Genetic Analysis of a Candidate Region for Psychiatric Illness on Chromosome 4p

Sarah Underwood

Thesis submitted for the degree of
Doctor of Philosophy
The University of Edinburgh

2004

Declaration

I declare that this thesis has been composed by myself, that all the work is my own unless otherwise clearly stated and that the work has not been submitted for any other degree or professional qualification.

Sarah Underwood

Acknowledgements

I would like to thank Kathy Evans and David Porteous for their unfailing support and guidance during this PhD. Special thanks must also go to fellow members of the chromosome 4 group, in particular, Helen Torrence, Susan Anderson, Stewart Morris, Stephanie LeHellard and Pippa Thomson. A big thanks also goes to all members of the Medical Genetics Lab, especially our long suffering lab managers and general do-gooders Heather Davidson, Ann Docherty and Sheila Christie, (and of course Susan and Helen again), who have helped enormously over the years with day to day life in the lab. Life would have been much harder without Alison Condie at the CRF, and before her Angie, tirelessly aiding in the continual sequencing and genotyping effort. Many thanks to all those at the Sanger Institute, especially, Gareth Howell, Alison Coffey and Jackie Bye, who lent me the use of their cDNA libraries and lab facilities and devoted many hours to helping me out. Life simply would not have been the same without my fellow inmates Simon and Jean. You made life bearable, if not a little fun. And last but definitely not least, Andrew and mum, my emotional crutches. And dad, I know you can see this somewhere; this ones for you.

Abstract

Psychiatric illnesses are debilitating conditions for those affected and place a significant burden on the National Health Service, the social services and the family. Here I describe genetic analysis, physical mapping and transcript mapping of a region of chromosome 4p that is linked to psychiatric illness, including bipolar and unipolar affective disorders and schizophrenia.

I have studied four families that show linkage of psychiatric illness to chromosome 4p. Linkage was first observed in a large family, F22, segregating bipolar affective disorder (BPAD) and recurrent major depression (RMD). Subsequently, a smaller family, F59, segregating affective disorders (Blackwood *et al*, 1996_a), and two families (F50 & F48) segregating schizophrenia (SCZ), schizoaffective disorder and BPAD confirmed this linkage.

Previously, comparison of the haplotypes inherited with illness in each family allowed prioritisation of two sub-regions for detailed study. Minimal Region One (MR1) is defined by overlap of the disease chromosomes from three Celtic families (F22, F59 & F50). Minimal Region Two (MR2) is defined by the two largest families F22 and F48, as well as F50. The sequence available from the human genome sequencing project for these two regions is largely complete. Here, I describe an extension to the BAC map in the repetitive telomeric end of MR1. The telomeric end of MR1 is defined by a recombination event in an individual from F50. I mapped clones, designed markers and refined the position of the recombination breakpoint. I also refined the position of the recombination breakpoint at the centromeric end of MR1, as defined by a member of F59.

I describe construction of a transcript map of MR1and 2 using bioinformatics methods, RT-PCR and cDNA library screening. I then selected two candidate genes from this region: orphan g-protein-coupled receptor 78 (GPR78) and superoxide dismutase 3 (SOD3), for further study. Firstly, I identified SNPs in the genes from the linked families, and then carried out a preliminary association study on 95 SCZ

patients, 93 BPAD patients 95 controls. The linkage disequilibrium (LD) between the markers was measured and, using a low stringency significant p-value cut off, revealed a positive association in GPR78. SNPs were then tested on a larger population for association. This work adds to the case for studying the role of chromosome 4 in the genetic susceptibility to affective disorder.

Contents

Chapter 1 Introduction

1.1.	**Psychiatric Illness: The Clinical Picture**	**2**
1.1.1.	Phenotypes and Diagnosis	2
1.1.2.	Pharmacological Treatments	6
1.1.2.1.	Schizophrenia	6
1.1.2.2.	Bipolar Affective Disorder	7
1.1.2.3.	Recurrent Major Depression	8
1.2.	**Environmental Aetiology**	**8**
1.3.	**Biological Aetiology**	**10**
1.3.1.	Neuropathology	10
1.3.2.	Theories of Aetiology	10
1.3.2.1.	The Pathogenesis of Mood Disorders	11
1.3.2.2.	The Pathogenesis of Schizophrenia	13
1.4.	**Genetic Aetiology**	**14**
1.4.1.	Family Studies	15
1.4.2.	Twin Studies	15
1.4.3.	Adoption Studies	16
1.4.4.	Mode of Transmission	17
1.5.	**Molecular Genetics**	**17**
1.5.1.	Linkage Analysis	17
1.5.1.1.	Methodology	17
1.5.1.2.	Methodological Considerations	19
1.5.1.3.	Findings	20
1.5.1.4.	Linkage Overlap in Bipolar Affective Disorder and Schizophrenia	22
1.5.2.	Association Analysis	24
1.5.2.1.	Methodology	24
1.5.2.2.	The Power of Association Analysis	27
1.5.2.3.	Linkage Disequilibrium and its Role in Association	30
1.5.2.4.	Findings of Association Analysis	31
1.5.3 .	Heterogeneity in Psychiatric Illness	35
1.5.4.	Characterising Linkage Disequilibrium in the Human Genome	37
1.5.5.	The Common Disease/Common Variant Hypothesis	40
1.6.	**The Human Genome Project**	**44**
1.6.1.	Why sequence the Human Genome?	44
1.6.2.	Characteristics of the Human Genome	44
1.6.3.	Annotating the Human Genome	47
1.7.	**The Chromosome 4 Linkage Families**	**48**
1.7.1.	The Families	48
1.7.2.	Disease Associated Haplotypes	50
1.8.	**Thesis Aims**	**53**

Chapter 2 Materials and Methods

2.1.	**Clinical Resources**	**55**
2.1.1.	Families	55
2.1.2.	Case and Control Samples	60
2.1.3.	DNA Panels	60
2.1.3.1.	Allele Sharing Panel	60
2.1.3.2.	Pools Plate	60
2.1.3.3.	Monochromosomal Hybrid Panel	62
2.2.	**Oligonucleotides**	**63**
2.2.1.	Oligonucleotide Design	63
2.2.2.	Oligonucleotide Synthesis	63
2.2.3.	Oligonucleotide Primer Sequences	63
2.3.	**Amplification of DNA by the Polymerase Chain Reaction (PCR)**	**64**
2.3.1.	PCR Reagents	64
2.3.2.	PCR Cycling	65
2.4.	**RT-PCR**	**65**
2.4.1.	RNA Extraction from Cells	65
2.4.2.	cDNA Synthesis and RT-PCR	66
2.5.	**cDNA Library Screening**	**66**
2.5.1.	cDNA Libraries	66
2.5.2.	cDNA Library Screening	70
2.5.2.1.	Pre-Screen	70
2.5.2.2.	Vectorette PCR	70
2.5.2.3.	Nested PCR	71
2.6.	**Agarose Electrophoresis**	**71**
2.6.1.	Solutions	71
2.6.2.	Size Markers	71
2.6.3.	Agarose Gel Electrophoresis	71
2.7.	**Purification and Concentration of DNA**	**72**
2.7.1.	PCR Purification	72
2.7.2.	Purification of DNA from Agarose Gels	72
2.7.2.1.	Spin Columns	73
2.7.2.2.	Filter Tip Purification	73
2.7.2.3.	Gel Purification Kit	73
2.7.3.	Ethanol Precipitation	73
2.8.	**Sequencing**	**74**
2.8.1.	Sequencing PCR Products	74
2.9.	**Genotyping**	**75**
2.9.1.	SNP Genotyping	75
2.9.1.1.	SNaPshot™	75
2.9.1.2.	TaqMan®	77
2.9.1.3.	Sequenom® MASSARRAY™	78
2.9.2.	Microsatellite Genotyping	78
2.10.	**Bacterial Cell Culture**	**79**
2.10.1.	Solutions	79
2.10.2.	Bacterial Cell Culture and Colony PCR	79

2.11. Bioinformatic Tools 80

Chapter 3 Contig Mapping in Minimal Region One

3.1. Introduction 83
3.2. Bioinformatic Analysis 89
3.2.1. Searching for Bioinformatic Evidence of Clone Overlap 89
3.2.2. The Genomic Landscape 90
3.3. Contig Mapping 92
3.3.1. Marker Design 92
3.3.1.1. RP11-751L19 94
3.3.1.2. RP11-180A12 94
3.3.1.3. Centromeric End of RP11-264E23 94
3.3.1.4. Telomeric End of RP11-264E23 95
3.3.2. Testing Marker Specificity 96
3.3.3. Colony PCR Results 104
3.4. Discussion 106

Chapter 4 Recombination Breakpoint Mapping of Minimal Region One

4.1. Introduction 111
4.2. F59 Recombination Breakpoint 115
4.2.1. Family Members 115
4.2.2. Markers 116
4.2.3. Genotyping 119
4.3. Definition of the F59 Recombination Breakpoint Interval 119
4.4. Phase 4: Further Refinement of the Recombination Breakpoint 124
4.4.1. Family Members 124
4.4.2. Markers 124
4.4.3. Genotyping 125
4.4.4. Results 125
4.5. F50 Recombination Breakpoint 127
4.5.1. Family Members 129
4.5.2. Markers 129
4.5.2.1. Microsatellite Markers 129
4.5.2.2. Single Nucleotide Polymorphism Markers 134
4.5.3. Genotyping 135
4.6. Definition of the F50 Recombination Breakpoint Interval 135
4.7. Investigating the homozygosity of F50-3 138
4.7.1. Assay to detect Chromosome Loss in F50-3 140
4.7.2. PCR Assay Limitations 144
4.8. Discussion 145

Chapter 5 Transcript Map of Two Candidate Regions for Psychiatric Illness on Human Chromosome 4p

5.1.	**Introduction**	**148**
5.2.	**Defining the Region**	**150**
5.2.1.	ACeDB	150
5.2.2.	Minimal Tiling Path	153
5.3.	**Sequence Composition of the Region**	**156**
5.4.	**Known Genes in MR1 and MR2**	**160**
5.4.1.	The Known Genes	160
5.4.2.	Evaluating the Known Genes	163
5.5.	**Identifying Novel Genes**	**165**
5.5.1.	cDNA Library Screening	165
5.5.1.1.	Primer Design	165
5.5.1.2.	cDNA Libraries	166
5.5.1.3.	STS Screening	166
5.5.1.4.	Vectorette Screening	166
5.5.1.5.	Vectorette PCR Problems	172
5.5.1.6.	Nested PCR Results	176
5.5.2.	RT-PCR	178
5.5.2.1.	Primer Design	178
5.5.2.2.	RT-PCR Results	181
5.6.	**Results from cDNA Library Screening and RT-PCR**	**183**
5.6.1.	Putative Genes without Evidence	183
5.6.2.	Novel Genes Identified	183
5.6.2.1.	74M11 Gene	183
5.6.2.2.	G-Protein-Coupled Receptor 125 (GPR125)	186
5.6.2.3.	LOC91050	189
5.6.2.4.	LOC202025	191
5.6.2.5.	LOC132895	193
5.7.	**Pseudogenes**	**194**
5.8.	**Discussion**	**199**

Chapter 6 Genetic Analysis of the Superoxide Dismutase 3 (SOD3) Gene

6.1.	**Introduction**	**203**
6.2.	**SNP Identification**	**208**
6.2.1.	Sample	208
6.2.2.	STS Design and SNP Detection	208
6.2.3.	SNPs Identified	209
6.3.	**SNP Analysis**	**211**
6.3.1.	Amino Acid Changes	211
6.3.2.	Allele Sharing	216
6.4.	**Association Study on Pooled DNA**	**222**
6.4.1.	Sample	222

6.4.2.	SNPs	222
6.4.3.	Genotyping	222
6.4.4.	Analysis	223
6.4.5.	Results	225
6.4.6.	Problems	226
6.5.	**Phase I Association Study on Individuals**	**227**
6.5.1.	Sample	227
6.5.2.	SNPs	227
6.5.3.	Genotyping and Analysis	227
6.5.4.	Results	228
6.6.	**Phase II Association Study on Individuals**	**232**
6.6.1.	Sample	232
6.6.2.	SNPs and Results	232
6.7.	**Discussion**	**233**

Chapter 7 Genetic Analysis of the G-Protein-Coupled Receptor (GPR78) Gene

7.1.	**Introduction**	**237**
7.2.	**SNP Identification**	**240**
7.2.1.	Sample	240
7.2.2.	STS Design and SNP Detection	240
7.2.3.	SNPs Identified	240
7.3.	**SNP Analysis**	**242**
7.3.1.	Amino Acid Changes	242
7.3.1.1.	SNP ih31	242
7.3.1.2.	SNP ih34	246
7.3.2.	Allele Sharing	248
7.4.	**Association Study on Pooled DNA**	**256**
7.4.1.	Sample	256
7.4.2.	SNPs	256
7.4.3.	Genotyping	259
7.4.4.	Results	259
7.4.5.	Problems	260
7.5.	**Phase I Association Study on Individuals**	**261**
7.5.1.	Sample	261
7.5.2.	SNPs	261
7.5.3.	Genotyping and Analysis	261
7.5.4.	Results	262
7.6.	**Phase II Association Study on Individuals**	**265**
7.6.1.	Sample	265
7.6.2.	SNPs and Genotyping	265
7.6.3.	Results	265
7.7.	**Discussion**	**266**

Chapter 8 Discussion

8.1.	**Summary**	**271**
8.2.	**Families**	**271**
8.3.	**Minimal Region One**	**272**
8.3.1.	Minimal Region One Contig	272
8.3.2.	Minimal Region One Recombination Breakpoints	274
8.4.	**Allele Sharing**	**276**
8.5.	**Identification of Genes**	**277**
8.6.	**Identification of Association**	**279**
8.7.	**Identifying the Causal Variant**	**281**
8.8.	**Future Work**	**283**
8.9.	**Finding Genes in Psychiatric Illness**	**284**
8.10.	**Conclusions**	**285**

References **286**

Appendix I Oligonucleotide Primer Sequences 308

Appendix II DNA Pool Peak Heights of SOD3 SNPs 321

Appendix III DNA Pool Peak Heights of GPR78 SNPs 324

Appendix IV Raw Data of the Phase I and Phase II Association Study 329

List of Figures

1-1	The diagnostic criteria for recurrent major depression	3
1-2	The diagnostic criteria for bipolar affective disorder I	4
1-3	The diagnostic criteria for schizophrenia	5
1-4	Overlapping haplotypes of families 22, 59, 50 and 48	52
2-1	Family 22 pedigree	56
2-2	Family 59 pedigree	57
2-3	Family 50 pedigree	58
2-4	Family 48 pedigree	59
2-5	DNA panels	61
3-1	Overlapping haplotypes of families 22, 59, 50 and 48	84
3-2	Contig assemblies of the human genome working draft sequence	87
3-3	Contig assembly of the human genome working draft sequence in ACeDB	88
3-4	The position of STS markers along the contig	92
3-5	Example of primer design around existing STS st26424.snp	93
3-6	Agarose gels showing MCHP PCR results	97
3-7	Agarose gels showing MCHP PCR results	98
3-8	Agarose gels showing MCHP PCR results	99
3-9	Agarose gels showing MCHP PCR results	100
3-10	Agarose gel showing MCHP PCR results	101
3-11	Agarose gels showing MCHP PCR results	103
3-12	The results of colony PCR for five markers	105
4-1	Overlapping marker haplotypes of families 22, 59, 50 and 48	114
4-2	Tiling path of nine BAC clones (RP11-) between recombinant and non-recombinant markers stba473m13.2 and st362C2-72J1	116
4-3	Fluorograms of microsatellite marker stD4S2906	120
4-4	Part of the sequence trace for STS st585128.snp	121
4-5	Four phases of genotyping	123
4-6	The marker haplotypes for 13 microsatellites and two SNPs in F59	126
4-7	The telomeric recombination breakpoint region of MR1 defined by F50	128
4-8	The marker haplotypes for six microsatellites and 35 SNPs in F50	137
4-9	The marker haplotypes for six microsatellites and 35 SNPs in F50 members, F50-1 and F50-2	139
4-10	Fluorograms for three chromosome 4p microsatellite markers genotyped on somatic cell hybrids of individual F50-3	141
4-11	Agarose gels showing amplification of three STSs from GPR78 and three STSs from CPZ	143
5-1	Example of an annotated known gene in ACeDB	152
5-2	Schematic of the G bands on chromosome 4p with respect to the position of Minimal Region One and Minimal Region Two	158

5-3	GC content in Minimal Region One and Minimal Region Two	159
5-4	Position of the seven known genes in Minimal Region One	161
5-5	Position of the 13 known genes in Minimal Region Two	162
5-6	Example of an STS screened on the cDNA library primary plate	167
5-7	Vectorette schematic	167
5-8	Two examples of non-specific amplification from the cDNA libraries	173
5-9	Results of the experiments to test the effect of experimenter error	174
5-10	Examples of the results obtained from a nested experiment	175
5-11	An example of a nested PCR using a 1:100 dilution of the vectorette PCR as a template	177
5-12	A screen shot from ACeDB showing several LOC genes	179
5-13	Gene 74M11	185
5-14	GPR125	188
5-15	LOC91050	190
5-16	LOC202025	192
5-17	LOC132895	193
5-18	A putative pseudogene in ACeDB	196
5-19	Pseudogene S27	198
6-1	The structure of the SOD3 gene and peptide	207
6-2	SNP identification in the SOD3 gene	210
6-3	The results of a PIX analysis (MRC-RFCGR) for part of the SOD3 protein	213
6-4	Genotyping results from the Allele Sharing DNA panel	218
6-5	Examples of the fluorograms for two DNA pools for SNP rs2536512	224
7-1	The structure of the GPR78 gene and peptide	239
7-2	SNP identification in the GPR78 gene	241
7-3	The results of a PIX analysis (MRC-RFCGR) for part of the GPR78 protein containing SNP ih31	243
7-4	The results of a PIX analysis (MRC-RFCGR) for part of the GPR78 protein containing SNP ih34	247
7-5	Genotyping results from the Allele Sharing DNA panel	250
7-6	Linkage disequilibrium calculations of 18 SNPs in the GPR78 gene from 11 chromosomes	257

List of Tables

2-1	Monochromosomal somatic cell hybrid DNA panel	62
2-2	cDNA libraries	68
2-3	ABI PRISM® 3730 and 3100 DNA sequencer sample dilutions	79
3-1	Clone names and sequence accession number	88
4-1	Details of the markers used to genotype the F59 recombination breakpoint	118
4-2	Details of the markers used to genotype the F50 recombination breakpoint	132
4-3	The results of genotyping 24 F50-3 hybrids for three microsatellite markers	140
4-4	The peak heights of the 'contaminating allele' from the F50-3 hybrid genotyping gel	144
5-1	List of programmes used to annotate ACeDB	151
5-2	The minimal tiling path of Minimal Region One	154
5-3	The minimal tiling path of Minimal Region Two	155
5-4	Known genes in Minimal Region One	161
5-5	Known genes in Minimal Region Two	162
5-6	Characterisation of the known genes in Minimal Region One and Minimal Region Two	164
5-7	cDNA library screening results	169
5-8	Nested vectorette PCR results	177
5-9	The classification of LOCs in Minimal Region One and Minimal Region Two	181
5-10	RT-PCR results	182
5-11	Troubleshooting RT-PCR results	183
5-12	Pseudogenes in Minimal Region One and Minimal Region Two	195
6-1	Results of a PROSITE scan of the SOD3 protein	214
6-2	The prediction of the phosphorylation status of the SOD3 protein	215
6-3	Haplotypes identified from the Allele Sharing DNA pane	220
6-4	The results of a Fishers exact test on each of the eight SNPs in the SOD3 gene	221
6-5	The results of a chi-square test for association on pooled DNA	226
6-6	The results of the Phase I association study	231
7-1	Results of a PROSITE scan of the GPR78 protein	244
7-2	The prediction of the phosphorylation status of the GPR78 protein	245
7-3	Haplotypes identified from the Allele Sharing DNA panel	250
7-4	The results of a Fishers exact test on each of the 21 SNPs in the GPR78 gene	254
7-5	The results of a chi-square test for association on pooled DNA	260
7-6	The results of the Phase I association study	263
7-7	The results of the Phase II association study	265

Abbreviations and Symbols

µg	Microgram
µl	Microlitre
µM	Micromolar
Ach	Acetylcholine
ACeDB	A *C.elegans* database
ACTH	Adrenocorticotropic hormone
AM	Adrenomedullin
AMP	Ampicilin
AMPT	alpha-methyl-para-tyrosine
ANAPC4	Anaphase promoting complex 4
AVP	Arginine-vasopressin
BAC	Bacterial artificial chromosome
BDNF	Brain derived neurotrophic factor
bp	Base pair
BLAST	Basic local alignment search tool
BPAD	Bipolar affective disorder
BSA	Bovine serum albumin
cAMP	cyclic adenosine monophosphate
CCKAR	Cholecystokinin A receptor
CDCV	Common disease common variant hypothesis
cDNA	Cloned deoxyribose nucleic acid
CEPH	Centre du Etude Polymorphisme Humain
CEN	Centromere
COMT	Catechol-O-methyltransferase
CPZ	Carboxypeptidase Z
CRH	Corticotrophin releasing hormone
CTL	Control
DA	Dopamine
dbSNP	SNP database
DMSO	Dimethyl sulfoxide
DNA	Deoxyribose nucleic acid
dNTP	Deoxynucleotide
ddNTP	Dideoxynucleotide
DDX15	DEAH (Asp-Glu-Ala-His) box polypeptide 15
DRD5	Dopamine receptor D_5
DSM	American diagnostic and statistical manual
DZ	Dizygotic
ECM	Extra cellular matrix
EDTA	Ethylenediamine tetra-acetic acid
EPS	Extra pyramidal side effects
ERP	Event related potential
EST	Expressed sequence tag
EXO	Exonuclease 1
FISH	Fluorescence insitu hybridisation
g	Gravity
GABA	Gamma amino butyric acid

GBA3	Glucosidase, beta, acid 3
GPR78	G protein coupled receptor 78
GPR125	G protein coupled receptor 125
GR	Glucocorticoid receptor
GSK-3	Glycogen synthase kinase 3
HGP	Human genome sequencing project
HPA	Hypothalamic pituitary adrenal axis
HS3ST1	Heparan sulphate (glucosamine) 3-O-sulfotransferase 1
5HT	Serotonin
ICD	International classification of diseases
IHGSC	International human genome sequencing consortium
IP_3	Inositol (1,4,5) tri-phosphate
kb	Kilobase
l	Litre
LC	Locus coeruleous
LD	Linkage disequilibrium
Li	Lithium
LINE	Long interspersed elements
LG12	Leucine-rich repeat LGI family, member 2
LOD	Logarithm of odds
M	Molar
MALDI-TOF	Matrix assisted laser desorption/ionisation-time of flight
Mb	Megabase
MCHP	Monochromosome hybrid panel
MIST	Mast cell immunoreceptor signal transducer
ml	Millilitre
MAO	Monoamine oxidase
mol	Mole
MR	Mineralocorticoid receptor
MR1	Minimal region one
MR2	Minimal region two
MRC-RFGRC	Medical Research Council-Rosalind Franklin Genome Research Centre
ms	Millisecond
mRNA	Messenger ribonucleic acid
MZ	Monozygotic
NA	Noradrenalin
NCBI	National Centre for Biotechnology Information
ncRNA	Non-coding ribonucleic acid
ng	Nanogram
NMDA	N-methyl-D-aspartate
NO	Nitric oxide
OR	Odds ratio
ORF	Open reading frame
PAC	P1 artificial chromosome
PCP	Phencyclidine
PCR	Polymerase chain reaction
PI4K2B	Phosphatidylinositol 4-kinase type-II beta

PKC	Protein kinase C
PPARGC1	Peroxisome proliferative activated receptor, gamma, coactivator 1
5'RACE	Rapid amplification of cDNA ends
RBPSUH	Recombining binding protein repressor of hairless (drosophila)
REM	Rapid eye movement
RF	Recombination fraction
RNPS1	RNA binding protein S1
RMD	Recurrent major depression
RR	Relative risk
RT-PCR	Reverse-transcriptase polymerase chain reaction
S27	40S ribosomal protein S27
SAD	Seasonal affective disorder
SAM	System for Assembling Markers
SAP	Shrimp alkaline phosphatase
SCZ	Schizophrenia
SELDI-TOF	surface enhanced laser desorption/ionisation-time of flight
SEM	Standard error of the mean
SINE	Short interspersed elements
SLA/LP	Soluble liver antigen/liver pancreas antigen
SLC2A9	Solute carrier family 2 (facilitated glucose transporter) member 9
SLC34A2	Solute carrier family 34 (sodium phosphate) member 2
SNP	Single nucleotide polymorphism
SSRI	Serotonin selective reuptake inhibitor
STS	Sequence tagged site
TAE	Tris-acetate EDTA
TBE	Tris-borate EDTA
TDT	Transmission disequilibrium test
TEL	Telomere
TET	Tetracycline
TIGR	The Institute for Genomic Research
Tm	Melting temperature
TMD	Transmembrane domain
UCSC	University of California Santa Cruz
USP17	Ubiquitin specific protease activity 17
UTR	Untranslated region
UV	Ultra violet
V	Volts
VFCS	Velofacial cardio syndrome
VPA	Valproate
WDR1	WD repeat domain 1
WGA	Whole genome amplification
ZCCHC4	Zinc finger, CCHC domain containing, 4

Chapter One

Introduction

Introduction

1.1. Psychiatric Illness: The Clinical Picture

1.1.1. Phenotypes and Diagnosis

Attempts to distinguish and classify psychosis and disorders of mood extend back several thousand years to Hippocrates (ca 460-377 B. C.). However, it wasn't until the late nineteenth century and the early twentieth century that the discipline of the diagnosis of mental illness established itself. Emil Kraepelin, regarded as the founder of the discipline, distinguished four different sub-types of 'dementia praecox', or schizophrenia, from 'manic–depressive insanity', which grouped together single episode depression, recurrent major depression (RMD) and bipolar affective disorder (BPAD) (Shean, 2004). Classification systems have changed and evolved over the years and are still open to debate even today.

Today, the most commonly used diagnostic criteria are the Diagnostic and Statistical Manual of Mental Disorders (DSM) (American Psychiatric Association, 2000), the Research Diagnostic Criteria (RDC) (Spitzer *et al*, 1978) and the Classification of Mental and Behavioural Disorders (ICD) (World Health Organisation, 1992). DSM is currently revising its 4^{th} edition (DSM-IV-RT), and ICD has published its 10^{th}. The reliability of the current diagnostic systems is open to debate since diagnosis can vary between psychiatrists, countries and measures used (Janca, 2001; Bertelsen, 2002; Benazzi, 2003_b). Figures 1-1, 1-2 and 1-3 detail the core diagnostic criteria used to diagnose RMD, BPAD and schizophrenia using the DSM-IV-RT manual.

Bipolar affective disorder is characterised by episodes of mania and depression. The pervading effect is on mood, not cognition. However, the manic phase of BPAD can disturb cognitive processes such as decision making and reasoning. DSM-IV-RT (American Psychiatric Association, 2000) classifies BPAD into four categories. Patients with BPAD I disorder have had a full manic episode and usually major

depression, where as patients with BPAD II disorder have a milder form of mania with episodes of major depression. A mixed disorder consists of symptoms that meet the criteria for both a manic episode and a major depressive episode nearly every day. Finally, cyclothymic disorder is distinguished by the absence of depression, but bipolar I or II will eventually develop in 15-30 % of patients (Manning *et al*, 1998).

The life time population prevalence of BPAD is 1.2% (Weissman, 1991), although higher rates of ~4% have been observed (Angst, 1995). The age of onset for BPAD is generally ten years earlier than RMD, usually between adolescence and 30 years (Angst and Sellaro, 2000). Episodes of un-medicated mania and depression have been estimated to last from six months to a year and the median cycling is 18 months. Residual symptoms are common between cycles and the proportions of mania and depression remain stable into old age (Angst and Sellaro, 2000).

DSM-IV-RT Diagnostic criteria for major recurrent depressive disorder

A. Presence of two or more major depressive episodes.
Note: To be considered separate episodes, there must be an interval of at least 2 consecutive months in which criteria are not met for a major depressive episode.

B. The major depressive episodes are not better accounted for by schizoaffective disorder and are not superimposed on schizophrenia, schizophreniform disorder, delusional disorder, or psychotic disorder Not Otherwise Specified.

C. There has never been a manic episode, a mixed episode, or a hypomanic episode. Note: This exclusion does not apply if all of the manic-like, mixed-like, or hypomanic-like episodes are substance or treatment induced or are due to the direct physiological effects of a general medical condition.

Figure 1-1: The diagnostic criteria for recurrent major depression. Adapted from The Diagnostic and Statistical Manual of Mental Disorders, (American and Psychiatric Association, 2000, 4th edition, text revision). See Figure 1-2 for the symptoms of a major depressive episode.

DSM-IV-RT Diagnostic criteria for bipolar I disorder

At least one Manic or Mixed episode, but there may be Hypomania or Major Depression as well.

Manic episode:

A. A distinct period of abnormally and persistently elevated, expansive, or irritable mood, lasting at least 1 week.

B. Three (or more) of the following symptoms have persisted:

(1) inflated self-esteem or grandiosity

(2) decreased need for sleep (e.g., feels rested after only 3 hours of sleep)

(3) more talkative than usual or pressure to keep talking

(4) flight of ideas or subjective experience that thoughts are racing

(5) distractibility (i.e., attention easily drawn to unimportant or irrelevant stimuli)

(6) increase in goal-directed activity or psychomotor agitation

(7) excessive involvement in pleasurable activities that have a high potential for painful consequences (e.g. buying sprees, sexual indiscretions, foolish investments)

Major depressive episode:

A. Five (or more) of the following symptoms have been present for 2-weeks and at least one of the symptoms is either (1) or (2).

(1) Depressed mood most of the day, nearly every day, as indicated by either subjective report (e.g., feels sad or empty).

(2) Markedly diminished interest or pleasure in all, or almost all, activities most of the day, nearly every day.

(3) Significant weight loss or weight gain, or decrease or increase in appetite nearly every day.

(4) Insomnia or hypersomnia nearly every day.

(5) Psychomotor agitation or retardation nearly every day.

(6) Fatigue or loss of energy nearly every day.

(7) Feelings of worthlessness or excessive or inappropriate guilt nearly every day.

(8) Diminished ability to think or concentrate, or indecisiveness, nearly every day.

(9) Recurrent thoughts of death, recurrent suicidal ideation without a specific plan, or a suicide attempt or a specific plan for committing suicide

Figure 1-2: The diagnostic criteria for bipolar affective disorder I. Adapted from The Diagnostic and Statistical Manual of Mental Disorders, (American and Psychiatric Association, 2000, 4th edition, text revision).

DSM-IV-RT diagnostic criteria for schizophrenia

A. Characteristic symptoms:

Two (or more) of the following. Only one Criterion A symptom is required if delusions are bizarre or hallucinations consist of a voice keeping up a running commentary on the person's behaviour or thoughts, or two or more voices conversing with each other.

(1) delusions

(2) hallucinations

(3) disorganized speech (e.g., frequent derailment or incoherence)

(4) grossly disorganized or catatonic behaviour

(5) negative symptoms, i.e., affective flattening, alogia, or avolition

B. Social/occupational dysfunction:

For a significant portion of the time since the onset of the disturbance, one or more major areas of functioning such as work, interpersonal relations, or self-care are markedly below the level achieved prior to the onset (or when the onset is in childhood or adolescence, failure to achieve expected level of interpersonal, academic, or occupational achievement).

C. Duration:

Continuous signs of the disturbance persist for at least 6 months. This 6-month period must include at least 1 month of symptoms (or less if successfully treated) that meet Criterion A (i.e., active-phase symptoms) and may include periods of prodromal or residual symptoms.

D. Schizoaffective and mood disorder exclusion:

Schizoaffective Disorder and Mood Disorder With Psychotic Features have been ruled out because either (1) no major depressive, manic, or mixed Episodes have occurred concurrently with the active-phase symptoms; or (2) if mood episodes have occurred during active-phase symptoms, their total duration has been brief .

E. Substance/general medical condition exclusion:

The disturbance is not due to the direct physiological effects of a substance (e.g., a drug of abuse, a medication) or a general medical condition.

F. Relationship to a pervasive developmental disorder:

If there is a history of autistic disorder or another Pervasive Developmental Disorder, the additional diagnosis of Schizophrenia is made only if prominent delusions or hallucinations are also present for at least a month.

Figure 1-3: The diagnostic criteria for schizophrenia. Adapted from The Diagnostic and Statistical Manual of Mental Disorders, (American and Psychiatric Association, 2000, 4th edition, text revision).

Recurrent major depression is the most common of the psychiatric illnesses. Prevalence ratings have been estimated from 7.8% (Weissman, 1991), to ~17% (Angst, 1995). In DSM-IV-RT (American Psychiatric Association, 2000), RMD is diagnosed after at least two separate episodes and by the absence of any manic or schizophrenic features. Core symptoms of RMD include a depressed mood for most of the day, nearly every day, and a marked loss of interest or pleasure in all or almost all activities. Cognitive deficits are due to a lack of energy and/or motivation, or the toxic effects of stress hormones rather than an active and specific defect.

Schizophrenia has a lifetime risk of ~1% in the population, and an average age of onset between early adolescence to 35 years (Frangou and Murray, 1997). In DSM-IV-RT (American Psychiatric Association, 2000) symptoms of schizophrenia are positive or negative. Positive symptoms include delusions, hallucinations, disorganized speech, and disorganized or catatonic behaviour. Negative symptoms include affective flattening, alogia (a lack of fluency and productivity of thought and speech) and avolition (absence of initiative or motivation to begin or maintain a behaviour). Symptoms must seriously interfere with day to day functioning and be present for at least six months. A well characterised set of cognitive impairments are also observed, involving verbal memory and learning and attention-vigilance (Saykin *et al*, 1994). There are a number of other disorders classified under the schizophrenia spectrum, including schizoaffective disorder. This has a phenotype that is a mix of BPAD and schizophrenia. It is characterised by a period of illness which corresponds to a major depressive episode, a manic episode, or both, with concurrent symptoms that meet the criteria for schizophrenia (American Psychiatric Association, 2000).

1.1.2. Pharmacological Treatments

1.1.2.1. Schizophrenia

Traditional treatments for schizophrenia are pharmacologically simple, targeting only one or two brain neurotransmitter systems. Typical antipsychotics (e.g. reserpine)

have an affinity for dopamine (DA) D_2 receptors, reducing DA release. They relieve positive symptoms, but have little effect on cognitive impairments and negative symptoms (Thaker and Carpenter, 2001). The atypical antipsychotics (e.g. clozapine) block DA D_2 receptors, but also block DA D_1 and D_4 receptors and seretonergic (5HT) receptors, whilst newer atypical antipsychotics (e.g. olanzapine) also have an affinity for acetylcholine (ACh) receptors (Frangou and Murray, 1997). Drugs that enhance glutamatergic transmission via the N-methyl-D-aspartate (NMDA) receptor have had some success at reducing negative symptoms (Thaker and Carpenter, 2001). Thus, symptoms can be relieved by targeting several brain neurotransmitter systems.

1.1.2.2. Bipolar Affective Disorder

The two commonest pharmacological treatments for BPAD are lithium (Li) and the anticonvulsant valproate (VPA). These stabilise mood by reducing the incidence of severe depression and mania. A variety of other medications, for example antipsychotics and antidepressants, can be used and multiple medications are common (Lloyd *et al,* 2003). Few studies directly address the efficacy of antidepressants since it is often assumed that the results of studies of RMD are generalizable (Thase and Sachs, 2000).

The effects of Li and VPA in the brain are diverse. For example, Li alters the activity of DA, 5HT, ACh, noradrenalin (NA), γ-aminobutyric acid (GABA), glutamate and various neuropeptides (Lenox and Hahn, 2000). Such diverse effects could be mediated by an effect on cellular signalling cascades, for example, neuronal inositol (1,4,5) tri-phosphate (IP_3) (Berridge *et al*, 1989; Williams *et al*, 2002), protein kinase C (PKC) and glycogen synthase kinase-3 (GSK-3) (Coyle and Manji, 2002). The PKC and GSK-3 signalling cascades regulate gene transcription of the immediate early genes, DNA binding transcription factors that in turn regulate a wide range of gene transcription (Ikonomov and Manji, 1999). A bimodal mechanism of action has been proposed where Li raises baseline cellular activity and attenuates peak activity, thereby reducing extremes and stabilising fluctuation (Jope, 1999_a; Jope, 1999_b).

1.1.2.3. Recurrent Major Depression

Most efficacious drugs for depression affect brain neurotransmitter systems, by either reducing their catabolism (e.g. monoamine oxidase (MAO) inhibitors) or blocking synaptic reuptake (e.g. trycyclic antidepressants). Most trycyclic antidepressants (e.g. imipramine) act upon both NA and 5HT, but more selective antidepressants (e.g. the serotonin selective reuptake inhibitor fluoxetine) are safer and easier to use (Skolnick *et al*, 2003). Broad spectrum antidepressants (e.g.DOV 216,303) inhibit the reuptake of DA in addition to NA and 5HT and single dopamine reuptake inhibitors are also antidepressant (Skolnick *et al*, 2003).

1.2. Environmental Aetiology

A number of environmental risk factors have been identified for schizophrenia and affective disorders, such as maternal malnutrition, maternal viral infection, foetal hypoxia and winter births (Bromet and Fennig, 1999; Moller, 2003). However, it is the pathological role of stress that has received considerable attention.

It is widely believed that a biological or genetic predisposition for psychiatric illness is pushed past a resistance threshold by environmental factors (McGue *et al*, 1985). The Holmes and Rahe Social Readjustment Rating Scale (Holmes and Rahe, 1967) rates common life events, such as the death of a loved one, for stressfulness. Findings suggest that, compared to controls, symptom onset in psychiatric illness is preceded by an increased number of life events (Brown and Birley, 1968; Fowles, 1992; Norman and Malla, 1993). However, no specifically psychiatric environmental factors can be identified (Walker and Diforio, 1997), and some of the increase in stressful life events may be bought on by illness itself (Frangou and Murray, 1997). Therefore, the precipitate role of environmental stress is unclear.

The hypothalamic-pituitary-adrenal (HPA) axis is the primary system involved in the regulation of the physiological stress response. In response to stress, corticotrophin

releasing hormone (CRH) and arginine-vasopressin (AVP) are released into the circulation from the hypothalamus. This causes the release of adrenocortico trophic hormone (ACTH) from the anterior pituitary which in turn promotes the secretion of cortisol from the adrenal cortex. Cortisol has a catabolic action on a variety of physiological systems. Negative feedback via glucocorticoid receptors (GR) and mineralocorticoid receptors (MR) in the hippocampus ensures that cortisol secretion occurs within a narrow time window of immediate need (Herman *et al*, 1996).

Animal studies provide evidence for the role of the HPA axis in the response to environmental stress, and also its role in the manifestation of stress related behaviour. In rats, stress results in significant increases in corticosterone (the cortisol equivalent in rats) (Lopez *et al*, 1998). Prenatal stress leads to an increased baseline corticosterone level and greater, prolonged and non-habituating corticosterone levels in response to stress in adult rats and monkeys (Weinstock, 1997). Adult prenatally stressed rats have fewer hippocampal GR and MR receptors compared to controls (Weinstock, 1997). Behavioural abnormalities have been observed in rats and monkeys that have been exposed to various forms of prenatal stress. These include reduced play and social interaction (Clarke and Schneider, 1993), attentional deficits (Schneider, 1992) and an increase in defecation and a reduction in ambulation (Thompson, 1957; Wakshlak and Weinstock, 1990). An injection of CRH results in many of these behavioural measures, providing a link between the response of the HPA axis to stress and the behavioural consequences of stress (Weinstock, 1997).

Increases in baseline urinary cortisol, CRH and ACTH levels have been observed in BPAD, RMD and schizophrenic patients (Thakore, 1998; Altamura *et al*, 1999; Pariante and Miller, 2001). In addition, depressed patients have been shown to exhibit enlarged pituitary and adrenal glands and an enhanced response to ACTH (Pariante and Miller, 2001). Elevation of CRH may be due to impaired negative feedback at GR since repeated stress results in a failure of negative feedback in the brain and antidepressants enhance GR function (Pariante and Miller, 2001).

In line with animal studies, maternal stress during pregnancy leads to delays in motor development and behavioural abnormalities in children such as clinging, crying, hyperactivity, unsociable behaviour (Stott, 1973; Meier, 1985) and a higher incidence of attention deficit hyperactivity disorder (Clements, 1992). Mothers suffering bereavement of spouses during pregnancy have children with a higher incidence of psychiatric disorders (Huttenen and Niskanen, 1978).

1.3. Biological Aetiology

1.3.1. Neuropathology

The lack of available living neural tissue from patients has hindered the elucidation of pathological mechanisms in psychiatric illness. Post-mortem studies represent one way to analyse neural tissue. However, sample preservation, post-mortem cellular changes and medication effects can complicate the interpretation of results.

An often reported finding from post-mortem studies of mood disorders is the loss of glial cells (Harrison, 2002) and increased activity of the cAMP and phosphoinositol signalling pathways (Vawter *et al*, 2000). Differences between BPAD and RMD patients are also observed. For example, the basal ganglia is smaller in RMD patients and BPAD patients show increased neuronal number in the locus coeruleous (LC) (Baumann and Bogerts, 2001). The most common pathology in schizophrenia appears to be enlarged ventricular volume and a reduction in grey matter, accounted for by changes in myelination rather than cell loss (Harrison, 1999; Halliday, 2001). However, similar changes are also observed in patients suffering from alcohol abuse and schizophrenics are more likely to abuse alcohol than controls (Halliday, 2001).

1.3.2. Theories of Aetiology

Theories of the aetiology abound, virtually every neurotransmitter has been studied and much attention has been devoted to receptors and other components of

neurotransmitter pathways. However, a full understanding of the pathological mechanisms of psychiatric illness remains elusive.

1.3.2.1. The Pathogenesis of Mood Disorders

The role of NA in mood disorders is well established. NA containing cell bodies in the LC in the brainstem receive input from other brainstem nuclei and project to most areas of the cortex, subcortex and the spinal chord (Cooper *et al*, 1996). Drugs that cause a decrease in NA (e.g. AMPT) cause depression, whilst elevated NA can cause manic symptoms (Ressler and Nemeroff, 1999). The NA transporter is inhibited by many antidepressants, thus increasing synaptic NA concentrations. However, a complimentary decrease in the β-adrenergic post-synaptic receptor is observed after chronic antidepressant treatment. Thus, the net effect of this is unclear (Ressler and Nemeroff, 1999). Some studies have shown a decrease in cell number in the LC in suicide victims (Arango *et al*, 1996), and chronic stress in rats decreases axonal density in the LC that is reversible by antidepressants (Kitayama *et al*, 1997). Thus, although dysregulation of NA systems occurs in depression, a simple increase or decrease of NA is not sufficient to explain findings (Ressler and Nemeroff, 1999).

