAAG
GAGAAACATCAAAAACCAAAGgtaaggtgatttcatacatagttaaataaaatcaaactatttaatagaaaataaacacatatgtgtgttt
gaacacatacaatgttctcatctgtcagtcataagtcaaacatagggtttaatgcctcatggcaaatacggaaaaattaaattgcaaagttagattt

G31B5-B
TTTCTGCTGTTACAGAAACAACTGAAAACAGTCAGCAGTGTGCTGTCTAATgcacagcatcagataag
gttgagcccagggacctgccacttcctccttcagtcaggatatgacttcaacatttacatgggtaatttttttttcttacagtaattctttggtaactctttattgtctt
gctattttctatgaggtatgttccaaatctcatatttagacatgcctgcagt

G31B5-C
TGATTCAACTGAGgttcgtagtttattccatttttatgaatgttatgcaaacacattatgaaatattatgaagacaatgtcacttttgcaaagtaaaat
aagaaaaatgaatacattttaaaattatagccttgcatatgtttttgtgataaatgaacaaaccattatttagtactgttgtggaaagcaatgggttcttgtcacac
gaccagg

G31B5-D
AAACTGCTGCTTCGctataaaatagaaaagaaaaagcaaaggaaacttagaatttattttcactgtccaaactgataagtatattagattatagtt
taataagcaccgtctcttgcaaatacaggctttaaaaagctcttctctctggaatatccaaaggctgtgaaggtatacctattaacatttttatatttacagttgat
caaatgtctttctacgctttt

G31B5-E
TTTCTGAGGAAAAGgtaaatattctctgataatcatttgtatttttctactaaaactttcttaaactattcagctannctatagaagtgcttggtacttt
atgaagtacttaagcatttgttttcttcctcagtcattatgcccagaacttgggtggaaggaagtattactnncattttaacagaagaaactgacactca

G31B5-F
CTGAAATAGCTCTCATTTTTTTACCCAAACCGCCTCCATTATTTGAAGTTTTACTTTGTTCTCC
ACTTCCATTTGTGGGATCTCCTTCATGTGCctagtttagaattgtttacaaaagacaaaataataaagcaataaaaaacattgat
ctgaaatcactgttacagagaacattctgagtttgtggattttgcatgcattctgtgcagcagcctccaaaatttctaacaatccggacataacacatatttttaa
agttgtccagccctggttgcttgtgttacgaactcagaaagtaaaattcgtgatgttaaaggacttgagcctcacatctgcaactcttttagagaattttgactat
ttttaaat

Table 5: Sequences of the RACE products and intron/exon boundaries of the first three exons are shown. The RACE products were used to identify the gene G31B5. The RACE products and the primers that were used to identify the six sequence are shown directly above each sequence. The position of the primers can be seen in Figure 6.3. The next two sequences (AA035703-M13F+ and AA741162-M13R+) were derived from sequencing the Image clones, 360025 and 1286563, repectively with the indicated M13 primers. These two sequences were used to assemble the G31B5 sequence; their position can be seen in Figure 6.3. The final six sequences were determined by direct sequencing from the BAC clone B39I12. The primers, indicated above the sequences, were designed from the G31B5 sequence, which contained an unknown number of introns. By sequencing over the intron (upper case)/exon (lower case) boundaries three additional exons were identified, resulting in eight exons in total for G31B5.

Appendix II Publications

Groet J., J.H. Ives, T.A. Jones, M. Danton, R.H. Flomen, D. Sheer, R. Hrascan, K. Pavelic and D. Nizetic. 2000. Narrowing of the region of allelic loss in 21q11-21 in squamous non-small cell lung carcinoma and cloning of a novel ubiquitin-specific protease gene from the deleted segment. *Genes, Chromosomes and Cancer* **27:** 153-161.

Groet J., J.H. Ives, A.P. South, P.R. Baptista, T.A. Jones, M-L Yaspo, H. Lehrach, M.-C. Potier, C. Van Broeckhoven and D. Nizetic. 1998. Bacterial contig map of the 21q11 region associated with Alzheimer's disease and abnormal myelopoiesis in Down syndrome. *Genome Research* **8:** 385-398.

Potier M.-C., A. Dutriaux, R. Orti, J. Groet, N. Gibelin, G. Karadima, G. Lutfalla, A. Lynn, C. Van Broeckhoven, A. Chakravarti et al. 1998. Two sequence-ready contigs spanning the two copies of a 200-kb duplication on human 21q: partial sequence and polymorphisms. *Genomics* **51:** 417-426.

Vidal-Taboada J.M., S. Bergonon, M. Sanchez, C. Lopez-Acedo, J. Groet, D. Nizetic, A. Egeo, P. Scartezzini, N. Katsanis, E.M.C. Fisher, J.M. Delabar and R. Oliva. 1998. High resolution physical mapping and identification of transcribed sequences in the Down Syndrome region. *Biochem. Biophys. Res. Comm.* **243:** 572-578.

Katsanis, N., J.H. Ives, J. Groet, D. Nizetic and E.M.C. Fisher. 1998. Localization of receptor interacting protein 140 (RIP 140) within 100kb of D21S13 on 21q11, a gene-poor region of the human genome. *Human Genetics* **102:** 221-223.

Osoegawa K., R. Susukida, S. Okano, S. Minoshima, N. Shimizu, P.J. de Jong, J. Groet, J. Ives, H. Lehrach, D. Nizetic and E. Soeda. 1996. An integrated map with cosmid/PAC contigs of a 4-Mb Down syndrome critical region. *Genomics* **32:** 375-387.

LETTER

Bacterial Contig Map of the 21q11 Region Associated with Alzheimer's Disease and Abnormal Myelopoiesis in Down Syndrome

Jürgen Groet,[1] Jane H. Ives,[1] Andrew P. South,[1] Pedro R. Baptista,[1] Tania A. Jones,[2] Marie-Laure Yaspo,[3] Hans Lehrach,[3] Marie-Claude Potier,[4] Christine Van Broeckhoven,[5] and Dean Nižetić[1,6]

[1]Centre for Applied Molecular Biology, School of Pharmacy, University of London, London WC1N 1AX, UK; [2]Human Cytogenetics Laboratory, Imperial Cancer Research Fund, London WC2A 3PX, UK; [3]Max-Planck-Institute for Human Genetics, Berlin-Dahlem, Germany; [4]Neurobiology Laboratory, Centre National de la Recherche Scientifique URA 2054, 75005 Paris, France; [5]Laboratory of Neurogenetics, Flanders Interuniversity Institute for Biotechnology (VIB), Born Bunge Foundation (BBS), University of Antwerp (UIA), Department of Biochemistry, Antwerp, Belgium

We present a high-resolution bacterial contig map of 3.4 Mb of genomic DNA in human chromosome 21q11–q21, encompassing the region of elevated disomic homozygosity in Down Syndrome-associated abnormal myelopoiesis and leukemia, as well as the markers, which has shown a strong association with Alzheimer's Disease that has never been explained. The map contains 89 overlapping PACs, BACs, or cosmids in three contigs (850, 850, and 1500 kb) with two gaps (one of 140–210 kb and the second <5 kb). To date, eight transcribed sequences derived by cDNA selection, exon trapping, and/or global EST sequencing have been positioned onto the map, and the only two genes so far mapped to this cytogenetic region, STCH and RIP140 have been precisely localized. This work converts a further 10% of chromosome 21q into a high-resolution bacterial contig map, which will be the physical basis for the long-range sequencing of this region. The map will also enable positional derivation of new transcribed sequences, as well as new polymorphic probes, that will help in elucidation of the role the genes in this region may play in abnormal myelopoiesis and leukemia associated with trisomy 21 and Alzheimer's Disease.

Chromosome 21 has one of the most advanced mapping states in the human genome project, which is probably attributable to a combination of its small size and the large concentration of genetic interest (Shimizu et al. 1995). The map proceeded from a very early detailed genetic map (Antonarakis et al. 1989; McInnis et al. 1993) and a whole chromosome *NotI* restriction map (Ichikawa et al. 1993), accompanied by YAC overlap maps (Chumakov et al. 1992; Patterson et al. 1993; Nizetic et al. 1994; Korenberg et al. 1995) and an integrated cosmid-pocket map of the entire chromosome (Nizetic et al. 1994), to megabase-size bacterial contigs of selected regions (Patil et al. 1994; Lafreniere et al. 1995; Eki et al. 1996; Osoegawa et al. 1996; Ohira et al. 1996; Stone et al. 1996; Hubert et al. 1997) as well as global and regional transcriptional maps (Cheng et al. 1994; Peterson et al. 1994; Lucente et al. 1995; Tassone et al. 1995, Yaspo et al. 1995; Chen et al. 1996). Though one of the most advanced among whole human chromosome maps, the chromosome 21 map is still not sufficiently complete in its proximal half to enable detailed transcriptional mapping and large-scale sequencing. Attempts at positioning cosmid-derived, cDNA-selected, or exon-trapped products on near-complete YAC contigs routinely result in ~25%–40% failure, despite proven 21q location of the same products (Cheng et al. 1994; Yaspo et al. 1995; Gardiner 1996). This failure illustrates the imperfection of the YACs as high-resolution mapping tools and the need to extend the map to fully overlapping sets of bacterial vector clones. So far, the published bacterial contig maps encompass approximately one-third of the long arm, and a large

[6]Corresponding author.
E-MAIL dnizet@pharm.lon.ac.uk; FAX 44-171-278 1939.

international sequencing consortium is nearing 5 Mb of completed genomic sequence in at least three distinct regions, which amounts to ~15% of the length of the long arm (see http://www-eri.uchsc.edu/chr21/dna/map.html and http://chr21.rz-berlin.mpg.de/consortium.html), all of these concentrated in the telomeric half. Studies of partial trisomies and correlations to phenotypic features of Down syndrome (DS) (Delabar et al. 1993) led to a widespread belief that practically all the dismorphic features were caused by trisomy of the gene(s) in the so-called Down syndrome critical region (DCR). This region, as well as the loci for other interesting genetic diseases all mapped to the distal third of the chromosome. More recent data show that trisomy of the proximal half of the chromosome can lead to DS features (for summary, see Korenberg et al. 1994) and demonstrate that a more complex model is probably more appropriate to explain DS, which makes genes throughout the entire length of the long arm interesting in the study of the molecular basis of various features of DS (Korenberg et al. 1994). Increased disomic homozygosity in the pericentromeric region of 21q has been found amongst DS subjects with transient abnormal myelopoiesis (TAM) and/or ANLL-M7 leukemia compared to DS individuals with normal myelopoiesis. This was observed by use of both cytogenetic (Abe et al. 1989) and molecular techniques (Niikawa et al. 1991; Shen et al. 1995; Dagna-Bricarelli et al. 1997), in particular around the markers D21S215, D21S258, D21S120, and D21S192. It has been proposed that this region may house a gene encoding a product with tumor suppressor or growth control functions (Niikawa et al. 1991; Ohta et al. 1996). The interval around the markers D21S13 and D21S16 (St.George-Hyslop et al. 1987; Goate et al. 1989; Van Broeckhoven et al. 1990; Schellenberg et al. 1991) has been designated in the past as the Alzheimer's Disease-1 (AD1) genomic region (Van Broeckhoven 1995) because of its high linkage (lod > 3) repeatedly found with Familial Early Onset Alzheimer's Disease (FEOAD). Because the same families subsequently showed mutations in presenilin genes, the 21q11 linkage was never explained.

We present a map of 3.4 Mb of genomic DNA encompassing both the ex-AD1 locus and the region of disomic homozygosity in DS leukemia. This work converts a further 10% of the long arm into a high-resolution bacterial contig map. The contigs will be sequenced by the chromosome 21 sequencing project within the framework of the German Human Genome Programme.

RESULTS

Construction of a PAC-Pocket Sublibrary Enriched for the Region

A YAC contig from the integrated cosmid-pocket map published previously (Nizetic et al. 1994) was used as a starting point. Eight YACs that form the minimal YAC tiling path (YACs: 1A2y21, 7A1y21, 255d1, 6A6y21, 23cb10ici, 2G2y21, 2C12y21, and 5A11y21) were excised after pulsed-field gel electrophoresis (PFGE), purified, and used as hybridization probes against a five genome equivalent PAC library (Ioannou et al. 1994) displayed on high-density membranes. A group of 1730 PACs were detected

Figure 1 The integrated map of the 21q11–q21 region. The *top* horizontal bars represent the cytogenetic (dark and light) bands, the direction of the centromere is indicated, and the region around D21S190 that is duplicated once more on 21q22.1 (Dutriaux et al. 1994) is shaded. The region of homology to other chromosomes is shown as a thin horizontal gray bar immediately underneath the cytogenetic bar level. The next horizontal level is the scale (in kb), followed by the line representing genomic DNA, containing symbols: (circles) STS or hybridization markers (D21 is omitted in their names); (solid rectangles) fully characterized genes; (bent arrows, pointing *left* and *right*) segments contained within somatic cell hybrids on rodent background (whose names are given above the arrows) that were checked against STSs by PCR; below is an open bar representing the restriction map, vertical lines are restriction sites for *Xho*I (no letters or X), *Sal*I (S), and *Not*I (N). Asterisks represent *Xho*I sites whose respective order in that segment could not be precisely determined. The *Mlu*I sites described in text are shown as bold M restriction mapped to exact positions within PACs used (see Results). The only two gaps in the map are shown as open squares labeled G1 and G2 for gap 1 and gap 2, respectively, as referred to in the text. Lines underneath this bar with solid triangles pointing down represent limits of localization of ESs derived by cDNA selection or exon trapping, described in Cheng et al. (1994), Yaspo et al. (1995), and Schuler et al. (1996). The next level represents the PACs (from the library by Iouannou et al. 1994), BACs (B in front of the clone name, from the library by Shizuya et al. 1992), or cosmids [c102 or Q in front of the clone name, for the library by Nizetic et al. (1991a) or the Lawrence Livermore library, respectively] drawn to scale. The clones contained in the MTP are shown in boldface type. The restriction map above refers only to these clones. The YAC clones described in Nizetic et al. (1994) are shown in the bottom horizontal level as thin gray bars.

initially, but after elimination of PACs that hybridized to more than three YACs and/or hybridized to unrelated YACs, 1591 PACs were left, which could be distributed in 13 pockets on the basis of their hybridization fingerprint. Some pockets were too large (PACs hit by single YACs 255d1 and 6A6y21, amounting to 526 and 283 positives, respectively), probably containing many spurious weakly hybridizing PACs, and were not considered further. Finally, 560 PACs distributed in 11 pockets were re-