Increasing evidence points to the role of glutamate and GABA. Several studies report altered glutamatergic function in depressed patients and the antidepressant activities of antagonists of the glutamate receptor NMDA (Sancora *et al*, 2003). Both animals and humans show decreased GABAergic transmission in response to stress (Petty, 1995), and the mood stabiliser VPA is a GABA agonist (Petty, 1995). Furthermore, the blockade of $GABA_A$ receptors results in learned helplessness, an animal model of depression, in non-stressed rats (Sancora *et al*, 2003).

Astrocytes are the main source of glutamate and GABA synthesis, taking up glutamate from the synapse to synthesise glutamine. Glutamine is then used to re-synthesise glutamate and synthesise GABA. Impaired glial function, as inferred from post-mortem studies, would result in decreased glutamate uptake and increased extra-

synaptic levels. Thus, NMDA antagonists may exert an anti-depressive action by blocking the increased activation that increased levels of extra-synaptic glutamate would cause. However, increased levels of extra-synaptic glutamate might decrease presynaptic glutamate release via negative feedback. A decrease in glutamate synthesis in astrocyctes and decreased presynaptic glutamate release would also decrease the amount of glutamine available for GABA synthesis (Sancora *et al*, 2003). As with NA, a simple explanation of altered glutamate and GABA functioning in mood disorders is not forthcoming.

Other neurotransmitters are implicated in the pathogenesis of mood disorders. For example, 5HT neurons in the raphe nucleus innervate hypothalamic areas and stimulate the release of ACTH and cortisol. Hyper-reactivity of 5HT and the related HPA axis pathways has been observed in RMD and BPAD (Pariante and Miller, 2001). Nitric oxide (NO), a soluble gas that can act as an atypical neurotransmitter, has been implicated in numerous behaviours including depression (Papageorgiou *et al*, 2001; Suzuki *et al*, 2001) and the regulation of the HPA axis (Bernstein *et al*, 1998; Shan and Krukoff, 2001).

It is clear that a wide variety of neurochemical changes are involved in the pathogenesis of mood disorders. Consequently, there is increasing interest in fundamental molecular changes. For example, there is some interest in the role of circadian rhythms in the pathogenesis of mood disorders (Sher, 2003). Circadian rhythms are biological rhythms (e.g. the sleep/wake cycle and hormone secretion) with a period close to 24 hours that are regulated by cells in the hypothalamus. Abnormalities in the sleep/wake cycle are observed in winter seasonal affective disorder (SAD). Patients show increased winter nocturnal melatonin secretion compared to controls and propanol, that shortens this secretion, alleviates symptoms (Schlager, 1994). Light treatment, increasing the light part of the day-night cycle, also alleviates symptoms (Lewy *et al*, 1982). Disruption also exists in non-seasonal mood disorders (Wehr *et al*, 1983; Reimann *et al*, 2001). For example, blunting of circadian rhythm amplitude, advanced phase and doubling of the sleep-wake cycle

are observed in BPAD patients (Reimann *et al,* 2001). The cyclic pattern of relapse may be linked to circadian disruption (Wehr *et al*, 1983), and evidence suggests that sleep alterations precede the onset of depression and mania (Sher, 2003).

Recent attention has also focused on possible common cellular signalling pathways underlying mood disorders, such as impairments in neuroplasticity and cellular resilience (Duman *et al*, 1997; Manji *et al,* 2000). Numerous neuroplastic events could be involved, such as alterations in dendritic function, synaptic remodelling, long term potentiation, axonal sprouting, neurite extension, synaptogenesis and neurogenesis (Manji and Lenox, 2000). The cell loss and/or atrophy observed post-mortem might indicate neuroplastic impairment. A possible mediator might be stress. Stress can cause cell atrophy and reduce neurogenesis in the hippocampus and reduce cellular resilience to insults such as glutamate excitotoxicity, possibly by the inhibition of neurotrophic factors (Manji *et al*, 2000). Lithium and VPA have been found to have significant neuroprotective properties after various cytotoxic insults *in vitro* and *in vivo* in rats, and Li, VPA and antidepressants increase neurogenesis in the hippocampus (Manji *et al*, 2000). Furthermore, Li or VPA treated patients exhibit less prefrontal cortex volume loss than untreated patients (Drevets *et al*, 1997).

1.3.2.2. The Pathogenesis of Schizophrenia

The hyper-dopaminergic hypothesis of schizophrenia has been widely studied. An excess of DA is thought to cause psychotic symptoms since DA agonists (e.g. cocaine and amphetamine) have marked psychotic effects (Thaker and Carpenter, 2001). Amphetamine causes excess synaptic DA release in schizophrenic patients compared to controls, and correlates with a worsening of psychotic symptoms (Breier *et al*, 1997). However, the theory does not explain the negative symptoms of schizophrenia. Chronic phencyclidine (PCP) administration, which reduces glutamate transmission at the NMDA receptor, induces psychotic symptoms, negative symptoms and cognitive impairment (Tamminga, 1998).

Other neurotransmitters with altered function in schizophrenia include 5HT, markers of the HPA axis (Pariante and Miller, 2001) and ACh (Araki *et al*, 2002). As with mood disorders, a wide range of systems appear to be involved and focus on molecular mechanisms will arguably be more informative. For example, there is some evidence for a decreased neuronal density in the prefrontal and occipital corticies and reduced brain volume in schizophrenic patients, possibly due to decreased neuronal branching and connectivity (Selemon and Goldman-Rakic, 1999).

There are three main hypotheses of schizophrenia: neurodegenerative, early developmental and progressive (De Haan and Bakker, 2004). The neurodegenerative hypothesis, a progressive loss of tissue, is generally unsupported since gliosis, an important marker of neurodegeneration, has not been confirmed in schizophrenic brains (De Haan and Bakker, 2004). The prenatal and early developmental hypothesis states that neuronal migration may be disturbed during pregnancy due to, for example, stress or viral infection. Evidence to support this includes abnormalities in the positioning of neurons in the cortex in schizophrenics, increases of schizophrenia after complications during pregnancy and the alterations observed in brain asymmetry and brain volume in schizophrenics and young individuals at risk (De Haan and Bakker, 2004). The long latency between damage and symptoms may be explained by the maldevelopment *in utero* of a late maturing system such as connectivity (Weinberger and Lipska, 1995). In early development, abnormal neuronal migration, cytoarchitecture and synapse formation occur in those predisposed to developing schizophrenia. Subsequently, the time at which schizophrenia develops in early adolescence is when pruning of neuronal connections reaches a significant phase (Lieberman *et al,* 1997).

1.4. Genetic Aetiology

The vast numbers of biological studies, some of which have been discussed, have failed to identify a unifying theory of the pathological mechanisms underlying

psychiatric illness. Genetics is a hypothesis free approach with perhaps more hope for identifying the genes and the cellular pathways involved.

1.4.1. Family Studies

Over the years, it has become widely accepted that there is a substantial genetic component in the liability to developing psychiatric illness. This was first indicated by an increased risk to relatives of probands with schizophrenia, BPAD or RMD. For example, Kendler and Diehl (1993) revealed an increased risk for schizophrenia (4.8%) compared to controls (0.5%) in a meta-analysis of seven studies. A review of 21 studies also suggested an increased risk of BPAD in the relatives of BPAD probands (Craddock and Jones, 1999). Fewer studies have looked specifically at RMD, but in five that did there was evidence for a genetic component (Sullivan *et al,* 2000). Only a couple of studies, for example Pope *et al* (1982), have failed to show a similar increased familial risk.

1.4.2. Twin Studies

Twins provide important evidence for a genetic component to psychiatric illness. By comparing monozygotic (MZ) and dizygotic (DZ) twins it is possible to distinguish the role of genetics from non-shared environment. In addition, incomplete concordance between MZ twins provides evidence for variable penetrance.

In a study of UK twin pairs, concordance rates for schizophrenia, schizoaffective disorder, mania and other related disorders were found to be between 10 and 44% for MZ twins and 0 and 11.6% for DZ twins (Cardno *et al,* 1999). Craddock and Jones (1999) reviewed six twin studies of BPAD. All showed an increased probandwise concordance rate in MZ compared to DZ twins and a pool of the data gave an MZ concordance of 50%. A meta-analysis of six studies of RMD identified concordance rates of 14-37% in DZ twins and 23-67% in MZ twins (Sullivan *et al,* 2000).

Concordance rates for schizophrenia and BPAD are often higher than for RMD, suggesting greater genetic heterogeneity in RMD. For example, a meta-analysis of all the major twin studies of BPAD and RMD revealed higher MZ twin concordance rates for BPAD (72%) than for RMD (40%), although concordance rates for BPAD and RMD were more similar in DZ twins (Allen, 1976). The variation of concordance with study design reveals further insight into the underlying genetics. Higher concordance rates observed with a broader diagnostic criterion provide evidence for variable penetrance. A narrow, *vs* a broad, diagnosis of schizophrenia revealed concordance rates of 36 and 56% in MZ twins and 18 and 26% in DZ twins respectively (Fischer *et al,* 1969). Bertelsen *et al* (1977) observed a concordance in MZ twins of 58% with a strict illness definition of BPAD and 84% with a broad definition, and a concordance of 17 and 35% respectively for DZ twins.

1.4.3. Adoption Studies

Family and twin studies cannot completely distinguish between concordance as a result of shared genetics or shared environment. Adoptee studies, where the genetic family is distinct from the environment shared with the adopted family, represent a key way of demonstrating a genetic component to psychiatric illness.

Studies have consistently shown that adopted children, whose biological mothers are diagnosed with schizophrenia, show an increased occurrence of schizophrenia or related disorders compared to controls (Kety *et al*, 1971; Kety, 1988; Tienari, 1991; Tienari *et al*, 1994). Mendlewicz and Rainer (1977) found that BPAD, RMD, schizoaffective disorder and cyclothymia in the biological parents of BPAD adoptees was in excess of that seen in the adoptive parents, and was the same as the rate of psychopathology in the parents of non-adopted BPAD individuals. Furthermore, the rate of psychopathology in the adoptive parents was the same as in the adoptive parents of normal offspring. Two studies using the modern diagnosis of BPAD showed a trend towards the increased risk of biological parents of BPAD probands

for affective disorder (Craddock and Jones, 1999). Adoptee studies of RMD also provide qualitative evidence for a genetic component to RMD (Sullivan *et al*, 2000).

1.4.4. Mode of Transmission

Family data shows that psychiatric illness is not inherited in a Mendelian manner, but the actual mode of inheritance is not immediately clear. Using large data sets, McGue *et al* (1985) assessed the pure genetic forms (i.e. no account taken of environmental factors) of a single major locus model (GSL) and a multiple genes model (MT) for the genetic transmission of schizophrenia. Only the MT model fitted the data, but it only narrowly achieved significance. Whilst their analysis rejects the idea that a single gene could account for all of the schizophrenia phenotype, it does not exclude the possibilities of a major gene contributing to the liability or polygenic bilinear inheritance (Mynett-Johnson and McKeon, 1996).

Genes can interact multiplicatively or additively. The best fit to morbid risk data is seen with a multiplicative model of 3-4 loci (Risch, 1990). Multiple additively interacting genes would predict a decrease in the risk to relatives by a factor of 0.5 for each degree of relationship. Family data does not fit this model and therefore suggests multiplicative interaction (Mynett-Johnson and McKeon, 1996). A combination of inheritance patterns seems likely, with major gene inheritance in some families, multiplicative polygenic inheritance in others and environmental phenocopies for some individuals (Mynett-Johnson and McKeon, 1996).

1.5. Molecular Genetics

1.5.1. Linkage Analysis

1.5.1.1. Methodology

Linkage analysis was designed to identify genes for monogenic disorders but it has been applied to complex disorders with some success. Linkage measures whether a

marker allele is inherited more often than would be expected by chance with a phenotype. The method relies on the principles of recombination. Alleles on different chromosomes are randomly assorted during meiosis and alleles on the same chromosome are assorted depending upon recombination whereby the chance of a recombination event separating the marker and disease allele increases with distance. The distance between two alleles is measured by the recombination fraction (RF): the number of recombinant genotypes per 100 meioses. RFs of <0.5 (50%) and >0.50 mean that two loci are on the same or different chromosomes respectively.

In families, the small numbers of progeny mean that the RF may not be reliable. Therefore, linkage measures the probability of obtaining a set of results in a family on the basis of independent assortment and a specific degree of linkage. The odds of these two probabilities (independent assortment versus linkage) are divided to calculate the ratio and the results are expressed as a logarithm (a Lod score). For example, in a family of two parents and six offspring, one parent and three offspring have disease X (non-diseased = x) and all are genotyped for a biallelic marker (M/m). Four offspring have the parental genotype (XM or xm) and two have the recombinant genotype (Xm or xM). The expected frequencies of parental and recombinant genotypes are calculated for a range of RFs. For example, for independent assortment (RF: 0.5), the probability of a parental genotype is 0.25 and the probability of a recombinant genotype is 0.25. Thus, the probability of four parental and two recombinant genotypes is 0.25 x 0.25 x 0.25 x 0.25 x 0.25 x 0.25 = 0.00024. For a RF of 0.30, the probability of a parental genotype is 0.35 and the probability of a recombinant genotype is 0.15. Thus, the probability of four parental and two recombinant genotypes is 0.35 x 0.35 x 0.35 x 0.35 x 0.15 x 0.15 = 0.00034. The odds ratio of these two probabilities is 0.00034/0.00024 = 1.42. This means that a RF of 0.30 is 1.42 times more likely than independent assortment. The Lod score is the logarithm of this ratio (0.15). The odds of obtaining the result are tested under different values of RF to see which value gives the highest probability. Since logarithms are exponents, Lod scores from different matings can be accumulated to support or not support a linkage value.

Odds for linkage of 1000:1 gives a Lod score of 3, the traditional threshold for a significant result. However, this does not correspond to a p-value of 0.001 (or 10^{-3}) since a Lod score is a ratio of two probabilities and a p-value is an absolute probability. A Lod of 3 means that the data is 10^3 fold more likely to occur under a hypothesis of linkage than under a null hypothesis of independent assortment. A p-value of 10^{-3} means that the probability of the observed data is 10^{-3} under the null hypothesis. In fact, a Lod of 2.1 equals a p-value of 10^{-3} and a Lod of 3 equals a p-value of 10^{-4} (Lander and Kruglyak, 1985). However, the corresponding p-value of a Lod score can vary depending on the linkage method used (Elston, 1998).

1.5.1.2. Methodological Considerations

Linkage analysis is a parametric method because it requires the strict definition of parameters such as mode of inheritance and therefore is best suited to monogenic disorders. Since psychiatric illnesses have uncertain inheritance patterns, incomplete penetrance and diagnostic ambiguity, they do not lend themselves easily to linkage (Mynett-Johnson and McKeon, 1996). However, there are a number of factors that enhance the power to detect susceptibility loci. For example, a number of phenotypes can be tested, bilinear inheritance can be ruled out and clinical diagnosis can be reviewed regularly to capture new and changed diagnoses (Mynett-Johnson and McKeon, 1996). There are, however, lower limits to the power to detect linkage. For example, allelic association identified the APO*E4 susceptibility allele for late onset Alzheimer's disease. Whilst the E4 allele was estimated to account for ~17% of the population variance, a linkage study of 32 pedigrees found only modest evidence for linkage to this locus (Mynett-Johnson and McKeon, 1996).

In the absence of an alternative strategy, linkage analysis has become a standard tool for identifying genomic susceptibility regions for psychiatric illness. The result has been vast numbers of suspected false positives and many genomic regions have been implicated. In order to make sense of the huge amount of data, Lander and Kruglyak

(1995) propose four levels of significance: A Lod of 2.2 for suggestive linkage, a Lod of 3.6 for significant linkage, a Lod of 5.4 for highly significant linkage and replication of an original result with a p-value of <0.01 for confirmed linkage.

To circumvent the problem of specifying the mode of inheritance, alternative model-free nonparametric linkage methods have been devised. In general, they rely on detecting the fact that affected subjects share marker alleles more often than would be expected by chance. Nonparametric methods differ in terms of the types of pedigree structure they are applied to, whether they rely on identity-by-descent (if parents are included) or identity-by-state (siblings only) information and the methods of analysis used (Elston, 1998). Results are analysed using the Chi-square statistic to determine the probability that alleles are shared greater than would be expected by chance, or a Lod score that, rather than being maximised over the RF, is maximised over the expected probability of allele sharing (0.25 versus 0.5) (Elston, 1998).

1.5.1.3. Findings

Linkage studies have implicated many genomic regions in the genetic susceptibility to schizophrenia, BPAD and RMD, but not all have been replicated. For example, the significant linkage of mood disorders to chromosome 11p reported by Egeland *et al* (1987) failed to be replicated in most subsequent studies (Tsuang *et al*, 2004). The regions of greatest interest are these which have been independently confirmed. As replication studies are conducted, several separate genomic regions are consistently emerging, such as those on 8p 13q and 22q for schizophrenia and 4p, 13q and 22q for BPAD (Berry *et al*, 2003; Tsuang, 2004). This provides evidence for substantial locus heterogeneity in psychiatric illness.

Some of the more consistent findings for schizophrenia include chromosomes 1, 6, 8, 13 and 22. Linkage evidence supported by a second type of genetic evidence, such as the chromosomal abnormalities observed on chromosomes 1 and 22 and the biological correlates on chromosome 15, are especially compelling (Berry *et al*, 2003).

Linkage has been observed to schizophrenia on chromosome 1, in the region of a chromosomal (1:11) translocation which is associated with schizophrenia in a large family (Berry *et al*, 2003). The translocation directly disrupts the DISC1 gene on chromosome 1 (Millar *et al*, 2000). The gene is expressed in multiple regions of the brain and is possibly involved in brain development and function (Millar *et al*, 2003). There are positive linkage results for both arms of chromosome 6, but most interest is focused on the 6p21-24 region, containing several candidate genes, including the spinocerebellar ataxia gene, the HLA region, NOTCH4 and dysbindin (Berry *et al*, 2003). The linkage to schizophrenia on chromosome 8p is especially compelling in the light of corresponding association results and haplotype analysis (Stefansson *et al*, 2002). This is discussed in more detail in Section 1.5.2.4. Linkage to Chromosome 15 is interesting since the α-7 nicotinic cholinergic receptor subunit gene is positioned here. This gene is implicated in the P50 sensory gating deficit, an attention deficit marked by an inability to filter out irrelevant stimuli, that is observed in schizophrenic patients (Freedman *et al,* 1983). Linkage of the P50 sensory gating deficit to markers within the α-7 nicotinic cholinergic receptor subunit gene has been replicated a number of times (Berry *et al*, 2003). Linkage to chromosome 22 has been identified by several studies in the region of the velo-cardio-facial syndrome (VCFS) deletion on 22q11 (Berry *et al*, 2003). Patients with VCFS show an elevated risk for schizophrenia and major affective disorders (Tsuang *et al*, 2004). COMT and proline dehydrogenase are two candidate genes in this region (Collier and Li, 2003).

Consistent evidence for linkage has been identified to BPAD on chromosomes 4, 12, 13, 18, 21 and 22. Significant linkage of BPAD and RMD to chromosome 4p, with a maximum Lod score of 4.8, was identified in a large Scottish pedigree (Blackwood *et al,* 1996), and independent confirmation was obtained from several studies (Asherson *et al*, 1998; Ewald *et al*, 1998; Detera-Wadleigh *et al*, 1999; Williams *et al*, 1999). These findings are discussed in greater detail in Section 1.7.1. Barden *et al* (1988; cited in Berrettini, 2000_b) identified significant linkage to BPAD on chromosome 12q, with a Lod score exceeding 8, in a sub-population of French origin from Quebec

(Berrettini, 2000_b). This locus has been confirmed by studies of Danish and American pedigrees (Detera-Wadleigh *et al*, 1999; Ewald *et al*, 1998). The confirmed susceptibility locus for schizophrenia on chromosome 13q has also been linked to BPAD in two independent reports (Berrettini, 2000_a; Berrettini, 2000_b). Linkage has been observed to both arms of chromosome 18 and the pericentromeric regions (Baron, 1997). Interestingly, the linkage observed for BPAD to the q arm of chromosome 18 by Stine *et al*, (1995) provides evidence for paternal inheritance (Baron, 1997; Berretinni, 1998). Linkage of BPAD to chromosome 21q22 has also been observed by several independent investigators (Baron, 1997; Aita *et al*, 1999). Interestingly, this is the same region as the Downs syndrome critical region and they may share common gene variants (Aita *et al*, 1999). The linkage to schizophrenia and VCFS on chromosome 22q, described above, has also been linked to BPAD by several independent investigators (Berrettini, 2000_a; Berrettini, 2000_b).

The range of chromosomal regions implicated in schizophrenia and BPAD from linkage studies may reflect both Type 1 (incorrect rejection of null hypothesis) and Type 2 (incorrect acceptance of null hypothesis) errors. Type 1 errors can arise from inappropriate study design, for example small sample sizes and incorrect disorder specification. Type 2 errors can arise for the same reason, but also as a result of locus heterogeneity (Tsuang *et al*, 2004).

1.5.1.4 Linkage Overlap in Bipolar Affective Disorder and Schizophrenia

Schizophrenia and BPAD share certain epidemiological characteristics such as age of onset, lifetime risk, worldwide distribution and risk of suicide (Berrettini, 2000_a). Both exhibit similar risk factors, including an excess of perinatal complications and winter births (Torrey, 1999). The clinical entity of schizoaffective disorder, an intermediate phenotype between BPAD and schizophrenia, adds support for a continuum theory between these illnesses (Moller, 2003). Benazzi (2003_a) found evidence to support the continuity between the symptoms of RMD and BPAD and

the psychotic symptoms more usually associated with schizophrenia can occur in both BPAD (Potash *et al*, 2001), and RMD (Carpenter *et al*, 1973).

The evidence for a continuum manifests itself in biological studies. Most neurochemical systems that are dysfunctional in schizophrenia, for example, brain catecholamine and HPA axis functioning, are also dysfunctional in BPAD (Moller, 2003). A common post-mortem characteristic of schizophrenia, the reduced activity of genes responsible for producing myelin, has recently been confirmed in BPAD (Tkachev *et al,* 2003). Furthermore, some medications for schizophrenia, such as the antipsychotic olanzipine, are approved as treatment for bipolar mania (Moller, 2003).

Schizophrenia and BPAD can occur together in single families (Detera-Wadleigh *et al*, 1999) and the risk for illness crosses diagnostic boundaries, although family studies cannot differentiate the role of shared environment from shared genetics. Relatives of schizophrenic probands have an increased risk of RMD and schizoaffective disorder (Gershon *et al*, 1988) whilst relatives of BPAD probands have been shown to have an increased risk of schizoaffective disorder (Maier *et al*, 1993) and RMD (Craddock and Jones, 1999). Concordance has also been observed between MZ twins for BPAD and RMD (Bertlesen *et al*, 1977; Allen *et al*, 1974).

On the basis of this evidence, it is not surprising that linkage studies highlight similar susceptibility regions. For example, the linkage to chromosome 4p in families segregating BPAD and RMD was supported by independent investigators in families segregating both schizophrenia (Detera-Wadleigh, 1999; Williams *et al*, 1999) and schizoaffective disorder (Asherson *et al*, 1998). The confirmed loci for BPAD on 18p11 and 22q11-13 have been observed in schizophrenia and the confirmed loci for schizophrenia on 13q32 has been observed in BPAD (Berrettini, 2000_a; Berrettini, 2000_b). Chromosome 10p is another region where positive linkage to both BPAD and schizophrenia has been observed (Berrettini, 2000_a).

Direct evidence for overlapping susceptibility loci comes from a meta-analysis of all published whole genome scans for BPAD and schizophrenia. The strongest evidence for linkage to BPAD lies on chromosomes 13q and 22q whilst the strongest evidence for linkage to schizophrenia lies on chromosomes 8p, 13q and 22q (Badner and Gershon, 2004). The emergence of common susceptibility loci might reflect the phenotypic similarities. For example, a subset of BPAD pedigrees with a high incidence of psychotic symptoms (hallucinations and delusions) showed the strongest linkage to chromosomes 13q and 22q, but no linkage was observed if this subset was included in a larger sample (Potash *et al*, 2003). The region of linkage reported to BPAD and schizophrenia on chromosome 22q11 includes several candidate genes, including COMT. Association between a COMT haplotype and schizophrenia has recently been replicated in BPAD. In addition, for both BPAD and schizophrenia, the haplotype relative risk was greater in women (Shifman, 2004).

Despite this, there are several genomic regions that appear to be unique. Evidence for linkage of BPAD to chromosome 12q has been identified, but there are no reports of linkage of schizophrenia to this region (Berrettini, 2000_a; Tsuang *et al*, 2004) and several studies report linkage of schizophrenia to chromosome 6p and 8p, but there are no reports of linkage of BPAD to these regions (Berrettini, 2000; Berry *et al*, 2003). In conclusion, the evidence for shared susceptibility is convincing, justifying the combination of linkage findings from BPAD and schizophrenia to help define susceptibility alleles. However, it cannot be assumed that linkage to one is applicable to the other without supporting evidence and allelic heterogeneity might underlie shared loci.

1.5.2. Association Analysis

1.5.2.1. Methodology

Since recombination events are relatively rare and occur once or rarely twice per chromosome per generation, studying one or even several families means that linkage

regions will be large, containing many candidate genes. A higher resolution mapping technique is the association study. This measures allele frequency in populations of unrelated individuals and therefore effectively utilises hundreds of recombination events. Consequently, alleles that are inherited with the phenotype will be much closer to the susceptibility allele. The method is non-parametric because no assumptions are made about the mode of inheritance.

Chi-squared contingency table tests can be used to test whether there is a significant difference in allele frequency between two groups. For example, values in a 2x2 contingency table could be the number of occurrences of allele 'M' *vs* allele 'm' in well *vs* unwell individuals. The null hypothesis is: no association between allele and mental health, and therefore the cells of the table would reflect the allele frequency of each allele observed in a group of controls. Each of the obtained values in the cells is compared to the values under the null hypothesis. The significance of this difference, accounting for type 1 errors (a false positive result when the null hypothesis is incorrectly rejected) and degrees of freedom (the number of cell entries required before the values of the rest are fixed, i.e. 1 degree of freedom in a 2x2 contingency table test), is then assessed. It is important to remember that the Chi-square test is a measure of association and says nothing about causality.

Results of association are often expressed as an odds ratio (OR) or a relative risk (RR), as in traditional epidemiological case-control studies. Strictly, RR is used in a prospective study and OR in a retrospective study. A prospective study measures whether two groups that differ in exposure to a risk factor do or do not develop the primary outcome (e.g. does having a parent with BPAD increase the risk of developing BPAD in the offspring?). A retrospective study used the primary outcome as the basis of grouping and the level of the risk factor is measured in each (e.g. at what level does a risk allele occur in BPAD and control individuals?). The RR measures the risk rate of developing the primary outcome in exposed or non-exposed individuals. The OR measures the odds of being exposed to the risk factor in cases or controls. Thus, case-control association studies are retrospective. However, it is

common for publications to refer to RR in retrospective studies because the values of OR and RR are approximately equal when the primary outcome is relatively rare.

The OR (discussed in relation to the calculation of Lod scores in Section 1.5.1.1) is obtained by calculating the ratio of the odds of being exposed in cases (i.e. the occurrence of the risk factor divided by non-occurrence of the risk factor) and the odds of being exposed in controls. The RR is obtained by calculating the ratio of the risk of developing the primary outcome when the risk factor is present (all individuals with the risk factor and the primary outcome divided by all individuals with the primary outcome) and the risk of developing the primary outcome when the risk factor is not present.

A RR of 2.00 can also be expressed as an increased risk of 100%. However, this obviously does not mean that individual with the risk factor has a 100% chance of developing the primary outcome, but that their risk is 100% higher. Relative risk refers to a population risk and not individual risk since the possession of a risk factor does not predict whether that individual will develop the primary outcome.

A number of different association study designs can be applied. The case-control design compares unrelated cases and unrelated controls from the same ethnic population. For some diseases it may be necessary to select controls on some criteria that may be associated with disease status, such as sex or age. Simplex family designs, the affected child with its parents, or affected sibling pairs, have been advocated to reduce the problem of population stratification that can be seen with the case-control design and the spurious associations that result from this. However, the family design can result in a loss of power. For example the transmission disequilibrium test (TDT), that analyses the inheritance of parental alleles to an affected offspring, is restricted to using only heterozygous parental genotypes. In addition, the control population in a case-control design can be tested for stratification using unlinked markers. Perhaps most importantly, a case-control design is easier to perform due to the ease of sample collection (Risch, 2000; Cardon

and Bell, 2001). A further advantage of the case control design is that individual DNA can be pooled together to compare allele frequency, significantly reducing the amount of genotyping (Breen *et al,* 1999; Le Hellard *et al,* 2002).

Association studies can be hypothesis driven, using candidate genes, or genome wide, using random markers. The application of genome wide association studies increased in popularity with the discovery of the commonality of the single nucleotide polymorphism (SNP) (Altshuler *et al*, 2000, Syvanen, 2001). Nevertheless, the limited availability of quick and cheap high throughput genotyping techniques, the unknown patterns of genome wide linkage disequilibrium (LD) and issues concerning which SNPs to choose and the question of whether common or rare variants operate in complex disease still makes the utilisation of SNPs and genome wide association studies problematic (Owen *et al*, 2000; Botstein and Risch, 2003).

1.5.2.2. The Power of Association Analysis

The power, i.e. the ability to detect what is being sought, of the association study design lends itself more favourably than linkage to detecting the genetic effects of common variants with low RRs (Risch and Merikangas, 1996; Risch, 2000). Most initial studies aim for 80% power, but a higher power is desirable in replication studies. In order to calculate power in a case-control association study, it is necessary to determine firstly, the level of significance desired. This will reflect the level of Type 1 errors to be tolerated. A p-value of at least <0.05 is standard (a 5% chance of a Type 1 error). Secondly, it is necessary to estimate the size of the effect. This can be parameterised in terms of the allele frequency in the controls and the increase in RR expected in the cases: the RR of a particular allele for developing illness. Unfortunately, RR is usually unknown and therefore it is prudent to design a study based on a small RR. The RRs and ORs that have been reported in the literature for psychiatric illness are variable but generally range from as low as 1.2 to ~3.0 (O'Donovan and Owen, 1999; Li *et al*, 2004_b).

Allele frequency can also impact upon power calculations. For example, for a RR of 2 and a Type 1 error rate of 0.05, approximately 600 individuals would be required to observe an association with a marker that has a susceptibility allele frequency 0.50, but this increases to approximately 1500 if the allele frequency is 0.10 (Glatt and Freimer, 2002). This is because a low susceptibility allele frequency means that the same RR is harder to detect. For example, a RR of 2 and a susceptibility allele frequency of 0.40 in controls would mean that the frequency of the allele in a case population would be 0.57 (because RR = (0.57/0.43)/0.40/0.60) = 2). However, if the susceptibility allele frequency is 0.10, the frequency in the case group would be 0.18. Furthermore, the lower the RR, the larger the discrepancy (Glatt and Freimer, 2002). For example, a RR of 1.2 and a susceptibility allele frequency of 0.10 in controls would give a frequency in the case population of 0.12. Genotyping errors can also significantly decrease the power of a test (Gordon *et al*, 2002) and it is possible therefore to account for an error rate in the power calculation. Once these parameters have been chosen, a formula (based on the sampling properties of the central Chi-squared distribution under the null hypothesis and the non-central Chi-squared distribution under the alternative hypothesis) determines the sample size required to attain 80% power.

Inevitably, an association study will test many markers, increasing the occurrence of Type 1 errors. A large number of simultaneous tests on the same data set is akin to placing a single bet at a casino, but spinning the roulette wheel numerous times. The chances of winning, or finding a positive result, increase. The Bonferroni correction is used to correct for multiple testing by raising the standard of proof required to identify a positive result, therefore reducing the Type 1 error rate. A Bonferroni adjusted p-value is the desired Type 1 error rate divided by the number of outcomes being tested. For example, testing 10 SNPs for association means that a p-value 10 times more stringent should be required to accept the hypothesis (0.05/10 = 0.005). Corrections for multiple testing should also be considered if the study groups are split for different factors (e.g. gender). Inappropriate correction results in either an

increase in Type 1 errors due to weak correction or a decrease in statistical power due to overly stringent correction (Cardon and Bell, 2001).

The application of correction for multiple testing in association analysis is controversial. If tests are correlated, the problem of multiple testing is less critical because each new test is not an independent opportunity for a Type I error. Markers in LD are not independent tests and a bonferroni correction is likely to be too conservative (Ott, 2000). However, Ott does not suggest what level of LD would be sufficient, especially in light of the fact that LD calculations will vary depending on the statistic used. Researchers also tend to carry on testing many markers in many locations regardless of the results obtained. This raises the question of whether a genome wide correction value is required and whether this should be based on how many markers will be tested in the future (Nyholt, 2001). Furthermore, a reported association study that has been corrected for a certain number of tests may have originally tested a great deal more (Tomassi, 2004). Due to these problems, several researchers stress the importance, not of correction per se, but of providing a detailed appraisal of the strategies used and the importance of replication studies (Malhotra and Goldman, 1999; Nyholt, 2001; Tomassi, 2004).

In a review, Cardon and Bell (2001) suggest a set of guidelines for performing association analysis. More than one control population should be used in each study and sample sizes should accommodate suboptimal scenarios, such as conservative RR estimates. Multiple testing should be accounted for and different sample panels (different ethnic backgrounds) should be considered to help refine regions of LD. Independent confirmation of an association should be obtained and a lack of association is meaningless. Very few studies conform to these guidelines and for most it is simply not practical. In view of the associated problems with linkage and association studies, the most effective strategy may be to combine them both, using association studies to analyse regions of suggested linkage (Baron, 2001).

1.5.2.3. Linkage Disequilibrium and its Role in Association

All measures of LD are based on the covariance measure D, where D is the difference in the observed and expected frequencies of AB gametes for two loci with alleles A,a and B,b. Measures of LD differ in the way that D is scaled and studies that use different measures cannot be compared (Nordborg and Tavare, 2002). The two most common methods used to calculate LD are D' and r^2 (Nordborg and Tavare, 2002). Values lie between zero and one, where one indicates complete LD. The r^2 measure is the squared correlation coefficient of the alleles between two loci where r is D that has been scaled by the standard deviations of the allele frequencies. The 2x2 Chi-square contingency table test of gamete frequencies provides a significance test for the null hypothesis $r^2=0$. A disadvantage is that r^2 is dependent on the difference in allele frequencies at the two loci, so that r^2 can only take on its full range of 0 to 1 when the allele frequencies at the two loci are the same. However, r^2 is not biased by small sample sizes and low allele frequency (Tishkoff and Verrelli, 2003). In contrast, D' scales D by the maximum value it could take given the allele frequencies at the two loci, so that D' can take on the full range of 0 to 1 whatever the allele frequencies. However, low allele frequency and small sample size often result in D' estimates of 1.

Most power calculations are based on the SNP in question being the susceptibility SNP, but this is not generally the case. Although LD between SNPs ought to maintain power, the relationship is complex and depends upon effect size, the disease allele frequency (DAF), the marker allele frequency (MAF) and the LD between the two. Using data from several susceptibility alleles reported in the literature, Zondervan and Cardon (2004) show that if a susceptibility allele has a high RR (4.0), a study will have high power to detect it with markers across a wide range of allele frequencies even with only moderate LD between the two (D' <0.5). A moderate RR (2.0) means that only common markers (>0.1) will have high power to detect common disease variants, even if their allele frequencies are not the same, as long as LD is high (D' >0.8). A low RR (1.2) means that common markers will only have

power to detect association if the allele frequencies match and LD is high (Zondervan and Cardon, 2004). It has also been found that haplotypes have greater power to detect association than single marker tests and are more robust (i.e. less susceptible to the effects of random drift, mutation and LD) (Akey *et al,* 2001).

The power to detect association is also affected by haplotype phase. For example, the T allele of SNP T/t has a RR of 2 and a frequency of 0.3 (t = 0.7). A case-control study genotypes two SNPs in this region: A/a and B/b. T occurs only on the AB haplotype. This haplotype has a frequency of 0.3 in the population. A is present only on the AB haplotype and therefore also has a frequency of 0.3. B occurs on haplotypes AB and aB with a total frequency of 0.7. Thus, A is in complete LD with T (according to both D' and r^2), has the same frequency (0.3) and therefore also has a RR of 2. B is also in complete LD with T in terms of D' (since T only occurs with A and B), but it has a different frequency (due to its presence on other haplotypes, indicating an r^2 of <1) and therefore has a reduced RR of 1.43. Detecting association with either A or B would appear to be optimal since there is complete LD and the allele frequencies are the same as T/t. However, because the haplotype phase is different (i.e. the frequency of T and t is 0.3 and 0.7 respectively and the frequency of B and b is 0.7 and 0.3 respectively, but it is allele B that shared a haplotype with T), the RR is different. For equal statistical power, a sample size 5.5 times greater would be required to detect an association with B than A. Therefore, matching allele frequency and high LD is not enough to ensure power. However, haplotype phase cannot be predicted (Zondervan and Cardon, 2004).

1.5.2.4. Findings of Association Analysis

Due to the problems discussed in the previous sections, association studies of psychiatric illness have been subject to a similar excess of false positive and false negative results as linkage analysis. In an analysis of 301 studies covering 25 different reported associations to a variety of complex disorders, under half (11) of the reported associations had strong evidence for replication. The main reason for

non-replication was underpowered studies due to small sample number (Lohmueller *et al*, 2003). This highlights the fact that false negatives are just as prevalent as false positive results (Owen *et al*, 2000). For example, the associations of schizophrenia to both the DA D_3 and the 5HT2A receptors are estimated to have small effect sizes (1.2) and therefore sample sizes in the region of 1000 individuals would be required for replication (O'Donovan and Owen, 1999).

Most association studies in psychiatric illness have been hypothesis driven studies of candidate genes. However, even after a linkage study, there will be multiple potential candidate genes and since the biological aetiology of psychiatric illness remains obscure, almost any brain expressed gene can be posed as a candidate. The neurotransmitter receptors and other components of their pathways have been the focus of much attention. The dopamine and serotonin family of receptors, acetylcholine and GABA receptors and neurotransmitter related enzymes such as MAO and COMT are among some of the examples (Mynett-Johnson and McKeon, 1996).

Significant association to schizophrenia was identified in a meta-analysis of all published results of the 5HT2A receptor gene 102T/C polymorphism (Williams *et al*, 1997, cited in Owen *et al*, 2000). However, the association found fails to explain the biological mechanism since the polymorphism is silent. No difference in receptor density in the frontal cortex was observed in those with the susceptibility allele. Another SNP in the 5' promoter region of the gene is in LD with 102T/C. However 102T/C does not show allele specific differences in promoter basal activity. Furthermore, the presumptive ancestral allele of 102T/C is, paradoxically, also the one that is associated with schizophrenia (Petronis, 2000). Epigenetic mechanisms, such as genomic imprinting, could account for some of the inconclusive findings from both linkage and association analysis. There is evidence to suggest that expression of the paternal 5HT2A receptor gene is sometimes inhibited. Therefore, it might be critical to determine how many individuals express both copies of the gene, how many express one copy and which one it is (Petronis, 2000).

The DA receptors have been the focus of numerous association studies. There has been a consistent lack of replication of the original association between schizophrenia and the DA D_2 and D_4 receptors (O'Donovan and Owen, 1999; Berry *et al*, 2003). Again, mixed findings have been observed for the DA D_5 receptor gene (Muir *et al*, 2001; Berry *et al*, 2003). However, many studies seem to converge on a consensus for an association between schizophrenia and the DA D_3 receptor (O'Donovan and Owen, 1999; Petronis, 2000).

The targets for association studies in mood disorders have been very similar to those for schizophrenia, in particular the DA and 5HT receptors and transporters. Collier *et al* (1996, cited in Tsuang *et al*, 2004) were the first to study the 5HT transporter in BPAD and RMD patients, finding an association between a low expression variant and illness. Replication studies have been variable (Tsuang *et al*, 2004). There appears to be a general failure to find association between mood disorders and the 5HT2A receptor, despite this receptor consistently being unregulated in depression. In addition, studies of the DA D_2 receptor have been similarly inconsistent, although results for the dopamine D_4 receptor look more promising (Tsuang *et al*, 2004).

The Neuregulin (NRG1) gene on chromosome 8p is an interesting example where linkage analysis and subsequent association analysis has provided compelling evidence for the location of a susceptibility variant for schizophrenia. Stefansson *et al* (2002) conducted a genome wide linkage study in 33 Icelandic families with schizophrenia and reported a Lod score of 3.02 on chromosome 8p. Haplotype analysis, with one marker every 75 kb, identified two risk haplotypes in a 600kb region that contained the NRG1 gene. SNPs were identified by sequencing and an association study was carried out on 394 controls and 478 schizophrenics. A few SNPs showed mild but significant single-marker association. However, a 290 kb seven marker risk haplotype (five SNPs and two microsatellites) was identified at the 5' end of the gene with a relative risk of 2.2. These findings were supported by a TDT test in the linkage population, where the risk haplotype was transmitted at a

ratio of 2:1. Since none of the individual markers showed the same association as the haplotype, they concluded that the causative marker was not one of the haplotype markers.

The same seven marker risk haplotype, conferring a risk of 1.8, was subsequently identified in ~600 Scottish schizophrenics (Stefansson *et al*, 2003), and the results have also been replicated in a British population by Williamson *et al* (Collier and Li, 2003). A further replication by Corvin *et al* (2004) in an Irish population identified a two marker risk haplotype, overlapping with Stefansson's core haplotype.

In 246 Chinese Han schizophrenic family trios, Yang *et al* (2003) found positive association for two randomly chosen SNPs in NRG1, but not for one of the SNPs from the seven marker risk haplotype identified by Stefansson *et al* (2002). Again, Li *et al* (2004_b), in 138 Chinese Han schizophrenic trios and 298 case and 336 unrelated controls, found that neither the seven marker risk haplotype, nor the individual alleles, were in excess in the case and control group. However, three alternative risk haplotypes were identified. One upstream of Stefansson's risk haplotype conferred a relative risk of 3.1. A second, overlapping Stefansson's risk haplotype conferred a relative risk of 2.1. A third risk haplotype, at the 3' end of the gene, was found to confer a risk only in the TDT test in the trios. Iwata *et al* (2004) found no association between Stefansson's risk haplotype and ~600 Japanese schizophrenics. The different haplotypes observed in Chinese and Northern Europeans is not surprising since a different relationship between marker haplotypes and underlying pathogenic variants could be expected to exist. It is possible that either the same variant operates, or that there is locus heterogeneity (Li *et al*, 2004_b). All of these results identify risk haplotypes that confer greater risk than any of the alleles alone, and therefore none claim to have identified the susceptibility variant itself (Collier and Li, 2003).

The NRG1 gene is a good candidate. The protein has multiple functions in a variety of tissues, including directing the migration of cortical neurons along radial glial cells, the regulation of NMDA, ACh and GABA receptor expression and long term

potentiation (Collier and Li, 2003). Stefansson *et al* (2002) showed that mice with reduced NRG1 activity or expression display hyperactivity that was reversible with antipsychotics, and impaired pre-pulse inhibition, a startle response that is impaired in schizophrenic patients. Hashimoto *et al* (2003) identified a 16-23% increase in expression of NRG1 type I isoform in the prefrontal cortex of schizophrenic post-mortem brains which was positively correlated with medication dose.

1.5.3. Heterogeneity in Psychiatric Illness

Locus heterogeneity in psychiatric illness can be inferred from the numerous genomic regions implicated by linkage analysis (Berry *et al*, 2003; Tsuang *et al*, 2004). This in turn implicates a number of candidate genes and replication studies from association analysis are beginning to consistently implicate several genes, for example NRG1, COMT and proline dehydrogenase (Collier and Li, 2003).

Locus heterogeneity does not preclude the possibility of allelic heterogeneity in psychiatric illness, where multiple mutations in one gene can lead to the same phenotype. This is common in monogenic disorders. For example, more than 963 mutations in the CFTR gene have been identified that cause cystic fibrosis (Wright and Hastie, 2001). However, allelic heterogeneity in psychiatric illness is largely hypothetical since the identification of susceptibility loci is in its early stages. Despite this, some indications for allelic heterogeneity are emerging. For example, the seven marker risk haplotype for schizophrenia in the NRG1 gene occurs in a number of European populations (Collier and Li, 2003) but is not present in Chinese populations (Li *et al*, 2004_b). However, the Chinese population possesses three alternative risk haplotypes, two of which do not overlap with the European risk haplotype (Li *et al*, 2004_b). This strongly suggests that different susceptibility alleles, from different regions of the NRG1 gene, operate in different ethnic populations.

Locus and allelic heterogeneity will mean that susceptibility variants will be harder to detect by linkage and association analysis and has important implications for study

design. Linkage studies might benefit from studying large families, where a more limited number of interacting genes and environmental influences will be operating. A number of genomic regions have been identified in such a manner. Linkage to BPAD from multiple independent studies on large pedigrees has been observed on chromosomes 4p and 12q (Blackwood *et al*, 1996; Blackwood *et al*, 2001). However, in a single family, the disease may be the result of a rare highly penetrant variant and therefore of limited application to the wider population (Blackwood *et al*, 2001). Both association and linkage analysis will benefit from strategies to reduce the proportion of a sample whose disease is caused by non-genetic factors, for example, by studying severe forms of disease or extreme clinical phenotypes (Wright and Hastie, 2001) and sample sizes should be large enough to detect small effects.

The complexity of the psychiatric phenotype is well known. DSM-IV-RT (American Psychiatric Association, 2000) allows for substantial phenotypic variation between individuals by diagnosing from a minimal number of core symptoms. Measuring phenotypic heterogeneity might aid the detection of susceptibility variants. For example, a psychotic form of BPAD characterised by delusions and hallucinations has been identified by several researchers (Leckman *et al,* 1984; Potash *et al,* 2001) and Potash (2003) identified stronger evidence for linkage of BPAD to chromosome 13q and 22q in pedigrees that were characterised by psychotic symptoms than pedigrees that were not.

Consideration of ethnicity as a source of locus heterogeneity is also important. For example, a meta-analysis of 24 studies of association of schizophrenia to an allele of the DRD3 gene did not support association. However, when the samples were divided into five different ethnic groups (African, Northern European Caucasian, Mediterranean, Asian, and American), an association was identified in the Caucasian group with an odds ratio of 1.23 for the homozygous genotype (Dubertret *et al*, 1998). These results suggest that the Northern Europeans possess a susceptibility allele not present in the other four ethnic groups. The small OR could explain the

lack of association identified in some of the individual studies but, without consideration of ethnicity, even the meta-analysis did not identified an association.

1.5.4. Characterising Linkage Disequilibrium in the Human Genome

The characterisation of genome wide patterns of LD is an important consideration for complex disease gene mapping. Consequently, recent attention has been on characterising LD decay over genomic distance. Population simulations estimated that LD would not extend much past 3 kb in a population and therefore ~500,000 SNPs would be required for whole genome association and haplotype studies (Kruglyak, 1999_b). However, other simulations report that in the order of 30,000 SNPs or less would be sufficient (Collins *et al*, 1999, cited in Ott, 2000).

Population differences in LD decay have been observed. The theory is that younger more genetically isolated populations, such as those founded recently in geographically isolated regions, will be less genetically diverse and will show greater LD around disease genes, making the disease gene easier to detect with fewer markers. Increased LD has been observed around rare disease mutations in isolated populations, but the situation for common disease is unclear (Jorde, 1995 cited in Kruglyak, 1999_a). Conversely, reduced LD was observed in the genetically heterogeneous African (Nigerian) population, extending over just ~5kb compared to ~60kb in Europeans (Reich *et al*, 2001). However, the |D'| measure of LD used here is biased inversely with sample size and the African sample was nearly twice the size of the European sample. If only half of the African samples are used, the LD patterns of Africans and Europeans are more similar (Weiss and Clark, 2002). Lonjou *et al*, (1999) performed a meta-analysis of three studies to compare LD in a number of sub-populations from the eight major geographic regions (Europe, Near East, India and Pakistan, Far East, sub-Saharan Africa, the Americas, Oceania and North Africa) for two blood group loci on chromosome 4 and chromosome 1. Six of the populations studied were genetic isolates. They found that while sub-Saharan Africa displayed consistently lower LD, the other seven populations showed little variation, with only

slightly higher LD in the genetic isolates (Kruglyak, 1999_b). However, these two genomic regions do not necessarily represent the entire genome and only four pairs of loci were considered. Other research has also found that LD levels on chromosome 18 and X in the isolated populations of Finland and Sardinia did not differ markedly from those in the more mixed populations of the UK and the US (Taillon-Miller *et al,* 2000; Eaves *et al*, 2000).

In 2001, researchers began to find that LD, rather than decaying gradually over genetic distance, occurred in blocks separated by recombination hotspots (Jeffreys *et al*, 2001; Daly *et al*, 2001; Rioux *et al*, 2001). Empirical evidence for recombination hotspots has come from studies of crossover in sperm (Jeffreys and May, 2004). The implication of LD blocks was that there needed to be a shift from an assessment of LD's dependence on genetic distance to the detection of block boundaries. Consequently, a small number of markers in each block could represent haplotype diversity. For example, Daly *et al* (2001) showed that over an 84kb stretch of sequence, two haplotypes accounted for more than 96% of the variation and that one SNP would capture most of the variation over this stretch. In a genome wide study, Patil *et al* (2001) identified haplotype blocks of limited diversity and found that greater than 80% of twenty globally diverse human chromosomes could be characterised by three common haplotypes (Patil *et al*, 2001). The implication of this for association analysis is that a minimal number of SNPs could be chosen to represent haplotype diversity.

There are many different methods, based essentially on two different types of analysis, to statistically determine the presence of blocks. The relative strengths and weaknesses of each depend upon the study aims. The first type of analysis defines blocks as regions of limited haplotype diversity. Low diversity is associated with increased LD. However, thresholds for the number of haplotypes and proportion of observations that define a block are often subjective and arbitrary and the algorithms used to identify blocks need further validation (Tishkoff and Veralli, 2003). For example, Patil *et al* (2001) analysed 24,047 SNPs from haploid copies of

chromosome 21 isolated in rodent-human somatic cell hybrids. The algorithm used identified 4,135 blocks using 4,563 SNPs. In contrast, Zhang *et al* (2002), using an alternative algorithm on the same data set, identified 2,575 blocks using 3,582 SNPs. If the goal for disease gene mapping is to minimise genotyping whilst maximising information content then no reference to the underlying LD is necessary (Van den Oord and Neale, 2004).

The second method of haplotype block detection uses pairwise disequilibrium to identify transition zones in which there is evidence for recombination, i.e. a contiguous set of SNPs in which the average LD is greater than some predetermined threshold (Van den Oord and Neale, 2004). Reich *et al* (2001) used a D' half-life value of <0.5 to detect blocks whilst Gabriel *et al* (2002) defined a block as a region where less than 5% of SNP pairs show evidence for recombination (D' >0.9). Genotyping errors mean that a D' of 1 is too stringent. For example, a genotyping error of 2% and a D' threshold of 1 would mean that only 55% of the true blocks would be identified, but a D' of 0.9 would identify 97% of blocks (Van den Oord and Neale, 2004). However, if values are too low, blocks will be missed. In a simulation study, a D' of 0.7 meant that 40% of markers with a recombination between them would be classified into a block, compared to 21% for a D' of 0.8 and 7% for a D' of 0.9 (Van den Oord and Neale, 2004). The measure of LD used will also affect the block structure detected and evidence suggests that D' is better than r^2. For example, between two SNPs (with alleles A,a and B,b), the strict definition of a haplotype block is where there has been no recombination and mutation is the only source of variation. Therefore, three of the four possible haplotypes will be observed (A,B; a,b and either A,b or a,B). D' in this case would be 1. However, r^2 will only be 1 if two of the four possible haplotypes variants are observed (A,B or a,b) (Van den Oord and Neale, 2004). Simulation studies suggest that even for a region which is a block according to the strictest definition, r^2 will often be close to zero (Van den Oord and Neale, 2004). However, D' is biased upward inversely with sample size. Therefore, if a sample is to be used to design a global mapping panel, it should be large enough that the observed LD structure is not an artefact of small sample size (Weiss and

Clark, 2002). Simulations suggest that at least 100 chromosomes should be used to adequately represent LD (Wang *et al,* 2002)

If blocks do exist and they are the result of recombination hotspots, they should be shared across populations. However, if they are due to other factors such as population bottlenecks, they will be population specific. Furthermore, they may not apply to the whole genome (Tishkoff and Verrelli, 2003; Wall and Pritchard, 2003; Van den Oord and Neale, 2004). A study using a selection of the SNPs from the dbSNP database (www.ncbi.nlm.nih.gov/SNP/) for chromosome 19 showed that only 32% of the chromosome exhibited a block like structure, and recombination hotspots were not required to fit the data (Phillips *et al*, 2003). Wall and Pritchard (2003) estimated that haplotype blocks only account for 50% of the genome. If LD is to be utilised for mapping, these issues need to be resolved.

Large scale projects are underway to utilise LD for disease gene mapping using SNPs. The dbSNP database is a reference database of human SNP variation. However, Johnson *et al* (2001) found that dbSNP did not hold sufficient SNPs to capture the haplotype diversity for a 135kb region spanning nine genes (Johnson *et al*, 2001). The HapMap project (The international HapMap consortium, 2003) (www.hapmap.org/) builds on dbSNP and aims to characterise human diversity and genomic LD patterns in populations from Africa, Asia and Europe, by detecting common variants, their frequencies and the correlations between them. However, this will only be useful if common SNPs underlie susceptibility to the diseases to be mapped (Tishkoff and Verrelli, 2003).

1.5.5. The Common Disease/Common Variant Hypothesis

Common complex diseases are polygenic (multiple genes) or multifactorial (multiple genes and environmental factors). Susceptibility variants have low penetrance because they are not sufficient to cause the disease and, therefore, exist in the general population. The Common Disease/Common Variant (CD/CV) hypothesis of common

complex disorders states that, at each disease loci, one, or a small number of, common (>0.10 frequency) susceptibility variants underlie disease. The implication of this is that disease gene mapping by association analysis will be a powerful tool for detecting loci of small effect and it underpins the applicability of the HapMap project (Pritchard and Cox, 2002).

For the majority of common complex diseases, disease variants have not been identified. However, a meta-analysis of 25 different reported associations to complex disease found more positive replications than would be expected by chance, suggesting that true positive associations make up a substantial portion of the literature. Since the analysis was restricted to common variants, the results also support the CD/CV hypothesis (Lohmueller *et al,* 2003). Common variants have been identified for several common diseases. The APOE*E4 allele is a common allele (frequency 0.04-0.49 in controls) that increases the risk of Alzheimer's disease, the PPARγ allele (frequency 0.85 in controls) is implicated in the risk for type 2 diabetes and the INS tandem repeat allele (frequency 0.75 in controls) is implicated in the risk for type 1 diabetes (Pritchard and Cox, 2002).

The possibility exists that these alleles were identified because they are more penetrant than most, or have a simpler allelic architecture than other disease alleles (Pritchard and Cox, 2002). Furthermore, association analysis might not have enough power to detect variants with a more complex allelic architecture. The NOD2 locus involved in Crohn's disease is fairly complex, with several variants identified, each with a frequency of ~1% in controls and a total frequency of susceptibility variants of ~10% in controls (Hugot *et al*, 2001; Ogura *et al*, 2001). However, these variants were identified because the gene made good biological sense and the mutations were obvious (non-synonomous and frameshift mutations). With lower penetrance, a gene of unknown function, or synonomous SNPs, variants might not be so easily found (Pritchard and Cox, 2002). A further implication is how applicable the CD/CV hypothesis really is. Substantial allelic complexity might underlie many currently

undiscovered common disease variants simply because association analysis does not have the power to detect them.

Several historical factors favour the CD/CV hypothesis. Humans underwent a large rapid population expansion between ~18,000-150,000 years ago (700-6000 generations) increasing in size from 10^4 to the present $6x10^9$ (Smith and Lusis, 2002). Alleles that negatively affected reproductive fitness in the founding population would have been at a low equilibrium frequency, whilst alleles with little or no effect on fitness, for example, diseases with late onset or a heterozygous advantage, would have been at a high equilibrium frequency. Furthermore, alleles that were common in the founder population will still be common today because there has not been enough time to dilute them out (Wright and Hastie, 2001). Therefore, today's common diseases are the ones that are caused by common alleles that were not under selection pressure in the past. In contrast, alleles that were rare in the founder population will today be a substantially rarer member of a diverse allelic set (Wright and Hastie, 2001). Today's monogenic diseases follow this. For example, Haemophilia B is a monogenic disease with 167 distinct mutations, hypothesised to be the result of at least 302 independent mutation events (Green *et al*, 1999).

Theoretical arguments are upheld by computer simulations. Modelling an ancestral population of 10,000, with subsequent rapid expansion to $6x10^9$, Reich and Lander (2001) analysed one disease locus with multiple disease alleles, where rare diseases were subject to strong selection and common diseases were subject to mild or no selection and/or a heterozygote advantage. The initial and final stages for a rare and a common disease are the same, with low allelic diversity in the initial population and high diversity in the expanded population, but the kinetics are different. Allelic diversity occurs rapidly (over 1000s of years) for rare diseases and slowly (over millions of years) for common diseases due to the effects of mutation and selection. The rate of new disease mutations per generation is the same for common and rare disease but the pool of disease chromosomes they enter for the rare disease is much smaller, resulting in a greater proportional effect on that class per generation and a

more rapid turnover. Selection against the rare disease is more intense than for the common disease, again leading to a more rapid turnover. However, genetic drift (meaning that allele frequencies are not at equilibrium) and non-random mating (in sub-populations and bottlenecks) were not accounted for. Genetic drift would lead to a more complex allelic structure and non-random mating would lead to a simpler allelic structure (Reich and Lander, 2001).

The CD/CV hypothesis is perhaps more of a guiding principle than a definitive rule. Some common complex diseases, such as heart disease, are common because of highly prevalent non-genetic influences, not because of common disease alleles and show a reduction in heritability with increasing age (Wright and Hastie, 2001). Some Mendelian disorders do not follow the predictions of allelic heterogeneity. Cystic fibrosis, a recessive disorder, does show a high level of allele diversity, with >963 rare alleles identified, but, 70% of patients possess the deltaF508 allele, an allele that is also found on 1.5% Caucasian chromosomes (Wright and Hastie, 2001). The skew might be due to a selective heterozygous advantage. For example, it is believed that deltaF508 allele confers resistance to cholera (Pritchard and Cox, 2002). Conversely, the CD/CV hypothesis does not rule out the possibility of a large number of rare alleles operating in complex disease (Smith and Lusis, 2002).

There are a number of implications of the CD/CV hypothesis for complex disease gene mapping and the HapMap project. Association requires the enrichment of a predisposing allele, but if substantial allelic heterogeneity exists, each new mutation will arise on an independent haplotype background and association will not have the power to detect a difference (Pritchard and Cox, 2002). It is also unknown what level of locus heterogeneity exits for most complex disorders. Extreme heterogeneity would make mapping difficult since the effect size of each locus would be reduced (Pritchard and Cox, 2002). Again, strategies to reduce the proportion of a sample whose disease is caused by non-genetic factors are sensible (Wright and Hastie, 2001).

1.6. The Human Genome Project.

1.6.1. Why Sequence the Human Genome?

The draft sequence of the human genome was published in February 2001 and the finished sequence was announced on the 14th April 2003. The coincident publication of the 'rival' public (Lander *et al,* 2001) and private (Venter *et al*, 2001) draft sequences represented a token gesture for the media and was fairly arbitrary. Even with the announcement of the finished sequenced, sequencing is still ongoing and a number of gaps remain. The current status (February 2004) reported at the Sanger Institute (www.sanger.ac.uk/HGP/) is that 88.99% of the genome sequence is finished and 4.23% of the sequence is of draft quality.

The benefits of the human genome sequence for biomedical research are phenomenal. The Sanger Institute, responsible for approximately one third of the finished sequence, is now beginning to focus on using the sequence to identify disease genes. They claim that the finished sequence of chromosome 20 has accelerated the search for susceptibility genes to diabetes, childhood eczema and leukaemia. The benefits of the finished human genome sequence and follow on projects such as the SNP databases and the HapMap project, promise to be most pronounced for gene mapping in complex diseases and disorders. As has been discussed, common multifactorial diseases, such as psychiatric illness, hypothesised to be caused by common gene variants, are proving notoriously difficult to map. A catalogue of human variation imposed upon the genes along a section of the genome will help to guide future studies to define disease genes. However, the exact benefits that these large scale projects will bring are unclear (Terwilliger *et al,* 2002).

1.6.2. Characteristics of the Human Genome

The most striking results of the human genome sequencing project (HGP) has been the increasing confirmation of the relatively few genes that make up the human

genome. The early estimates 100,000 have been curbed dramatically to the order of 30,000 or less (Pennisi, 2003). The fact that the nematode worm *Caenorhabditis elegans* has approximately 20,000 genes means that there is not as much difference between humans and simple life forms as previously thought. That the fruit fly *drosophila* has less, at 14,000 than the distinctly less sophisticated *C.elegans* is a further puzzle (Claverie, 2001). The belief is that increasing organismal complexity must therefore come from alternative splicing or post-translational modifications. It has been estimated that approximately 30-50% of genes might be alternatively spliced (Lander *et al*, 2001; Venter *et al*, 2001). However, there is evidence that the rates of alternative splicing and the number of alternative transcripts produced per gene are comparable in seven organisms, including humans, using a bioinformatics protocol to compare ESTs and mRNA sequences (Brett *et al*, 2002). Therefore, post-translational modification, rather than alternative splicing, may account for the increased complexity of humans. Post translational modification, of which there are over 200 different types known, for example phosphorylation and glycosylation, is an important determinant of protein function (Banks *et al*, 2000).

Another puzzling finding identified as a result of the genome projects is that genome size does not correlate with organism complexity. For example, the human genome is 200 times larger than that of the yeast *S. cerevisiae*, but is 200 times smaller than that of *amoeba dubia* (Lander *et al,* 2001). This can be explained by the fact that less than 5% of the human genome contains the coding and regulatory sequences of genes, whereas approximately 50% of it is composed of five classes of repetitive sequence: transposon-derived repeats (accounting for 45% of the genome), processed pseudogenes, simple sequence repeats (for example microsatellite repeats), segmental duplications and blocks of tandemly repeated sequences (Lander *et al*, 2001).

The transposon-derived repeats can be divided into four types: long interspersed elements (LINEs), short interspersed elements (SINEs), long terminal repeat (LTR) retrotransposons and DNA transposons. LINEs are 6-8kb in length and consist of a polymerase II promoter and two ORFs. A translated LINE RNA, with its protein, will

reverse transcribe itself from the 3' end into DNA; often not successfully completing reverse transcription to the 5' end and creating a truncated insertion (Lander *et al*, 2001). SINEs are between 100 and 400 bp and do not encode proteins and are thought to use the LINE machinery to retrotranspose themselves into DNA. The only active SINE in the human is the Alu element (Batzer and Deininger, 2002). LTR retrotransposons are DNA elements characterised by long terminal repeats and DNA transposons are DNA elements that move about the genome by excision and reintegration without an RNA intermediary (Smit and Riggs, 1996).

Pseudogenes are non-functional genes that result from a failure of transcription, translation or result in an altered protein function. Processed pseudogenes arise by reverse transcription of the mature mRNA into the genome sequence, utilising the reverse transcription machinery of the transposon-derived repeats (Lander *et al*, 2001; Venter *et al*, 2001). Pseudogenes can also arise by genomic segmental duplication. Evidence from chromosome 22 suggests that 19% of the genes are pseudogenes and that 82% of these arose by retrotransposition (Mighell *et al*, 2000).

Pseudogenes and segmental duplication are considered to be one mechanism by which genomes evolve, whereby mutation in the duplicated genes allows them to take on new functional roles (Samonte and Eichler, 2002). The HOX gene cluster is one example of a gene cluster that appears to have been duplicated four times in the human genome, although it is unclear whether individual gene duplications or whole genome duplication explain the findings (Spring, 2002).

In addition to the proteome, a significant proportion of human genes encode non-coding RNAs (ncRNAs): RNAs that do not code for protein. These include transfer RNAs that translate RNA into protein, ribosomal RNAs that are also involved in the translational machinery, small nucleolar RNAs that are required for ribosomal RNA processing, small nuclear RNAs that are a component of spliceosomes, and finally ncRNAs that have known or unknown biochemical function (Lander *et al*, 2001).

1.6.3. Annotating the Human Genome

It is increasingly important for the identification of disease genes to annotate the finished human genome sequence. Genes can be identified in three different ways: direct evidence by the alignment of sequenced ESTs and mRNA to the genomic sequence, indirect evidence from the homology to known proteins and DNA in humans and other species and programmes that calculate statistical information about gene structure such as splice sites and ORFs (Venter *et al*, 2001; Lander *et al*, 2001).

Gene and exon prediction programmes make predictions based on the characteristics of known genes and their structure. This is made especially difficult in the human genome due to the large amount of 'noise' from non-coding sequence, making ORFs harder to distinguish (Lander *et al*, 2001). Computational gene identification is composed of two non-trivial steps. First, they need to divide the sequence into segments that are likely to correspond to an individual gene and secondly, they need to construct a gene model that reflects the structure of the transcripts (Lander *et al*, 2001). Gene prediction programmes have been estimated to correctly predict approximately 70-75% of individual exons (Burge and Karlin, 1997; Guigo *et al*, 2000). However, these figures may be overestimated since the training sets are known genes which may have more standard gene structures.

Non-coding RNAs are often small and do not have a translated ORF or a polyadenylation signal. They are therefore hard to identify by standard gene finding programmes or experiments involving cDNA (since cDNA is generated by targeting the poly-A tail of mature mRNAs that is generated after recognition of the polyadenylation signal) (Lander *et al*, 2001).

In addition to gene prediction programmes and homology based matches, CpG islands represent another way of identifying genes. CpG islands are stretches of unmethylated DNA with a high frequency (greater than 50%) of CpG dinucleotides that occur at the transcriptional start site of genes. It has been shown that most

housekeeping genes possess a CpG island (Larsen *et al*, 1992) and that CpG island methylation is correlated with gene inactivation (Venter *et al*, 2001). Venter *et al* (2001) estimated a total of 26,826 CpG islands in the human genome by using a computational method that used a ratio of the observed versus expected frequency of CG dinucleotides that was equal to or greater than 0.8. Interestingly, this figure is comparable to the estimated total number of genes in the genome (Pennisi, 2003). However, it does not account for how many CpG islands are not associated with genes. The finding of Antequera and Bird (1993), that 56% of human genes are associated with CpG islands and the estimate of Venter *et al* (2001), that there are ~27,000 CpG islands in the human genome, do not follow if there are only ~30,000 genes in the human genome. They could be reconciled if some CpG islands are not associated with genes.

1.7. The Chromosome 4 Linkage Families

1.7.1 The Families

Blackwood *et al* (1996) described a genome wide linkage study in a family of 120 individuals, F22, segregating BPAD and RMD (Figure 2-1). A three point linkage analysis gave a maximum multipoint Lod score of 4.8 in the region D4S431-D4S403 on chromosome 4p. A 22Mb haplotype is inherited from one founder individual to all 11 cases of BPAD I and II and 14 out of 16 cases of RMD. This haplotype defines the disease associated chromosome in F22. Two individuals with early onset RMD do not possess the haplotype, suggesting that they are phenocopies. Nine individuals without a psychiatric diagnosis do possess the haplotype, suggesting that there is incomplete penetrance of the susceptibility locus. As previously discussed, psychiatric illnesses are complex disorders where multiple genes and environmental factors contribute to the phenotype. Therefore, phenocopies and incomplete penetrance are expected. In addition, there were six individuals with the haplotype, including the founder, whose diagnostic status was unknown. A reanalysis of the

family 22 data, using a method that does not require specification of the mode of inheritance, confirmed the initial findings (Visscher *et al*, 1999).

A further 11 Scottish families were typed by this group for marker D4S394 in the linked region. Linkage was tested on all 11 families together (a total of 40 cases of BPAD I/II or schizoaffective disorder and 17 cases of RMD) and a combined Lod score of 3.3 with a broad phenotype was identified. It was subsequently shown that only three families (including F59) contributed positive Lod scores. Analysed alone, F59 had a Lod score of 1.2. From subsequent unpublished laboratory work by others in the group, F59 has been shown to exhibit a chromosome 4p haplotype that is transmitted from an unknown founder to all five cases of BPAD I and II disorder and one undiagnosed individual. The individual without a diagnosis who carries the disease haplotype again suggests that there is incomplete penetrance of the susceptibility locus. The Lod score contributed by this family is small (the odds for a Lod of 1.2 are a little higher than 10:1 in favour of linkage), but the small size of the family (n = 6) imposes limits on the maximum Lod score that can be achieved by this family. Therefore, it is possible that the linkage result for F59 is spurious. Family 59 was included in the analysis of the chromosome 4p candidate region with these limitations in mind.

Two remaining families, F50 and F48 (Figures 2-3 and 2-4), were reported in the literature as showing linkage of psychiatric illness to chromosome 4p (Asherson *et al*, 1998; Detera-Wadeligh *et al*, 1999) and access to these samples has been obtained via collaboration. Family 50 segregates schizophrenia and schizoaffective disorder. However, one individual has a form of psychosis that could not be further classified due to incomplete data. The results of a genome wide linkage study in F50 found a Lod score of 2.0 on chromosome 4p, and a haplotype is inherited with the disorder (Asherson *et al*, 1998). This is below the level for genome wide significance, but the maximum attainable Lod score is limited by the small size of the family. Family 48 is a large family of 82 individuals who segregate BPAD, RMD and schizophrenia. A genome wide linkage study identified a Lod score of 3.2 on chromosome 4p. There is

also a haplotype that is inherited with the illness (Detera-Wadleigh *et al*, 1999). The fact that families 48 and 50 segregate different psychiatric illnesses to families 22 and 59 was recognised. However, there is evidence for other overlapping genomic susceptibility regions in schizophrenia and BPAD and clinical, epidemiological and biological evidence suggests that RMD, BPAD, schizoaffective disorder and schizophrenia should be seen along a continuum rather than separate diseases (discussed in Section 1.5.1.4). These findings supported the consideration of these linkage findings from families 50 and 48 together with those of families 22 and 59.

Other studies have replicated the linkage to chromosome 4p. Ewald *et al*, (1998) studied 16 markers between 4pter-4p12 in two Danish families. They identified a Lod score of 2.0 for marker D4S394 under a recessive model, suggesting bilinear inheritance. Although the linkage from Blackwood *et al* (1996) was found using a dominant model, the results are not contradictory if allelic heterogeneity operates. In a genome wide scan of 196 affected sibling pairs with schizophrenia, Williams *et al* (1999), identified suggestive linkage on chromosome 4p, 18q and Xcen. The Lod score on 4p was 1.7 for marker D4S2983 which lies inbetween D4S431 and D4S403, the region implicated by Blackwood *et al* (1996). Lerer *et al* (2003) have also identified suggestive nonparametric linkage to chromosome 4p15-16 in a population of Arab Israeli families. A recent study of schizophrenic and BPAD individuals from the Faroe Islands found an association to several overlapping two and three marker haplotypes on chromosome 4p16.1 (Als *et al*, 2004). This is interesting in view of the possible common ancestry between the Scottish and the Faroese populations. However, this study was based on very small sample sizes. Interestingly, Ginns *et al* (1998) identified a region on chromosome 4p linked to a 'mental health wellness' phenotype in several large Old Order Amish families segregating BPAD.

1.7.2 Disease Associated Haplotypes

The haplotype inherited with illness in each of the four families studied defines the region that contains the susceptibility polymorphism. Haplotypes are defined by

tracing the inheritance of marker alleles from parent to offspring in a family. An interruption in the inheritance of a contiguous haplotype from a parent to their offspring with the commencement of the inheritance of the alternate haplotype from the second chromosome of that parent constitutes a recombination breakpoint.

Haplotypes define the linkage regions, but they can also be studied for allele sharing since this might be an indication of a founder effect. There are a number of reasons why a common ancestor may be hypothesised. Firstly, three of the families are Celtic. However, this is merely a cultural observation and does not have a genetic basis. Secondly, the disease haplotypes of each of the four families overlap. This overlap may point to a common susceptibility locus. Other studies have found a similar founder effect, such as the seven marker haplotype in the NRG1 gene observed in several Northern European populations (Stefansson *et al*, 2002; Stefansson *et al*, 2003) but not in Chinese or Japanese populations (Li *et al*, 2004_b; Iwata *et al*, 2004).

The disease haplotypes of the four families do not fully overlap. There are two main regions of interest: Minimal Region One (MR1), defined by the overlap of F22, F59 and F50 and Minimal Region Two (MR2), defined by the overlap of F22, F50 and F48 (Figure 1-4). MR1 is a good candidate region because the families are Celtic and association has been found between MR1 markers and schizophrenia (Muir *et al*, 2001). However, MR2 comprises the two largest families, F22 and F48. As mentioned previously, the Lod scores in F59 and F50 are small. However, a large linkage region was identified from F22, containing many potential candidate genes. Financial and time constraints meant that definition of the minimal regions helped prioritise the research efforts into chromosomal regions of more manageable sizes. However, it is recognised that the minimal region boundaries are dependent on the validity of the linkage results of each family and on specific individuals within families. Therefore, genes outside MR1 and MR2, but inside the F22 linkage region, certainly cannot be discounted as potential candidates for future work.

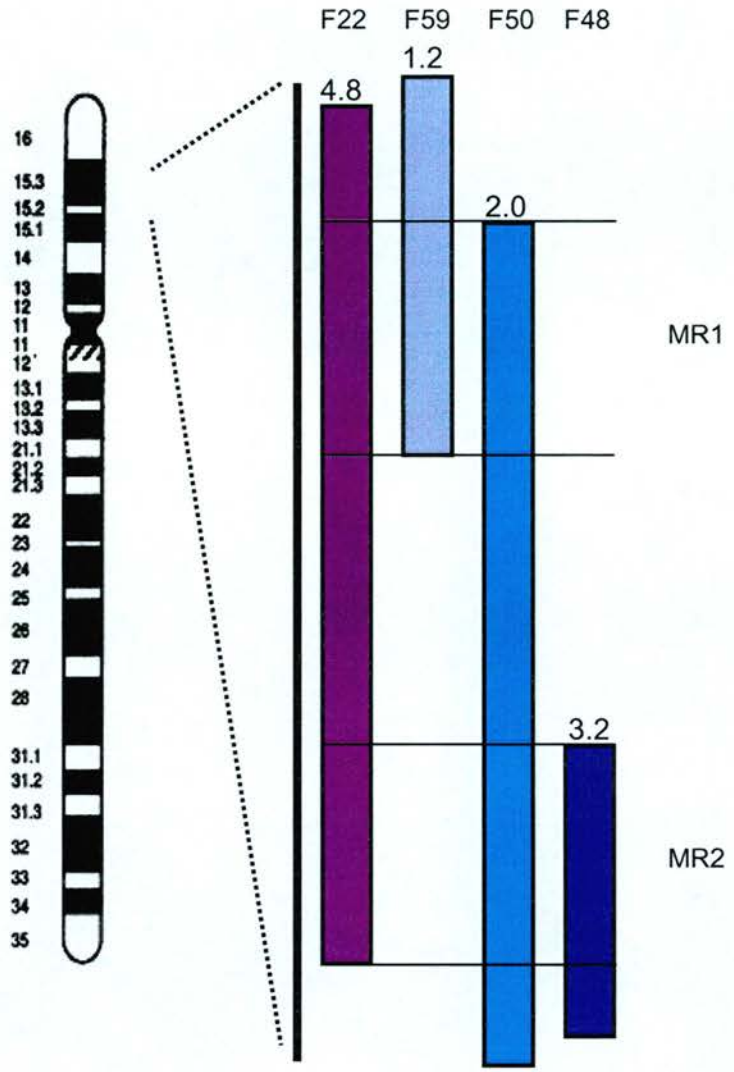


Figure 1-4: Overlapping haplotypes of families 22, 59, 50 and 48. The coloured bars represent a marker haplotype on chromosome 4p inherited with psychiatric illness in each family. The family number is marked along the top. The number above the bar shows the Lod score for that family (F22 = a multipoint Lod score). The four haplotypes overlap, enabling the delineation of smaller regions of interest. The solid horizontal lines mark two such regions: Minimal Region One (MR1) and Minimal Region Two (MR2).

1.8. Thesis Aims

At the start of this project (October 2000) the chromosome 4p disease associated haplotypes of families 22, 59, 48 and 50 in the regions of linkage were in the process of being defined and a BAC/PAC contig of part of this region had been constructed (Evans *et al*, 2001_a). The general aim of this PhD was to define the regions of interest further and use this to guide association analysis of candidate genes. Four specific aims were:

1. To define the genomic region that lies in the recombination breakpoint interval of MR1.
2. To refine the two recombination breakpoint intervals that define MR1 by microsatellite and SNP identification and genotyping.
3. To construct a transcript map of MR1 and MR2. In order to direct and interpret future association analysis of the two best candidate regions, it was important to know the position of all the genes.
4. To carry out association analysis of two positional and functional candidate genes, the orphan G-protein-coupled receptor 78 (GPR78) gene and the superoxide dismutase 3 (SOD3) gene, within MR1 and MR2, respectively.

Chapter Two

Materials and Methods

Materials and Methods

2.1 Clinical Resources

2.1.1. Families

DNA from Family 22 and 59 was collected as a blood sample by Professor D. Blackwood and Professor W. Muir (Department of Psychiatry, Royal Edinburgh Hospital, The University of Edinburgh, Edinburgh). Individuals were ascertained through contact with general practitioners and local hospitals. Each individual was interviewed using the Schedule for Affective Disorders and Schizophrenia (lifetime version). Figure 2-1 and 2-2 detail the family pedigree and includes the diagnosis and ID number of each individual. The sex of each individual was unavailable for reasons of patient confidentiality. DNA from family 50 was obtained by collaboration with Professor P. Asherson (Neuropsychiatric Genetics Unit, Divisions of Psychological Medicine and Medical Genetics, University of Wales, College of Medicine, Cardiff). Figure 2-3 details the family pedigree, and includes the sex, diagnosis and ID number of each individual. DNA from family 48 was obtained by collaboration with Professor S. D. Detera-Wadleigh (Clinical Neurogenetics Branch, National Institute of Mental Health, National Cancer Institute, National Institutes of Health, Bethesda, MD 20892). Figure 2-4 details the family pedigree and the diagnosis and ID number of each individual. Again, the sex of each individual was unavailable for reasons of patient confidentiality.

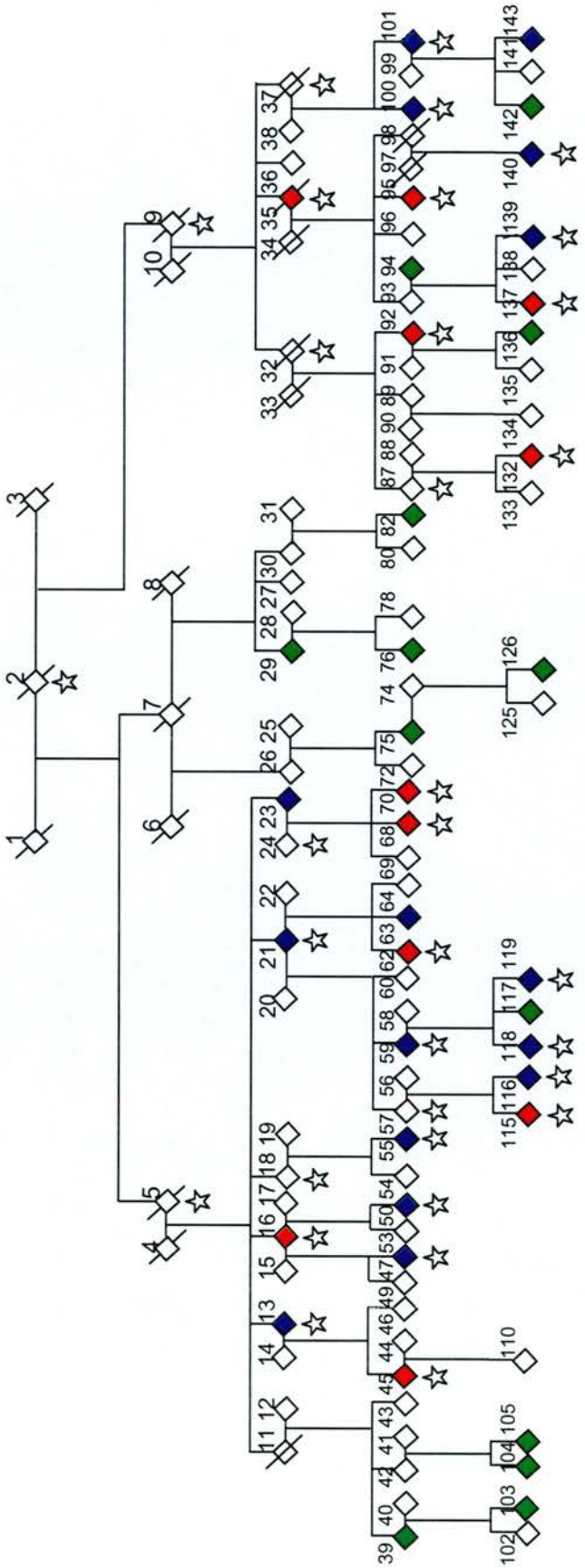


Figure 2-1: Family 22 pedigree. Number = Lab ID number. ◇= deceased. No infill = no diagnosis. Red infill = Bipolar affective disorder I and II. Blue infill = recurrent major depression. Green infill = other psychiatric diagnosis. ☆= disease associated haplotype carrier. Individual gender not available.

Figure 2-2: Family 59 pedigree. Number = Lab ID number. □ = male. ○= female. ⊘ = deceased. No infill = no diagnosis. Black infill = bipolar affective disorder. ☆ = disease associated haplotype carrier.

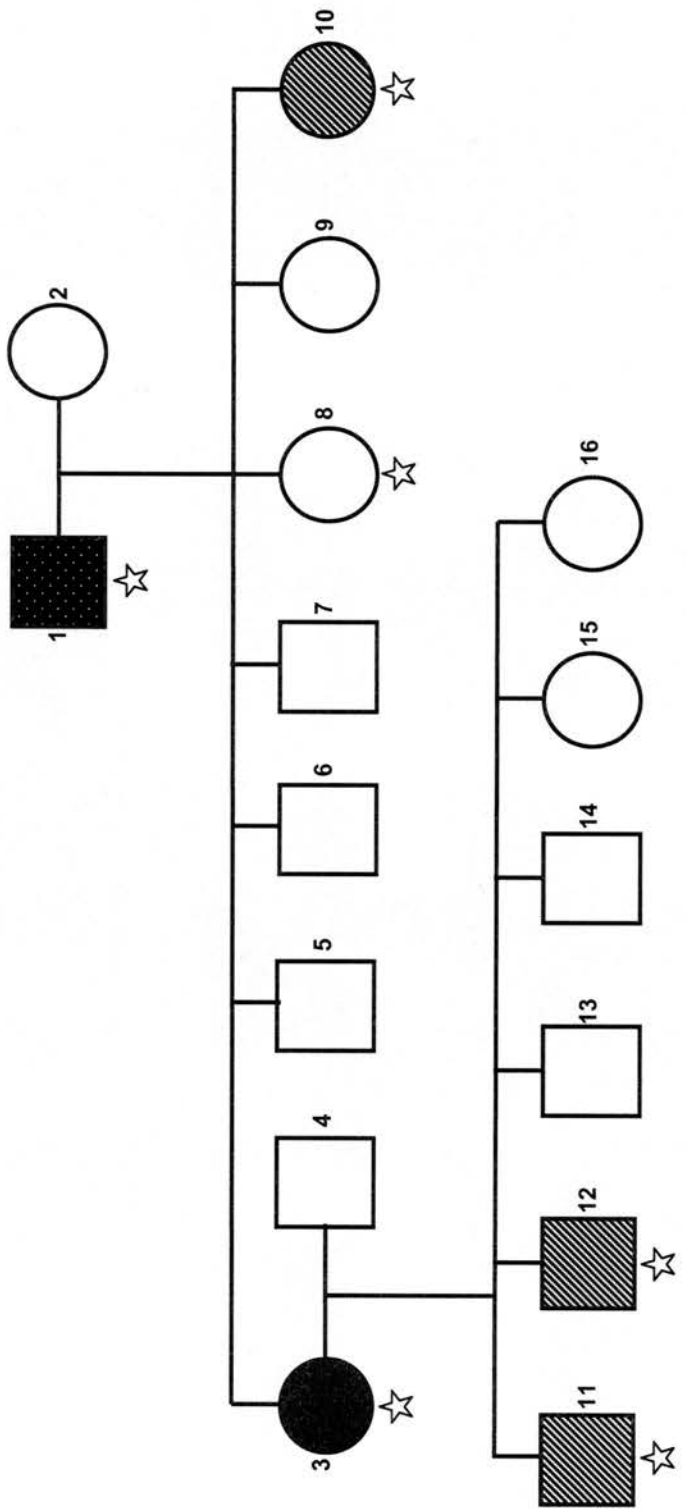


Figure 2-3: Family 50 pedigree. Number = Lab ID number. □ = male. ○ = female. No infill = no diagnosis. Black = schizophrenia. Hatched = schizoaffective. Dotted = Unspecified psychosis, due to incomplete records. ☆ = disease associated haplotype carrier.