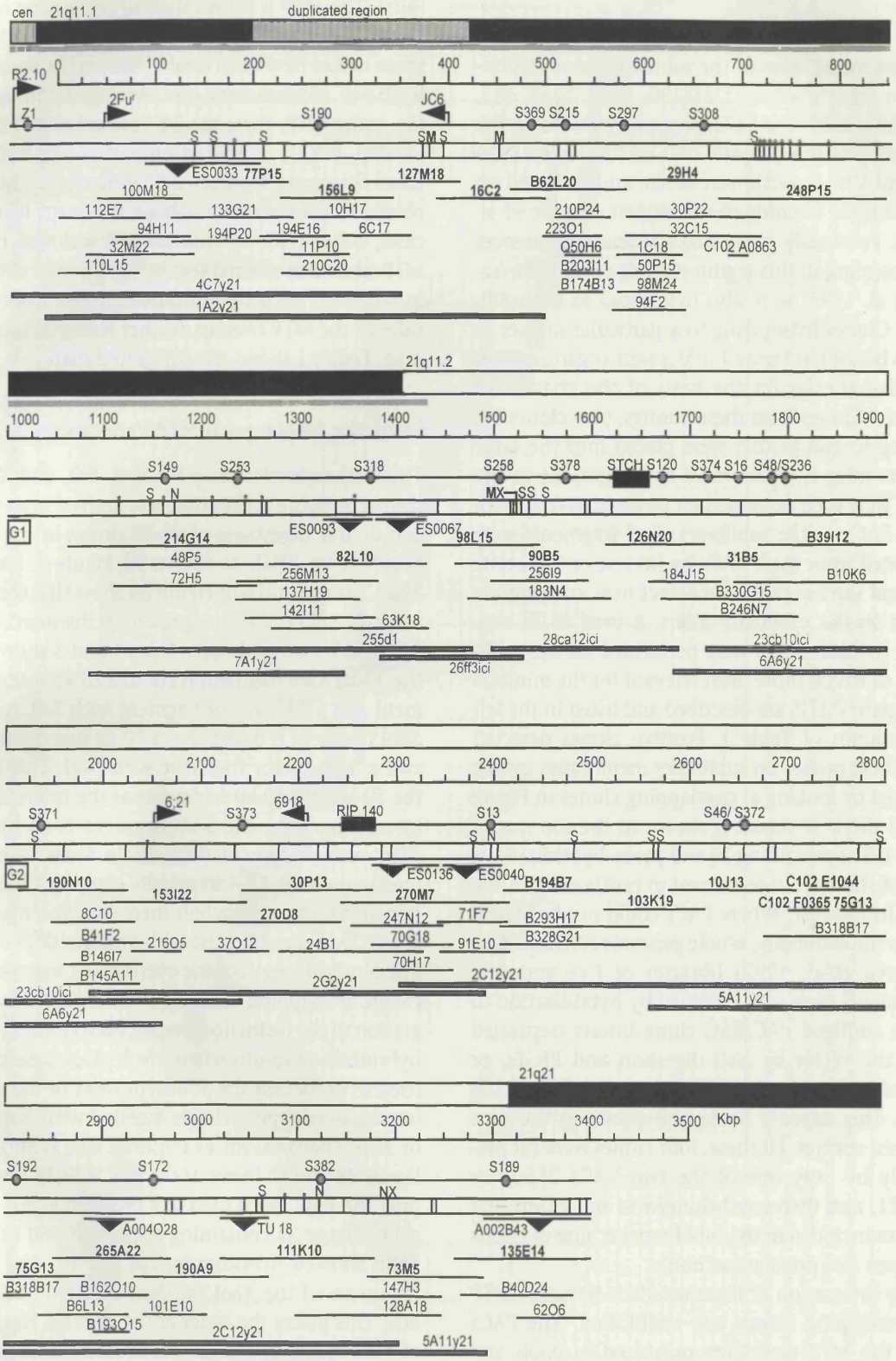


Figure 1 *(See facing page for legend.)*

gridded into six microtiter plates as a sublibrary enriched for PACs from the 21q11–q21 region. The PAC clones of the sublibrary were spotted and grown as colonies on nylon membranes for hybridization.

Construction of Contigs

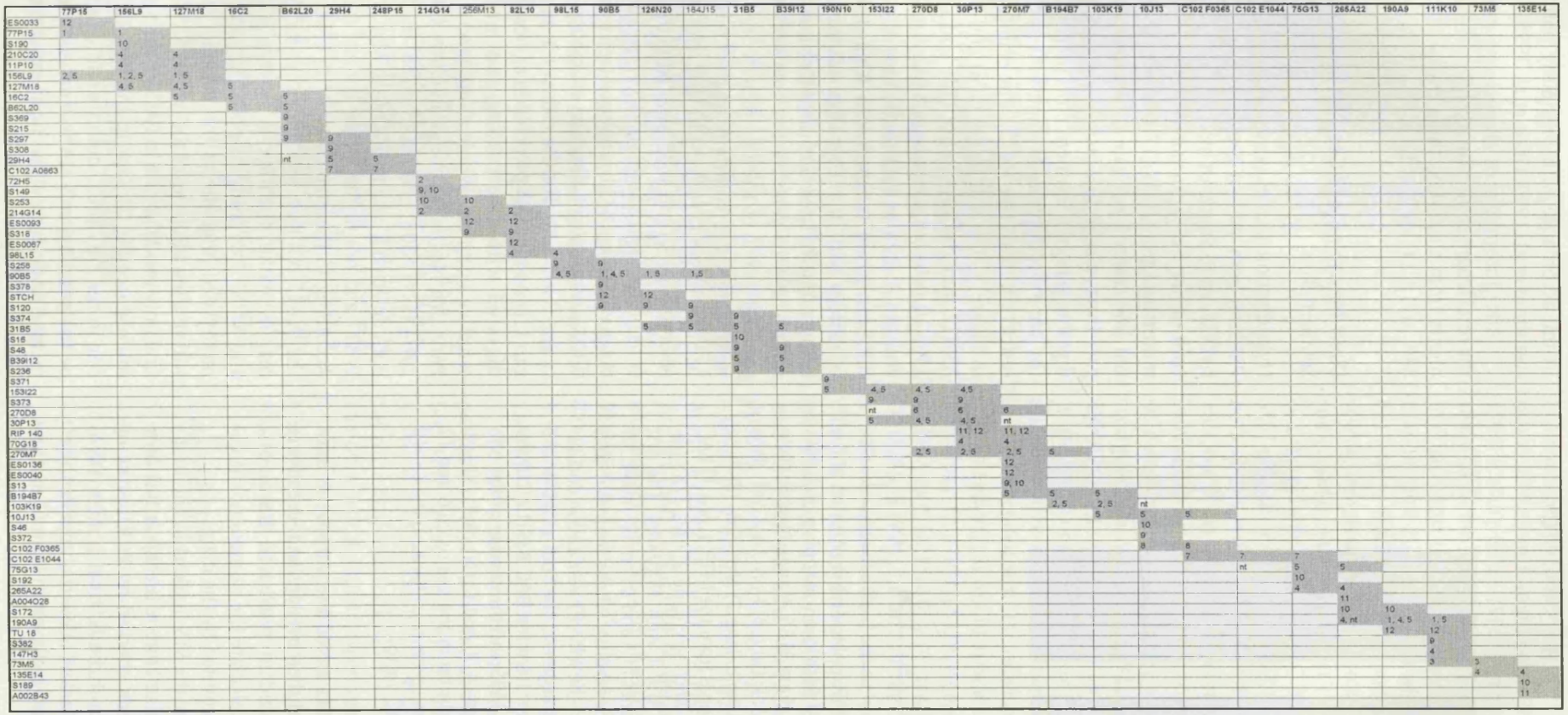
Gridded membranes of the sublibrary were hybridized to eight markers, D21S190, S149, S253, S13, S46, S192, S189, and S172, previously placed on the PFGE map of the genomic DNA of this region (Van Hul and Van Broeckhoven 1993), and/or found on YACs and/or cosmids of this region (Nizetic et al. 1994). Previously published expressed sequences (ES) mapping in this region (Cheng et al. 1994; Yaspo et al. 1995) were also hybridized to the sublibrary. Clones hybridizing to a particular marker or ES can be seen in Figure 1. PACs were organized into tentative contigs on the basis of the content of markers. Clones from these contigs, plus clones belonging to groups that were placed into the same pockets from the pattern of hybridizations to the YACs, were then examined for physical overlap with other PACs in the sublibrary. End fragments were generated from these PACs by inverse or vectorette PCR and used as probes to detect new, overlapping clones on the sublibrary filters. A total of 28 end-probe hybridizations were performed on the sublibrary, of which those most relevant for the minimal tiling path (MTP) are described and listed in the left end column of Table 1. Positive clones detected with these probes on sublibrary membranes can be deduced by looking at overlapping clones in Figure 1, and the MTP subset is shown in the top row of Table 1. The remaining 10 end-probe hybridizations are not shown, but were used in positioning of the PACs in Figure 1. Where PACs could not be identified in the sublibrary, whole genomic PAC and BAC (Shizuya et al. 1992) libraries of five genomic equivalents each were screened by hybridization to whole contiged PAC/BAC clone inserts (separated from the vector by *Not*I digestion and PFGE), or screened as pools of gridded clones by PCR (PACs only). This added a further 28 clones in the gaps between contigs. Of these, four clones were hit previously by only one of the two YACs 255d1 or 6A6y21, and therefore belonged to the group that was not included in the sublibrary because of excess numbers (see previous section).

By integration of these results, a tentative MTP of overlapping clones was established. The PACs from the MTP were then organized in pools and screened by use of sequence-tagged sites (STSs) from the CEPH–Genethon YAC contig of chromosome 21 (Chumakov et al. 1992; Patterson et al. 1993; Graw et al. 1995; Korenberg et al. 1995), adding an additional 13 markers onto the map (summarized in Table 1). Table 1 contains the summary of 12 different types of hybridization or PCR-based results (hits). The MTP is shown in bold on the top row of Table 1 and in Figure 1. Most overlaps between those clones have been established on the basis of at least two confirmatory hits. Averaged throughout the entire MTP, there are 2.5 confirmatory hits per overlap. If YAC hybridization patterns to PACs are taken into consideration, 3.5 confirmatory hits are obtained, averaged throughout the map. In many cases, other PACs or cosmids not included in the MTP (shown as normal text in Fig. 1), were detected in common with hybridization probes from both sides of the MTP overlap, further strengthening the map. Figure 1 shows the integrated map.

Restriction Mapping and MTP Verification

The final restriction map for *Not*I, *Sal*I, and *Xho*I is shown in Figure 1. This map was derived from single and double digestions of all 30 clones in the MTP, followed by PFGE as shown in Figure 2 for PAC 31B5. Lanes 2 and 6 in Figure 2A show that there are no extra *Sal*I or *Not*I sites present in the insert of the PAC 31B5. Comparison of lanes 1 and 3 shows that the 30-kb *Xho*I fragment is cut to a 23-kb insert fragment and a 7-kb vector fragment with *Sal*I. A 15-kb *Xho*I fragment is digested to a 10-kb insert fragment and a 5-kb vector fragment with *Not*I. This leaves the 50-kb and 40-kb fragments as the internal *Xho*I insert fragments, the 5-kb fragment is an internal *Xho*I vector fragment (lane 1). In some cases, the precise order of *Xho*I fragments could not be determined, for example, when there were too many internal *Xho*I sites (marked with an asterisk in Fig. 1). The physical length of the overlapping segments between overlapping clones was determined by integration of the restriction map of each clone with the hybridization results when whole clone inserts were used as probes on the Southern blots of the neighboring, overlapping clones digested with *Not*I/*Xho*I or *Not*I/*Xho*I/*Sal*I. An example of this is shown in Figure 2B and C. Insert of the BAC B39I12 (overlapping the PAC 31B5) used as a probe to a blot of the gel in Figure 2B containing both itself and the PAC 31B5 showed hybridization to the 10- and 50-kb fragments of the *Xho*I/*Sal*I/*Not*I-digested 31B5 (Fig. 2C). This places the order of the internal fragments of 31B5 and the overlap with B39I12 as indicated in Figure 2D.

Table 1. Summary of the Hybridization and PCR-based Hits That Establish the Overlaps within the MTP



(Left) Sources of probes; (top) target recombinant clones. All clones belong to the MTP (see Fig. 1) except 256M13 and 184J15. The exact type of probe used in each probe target combination was (1) inverse PCR T7 end; (2) inverse PCR SP6 end; (3) vectorette T7 end; (4) vectorette SP6 end; (5) PAC/BAC clone insert; (6) PAC *Alu* PCR products; (7) cosmid insert fragments; (8) T7 riboprobe; (9) marker by PCR; (10) marker by hybridization; (11) gene or EST by PCR; (12) gene or ES by hybridization. ES sequences are described in the text and in Cheng et al. (1994), Yaspo et al. (1995), and Schuler et al. (1996); sequences for gene probes (see Methods) for STCH and RIP 140 are described in Brodsky et al. (1995) and Cavailles et al. (1995), respectively; markers from the GDB start with an S, the prefix D21 is omitted; cosmids start with the prefix c102, BACs start with the letter B, and the rest are PACs.

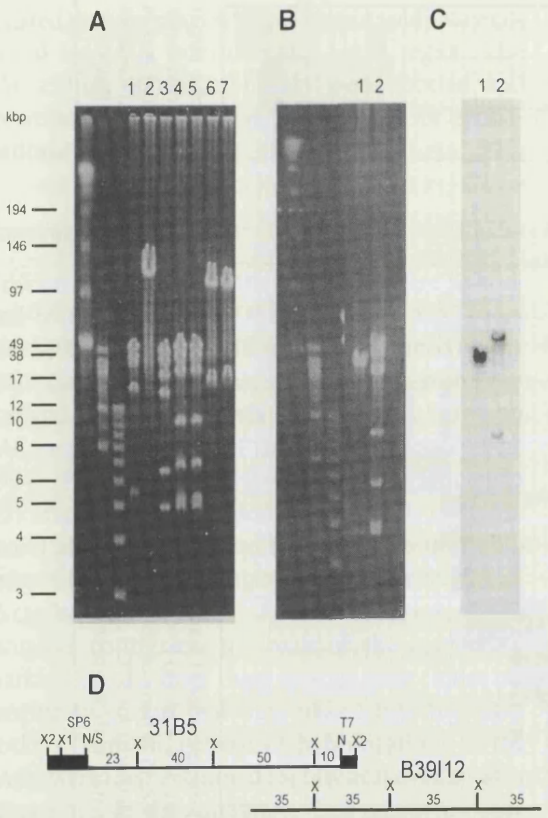


Figure 2 Restriction mapping and verification of the degree of overlap. (*A*) Restriction mapping of a PAC. The PFGE of the PAC clone 31B5. (Lanes *1*) *Xho*I; (lane *2*) *Sal*I; (lane *3*) *Xho*I–*Sal*I; (lane *4*) *Not*I–*Xho*I–*Sal*I; (lane *5*) *Not*I–*Xho*I; (lane *6*) *Not*I; (lane *7*) *Not*I–*Sal*I. (*B*) PFGE of two overlapping clones from the MTP. (Lane *1*) BAC B39I12 cut with *Not*I–*Xho*I–*Sal*I; (lane *2*) PAC 31B5 (restriction mapped in *A*) cut with *Not*I/*Xho*I. The first three un-numbered lanes in *A* and *B* contain molecular weight markers (see Methods), whose fragment sizes in kb are indicated on the *left*. (*C*) Verification and sizing of the overlapping segment; the Southern blot of *B* hybridized to the insert from the BAC B39I12. (*D*) Restriction map of PAC 31B5 and the overlap of the two clones as deduced from *A, B,* and *C*. (N) *Not*I; (S) *Sal*I; (X1 and X2) two *Xho*I sites in the vector. SP6 and T7 are the RNA polymerase promoters flanking the *Bam*HI cloning site of the PAC vector.

These experiments served the purpose of both confirming the established overlaps and identifying the list of nonredundant restriction fragments, which was used to calculate the nonredundant genomic DNA length covered by the map.

The positions for the STS markers in the intervals between somatic cell hybrid breakpoints shown in Figure 1 have been checked by PCR by use of genomic DNA of the hybrids, total human DNA, and rodent DNA. The data allow the positioning of the hybrid breakpoints as shown in Figure 1 and are in full accordance with previously published data (Gardiner et al. 1990; Graw et al. 1995).

After integration of all data and estimation of the total length of all nonredundant restriction fragments from the PFGE, three contigs emerge: 850, 850, and 1500 kb (Fig. 1).