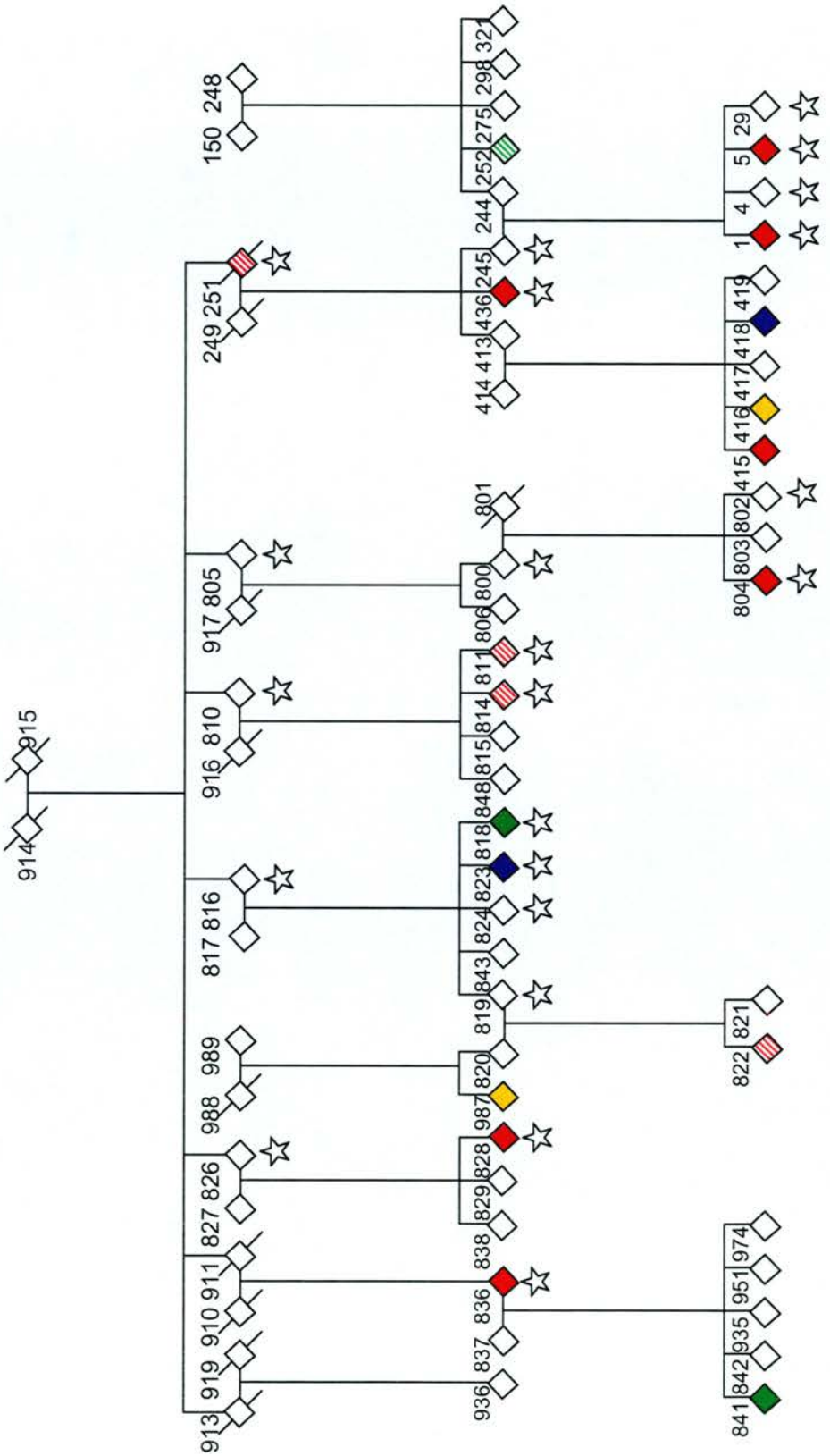


Figure 2-4: Family 48 pedigree. Number = Lab ID number. ◇= deceased. No infill = no diagnosis. Red = bipolar affective disorder. Green = schizophrenia. Blue = cyclothymia. Yellow = Hypomania. Red hatched = recurrent major depression. Green hatched = other psychiatric diagnosis. ☆ = disease associated haplotype carrier. Individual gender not available.

2.1.2. Case and Control Samples

Individuals suffering from schizophrenia (SCZ), bipolar affective disorder (BPAD) and recurrent major depression (RMD) were recruited from the Royal Edinburgh Hospital by Professor D. Blackwood and Professor W. Muir. Control patients were recruited from the blood transfusion service, Edinburgh.

2.1.3. DNA Panels

2.1.3.1. Allele Sharing Panel

The Allele Sharing (AS) DNA Panel contains 46 DNAs, a control DNA pool (Section 2.1.3.2.) and a negative control. The DNA was at a concentration of 10ng/μl. The plate contained thirty-one members of F22, five members of F59, seven members of F50 and three members of F48. This represents multiple copies of the four disease chromosomes, to enable unambiguous determination of the disease haplotype, and 41 control chromosomes (defined as all non-disease chromosomes). The control pool was included to give a measure of the range of alleles for microsatellite markers and the allele frequency in a control population. See Figure 2-5: A, for the plate layout and family IDs.

2.1.3.2. Pools Plate

DNA samples from the case and control groups (Section 2.1.2.) were pooled by S. Le Hellard. Four DNA pools corresponding to diagnosis were constructed: 192 individuals with schizophrenia, 271 with BPAD I and II, 87 individuals with RMD and 383 controls.

The pools plate was set up with either four or six replicates of each pool and 23 individual control DNAs. The 23 individual DNAs ensure that for SNPs with a minor allele frequency of 0.20, there will be approximately five individuals with the minor

allele. This provides enough individuals from which to calculate the *K* ratio (Section 6.4.4.). See Figure 2-5: B, for the plate layout with six pool replicates.

A

50 1	50 2	50 3	50 4	50 7	50 12	50 15	59 208	59 209	59 200	59 268	59 374
22 9	22 30	22 77	22 32	22 78	22 33	22 4	22 36	22 6	22 38	22 39	22 40
22 2	22 9	22 46	22 47	22 8	22 49	22 55	22 57	22 97	22 98	22 45	22 146
22 49	22 50	22 99	22 47	22 53	22 103	22 43	48 23	48 5	48 36	pool	neg.

B

690	698	620	621	685	687	684	578	792	753	777	638
630	631	637	640	644	645	646	697	694	692	671	neg.
C	C	C	C	C	C	BP	BP	BP	BP	BP	BP
S	S	S	S	S	S	RMD	RMD	RMD	RMD	RMD	RMD

Figure 2-5: DNA panels used (96 well plate layout). **A:** Allele sharing panel. Top left = family identification number (ID), bottom right = individual lab ID. pool = DNA pool of 383 control individuals. neg = no template control. **B:** Pools plate. Number = control individual lab ID. neg = no-template control. C = DNA pool of 383 control individuals, BP = DNA pool of 271 recurrent bipolar affective disorder I and II patients, S = DNA pool of 192 schizophrenia patients, RMD = DNA pool of 87 patients with recurrent major depression.

2.1.3.3. Monochromosome Hybrid Panel

A monochromosomal somatic cell hybrid DNA panel (MCHP) was obtained from the MRC-RFGRC (Kelsall *et al*, 1995) (Table 2-1). This is a panel of rodent (mouse or hamster) somatic cells that each retains a single human chromosome on a rodent background. The panel is obtained in a 48 well microtitre plate containing 27 samples: 1μg of each human chromosome in 25μl of sterile water, and 1μg of human, hamster and mouse control DNA. The MCHP was diluted to 30ng/μl, and 15ng was used in a 10μl PCR (Section 2.3).

Chromosome	DNA	Rodent Origin
1+X	GM07299	Hamster
2	GM10826B	Hamster
3	GM10253	Hamster
4	HHW416	Hamster
5	GM10114	Hamster
6	MCP6BRA	Mouse
7	CLONE21E	Mouse
8	C4A	Hamster
9	GM10611	Hamster
10	762-8a	Mouse
11	JICL4	Hamster
12	1Aa9602+	Mouse
13	289	Mouse
14	GM10479	Mouse
15	HORLI	Mouse
16	2806H7	Mouse
17	PCTBA1.8	Mouse
18	DL18TS	Hamster
19	GM10612	Hamster
20	GM10478	Mouse
21	THYB1.3	Mouse
22	PgME25NU	Mouse
X	HORL9X	Mouse
Y	853	Hamster

Table 2-1: Monochromosomal somatic cell hybrid DNA panel (MRC-RFGRC). Table shows the DNA and the rodent background of each human chromosome.

2.2. Oligonucleotides

2.2.1. Oligonucleotide Design

Primers were designed using a local copy of the Whitehead Institute for Biomedical Research primer3 programme (http://www.broad.mit.edu/cgi-bin/primer/primer3/www.cgi) (Rozen and Skaletsky, 2000), with inhouse defaults built in. Where the primers amplified coding sequences, a BLASTn similarity search was performed to check that there was no priming from related gene family members. Typically, oligonucleotides were 18-30bp long with approximately 50% GC content and a melting temperature (Tm) of between 50°C and 70°C. They were chosen to be stable, and not predicted to form hairpin structures and primer dimerism where the primers were used as a pair.

2.2.2. Oligonucleotide Synthesis

Oligonucleotides were commercially synthesised from MWG Biotech, Sigma Genosys or Invitrogen, typically at the 0.03 μmol scale with standard purification. One exception was the 224 vectorette primer which was synthesised at the 0.05 μmol scale and purified by reverse phage cartridge. If a primer pair was used for microsatellite genotyping, typically the forward primer was labelled at the 5' end with one of hex, tet or fam fluorescent tags.

2.2.3. Oligonucleotide Primer Sequences

See Appendix I for details of the STSs used in each chapter. Primer sequences, the resultant STS size, the optimal PCR cycling annealing temperature, the optimal PCR buffer, additives included in the PCR and further comments about the PCR are included.

2.3. Amplification of DNA by the Polymerase Chain Reaction (PCR)

The polymerase chain reaction allows the rapid and specific amplification of a DNA template from a highly complex mixture of DNA. The specificity is provided by oligonucleotide primers complementary to the 5' ends of the sequence to be amplified. The PCR reaction is a three step process. In the first step the DNA is denatured at a high temperature. The next step reduces the temperature to allow the oligonucleotides to anneal to the specific site on the DNA. In the final step the temperature is adjusted to allow amplification of the desired sequence by a thermostable DNA polymerase. Multiple cycles of these three steps results in exponential amplification of the desired sequence. Numerous optimisations and modifications to the basic technique have been used during this work.

2.3.1. PCR Reagents

Typically, a master mix was prepared on ice consisting of 1X PCR buffer (supplied by Roche, Applied Biosystems, Gibco BRL, NEB or Invitrogen), 1.5mM $MgCl^2$, between 100μM and 167μM of each deoxy-nucleotide (dNTP) (Sigma), 0.5-1U *Taq* DNA polymerase (*Taq* was either made inhouse or bought from Sigma) and between 0.33μM and 0.67 μM of each oligonucleotide primer. Optional additives to this basic PCR were 1X PCR Enhancer (Invitrogen) or 0.5μl PCR grade dimethyl sulfoxide (DMSO) (Sigma, cat.no. D8418). DNA was prepared on a 96 well plate or in 0.5ml eppendorfs. Typically, 20-40ng DNA was used, depending on the PCR final volume, and then the master mix was alliquotted into the tubes or plate. PCRs were run on a MJR-PTC 220 or MJR-PTC 200 thermal cycler (MJ Research). Heated lids were used to minimise evaporation.

2.3.2. PCR Cycling

PCR's consisted of an initial denaturing step of 93°C for 1-2 minutes, followed by 25-35 cycles of the three steps: denaturation, annealing and extension. The denaturation step typically lasted 5-20 seconds. The annealing step temperature is determined by the predicted Tm of the oligonucleotide primers. Primers were designed with a Tm of approximately 60°C and consequently the annealing temperature step in the PCR was typically performed 5°C below this at 55°C. The annealing step was 20-30 seconds. The extension step is carried out at 72°C, the optimal temp for *Taq* DNA polymerase. The length of extension was estimated depending on the size of the product; approximately 1 minute per 1kb of sequence was used. After the final cycle, a further elongation step of 5-10 minutes was used to ensure fragments were complete. This was increased to 50 minutes for microsatellite markers.

The 'Touch down' PCR was used to increase the specificity of the PCR reaction (Don *et al*, 1991). An initial annealing temperature is used that is 10°C above that of the final annealing temperature required. With every cycle, this is reduced by 1°C, until after 10 cycles the desired annealing temperature is reached. The normal 25-35 cycles at the desired annealing temperature can then proceed. This results in only highly specific primer-template reactions occurring during the first few cycles, thus selecting for the correct product during the remainder of the reaction. See Appendix I for individual oligonucleotide sequences with the required PCR additives and cycling conditions.

2.4. RT-PCR

2.4.1. RNA Extraction from Cells

RNA was extracted from a human lymphoblastoid cell line, using the RNeasy® Mini kit (Quiagen) as per the manufacturers instructions, and stored at -70°C.

2.4.2. cDNA synthesis and RT-PCR

cDNA was synthesised using the First Strand cDNA Synthesis kit (Boehringer Mannheim) as per the manufacturers instructions, using the $p(dN)_6$ primers. Control tubes without reverse transcriptase and without RNA were also made at the same time. RT-PCR was carried out as in Section 2.3, using 2µl RNA in a 15µl total volume PCR.

Alternatively, Universal QUICK-Clone™ cDNA II (Clontech) was used to amplify putative exons. This is a ready made human multiple tissue cDNA. The cDNA was supplied as 2ng/µl, and was diluted to 0.4ng/µl for RT-PCR. RT-PCR was carried out as in Section 2.3, using 1µl of cDNA in a 25µl total volume RT-PCR.

STSs for RT-PCR were designed as in Section 2.2 (see Appendix I for details of each primer). Where the genomic DNA product was larger than ~1-2kb, a number of different buffering systems were used at a variety of different annealing temperatures to produce product. Where the genomic DNA product size was smaller than ~1-2kb, the RT-PCR was first optimised on genomic DNA for buffering system and annealing temperature. In this case, the buffer, additive and annealing temperature of the RT-PCR are noted in Appendix I.

2.5. cDNA Library Screening

cDNA library screening was carried out at the Sanger Institute, Cambridge.

2.5.1. cDNA Libraries

Twenty-two cDNA libraries (Table 2-2) were used. Each library consisted of ~100,000 clones that were divided into five superpools of ~25,000 clones each. The 22 cDNA libraries were arranged on three 96 well plates: A primary plate of nine libraries, a secondary plate of nine libraries and a tertiary plate of four libraries.

Table 2-2 provides details of each cDNA library. Each plate contained the five superpools of each library and a positive and a negative control. This was replicated twice on each 96 well plate, enabling two STSs to be screened simultaneously.

The clones were prepared with a vectorette ligated onto the ends (Riley *et al*, 1990). A vectorette is composed of two strands of non-basepairing sequence. The sequence of each strand is the same and is in the same orientation (i.e. not reversed or complimented). The vectorette primer (224) (see Appendix I) is a 30 base pair long oligonucleotide that is identical in sequence (i.e. not reversed or complimentary) to the vectorette sequence and therefore is not able to anneal to the vectorette. The vectorette primer is used in conjunction with a specific oligonucleotide that anneals to the cDNA insert sequence. The specific oligonucleotide is required to anneal to the cDNA and amplify into the vectorette, thus creating the complementary sequence for the vectorette primer to anneal and amplify back. Therefore, the vectorette is used as a specificity regulator. However, non-specific product would be possible if the cDNA library contains sequence complimentary to the vectorette primer sequence.

Table 2-2: cDNA libraries. Twenty-two cDNA libraries were arranged on three 96 well plates: primary (P), secondary (S) and tertiary (T). The table shows which tissue each cDNA library comes from, the name of the vector, host and antibiotic used (Amount of antibiotic is quoted in μg/ml. AMP = ampicilin, TET = tetracycline.) and the source of the library.

Library Code	Plate	Description	Vector	Host	Antibiotic	Source
AK	P	Adult kidney	pcDNA3.1 [BstX1+Not1 site]	TOP10F'	AMP [50] TET [12.5]	Invitrogen
FLU	P	Fetal lung	pcDNA1 [BstX1 site]	MC1061/p3	AMP [30] TET [10]	Invitrogen
AH	P	Adult heart	pcDNA3-Uni [BstX1+Not1 site]	TOP10F'	AMP [50] TET [10]	Invitrogen
AB	P	Adult brain [temporal lobe, normal]	pcDNA3.1-Uni [BstX1+Not1 site]	TOP10F'	AMP [50] TET [12.5]	Invitrogen
HeLa	P	Cervical carcinoma	pcDNA3.1-Uni [BstX1+Not1 site]	TOP10	AMP [50] TET [12.5]	Invitrogen
T	P	Adult testis	pCDM8 [BstX1 site]	MC1061/p3	AMP [50] TET [10]	Clontech
U	P	Monocyte from a patient with promonocytic leukaemia [NOT activated].	pCDM8 [BstX1 site]	MC1061/p3	AMP [12.5] TET [7.5]	D. Simmons
HPB	P	HPBall [T cell]	pπH3M/pCDM8 [BstX1]	MC1061/p3	AMP [12.5] TET [7.5]	D Simmons
SK	P	SK-N-MC cell line [neuroblastoma, metastasis to sub-orbital area]	pcDNA1 [BstX1 site]	MC1061/p3	AMP [30] TET [10]	Invitrogen
HIS	S	Small intestine	pcDNA3 [BstX1 site]	XL-10 Gold	AMP [50] TET [10]	M Stammers
HL	S	HL60 cell line [peripheral blood]	pcDNA [BstX1 site]	MC1061/p3	AMP [30] TET [10]	Invitrogen
FL	S	Fetal liver	pcDNA1	MC1061/p3	AMP [30]	Invitrogen

			[BstX1 site]		TET [10]	
FB	S	Fetal brain	pcDNA1 [BstX1 site]	MC1061/p3	AMP [30] TET [10]	Invitrogen
ALU	S	Adult lung	pcDNA1 [BstX1+Not1]	MC1061/p3	AMP [50] TET [10]	Clontech
H9	S	Full term placenta, normal pregnancy	pπH3M/pCDM8 [BstX1]	MC1061/p3	AMP [12.5] TET [7.5]	D Simmons
Dau	S	Daudi [B lymphoma]	pπH3M/pCDM8 [BstX1]	MC1061/p3	AMP [12.5] TET [7.5]	D Simmons
YT	S	HTLV-1 +ve adult leukaemia T cell	pπH3M/pCDM8 [BstX1]	MC1061/p3	AMP [12.5] TET [7.5]	D Simmons
BM	S	Bone marrow	pπH3M/pCDM8 [BstX1]	MC1061/p3	AMP [12.5] TET [7.5]	D Simmons
NK	T	Natural killer T cell	pπH3M/pCDM8 [BstX1]	MC1061/p3	AMP [12.5] TET [7.5]	D Simmons
Uact	T	Monocyte form patient with promonocytic leukaemia [activated]	pπH3M/pCDM8 [BstX1]	MC1061/p3	AMP [12.5] TET [7.5]	D Simmons
DX3	T	Melanoma	pπH3M/pCDM8 [BstX1]	MC1061/p3	AMP [12.5] TET [7.5]	D Simmons
P	T	Pfizer adult brain	pcDNA1 [BstX1 site]	MC1061/p3	AMP [12.5] TET [7.5]	Pfizer (private)

2.5.2. cDNA Library Screening

2.5.2.1. Pre-screen

STSs were designed as in Section 2.2 (see Appendix I for details of each STS). PCR with a pair of specific primers confirmed if the STS was represented as a cDNA. The primary plate was screened first, and if no positive libraries were identified, the secondary and tertiary plates were screened. The aim was to identify three positive superpools, preferably from different libraries. PCR was carried out as in Section 2.3, using 1X NEB PCR buffer and including 0.2µl of 1/20 β mercapto ethanol (BDH) and 0.25µl 10mg/ml bovine serum albumin (BSA) (NEB, cat.no. A7517) in a 15µl total volume PCR. The PCR cycling conditions were as follows: 93°C for 1 minute, followed by 10 cycles of 93°C for 15 seconds, 65°C for 30 seconds (- 1 second per cycle) and 72°C for 45 seconds. This was followed by 30 cycles of 93°C for 15 seconds, 55°C for 30 seconds and 72°C for 45 seconds. Finally, an extension step of 72°C for 10 minutes. The product was run on a 2.5% 0.5X TBE agarose gel and visualised under UV light.

2.5.2.2. Vectorette PCR

A vectorette PCR was performed with each STS primer and the vectorette primer (224) from the pre-screen on three positive superpools. Vectorette PCR was performed as in Section 2.3 except 0.12µl of each of Perfect Match, *Taq* extender and Advantage-2 *Taq* (Clontech) was used instead of *Taq* DNA polymerase and 0.2µl of 1/20 β mercapto ethanol and 0.5µl 5mg/ml BSA were added in a 15µl total volume PCR. The PCR cycling conditions were as follows: 95°C for 2 minutes, followed by 17 cycles of 94°C for 5 seconds, 65°C for 30 seconds and 72°C for 3 minutes. This was followed by 18 cycles of 94°C for 5 seconds, 60°C for 30 seconds and 72°C for 3 minutes. Finally, an extension step of 72°C for 10 minutes. The product was run on a 2.5% 0.5X TBE agarose gel and visualised under UV light.

2.5.2.3. Nested PCR

The vectorette PCR was also amplified with nested primers, designed as described in Section 2.2 (see Appendix I for details of each nested primer). This was to increase specificity of the product. A 1:100 dilution of the vectorette PCR was used as a template, and PCR was performed as in Section 2.3, with 1X Perkin Elmer PCR buffer and 1X PCR enhancer (Invitrogen) in a 15-25µl total volume PCR.

2.6. Agarose Electrophoresis

2.6.1. Solutions

20X TBE	1.8M Tris.HCL, 40mM EDTA, 1.8M boric acid, pH 8.0
50X modified TAE	40mM Tris-acetate, 0.1 mM Na_2EDTA, pH 8.0
10X DNA loading buffer	20% ficoll (Sigma), 100mM EDTA, orange G (Sigma)

2.6.2. Size Markers

The marker used depended on the expected band sizes to be resolved. Typically 250-500ng of DNA was loaded per marker lane and at least two marker lanes were run per gel to check for even migration across the gel.

Size markers

1. lamda DNA digested with Hind III (Boehringer Mannheim).
2. Ready Load™ 1 kb DNA ladder (Invitrogen).
3. Ready Load™ 100 bp DNA ladder (Invitrogen).
4. Ready Load™ PhiX174 RF DNA /HaeIII fragments (Invitrogen)

2.6.3. Agarose Gel Electrophoresis

Size fractionation of DNA was achieved by migration through agarose gels. Agarose was dissolved in 0.5X TBE or 0.5X modified (low salt) TAE buffer by heating. The

final concentration of agarose depended on the size of DNA fragments to be resolved: lower concentrations for larger fragments. Most PCR products were run on 1-2% agarose gels. 1µl per 100ml of buffer of 10ng/µl Ethidium bromide (Sigma) was added to all gels to stain the DNA. Ethidium bromide intercalates between bases of nucleic acid and fluoresces upon stimulation by UV light. 1X loading buffer was added to the DNA prior to loading. Gels were placed in an electrophoresis tank of appropriate size and covered with buffer to match the gel. An electrical current of 50-120 V was passed through the buffer, resulting in migration of the negatively charged DNA and loading dye towards the anode. DNA fragments were visualised on a UV transilluminator and photographed.

2.7. Purification and Concentration of DNA

2.7.1. PCR Purification

Prior to both SNaPshot^{TM} genotyping and sequencing of PCR products it was necessary to remove excess primers and dNTPs from the PCR. Therefore, the PCR was treated with EXOSAP-IT (USB), a mixture of two exonucleases. The enzyme exonuclease 1 (EXO) degrades the unincorporated oligonucleotides remaining in the reaction and the enzyme shrimp alkaline phosphatase (SAP) removes any unincorporated dNTPs. Typically, 1-2µl of enzyme was used per 5µl PCR. If the PCR was being sequenced, 10-40ng of the PCR was used in a final volume of 5µl. The PCR concentration was estimated by running an aliquot of the PCR on an agarose gel with 250ng and 500ng of PhiX174 RF DNA /HaeIII fragments marker. The enzyme and PCR mix was incubated on a hybaid thermocycler at 37°C for 1 hour, and the enzymes were inactivated by heating for 20 minutes at 80°C.

2.7.2. Purification of DNA from Agarose Gels

subsequent to RT-PCR, product that was to be sequenced was first excised from an agarose gel and purified in one of the following ways.

2.7.2.1. Spin Columns

RT-PCR product was excised from a 1.2% 0.5X Modified TAE (Millipore) agarose gel. The gel slice was then placed in an Ultrafree®-DA gel extraction column (Millipore) and centrifuged at 5000 g for 10 minutes on a Biofuge Fresco desk top centrifuge. The resulting product was ready for direct sequencing.

2.7.2.2. Filter Tip Purification

RT-PCR product was excised from a 1.2-2% 0.5X TBE agarose gel. The gel slice was then placed at the top end of a 200µl filter tip (Rainin), placed in a 1.5ml eppendorf tube and spun at 10,000 g for 15 minutes on a Biofuge Fresco desktop centrifuge. A 0.025µM drop dialysis filter (Millipore) was placed on the surface of a pool of sterile water in a petri dish. The eluted PCR product was placed as a drop onto the surface of the filter and left for 1 hour. The concentration of salts in the sample equilibrates with the water, but the filter prevents the movement of macromeolecules across the surface. The drop was then removed by pipetting.

2.7.2.3. Gel Purification Kit

RT-PCR products obtained from the vectorette PCR (Section 2.5.2.2.) were excised from a 2.5% TBE agarose gel. The gel slice was added to 30µl sterile water and left overnight so that product leached out into the water. This was used as a template for a second round of PCR. Product from this second round was excised from a 2.5% 0.5X TBE agarose gel. The DNA was extracted from the gel slice using the QIAquick Gel Extraction Kit (Qiagen) as per the manufacturer's instructions.

2.7.3. Ethanol Precipitation

Ethanol precipitation was used subsequent to sequencing products obtained from a

PCR. A 1/5 volume of the PCR of 25mM EDTA and 2-3 volumes of 95-99.9% ethanol (Fisher analysis grade) were added to the PCR to give a final ethanol concentration of 67-71%. An optional 1/20 volume of pellet paint was also added. This was incubated at room temperature for 15 min. If the precipitation was carried out in 0.5ml eppendorf tubes, the mixture was centrifuged at 10,000 g on a Biofuge Fresco desk top centrifuge for 40 minutes. If precipitation was carried out in a 96 well plate, the mixture was centrifuged at 10,000 g for 30-40 minutes on a Jouan CR 422 centrifuge. The supernatant was removed from eppendorfs by pipeting or from 96 well plates by inversion of the plate on a tissue and careful tapping. The pellet was washed once or twice with 70% ethanol. In eppendorfs, 500µl of 70% ethanol was added to the pellet, which was left at room temperature for five minutes and then centrifuged at 10,000 g for five minutes. In a 96 well plate, 30µl of 70% ethanol was added to the pellet and the plate was centrifuged for 15 minutes at 10,000 g. The supernatant was then removed from eppendorfs by pipeting or from 96 well plates by inversion of the plate on a tissue and careful tapping. If a second wash was performed the process was repeated. If precipitation was being carried out in a 96 well plate, a final gentle spin of the inverted plate onto a tissue, by letting the centrifuge reach 3000 g, removed the last of the ethanol not removed by inversion of the plate alone. Pellets were left to air dry in the dark.

2.8. Sequencing

2.8.1. Sequencing PCR Products

Sequencing was carried out using dye terminator chemistry. Dideoxy-nucleotides (ddNTPs) terminate the sequencing reaction at each base to generate a ladder of fragments. Each of the four ddNTPs is labelled with a different dye, enabling a computer to read the sequence. The sequencing reactions were performed using the BigDye® Terminator v3.1. cycle sequencing kit (Applied Biosystems). Typically, 10-40ng DNA, 3.2µM of the appropriate primer and 0.5-1µl of the BigDye terminator cycle sequencing kit was used in a total volume of 10µl. The pGEM

control vector supplied in the kit was sequenced with the M13 primer to control for the quality of template.

The cycle sequencing reaction was performed on a hybaid DNA thermal cycler. The cycling conditions were: 96°C for 1 minute and then 25 cycles of 96°C for 10 seconds, 50°C for 10 seconds and 60°C for 4 minutes.

Sequences were purified by ethanol precipitation (Section 2.7.3) and sequenced on an ABI PRISM® 377 or 3730 Genetic Analyser by Alison Condie. The output file is directly amenable to computer analysis. Multiple sequence chromatograms were aligned using the phredPhrap software and visualised with the Consed programme. Iindividual sequence chromatograms were visualised with the Chromas programme.

2.9. Genotyping

All ABI PRISM® gel electrophoresis was performed by Alison Condie at the Welcome Trust Clinical Researach Facility, Western General Hospital, Edinburgh. For both SNP and microsatellite genotyping, the capillary system (ABI PRISM® 3730 or 3100 Genetic Analyser) was used in preference over the slab gel system (ABI PRISM® 377 Genetic Analyser). The capillary system means that the individual sample fluorescence cannot be distorted by neighbouring samples as it can on a slab gel.

2.9.1. SNP Genotyping

2.9.1.1. SNaPshot™

SNP oligonucleotides were designed to the 16-24bp of sequence immediately flanking the SNP. This consequently restricts the choice of oligonucleotide to either the forward or reverse orientation. The decision of which orientation to choose was based on the folding characteristics of the PCR template and the SNP oligonucleotide. The *m*fold (http://www.bioinfo.rpi.edu/applications/mfold/old/dna/)

programme (Zuker, 2003) at the Bioinformatics Centre at Rensselaer and Wadsworth makes predictions about how a length of DNA is likely to fold up at different temperatures, and allows an assessment of the accessibility of the SNP oligonucleotide to the PCR template. The decision to design the forward or reverse orientation was made based on this.

The SNaPshot™ ddNTP primer extension lit (PE Biosystems) includes DNA polymerase and fluorescently labelled ddNTPs, each labelled with a different dye. The SNP oligonucleotide anneals to the DNA template and an extension reaction of one base pair takes place, the ddNTP blocking the addition of further nucleotides. Therefore, for example, an adenine nucleotide on the template DNA strand would result in the addition of a tyrosine ddNTP to the SNP oligonucleotide, which fluoresces red. A homozygote would be represented by a single dye peak, a heterozygote by two dyes peaks.

SNP oligonucleotides were designed to be multiplexed together. A size difference of four base pairs could be distinguished on a polyacrylamide gel and therefore, oligonucleotides were designed to be 16, 20 or 24 bp in length. In addition, SNPs that gave complementary genotypes could be distinguished by colour. Therefore, a maximum of six SNP oligonucleotides could be multiplexed in one reaction. Reactions were multiplexed after the SNaPshot™ reaction to prevent differential amplification of the SNP oligonucleotides during the SNaPshot™ reaction.

Typically, 2μM of oligonucleotide was used with 2μl of SNaPshot™ mix and 3μl of the PCR template in a final volume of 10μl. Thermal cycling of SNaPshot™ reactions increases signal intensity and decreases sensitivity to reaction conditions. Typically, 25-30 cycles were performed with a short 10 second denaturation step at 95°C, an annealing step of 50°C for 5 seconds, and an extension step at 60°C for 30 seconds. The SNapShot™ reaction was treated with 1μl SAP enzyme (USB) to remove unincorporated ddNTPs, incubated at 37°C for 1 hour and inactivated at 80°C for 15 minutes. A 1:15-1:40 dilution of the product underwent polyacrylamide gel electrophoresis on an ABI PRISM 310 or 3730 Genetic Analyser, and was

analysed using the GeneScan version 3.0, or the GeneMapper version 3.0, software. Genotypes were scored blind to phenotype on two separate occasions.

DNA pooling was used as a means of dramatically reducing the number of genotyping reactions required to perform association studies (Section 2.1.3.2.). Genotyping was performed using the SNaPshot™ method. It was very important to obtain the optimal dilution for the pools before running the sample on an ABI PRISM platform. This is because an overloaded lane, where the signal exceeds the maximum fluorescence readable, reduces the reliability of the results. Therefore, a test run was performed, on a few control DNAs, to determine the dilution factor. The same PCR template was used to perform a second SNaPshot™ reaction on the pools plate.

2.9.1.2 TaqMan®

The probes and primers used in the TaqMan® genotyping assays were designed by Applied Biosytems and the genotyping was carried out by A. Condie at the Welcome Trust Clinical Research Facility, Western General Hospital on an ABI PRISM® 9700 sequence detection system.

The TaqMan® method requires a primer pair and an internal probe. The internal probe consists of two types of fluorophores. The quencher fluorophore, on the 3' end of the probe, reduces the fluorescence from the reporter fluorophore, on the 5' end of the probe. The probe and the primers anneal to the DNA and *Taq* DNA polymerase adds nucleotides to the primers and removes the probe from the template DNA. This separates the quencher from the reporter, and allows the reporter to emit its energy. The two alleles of a SNP are tagged with different coloured reporters. Therefore detection of one colour in the assay represents a homozygous genotype and the detection of an equal amount of both colours represents a heterozygous genotype. The colour reported is plotted on a graph of the colour spectrum. Individuals with either colour, or an equal mix of the two, cluster together. Samples that have been

contaminated, or in which the DNA quality is poor are often seen as an outlier and the data can be discarded.

2.9.1.3. Sequenom®

A number of SNPs were genotyped by the Sanger Institute using the Sequenom® MassARRAY™ system. A PCR is performed of the region surrounding the SNP, and this is used as a template for a primer extension reaction containing the primer, DNA polymerase, a mix of dNTPs and one of the four possible ddNTPs. This generates allele specific extension products that are generally 1-4bp longer than the original primer. For example, a SNP with variants cytosine/tyrosine would include a guanine ddNTP in the reaction mix. This would stop the primer extension reaction at either the cytosine variant of the SNP, or the next cytosine nucleotide in the sequence. Genotypes are determined based on molecular weight using Mass Spectrometry (MALDI-TOF). Homozygotes would generate one of two molecular weights depending on the allele, and heterozygotes would generate both molecular weights.

2.9.2. Microsatellite Genotyping

Di- and tri-nucleotide repeats were identified by the repeat finder programme Sputnik (http://espressosoftware.com/pages/sputnik.jsp) developed by Chris Abajian (The University of Washington department of molecular biotechnology). The results of this are displayed in the inhouse database ACeDB. Oligonucleotide primers were designed as in Section 2.2 to flank the repeat and the 5' end of the forward primer was labelled with a fluorescent tag. After PCR, a dilution of the PCR was run on the ABI PRISM® 3730 or 3100 Genetic Analyser. The product was run with HiDi™ formamide (Applied Biosytems) with either the GeneScan®-500 [Rox]™ or [tamra]™ size standard (Applied Biosystems) (Table 2-3). Microsatellite markers were multiplexed based on STS size and fluorescent tag. Results were analysed using the GeneScan version 3.0 or the GeneMapper version 3.0 analysis software.

Genetic Analyser	PCR Dilution	Sample Preparation	Size Standard
3730	1:100	10μl HiDi formamide 1μl sample dilution	0.06μl
3100	1:10-1:40	10μl HiDi formamide 1μl sample dilution	1μl

Table 2-3: ABI PRISM® 3730 and 3100 Genetic Analyser sample dilutions.

A characteristic of microsatellite genotyping can be stutter peaks and the addition of an adenine to the end of the product. Stutter peaks occur when the *Taq* DNA polymerase slips as it moves along the repeat, missing out one or more of the repeat blocks and therefore producing product that is one or more repeat blocks smaller than the actual allele. The real allele is usually the largest and the strongest signal. Therefore, microsatellite markers were genotyped on a minimum of 20 individuals to aid the correct interpretation of the marker characteristics. The *Taq* DNA polymerase also tends to add an adenine onto the end of the PCR product, making the product one base pair larger where this occurs. To ensure the uniform addition to all PCR products, all microsatellite PCR reactions included a final 72°C extension step of 50 minutes. See appendix I for primer sequences, PCR mix and PCR cycling conditions. Genotypes were scored blind to phenotype on two separate occasions.

2.10. Bacterial Cell Culture

2.10.1. Solutions

Luria-Bertani (LB)-broth	1% (w/v) Bacto-Tryptone (Difco); 0.5% (w/v) Bacto-yeast extract (Difco); 0.1% (w/v) NaCl; pH 7.
LB plates	LB-broth with 15g of Bacto Agar/litre of liquid media.
Freeze Solution	15% glycerol in LB-broth with Mg.

2.10.2. Bacterial Cell Culture and Colony PCR

Five BACs were used for colony PCR. BACs RP11-264E23, RP11-751L19 and RP11-626O20 were stored at -70°C as a glycerol stock inhouse. CTD-2205P10 was

obtained from the Research Genetics (Invitrogen) as a glycerol stock. RP11-180A12 was obtained from the Sanger Institute as an agar stab. These two clones were grown on a plate containing L-agar and 12.5μg/ml chloramphenicol antibiotic and cultured at 37°C overnight. Colonies were picked with a cocktail stick and twirled in 15ml sterile water. This was used as a template for colony PCR. Colony PCR was performed using 3μl of the template in a 15μl PCR, as described in Section 2.3. The cocktail stick was then cultured in a 25ml falcon tube in 3ml LB-broth with Magnesium, and incubated at 37°C overnight. Glycerol stocks were made by spinning down 1ml of the culture at 10,000 g for 1 minute, to form a pellet. The supernatant was removed and was re-suspended in 1ml freeze solution and stored at -70°C.

2.11. Bioinformatic Tools

Websites that are referred to during the thesis are listed here.

ACeDB	www.acedb.org/
Analyze-it® for excel	www.analyze-it.com/
CEPH	www.cephb.fr
Chromas	www.technelysium.com.au/chromas.html
CpG island	www.rfcgr.mrc.ac.uk/Registered/Help/alfresco/#cpg
dbSNP	www.ncbi.nlm.nih.gov/SNP/
DNA clone sequences	www.ncbi.nlm.nih.gov/
Ensembl	www.ensembl.org
EST database	www.ncbi.nlm.nih.gov/dbEST/
Fgenes	www.softberry.com
GenScan	genes.mit.edu/GENSCAN.html
International HapMap project	www.hapmap.org/
*M*fold	http://www.bioinfo.rpi.edu/applications/mfold/old/dna/)
MRC-RFCGR	www.rfcgr.mrc.ac.uk
MZEF	sciclio.cshl.org/genefinder/
NCBI	www.ncbi.nlm.nih.gov/
NetPhos	www.cbs.dtu.dk/services/NetPhos/
Primer3	http://www.broad.mit.edu/cgi-bin/primer/primer3/www.cgi
PROSITE	www.expasy.org/prosite

RepeatMasker	repeatmasker.genome.washington.edu/
Sanger Institute	www.sanger.ac.uk/HGP/
Sputnik	cbi.labri.fr/outils/Pise/sputnik.html
Stack	www.sanbi.ac.za/Dbases.html
SwissProt	www.ebi.ac.uk/swissprot/index.html
Tigr	www.tigr.org/
UCSC	www.genome.ucsc.edu
Unigene	www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene

Chapter Three

Contig Mapping in Minimal Region One

Contig Mapping in Minimal Region One

3.1 Introduction

In the early stages of the Human Genome Project (HGP), sequence information was sparse. The group required a reliable physical map in order to proceed with recombination breakpoint mapping, gene identification and to be able to direct and interpret future association studies. A BAC/PAC contig of Minimal Region One (MR1), which comprises the overlap between the disease-associated haplotypes of three families (Figure 3-1), was constructed by a combination of STS content analysis and restriction fragment fingerprinting (Evans *et al*, 2001_a). The contig was displayed inhouse using the SAM (system for assembling markers) database (Soderlund, 1995), a programme to aid marker ordering. The clone coverage of the inhouse contig was more extensive than the publicly available map. In addition, the results were regarded as more reliable because the degree of experimental input for this small region outweighed that of a comparable region for the HGP.

Clones that are being sequenced as part of the HGP are displayed in an inhouse ACeDB database (www.acedb.org/). ACeDB is a genome database, developed by J. Thierry-Mieg (CNRS, Montpellier) and R. Durbin (Sanger Institute) in 1989, that was originally used in the C.elegans genome project, from which its name was derived (A C. elegans DataBase). The database is constructed and managed inhouse by S. Morris and displays the sequence of the clones. It is then masked for simple repeats and CpG islands, and annotated with the results of BLASTn similarity searches to human clones, vertebrate and invertebrate genomic DNA, vertebrate mRNA and protein, and the results of the gene and exon prediction programmes GenScan, FGenes and MZEF. In addition, it holds inhouse data. This gives the group a central resource with which to enter and manage contig, microsatellite, SNP and other data generated in the group, and also to study existing genes and other features of the genomic landscape, including novel transcript evidence. The ordering of the clones is not just a reproduction of the HGP, but the available data is inspected at a

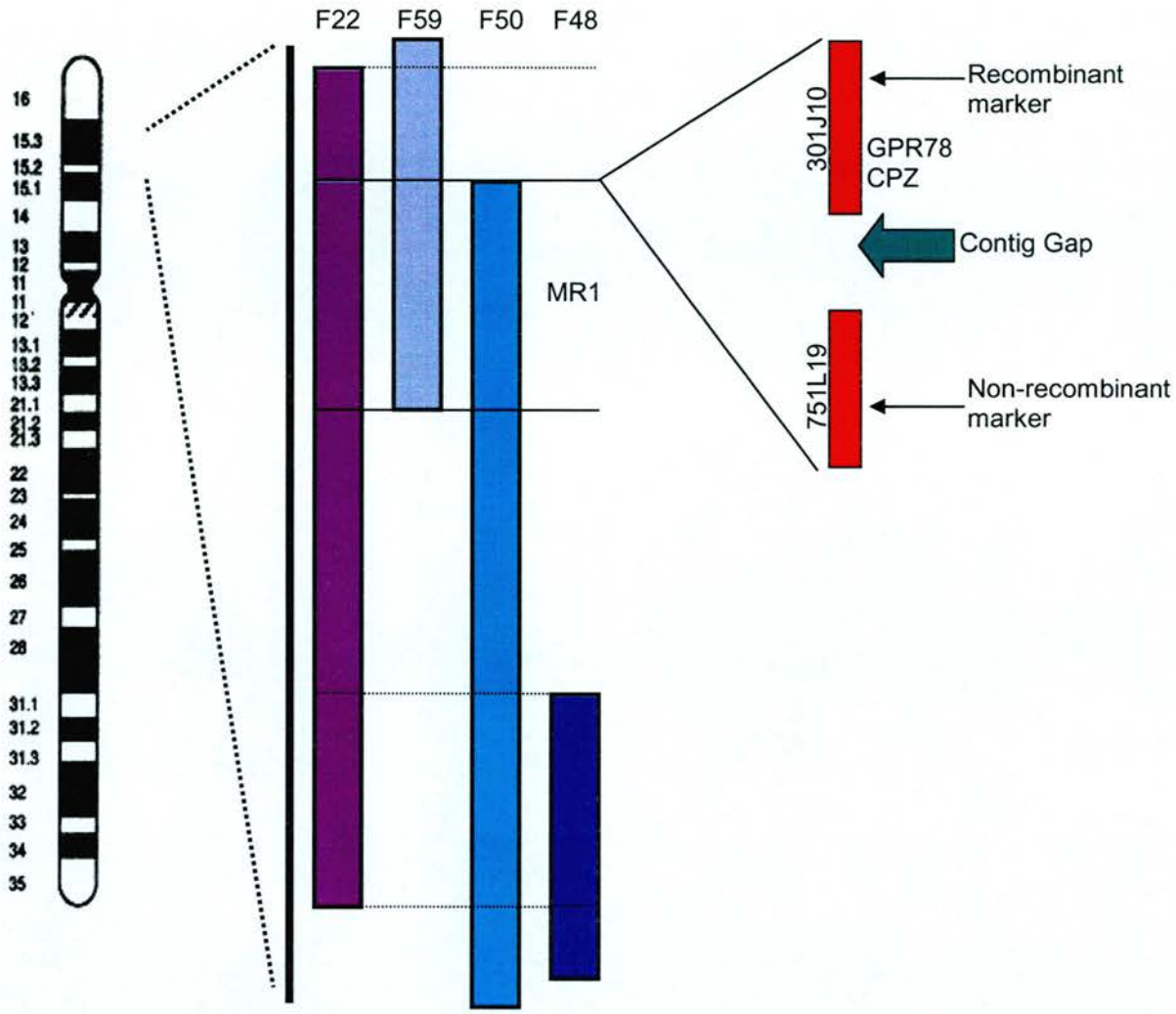


Figure 3-1: Overlapping haplotypes of families 22, 59, 50 and 48. The coloured bars represent a marker haplotype on chromosome 4p inherited with psychiatric illness in each family (the position of the centromeric recombination breakpoint in F50 has not been determined). The family number is marked along the top. The four haplotypes overlap, enabling the delineation of smaller regions of interest (horizontal lines). The solid horizontal lines mark the delineation of Minimal Region One (MR1). The telomeric recombination breakpoint region of MR1, defined by a member of family 50, is expanded. The red bars represent the BAC clones (RP11-) containing the two markers defining the recombination breakpoint (arrows). The region contains a contig gap (green arrow) and two genes: orphan G-protein-coupled receptor 78 (GPR78) and Carboxypeptidase Z (CPZ).

detailed level to iron out any inconsistencies and give a better quality of data than is available publicly for our small region of interest. This was essential in the early stages of the project, but has become less so as the HGP moves towards a finished product. However, even today, there still remain small regions of incomplete sequence.

The physical map and the sequence available at the start of my project, both from our inhouse contig and the HGP, had numerous sequence and contig gaps. Over the three years many of these have been closed. Both minimal regions currently have four gaps (February 2004). One of the gaps in MR1 lies in between the existing recombinant and non-recombinant markers defining the telomeric end of the disease haplotype in F50. This is also the boundary which defines the telomeric end of MR1. This gap was also observed in the inhouse contig and estimated to be ~300kb by fluorescence in situ hybridisation (FISH) analysis (Evans *et al*, 2001_a). The HGP has not closed the gap either, and in fact the clone map published around the gap changed significantly with every release of the genome browser at the University of California, Santa Cruz (UCSC) (genome.cse.ucsc.edu). The region on the telomeric side of the gap is not represented in a number of different libraries (Evans *et al*, 2001_a) and the sequence at the centromeric side of the gap appears to be a highly repetitive region. For this reason the in-house contig around this interval was not considered to be reliable. However, closing this gap remains an important aim due to the presence of the recombination breakpoint. At the start of my project the recombinant marker lay on clone RP11-301J10 and the non-recombinant marker was positioned on clone RP11-751L19 (Figure 3-1).

The December 2001 UCSC Golden Path release contained four clones between RP11-751L19 and the gap (Figure 3-2; A). These overlapped with each other but not with RP11-751L19. This included two clones not seen again in later releases: RP11-423D16 was removed and is still unfinished, and RP11-30E22 is also still unfinished and has now been localised to chromosome 11.

The April 2002 UCSC Golden Path release inserted ~9.3 million base pairs in between RP11-751L19 and the gap (data not shown). This was then removed from the next release in June 2002 (Figure 3-2; B). The June 2002 release included two fully sequenced clones, three partially sequenced clones and two unfinished clones extending into the gap from RP11-751L19. All of these clones appear to overlap with each other and RP11-751L19.

I started working with this region when the November 2002 UCSC Golden Path release became available (Figure 3-2; C). The release contained five fully sequenced clones extending into the gap from RP11-751L19, with two unfinished clones spanning one gap and introducing two new sequence gaps. It contained all the same clones as the June 2002 release, except in a different order and with two additional gaps.

The April 2003 UCSC Golden Path release (see Figure 3-2; D), released after I had started to work on this region, improved upon the previous release. It displayed three of the same fully finished clones from the last release in essentially the same position. However, these three clones now appeared to overlap and a fourth clone replaced CTD-2205P10. This resembled the June 2002 release except that CTD-2205P10 had been replaced and three clones have been removed. Two of the BACs positioned in the gap by the April 2003 UCSC Golden Path release, RP11-264E23 and RP11-180A12, had been previously screened with markers and placed in this region as part of our inhouse contig construction. However, the markers used in this screening also hybridised to other chromosomal regions and therefore could not be relied upon.

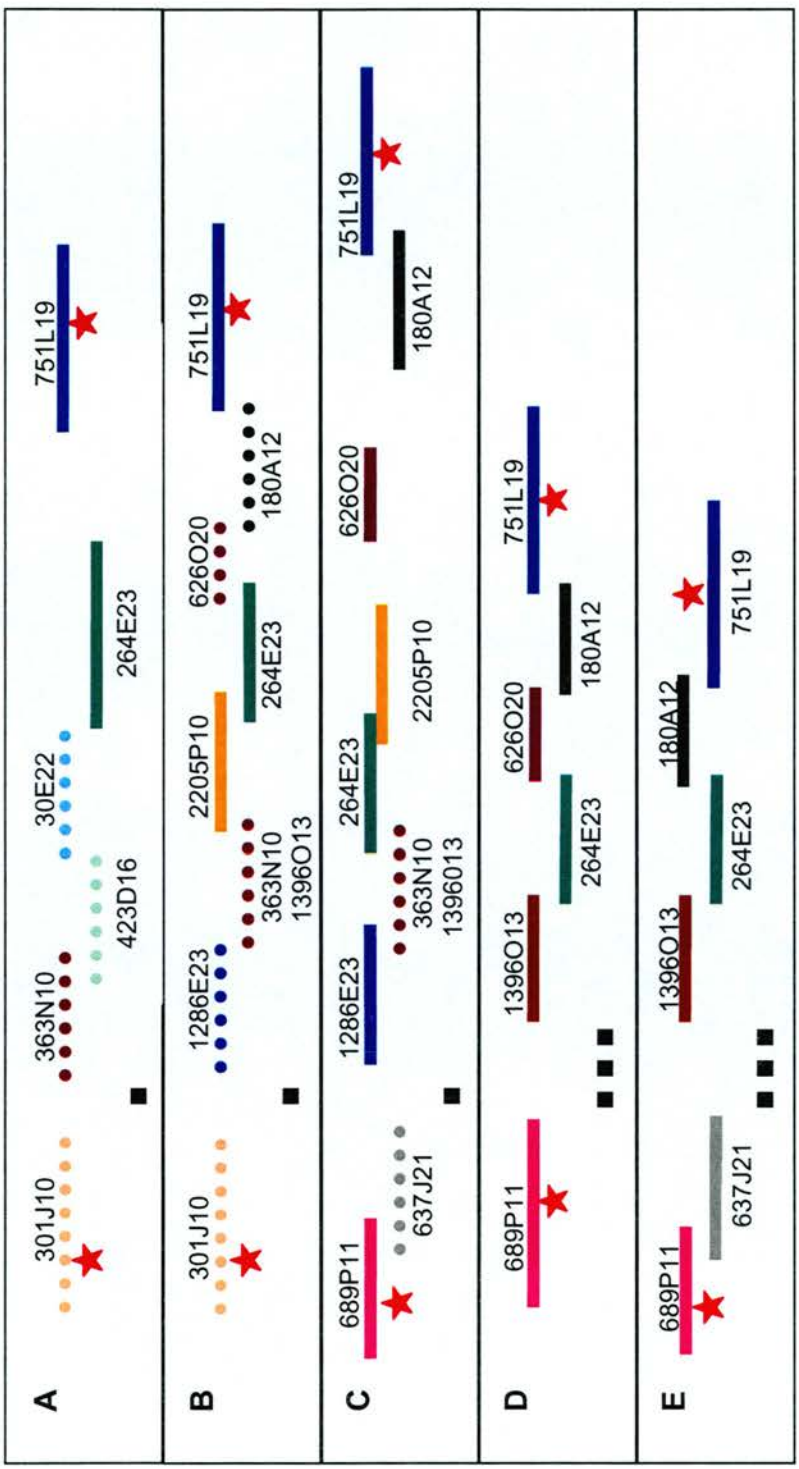


Figure 3-2: Five subsequent contig assemblies of the human genome working draft sequence at the University of California, Santa Cruz (genome.cse. uscs.edu) in the region of the telomeric recombination breakpoint defining Minimal Region One (MR1). Red stars: recombinan and non-recombinant markers defining MR1. Black dotted line: contig gap (each dot = ~50kb). Coloured bars: sequence clones (solid line = finished sequence, dotted line = unfinished sequence). See Table 3-1 for clone details. **A:** December 2001 **B:** June 2002 **C:** November 2002 **D:** April 2003 **E:** July 2003.

Figure 3-3: Contig assembly in the region of the telomeric recombination breakpoint defining Minimal Region One from the inhouse sequence database ACeDB. Contig is based on the November 2002 human genome working draft sequence at the University of California, Santa Cruz (genome.cse.ucsc.edu). Black dotted line: contig gap (each dot = ~50 kb). Coloured bars: sequence clones. See Table 3-1 for clone details.

Clone ID	Sequence ID	Status at February 2004
RP11-751L19	AC098976	Finished
RP11-180A12	AC105916	Finished
RP11-626O20	AC118279	Finished
RP11-264E23	AC097493	Finished
CTD-2205P10	AC117179	Finished
RP11-1396O13	AC116655	Finished
RP11-363N10	AC022770	32 unordered pieces
RP11-1286E23	AC108519	Finished
RP11-301J10	AC007104	Replaced with RP11-689P11
RP11-689P11	AC105345	Finished
RP11-637J21	AC116643	Finished
RP11-423D16	AC096572	16 unordered pieces
RP11-30E22	AC073507	Ch11: 38 unordered pieces

Table 3-1: Full details of the sequence clones located in the telomeric recombination breakpoint region that defines Minimal Region One. Clones form part of the contig from the human genome working draft sequence at the University of California, Santa Cruz (gneome.cse.ucsc.edu). Table shows full clone name, sequence accession number and current sequencing status.

I started working on this region after S. Morris had used the November 2002 release to construct the ACeDB contig (Figure 3-3). The current resolution of the telomeric recombination breakpoint of MR1 had reached an impasse, as all putative microsatellite repeats from the clones within the recombinant region had been genotyped. The interval contained two known potential candidate genes and possibly other genes. It was important to identify all the genes in our region of interest in order to carry out further studies on them, such as association analysis. Therefore, not only did these four clones provide a resource of novel polymorphic repeats to refine the breakpoint, but they also provided a potential resource from which to identify novel genes. However, the nature of this region and the variability of the publicly available sequence assembly lead me to carry out experimental work to confirm their locations before they were used to map transcripts and to refine the telomeric recombination breakpoint interval of MR1.

3.2 Bioinformatic Analysis

3.2.1. Searching for Bioinformatic Evidence of Clone Overlap

Clone RP11-751L19 contains single copy chromosome 4 sequence. If clone RP11-180A12 overlaps clone RP11-751L19, it may be that this overlap extends into this chromosome 4 specific region. This overlap would not be displayed on the UCSC Golden Path assembly. At the National Centre for Biotechnology Information (NCBI) (www.ncbi.nlm.nih.gov/) it is possible to obtain the whole sequence of a clone. These are in fragments if the sequence is incomplete. Each successive version of a clone sequence at NCBI tends to be trimmed as clones are overlapped with their neighbours. Therefore previous versions can be considerably longer than the full length contiguous version. Neighbouring clones are trimmed to a standard 1000bp overlap and a genuinely longer overlap is lost because it is redundant data. However, for my purposes, a longer overlap could be extremely useful since it might show that clones RP11-180A12 and RP11-751L19 overlap in the single copy sequence region.

Previous sequence versions of clones are also available from NCBI. The latest version of RP11-180A12 is sequence AC105916.4. This constitutes one continuous sequence of length 104,269bp. The previous version, sequence AC105916.3, constitutes two unordered pieces of total length 143,787bp. The ~40kb that has been lost between versions three and four may include sequence that overlaps with RP11-751L19. I took 60bp of sequence from the beginning and end of each fragment from AC105916.3 and tested it for a match to RP11-751L19 in ACeDB. I did not find a match. However, the end of one fragment did match part of RP11-264E23. This suggested that RP11-180A12 overlaps with RP11-264E23, but it does not anchor these clones to this specific locality of chromosome 4 because they do not definitely overlap with RP11-751L19.

3.2.2. The Genomic Landscape

In order to localise the clones relative to each other and to chromosome 4p I carried out bioinformatic analysis to identify single copy chromosome 4p sequences. This information could then be used to direct the development of chromosome 4 specific STSs, which could then be used to confirm overlaps between clones.

The clone RP11-751L19 was known to be in the correct local position on chromosome 4 since this clone contains a number of chromosome 4 specific genes. It was the clones telomeric to RP11-751L19 into the contig gap (Figure 3-3) that I wanted to determine the position of. Analysis of the region carried out by S. Morris, revealed that the first 43kb of clone CTD-2205P10 contains a chromosome 4p specific repeat identified by Kogi *et al* (1997). They identified cosmid clone, CRS447, and located it to chromosome 4p15.1-15.3 by FISH and human CHO-hybrid cell panel mapping. This clone consists of 4.7kb units repeated in tandem, and it was estimated that there are 50-70 repeats units per haploid genome. This was good evidence for CTD-2205P10 being positioned in the correct chromosomal region, but not definitive evidence for it being correctly positioned in the immediate local environment. A 4.7kb repeat unit repeated 50-70 times gives an estimated size of

235-329kb. Therefore, approximately 192-286kb of CRS447 repeat containing DNA does not appear in the public database. CTD-2205P10 is positioned adjacent to the gap, and if this is correct then presumably the remaining repeat blocks are located in the gap.

ACeDB displays the results of BLASTn matches to clones in the HGP that have a identity of greater than 99%. Historically, this was a useful tool to identify overlapping clones that had not been overlapped in the HGP. However, it therefore also displays fragments of clones from other chromosomes that are also greater than 99% identical to chromosome 4. Therefore, I used this feature of ACeDB to detect regions of the chromosome 4 clones that display greater than 99% identity to clones on other chromosomes. Clone CTD-2205P10 displayed regions of homology to chromosome 8 clone RP11-354G4, and to two clones on chromosomes 11: CTD-2313N18 and RP11-757C15. Clone RP11-264E23 also showed regions of homology to the same two clones on chromosome 11 and in addition clone RP11-684B2 on chromosome 11 and clone RP11-489M13 on chromosome 4p16.3; a different region of chromosome 4. Clone RP11-180A12 showed regions of homology with clones RP11-540E4 and RP11-354G4 on chromosome 8 and clone RP11-167J8 on chromosome 11. This sequence similarity supported the findings that this region is highly repetitive.

Whilst ACeDB displays regions of 99-100% similarity, it does not display clones from other genomic regions that are less that 99% similar, nor does it display regions that are unique to chromosome 4. In order to design markers to confirm overlaps between adjacent clones I used BLASTn sequence similarity searches against the HGP draft sequence of clones CTD-2205P10, RP11-264E23, RP11-606O20 and RP11-180A12. I used fragments from ACeDB of approximately 0.5-1kb, covering the parts of the sequence that were not masked by the Sputnik or the Repeat Masker simple repeat finder programmes or that already displayed 99-100% sequence similarity with other clones. Unfortunately, I did not find any region that was chromosome 4 specific. Every BLASTn result revealed multiple similar clones

positioned on different chromosomes, and every match showed between 92 and 97% identity. Therefore, I had identified that the four clones were highly similar to many clones in other genomic regions, and that not one region was unique to chromosome 4. Therefore, I was not able to identify any regions suitable for marker design.

3.3. Contig Mapping

In the absence of any chromosome 4 specific regions to guide my experiments I decided to design STSs to the sequence at the ends of each clone and screen neighbouring clones with these markers using colony PCR.

3.3.1. Marker Design

I designed markers at the telomeric end of RP11-751L19 and RP11-180A12 and at the telomeric and centromeric end of RP11-264E23. The two markers on RP11-264E23 involved redesigning existing inhouse markers that were non-specific to chromosome 4 (Figure 3-4). Primers were designed by performing a BLASTn similarity search, identifying those clones for which there was a high degree of similarity, and designing primers that matched only the chromosome 4 clone at the 3' end (Figure 3-5).

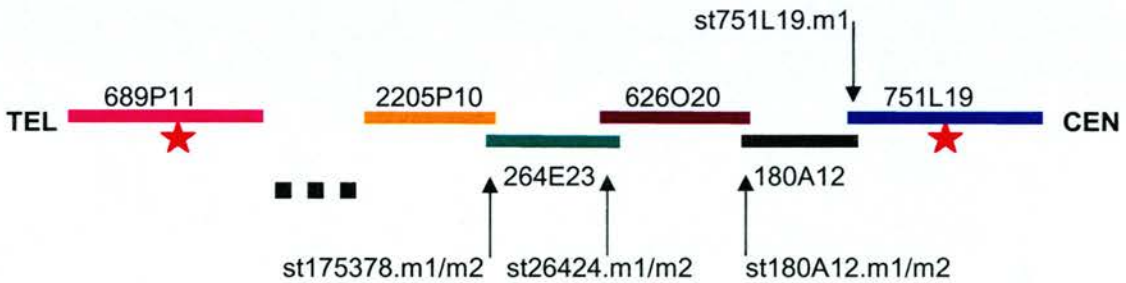


Figure 3-4: The position of STS markers (arrows) along the contig (from the inhouse sequence database ACeDB) that spans the telomeric recombination breakpoint of Minimal Region One (MR1). The ACeDB contig is based on the November 2002 human genome working draft sequence at the University of California, Santa Cruz (genome.cse.ucsc.edu). Coloured bars: sequence clones (see Table 3-1 for full clone details). Black dotted line: contig gap (~150 kb). Red stars: markers defining the telomeric recombination breakpoint of MR1. TEL = telomeric. CEN = centromeric.

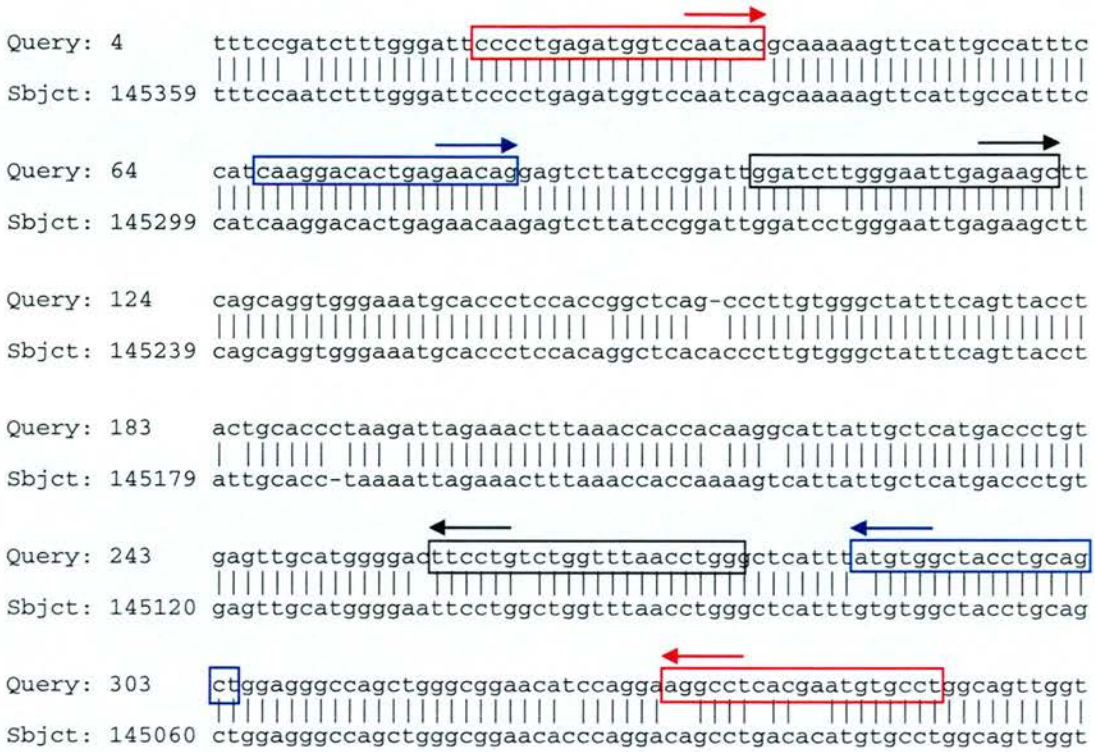


Figure 3-5: Example of primer design around existing STS st26424.snp. Figure shows the results of a BLASTn carried out at the national Centre for Biotechnology Information (www.ncbi.nlm.nih.gov). Query sequence: sequence from clone RP11-264E23. The results revealed matches to 20 sequence clones from other chromosomes. One of the 20 matches is shown here (subject sequence: chromosome 16 clone, RP11-10K17). The black boxes show an existing primer pair designed from RP11-264E23 sequence that are an exact match at their 3' end to RP11-10K17. The red boxes delineate one pair of alternative primers (st26424.m1), and the blue boxes delineate a second pair of alternative primers (st26424.m2) that were designed from RP11-264E23. Both primer pairs have 3' mismatches to RP11-10K17 and to the other 19 similar clones identified from the BLASTn.

3.3.1.1. RP11-751L19

I obtained the full sequence of RP11-751L19 (AC098976) from the NCBI website and performed similarity searches using BLASTn on Homosapiens. The first 900bp matched five clones with between 93 and 95% identity: RP11-119D9 and RP11-655M14 on chromosome 11, RP111-93K22 on chromosome 3, RP11-740N7 on chromosome 7 and a section of chromosome 21q. Upon visual inspection, there was no mismatch between RP11-751L19 and the other clones of more than one or two base pairs in length. It was possible however, to design a pair of primers, named STS st175L19m1, with a mismatch of one or two base pairs at the 3' end to each of the four similar clones.

3.3.1.2. RP11-180A12

I performed a BLASTn of 500-1000bp fragments along the telomeric end of RP11-180A12. I identified a region that matched only one clone from chromosome 16 (RP11-10K17) with any significance. This region was at the end of clone RP11-180A12, near the overlap with RP11-626O20, and therefore could be utilised in mapping. When the chromosome 4 and chromosome 16 sequence fragments were aligned it was possible to see that there was no region of mismatch greater than two or three base pairs long in which to design primers. Therefore I designed two STSs, st180A12m1 and st180A12m2, with 1-2bp mismatches to the chromosome 16 clone at the 3' end of each primer.

3.3.1.3. Centromeric end of RP11-264E23

Clone RP11-264E23 had previously been placed in this region in ACeDB, and markers had been designed from its sequence. However, some of these markers had been shown to amplify other chromosomes by others in the group and data from this clone was considered to be unreliable. I chose a marker that had been designed to a portion at the centromeric end of clone RP11-264E23 and performed a BLASTn of the primers. The results revealed multiple matches at the 3' end of the primer to

clones on other chromosomes. Therefore I performed a BLASTn of the entire STS and 100bp either side (total 379bp). The results revealed 20 matches of greater than 93% similarity to clones on chromosomes 3 (RP11-803B1/379B18/77P16/CTB-182H21), 4 (RP11-747H12/489M13), 7 (RP11-740N7), 8 (RP11-419I17/540E4/483N3/55605), 11 (RP11-167J9/138N3/119D9/RP13-726E6/RP5-903G2), 12 (RP11-90D4),16 (RP11-10K17/382B13) and 17 (RP13-383M5). Clone RP11-489M13 is located elsewhere on 4p16.3 in the July 2003 UCSC Golden Path, and clone RP11-747H12 has not been localised to a position on chromosome 4. In all but one instance (chromosome 16 clone RP11-382N13) both existing primers matched the clones in a position that would result in either the same size product or a product one base pair larger from the other genomic region after PCR. However, by manual inspection of these matches it was possible to design two primer pairs with 3' mismatches to every non-chromosome 4 clone. These STSs were named st26424.m1 and st26424.m2.

3.3.1.4. Telomeric end of RP11-264E23

Marker st175378snp had also been previously designed to clone RP11-264E23. I performed a BLASTn of the primer pair, and considered only matches at the 3' end of the primers. The forward primer matched exactly clone RP11-93K22 on chromosome 3, and clones RP11-757C7 and RP11-655M14 on chromosome 11, and matched clone RP5-1173A5 on chromosome 11 with one base pair mismatch in the middle. The reverse primer did not have any 3' matches. However the first 15bp of the primer did exactly match clones on chromosome 3 and 11 including RP11-93K22 and RP5-1173A5. This would produce product of the same size as product from chromosome 4 after PCR. The primers also matched exactly clone RP11- 1396O13 on chromosome 4. This clone has been localised to this region in some UCSC sequence versions and it was also represented in the Ensembl human genome working draft release (www.ensembl.org) and overlaps with the end of RP11-264E23.

I performed a BLASTn of 600bp around the existing STS st175378snp, and designed two further sets of primers with 3' mismatches to the two clones RP11-93K22 and RP5-1173A5. These STSs were named st174378.m1 and st174378.m2.

3.3.2. Testing Marker Specificity

Markers were initially tested on three control DNAs. My aim was to find the highest stringency PCR conditions that gave a clear single band with no evidence of smearing or bands of the wrong size. Stringency was achieved by performing a test PCR at three annealing temperatures: 55°C, 60°C and 65°C. If a clean band was obtained in one of the conditions, no further modifications were made. If bands of the incorrect size or smearing were obtained, the annealing temperature was increased to a maximum of 70°C and/or the Mg^{+} concentration was reduced to 1-1.25mM. If no band was obtained, the annealing temperature was decreased and/or additives were tried in the PCR (e.g.DMSO). For example, if a PCR at 55°C amplified multiple sized products, but a PCR at 60°C did not amplify any product, the PCR was tried at 57°C. Once these conditions (i.e. PCR conditions that gave a single clear band) were satisfied, primer pairs were tested for specificity to chromosome 4 by PCR of a somatic cell monochromosomal hybrid panel (MCHP). This is a panel of rodent somatic cells that each retain a single human chromosome on a rodent background. The exception is the chromosome 20 hybrid, which as well as chromosome 20 also retains chromosome 4p. Therefore all markers would be expected to be positive both in the chromosome 20 and the chromosome 4 hybrids. If markers were found to amplify from additional hybrids, PCR conditions were further optimised by changing the annealing temperature of the PCR or by altering the reaction mix, for example, using a different PCR buffer, altering the magnesium concentration or varying an additive such as dimethyl sulfoxide (DMSO).

STS markers st180A12.m1 and st180A12.m2 amplified from chromosomes 4 and 20 only using the PCR conditions determined from the genomic DNA test. Therefore no further modifications were necessary (Figure 3-6).

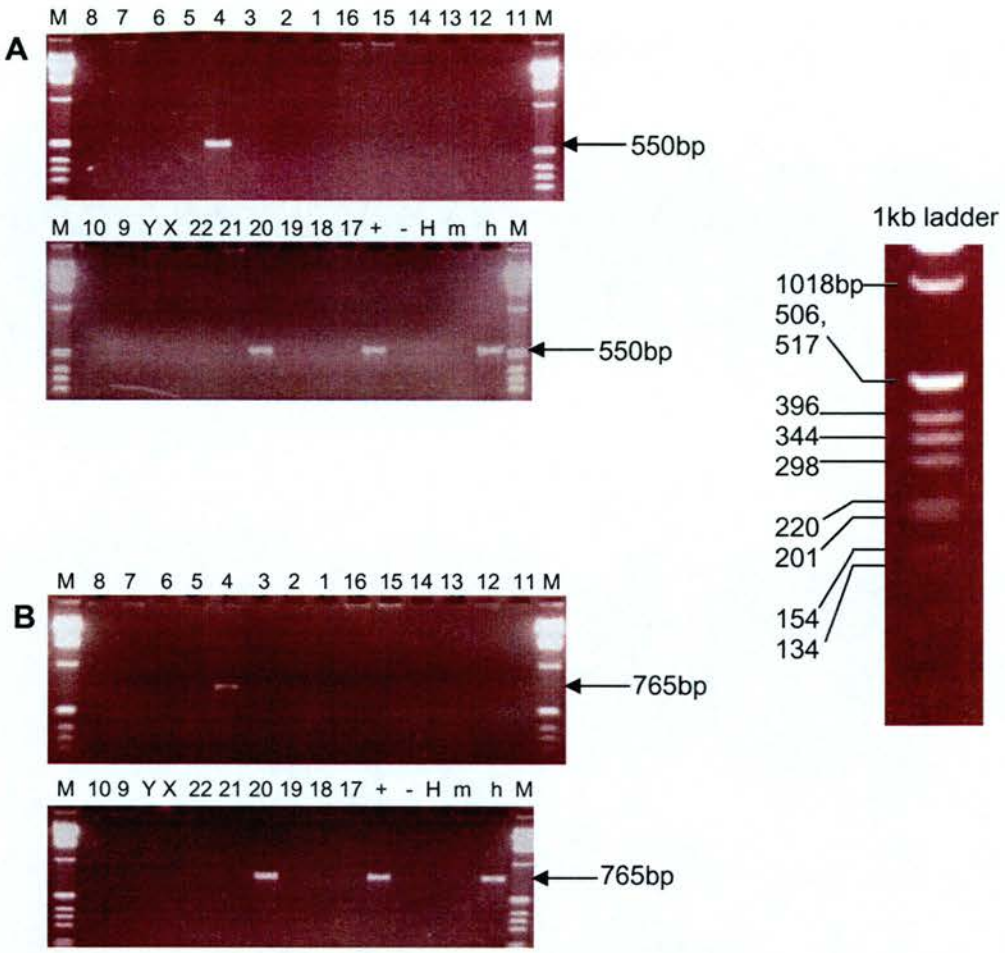


Figure 3-6: Agarose gels showing the results of a PCR on a monochromosome hybrid panel (MCHP). Human chromosomes in the MCHP are on a background of hamster and mouse DNA. Each lane on the gel is labelled with the corresponding human chromosome. The human chromosome 20 hybrid also contains a fragment of human chromosome 4p. The marker (M) = Ready Load™ 1 kb DNA ladder (Invitrogen). The MCHP provides human (H), mouse (m) and hamster (h) genomic DNA controls. An inhouse human genomic DNA control (+) and a no-template control (-) was also included. **A:** Marker st180A12.m1 amplifies product of the correct size (550 bp) from chromosomes 4 and 20 and the two human genomic DNA controls (H and +). **B:** Marker st180A12.m2 amplifies product of the correct size (765 bp) from chromosomes 4 and 20 and the two human genomic DNA controls.

STS marker st751L19m1 was not specific to chromosome 4 using the conditions determined from the control DNA test (Figure 3-7; A). However, as can be seen, the greatest intensity of signal was seen in the correct lanes at the correct size and therefore more stringent PCR conditions were tried. Increasing the annealing temperature by five degrees from 55°C to 60°C, whilst also decreasing the $MgCl^2$ from 1.5mM to 1.25mM was too stringent and the STS did not work (data not shown). Increasing the annealing temperature by five degrees from 55°C to 60°C was enough to make the STS specific (see Figure 3-7; B).

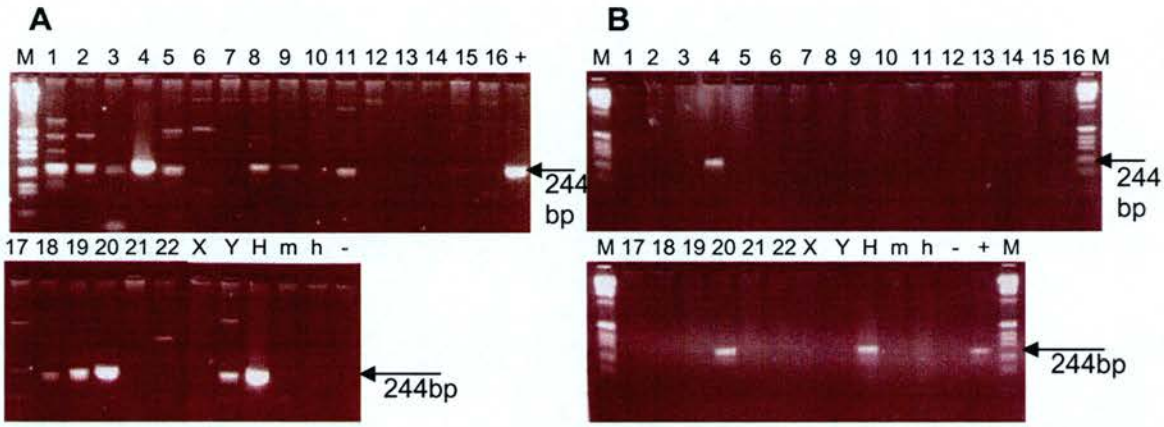


Figure 3-7: Agarose gels showing the results of a PCR on a monochromosome hybrid panel (MCHP). Human chromosomes in the MCHP are on a background of hamster and mouse DNA. Each lane on the gel is labelled with the corresponding human chromosome. The human chromosome 20 hybrid also contains a fragment of human chromosome 4p. The marker (M) = Ready Load™ 1 kb DNA ladder (Invitrogen). The MCHP provides human (H), mouse (m) and hamster (h) genomic DNA controls. An inhouse human genomic DNA control (+) and a no-template control (-) was also included. **A:** Marker st751L19.m1, after a PCR at a 55°C annealing temperature, amplifies product of the correct size (244 bp) and bands of incorrect sizes, from multiple chromosomes, but amplifies the correct sized product only form the two genomic DNA controls. **B:** Marker st751L19.m1, after a PCR at a 60°C annealing temperature, amplifies product of the correct size from chromosomes 4 and 20 only and the genomic DNA controls.

STS marker st26424.m2 displayed some low level amplification on the MCHP for chromosomes 3, 4, 5, 11, 18 and 19, although the background bands were a different size to the specific product (Figure 3-8; A). However, st26424.m1 was specific (Figure 3-8; B). It was noted from the primer design stage (Section 3.3.1.3) that there was a high degree of similarity to chromosome 4 clones RP11-747H12 and RP11-489M13 from other regions of the chromosome. Whilst the primers were designed to have 3' mismatches to these clones, it is not possible to determine regional specificity from the MCHP. However, since the marker did not amplify the highly similar clones from other chromosomes, there was no reason to suspect that it would also amplify these two clones from chromosome 4. Furthermore, the marker used to localise the other end clone RP11-264E23 to CTD-2205P10 acts as an independent test for correct localisation. Therefore, it was not necessary to screen RP11-747H12 and RP11-489M13 with this marker.

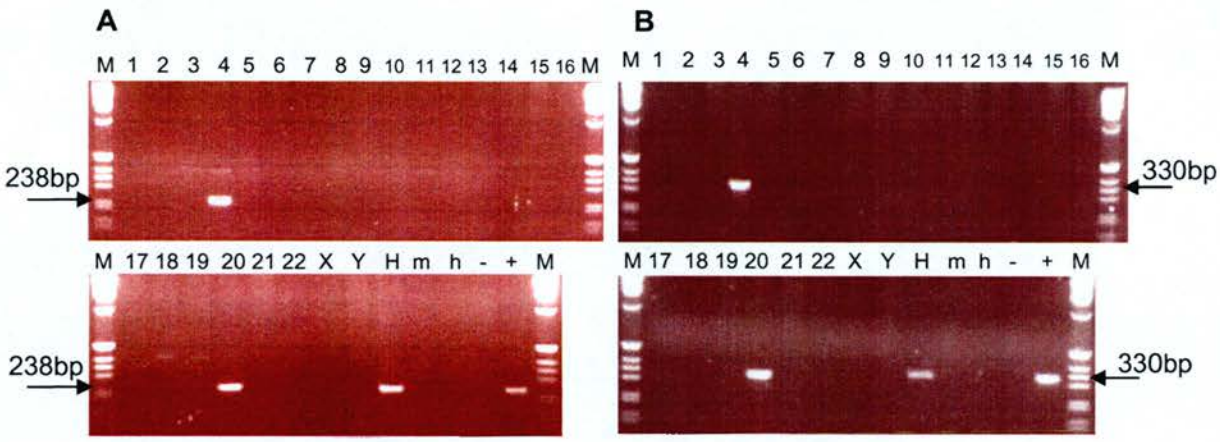


Figure 3-8: Agarose gels showing the results of a PCR on a monochromosome hybrid panel (MCHP). Human chromosomes in the MCHP are on a background of hamster and mouse DNA. Each lane on the gel is labelled with the corresponding human chromosome. The human chromosome 20 hybrid also contains a fragment of human chromosome 4p. The marker (M) = Ready Load™ 1 kb DNA ladder (Invitrogen). The MCHP provides human (H), mouse (m) and hamster (h) genomic DNA controls. An inhouse human genomic DNA control (+) and a no-template control (-) was also included. **A:** Marker st26324.m2 amplifies product of the correct size from chromosomes 4 and 20 and the genomic DNA controls, but also shows faint amplification of a larger product from chromosomes 3, 4, 5, 11, 18 and 19. **B:** Marker st26424.m1 amplifies product from chromosomes 4 and 20 only and the genomic DNA controls.

Using PCR conditions optimised from control DNA tests, st175378snp was not specific on the MCHP, amplifying chromosomes 2, 3, 4, 6, 8, 11, 12, 13, 15 and 20 (Figure 3-9; A). A 5°C increase in annealing temperature still resulted in amplification from chromosomes 3, 4, 11, 15 and 20 (Figure 3-9; B). The BLASTn of the primers had previously revealed similarities to clones on chromosomes 3 and 11 but not on chromosome 15 (Section 3.3.1.4). This could be because this region of chromosome 15 had not been sequenced in the human genome sequencing project at that time. Further attempts to increase the specificity of the PCR by reducing the magnesium concentration and increasing the annealing temperature were not successful (data not shown). Therefore, the primers were redesigned, and two novel STSs were tested: st175378.m1 and st175378.m2.

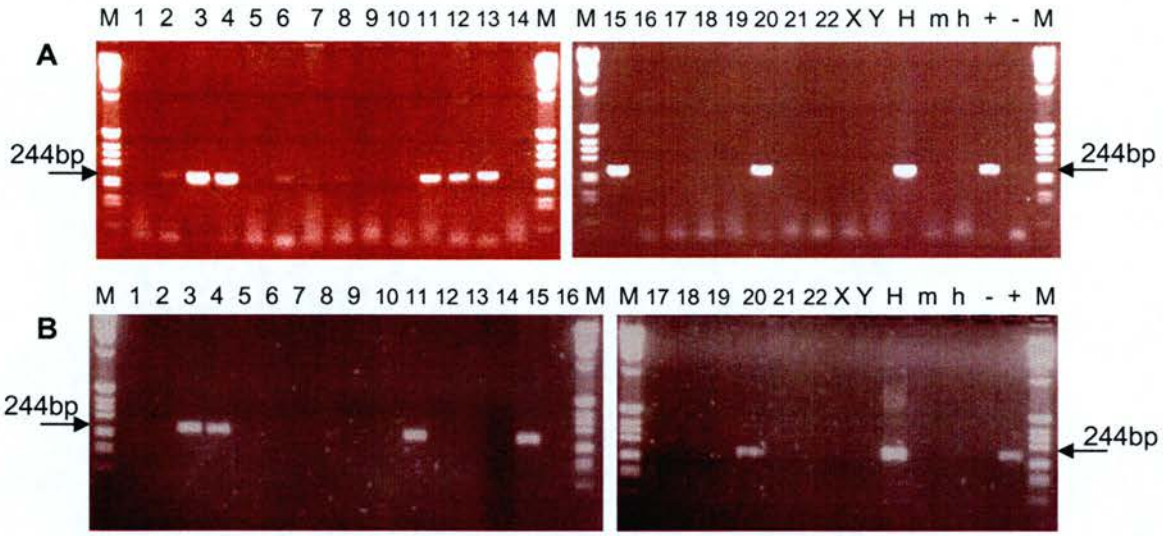


Figure 3-9: Agarose gels showing the results of a PCR on a monochromosome hybrid panel (MCHP). Human chromosomes in the MCHP are on a background of hamster and mouse DNA. Each lane on the gel is labelled with the corresponding human chromosome. The human chromosome 20 hybrid also contains a fragment of human chromosome 4p. The marker (M) = Ready Load™ 1 kb DNA ladder (Invitrogen). The MCHP provides human (H), mouse (m) and hamster (h) genomic DNA controls. An inhouse human genomic DNA control (+) and a no-template control (-) was also included. **A:** Marker st175378snp amplifies the correct sized product (244 bp) from chromosomes 2, 3, 4, 6, 11, 12, 13, 15, 20 and the genomic DNA controls. **B:** After a PCR with a 5°C increase in annealing temperature, st175378snp amplifies the correct sized product from chromosomes 3, 4, 11, 15, 20 and the genomic DNA controls.

Tests on control DNA for STS st175378.m1 were variable (data not shown). For some samples there was a high degree of smearing. However, when it did work it appeared to be specific, although at quite a low annealing temperature. Under these conditions the STS did not work at all on the MCHP except in the human DNA control, and gave some very low level amplification of a smaller product from chromosomes 1, 4 and Y, a larger product from chromosome 7 and the correct sized product from chromosome 6 (Figure 3-10).