Direct Measurement of Gap Sizes on Genomic DNA Southern Blots

Hybridization probes were prepared from selected PACs on both sides of gap 1 and gap 2 (Fig.1) and hybridized to Southern blots of the hybrid cell line containing only human chromosome 21 on a mouse background (WA17) and the parental mouse cell line control. For gap 2, end probes facing the gap from either side were hybridized to *Xho*I-digested genomic Southern blots prepared by use of the same electrophoresis conditions as for the PAC restriction maps (see Fig. 2). The results are shown in Figure 3. The gap facing (SP6) end of the BAC B39I12 and the gap facing (T7) end of the PAC 190N10 detected a common 83-kb band in WA17 and not in mouse control. Because the clones are fully restriction mapped with *Xho*I, this result is consistent with the *Xho*I sites nearest to the gap being unmethylated and therefore cut in the genomic DNA to generate the common 83-kb fragment [if either of the two nearest sites had been methylated, the common fragment would have been far bigger (116 or >140 kb)]. The two nearest *Xho*I sites are 34 and 46 kb away from the gap facing ends of the B39I12 and 190N10, respectively, which means that 80 of 83 kb is covered by the clones. This direct measurement allows an estimation of the size of gap 2 as <5 kb. For the estimation of the size of gap 1, genomic DNA prepared in agarose blocks was digested with rare cutting and methylation-sensitive enzymes. Whole clone inserts from two PACs that belong to contigs on either side of gap 1, 29H4 (S297,S308) and 98L15 (S258) used as probes both detected a common 1060-kb *Mlu*I fragment in WA17 and not in the mouse control (not shown). According to the map by van Hul et al. (1993), in the WA17 cell line, one unmethylated *Mlu*I site is mapped within the interval between markers S190 and S215. This study mapped the next distal unmethylated *Mlu*I site at the distance of 1050 kb, within a few kilobases of the marker S258. We could find only two *Mlu*I sites in PACs in the S190 to S215 interval (see Fig. 1): One in the PAC 127M18, and one in the PAC 16C2, both therefore candidates to be the proximal end of the observed 1060-kb fragment. The first site distal from gap 1 (common to

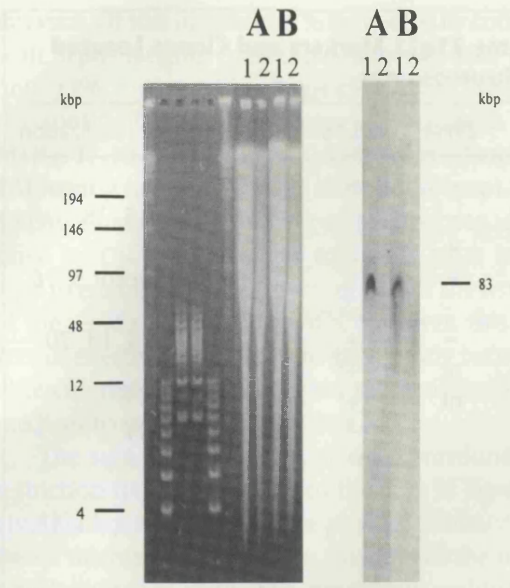


Figure 3 Direct measurement of gap 2 with genomic Southern blot hybridization. (*Left*) Photograph of a pulsed-field gel stained with ethidium bromide. Unlabeled lanes contain markers with some band sizes (in kb) are indicated. The next four lanes contain genomic DNA prepared in agarose blocks and digested with *Xho*I. (Lane *1*) WA 17 (hybrid line containing only human chromosome 21 on mouse A9 background); (lane *2*) A9 control. A and B are identical pairs of lanes run on the same gel, blotted, separated by cutting of the blotted membrane, and hybridized to end probes from B39I12 (SP6 end) and 190N10 (T7 end), respectively. (*Right*) Printout of a PhosphorImager scan of the hybridized membranes (B39I12, *A;* 190N10 *B*). Both ends were generated with a modified vectorette–PCR technique (see Methods); B39I12 SP6 end was a 750-bp fragment amplified from the vectorette library generated with *Pst*I; the 190N10 T7 end was a 420-bp fragment amplified from the vectorette library generated with *Hin*dIII.

PACs 98L15 and 90B5) was found in the cluster of *Xho*I fragments adjacent to the marker S258. When the PAC coverage between the *Mlu*I sites likely to have generated the 1060-kb fragment is measured from restriction-sized PAC inserts of the contigs, a figure of 850 kb or 920 kb is obtained, respectively, for the two *Mlu*I sites as proximal ends, in PACs 16C2 or 127M18. This result allows an estimation of the minimal and maximal size of gap 1 as 140 and 210 kb, respectively.

Regions of Homology to Other Chromosomes

Sequences homologous to other chromosomes (especially chromosome 13 and all acrocentrics) have been repeatedly observed in the proximal 21q (Choo 1990; van Camp et al. 1992; van Hul et al. 1993; Chiang et al. 1995; Graw et al. 1995). The study by Graw et al. (1995) found additional homologies other than to chromosome 13 (see last column in Table 2), in particular with chromosome 18, but also to many other chromosomes. We present several new results of chromosomal locations for some markers (summarized in Table 2). Integration of data in Table 2 for the BAC B62L20 (S369, S215, S297) and the PAC 29H4 (S297, S308) indicates that both clones must contain DNA cloned from chromosome 21, as it is the only common chromosome localization to the particular combination of three or two STSs present in these two clones, respectively. Similarly, the PAC 82L10 (S318, ES0093, ES0067) must originate from chromosome 21, as it is the only common chromosome localization for all three markers (Table 2). When these localizations are all integrated with the order of markers on the map in Figure 2, it would appear that the region of homology to other chromosomes ends between ES0067 and D21S258. Any marker telomeric to D21S258 checked by any of the above-mentioned studies was indeed found to be chromosome 21 specific. To confirm this border, FISH was performed with PACs near this border: 90B5 (S258,S378, STCH, and S120) and the next two PACs overlapping the 90B5 in the proximal direction, 98L15 (S258) and 63K18 (S318,ES0067,ES0093), plus two more clones from the proximal contig. The FISH results (summarized in Table 2) place the border between the chromosome 21-specific and shared regions near the proximal end of the PAC 98L15. The region of homology to other chromosomes is shown as a thin horizontal gray bar immediately underneath the cytogenetic bar level in Figure 1.

To further confirm the orientation of the most proximal contig, a three-loci comparison hybridization with two-color FISH (in different combinations) to interphase nuclei was performed (not shown). The order Z1–Q50H6(S215)–c102A0863 was obtained. A similar experiment confirmed the order of three contigs with respect to each other; the FISH probe order c102A0863–90B5(S258)–71F7(S13) was confirmed (not shown).

Genes and ESTs

Several global efforts to enrich the transcriptional map of the chromosome have placed some selected cDNAs and ESTs into this region by use of YACs or somatic cell hybrid panels. We show the precise localization of these transcribed sequences (Fig. 1):

Table 2. Summary of the PCR and FISH Analyses of Some 21q11 Markers and Clones Located in the Region Showing Sequence Homology to Other Chromosomes

	WA17	A9	GB3	CHO	B62L20	29H4	82L10	Chromosome localization
D21S369	+	−	−	−	+	−	−	[a]21, 18
D21S215	+	−	−	−	+	−	−	[a]21, 10, 14, 15, 18
D21S297	+	−	−	−	+	+	−	[a]21, 17
D21S308	+	−	−	−	−	+	−	[a]21 and all except 5, 16, 17, X, Y
D21S318	+	−	+	−	−	−	+	[a]21, 2, 9, 9/13, 18, 20
21ES0093 (PT4)	+*	−*	−*	−*	nt	nt	+*	[b]3, 16
21ES0067 (PT5)	nt	nt	nt	nt	nt	nt	+*	[b]21, 22
D21S258	+	−	−	−	−	−	−	21 (this paper)
FISH: 90B5								21q11.2
FISH: 98L15								21q11.2
FISH: 63K18								21, 2, 9, 13, 14, 15, 18, 19, 22
FISH: C102 A0863								21, 13
FISH: B62L20								21, 18p11.2
FISH: 223O1								21, 18p11.2

BAC B62L20 and PAC 29H4 are a pair of overlapping clones, and PAC 82L10, a third clone from the MTP crossing parts of this region (see Fig. 1). Two rodent background hybrids containing (as the only human chromosomes) the whole chromosome 21 (WA17) and the whole chromosome 13 (GB3) were used. The cell line A9 was used as the mouse background for WA17, and the CHO as the hamster background control for GB3. All marker primer combinations and PCR conditions were as in GDB. Asterisks indicate that results were obtained by genomic Southern blot hybridization; (FISH:clone name) result obtained using the clone as a probe for FISH on normal human metaphases. The chromosome localization column shows data taken from Graw et al. (1995)[a] and Chiang et al. (1995)[b], for comparison, which is necessary to deduce the chromosome 21 origin of the three BAC/PAC clones shown. The result for D21S258 was performed in this study, on the whole genome panel of monochromosomal hybrids.

(H21)—ES0033, ES0093, ES0067, ES0136, and ES0040 were cDNAs selected from fetal brain on random cosmids from a flow-sorted chromosome 21 library and subsequently mapped to 21q11 YACs (Cheng et al. 1994). Their multiple chromosomal localizations and expression patterns have been studied by Chiang et al. (1995). For one of these sequences (ES0136), Chiang et al. could not detect any clones in multiple tissue cDNA library screens. We detected RT-PCR products for this EST in fetal brain and fetal lung RNA, but not in fetal heart, fetal liver, or adult bone marrow. The TU18 was a trapped exon from a random chromosome 21 cosmid (Yaspo et al. 1995), whereas A004O28 and A002B43 were ESTs mapped onto the human genome map of radiation hybrid fragments (Schuler et al. 1996). We could obtain the correct RT-PCR products for A002B43 in all examined tissues (fetal heart, brain, liver, and lung, and adult bone marrow), for A004O28 only in fetal brain and fetal liver, whereas for TU18 we could not amplify any RT-PCR prducts in this panel of tissues. The cDNA H21ES0093 (Cheng et al. 1994) was later found to contain a portion of the newly characterized gene encoding for the human nonsmooth muscle neutral calponin (Masuda et al. 1996). However, whether the full gene is located in 21q11 remains to be determined, because this EST seems to be repeated on several chromosomes [Chiang et al. (1995) found it only on chromosomes 3 and 16, not on 21]. We mapped at least one of its copies to chromosome 21 (see Table 2) by detecting several bands on the genomic Southern blot, common to the total human genomic DNA and 21-only hybrid WA17, not the rodent background. We could also see five additional bands in total human DNA, consistent with the colocalization on additional chromosomes. Very recently, a pair of primers designed as specific for the 3′ UTR of the published functional gene shows clearly that this copy of the gene is not on chromosome 21 (J.H. Ives, unpubl.). The precise positions of the only other two genes so far mapped to this region: STCH (Brodsky et al. 1995) and RIP140 (Cavailles et al. 1995; Katsanis et al. 1998) are also shown in Figure 1.

DISCUSSION

The overall size of the region mapped is 3400 kb. The bacterial contigs shown cover 94%–95% of this

distance. Of this distance, 61% is covered by contigs with depth ranging from three- to seven-fold, and only 13% is covered by single-clone-deep parts of the contig. The average degree of overlap in the MTP is 17.3% of clone length. In the part where the YAC contig had been reliable (80% of the map), 55/59 PAC clones built into the contigs have been identified by YACs. The strategy using YACs has managed to enrich for clones covering 93% of the length of the region spanned by YACs. However, this degree of effectiveness is likely to vary greatly between different regions depending on the abundance of medium-to low-copy repeats in a YAC.

The sum of the lengths of the nonredundant restriction fragments shown in the map in Figure 1 gives the measurement of the physical distance between the markers for various intervals of the map, which matches well with previously published physical maps of large genomic DNA fragments separated by PFGE (van Hul and Van Broeckhoven 1993).

We could not find a single clone corresponding to the gap 1 and gap 2 DNA in the total of 16 genome equivalent libraries of flow-sorted cosmids, PACs, and BACs with probes from the contig ends. This probably means that the gaps contain sequences that are very difficult to clone in bacterial systems, especially in the case of gap 2. Closing gap 1 may be complicated further by the concentration of medium-copy repeats, making it impossible to detect specific clones. The published distance of 732 kb for the interval between D21S13 and D21S382 (called Eag101 on the genomic long-range map; van Hul and Van Broeckhoven 1993) matches almost exactly our PAC contig distance of 725 kb. These two markers are also the *Not*I-linking clones for the two *Not*I sites found on the global chromosome 21 *Not*I restriction map (Ichikawa et al. 1993), which gives a distance of 720 kb between them. On the restriction map of PACs, *Not*I sites have indeed been found at both D21S382 and D21S13 positions, which helped anchor our contigs to the *Not*I map of chromosome 21 (Ichikawa et al. 1993).

Integration of our data and those previously published (some shown in Table 2; Choo 1990; van Camp et al. 1992; van Hul et al. 1993; Graw et al. 1995; Chiang et al. 1995) with the physical and overlap map in Figure 1, shows several distinct regions of homology. The first 200 kb of the contig seems to be the region shared by all acrocentric chromosomes, the next 220 kb is occupied by the duplicated region (van Camp et al. 1992; Dutriaux et al. 1994; Potier et al. 1996), which is duplicated in 21q22.1 but is also homologous to chromosome 13 and not to other acrocentric chromosomes. Throughout the next segment (approximate positions 420–1330 on the kilobase scale, Fig. 1) extends the region containing sequences shared between chromosome 21 and multiple other chromosomes, with predominant homology to chromosomes 13 and 18. It is possible that translocations, duplications, and inversions precipitated by the presence of low- and medium-copy repeats present in many heterochromatin regions of the genome, may have been the events shaping the evolution of the DNA in 21q11.1. Many somatic cell hybrid breakpoints (Gardiner et al. 1990; Graw et al. 1995) deriving from translocation and deletion events map in this area. The region around D21S190 is duplicated once more about 15 Mb further distal in 21q22.1 (Dutriaux et al. 1994). The PAC 156L9 from the MTP (Fig. 2) is completely included in the duplicated region, but it has been assigned to the proximal copy of the region on the basis of polymorphisms (compared to the distal, 21q22.1 located copy of the same region) detected with two new polymorphic markers derived from the duplicated region (D21F127 and D21F158), which is the subject of a parallel study (M.-C. Potier, A. Dutriaux, R. Orti, J. Groet, N. Gibelin, G. Karadima, G. Lutfalla, A. Lynn, C. Van Broeckhoven, A. Chakravarti et al., in prep.). Constitutional inversions have also been seen at these loci. One study has mapped an inversion breakpoint in a DS child with TAM (Ohta et al. 1996) very near the clones that are at the border of the D21S190 duplicated region and the chromosome 21-specific (S215, S369) segment. Genomic sequencing of these border regions may reveal sequences that play important roles in the mammalian chromosome evolution.

There is widespread evidence pointing at the extremely uneven distribution of transcribed sequences over 21q (for summary, see Gardiner 1996). This analysis predicts a gene density of one gene every 6 Mb in the proximal part of 21q21, contrasted with one gene every 10 kb in the telomeric 10% of the long arm. The level of transcripts could be too low for the global cDNA selection approaches to be successful in this region. For example, none of the attempted global approaches so far (for summary, see Gardiner 1996) was able to detect ESTs corresponding to any of the parts of the two genes mapped by other means to 21q11.2 (STCH and RIP140), despite their ubiquitous tissue expression pattern (Brodsky et al. 1995; Cavailles et al. 1995). This result could mean that there are more genes than initially predicted in this region, especially in the R-band 21q11.2, but a combination of ap-

proaches [including one recently described (Yu et al. 1997)] may be necessary to detect them. Any new genes, as well as the existing two (three) genes, have to be analyzed for potential roles in disease phenotypes associated with the region. Increased disomic homozygosity has been found amongst DS subjects with TAM and/or ANLL-M7 leukemia compared to DS individuals with normal myelopoiesis by use of both cytogenetic (Abe et al. 1989) and molecular techniques (Niikawa et al. 1991; Shen et al. 1995; Dagna-Bricarelli et al. 1997), in particular around the markers D21S215, D21S258, D21S120, and D21S192. It has been proposed (Niikawa et al. 1991; Ohta et al. 1996) that this region may house a gene encoding a product with tumor suppressor or growth control functions. Both the STCH and RIP140 genes map inside the wider region of elevated disomic homozygosity. The chromosome 21q11 linkage in FEOAD families has never been explained, and it has been speculated that a modifier gene located in the ex-AD1 locus could act together with the PSEN1 gene to generate or modify the AD phenotype (St.George-Hyslop et al 1992; Van Broeckhoven 1995). So far, it has not been possible to test such hypotheses, because no genes have been mapped into this gene-poor region.

The high-resolution bacterial clone overlap map presented in this report will be the basis for the derivation of a more complete transcriptional map of this genetically interesting, but hitherto gene-poor region, as well as the starting substrate for the large-scale sequencing of the proximal 10% of the length of the chromosomal arm 21q.

METHODS

Libraries Used

Cosmid ICRF c102/ 103, flow-sorted chromosome 21 libraries (Nizetic et al. 1991a). PAC Library RPCI-1 (Ioannou et al. 1994). Total Human BAC library (Shizuya et al. 1992; Research Genetics). A single cosmid Q50H6 was from the Lawrence Livermore flow-sorted chromosome 21 cosmid library.

Hybridization of YACs to High Density PAC Membranes

YAC DNA was isolated as described previously (Nizetic and Lehrach 1995). Hybridization was carried out as described by Baxendale et al. (1991).

Picking and Spotting of PAC clones

Positive clones were regridded into 96-well plates containing 2×YT (Sambrook et al. 1989), 1× HMFM (Nizetic and Lehrach 1995), and 30 µg/ml kanamycin, grown overnight at 37°C. Duplication of the clones was performed in 384-well plates. Spotting and growing of clones onto Hybond N+ and subsequent processing was carried out as described (Nizetic et al. 1991b; Nizetic and Lehrach 1995). For the DNA isolation, PAC/ BAC clones were grown for 14 hr in 15 ml of superbroth (32 grams of tryptone, 20 grams of yeast extract, 5 grams of NaCl, 5 ml of 1 N NaOH per liter) with 30 µg/ml kanamycin. Cosmids were grown in 6 ml of 2× YT. DNA was isolated by use of a modified alkaline lysis method (Sambrook et al. 1989).

PCR Reactions

Standard reactions were performed in a 50-µl total volume, at a final concentration of 1× reaction buffer [10× reaction buffer: 100 mM Tris-HCl, pH 9.0, 15 mM $MgCl_2$, 500 mM KCl, 1% Triton X-100, 0.1% (w/vol) gelatin], 25 pmoles of each primer, 10 nmoles of each dNTP (Promega), and 0.1 units of Supertaq (HT Biotechnology).

PAC *Alu* PCR was carried out as described previously (Cole et al. 1991).

(PAC) Inverse PCR was carried out as follows: 1 µl of DNA (~300 ng) was digested with either *Bfa*I or *Nla*III (NEB) for 2 hr. Phenol and chloroform extraction and precipitation of the DNA were carried out, after which the samples were ligated overnight at 14°C in a 100-µl total volume with 1 unit of T4 DNA ligase (Life Technologies). PCR was performed with primer 1, CATTTAGGTGACACTATAGAGGATC, and primer 2, CGGCCAATTAGGCCTACCCAC, for the *Bfa*I-digested samples and primer 3, CGACTCACTATAGGGAGAGGATC, and primer 4, CAACCCAGTCAGCTCCTTCC, for the *Nla*III-digested samples. The annealing temperature was 60°C. Samples were resolved on a 1% (w/vol) agarose gel and PCR products isolated by gel purification.

(PAC) vectorette PCR was carried out as described previously (Riley et al. 1990) with primers 1 and 3 as the vector primers.

Gene and EST PCR was carried out as follows: cDNAs cloned into M13 (Cheng et al 1994) were amplified by PCR with M13 primers: Forward, CCCAGTCACGACGTTGTAAAACG and reverse, -AGCGGATAACAATTTC ACACAGG. The annealing temperature was 55°C. Primers for RIP140 were: Forward, -TCATTTTTCTTGATCGTTGTGG and reverse, -TGTAAAAAGGTCATTTCCCCC. The annealing temperature was 55°C. Genbank accession no. X84373. Primers for STCH were: Forward, -GTATTGAAAGAAGGCCAC, and reverse, -CTAAAGCACTGACTTGGAG. The annealing temperature was 50°C. Primers for PT12 (Yaspo et al. 1995) were: Forward, -AACCTTGGAAGCAAGTTA and reverse, -AATATGCAGATTCTCCTC. The annealing temperature was 42°C. For PT14 (Unigene A004O28) primers were: Forward, -TACAAGCTTATAGAGAAAGTCAA and reverse, -TATGTATAAAATAGGCAACTA GG. The annealing temperature was 57°C. For PT15 (Unigene A002B43) primers were: Forward, -AATATTATGTATCGTACACAGTG and reverse, -AGACAAAATGACTATACAGACTT. The annealing temperature was 57°C.

Preparation of Probes from Cosmid Inserts

Five microliters of DNA (see cosmid DNA isolation) was digested with 1 µl of *Sau*3AI (10 U/µl, Life Technologies). Aga-

rose gel electrophoresis was performed with 1% agarose. Bands exceeding 1200 bp were cut out of the gel, and DNA was isolated with Prep-a-Gene according to manufacturer's recommendations (Bio-Rad).

T7 riboprobe was prepared as follows: To 2 µl of cosmid DNA, 3 µl of buffer (40 mM Tris HCl at pH 8.0, 10 mM $MgCl_2$, 10 mM spermidine, 50 mM NaCl), 2 µl of 100 mM DTT, 1 µl of rNTP (5 mM each, Promega), 1.5 µl [^{32}P]UTP (ICN), 8 µl of tRNA (10 mg/ml, Sigma), 1 µl of RNAsin (Promega), 0.5 µl of T7 RNA Polymerase (NEB), and 11 µl of DEPC-treated H_2O was added. The reaction was incubated for 75 min at 37°C. After precipitation, the pellet was resuspended in 50 µl of H_2O and added to the hybridization.

Southern Blotting and Hybridization

Alkaline capillary blotting was performed on Hybond N+ (Amersham) and carried out according to the manufacturer's recommendations. Southern blots were prehybridized (4–6 hr) and hybridized (overnight) at 65°C in 25 ml of 6× SSC (20× SSC: 3 M NaCl, 0.3 M Na_3citrate, pH 7.0), 5× Denhardt's [100× Denhardt's: 2% (w/vol) BSA, 2% (w/vol) Ficoll, and 2% (w/vol) PVP], 0.5% SDS, and 100 µg/ml herring sperm DNA. Filters spotted with clones were prehybridized and hybridized in 25 ml (total human PAC/BAC library filters) and 8 ml (21q11 sublibrary filters) of Church solution (0.5 M Na-phosphate buffer at pH 7.2, 7% SDS, 1 mM EDTA, 0.1 mg/ml tRNA). Blots were washed in 2× SSC, 0.1% SDS for 15 min at room temperature and subsequently with 0.2× SSC, 0.1% SDS at 65°C for 30 min. Autoradiography was performed with X-ray film or images were visualized with PhosphorImager screens (Molecular Dynamics).

PFGE

PAC DNA (0.50–1 µg) was digested at 37°C with one or a combination of the following restriction enzymes, *NotI*, *SalI* (NEB), *XhoI* (Life Technologies). Genomic DNA in agarose blocks was prepared by use of standard protocols (Anand 1995). PFGE was carried out with a CHEF-DR II system (Bio-Rad) according to the manufacturer's recommendations with 0.5× TBE (10× TBE: 108 grams of Tris base, 55 grams of boric acid, 40 ml of 0.5 M EDTA at pH 8.0 per liter) and 1% rapid agarose (Life Technologies). Electrophoresis conditions were as follows: For the 4- to 200-kb window, 5.2 V/cm, 14 hr, switch time 3–15 sec; for the 100- to 1600-kb window, 6 V/cm, 22 hr, switch time 40–80 sec. DNA size markers used (see Fig. 3, from left to right) were Lambda ladder (Bio-Rad), high molecular weight DNA markers (Life Technologies), 1-kb DNA ladder (Life Technologies).

PAC/BAC Insert Isolation

PAC/BAC clones were digested at 37°C with *NotI* (NEB). PFGE was carried out as described above with 1% low-melting-point agarose gels (Life Technologies). Inserts were cut out of the gel and weighed. To an average slice that weighed 250 mg, 150 µl of H_2O, 40 µl of 1 M NaCl, and 8 µl of 0.5 M EDTA (pH 8.0) were added. The agarose slice was dissolved at 68°C for 15 min, followed by incubation at 37°C for 15 min. β-Agarase I (2 µl) was added (Calbiochem, 1 U/ml) and the sample incubated at 37°C for 14 hr. After phenol/chloroform/isoamyl alcohol (24:1) extraction, 360 µl of sample was precipitated with 5.4 µl of 10 mg/ml Dextran T40, 95 µl of 1 M NaCl, and 1062 µl of 100% ethanol. The pellet was dissolved in 20 µl of H_2O.

FISH Analysis

PAC or cosmid DNA was labeled with biotin-11-dATP (Gibco BRL) (for green signals) or with digoxygenin-11-dUTP (Boehringer Mannheim) (for red signals). Approximately 0.2 µg of each labeled PAC DNA sample was mixed with 5 µg of Cot1 DNA (Gibco BRL), precipitated, denatured, allowed to preanneal, and then applied to a denatured slide and hybridized overnight. Slides were washed and biotin/digoxygenin presence was detected with avidin-FITC/Anti-digoxygenin-rhodamine, yielding green and red signals, respectively. Chromosome images were captured by use of a Zeiss Axioskop microscope equipped with a charge coupled device (CCD; Photometrics) connected to an Apple Powermac 8100 computer. Separate images of probe signals and DAPI banding patterns were pseudocolored and merged by use of SmartCapture software (Vysis, Inc., Chicago, IL, USA). Results were derived from at least 20 separate and clearly readable metaphases, or at least 30 interphase nuclei (for consistant probe-ordering purposes).

Human–rodent hybrid cell lines were grown in RPMI 1640 (Life Technologies), with antibiotics and 10% fetal calf serum. Genomic DNA was isolated by standard methods.

ACKNOWLEDGMENTS

We thank Pieter de Jong for the access to the PAC library, Colin James for help in electronic analysis of data, students Ana Campos, Olusola Akinwunmi, Julius Adansola, and Joanne Symons for help in the isolation of some cosmid and PAC clones, Denise Sheer for advice in planning FISH experiments, John Crolla for some unpublished confirmatory FISH results, Barbara Groet for critical reading of the manuscript, Charles Buys and Elizabeth Fisher for hybrid cell lines, Jan-Fang Cheng for the collection of published cDNAs, Nico Katsanis for the RIP140 EST and the MRC–Human Genome Mapping Project Resource Centre UK for access to various molecular resources. A.P.S. has been partly supported by the Constance Bequest Fund and P.R.B. by the Science and Technology Programme JNICT-Portugal. We thank the Central Research Fund of the University of London for help in purchase of a machine. This work has been supported by the BMH4-CT96-0554 grant from the Commission of European Communities and partly by the Human Genome Mapping Project Strategic grant G94226164 from the Medical Research Council, UK.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Abe, K.,T. Kajii, and N. Niikawa. 1989. Disomic homozygosity in 21-trisomic cells: A mechanism responsible for transient myeloproliferative syndrome. *Hum. Genet.* **82:** 313–316.

Anand, R. 1995. Cloning into yeast artificial chromosomes. In *DNA cloning 3—Complex genomes: A practical approach* (ed. D. Hames and D. Glover), pp. 103–127. IRL, Oxford University Press, Oxford UK.

Antonarakis, S.E., A.C. Warren, M.K. McCormick, J.G. Lewis, P.A. Hieter, and A. Chakravarti. 1989. Molecular mapping of chromosome 21 and the region responsible for Down syndrome. *Prog. Clin. Biol. Res.* **311:** 29–43.

Baxendale, S., G.P. Bates, M.E. MacDonald, J.F. Gusella, and H. Lehrach. 1991. The direct screening of cosmid libraries with YAC clones. *Nucleic Acids Res.* **19:** 6651.

Brodsky, G., G.A. Otterson, B.B. Parry, I. Hart, D. Patterson, and F.J. Kaye. 1995. Localization of STCH to human chromosome 21q11.1. *Genomics* **30:** 627–628.

Cavailles, V., S. Dauvois, F. L'Horset, G. Lopez, S. Hoare, P.J. Kushner, and M.G. Parker. 1995. Nuclear factor RIP140 modulates transcriptional activation by the estrogen receptor. *EMBO J.* **14:** 3741–3751.

Chen, H., R. Chrast, M.A. Morris, M.D. Lalioti, and S.E. Antonarakis. 1996. Cloning of 559 potential exons of genes of human chromosome 21 by exon trapping. *Genome Res.* **6:** 747–760.

Cheng, J-F., V. Boyartchuk, and Y. Zhu. 1994. Isolation and mapping of human chromosome 21 cDNA: Progress in constructing a chromosome 21 expression map. *Genomics* **23:** 75–85.

Chiang, P-W., G. Dzida, J. Grumet, J-F. Cheng, W-J. Song, E. Crombez, M.L. Van Keuren, and D.M. Kurnit. 1995. Expressed sequence tags from the long arm of human chromosome 21. *Genomics* **29:** 383–389.

Choo, K.H. 1990. Role of acrocentric cen-pter satellite DNA in Robertsonian translocation and chromosomal non-disjunction. *Mol. Biol. Med.* **7:** 437–449.

Chumakov, I., P. Rigault, S. Guillou, P. Ougen, A. Billaut, G. Guascono, P. Gervy, I. Legall, P. Soularue, L. Grinas, et al. 1992. Continuum of overlapping clones spanning the entire human chromosome 21q. *Nature* **359:** 380–387.

Cole, C.G., P.N. Goodfellow, M. Bobrow, and D.R. Bentley. 1991. Generation of novel sequence tagged sites (STSs) from discrete chromosomal regions using ALU-PCR. *Genomics* **10:** 816–826.

Crete, N., J-M. Delabar, Z. Rahmani, M-L. Yaspo, J. Kraus, A. Marks, P.M. Sinet, and N. Creau-Goldberg. 1993. Partial physical map of human chromosome 21 from fibroblast and lymphocyte DNA. *Hum. Genet.* **91:** 245–253.

Dagna-Bricarelli, F., A. Argusti, S. Cavani, M. Pierluigi, C. Perfumo, L. Perroni, P. Strigini, F. Cotter, and D. Nizetic. 1997. Molecular study of the 17 DS families with leukemia using polymorphic markers from chromosome 21. International Conference on Chromosome 21 and Medical Research on Down Syndrome. *Cytogenet. Cell Genet.* **77:** 31.

Delabar, J.M., D. Theophile, Z. Rahmani, Z. Chettouh, J.L. Blouin, M. Prieur, B. Noel, and P.M. Sinet. 1993. Molecular mapping of twenty-four features of Down syndrome on chromosome 21. *Eur. J. Hum. Genet.* **1:** 114–124.

Dutriaux, A., J. Rossier, W. van Hul, D. Nizetic, D. Theofille, J.-M. Delabar, C. van Broeckhoven, and M.-C. Potier. 1994. Cloning and characterization of a 135- to 500-kb region of homology on the long arm of human chromosome 21. *Genomics* **22:** 472–477.

Eki, T., M. Abe, K. Furuya, N. Fujishima, H. Kishida, A. Shiratori, K. Yokoyama, D. Lepaslier, D. Cohen, and Y. Murakami. 1996. 1.8-Megabases fine physical map encompassing IFNAR and AML1 loci on human-chromosome 21q22.1. *DNA Sequence* **6:** 95–108.

Gardiner, K. 1996. Base composition and gene distribution: Critical patterns in mammalian genome organization. *Trends Genet.* **12:** 519–524.

Gardiner, K., M. Horisberger, J. Kraus, U. Tantravahi, J. Korenberg, V. Rao, S. Reddy, and D. Patterson. 1990. Analysis of human chromosome-21—correlation of physical and cytogenetic maps—gene and CpG island distribution. *EMBO J.* **9:** 25–34.

Goate, A.M., A.R. Haynes, M.J. Owen, M. Farrall, L.A. James, L.Y.C. Lai, M.J. Mullan, P. Roques, M.N. Rossor, R. Williamson, and J.A. Hardy. 1989. Predisposing locus for Alzheimer disease on chromosome 21. *Lancet* **8634:** 352–355.