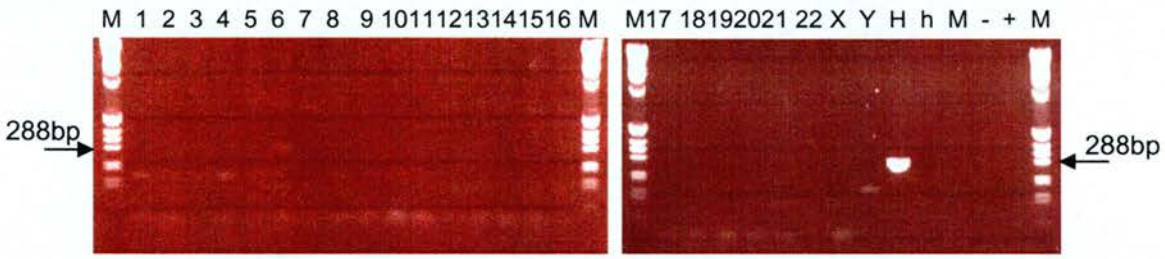


Figure 3-10: Agarose gels showing the results of a PCR on a monochromosome hybrid panel (MCHP). Human chromosomes in the MCHP are on a background of hamster and mouse DNA. Each lane on the gel is labelled with the corresponding human chromosome. The human chromosome 20 hybrid also contains a fragment of human chromosome 4p. The marker (M) = Ready Load™ 1 kb DNA ladder (Invitrogen). The MCHP provides human (H), mouse (m) and hamster (h) genomic DNA controls. An inhouse human genomic DNA control (+) and a no-template control (-) was also included. Marker st175378.m1 amplifies incorrect sized product from chromosomes 1, 4, 6, 7 and Y. It does not amplify the correct sized product (288 bp) from chromosome 4, or the in house genomic DNA control (+), but does amplify the correct sized product from the genomic DNA control provided with the MCHP (H).

STS marker st175378.m2 did not work consistently on all control DNA either, but produced a clean band for some samples (data not shown). Under these conditions it did not amplify from the chromosome 4 or the chromosome 20 hybrid, but did amplify from the positive human controls, producing a clean band (Figure 3-11; A). Amplification was observed from other chromosomes. The chromosome 1-5, 18 and 19 hybrids all show a similar pattern of amplification and all have hamster DNA background. This suggested that at least some of the amplification was from hamster DNA. Increasing the annealing temperature from 55°C to 60°C, altering the buffer and using DMSO additive did not get rid of all of the spurious amplification (Figure 3-11; B). Using the original buffer at an annealing temperature of 60°C eliminated the background but also reduced the efficiency of the STS. It worked poorly on the positive controls (Figure 3-11; C). There is also a non-specific band in lane 18, but this is a different size compared to the specific product.

These experiments lasted over some time. After the above experiments had been carried out, I again performed a BLASTn of the first 1000bp of RP11-264E23 and the results revealed a 100% match to chromosome 4 clones CTD-2205P10 and RP11-1396013. This would be due to the increased amount and/or quality of the sequence in the public database, and provided very strong evidence that the clones overlap. Therefore, further optimisation was not carried out and marker st175378.m2 was tested using the original buffer and an annealing temperature of 60°C (Figure 3-11: C).

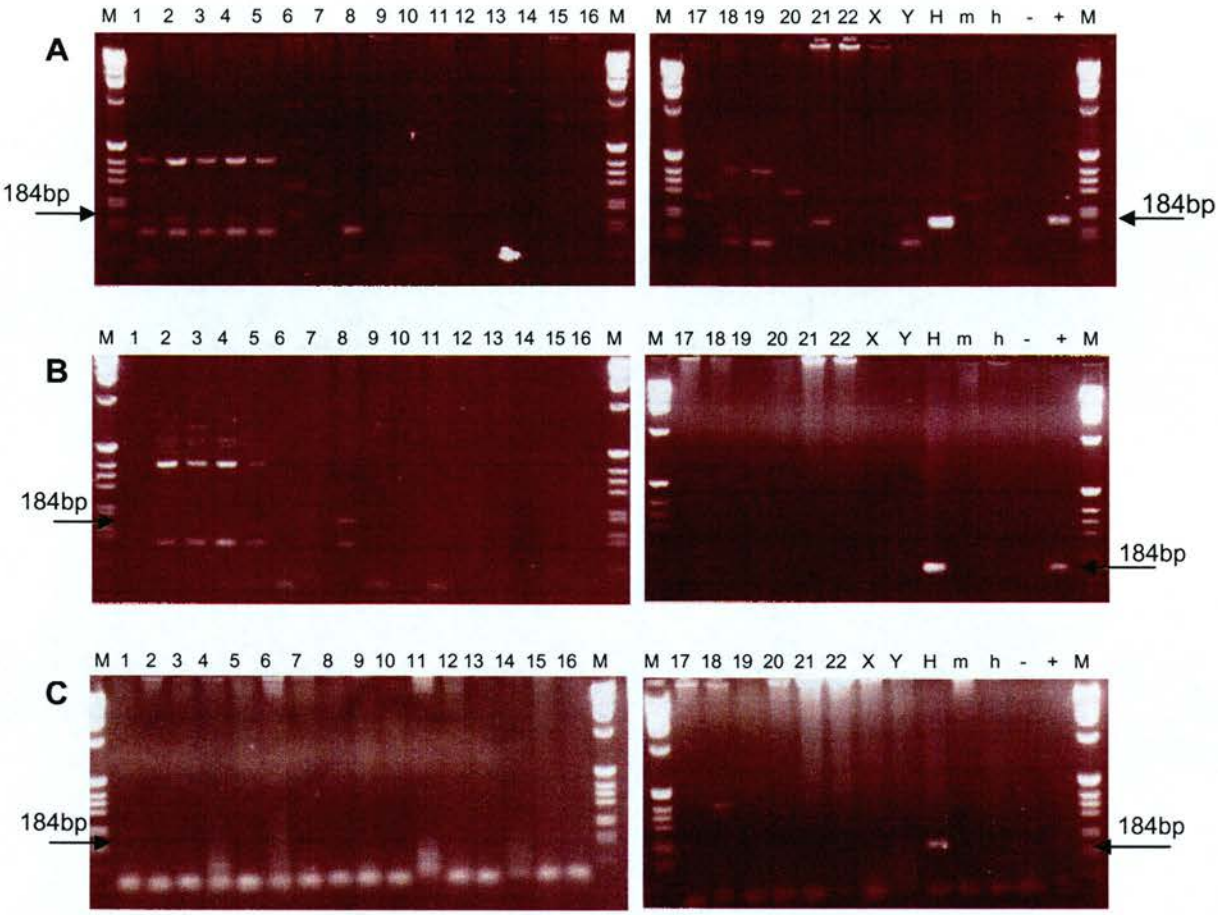


Figure3-11: Agarose gels showing the results of a PCR on a monochromosome hybrid panel (MCHP). Human chromosomes in the MCHP are on a background of hamster and mouse DNA. Each lane on the gel is labelled with the corresponding human chromosome. The human chromosome 20 hybrid also contains a fragment of human chromosome 4p. The marker (M) = Ready Load™ 1 kb DNA ladder (Invitrogen). The MCHP provides human (H), mouse (m) and hamster (h) genomic DNA controls. An inhouse human genomic DNA control (+) and a no-template control (-) was also included. **A:** After a PCR at an annealing temperature of 55°C, marker st175378.m2 amplifies the correct sized product (184 bp) from the human genomic DNA controls (H and +) only. Product of the incorrect size is amplified from chromosomes 1, 2, 3, 4, 5, 6, 10, 17, 18, 19. 20, 21, X ,Y and the mouse and hamster genomic DNA controls. **B:** After a PCR with an annealing temperature of 60°C and different PCR reagents, marker st175378.m2 amplifies the correct sized product from chromosome 8 and the human genomic DNA controls (H and +) and product of the incorrect size from chromosomes 2, 3, 4, 5 and 8. **C:** After a PCR at an annealing temperature of 60°C and the original PCR reagents, marker st175378.m2 amplifies the correct size product from only the human genomic DNA controls (H and +) and a larger product from chromosome 18.

3.3.3. Colony PCR Results

The four STSs were screened on the BACs by colony PCR and each STS was found to amplify the BAC screened. Positive genomic DNA and no template controls were included in each PCR.

Marker st751L19.m1 amplified from both RP11-751L19 and RP11-180A12 (fig 3-12; A), confirming that these two clones overlap. Both st180A12m1 and st180A12m2 amplified from RP11-180A12 and RP11-626O20 (Figure 3-12; B, C & D) highlighting the error in the November 2002 Golden path release which leaves a gap between them. This has been rectified in the April 2003 release. Marker st26424.m1 was positive for both RP11-626020 and RP11-264E23 (Figure 3-12; E) confirming the Golden Path April 2003 release. Marker st175378.m2 amplified from both RP11-264E23 and CTD-2205P10 but did not amplify from the genomic DNA control (Figure 3-12; F & G). As observed in the previous section, the STS did not work very efficiently. However, the STS could be expected to work better in a colony PCR, where the amount of template far exceeds that of the 20ng used in the genomic DNA control. The April 2003 sequence assembly was released subsequent to this work being carried out. After the above experimentation, a BLASTn sequence similarity search of st175378.m2 revealed a sequence identity of 99-100% to RP11-264E23, CTD-2205P10 and RP11-1396O13. Therefore it was not deemed necessary to test the STSs on RP11-1396O13 as well.

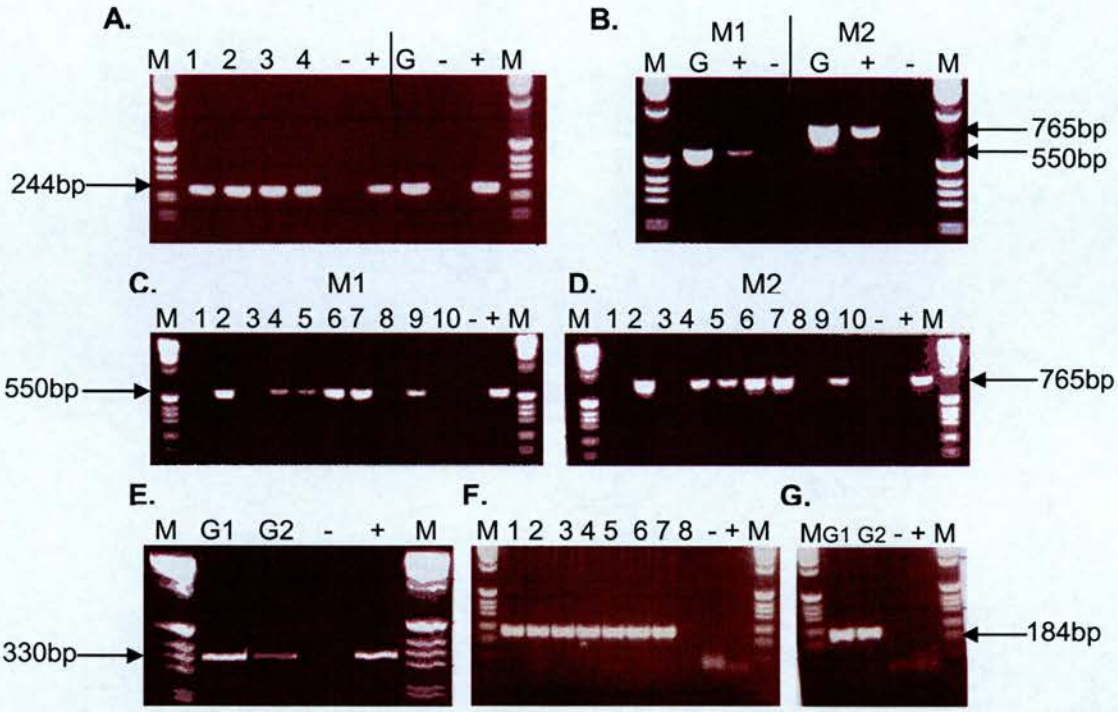


Figure 3-12: Agarose gels showing the results of colony PCR for five markers. Each marker, except st751L19.m1, was tested on two BAC clones. Colony PCR was performed on either colonies grown overnight from a glycerol stock, or from a crystal taken directly from a frozen glycerol stock of the clone. The size of each marker (in base pairs) is shown by the arrow. The DNA marker (M) = Ready Load™ 1 kb DNA ladder (Invitrogen). Human genomic DNA (+) and no-template controls (-) are included for each PCR. **A.** Marker st751L19.m1 is positive for four BAC colonies (1-4) grown from a glycerol stock of RP11-180A12, and a glycerol stock crystal of RP11-751L19 (G). **B.** Markers st180A12.m1 (M1) and st180A12.m2 (M2) are positive for a glycerol stock crystal of RP11-626O20 (G). **C.& D.** Markers st180A12.m1 (M1) and st180A12.m2 (M2) are positive for six out of 10 (1-10) colonies of RP11-180A12 grown from a glycerol stock. **E.** Marker st26424.m1 is positive for a glycerol stock crystal of RP11-264E23 (G1) and RP11-626O20 (G2). **F.** Marker st175378.m2 is positive for seven out of eight (1-8) colonies of CTD-2205P10 grown from a glycerol stock. The genomic DNA control (+) failed to amplify. **G.** Marker st175378.m2 is positive for glycerol stock crystals of RP11-264E23 (G1, G2). The genomic DNA control (+) failed to amplify.

3.4. Discussion

The publicly available contigs from UCSC in the region between RP11-301J10 and RP11-751L19 on chromosome 4 have varied, sometimes significantly, with each subsequent release. The inhouse contig did not reliably extend past RP11-751L19, therefore novel clones in this region would be a potential source of novel genes to include in future association studies, and also a source of novel microsatellite and SNP markers to use in refining the F50 recombination breakpoint. Therefore it was important to determine unequivocally which clones were positioned here and their relative order.

The inhouse contig built prior to the start of this project, and every release of the human sequence to date, has a gap that lies between RP11-301J10 and RP11-751L19. This interval also defines the telomeric recombination breakpoint interval of MR1. The continuing presence of this gap could be due to the repetitive nature of the DNA, as has been observed on the centromeric side and/or, due to the low representation of this region in genomic libraries, as has been observed on the telomeric side. This was confirmed by others by the fact that some of the markers in the region, when attempting to construct the inhouse contig past RP11-751L19, were non-specific to chromosome 4. I have also confirmed this by noting that ~43kb of the chromosome 4p specific tandem repeat CRS447 positioned at the end of CTD-2205P10 suggests that the remainder of this repeat lies within the gap, and also by the results of the series of BLASTn similarity searches I performed across four clones telomeric to RP11-751L19.

Evidence from FISH mapping (Evans *et al*, 2001_a), and from the six releases studied of the UCSC Golden Path suggest that the gap is in the order of 50-300kb in size. An average BAC size of ~150kb means that only one or two clones are required to span the gap. As mentioned above, a significant proportion of the sequence in the gap may be composed of the CRS447 repeat. It has been estimated that this is 235-329kb in length (Kogi *et al*, 1997), and approximately 192-286kb of this is therefore missing

from the sequence of the clone flanking the gap, CTD-2205P10. However, this is dependent on how many copies of the repeat occur in the individuals forming the HGP, as the number of repeats has been shown to be highly polymorphic between individuals (Gondo *et al*, 1998). This repeat possesses a 1590bp ORF encoding a deubiquinating enzyme (designated USP17), shown to have a functional promoter and mRNA expression in a number of tissues (Saitoh *et al*, 2000). Despite this, USP17 could not be located in either the July 2003 UCSC Golden Path or the 34b Ensembl genome browsers and was not an entry at NCBI.

The repetitive nature of the clones telomeric to RP11-751L19 made it difficult to design chromosome 4 specific markers. This had hindered those working on the inhouse contig in the past as mentioned above, and meant that despite the fact that the inhouse contig extended past RP11-751L19, it was considered to be very unreliable. In order to design markers I used the results of BLASTn similarity searches to identify which clones had a high level of similarity with the end of the clone of interest and then used these to design primer pairs specific to chromosome 4 at their 3' end. I then used increasing levels of PCR stringency to attain specificity to chromosome 4 on a monochromosomal cell hybrid panel. It is the 3' end of the primer that is the most important for the specificity of PCRs and this was bourne out by my results. A difference of only one or two base pairs was enough to confer specificity to chromosome 4.

My results also highlight the fact that the human genome sequence is not yet complete. Marker st175378snp amplified a portion of chromosome 15, but a BLASTn similarity search did not reveal any similarity to chromosome 15 clones.

In addition, my results also showed that STSs do not necessarily behave the same in a MCHP situation compared to whole genomic DNA. In the absence of the correct position to anneal to it appears that other non-specific locations can be amplified. However in the presence of the correct position these alternative sites are not amplified at all. This was seen for example in marker st175378.m1 (see Figure 3-11).

Therefore some caution is needed when interpreting the results of a PCR on one type of template if generalising it to another type of template. In addition, when designing the markers I did not perform the BLASTn against mouse or hamster sequence, but human only. Therefore some background could be due to homology between the species. A BLASTn of the primer sequences against hamster genomic sequence is not possible since the hamster genome has not been sequenced. However, subsequent BLASTn against the mouse genome for STSs st175378.m1, st175378.m2, st26424.m1 and st26424.m2 (the STSs that amplified bands of various sizes from multiple chromosomes) revealed several matches to mouse genomic DNA at the 3' end of each primer. The forward and reverse primer for each STS did not match the same clone and product of the expected size from human genomic DNA would not have been amplified. It does suggest, however, that some of the amplification of product of a different size to that expected from human genomic DNA could be due to amplification from mouse and/or hamster DNA. It seems reasonable to suggest that a low level of non-specificity, when the bands are of a different size to the specific band, can be tolerated on the MCHP when the specific chromosome amplifies from the appropriate hybrid only.

Since this work was carried out a subsequent sequence release from the HGP (July 2003) is available at UCSC. In this release, clone RP11-626O20 has been removed from the latest sequence assembly (Figure 3-2: E), despite there still being a gap between the flanking clones. However, I have shown in my results that it is correctly positioned between clones RP11-180A12 and RP11-264E23.

There were some clones that were not included in this work. Clone RP11-1286E23, overlapping RP11-1396013 (Figure 3-2), has been removed from the April and July 2003 releases. It also appears to have been completely removed from the UCSC Golden Path; searching with either the clone or the sequence name produces no results. Therefore, I did not include this clone in my work. There is also an additional clone, RP11-637J21, that in the July 2003 release was positioned into the gap from

the telomeric side. This clone could also be another source of novel microsatellite and SNP markers to refine the recombination breakpoint.

In conclusion, I have succeeded in accurately mapping four clones into the contig gap within the recombination breakpoint interval of the telomeric boundary of MR1. These clones can now be used as a source of markers to refine the breakpoint, and as a potential source of novel genes.

Chapter Four

Recombination Breakpoint Mapping Of Minimal Region One

Recombination Breakpoint Mapping of Minimal Region One

4.1. Introduction

Blackwood *et al* (1996) described a genome wide linkage study in a large family of 120 individuals, F22, segregating bipolar affective disorder (BPAD) and recurrent major depression (RMD) (Figure 2-1). A three point linkage analysis gave a maximum multipoint LOD score of 4.8 to a 14cM region on chromosome 4p. A 22Mb haplotype is inherited from one founder individual to all 11 cases of BPAD I and II and 14 out of 16 cases of RMD. This haplotype defines the disease associated chromosome in F22 (Figure 4-1). Two individuals with early onset RMD did not possess the haplotype, which suggests that they are phenocopies. Nine individuals without a psychiatric diagnosis did possess the haplotype, which suggests that there has been incomplete penetrance of the susceptibility locus. Since psychiatric illness is a complex disorder where multiple genes and environmental factors contribute to the phenotype, phenocopies and incomplete penetrance are expected when a single genetic locus is analysed.

A further family, F59, also showed evidence for linkage to the region (Figure 2-2). From subsequent unpublished laboratory work by others in the group, F59 has been shown to exhibit a ~8Mb disease associated haplotype (Figure 4-1). This is transmitted from an unknown founder to all five cases of BPAD I and BPAD II disorder, and one undiagnosed individual. The individual without a diagnosis who carries the disease associated haplotype again suggests that there is incomplete penetrance of the susceptibility locus.

Two remaining families, F50 and F48 (Figures 2-3 and 2-4), were reported in the literature as showing linkage of psychiatric illness to chromosomes 4p (Detera-Wadeligh *et al*, 1999; Asherson *et al*, 1998) and were obtained via collaboration. The results of the genome wide linkage study in F50 found a LOD score of 2.0 on

chromosome 4p. This is small, but is limited by the small size of the family. The genome wide linkage study in F48 identified a LOD score of 3.2 on chromosome 4p. These two families also show marker haplotypes that are inherited with the disorder. Family 50 segregates schizoaffective disorder, and was shown to have a disease associated haplotype greater than 20Mb (Figure 4-1). The six individuals who carry the disease haplotype are diagnosed with schizophrenia (SCZ) or schizoaffective disorder. In addition, one individual has a form of psychosis that could not be further classified due to incomplete data (Asherson *et al*, 1998). Family 48 is a large family of 82 individuals who segregate BPAD, RMD and SCZ. Subsequent unpublished work by others in the group has identified a ~14Mb disease associated haplotype (Figure 4-1).

The disease associated haplotype in each family defines the region that contains the susceptibility polymorphism. Haplotypes are defined by tracing the inheritance of marker alleles from parent to offspring in a family. An interruption in the inheritance of a contiguous haplotype from a parent to their offspring with the commencement of the inheritance of the alternate haplotype from the second chromosome of that parent constitutes a recombination breakpoint.

All of the above work was carried out before the start of this PhD project, when it was also observed that the disease associated haplotypes from the four families overlap (Figure 4-1). If the families have a common causative ancestral allele, or a different causative ancestral allele affecting the same gene, this overlap provides an informative way to significantly reduce the size of the susceptibility locus detected in F22. The overlapping haplotypes of the four families divide the linked region in F22 into four main candidate regions. Those of most interest are minimal region one (MR1) and minimal region two (MR2), where three of the four family haplotypes overlap. MR1 is defined by F22, F59 and F50. This is a good candidate region because all three families are of Celtic origin, making a founding variant more likely. In addition, an association has been found between MR1 markers and SCZ (Muir *et al*, 2001). However, families 59 and 50 are small and can contribute only small LOD

scores. MR2 is a good candidate region because it is defined by the two largest families, F22 and F48, who have the largest LOD scores. However, F48 originates from a different ethnic background; they are an Ashkanazi Jewish family from the USA. This makes a common ancestral variant less likely, but it does not preclude it.

Here I describe the analysis of MR1. The recombination breakpoints defining the other minimal regions have been studied and refined by other members of the group, and the current resolution of these are detailed in Figure 4-1. At the start of this project, the recombination breakpoints defining MR1 were defined by widely spaced markers (those detailed in Figure 4-1). When I started, the recombination breakpoint interval of the centromeric end of MR1 was ~1.8Mb and the telomeric end of MR1 was ~800kb. This meant that MR1 was ~4.8Mb. Previous marker genotypes had shown that the centromeric end of MR1 is defined by a recombination breakpoint in individual 2081 from F59 and that the telomeric end is defined by a recombination breakpoint in individual 12 from F50. It is important to refine the disease associated haplotypes to delineate the locus to a high resolution, thus eliminating unlinked genes from further study. Therefore more markers needed to be typed within the breakpoint intervals. I refined the recombination breakpoints by identifying novel microsatellite and SNP markers within the recombination breakpoint intervals and genotyping these in the families.

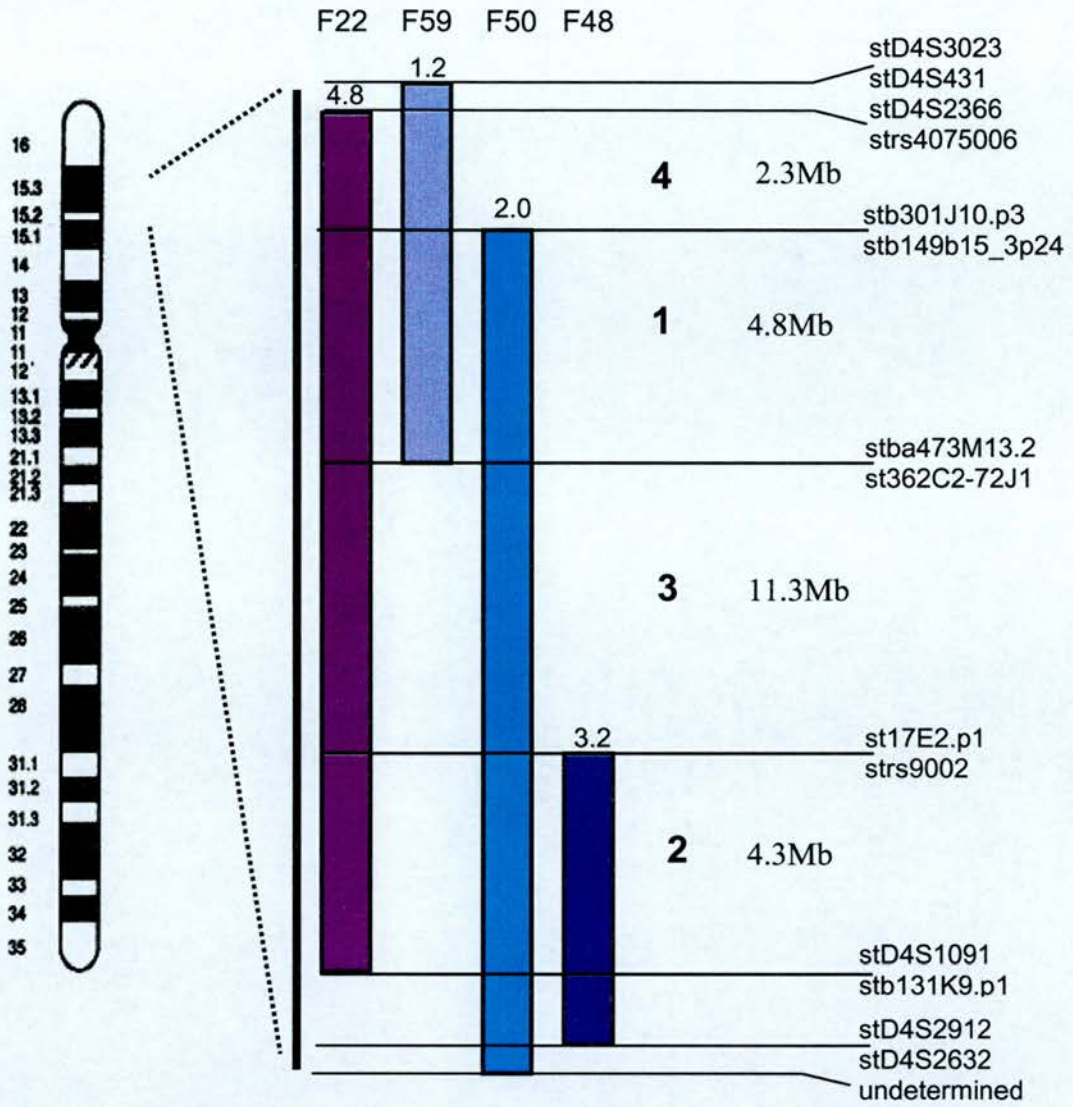


Figure 4-1: Overlapping marker haplotypes of families 22, 59, 50 and 48. The coloured bars represent a marker haplotype on chromosome 4p inherited with psychiatric illness in each family. The family number is marked along the top. The number above the bar shows the LOD score for that family (F22 = a multipoint LOD score). The four haplotypes overlap, enabling the delineation of smaller regions of interest. The horizontal lines mark the delineation of each minimal region (MR), with the corresponding region number (1-4) and the markers flanking the recombination breakpoints alongside. The size of each MR is noted. The markers delineating the breakpoint and corresponding size of MR's 2, 3 and 4 represents the current resolution (February 2004). The markers and corresponding size of MR1 represents the resolution at the start of my project (October 2000).

4.2. F59 Recombination Breakpoint

A recombination breakpoint in individual 2081 from F59 constitutes the centromeric boundary of MR1. It was important to refine this interval as far as possible to include or exclude genes from this high priority region. At the start of the project, inhouse markers stba473M13.2 (non-recombinant) and st362C2-72J1 (recombinant) defined the breakpoint. These markers were derived from BACs RP11-473M13 and RP11-11C13 respectively. The exact distance between these markers was not known because the intervening sequence was incomplete. However, the human genome sequencing project reported an average BAC insert size of 150kb, suggesting that the two markers were approximately 1.4Mb apart. Now that the sequence is complete, it has been possible to determine that they are 1.8Mb apart. At the start of the project, there were no known genes within this recombination interval. However, a formal transcript map of the region had not been constructed. Subsequently, I have identified a novel gene, 74M11, of unknown function within this interval on RP11-74M11 (Chapter 6).

4.2.1. Family Members

In order to map the recombination breakpoint, all eight members of F59 were genotyped. Data was obtained for seven members, not including sample 2082 where PCR amplification failed, probably due to poor DNA quality. I genotyped the entire family for all new markers in order to ensure that unambiguous determination of the phase of all chromosomes was possible. In addition, nine members of F22 (93, 94, 137, 139, 16, 17, 50, 51, 53) and five members of F50 (3, 4, 11, 12, 15) were genotyped. This was to determine the disease associated haplotype in these families, allowing the assessment of allele sharing between the three families, and to ensure that individual marker characteristics are accurately interpreted and the correct genotypes are scored. It is important to genotype enough individuals to deal with stutter and plus-A peaks, two common characteristics observed when genotyping microsatellite markers.

4.2.2. Markers

At the start of my PhD the recombination breakpoint was defined as falling between telomeric marker stba473m13.2 and centromeric marker st362C2-72J1 (Figure 4-1). Neither the contig produced by the human genome mapping project, nor the sequence of the BACs mapped to this interval, were complete at this time. However, previous work in the lab had resulted in the construction of a contig spanning MR1 (Evans *et al*, 2001_a). The contig data is displayed in the System for Assembling Markers (SAM) database (Soderlund, 1995). This programme aids the ordering of BACs on the basis of marker hybridisation and/or BLAST matches of STSs to partially sequenced BACs. A minimal tiling path of nine BACs that spanned the recombination interval between markers stba473m13.2 and st362C2-72J1 was chosen from this contig (Figure 4-2). The nine BACs were being sequenced as part of the human genome sequencing project, allowing new marker design. Where possible, BACs were chosen that contained one or more markers that hybridised to both BACs. Where BACs did not directly overlap it was possible to determine their relative position based on their overlap with surrounding BACs and PACs.

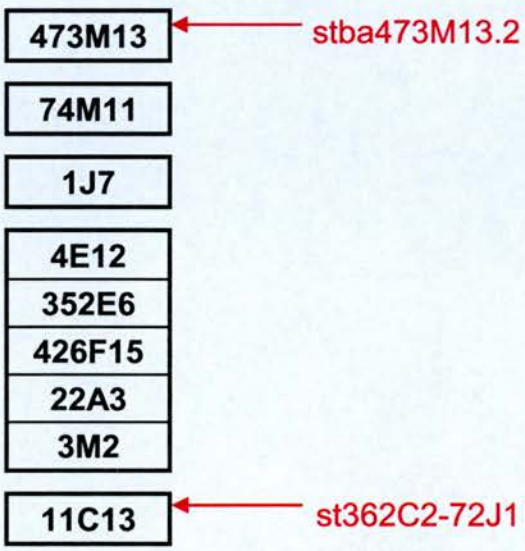


Figure 4-2: Tiling path of nine BAC clones (RP11-) between recombinant and non-recombinant markers stba473m13.2 and st362C2-72J1 that mark the centromeric recombination breakpoint of the chromosome 4p disease associated haplotype in family 59 and the centromeric boundary of Minimal Region One. BACs 4E12 to 3M2 overlap. The remaining BACs are positioned with respect to other BACs (not shown) in the region.

There were four phases of genotyping. The aim of the first phase was to identify and genotype an average of one marker per BAC. Primers were designed to flank putative polymorphic dinucleotide, pentanucleotide or tetranucleotide repeats identified by the software programme Sputnik (cbi.labri.fr/outils/Pise/sputnik.html). Although a repeat length of greater than 12 was used to judge whether a marker was likely to be polymorphic, each marker had to be tested on a set of control individuals to check that the repeat was polymorphic. In phases one and two of genotyping, ten putative polymorphic repeats were tested in this way (Table 4-1, page 115, details all the markers used). The PCR assay did not work for markers 19 and 20, and therefore I redesigned the primers. Markers 2, 4, 5, 6, 16 and 20 were polymorphic and marker 14 was monomorphic. An eleventh publicly available microsatellite marker (3) was also tried but the PCR repeatedly failed and therefore it was dropped. Therefore, nine microsatellite markers were typed in Phase one and two. In phase three of genotyping, four putative single nucleotide polymorphisms (SNP) identified in dbSNP (www.ncbi.nlm.nih.gov/SNP/) were tested for polymorphism in the family members. Primers were designed to flank the SNP's and genotypes were determined by sequencing. Two STSs were designed with two SNPs in each STS. Two of the SNP's (markers 8 and 11) were polymorphic and used for recombination breakpoint mapping. Phase four is discussed in greater detail in Section 4.4.

Clone (RP11-)	Marker Code	Marker Name	Marker Type	Note	Distance (bp)	Phase
473M13	1	stba473M13.2	Microsat		-	-
74M11	2	stba74M11.p1	Microsat		181 768	1
74M11	3	stD4S3352	Microsat	PCR failed	54 838	1
1J7	4	Stba1J7.p1	Microsat		236 640	1
4E12	5	stba4E12.p1	Microsat		185 458	1
352E6	6	stba352E6.p1	Microsat		363 429	1
352E6	7	stb352E6.p2	Microsat		14 345	4
352E6	8	rs1402045	SNP		18 289	3
352E6	9	rs545029	SNP	monomorphic	68	3
352E6	10	rs1402034	SNP	monomorphic	8571	3
352E6	11	rs1520285	SNP		132	3
426F15	12	stb426F15.p4	Microsat		20 908	4
426F15	13	D4S2906	Microsat		34 087	4
426F15	14	stb426F15.p2	Microsat	monomorphic	14 516	2
426F15	15	stb426F15.p5	Microsat		2957	4
426F15	16	stb426F15.p3	Microsat		19 436	1
426F15	17	stb426F15.p1	Microsat		21 336	2
22A3	18	stba22A3.p1	Microsat		194 494	1
3M2	19	stba3M2.p1b	Microsat		86 266	2
11C13	20	stba11C13.p1b	Microsat		193 861	2
11C13	21	st362C2-72J1	Microsat		132 536	-

Table 4-1: Details of the markers used to refine the family 59 recombination breakpoint. The breakpoint marks the centromeric end of the haplotype inherited with psychiatric illness in family 59 and delineates the centromeric boundary of Minimal Region 1. The table shows a tiling path of nine clones identified between the current recombinant and non-recombinant markers and the markers designed from them. The marker number refers to Figure 4-6. The distance alongside a marker refers to the distance between it and the preceding marker. Distance is measured from the first nucleotide of the forward primer for each marker. Four phases of genotyping were carried out. Microsat: microsatellite. SNP: single nucleotide polymorphism.

4.2.3. Genotyping

The nine microsatellite markers were tested using a variety of different conditions on three control DNAs to establish optimal PCR amplification. Each marker was then tested for polymorphism on the family members. A PCR was carried out using the optimal conditions and a dilution of the product was run on the ABI 3100 Genetic Analyser. Genotypes were analysed using the GeneScan version 3.0 software (Figure 4-3). Markers were multiplexed where possible. Genotypes were scored blind to phenotype and family relationship and then checked for Mendelian segregation. SNPs were genotyped by manual inspection. Sequence chromatograms were aligned using the phredPhrap software and visualised with the Consed programme (Figure 4-4)

4.3. Definition of the F59 Recombination Breakpoint Interval

Figure 4-5 details the four phases of genotyping and the relative position of the markers in each phase, and Figure 4-6 shows the genotyping results in the family. Six markers were used in phase one: 2, 4, 5, 6, 16 & 18 (Table 4-2 and Figures 4-5 and 4-6). Markers 2 and 4 are uninformative because the transmitting parent, F59-2078, is homozygous. This makes it impossible to distinguish between the disease associated and the non-disease associated chromosome. Both offspring, 2080 and 2081, share the same parental chromosome for markers 5 and 6, and therefore these are non-recombinant. Markers 16 & 18 are uninformative because again the parental genotype is homozygous. From these results it is possible to say therefore that the recombination breakpoint is centromeric to marker 6. Four markers were genotyped in phase two: 14, 17, 19 and 20. Marker 14 was not polymorphic. Markers 17, 19 and 20 show that offspring 2080 and 2081 inherit different chromosomes from parent 2078. Therefore, there has been a recombination event between marker 6 and marker 17, meaning that markers 17-20 are on the centromeric side of the recombination breakpoint.

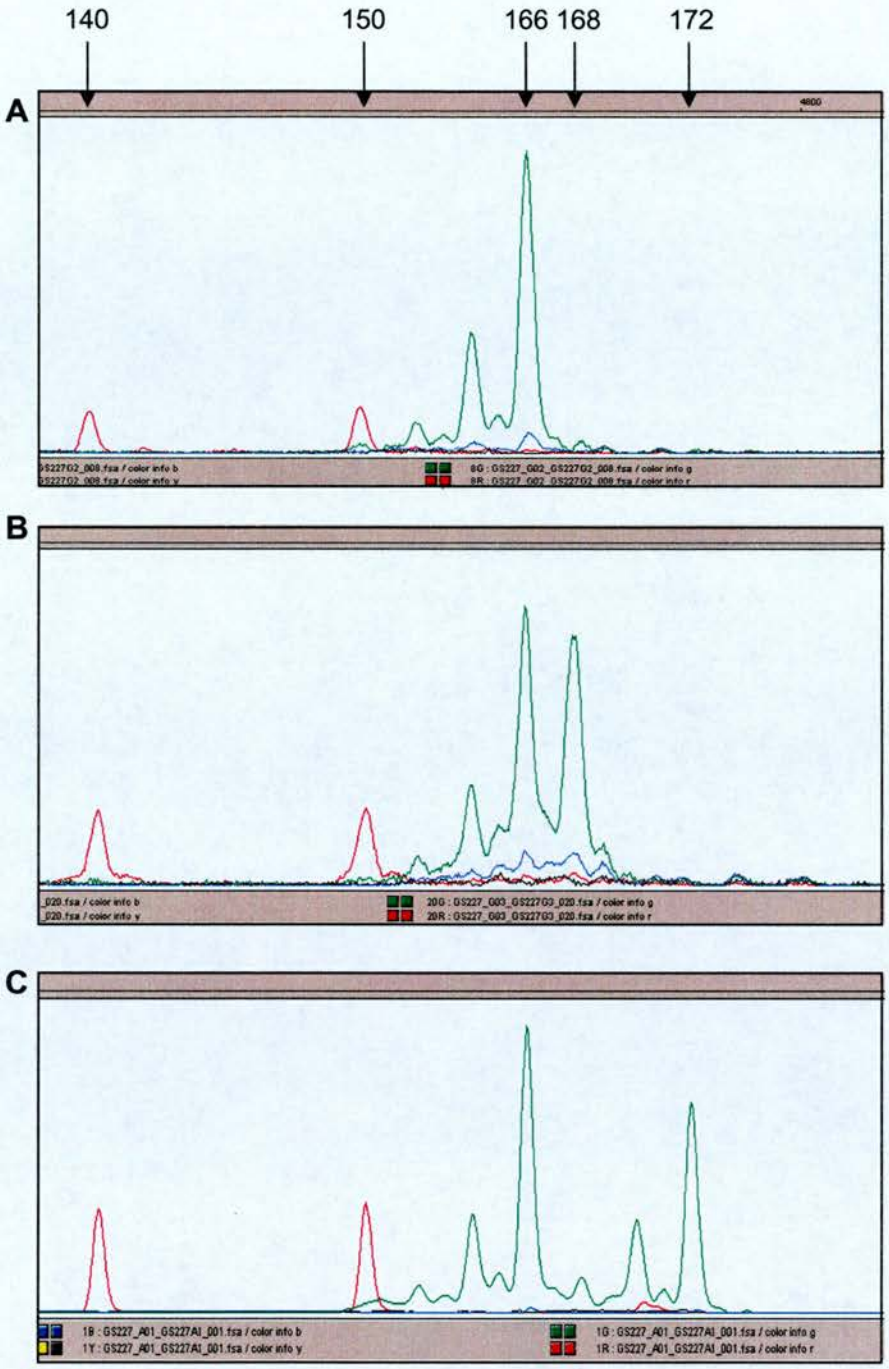


Figure 4-3: Fluorograms of microsatellite marker stD4S2906 (green) (visualised using GeneScan version 3.0 software). Size standard peaks are seen in red. Allele size (in base pairs) is marked along the top. The marker shows two major stutter peaks two base pairs apart, with minor background peaks in between. The genotypes of three individuals are shown **A:** Individual with homozygous 166 genotype. **B:** Individual with heterozygous 166/168 genotype. **C:** Individual with heterozygous 166/172 genotype.

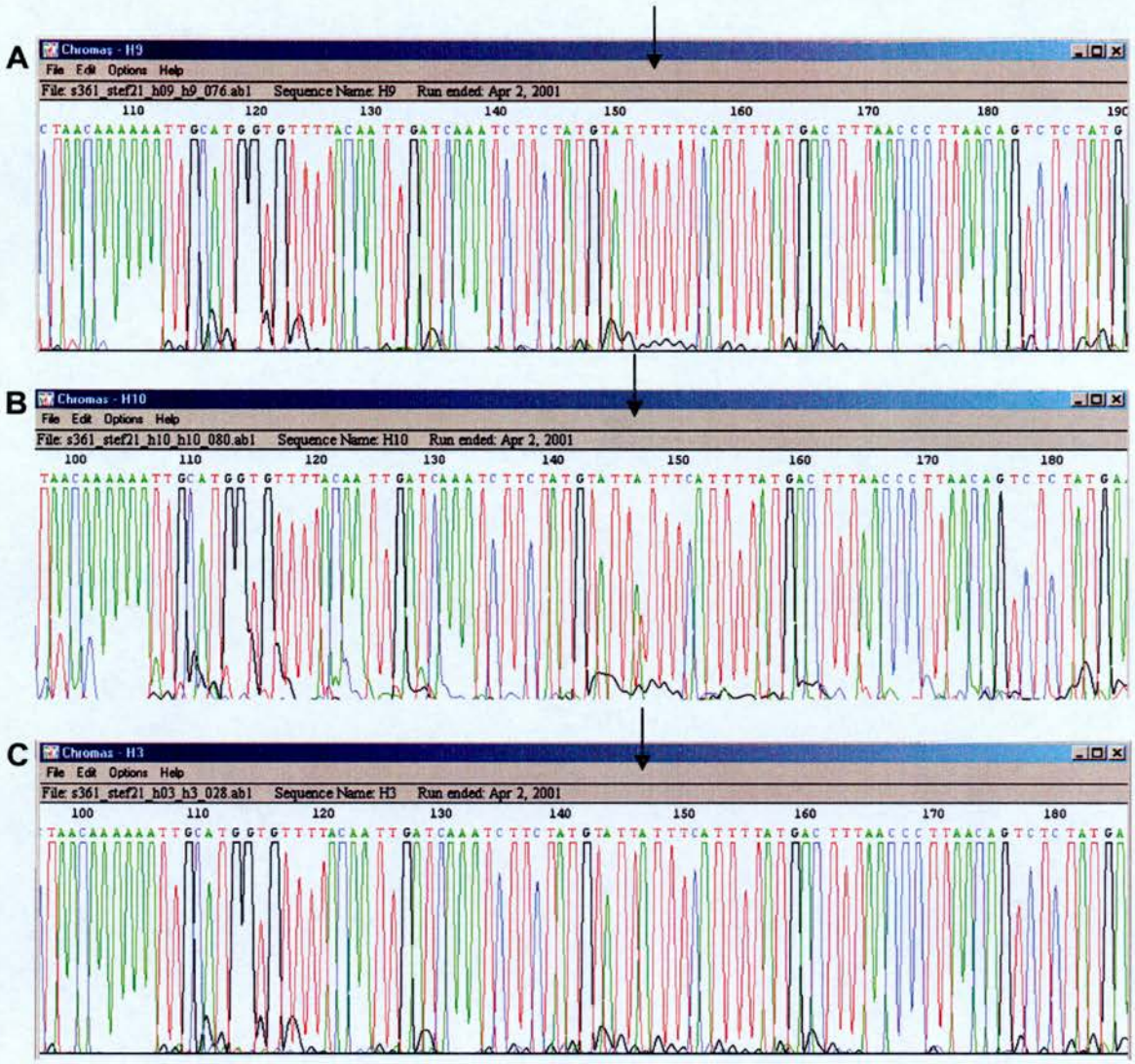


Figure 4-4: Part of the sequence trace for STS st585128.snp. The sequence trace is displayed using the Chromas software (www.technelysium.com.au/chromas. html). A single nucleotide polymorphism (rs1520285) is observed (arrow). The sequence traces of three individuals are shown with genotypes **A:** homozygous t/t, **B:** heterozygous t/a, **C:** homozygous a/a.