Graw, S.L., K. Gardiner, K. Hall-Johnson, I. Hart, A. Joetham, K. Walton, D. Donaldson, and D. Patterson. 1995. Molecular analysis and breakpoint definition of a set of human chromosome 21 somatic cell hybrids. *Somat. Cell Mol. Genet.* **21:** 415–428.

Hubert, R.S., S. Mitchell, X.N. Chen, K. Ekmekji, C. Gadomski, Z.G. Sun, D. Noya, U.J. Kim, C.R. Chen, H. Shizuya, M. Simon, P.J. de Jong, and J.R. Korenberg. 1997. BAC and PAC contigs covering 3.5Mb of the Down syndrome congenital heart disease region between D21S55 and MX1 on chromosome 21. *Genomics* **41:** 218–226.

Ichikawa, H., F. Hosoda, Y. Arai, K. Shimizu, M. Ohira, and M. Ohki. 1993. A *Not*I restriction map of the entire long arm of human chromosome 21. *Nature Genet.* **:** 361–365.

Ioannou, A.P., C.T. Amemiya, J. Garnes, P.M. Kroisel, H. Shizuya, C. Chen, M.A. Batzer, and P.J. de Jong. 1994. A new bacteriophage P1-derived vector for the propagation of large human DNA fragments. *Nature Genet.* **6:** 84–89.

Katsanis, N., J.H. Ives, J. Groet, D. Nizetic, and E.M.C. Fisher. 1998. Localisation of receptor interacting protein 140 (RIP 140) within 100kb of D21S13 on 21q11, a gene-poor region of the human genome. *Hum. Genet.* **102:** 221–223.

Korenberg, J.R., X-N. Chen, R. Schipper, Z. Sun, R. Gonsky, S. Gerwehr, N. Carpenter, C. Daumer, P. Dignan, C. Disteche et al. 1994. Down syndrome phenotypes: The

consequences of chromosomal imbalance. *Proc. Natl. Acad. Sci.* **91:** 4997–5001.

Korenberg, J.R., X-N. Chen, S. Mitchell, S. Fannin, S. Gerwehr, D. Cohen, and I. Chumakov. 1995. A high-fidelity physical map of human-chromosome 21q in yeast artificial chromosomes. *Genome Res.* **5:** 427–443.

Lafreniere, R.G., P.J. de Jong, and G.A. Rouleau. 1995. A 405-kb cosmid contig and HindIII restriction map of the progressive myoclonus epilypsy type 1 (EPM1) candidate region in 21q22.3. *Genomics* **29:** 288–290.

Lucente, D., H.M. Chen, D. Shea, S.N. Samec, M. Rutter, R. Chrast, C. Rossier, A. Buckler, S.E. Antonarakis, and M.K. McCormick. 1995. Localization of 102 exons to a 2.5Mb region involved in Down syndrome. *Hum. Molec. Genet.* **4:** 1305–1311.

Masuda, H., T. Kazuhiko, M. Takagi, K. Ohgami, T. Sakamaki, N. Shibata, and K. Takahashi. 1996. Molecular cloning and characterizion of human non-smooth muscle calponin. *J. Biochem.* **120:** 415–424.

McInnes, M.G., A. Chakravarti, J. Blaschak, M.B. Petersen, V. Sharma, D. Avramopoulis, J.L. Blouin, U. Konig, C. Brahe, T.C. Matise et al. 1993. A linkage map of human chromosome 21: 43 PCR markers at average intervals of 2.5cM. *Genomics* **16:** 562–571.

Niikawa, N., H.-X. Deng, K. Abe, N. Harada, T. Okada, H. Tsuchiya, I. Akaboshi, I. Matsuda, Y. Fukushima, Y. Kaneko, A. Kuwano, and T. Kajii. 1991. Possible mapping of the gene for transient myeloproliferative syndrome at 21q11.2. *Hum. Genet.* **87:** 561–566.

Nizetic, D. and H.Lehrach. 1995. Chromosome specific gridded cosmid libraries : Construction, handling and use in parallel and integrated mapping. In *DNA cloning 3—Complex genomes: A practical approach* (ed. D. Hames and D. Glover), pp. 49–79. IRL, Oxford University Press, Oxford UK.

Nizetic, D., G. Zehetner, A.P. Monaco, L. Gellen, B.D. Young, and H. Lehrach. 1991a. Construction, arraying, and high-density screening of large insert libraries of human chromosome X and 21: Their potential use as reference libraries. *Proc. Natl. Acad. Sci.* **88:** 3233–3237.

Nizetic, D., R. Drmanac and H. Lehrach. 1991b. An improved colony lysis procedure enables direct DNA hybridisation using short (10,11 bases) oligonucleotides to cosmids. *Nucleic Acids Res.* **19:** 182.

Nizetic, D., L. Gellen, R. Hamvas, R. Mott, A. Grigoriev, R. Vatcheva, G. Zehetner, M.L. Yaspo, A. Dutriau, C. Lopes et al. 1994. An integrated YAC-overlap and "cosmid pocket" map of the human chromosome 21. *Hum. Mol. Genetics* **3:** 759–770.

Ohira, M., H. Ichikawa, E. Suzuki, M. Iwaki, K. Suzuki, F. Saitoohara, T. Ikeuchi, I. Chumakov, H. Tanahashi, K. Tashiro, Y. Sakaki, and M. Ohki. 1996. A 1.6 Mb P1-based physical map of the down syndrome region on chromosome 21. *Genomics* **33:** 65–74.

Ohta, T., M. Nakano, T. Tsujita, K. Abe, K. Osoegawa, T. Yamagata, K-I Yoshiura, Y. Jinno, E. Soeda, Y. Nakamura, and N. Niikawa. 1996. Isolation of a cosmid clone corresponding to an inv(21) breakpoint of a patient with transient abnormal myelopoiesis. *Am. J. Hum. Genet.* **58:** 544–550.

Osoegawa, K., R. Susukida, S. Okano, J. Kudoh, S. Minoshima, N. Shimizu, P.J. de Jong, J. Groet, J. Ives, H. Lehrach, D. Nizetic, and E. Soeda. 1996. An intergrated map with Cosmid/PAC contigs of a 4-Mb Down symdrome critical region. *Genomics* **32:** 375–387.

Patil, N., A. Peterson, A. Rothman, P.J. de Jong, R.M. Mayers, and D.R. Cox. 1994. A high resolution physical map of 2.5 Mbp of the Down syndrome region on chromosome 21. *Hum. Mol. Genet.* **3:** 1811–1817.

Patterson, D., Z. Rahmani, D. Donaldson, K. Gardiner, and C. Jones. 1993. Physical mapping of chromosome 21. In *The phenotypic mapping of Down syndrome and other aneuploid* conditions (ed. C.J. Epstein), pp. 33–50. Wiley-Liss, New York, NY.

Peterson, A., N. Patil, C. Robbins, L. Wang, D.R. Cox, and R.M. Mayers. 1994. A transcript map of the Down syndrome critical region on chromosome 21. *Hum. Mol. Genet.* **3:** 1735–1742.

Potier, M.C., A. Dutriaux, and R. Reeves. 1996. Use of YAC fragmentation to delimit a duplicated region on human chromosome 21. *Mamm. Genome* **7:** 85–88.

Riley, J.A., R. Butler, D.J. Ogilvie, R. Finniear, D. Jenner, R. Anand, J.C. Smith, and A.F. Markham. 1990. A novel, rapid method for the isolation of terminal sequences from yeast artificial chromosome (YAC) clones. *Nucleic Acids Res.* **18:** 2887–2890.

Sambrook, J., E.F. Fritsch, and T. Maniatis. 1989. *Molecular cloning: A laboratory manual.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Schellenberg, G.D., M.A. Pericak-Vance, E.M. Wijsman, D.K. Moore, P.C. Gaskell, L.A. Jamaoka, J.L. Bebut, L. Anderson, K.L. Walsh, C.M. Clark, G.M. Martin, A.D. Roses, and T.D. Bird. 1991. Linkage analysis of familial Alzheimer disease using chromosome 21 markers. *Am. J. Hum. Genet.* **48:** 563–583.

Schuler, G.D., M.S. Boguski, E.A. Stewart, L.D. Stein, G. Gyapay, K. Rice, R.E. White, P. Rodriguez-tomé, A. Aggerwal, E. Bajorek et al. 1996. A gene map of the human genome. *Science* **274:** 540–546.

Shen, J.J., B.J. Williams, A. Zipursky, J. Doyle, S.L. Sherman, P.A. Jacobs, A.L Shugar, S.W Soukup, and T.J. Hassold. 1995. Cytogenetic and molecular studies of Down Syndrome individuals with leukemia. *Am. J. Hum. Genet.* **56:** 915–925.

Shimizu, N., S.E. Antonarakis, C. van Broeckhoven, D. Patterson, K. Gardiner, D. Nizetic, N. Créau, J-M. Delabar, J. Korenberg, R. Reeves, J. Doering, A. Chakravati, S. Minoshima, O. Ritter, and J. Cuticchia. 1995. Report of the

Fifth International Workshop on Human Chromosome 21 Mapping 1994. *Cytogenet. Cell Genet.* **70:** 157–182.

Shizuya, H., B.I. Birren, U-J. Kim, B. Mancino, T. Slepak, Y. Tachiiri, and M.I. Simon. 1992. Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in Escherichia coli using an F-factor-based vector. *Proc. Natl. Acad. Sci.* **89:** 8794–8797.

St. George-Hyslop, P.H., R.E. Tanzi, P.J. Polinsky, J.L. Haines, L. Nee, P.C. Watkins, R.H. Myers, R.G. Feldman, D. Pollen, D. Drachman et al. 1987. The genetic defect causing familial Alzheimer's disease maps on chromosome 21. *Science* **235:** 885–890.

St.George-Hyslop, P.H., J.L. Haines, E.I. Rogaev, M. Motilla, G. Vaula, M.A. Pericak-Vance, J.F. Foncin, M.P. Montesi, A.C. Bruni, S. Sorbi et al. 1992. Genetic evidence for a novel familial Alzheimer's disease locus on chromosome 14. *Nature Genet.* **2:** 330–334.

Stone, N.E., J-B. Fan, V. Willour, L.A. Pennacchio, J.A. Warrington, A. Hu, A. de la Chapelle, A-E. Leheshoki, D.R. Cox and R.M. Myers. 1996. Construction of a 750-kb bacterial clone contig and restriction map in the region of human chromosome 21 containing the progressive myoclonus epilepsy gene. *Genome Res.* **6:** 218–225.

Tassone, F., H. Xu, H. Burkin, S. Weissman, and K. Gardiner. 1995. cDNA selection from 10 Mb of chromosome 21 DNA: Efficiency in transcriptional mapping and reflections of genome organization. *Hum. Mol. Genet.* **4:** 1509–1518.

Van Broeckhoven, C.L. 1995. Molecular genetics of Alzheimer disease: Identification of genes and gene mutations. *Eur. Neurol.* **35:** 8–19.

Van Broeckhoven, C., H. Backhovens, W. Van Hul, G. Van Camp, P. Stinissen, A. Wehnert, P. Raeymaekers, G. De Winter, M. Bruyland, J. Gheuens, J. Martin, and A Vandenberghe. 1990. Genetic linkage analysis in early onset familial Alzheimer's dementia. In *Neuropsychopharmacology* (ed. W.E. Bunney, H. Hippius, G. Laakmann, and N. Schmauss), pp. 86–91. Springer, Berlin, Germany.

van Camp, G., M. Cruts, H. Backhovens, A. Wehnert, and C. Van Broeckhoven. 1992. Unique sequence homology in the pericentromeric regions of the long arms of chromosomes 13 and 21. *Genomics* **12:** 158–160.

van Hul, W. and C. Van Broeckhoven. 1993. A long range restriction map from the APP gene to te centromere of chromosome 21. In Alzheimer's disease: Advances in clinical and basic research (ed. B. Corain, K. Iqbal, M. Nicolini, B. Winblad, H. Wisniewski, and P. Zatta), pp. 235–242. John Wiley & Sons, New York, NY.

van Hul, W., G. Van Camp, L. Stuyver, J.-M. Delabar, M. G. McInnes, A.C. Warren, S.E. Antonarakis, and C. van Broeckhoven. 1993. A contiguous physical map of the pericentric region of chromosome 21q between D21Z1 and D21S13E. *Genomics* **15:** 626–630.

Yaspo, M.-L., L. Gellen, R. Mott, B. Korn, D. Nizetic, A. Poustka, and H. Lehrach. 1995. Model for a transcriptional map of human chromosome 21: Isolation of new coding sequences from exon and enriched cDNA libraries. *Hum. Mol. Genet.* **4:** 1291–1304.

Yu, J., S. Tong, Y. Shen and F.-T. Kao. 1997. Gene identification and DNA sequence analysis in the GC-poor 20 megabase region of human chromosome 21. *Proc. Natl. Acad. Sci.* **94:** 6862–6867.

Received August 4, 1997; accepted in revised form February 19, 1998.

Narrowing of the Region of Allelic Loss in 21q11-21 in Squamous Non-Small Cell Lung Carcinoma and Cloning of a Novel Ubiquitin-Specific Protease Gene From the Deleted Segment

Jürgen Groet,[1] Jane H. Ives,[1] Tania A. Jones,[2] Malcolm Danton,[1] Rachel H. Flomen,[1] Denise Sheer,[2] Reno Hrašćan,[3] Krešimir Pavelić,[3] and Dean Nižetić[1]*

[1]Centre for Applied Molecular Biology, School of Pharmacy, University of London, London, United Kingdom
[2]Human Cytogenetics Laboratory, Imperial Cancer Research Fund, London, United Kingdom
[3]Division of Molecular Medicine, Rudjer Bošković Institute, Zagreb, Croatia

We examined 42 fresh non-small cell lung carcinomas for allelic loss using 4 microsatellite markers located in a 4.5 Mb region in 21q11-21, a gene-poor interval recently found by others to be homozygously deleted and exhibiting frequent allelic loss in lung cancer. We found allelic loss across the entire segment in 13/34 informative squamous carcinomas, with 2 cases showing loss in only part of the region. Analysis by fluorescence in situ hybridization of P1-derived artificial chromosomes from the region directly on paraffin sections of the tumor is in concordance with the loss of heterozygosity (LOH) results, and tentatively excludes a 2 Mb segment bearing 2 of the only 3 known genes in the area. Exon trapping in the remaining segment of loss led to identification and cloning of a novel gene spanning 150 kb within the deletion. The full-length gene encodes a protein of 1,055 amino acids with homology to ubiquitin-specific proteases across the eukaryotic evolutionary spectrum. The expressed protein acts as a de-ubiquitinating enzyme as proved by the ability to cleave ubiquitin from a model fusion protein. We found no mutations in the sequence of the functional domains of this gene in any of the LOH-exhibiting tumor DNA samples. It is, however, interesting that genes of the same superfamily have been reported on 3p21, a locus showing the most frequent allelic instability and deletions in lung cancer. *Genes Chromosomes Cancer 27:153–161, 2000.* © 2000 Wiley-Liss, Inc.

INTRODUCTION

Lung cancer is the most frequent neoplastic cause of death among men, and the second most frequent neoplastic cause of death among women in the United Kingdom. A number of genes with known functions in growth control have been found altered in lung cancer (Xu et al., 1991; Rodenhuis and Slebos, 1992; Sameshima et al., 1992) and a number of sites on the genomic map have shown non-random and significant loss, including homozygous deletions in lung cancer cells. Deletions as well as loss of heterozygosity (LOH) have been detected at 2q33, 3p12-13, 3p14, 3p21-22, 9p21, and 16q24 (Rabbitts et al., 1990; Daly et al., 1993; Yokota and Sugimura, 1993; Sato et al., 1994; Wiest et al., 1997; Sato et al., 1998), indicating that multiple tumor suppressor genes (TSGs) are likely to be involved in its development and progression. In genome-wide analysis of chromosomal alterations and marker allelotyping, chromosome 21 gains or losses were not among the most prominent non-random changes (Virmani et al., 1998). In these studies, however, typically one marker representing chromosome 21 was used, from 21q22.3. Some studies of non-small cell lung carcinoma (NSCLC) nevertheless found 21q loss among the frequently recurrent losses (Sato et al., 1994; Testa et al., 1994). A recent study (Kohno et al., 1998) described a homozygous deletion encompassing more than 3 Mbp of genomic DNA in 21q11-21 in a case of NSCLC. Three different markers in this region showed from 42% to 62.5% incidence of LOH on fresh tumors. A gene called *ANA* with homology to the *TOB/BTG1* family of TSGs has been positioned in 21q21.1 and found to be within the deleted interval (Kohno et al., 1998). However, the same authors excluded this gene as the suppressor, because no mutations of the *ANA* gene were found in any of the cell lines or fresh tumors showing LOH. In this same region (the most proximal 10% of the length of 21q or 21q11-21), frequent chromosomal

The current address of Jane H. Ives is Department of Biological Sciences, University of Durham, Durham, UK.

Supported by: Cancer Research Campaign, UK; Grant number: SP 2475/0101; Leukaemia Research Fund, UK; Grant number: 98/38; Human Genome Mapping Project, Medical Research Council, UK; Grant number: G9422614; Imperial Cancer Research Fund.

*Correspondence to: Dean Nižetić, M.D., Ph.D., Centre for Applied Molecular Biology, School of Pharmacy, University of London, 29-39 Brunswick Square, London, WC1N 1AX, UK.

Received 19 May 1999; Accepted 12 August 1999

© 2000 Wiley-Liss, Inc.

rearrangements and an increased disomic homozygosity of DNA markers have been found repeatedly in Down syndrome (DS)-associated cases of acute myeloid leukemia (AML-M7) and DS-specific cases of transient abnormal myelopoiesis (TAM), compared with DS without leukemia (Abe et al., 1989; Niikawa et al., 1991; Cavani et al., 1998). Based on these findings, a hypothetical model has been proposed (Niikawa et al., 1991) by which disomic homozygosity of a loss-of-function allele of a putative pericentromeric TSG, which physiologically suppresses an unknown chromosome 21-encoded oncogene, could explain the occurrence of leukemia in DS. A high-resolution (sequence ready) physical map of 21q11 has recently been constructed by us (Groet et al., 1998) and chosen as the substrate for sequencing of the genomic DNA in this region in the Human Genome Sequencing Project. The map spans 3.4 Mbp encompassing more than half of the region reported as homozygously deleted in NSCLC (Kohno et al., 1998), and also contains the DNA marker with the highest NSCLC-associated LOH from the same report. In the study presented here, we identify the shared region of overlap (SRO) for the hemizygous loss in squamous NSCLC, which is approximately half the size of the homozygous deletion described by Kohno and coworkers (1998), and which excludes 2 of the only 3 known genes in the proximal third of 21q. In addition, we report the cloning of a novel full-length gene, encoding a protein with structural and functional features of ubiquitin-processing enzymes, located within the SRO.

MATERIALS AND METHODS

Tumor and Normal Tissue Samples

A total of 42 fresh NSCLC cases have been analyzed from the Croatian Tumour Bank (CTB), an initiative with set rules and criteria for accumulation of fresh clinical tumor specimens for molecular studies (Spaventi et al., 1994). Of these, half (including tumors #47 and #61) were samples that were recently studied for LOH of the *NM23-H1* gene (Bosnar et al., 1997), and the other half were fresh tumors (data obtainable from CTB, which also lists the tumor stage in the TNM system, grade, size, and survival data). In each case, the tumors and normal lung tissue specimens (as evaluated by the surgeon) were frozen in liquid nitrogen in the operating room and further stored at −70°C. Genomic DNA was isolated using standard procedures (Sambrook et al., 1989). For each sample, 4 μm serial frozen sections were cut, mounted on glass slides, and stained with hematoxylin-eosin (H&E). A pathologist confirmed the histologic type of the tumor and evaluated the percentage of normal cells within the tumor. Only samples with less than 20% non-tumor cells were used in this study.

Markers and LOH Analysis

Microsatellite analysis was performed using polymerase chain reaction (PCR) with appropriate primer pairs (sequences and PCR conditions as in Genome Data Base, Johns Hopkins University, Baltimore, MD), where the forward primer only from each pair was 5′ fluorescently labeled with Applied Biosystems (ABI; Foster City, CA) Big Dyes™ (6-FAM or HEX). Amplification products were analyzed using an ABI 310 Genetic Analyzer. Size standards (GeneScan 350) were mixed with every sample for accurate sizing; the separation of the mixture of denatured fragments was achieved by electrophoresis through a 47 cm capillary (module GS STR POP4 C) for approximately 30 min. Raw data were analyzed using GeneScan and Genotyper software. LOH ratios were calculated exactly as described in the GeneScan Applications manual provided by ABI. For each individual allele's fluorescence level, an average of 3 independent electrophoresis-analysis cycles on the ABI 310 was used for calculation.

Fluorescence In Situ Hybridization (FISH)

Unstained 4-μm-thick paraffin block sections were fixed to glass slides, and a standard pretreatment protocol was followed for formalin-fixed, paraffin-embedded slides. P1-derived artificial chromosome (PAC) DNA was labeled with digoxygenin-11-dUTP (Boehringer Mannheim, Germany). Approximately 0.5 μg of each labeled PAC DNA sample was mixed with 5 μg of Cot1 DNA (Gibco BRL, Gaithersburg, MD), precipitated, denatured, allowed to preanneal, and then applied to a denatured slide and hybridized overnight. Slides were washed and signal detected using anti-digoxygenin-rhodamine, followed by DAPI counterstain. Images were captured using a Zeiss Axioskop microscope equipped with a charge-coupled device (CCD) (Photometrics, Tucson, AZ) connected to an Apple Powermac 8100 computer. Images were captured on 3 levels of focus, and each level was examined for signals using SmartCapture software (Vysis, Inc., Chicago, IL). Only nuclei with signals were counted in each level, and the number of signals in each cell was determined.

Exon Trapping

DNA from the 2 PACs (73M5 and 135E14) that were shown to be hemizygously deleted in tumor #61 (Fig. 1) was digested using *Bam*HI and *Bgl*II to completion, or partially using *Sau*3AI, and exon trapping was performed with the resulting fragments in pLSP3B vector using standard technology.

Northern Blot Analysis

The cDNAs were labeled by random priming and hybridized to human multiple tissue Northern blots (Clontech, Palo Alto, CA) containing 2 μg polyA+ RNA per lane using the protocol recommended by the manufacturer. The exposure was for 14 hr to Molecular Dynamics (Sunnyvale, CA) Phosphorimager screens. The I.M.A.G.E. Consortium (Lennon et al., 1996) cDNA clone ID 824710 and the Unigene clone A002B43 have been used as labeled probes in separate experiments.

Cleavage Analysis of Ubiquitin-Met-β-Galactosidase Fusion Protein

This analysis was performed essentially as described (Everett et al., 1997). Model fusion protein ubiquitin-Met-β-galactosidase in a pACYC184 (Cm^r replicon) was represented by the plasmid pACYC-Ub-Met-β-gal, a kind gift of R. Everett. Plasmid pRB105 containing a *Saccharomyces cerevisiae* ubiquitin-specific protease UBP2 in an IPTG-inducible pBR322 (Amp^r replicon) was a kind gift of R. Baker, and was used as a positive control. The new gene *USP21* was cloned from nucleotide position 203 to nucleotide position 3367 (numbering as in GenBank AF134213) into *Sac*I/*Sal*I cloning sites of the IPTG-inducible *Escherichia coli* expression vector pQE30 (Qiagen, Chatsworth, CA). The *E. coli* XL-1 blue cells were transformed using a standard rubidium chloride-heat shock method with the combination of pACYC-Ub-Met-β-gal and either pQE30 vector, pQE30-*USP21*, or pRB105, and each of the 3 cotransformants was selected on medium containing chloramphenicol (42 μg/ml) and carbenicillin (75 μg/ml). Western blots were prepared by electrotransfer to a nitrocellulose membrane (Schleicher & Schull, Keene, NH). The β-galactosidase-containing bands were detected by an anti-β-galactosidase polyclonal rabbit antiserum (a kind gift of R. Everett) using an enhanced chemiluminescence (ECL) assay kit (rpn 2132; Amersham, Arlington Heights, IL) under conditions recommended by the manufacturer.

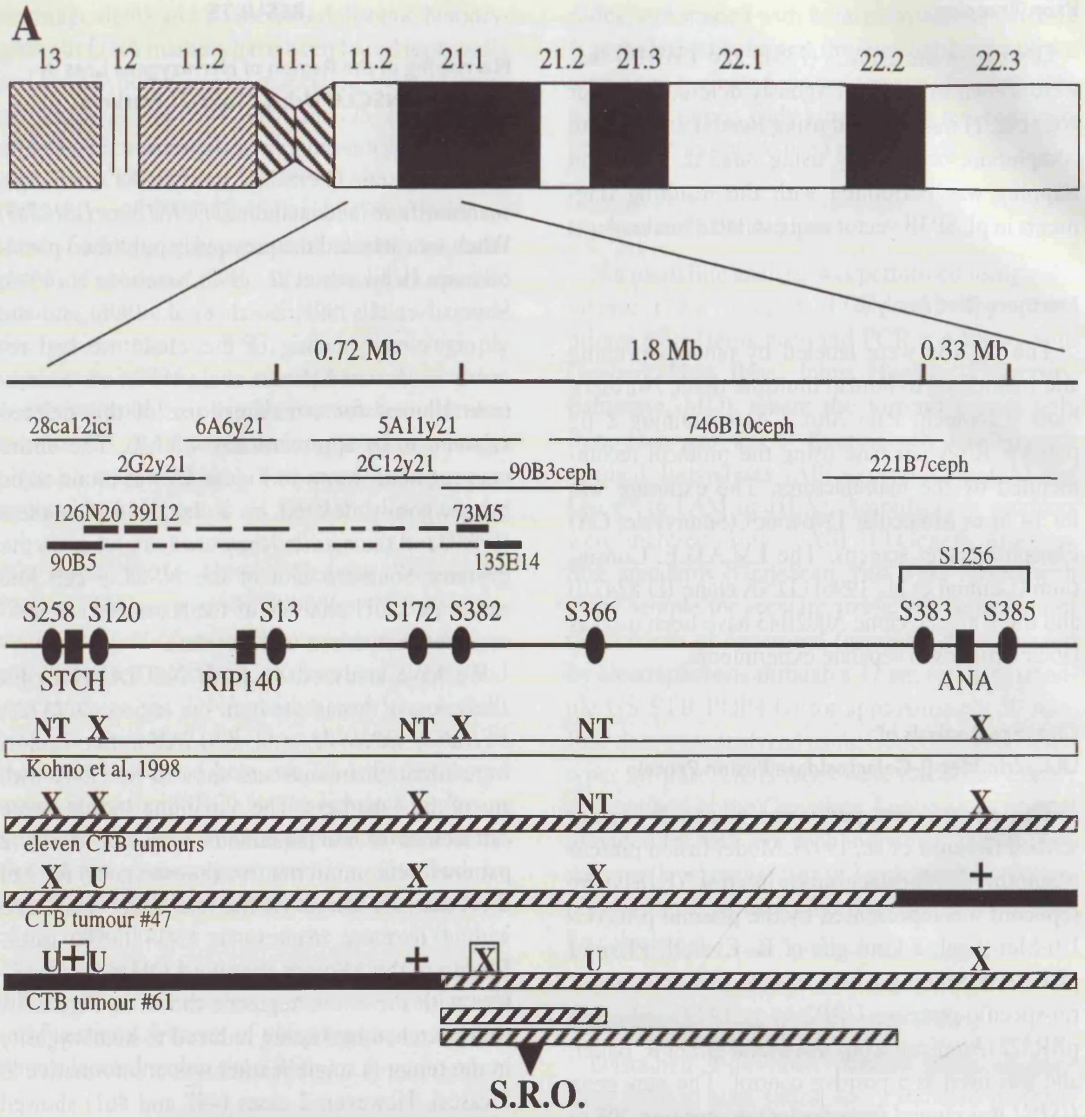
RESULTS

Narrowing of the Region of Hemizygous Loss in Squamous NSCLC Using 21q11-21 Markers

A homozygous deletion was discovered by Kohno and coworkers (1998) in an NSCLC cell line spanning from (and including) *D21S120* to *D21S385*. When we compiled the previously published physical maps (Ichikawa et al., 1993; Nizetic et al., 1994; Shimizu et al., 1995; Bosch et al., 1996) and the sequence-ready contig of the 21q11 we had recently constructed (Groet et al., 1998), our estimations allowed for a maximal size of this deleted segment to be approximately 4.5 Mb. The entire map segment shown in Figure 1A was found to be homozygously deleted by Kohno and coworkers (1998), and the markers they used as probes on the genomic Southern blot of the NSCLC cell line and/or in LOH analysis of fresh tumors are indicated for comparison with our data.

We have analyzed 42 fresh NSCLC cases for allelic loss of 4 markers from the region: *D21S120*, *D21S258*, *D21S172*, and *D21S1256*. Six tumors were adenocarcinomas and showed no LOH with any of the 4 markers. The remaining 36 cases were carcinomas of the squamous type. Of these, 2 patients were uninformative (homozygous) for 3 of the 4 markers tested. Of the rest, LOH was found with 1 or more markers in 13/34 (38%) cases. Eleven of the 13 cases showing LOH were consistent with the entire segment shown in Figure 1A (fully hatched bar) being reduced to homozygosity in the tumor (a single marker was uninformative in 5 cases). However, 2 cases (#47 and #61) showed LOH (X in Fig. 1A) and a lack of LOH (+ in Fig. 1A) with neighboring markers, respectively, suggesting an overlapping SRO. Using these 2 samples we attempted to map the borders of the SRO more precisely. To map the telomeric border, we analyzed tumors #47 and #61 additionally with the marker *D21S366*. This marker is present on the YAC 90B3ceph (Bosch et al., 1996), which overlaps with our bacterial contig (Groet et al., 1998) sharing *D21S382*, extending approximately 350 kbp telomerically (Shimizu et al., 1995; Bosch et al., 1996). This marker was found to be uninformative in #61, but showed LOH in #47, which limits the telomeric border of the SRO to the interval between *D21S366* and *D21S1256*. This latter marker resides on the same YAC as *D21S383*, *ANA*, and *D21S385* (Fig. 1A), but their relative order within the YAC is not known.

We further examined the most centromeric overlapping PACs from the bacterial contig (shown as


A
13
12
11.2
11.1
11.2
21.1
21.2
21.3
22.1
22.2
22.3
0.72 Mb
1.8 Mb
0.33 Mb
28ca12ici
6A6y21
5A11y21
746B10ceph
2G2y21
2C12y21
90B3ceph
221B7ceph
126N20 39I12
73M5
90B5
135E14
S1256
S258 S120
S13
S172 S382
S366
S383 S385
STCH
RIP140
ANA
NT X
NT X
NT
X
Kohno et al, 1998
X X
X
NT
X
eleven CTB tumours
X U
X
X
+
CTB tumour #47
U+U
+
X
U
X
CTB tumour #61
S.R.O.