There are two ways of interpreting who the recombinant individual is. I have assumed that the recombination event occurred in individual 2081. However, it could have occurred in both 2080 and 3688 in the same region and not 2081. However, the simplest explanation is that only one recombination event has occurred. Recombination is a relatively rare genetic event, and it is very unlikely to happen independently in two individuals in the same place.

Markers 6 and 17 that flank the recombination breakpoint are positioned on overlapping BACs. However, the sequence for both BAC's was unfinished at this time (February 2001), and therefore the extent to which I had resolved the recombination interval was impossible to determine accurately. Both BACs were known to contain a T7 end clone marker called stb272L7.t7, from BAC RP11-272L7. Therefore, the sequence fragment of RP11-3523E6 and RP11-426F15 that contained this marker could be unambiguously placed between markers 6 and 17.

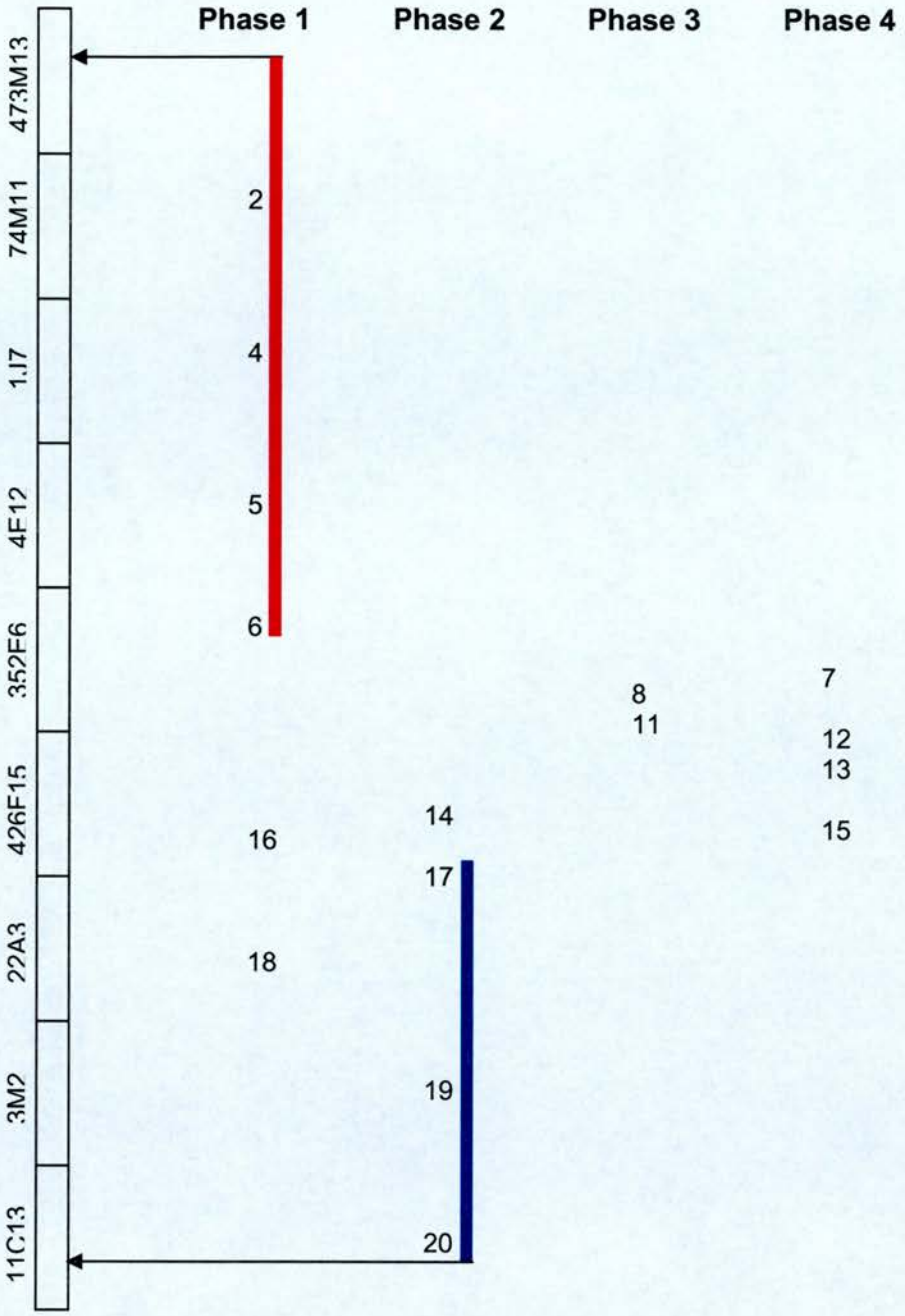


Figure 4-5: Four phases of genotyping were employed to refine the recombination breakpoint marking the centromeric end of the haplotype inherited with psychiatric illness in family 59 (delineating the centromeric boundary of Minimal Region 1). A minimal tiling path of nine BACs (RP11) between RP11-473M13 and RP11-11C13 was chosen between the current recombinant and non-recombinant markers (arrows). The relative position of each marker (number) in each phase is shown. Phase 1 resulted in the extension of the disease associated haplotype (red). Phase 2 resulted in the extension of the recombinant haplotype (blue). Phases 3 and 4 did not result in further refinement of the breakpoint.

Unfortunately, this sequence fragment did not contain any putative microsatellite repeats. It did, however, contain a number of putative SNPs identified in dbSNP.

In phase three of genotyping, two STSs were designed to flank four of these SNPs. STS st585142.snp contained markers 8 and 9, and STS st585128.snp contained markers 10 and 11. After PCR and sequencing, marker 8 and marker 11 were found to be polymorphic in the sample. However, neither SNP is informative for the recombination breakpoint because the genotype of parent 2078 is homozygous. At this point the recombination breakpoint could not therefore be resolved any further. The best option was to wait for the BAC sequences to be completed by the human genome sequencing project, and then to identify and test further putative polymorphic repeats in the interval.

4.4. Phase 4: Further Refinement of the Recombination Breakpoint

The sequence of both BACs was complete at February 2003. Markers 6 and 17 that flank the recombination breakpoint were found to be ~155kb apart.

4.4.1. Family Members

The markers were tested directly on the Allele Sharing (AS) panel (Figure 2-5) and F59-2081 which is not on the AS panel. This is a panel of 46 individuals from each of the four families. It enables the recombination breakpoint to be analysed but also enables allele sharing between the families to be assessed as well as providing a large population to interpret individual marker characteristics from (Section 4.2.1).

4.4.2. Markers

Four markers were genotyped in phase four. I designed markers 7, 12 and 15 to putative polymorphic repeats identified by Sputnik. In addition, I typed a publicly

available polymorphic marker (13) that was in the interval. I also inspected the region for SNPs identified by dbSNP that were located within existing STSs in ACeDB. Unfortunately there were none.

4.4.3. Genotyping

Microsatellite markers 7, 12, 13 and 15 were tested on three control DNAs to establish PCR conditions. They were diluted appropriately and genotyped on the ABI 3730 DNA sequencer and analysed using GeneScan version 3.0 software. Markers were multiplexed where possible. Genotypes were scored blind and segregation was checked before using the results.

4.4.4. Results

Figure 4-5 details the position of these phase 4 markers, and Figure 4-6 shows the genotyping results in the family. Marker 7 was uninformative in F59. From the AS panel, only two alleles were observed for this marker, and the minor allele was only observed four times in F22 and once in F59. Markers 12, 13 and 15 were polymorphic in the family but unfortunately the genotype of the transmitting parent 2078 was homozygous and therefore uninformative. Therefore the recombination interval could not be refined further as a result of genotyping these four markers.

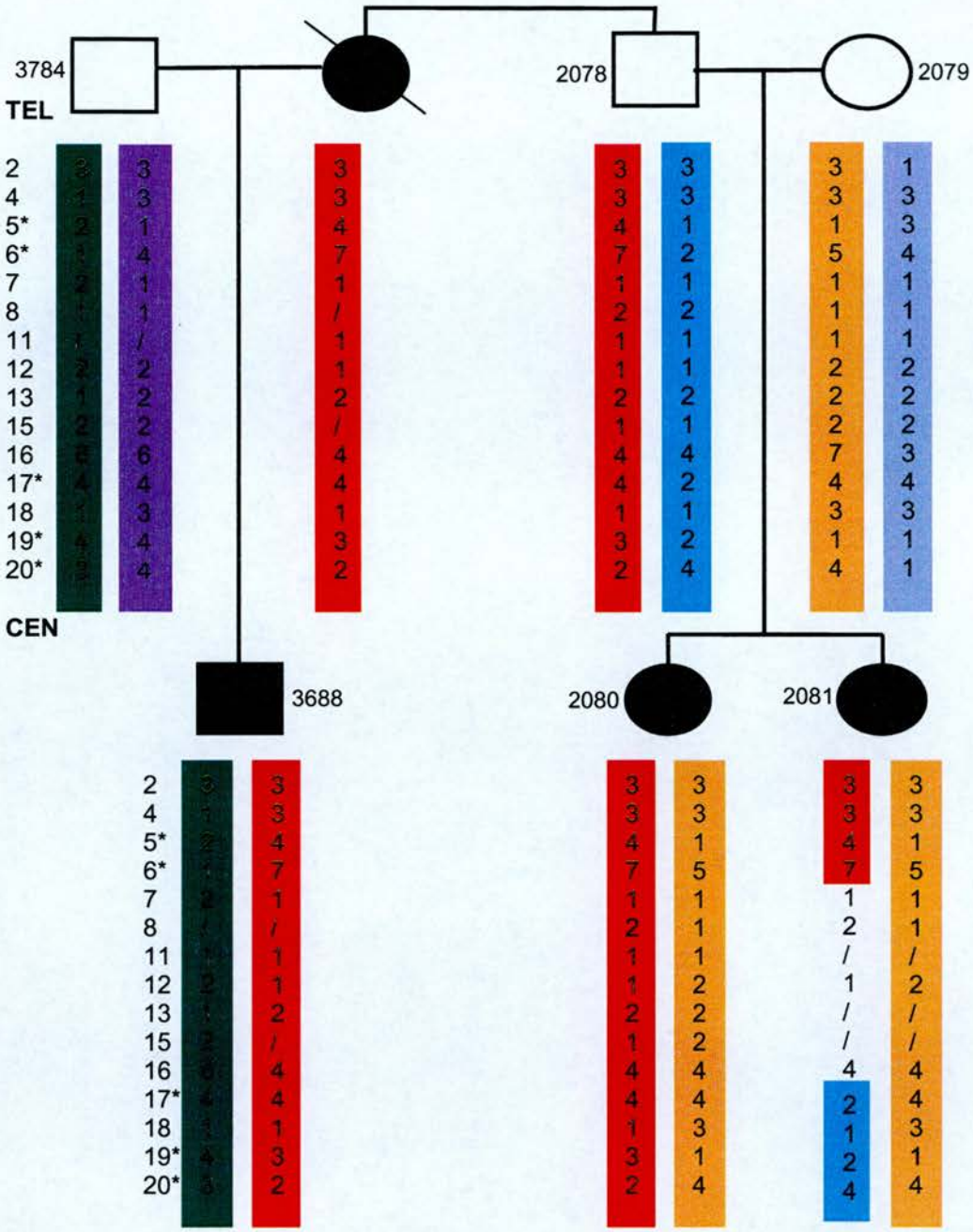


Figure 4-6: The marker haplotypes for 13 microsatellites and two single nucleotide polymorphisms in family 59. A recombination breakpoint in 2081 is revealed. Black infill = BPAD. Unfilled = no diagnosis. The crossed through individual is deceased, and the haplotype inferred. Markers with a star are fully informative for the breakpoint (marker number refers to Table 4-2). Coloured bars represent each of the different haplotypes and show the inheritance pattern. Red = disease associated haplotype because it is inherited in all cases of psychiatric illness. Offspring 2081 inherits part of the disease and part of the non-disease associated haplotype. It is not possible to determine which haplotype 2081 inherits between markers 6 and 17 (~155kb) with these marker genotypes. / = missing data.

4.5. F50 Recombination Breakpoint

The recombination breakpoint in F50 constitutes the telomeric boundary of MR1. At the start of my PhD the recombination interval lay between markers stb301J10.p3 and stb149b15_3p24 (Figure 4-1) that map to clones RP11-301J10 and RP11-751L19 respectively. The exact distance between these markers is not known because there is a gap both in our inhouse contig and in the publicly available sequence. However, FISH mapping estimated the gap to be ~300kb (Evans *et al*, 2001) and therefore the distance between the two markers was approximately 800kb.

The November 2002 release of the UCSC Golden Path positioned four clones (RP11-180A12, RP11-626O20, RP11-264E23 and CTD-2205P10) into the gap on the centromeric side (Figure 4-7). In chapter 3 I describe how I confirmed that they had been positioned correctly by bioinformatic analysis and a series of colony PCR experiments. These four clones provided a new resource from which to design microsatellite markers to refine the recombination breakpoint.

The F50 recombination breakpoint interval is important because it contains two known genes. The carboxypeptidase Z (CPZ) and orphan g-protein-coupled receptor 78 (GPR78) genes could either be included or excluded from MR1 depending on where recombination has occurred. CPZ, GPR78 and the recombinant marker st301J10.p3 are all positioned on BAC RP11-301J10. At the start of this project the sequence of this BAC was incomplete. However, by analysing the contig in SAM it was possible to determine the location of GPR78 and CPZ relative to the recombinant marker st301J10.p3. In the contig, marker SHGC18179 was telomeric to every other marker that had been tested for hybridisation to clone RP11-301J10. This marker lay within the same sequence fragment as the non-recombinant marker stb301J10.p3. Therefore, both genes were likely to be centromeric to this non-recombinant marker. This positioning was upheld with the release of the completed sequence of RP11-301J10.

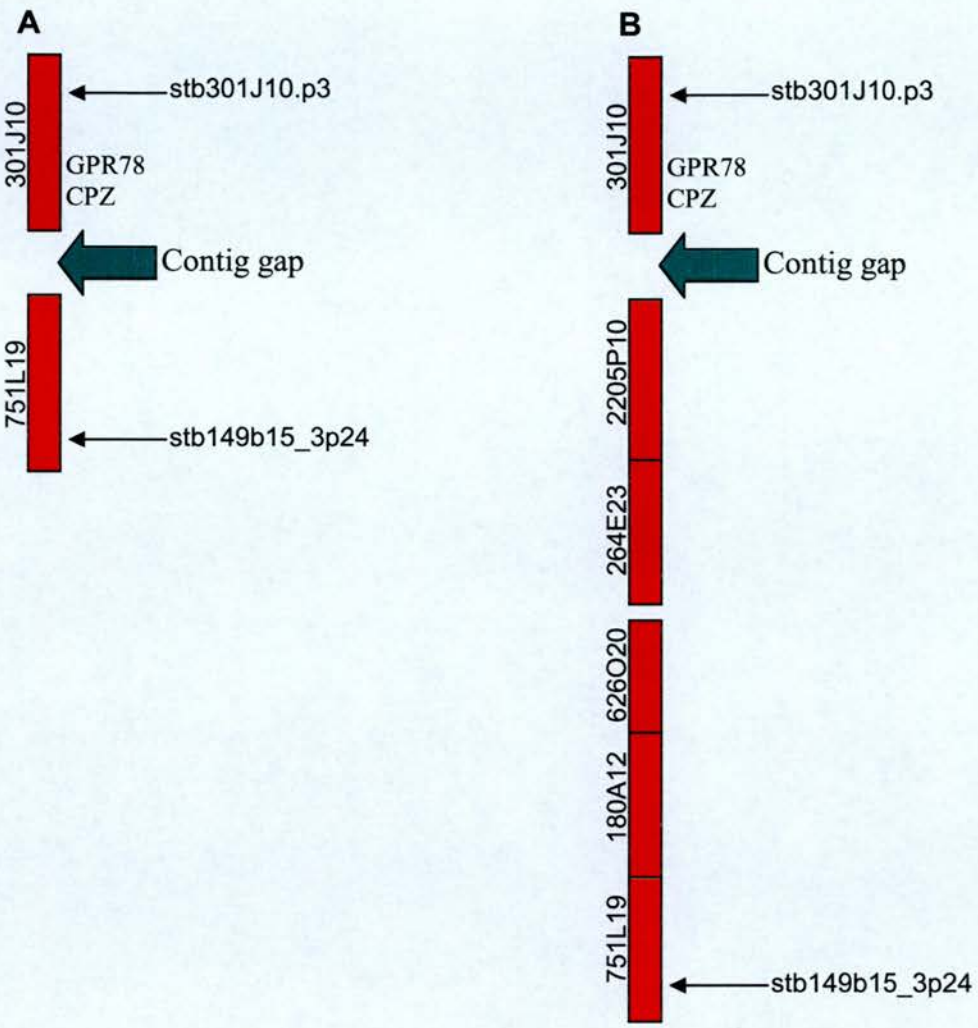


Figure 4-7: The telomeric recombination breakpoint region of Minimal Region One defined by family 50. **A.** At the start of the project only the two clones containing the non-recombinant (stb149b15_3p24) and recombinant (stb301J10.p3) marker flanking the recombination breakpoint were correctly positioned. **B.** Subsequently, the position of four clones was confirmed in the gap on the centromeric side. All clones are prefixed RP11-, except CTD-2205P10.

4.5.1. Family Members

The AS DNA panel was developed during these experiments, and therefore each marker was either typed on a selection of family members or the AS panel. Details for each marker can be found in Table 4-2. Eight of the 10 markers were typed on the AS panel, and two were typed on a selection of family 50, 59 and 22 members.

4.5.2. Markers

4.5.2.1. Microsatellite markers

As mentioned, the recombination breakpoint had previously been defined by others to between telomeric non-recombinant marker stb301J10.p3 and centromeric recombinant marker stb149b15-3p24 (Figure 4-1). The publicly available BAC contig spanning this interval was incomplete at the start of this project. Today (February 2004), the BACs within the interval have been fully sequenced but one contig gap remains.

The details of the markers can be seen in Table 4-2. In summary, ten microsatellite repeats were typed. Markers were designed to flank putative polymorphic repeats identified by Sputnik and Repeat Masker, except marker 34. This marker was identified by bioinformatic analysis carried out by Dr. Colin Semple from a fragment of Celera sequence Ce:GA_x54KRCDMTNS.1-112114. This lies beyond the limits of the public ally available and inhouse contig on the telomeric side. Two of the markers (3 and 38) had been previously identified, designed and tested for polymorphism by others in the group.

I described in Chapter 3 how the sequence on the centromeric side of the gap is highly repetitive, and it may be that this is prohibiting the identification of BACs that span the gap. As discussed previously, the human genome sequencing project sequence releases at UCSC have varied with respect to one another in this region.

The April 2003 release positioned four clones on the centromeric side of the gap and I confirmed their position by colony PCR (Chapter 3). This provided me with an additional resource for identification of microsatellite markers. However, the repetitive nature of these four clones makes it difficult to design any markers around putative polymorphic sequences, as many potential primers are likely to amplify from more than one chromosome.

Six of the ten markers (35, 36, 37, 38, 43 and 44) are in this highly repetitive region. Primers were designed by performing a BLASTn similarity search of the region surrounding the putative microsatellite repeat. The BLASTn alignment was used to identify highly similar clones from other genomic regions and then to design primers with mismatches to these highly similar regions at their 3' ends. The markers were then tested for specificity to chromosome 4 on a somatic cell monochromosomal hybrid (MCHP) panel. For example, for marker 38, I performed a BLASTn of a 769bp region around the microsatellite repeat. The results showed that it was greater than 94% similar to 11 clones from other chromosomes. I designed two pairs of primers with mismatches at their 3' end to every similar clone. Primers to five other markers in the repetitive region were designed in a similar way.

These six markers were then tested to obtain optimal PCR conditions on three control DNAs and then tested on the MCHP for specificity to chromosome 4. Marker 35 amplified from only chromosome 4 using the optimal PCR conditions determined. Marker 36 was not specific to chromosome 4. However, it is very close to marker 37 and therefore I did not pursue this marker. Marker 37 was not specific to chromosome 4 using the optimal PCR conditions. I redesigned the primers and these were specific to chromosome 4 under PCR conditions established from a control DNA test. However, under these conditions, the marker did not amplify efficiently from the AS panel. I performed a second PCR, lowering the annealing temperature by 2°C, and ran both the high and the low stringency PCR products on the ABI 3730 Genetic Analyser. The samples that had worked in both PCRs had the same

genotypes. Therefore I decided that the samples that had only worked in the lower stringency PCR were also reliable.

Marker 38 amplified only from chromosome 4 using the PCR conditions determined from the control DNA test. However, after PCR on the AS panel, the signal intensity of the product on an agarose gel was very low. Therefore I performed another PCR on the AS panel with 40ng of DNA in the PCR instead of 20ng. The signal intensity of the product on an agarose gel was much stronger. However, it was possible that providing excess DNA in the PCR could have enabled the primers to anneal to other chromosomes. Therefore I ran both high and low stringency PCRs on the ABI 3730 Genetic Analyser and compared the results of the samples that had worked in both conditions. There was complete agreement of the results from both PCRs and therefore the extra genotypes obtained from the PCR with double the amount of DNA in the reaction were deemed reliable.

Using the PCR conditions determined from the control DNA test, marker 43 amplified from only chromosome 4 using the MCHP. After PCR on the AS panel, these conditions resulted in non-specific bands when the product was run on an agarose gel. Therefore this marker was discarded. Exactly the same was observed for marker 44. However, on an agarose gel, only one extra band of approximately double the expected size was observed. Therefore I ran this on the ABI 3730 Genetic Analyser and was careful to check segregation before accepting the results as reliable.

Table 4-2 (continued overleaf): Details of the markers used to refine the family 50 recombination breakpoint. The breakpoint marks the telomeric end of the haplotype inherited with psychiatric illness in family 50 and delineates the telomeric boundary of Minimal Region 1. The table shows a tiling path of clones identified between the current recombinant and non-recombinant markers and the markers designed from them. The marker number refers to Figure 4-7. The distance alongside a marker refers to the distance between it and the preceding marker. Distance is measured from the first nucleotide of the forward primer for each marker. It is not possible to determine the distance between markers 31 and 32 because there is a contig gap in between them. Microsat: microsatellite. SNP: single nucleotide polymorphism. AS: Allele Sharing DNA panel. Le Hellard: SNPs genotyped by Dr S. Le Hellard. Not specific: primer pair amplifies other chromosomes in addition to chromosome 4.

Clone (RP11-)	Marker name	Marker Number	Marker Type	Population Typed	Note	Distance (bp)
689P11	St301J10.p3	1	Microsat	-		-
689P11	St689P11.p2b	2	Microsat	AS		54 055
689P11	SG4961	3	Microsat	F50/59/22	Not polymorphic	1321
689P11	ih164	4	SNP	AS		2428
689P11	Ih163	5	SNP	AS		111
689P11	Ih162	6	SNP	AS		28
689P11	Ih165	7	SNP	AS		173
689P11	Ih161	8	SNP	AS		106
689P11	Ih33	9	SNP	AS		229
689P11	Ih32	10	SNP	AS		114
689P11	Ih31	11	SNP	AS		456
689P11	Ih48	12	SNP	AS		777
689P11	Ih47	13	SNP	AS		6
689P11	Ih46	14	SNP	AS		4534
689P11	Ih45	15	SNP	AS		56
689P11	Ih34	16	SNP	AS		338
689P11	Ih44	17	SNP	AS		132
689P11	Ih43	18	SNP	AS		51
689P11	Ih42	19	SNP	AS		35

689P11	Ih41	20	SNP	AS		22
689P11	Ih40	21	SNP	AS		19
689P11	Ih36	22	SNP	AS		66
689P11	Ih37	23	SNP	AS		68
689P11	Ih39	24	SNP	AS		20
689P11	Ih38	25	SNP	AS		41
689P11	Ih35	26	SNP	AS		237
689P11	rs2302583	27	SNP	Le Hellard		4860
689P11	Ih17	28	SNP	Le Hellard		3
689P11	Ih18	29	SNP	Le Hellard		8540
689P11	Ih27	30	SNP	Le Hellard		2816
689P11	rs2302578	31	SNP	Le Hellard		3046
689P11	Ih25	32	SNP	Le Hellard		7029
689P11	st689P11.p1	33	Microsat	AS	Not polymorphic	29 647
-	stCeGax54.p1	34	Microsat	F50/59/22		-
CTD-2205P10	st2205P10.3	35	Microsat	AS		-
CTD-2205P10	st2205P10.p2	36	Microsat	-	Not specific	316
CTD-2205P10	St2205P10.p1b	37	Microsat	AS		25612
264E23	st264E23.p1b	38	Microsat	AS		99 561
180A12	rs7693695	39	SNP	F50		-
180A12	Ih158	40	SNP	F50		38
180A12	Ih159	41	SNP	F50		9
180A12	Ih160	42	SNP	F50		85
180A12	st180A12.p2	43	Microsat	AS	Not specific	409
180A12	st180A12.p1	44	Microsat	AS		44 259
180A12	rs6448969	45	SNP	F50		25 876
180A12	rs7434710	46	SNP	F50		6
752L19	st149b15_3p14	47	Microsat	-		65 139

4.5.2.2. Single Nucleotide Polymorphism Markers

Details of the SNP markers can be found in Table 4-2. The GPR78 gene lay within the recombination interval as defined prior to the start of my project. I identified 23 SNP's in the GPR78 gene (markers 4-26) by sequence analysis of family members using the AS panel. SNPs were genotyped by direct inspection of the sequence traces. The CPZ gene is the second known gene in the recombination interval. Stephanie Le Hellard had previously identified six SNP's from the CPZ gene (markers 27-32) by sequence analysis of members of F22, F50 and F59.

Seventeen primer pairs developed in house by others for other experiments contained putative SNPs in dbSNP. The SNPs in five STSs had been previously tested by others for polymorphism in F50 and they were either uninformative or not polymorphic. Therefore PCR conditions were tested for twelve STSs on three control DNAs. These conditions were then used to test for specificity to chromosome 4 on the MCHP since they were in the repetitive region. Seven STSs were specific to chromosome 4 using the PCR conditions optimised on control DNA.

These seven STSs contained a total of 17 SNPs identified in dbSNP. A PCR of each STS was performed on seven individuals from F50: -1, -2, -3, -4, -11, -12 and -15. The PCR products were sequenced and the sequences examined for the presence of SNPs. I identified six SNPs in F50. SNPs 39-42 were identified in STS st180A12.m1, and SNPs 45 and 46 were identified in STS stc202024.2. A BLASTn of these two STSs revealed that they were highly similar to other genomic regions. The BLASTn of the 736bp STS st202024.2 revealed that there were multiple clones showing 90% similarity or greater but that none of these aligned to more than a 652bp stretch of continuous sequence. A BLASTn of the 550bp STS st180A12.m1 revealed that clone RP11-10K17 from chromosome 16 was 93% similar; a difference of 39bp. Therefore, this result in addition to the specificity of the STSs to chromosome 4 on the MCHP led me to conclude that the six SNPs identified were real rather than the result of amplification of a different genomic region.

4.5.3. Genotyping

Markers were tested on three control DNAs to establish PCR conditions. They were diluted appropriately and run on an ABI 310, 377 or 3730 Genetic Analyser and analysed using the GeneScan version 3.0, or GeneMapper version 3.0, software. Markers were multiplexed where possible. SNPs were genotyped by direct sequencing. After PCR, sequence chromatograms were aligned using the phredPhrap software, visualised with the Consed programme and manually inspected for polymorphism. Genotypes were scored blind to phenotype and family relationship and then Mendelian segregation was checked.

4.6. Definition of the F50 Recombination Breakpoint Interval

Six microsatellite markers and 35 SNPs were used to analyse the telomeric recombination breakpoint of MR1 as defined by F50-12 (Figure 4-8). The results showed that individual 12 inherits the disease associated chromosome from marker 37 to marker 47. This extends the disease associated haplotype as defined at the start of my project by a further 336kb.

The results were interpreted as follows. The transmitting parent F50-3 is heterozygous for marker 37. F50-4 is also heterozygous for marker 37. The genotypes of the two offspring 12 and 15 allow for the correct interpretation of which chromosome the recombinant individual F50-12 has inherited. F50-15 inherits the green and the grey chromosomes, and has a homozygous genotype for marker 37. Therefore, the grey and the green chromosome must each contain allele 1. F50-12 has a heterozygous genotype for marker 37, and inherits the grey chromosome from parent F50-4, which as has been seen from F50-15 must contain allele 1. Therefore, allele 2 must have been inherited from parent F50-3. It can also be seen that F50-15 inherits the green chromosome from parent F50-3 and that this also contains allele 1. Therefore, allele 2 in F50-12 must be on the red disease associated chromosome. The argument also follows for markers 38, 39, 41, 42 and 45.

The results mean that the GPR78 and CPZ genes in the recombination interval cannot be ruled in or out of MR1. In addition, there are a number of putative genes in the region centromeric to marker 37 which are now included in MR1 as they lie within the newly defined region of disease associated chromosome.

By sequencing seven members of F50, seven STSs, that contained 17 putative SNPs, actually contained only six that were polymorphic in the family. This was probably due to two reasons. Firstly, my sample number is too small to represent all variation. Secondly, five of the STSs are in the repetitive region centromeric to the gap. Therefore the SNPs in the public database that have been localised to this region of chromosome 4 could in fact be due to the amplification of other chromosomal regions.

The homozygosity of F50-3 for markers 2-35 makes the determination of a recombination breakpoint in F50-12 impossible. This region of homozygosity extends over 203kb of known sequence between heterozygous markers 1 and 37, including an undetermined distance in the contig gap. This is investigated further in the following section.

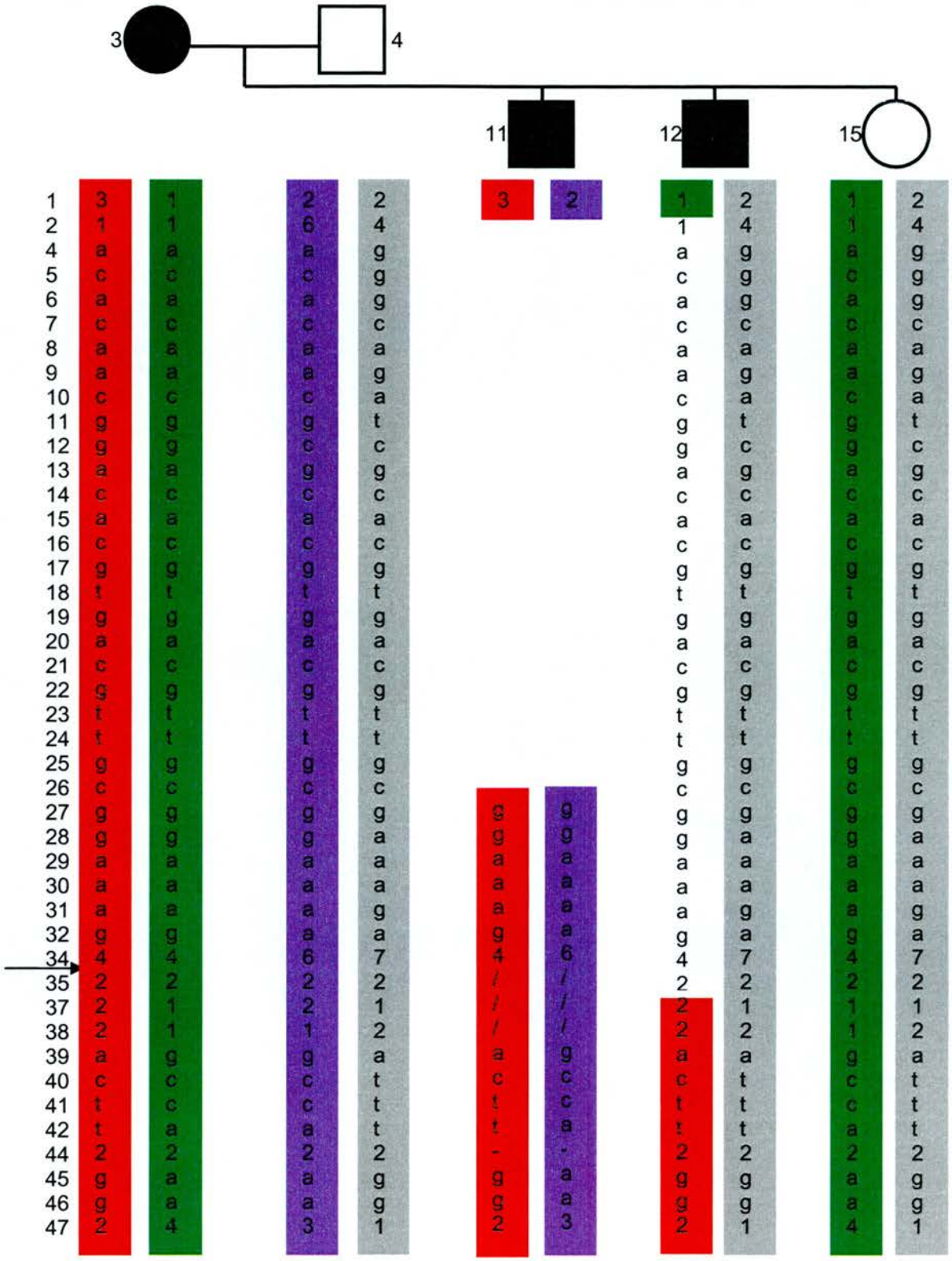


Figure 4-8: The marker haplotypes for six microsatellites and 35 single nucleotide polymorphisms in family 50. Black infill: schizoaffective disorder. Unfilled: well. Colours represent each of the different haplotypes to show inheritance pattern. Red: disease associated haplotype as it is inherited with a psychiatric illness. F50-12 inherits part of the disease and non-disease haplotypes, but it is not possible to determine inheritance between markers 2 and 35. / = missing data. Marker number as Table 4-3. Arrow: sequence gap.

4.7. Investigating the Homozygosity of F50-3

It seems unlikely that the genotypes of markers 2 to 35 of the transmitting parent F50-3 should all be homozygous. It is possible that there has been a loss of heterozygosity in this region. Any such region would have to extend over 203kb of known sequence and possibly includes the sequence in the contig gap.

The current data shows that a loss of heterozygosity would have to have occurred spontaneously in F50-3 rather than having been inherited from the parents F50-1 or -2. This is because over the length of homozygosity in F50-3, both F50-1 and -2 have heterozygote genotypes at a number of markers (Figure 4-9).

By inspecting the genotypes of F50-3, -4, -12 and -15 in Figure 4-8, it is possible to see that F50-15 inherits the green chromosome from F50-3, and is heterozygous for markers 2, 4, 5, 6, 9, 10, 11, 12, 13, 27, 31 and 32. This would not be possible if the green chromosome had been lost from F50-3. Therefore, if a loss of heterozygosity has occurred it must be on the disease associated chromosome. F50-12 also has heterozygous genotypes for markers 2, 4, 5, 6, 9, 10, 11, 12, 13, 27, 31 and 32, but, because we don't know if this individual inherits the red or the green chromosome, we cannot say for certain that the red disease associated chromosome is not missing in the F50-3. Furthermore, even if F50-12 has inherited a red disease associated chromosome from F50-3 for markers 2, 4 -3, and 28-46 it still leaves a section of homozygosity, from marker 14-27, which could still, theoretically, be lost in F50-3. However, this is unlikely because 13 of these 14 markers only span ~1kb (Table 4-2). Markers 14-26 were typed on the AS panel and the minor alleles of these 13 SNPs form a single haplotype, occurring on only five chromosomes in family 22 and not once in family 50, 59 or 48, giving a haplotype frequency of 11.4%. Therefore, a homozygous genotype is not unusual for these SNPs.

The region of homozygosity in F50-3 from marker 2-34 flanks the telomeric side of the contig gap and therefore could extend into this gap. One side of the contig gap

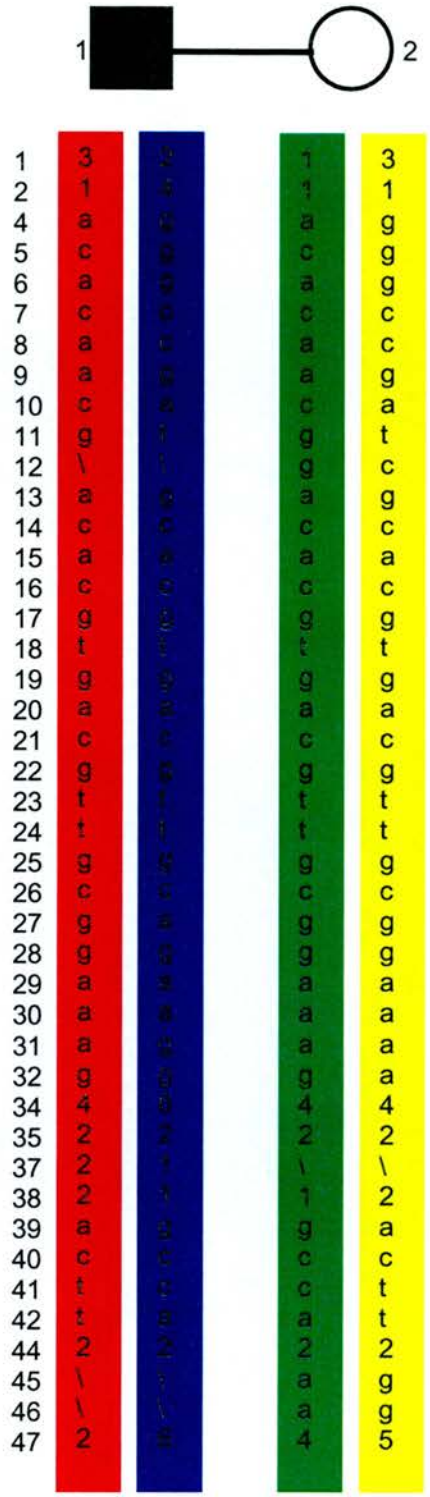


Figure 4-9: The marker haplotypes for six microsatellites and 35 single nucleotide polymorphisms in family 50 members, F50-1 and F50-2, the parents of F50-3 (Figure 4-6). Black infill: schizoaffective disorder. Unfilled: no diagnosis. Colours represent each of the different haplotypes. Red: disease associated haplotype since it is inherited in all cases of psychiatric illness. Marker number refers to Table 4-3. / = missing data.

is a region that appears to be poorly represented in genomic libraries (Evans *et al*, 2001_a). The other side of the contig gap is highly repetitive. It might be in just such regions that during DNA replication errors could lead to the elimination of stretches of DNA.

4.7.1. Assay to Detect Chromosome Loss in F50-3

Two somatic cell hybrids, one derived from F50-1 and another derived from F50-3, had previously been made in the lab by Ian Anderson. Four hybrids which retained one copy of chromosome 4 for individual F50-1 had already been identified. The hybrids retaining one copy of chromosome 4 for F50-3 had not been identified. Therefore I genotyped 24 hybrids derived from F50-3 with five chromosome 4 microsatellite markers for which F50-3 is heterozygous. Three markers showed clearly which hybrids were haploid and which were diploid (Figure 4-10). The PCR failed for one marker and the other marker appeared to be completely homozygous suggesting a mistyping. Out of 24 hybrids, three hybrids had segregated one chromosome 4 (A) and one hybrid had segregated the other chromosome 4 (B) (Table 4-3). I amplified six STSs from GPR78 and CPZ, in the region of homozygosity in F50-3, from the four haploid hybrids in F50-3 and the four haploid hybrids in F50-1 (Figure 4-11). Each of the STSs amplified both chromosomes and therefore did not support the hypothesis of a loss of heterozygosity.

Marker	Chromosome A	Chromosome B
stD4S2633	204	208
st473mca.5b	289	293
st481L13	245	243

Table 4-3: Genotyping three chromosome 4p microsatellite markers on 24 somatic cell hybrids of individual F50-3. F50-3 is heterozygote for the three markers after genomic DNA PCR. Therefore, the hybrids retaining one chromosome 4 are determined from a homozygous genotype. Of the 24 somatic cell hybrids, four hybrids retained one copy of each chromosome 4 (Chromosome A and B).