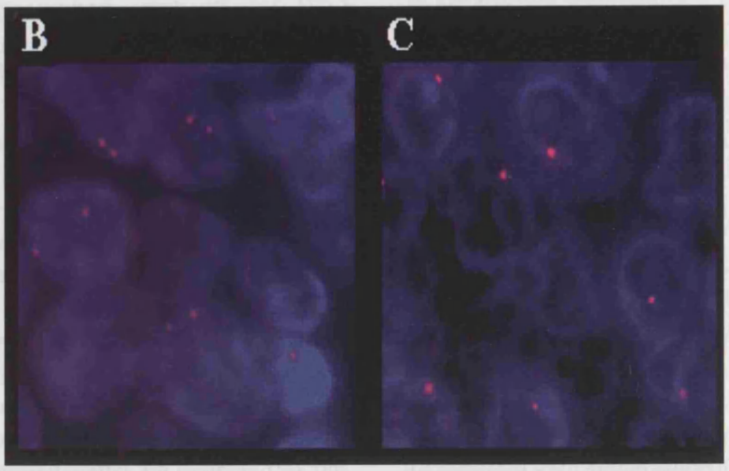

B
C

Figure 1.

short bold overlapping lines in Fig. 1A) by FISH. Tumor sample #47 was not of sufficient quantity or quality for FISH analysis. We performed FISH directly on paraffin sections of tumor #61 using a pool of 3 PACs as a probe (single PACs did not yield sufficiently strong FISH signals): 90B5, 126N20, and B39I12, which define the stretch of 350 kbp containing markers *D21S258*, *D21S120*, and the gene *STCH* (Groet et al., 1998) (Fig. 1A). This pool-probe gave 2 signals per nucleus in 71% of interphase nuclei with signals (a total of 301 nuclei counted) in tumor #61 (Fig. 1B). This result, combined with the lack of LOH at *D21S172* (Fig. 1A), suggests the presence of 2 copies of the segment centromeric to, and including, *D21S172*, in tumor #61. The 29% of nuclei giving a single signal is compatible with an expected level of artifact caused by the thickness of the tissue section (4 μm), cutting some nuclei in half, making it appear as if 1 signal were lost (T.A.J. and D.S., unpublished observations). In contrast, when the telomeric end of the contig was examined (a pool of PACs 73M5 and 135E14, as a probe), 61% of the nuclei gave a single signal out of the total of 294 counted nuclei with signals, and the rest gave 2 signals (Fig. 1C). Assuming the same proportion of artifact as in Figure 1B (double signals reduced to a single, and single signal nuclei reduced to "signal-less" nuclei by the cut, and therefore not counted), the predicted real number of "single signal" nuclei can be calculated as 77%, for the pool of PACs 73M5 and 135E14. This result is compatible with a hemizygous deletion in the tumor for the whole DNA segment represented by the 2 overlapping PACs, with the presence of normal cells infiltrating the cancer cell population at the level of 23%. A very similar result, though with fainter signals, was obtained when each of the 2 PACs (73M5 and 135E14) was used separately as a single probe on tumor #61 (not shown). These 2 PACs define the segment of 210 kbp starting 150 kbp telomeric from *D21S172*. The first PAC 73M5 contains a CpG island (Groet et al., 1998) at *D21S382* (LL56; Ichikawa et al., 1993; Kohno et al., 1998). As the heterozygosity of marker *D21S172* was clearly not lost in tumor #61, this places the centromeric border of the SRO within the 150 kbp interval (Groet et al., 1998) between *D21S172* and the proximal end of the PAC 73M5, near *D21S382*. Taken together, these mapping data allow for the SRO, shown in Figure 1A in its minimal and maximal allowed sizes (not drawn completely to scale) to be between 250 and 2,500 kbp, and definitely including PACs 73M5 and 135E14 and marker *D21S366*.

Figure 1. Identification of the SRO for hemizygous deletions in 21q11-21 in NSCLC. **A:** Cytogenetic map, *Not*I long-range physical map (Ichikawa et al., 1993), YAC contig (Nizetic et al., 1994; Shimizu et al., 1995; Bosch et al., 1996), and bacterial contig (Groet et al., 1998) are shown in consecutive horizontal layers, respectively, above the line showing the markers used in the LOH analysis (oval symbols) and known genes (rectangular symbols). Markers are named as in Genome Data Base (prefix "D21" omitted). In the column under each marker, an X (symbolizing LOH), + (an absence of LOH), and U (uninformative homozygous result) for that marker in the set of 11 Croatian Tumour Bank (CTB) tumors, or in individual tumors #47 and #61, are shown. NT, not tested. For comparison with our data, markers used as probes on the genomic Southern blot of the NSCLC cell line and/or in LOH analysis of fresh tumors in the study by Kohno and coworkers (1998) are indicated above the empty bar symbolizing the homozygous deletion they found. In our data, hatched bars indicate hemizygous deletions, and black filled bars indicate segments showing absence of LOH or deletions. Squared symbols X and + stand for predominantly single and predominantly double signals, respectively, detected by fluorescence in situ hybridization (FISH) on interphase nuclei of the paraffin-embedded sections of tumor #61, when PAC clones named and indicated as bold lines in the PAC contig above the marker line were used as probes. **B:** FISH using a pool of PACs 90B5, 126N20, and BAC 39I12 as a probe on the paraffin-embedded sections of tumor #61. Two signal nuclei are predominant. **C:** FISH using a pool of PACs 73M5 and 135E14 as a probe on the paraffin-embedded sections of tumor #61. Single signal nuclei are predominant.

Exon Trapping From PACs Within the Narrower Region of Allelic Loss Leads to Cloning of a Novel Gene Encoding a Product That Cleaves Ubiquitin From Carboxy-Terminal Fusion Proteins

We restriction-digested the DNA from the 2 PACs (73M5 and 135E14) that were shown in Figure 1 to be within the SRO (hemizygously deleted in tumor #61), and performed exon trapping with the resulting fragments. Twelve sequenced trapped exons, when analyzed using BLAST-N against public sequence databases, revealed clusters of overlapping cDNA clones, which matched exactly the sequences of the cDNAs forming contigs with a large open reading frame (ORF). In 3 cases (see Fig. 2)—from EST 824710 to AA209364, from AA307805 to AA081200, and from N92952r to Z45010—our trapped exon sequences served to bridge the gaps in the gene sequence. The combined sequence (GenBank accession No. AF134213) revealed a 199 bp 5′ untranslated region (UTR), start codon, an ORF of 3,165 nucleotides encoding a protein of 1,055 amino acids, a stop codon, a 3′UTR of 435 nucleotides, and a polyadenylation signal. The total length (without the polyA) assembled is 3,803 nucleotides. On multiple human tissue Northern blots (Fig. 3) a band of 4.1 kb is visible in all 16 tissues tested (including the normal human lung tissue) with a varying intensity. It is most prominent in skeletal muscle and testis, and the latter tissue also reveals a

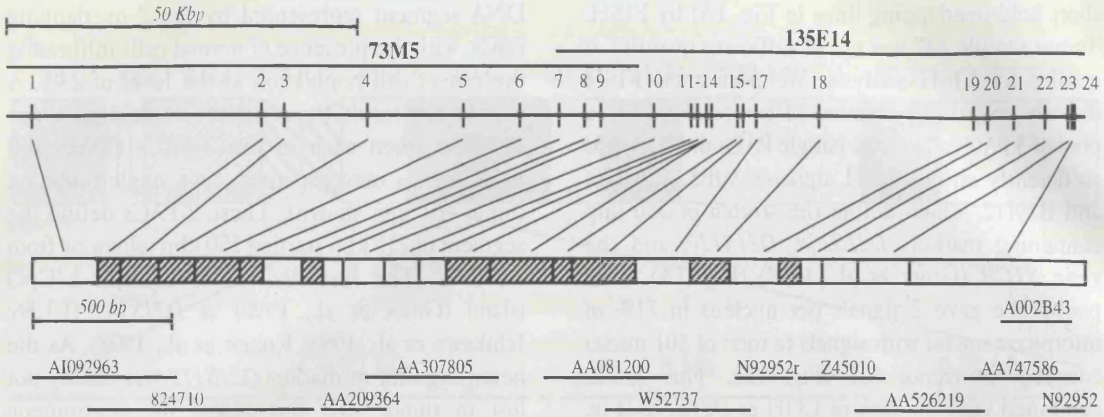


Figure 2. Trapped exons (hatched rectangles) and exons deduced from overlapping sequence analysis (white rectangles) defining the exon-intron structure of the new gene *USP21*. Top half shows 2 PACs 73M5 and 135E14, also used as FISH probes in Figure 1, which were the source of genomic DNA for exon trapping. Exon locations on the PACs are shown with vertical bars, and the 50 kbp scale bar refers to this part. Bottom half consists of overlapping cDNA fragments corresponding to exons above them, drawn in the same scale (500 bp scale bar is shown). Names of cDNA clones are as in dB-EST and Unigene databases; 824710 is the address of the clone in the I.M.A.G.E. Consortium collection. The complete cDNA sequence for the whole gene is the new GenBank entry with the accession No. AF134213.

prominent shorter hybridizing transcript of 1.4 kb. In the course of this analysis, the whole genomic sequence of the 2 PACs (73M5 and 135E14) became publicly available by the German Human Genome Sequencing Consortium (EMBL accession Nos. AJ010597 and AJ010598). Comparison of the genomic sequence with the overlapping cDNA clones and exon sequences revealed that 12 of the 24 exons had been exon trapped (hatched rectangles in Fig. 2). It also became apparent that the region immediately preceding the first exon of the gene comprises the known chromosome 21 CpG island at *D21S382* (also known as LL56 *Not*I linking clone on the *Not*I physical map of 21q; Ichikawa et al., 1993). When the deduced polypeptide sequence was compared with Swissprot and other public databases using BLAST-P, a clear pattern of significant homologies ($e = 10^{-6}$ to 10^{-29}) to proteins across the evolutionary spectrum of eukaryotes was found (Fig. 4): all of these proteins belonged to the superfamily of ubiquitin-specific proteases (USPs) or ubiquitin carboxy-terminal hydrolases (UCHs) (Baker et al., 1992; Swanson et al., 1996; Everett et al., 1997; Fujiwara et al., 1998; Hansen-Hagge et al., 1998; Jensen et al., 1998; Wilkinson and Hochstrasser, 1998; The C. elegans Sequencing Consortium, 1998). The polypeptide sequences were most highly conserved around the 3 domains (the Cys box, QQD box, and His box; Fig. 4) known to be essential for the main function of these enzymes: the cleavage of ubiquitin at its carboxy terminus from extension proteins (ubiquitin precursors) and from ubiquitinated proteins and protein fragments targeted for the degradation by the 26S proteasome pathway (Wilkinson and Hochstrasser, 1998). The cysteine residue at position 178 and the histidine residues at positions 599 and 607 (marked with arrowheads in Fig. 4), which were shown to be an absolute requirement for the function of USPs and UCHs (Hansen-Hagge et al., 1998; Wilkinson and Hochstrasser, 1998), were found in the correct positions in the sequence of the new gene. Because this is a first member of the USP family known to map onto human chromosome 21, we named this protein *USP21*, for ubiquitin-specific protease on chromosome 21. The ability of *USP21* to cleave a model ubiquitin fusion protein substrate was investigated by coexpression in *E. coli*. The complete coding sequence of *USP21* was cloned into a T5-driven, IPTG-inducible ex-

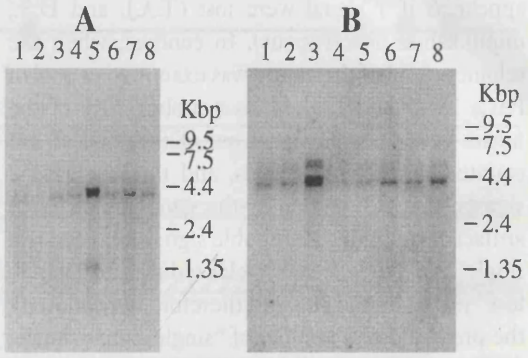


Figure 3. Multiple tissue Northern blots (Clontech) containing 2 μg of polyA+ RNA per lane, hybridized using parts of *USP21* cDNA sequence as probes. **A:** The Unigene clone A002B43, which corresponds to nucleotides 3492–3753 of the *USP21* sequence (GenBank AF134213), and mostly to the 3′ end of the gene (see Fig. 2), was the probe. **Lanes: 1,** peripheral blood leukocyte; **2,** colon; **3,** small intestine; **4,** ovary; **5,** testis; **6,** prostate; **7,** thymus; **8,** spleen. **B:** The cDNA 824710 corresponding to the 5′ end of the gene (nucleotides 200–1017; see also Fig. 2) was the probe. **Lanes: 1,** pancreas; **2,** kidney; **3,** skeletal muscle; **4,** liver; **5,** lung; **6,** placenta; **7,** brain; **8,** heart.

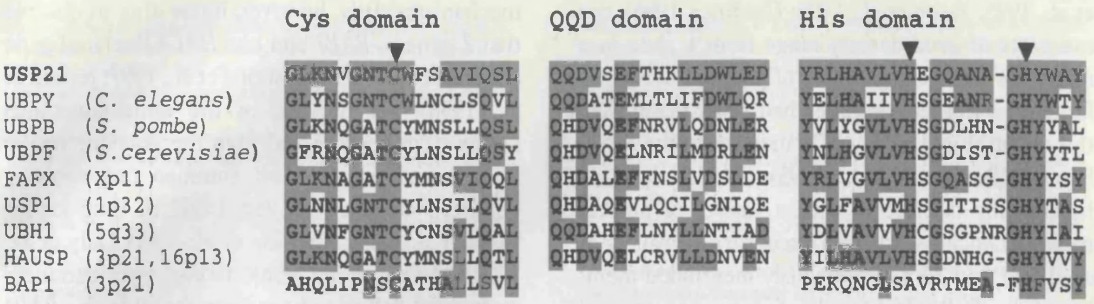


Figure 4. Comparison of protein sequences of *USP21* with other eukaryotic members of the superfamily of USPs (see Results). The protein BAP1 is actually from the family of ubiquitin C-terminal hydrolases, a distinct subfamily of this superfamily, showing homology only in the single key amino acids in the cys and his domains (arrowheads). Two reports show the localizations of the highly homologous sequences for the *HAUSP* gene to 3p21 (Kashuba et al., 1997) and 16p13 (Robinson et al., 1998), respectively.

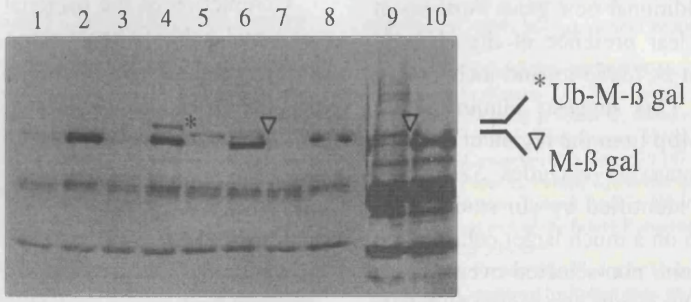


Figure 5. Demonstration of the deubiquitinating activity of *USP21* on a model ubiquitin fusion protein. Western blot of an SDS-PAGE was detected using an anti-β-galactosidase antiserum. **Lanes 1, 2:** *E. coli* XL-1 blue cells alone (in all cases, the second line of the pair is +IPTG); **lanes 3, 4:** same cells cotransfected with the model fusion protein encoding plasmid pACYC-Ub-Met-β-gal and pQE30 plasmid vector only (note the presence of the model ubiquitin-Met-β-galactosidase protein, band labeled with an asterisk); **lanes 5, 6:** as lanes 3, 4, except pQE30-*USP21* (full-length *USP21* gene cloned in the pQE30 expression vector) was added instead of pQE30; **lanes 7, 8:** same as lanes 3, 4, except pRB105 (yeast deubiquitinating enzyme UBP2) was transfected instead of pQE30; **lanes 9, 10:** overexposure of lanes 7, 8. Note the presence of the 8 kDa shorter, deubiquitinated Met-β-galactosidase (band labeled with a triangle).

pression vector (pQE30). As can be seen in Figure 5, the uncleaved Ub-Met-β-gal substrate (band labeled with an asterisk in Fig. 5, lane 4) converts to an 8 kDa shorter band (triangle in Fig. 5) in the cells cotransformed with either *USP21* (lanes 5, 6) or the yeast UBP2-expressing plasmid (lanes 9, 10). This result demonstrates that the novel gene product named *USP21* can efficiently function as a deubiquitinating enzyme.

Analysis of *USP21* Sequence in DNA Samples of NSCLC

We performed direct cycle sequencing of segments of the *USP21* coding sequence. The segments of sequence (GenBank accession No. AF134213) 619–734, 755–841, 980–1027, 1504–1665, 1922–2077, and 2193–2349 were analyzed by direct sequencing from tumor and normal tissue DNA for all 13 NSCLC cases exhibiting LOH (Fig. 1A). These include all 3 functional domains (Fig. 4), plus sequences around "cga" codons, known to be the most frequent sites of causative point mutations in tumor suppressors (Mancini et al., 1997) due to deaminations of CpG cytosines following demethylation, converting them to thymines, and creating "tga" stop codons. Sequences obtained for 13 tumor and normal DNA samples from the 13 cases showing LOH (Fig. 1A) were all identical, and all corresponded exactly to the sequence of *USP21* as in GenBank. For 2 LOH-exhibiting tumors (#5 and #13 from the CTB) we sequenced the entire length of the *USP21* gene from genomic DNA (nucleotides 245–3803), except exon 1, which is very G/C rich. Again, no mutations were detected and the sequence corresponded exactly to the sequence of *USP21* as in GenBank.

DISCUSSION

The findings of high incidence of LOH with 3 markers and a homozygous deletion in a 4.5 Mb region in 21q11-21 (Kohno et al., 1998) set the scene for a search for a potential gene with a tumor suppressive role in lung cancer in this region. The proximal third of the chromosome 21 long arm is an exceptionally gene-poor region of the human genome as estimated by a number of criteria (Shimizu

et al., 1995; Yaspo et al., 1995; Gardiner, 1996); the estimates of gene density range from 1 gene in a megabase to 1 gene in 6 Mb of genomic DNA. So far, only 3 full-length genes had been mapped in this region (Fig. 1A): *STCH* (a member of the hsp70 family) (Brodsky et al., 1995), *RIP140* (protein functionally interacting with a variety of nuclear receptors such as estrogen receptor) (Cavailles et al., 1995), and *ANA* (a previously mentioned member of the *TOB/BTG1* family of tumor suppressors) (Kohno et al., 1998). The results presented in our report, although on a small number of cases, suggest elimination of 2 of the 3 genes from the region hemizygously deleted in squamous lung carcinomas, and add an additional new gene. Analysis of tumor #61 shows clear presence of the chromosomal segment from *D21S258* to, and inclusive of, *D21S172* (Fig. 1). This suggests elimination of approximately 2.0 Mbp from the region of deletion (Fig. 1A), and tentatively excludes *STCH* and *RIP140*. The SRO identified by our study needs further confirmation on a much larger collection of tumors before random, non-selected events could be fully excluded. Our results also cannot eliminate the *ANA* gene from the SRO. However, Kohno and coworkers (1998) found no mutations in this gene in any of the LOH-bearing cases they studied. The centromeric border of the hemizygous deletion in tumor #61 is just centromeric of the PAC 73M5. This PAC, and 135E14 overlapping with it, contains (Fig. 2) the new gene we cloned and characterized in this work, *USP21*. From the homologies in the functional domains and from its ability to hydrolyze the bond between the C-terminal double glycine of ubiquitin and the linking methionine residue (Fig. 5), it can be concluded that *USP21* is a member of ubiquitin-specific proteases (Wilkinson and Hochstrasser, 1998). One of the main functions of this group of enzymes is deubiquitination resulting in rescue from destruction of proteins labeled for degradation in the 26S proteasome. This proteolytic pathway plays an important role in the regulation of levels of proteins which are critical for cell-cycle signaling and programed cell death, such as cyclin B, *CDKN1B*, and *TP53* (King et al., 1995; Lane, 1998; Scheffner, 1999). We found no mutations either in the functionally important regions of this gene or in the segments containing CpG codons (known mutation hot spots) in any of the LOH-bearing cases in our collection. This makes this gene less likely to be a tumor suppressor fitting a classical 2-hit model (Knudson, 1971), although we cannot exclude the presence of other, less predictable mutations, or methylation-silencing mechanisms. It is, however, interesting to observe that 2 genes—*BAP1* and one *HAUSP*-related gene (Everett et al., 1997; Kashuba et al., 1997; Jensen et al., 1998)—for proteins of the same functional family have been cloned from the genomic region 3p21, which is the most common one showing homozygous deletions and LOH in lung cancer (Kok et al., 1987; Rabbitts et al., 1990; Daly et al., 1993). Moreover, intragenic homozygous rearrangements and deletions have been found in the *BAP1* gene in lung carcinoma cell lines (Jensen et al., 1998). These findings may point to a need to investigate this family of proteins further for a potential suppressive role in lung cancer.

Completion of the bacterial contig maps of the proximal half of chromosome 21, and large-scale sequencing must be awaited to rule out completely the presence of any other TSGs which could explain deletions in this region in NSCLC.

ACKNOWLEDGMENTS

We are grateful to J. Karadža, Š. Križanac, S. Levanat, and J. Pavelić for help in obtaining clinical material and tissue sections, R. Everett and R. Baker for the ubiquitin model plasmids and antibodies, P. Chazot for help with the Western blot technique, and the MRC-Human Genome Mapping Resource Centre (Hinxton, UK) for various molecular resources. The Croatian Tumour Bank is supported by grant 101751 from the Ministry of Science and Technology of the Republic of Croatia.

REFERENCES

Abe K, Kajii T, Niikawa N. 1989. Disomic homozygosity in 21-trisomic cells: a mechanism responsible for transient myeloproliferative syndrome. Hum Genet 82:313–316.

Baker RT, Tobias JW, Varshavsky A. 1992. Ubiquitin-specific proteases of *Saccharomyces cerevisiae*. J Biol Chem 267:23364–23375.

Bosch A, Guimera J, Graw S, Gardiner K, Chumakov I, Patterson D, Estivill X. 1996. Integration of 30 CA-repeat markers into the cytogenetic, genetic and YAC maps of human chromosome 21. Eur J Hum Genet 4:135–142.

Bosnar MH, Pavelic K, Krizanac S, Slobodnjak Z, Pavelic J. 1997. Squamous cell lung carcinomas: the role of nm23-H1 gene. J Mol Med 75:609–613.

Brodsky G, Otterson GA, Parry BB, Hart I, Patterson D, Kaye FJ. 1995. Localization of STCH to human chromosome 21q11.1. Genomics 30:627–628.

Cavailles V, Dauvois S, L'Horset F, Lopez G, Hoare S, Kushner PJ, Parker MG. 1995. Nuclear factor RIP140 modulates transcriptional activation by the estrogen receptor. EMBO J 14:3741–3751.

Cavani S, Perfumo C, Argusti A, Pierluigi M, Perroni L, Schmiegelow K, Petersen MB, Cotter FE, Strigini P, Dagna-Bricarelli F, Nizetic D. 1998. Cytogenetic and molecular study of 32 Down syndrome families: potential leukaemia predisposing role of the most proximal segment of chromosome 21q. Br J Haematol 103:213–216.

Daly MC, Xiang RH, Buchhagen D, Hensel CH, Garcia DK, Killary AM, Minna JD, Naylor SL. 1993. A homozygous deletion on chromosome 3 in a small cell lung cancer cell line correlates with a region of tumour suppressor activity. Oncogene 8:1721–1729.

Everett RD, Meredith M, Orr A, Cross A, Kathoria M, Parkinson J. 1997. A novel ubiquitin-specific protease is dynamically associated with the PML nuclear domain and binds to a herpesvirus regulatory protein. EMBO J 16:1519–1530.

Fujiwara T, Saito A, Suzuki M, Shinomiya H, Suzuki T, Takahashi E, Tanigami A, Ichiyama A, Chung CH, Nakamura Y, Tanaka K. 1998. Identification and chromosomal assignment of USP1, a novel gene encoding a human ubiquitin-specific protease. Genomics 54:155–158.

Gardiner K. 1996. Base composition and gene distribution: critical patterns in mammalian genome organization. Trends Genet 12:519–524.

Groet J, Ives JH, South AP, Baptista PB, Jones TA, Yaspo M-L, Lehrach H, Potier M-C, Van Broeckhoven C, Nizetic D. 1998. Bacterial contig map of the 21q11 region associated with Alzheimer's disease and abnormal myelopoiesis in Down syndrome. Genome Res 8:385–398.

Hansen-Hagge TE, Janssen JWG, Hameister H, Papa FR, Zechner U, Seriu T, Jauch A, Becke D, Hochstrasser M, Bartram CR. 1998. An evolutionarily conserved gene on human chromosome 5q33-q34, UBH1, encodes a novel deubiquitinating enzyme. Genomics 49:411–418.

Ichikawa H, Hosoda F, Arai Y, Shimizu K, Ohira M, Ohki M. 1993. A NotI restriction map of the entire long arm of human chromosome 21. Nat Genet 4:361–366.

Jensen DE, Proctor M, Marquis ST, Gardner HP, Ha SI, Chodosh LA, Ishov AM, Tommerup N, Vissing H, Sekido Y, Minna J, Borodovsky A, Schultz DC, Wilkinson KD, Maul GG, Barlev N, Berger SL, Prendergast GC, Rauscher FJ III. 1998. BAP1: a novel ubiquitin hydrolase which binds to the BRCA1 RING finger and enhances BRCA1-mediated cell growth suppression. Oncogene 16:1097–1112.

Kashuba VI, Gizatullin RZ, Protopopov AI, Allikmets R, Korolev S, Li J, Boldog F, Tory K, Zabarovska V, Marcsek Z, Sumegi J, Klein G, Zabarovsky ER, Kisselev L. 1997. NotI linking/jumping clones of human chromosome 3: mapping of the TFRC, RAB7 and HAUSP genes to regions rearranged in leukemia and deleted in solid tumours. FEBS Lett 419:181–185.

King RW, Peters J-M, Tugendreich S, Rolfe M, Hieter P, Kirschner MW. 1995. A 20S complex containing CDC27 and CDC16 catalyzes the mitosis-specific conjugation of ubiquitin to cyclin B. Cell 81:279–288.

Kohno T, Kawanishi M, Matsuda S, Ichikawa H, Takada M, Ohki M, Yamamoto T, Yokota J. 1998. Homozygous deletion and frequent allelic loss of the 21q11.1-q21.1 region including the ANA gene in human lung carcinoma. Genes Chromosomes Cancer 21:236–243.

Kok K, Osinga J, Carritt B, Davis MB, van der Hout AH, van der Veen AY, Landsvater RM, de Leij LFMH, Berendsen HH, Buys CHCM. 1987. Deletion of a DNA sequence at the chromosomal region 3p21 in all major types of lung cancer. Nature 330:578–581.

Knudson AG Jr. 1971. Mutation and cancer: statistical study of retinoblastoma. Proc Natl Acad Sci USA 68:820–823.

Lane D. 1998. Awakening angels. Nature 394:616–617.

Lennon GG, Auffray C, Polymeropoulos M, Soares MB. 1996. The I.M.A.G.E. Consortium: an integrated molecular analysis of genomes and their expression. Genomics 33:151–152.

Mancini D, Shiva S, Ainsworth P, Rodenhiser D. 1997. Constitutively methylated CpG dinucleotides as mutation hot spots in the retinoblastoma gene (RBI). Am J Hum Genet 61:80–87.

Niikawa N, Deng H-X, Abe K, Harada N, Okada T, Tsuchiya H, Akaboshi I, Matsuda I, Fukushima Y, Kaneko Y, Kuwano A, Kajii T. 1991. Possible mapping of the gene for transient myeloproliferative syndrome at 21q11.2. Hum Genet 87:561–566.

Nizetic D, Gellen L, Hamvas R, Mott R, Grigoriev A, Vatcheva R, Zehetner G, Yaspo M-L, Dutriaux A, Lopes C, Delabar JM, Van Broeckhoven C, Potier MC, Lehrach H. 1994. An integrated YAC overlap and "cosmid-pocket" map of the human chromosome 21. Hum Mol Genet 3:759–770.

Rabbitts P, Bergh J, Douglas J, Collins F, Waters J. 1990. A submicroscopic homozygous deletion at the D3S3 locus in a cell line isolated from a small cell lung carcinoma. Genes Chromosomes Cancer 2:231–238.

Robinson PA, Lomonte P, Leek P, Markham AF, Everett RD. 1998. Assignment of herpesvirus-associated ubiquitin-specific protease gene HAUSP to human chromosome band 16p13.3 by in situ hybridisation. Cytogenet Cell Genet 83:100.

Rodenhuis S, Slebos RJ. 1992. Clinical significance of ras oncogene activation in human lung cancer. Cancer Res 52:2665s–2669s.

Sambrook J, Fritsch EF, Maniatis T. 1989. Molecular cloning: a laboratory manual. 2nd ed. Cold Spring Harbor, NY: Cold Spring Harbor Press.

Sameshima Y, Matsuno Y, Hirohashi S, Shimosato Y, Terada M, Yokota J. 1992. Alterations of the p53 gene are common and critical events for the maintenance of malignant phenotypes in small-cell lung carcinoma. Oncogene 7:451–457.

Sato M, Mori Y, Sakurada A, Fukushige S, Ishikawa Y, Tsuchiya E, Saito Y, Nukiwa T, Fujimura S, Horii A. 1998. Identification of a 910-kb region of common allelic loss in chromosome bands 16q24.1-q24.2 in human lung cancer. Genes Chromosomes Cancer 22:1–8.

Sato S, Nakamura Y, Tsuchiya E. 1994. Difference of allelotype between squamous cell carcinoma and adenocarcinoma of the lung. Cancer Res 54:5652–5655.

Scheffner M. 1999. Moving protein heads for breakdown. Nature 398:103–104.

Shimizu N, Antonarakis S, Van Broeckhoven C, Patterson D, Gardiner K, Nizetic D, Creau N, Delabar J-M, Korenberg J, Reeves R, Doering J, Ritter O, Cuticchia J. 1995. Report of the Fifth International Workshop on Human Chromosome 21 Mapping 1994. Cytogenet Cell Genet 70:147–182.

Spaventi R, Pecur L, Pavelic K, Pavelic ZP, Spaventi S, Stambrook PJ. 1994. Human Tumour Bank in Croatia: a possible model for a small bank as part of the future European Tumour Bank Network. Eur J Cancer 30A:419.

Swanson DA, Freund CL, Ploder L, McInnes RR, Valle D. 1996. A ubiquitin C-terminal hydrolase gene on the proximal short arm of the X chromosome: implications for X-linked retinal disorders. Hum Mol Genet 5:533–538.

Testa JR, Siegfried JM, Liu Z, Hunt JD, Feder MM, Litwin S, Zhou J, Taguchi T, Keller SM. 1994. Cytogenetic analysis of 63 non-small cell lung carcinomas: recurrent chromosome alterations amid frequent and widespread genomic upheaval. Genes Chromosomes Cancer 11:178–194.

The C. elegans Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans:* a platform for investigating biology. Science 282:2012–2018.

Virmani AK, Fong KM, Kodagoda D, McIntire D, Hung J, Tonk V, Minna JD, Gazdar AF. 1998. Allelotyping demonstrates common and distinct patterns of chromosomal loss in human lung cancer types. Genes Chromosomes Cancer 21:308–319.

Wiest JS, Franklin WA, Otstot JT, Forbey K, Varella-Garcia M, Rao K, Drabkin H, Gemmill R, Ahrent S, Sidransky D, Saccomanno G, Fountain JW, Anderson MW. 1997. Identification of a novel region of homozygous deletion on chromosome 9p in squamous cell carcinoma of the lung: the location of a putative tumor suppressor gene. Cancer Res 57:1–6.

Wilkinson KD, Hochstrasser M. 1998. Deubiquitinating enzymes. In: Peters JM, Finley D, Harris R, editors. Ubiquitin and the biology of the cell. New York: Plenum Press.

Xu HJ, Hu SX, Cagle PT, Moore GE, Benedict WF. 1991. Absence of retinoblastoma protein expression in primary non-small cell lung carcinomas. Cancer Res 51:2735–2739.

Yaspo M-L, Gellen L, Mott R, Korn B, Nizetic D, Poustka A, Lehrach H. 1995. Model for a transcript map of human chromosome 21: isolation of new coding sequences from exon and enriched cDNA libraries. Hum Mol Genet 4:1291–1304.

Yokota J, Sugimura T. 1993. Multiple steps in carcinogenesis involving alterations of multiple tumour suppressor genes. FASEB J 7:920–925.