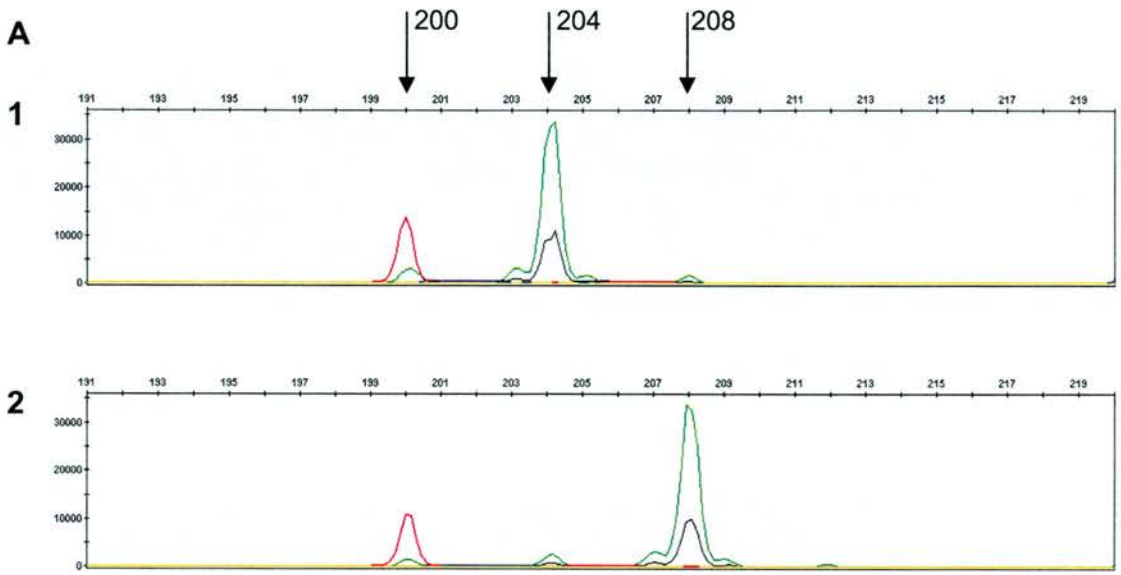
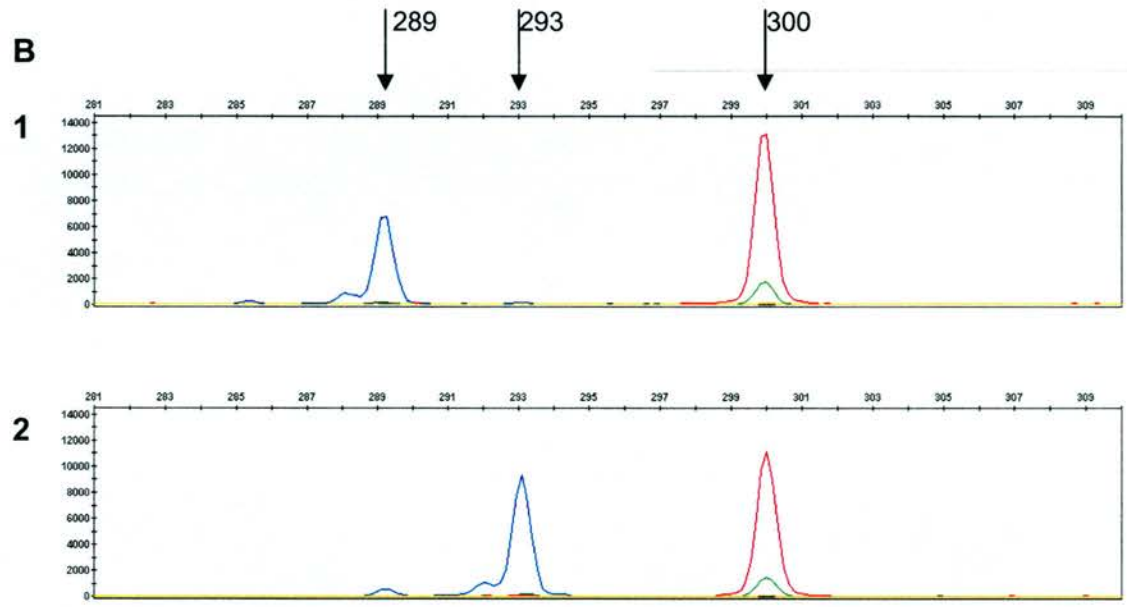
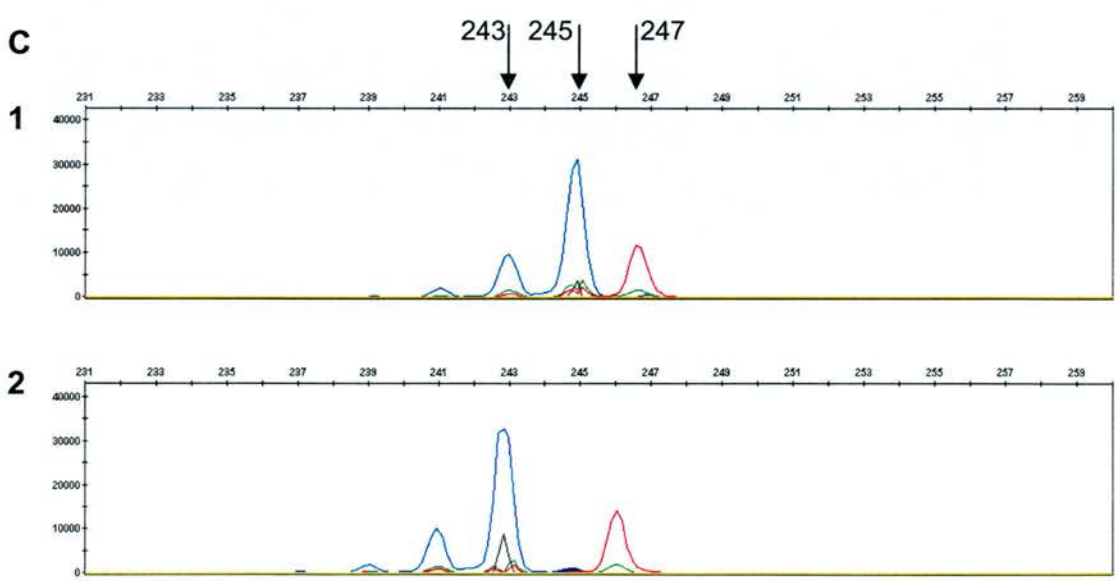


Figure 4-10 (continued overleaf): Fluorograms (displayed on the GeneMapper 3.0 software, Applied Biosystems) for three chromosome 4p microsatellite markers genotyped on somatic cell hybrids of individual F50-3. F50-3 is heterozygote for the three markers after genomic DNA PCR. Therefore, the hybrids retaining only one chromosome 4 are determined from a homozygous genotype. Of the 24 somatic cell hybrids, four hybrids retained one copy of each chromosome 4. The size standard marker is shown by the red peak in each case. The size of each allele (base pairs) is marked above. The figure shows the results for two of the hybrids, each retaining one chromosome 4. **A.** Marker stD4S2633. **A1.** hybrid displaying chromosome A genotype. **A2.** Hybrid displaying chromosome B genotype. **B.** Marker st473mca.5b. **B1.** Hybrid displaying chromosome A genotype. **B2.** Hybrid displaying chromosome B genotype. **C.** Marker st481L13. **C1.** Hybrid displaying chromosome A genotype. **C2.** Hybrid displaying chromosome B genotype. Despite the single predominant genotype in each fluorogram, it is still possible to see the other allele at a low level (except C1 where the stutter peaks of the major allele masks the region).


B
289
293
300
1
2


C
243
245
247
1
2

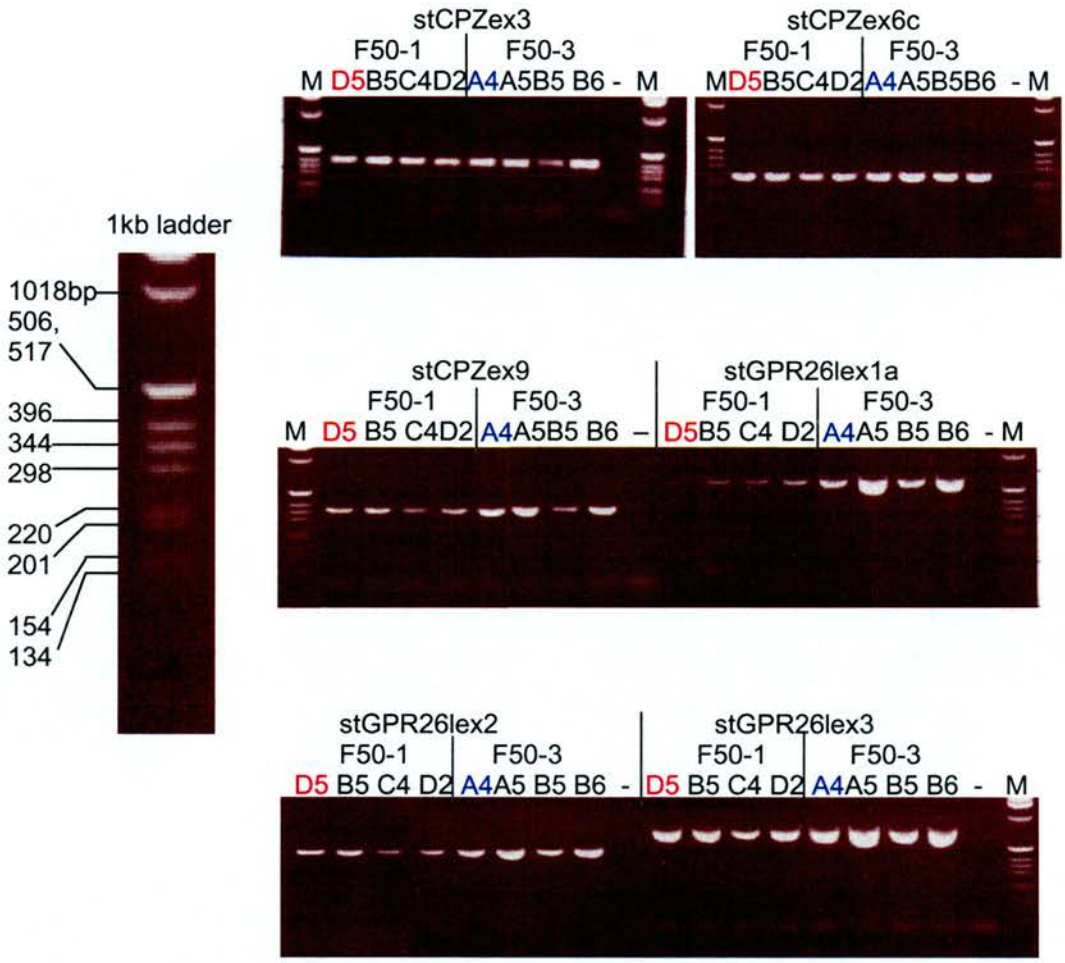


Figure 4-11: Agarose gels showing amplification of three STSs from the orphan G-protein-coupled receptor 78 gene and three STSs from the carboxypeptidase Z gene from four F50-1 and four F50-3 somatic cell hybrids that each retain only one chromosome 4. Since F50-1 is the parent of F50-3, this is a total of three chromosomes (A. B or C). Chromosome A is inherited from parent F50-1 to offspring F50-3. Three F50-1 hybrids retain chromosome A (B5, C4 and D2) and one retains chromosome C (D5). One F50-3 hybrid retains chromosome B (A4), and three retain chromosome A (A5, B5 and B6). Amplification was performed to test for the presence of both copies of chromosome 4p in the region containing these two genes in both individuals. A lack of amplification would suggest the deletion of this part of the chromosome. As can be seen, all hybrids amplify for each individual, although not equally. M = ReadyLoad™ 1 kb DNA ladder (Invitrogen).

4.7.2. PCR Assay Limitations

Although the majority of cells in a hybrid cell line will be monovalent for a given chromosome, a small number of cells will be bivalent. Therefore, the use of PCR is not ideal because sometimes only a very small amount of template DNA will result in successful amplification. I noticed that for four of the six STSs, the PCRs did not amplify equally for each hybrid, and that the same amplification pattern was observed for each STS. In Figure 4-11 it is possible to see that the two F50-3 hybrids A4 and B5 (chromosome A) amplify less efficiently than samples A5 or B6 for four of the six STSs. Therefore a re-examination of the genotyping gel for F50-3 (Figure 4-10) was undertaken to see if there was a correlation between how well the hybrid PCR worked on the STS compared to how large the peak of the 'contaminating' chromosome was in the genotyping gel. As can be seen from Figure 4-10, this was only possible for markers A and B, since the 'contaminating' peak of marker C was masked by the stutter peak of the major allele. Table 4-4 notes the peak height of the 'contaminating' peak. I found that the 'contaminating' peaks for hybrid A4 and B5 were smaller than A5 and B6 for marker B but not marker A. In the absence of more data, it was not possible to determine from this whether the presence of the other chromosome in the hybrid in Figure 4-10 could account for the efficiency of the PCR assay in Figure 4-11.

	Peak Height of Second Allele	
F50-3 Chromosome	**st473mca.5b**	**stD4S2633**
A (A4)	61	1667
B (A5)	586	12 224
B (B5)	308	10 601
B (B6)	1666	2528

Table 4-4: Three chromosome 4p microsatellite markers were genotyped on somatic cell hybrids of individual F50-3. F50-3 is heterozygote for the markers on genomic DNA. Hybrids retaining only one chromosome 4 will have a homozygous genotype. Of 24 somatic cell hybrids, four retained one copy of each chromosome 4 (A4, A5, B5 and B6). However, on a fluorogram (Figure 4-10), a peak the size of the second allele is present, suggesting the presence of the second chromosome. The table shows the peaks heights of the second allele for two of the three markers.

4.8. Discussion

The underlying biology of psychiatric disorders is largely unknown and the mechanism by which current treatments operate is also unclear. Therefore, virtually any gene can be proposed as a candidate, making candidate gene selection from large genetic susceptibility regions problematic. The disease associated haplotypes of the four families studied overlap. This has provided a suitable way of reducing the candidate intervals, assuming that the same susceptibility gene operates in the families. This has revealed four priority regions. It is very important to define the recombination breakpoints and these overlapping regions as precisely as possible for inclusion and exclusion purposes. It is also important to define each family disease associated haplotype as precisely as possible because they may have different ancestry.

Here I describe the genetic analysis of MR1, a candidate region for susceptibility to psychiatric illness. MR1 is defined by the overlapping haplotypes from families F22, F59 and F50. This is an important region because all three families are of Celtic ethnicity, which makes a common ancestor hypothesis more likely. In addition, association has previously been found between MR1 markers in the region of the dopamine D_5 receptor (DRD5) candidate gene and SCZ (Muir *et al*, 2001).

As a result of typing markers in family members I have refined the centromeric recombination breakpoint of MR1 to within 155kb. At present there are no known genes within this interval and therefore further refinement is unnecessary. I have also extended the disease associated haplotype in the telomeric recombination breakpoint of MR1 by a further 336kb. The exact size of the new recombination interval is unknown because there is a contig gap between the recombinant and non-recombinant marker. However, it can be estimated at approximately 500kb. At present there are two known genes, CPZ and GPR78, within this interval, and therefore further refinement is desirable. However, this will be made difficult by the repetitive nature of the sequence in the region.

A disadvantage of recombination breakpoint mapping is that recombination is a relatively rare event. The breakpoints the create MR1 are defined by only one individual in each family. Consequently, the size of this region is still large. Furthermore, in a smaller family the recombinant individual could be a phenocopy, sharing a haplotype by chance. With these limitations in mind it is still worthwhile refining an already large disease associated haplotype in a family to include or exclude potential candidate genes.

The possibility that part of one of the chromosomes in individual F50-3, the parent that transmits the recombination breakpoint, has been lost has been shown to be extremely unlikely. Furthermore, the possibility of loss of heterozygosity does not impinge on the disease process because the chromosome loss could not have arisen in individual F50-1. Therefore it is not the cause of illness in the family.

There is evidence that whole genome amplification (WGA) increases homozygosity (S. LeHellard, personal communication) which, whilst not desirable normally, would in this case perhaps reduce the problem associated with using a PCR based assay on a hybrid cell line. An interesting future experiment could be to subject the somatic cell hybrid of F50-3 to WGA, as this may produce a more reliable sample for PCR analysis of partial chromosome loss.

In conclusion, I have refined the recombination breakpoints of MR1, one of the two best candidate regions for the susceptibility to psychiatric illness in four families. Future work should aim to refine the telomeric boundary further because it contains two known genes.

Chapter Five

Transcript Map of Two Candidate Regions for Psychiatric Illness on Human Chromosome 4p

Transcript Map of Two Candidate Regions for Psychiatric Illness on Human Chromosome 4p

5.1. Introduction

The production of sequence from the human genome sequencing project (HGP) facilitates the identification of genes and their regulatory regions. This will not only help to identify disease genes but also the biological processes underlying normal cellular function.

It has been possible to begin genome annotation with draft sequence. However, because draft sequence is inherently fragmentary, high quality finished sequence is preferable. The finished sequence was announced on the 14th April 2003. However, sequencing is still ongoing and a number of gaps remain. The Sanger Institute (February 2004) (www.sanger.ac.uk/HGP/) reports that 88.99% of the genome sequence is finished and that 4.23% is of draft quality, totalling 93.22% of the human genome. The latest University of California Santa Cruz (UCSC) (www.genome.ucsc.edu) sequence release (July 2003) claims to cover 99% of gene containing regions. The aim of the sequencing project is to achieve an accuracy of greater than 99.99%. In addition to aiding the discovery of all the genes in the human genome, this level of coverage and accuracy will permit the identification of variation between human genomes to identify disease genes and allow comparison with other species to identify regulatory elements and elucidate evolutionary processes.

Known and predicted genes are annotated by a mixture of expression evidence and bioinformatic prediction methods. Bioinformatic tools are informative for gene discovery, providing support to expression evidence and suggesting the presence of genes where no expression evidence is available. Three common gene and exon prediction programmes are Fgenes, GenScan and MZEF. The results of these are displayed in our inhouse database ACeDB. The gene prediction programme Fgenes (www.softberry.com), developed at the Sanger Institute uses pattern recognition to

detect promoters, exons and poly-adenylation signals based on the linear discriminant function (LDF) method. This calculates a score of any open reading frame (ORF), between a start and a stop codon and predicts an exon if the LDF is above a threshold. Each exon is considered in the context of possible 5' and 3' exons with respect to splicing donor and acceptor sites, the ORF, start and stop codons and poly-adenylation and promoter sequences. The objective is to find the arrangement with the maximal value from all possibilities. Fgenes does not perform well if there are several genes within the sequence being analysed, when the GC content is low, and if there are short (<25bp) exons. Furthermore, the prediction of terminal exons is poor resulting in genes being split or joined together. Therefore, Fgenes is most likely to correctly identify genes that have a standard intron and exon length and standard TATA boxes and poly-adenylation signals.

GenScan, developed at Stanford University (Burge and Karlin, 1997) is another commonly used gene prediction programme. It uses a similar method of gene prediction as Fgenes and suffers from a similar set of problems. MZEF, developed by Michael Zhang, is an internal coding exon prediction program. It identifies a potential exon by the presence of splice sites and an ORF, measures a set of discriminant variables and then calculates the exon probability. If the probability exceeds a threshold an exon will be predicted. As with GenScan and Fgenes, false positives are a problem. Comparison of the overlap between multiple programmes, particularly those that use different methodologies, is useful. However, evidence from gene and exon prediction programmes should not generally be taken alone.

With the HGP nearing completion, large scale cDNA sequencing projects are increasingly common and provide a similar large scale genomic approach to gene identification. Comparative genomics is also proving to be a valuable tool for gene identification since cross species sequence homology is often likely to reflect the evolutionary conservation of genes and regulatory regions. Gene prediction programmes can be validated where these lines of evidence co-occur.

It is very important in this project to identify all the genes in our linkage regions, especially our best candidate regions, minimal regions one (MR1) and two (MR2). These candidate regions are large and because a definitive theory of the pathological processes in psychiatric illness is lacking, candidate genes are hard to define. The aim is to perform association studies in MR1 and MR2 in order to focus future protein based studies. Therefore, in turn, a map of the genes in these regions would help to focus these association studies.

This chapter describes efforts to build a transcript map of MR1 and MR2. To achieve this I used a combination of bioinformatics methods, cDNA library screening and RT-PCR.

5.2. Defining the Region

5.2.1. ACeDB

ACeDB, described in chapter three, is the inhouse database that displays the genomic sequence of MR1 and MR2. The database is constructed and managed inhouse by S. Morris and displays the sequence of the clones that span MR1 and MR2. Sequenced clone are obtained from the HGP and are re-analysed for sequence overlap. Where necessary the clone order is adjusted to satisfy this more rigourus inhouse analysis. Therefore, ACeDB represents a further layer of analysis to that of the HGP. This was of greater importance in the past when the HGP sequence was fragmentary.

The sequence in ACeDB is masked for simple repeats, annotated with CpG islands, and the results of BLASTn similarity searches to human clones, vertebrate and invertebrate genomic DNA, vertebrate mRNA and protein, and the results of the gene and exon prediction programmes GenScan, Fgenes and MZEF (Table 5-1). Figure 5-1 shows an example of an ACeDB map view of a known gene. As can be seen, it provides a fully integrated view of a gene, displaying the BLASTn alignments of ESTs and mRNA, the results of gene prediction programmes and in inhouse data.

Annotation	Website
Simple repeat finder RepeatMasker	repeatmasker.genome.washington.edu/
Simple repeat finder Sputnik	cbi.labri.fr/outils/Pise/sputnik.html
CpG island	www.rfcgr.mrc.ac.uk/Registered/Help/alfresco/#cpg
Human clones	www.ncbi.nlm.nih.gov/
Human ESTs	www.ncbi.nlm.nih.gov/dbEST/
Human EST database Unigene	www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene
Human EST database Tigr	www.tigr.org/
Human EST database Stack	www.sanbi.ac.za/Dbases.html
Human protein database SwissProt	www.ebi.ac.uk/swissprot/index.html
Mouse DNA	www.ncbi.nlm.nih.gov/
Mouse ESTs	www.ncbi.nlm.nih.gov/dbEST/
Mouse Unigene	www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene
Exon predictor programme MZEF	sciclio.cshl.org/genefinder/
Gene predictor programme GenScan	genes.mit.edu/GENSCAN.html
Gene predictor programme Fgenes	www.softberry.com

Table 5-1: List of the programmes and databases used to annotate ACeDB, the inhouse database of annotated chromosome 4p sequence and the website from which they are available.

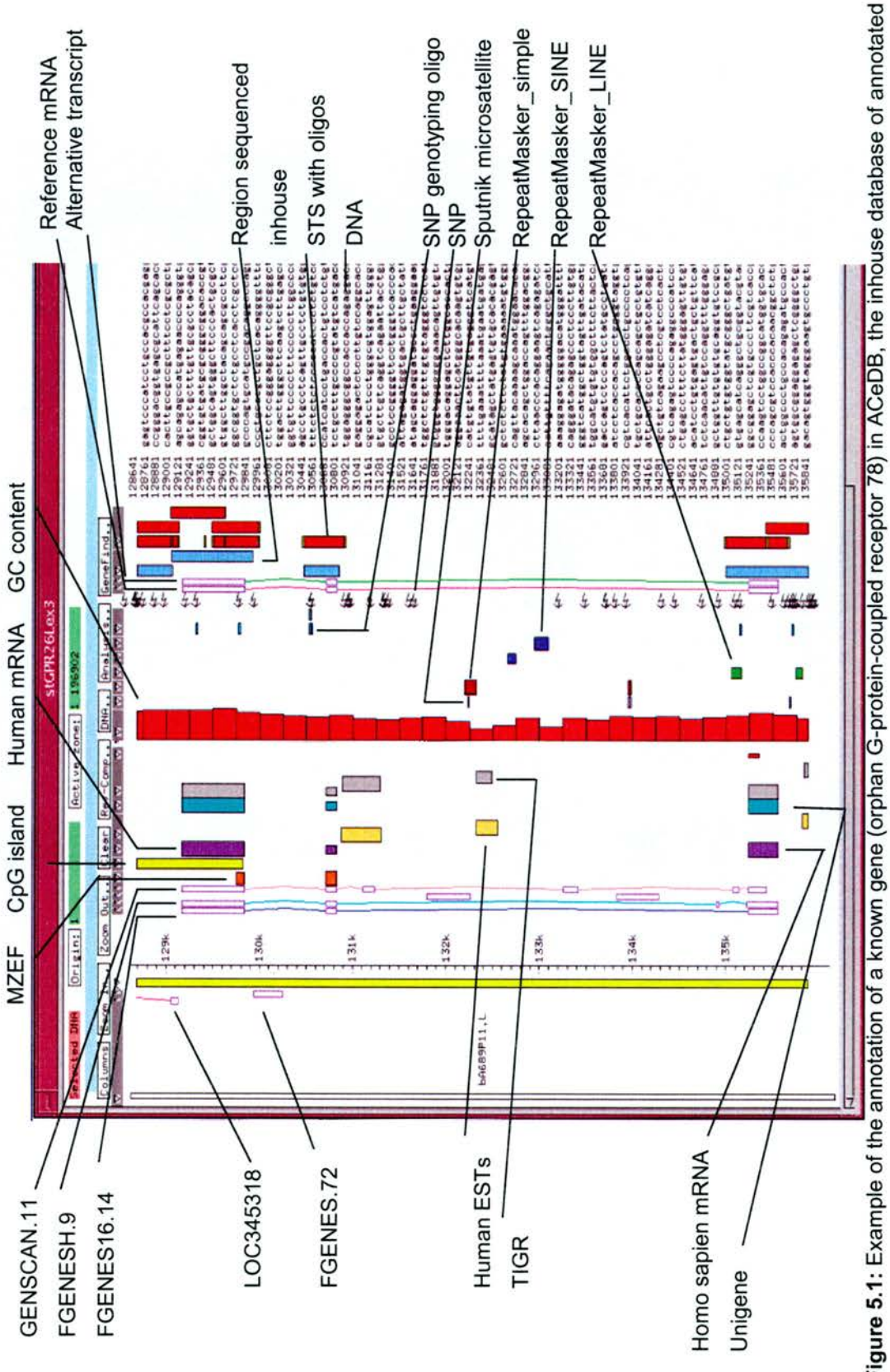


Figure 5.1: Example of the annotation of a known gene (orphan G-protein-coupled receptor 78) in ACeDB, the inhouse database of annotated chromosome 4p sequence.

5.2.2. Minimal Tiling Path

In order to position known genes and identify novel genes in the two minimal regions, I first needed to identify a clone contig. The group had previously described a BAC/PAC contig of MR1 (Evans *et al*, 2001_a). However, the clones that form the MR1 contig were not necessarily being sequenced in the HGP. Therefore, I used the inhouse contig, displayed in the SAM database (system for assembling markers) (Soderlund, 1995) in conjunction with the HGP sequence displayed in ACeDB. First, a minimal tiling path of BACs across MR1 was constructed using SAM. I determined which clones were part of both the SAM and the ACeDB contigs, referring back to SAM to bridge sequence gaps in ACeDB. At this time, many of the BACs displayed in ACeDB from the HGP were not fully sequenced and in many fragments, and contiguous sequence contigs were short. Therefore, the inhouse contig formed a useful resource to order clones with respect to one another. An *insilico* contig of MR2 consisting of clones being sequenced as part of the HGP, had been constructed previously in SAM by S. LeHellard. I chose a minimal tiling path from this contig to work from in ACeDB. Tables 5-2 and 5-3 detail the minimal tiling paths I constructed of MR1 and MR2.

Clone (RP11)	Sequence	Fragments (Dec 2001)
301J10	AC007104	16
	GAP	
751L19	AC068403	1
690D17	AC067775	9
448G15	AC005674	1
480F3	AC009528	21
494H11	AC006499	1
PACs and 1 non-sequenced BAC		
26P5	AC005599	1
61G19	AC084048	1
709L9	AC023150	1
270I3	AC027625	2
669F3	AC060784	1
512I20	AC073991	1
287J14	AC006230	1
281P23	AC025539	1
473M13	AC005699	1
74M11	AC024968	1
168E17	AC024335	9
1J7	AC015750	1
437G1	AC022313	1
4E12	AC019263	11
352E6	AC022769	1

Table 5-2: The minimal tiling path (December 2001) of Minimal Region One, the region of chromosome 4p identified by the overlapping disease associated haplotypes of families 22, 50 and 59. A contig built inhouse was used to select a tiling path of clones that were also being sequenced as part of the human genome project (HGP) and available on the University of California, Santa Cruz genome browser (genome.csi.ucsc.edu). Boxes identify overlapping clones. Clone names, sequence accession numbers and sequencing status were obtained from the National Centre for Biotechnology Information (www.ncbi.nlm.nih.gov). The gap between clones RP11-301J10 and RP11-751L19 represents a contig gap in the inhouse contig and the HGP contig. The gap between clones RP11-494H11 and RP11-26P5 represents a contig gap in the HGP contig only.

Clone (RP11)	Sequence	Fragments (Dec 2001)
17E2	AC026494	1
30A16	AC068465	MANY
362I16	AC021942	1
10G12	AC013609	1
800K23	AC023578	1
108G15	AC016782	13
412P11	AC027056	1
453O5	AC092440	1
406E22	AC036129	1
281M17	AC006052	1
654J13	AC093607	1
380P13	AC024670	1
1004L19	AC020746	1
688P1	AC068410	1
11D19	AC006229	1
735L15	AC041010	1
617A17	AC092846	1
336H2	AC005769	5
192P23	AC027517	40
106M4	AC006390	1
401G6	AC073829	2
310G15	AC007073	1
660M5	AC026257	3
751O1	AC044900	24
778B12	AC058821	46
470D11	AC093660	26
421A17	AC020706	24
302F12	AC092436	1
-	AC006393	28
239C17	AC006391	5
330J16	AC024360	1
131K9	AC022747	MANY

Table 5-3: The minimal tiling path (December 2001) of Minimal Region Two (MR2) the region of chromosome 4p identified by the overlapping disease associated haplotypes of families 22, 48 and 50. An *in silico* contig, built inhouse by S. Le Hellard using the clone contig created by the human genome project (HGP) available on the University of California, Santa Cruz genome browser (genome.csi.ucsc.edu), was used to select a minimal tiling path of clones being sequenced as part of the HGP. Boxes identify overlapping clones. Clone names, sequence accession numbers and sequencing status were obtained from the National Centre for Biotechnology Information (www.ncbi.nlm.nih.gov). A clone name for sequence AC006393 was not available.

5.3. Sequence Composition of the Region

Giemsa staining (G-banding) of chromosomes is a technique that enables individual chromosomes to be identified. The giemsa dye stains regions of the chromosome that are rich in adenine and thymine nucleotides, producing a dark band under a microscope. Due to these properties of the Giemsa stain, the banding patterns reflect the GC content (Bickmore and Sumner, 1989).

To determine the GC content of MR1 and MR2, I first had to localise our inhouse markers defining the minimal regions to the public map at UCSC. MR1 is defined by inhouse telomeric marker st301J10.p3 and inhouse centromeric marker st426F15.p1. I positioned the centromeric marker very near to public marker SHGC-50831 at position 8,613,965 in the UCSC July 2003 release, and the telomeric marker ~50kb from public marker D4S504 at position 12,450,000. MR2 is defined by inhouse telomeric marker st17E2.p1 and inhouse centromeric marker st131K9.p1. I positioned the telomeric marker ~70kb from the 3' end of gene GPR125 at position 22,090,000, and the centromeric marker ~70kb before public marker MFD281 at position 26,300,000. In this way I could locate the recombination breakpoint boundaries of MR1 and 2 and calculate the GC content of the regions.

The first ~2.1Mb of MR1 corresponds to a light band, and the last ~1.4Mb corresponds to a dark band. The first ~720kb of MR2 corresponds to a dark band and the last ~3.7Mb corresponds to a light band (Figure 5-2). The GC content of MR1 reduces as you move from the telomeric to the centromeric end (Figure 5-3: A). This corresponds to the move from a light to a dark band. The GC content of MR2 does not change as clearly, but does show an increasing trend from the telomeric to the centromeric end (Figure 5-3: B). Thus, the GC content corresponds to the banding pattern.

The GC rich light bands are thought to represent gene rich regions whilst the GC poor dark bands are thought to represent gene poor regions. This is bourne out by the

gene density and location of the genes in MR1 and MR2 (Section 5-4). MR1 is relatively gene poor compared to MR2, consistent with a larger proportion of it being in a dark band.

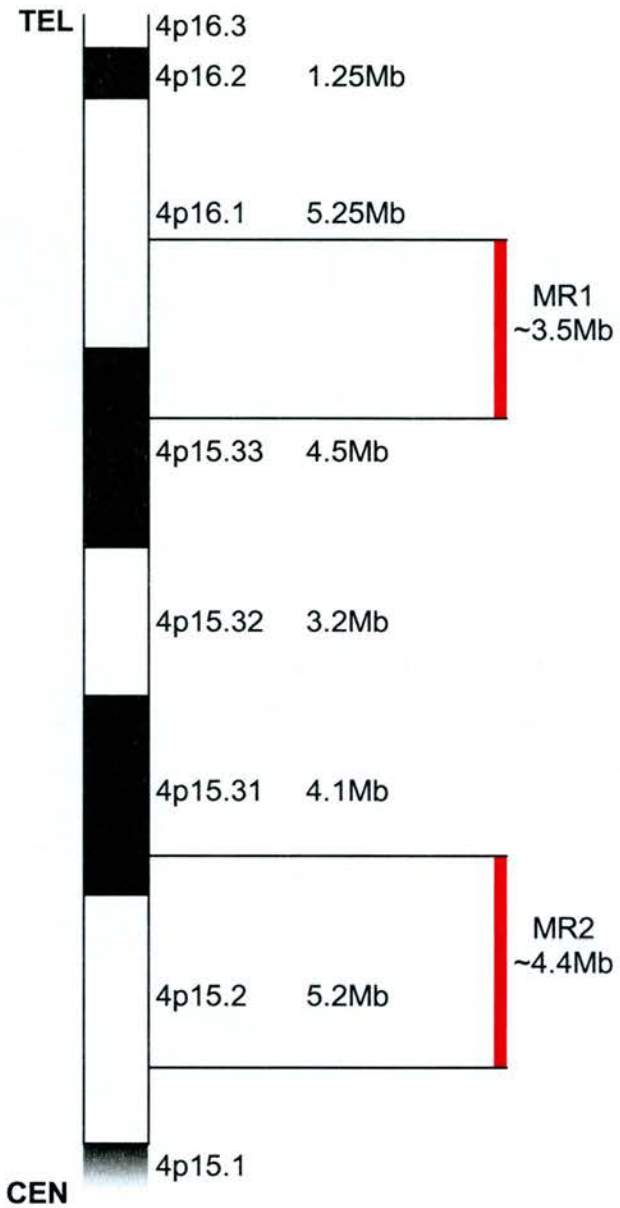


Figure 5-2: Schematic of the Giesma bands (Black and white) on chromosome 4p with respect to the position of both Minimal Region One (MR1), identified by the overlapping disease associated haplotypes of families 22, 50 and 59 and Minimal Region Two (MR2), identified by the overlapping disease associated haplotypes of families 22, 48 and 50 The size (Megabases) and name of each band is shown. Giesma stains regions rich in the nucleotides AT, and thus bands are dark (black) or light (white), representing GC poor and GC rich regions respectively. TEL = telomeric, CEN = centromeric.

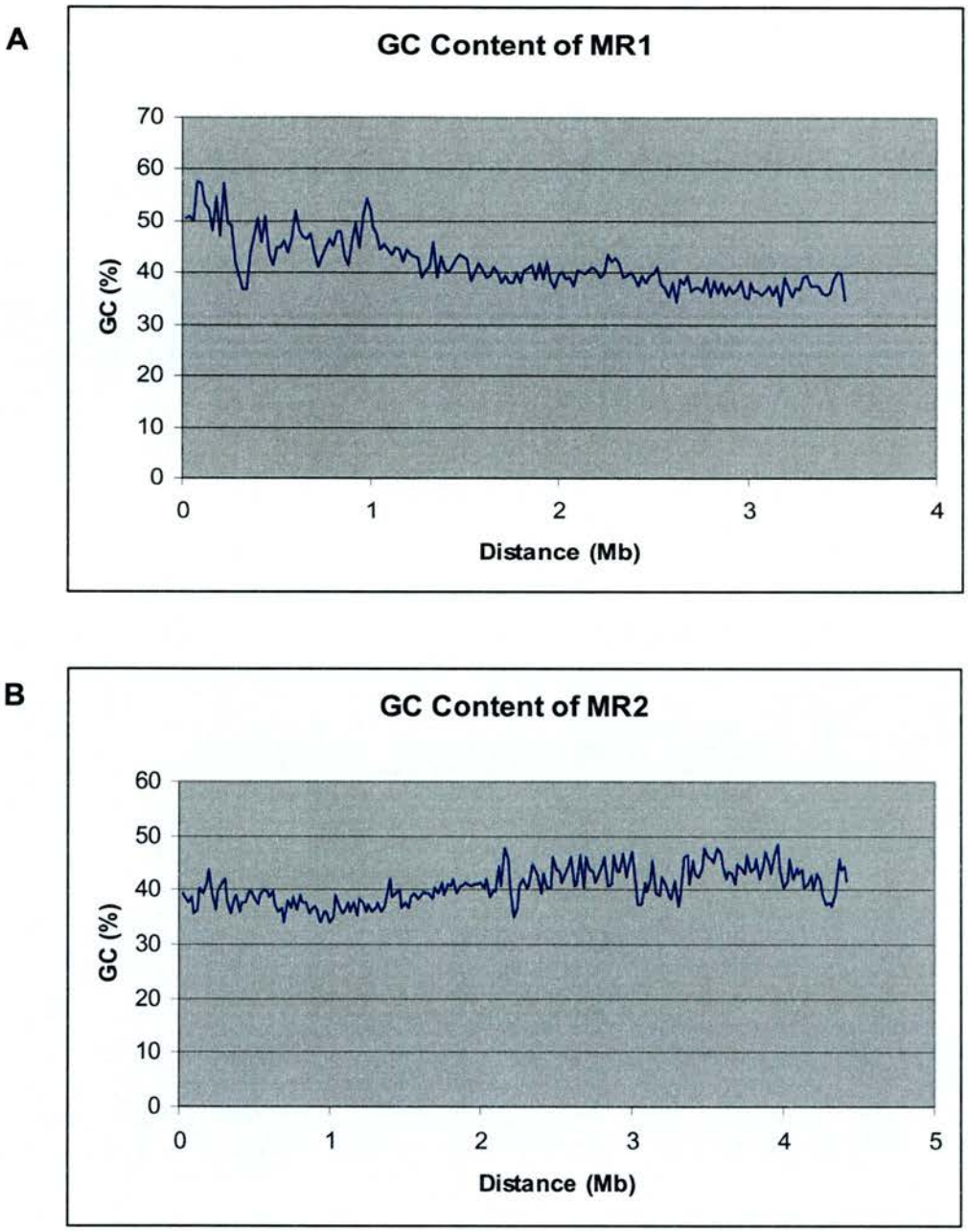


Figure 5-3: Graphical representation of the percentage of GC nucleotides in **A:** Minimal Region One (MR1), the region identified by the overlapping disease associated haplotypes of families 22, 50 and 59 and **B:** Minimal Region Two (MR2), the region identified by the overlapping disease associated haplotypes of families 22, 48 and 50, on chromosome 4p. Data is taken from the July 2003 release of the University of California, Santa Cruz genome browser (genome.csi.ucsc.edu). Each data point represents the percentage of GC nucleotides in a 20 kilobase window of genome sequence. Distance is measured in Megabases (Mb). The telomeric end of each minimal region corresponds to 0Mb.

5.4. Known Genes in MR1 and MR2

5.4.1. The Known Genes

Today (February 2004) there are seven known genes in MR1 (Table 5-4) and 13 known genes in MR2 (Table 5-5), known genes being defined as 100% alignment of the mRNA to the DNA sequence. The tables show the gene symbol and the full name with accession numbers of the mRNA and protein sequence, and the National Centre for Biotechnology Information (NCBI) (www.ncbi.nlm.nih.gov/) reference sequence (RefSeq) validation status. Predicted and provisional sequences constitute mRNA that has been through a round of BLASTn and for which there is a full-length coding sequence. All provisional sequences undergo a manual review step which makes corrections to the sequence, adds additional publications, adds a summary of gene function, provides additional records for splice variants, adds feature annotations to the nucleotide or protein sequence, and provides an indication of completeness.

The position of each known gene, in each minimal region, can be seen in Figures 5-4 and 5-5. Every gene except HS3ST1 in MR1 is located in the light band 4p16.1, and every gene in MR2, except GPR125 and GBA3, are located in the light band 4p15.2. This follows with the observations in the previous section that G-banding corresponds to both GC content and gene density.

Gene Symbol	Gene Name	mRNA	Protein	NCBI Status
Gpr78	G-protein coupled receptor 78	NM_080819	NP_543009	Validated
CPZ	Carboxypeptidase Z	NM_003652	NP_003643	Reviewed
DRD5	Dopamine receptor D5	NM_000798	NP_000789	Reviewed
SLC2A9	Solute carrier family 2 (facilitated glucose transporter) member 9	NM_020041	NP_064425	Provisional
WDR1	WD repeat domain 1	NM_005112	NP_005103	Reviewed
MIST*	Mast cell immunoreceptor signal transducer*	XM_093920	XP_093920	NCBI genome annotation
HS3ST1	Heparan sulphate (glucosamine) 3-O-sulfotransferase 1	NM_005114	NP_005105	Reviewed

Table 5-4: Known genes in Minimal Region One (MR1) (February 2004). MR1 is the region of chromosome 4p defined by the overlapping disease associated haplotypes of families 22, 50 and 59. The known genes are those annotated on the University of California, Santa Cruz genome browser (genome.csi.ucsc.edu). mRNA and protein accession numbers and the National Centre for Biotechnology Information (NCBI) status were obtained from the NCBI locuslink browser (www.ncbi.nln.nih.gov/locuslink). * interim gene symbol and name.

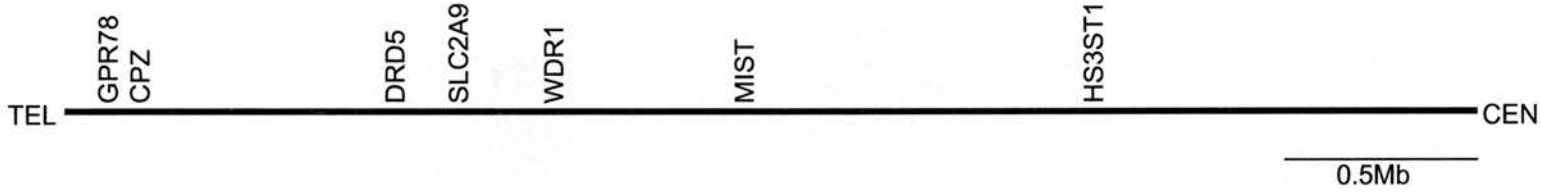


Figure 5-4: Position of the seven known genes in Minimal Region One (MR1), a region of chromosome 4p defined by the overlapping disease associated haplotypes of families 22, 50 and 59. Black bar: MR1. TEL: telomeric. CEN: centromeric. See Table 5-4 for full gene names. Mb: Megabase

Gene Symbol	Gene Name	RNA	Protein	NCBI Status
GPR125	G-protein coupled receptor 125	XM_291111	XP_291111	NCBI genome annotation
GBA3	Glucosidase, beta, acid 3	NM_020973	NP_066024	Provisional
PPARGC1A	Peroxisome proliferative activated receptor, gamma, coactivator 1	NM_013261	NP_037393	Reviewed
DDX15	DEAH (Asp-Glu-Ala-His) box polypeptide 15	NM_001358	NP_001349	Reviewed
SOD3	Superoxide dismutase 3, extracellular	NM_003102	NP_003093	Provisional
LGI2	Leucine-rich repeat LGI family, member 2	NM_018176	NP_060646	Provisional
SLA/LP	Soluble liver antigen/liver pancreas antigen	NM_016955	NP_058651	Provisional
PI4K2B	Phosphatidylinositol 4-kinase type-II beta	NM_018323	NP_060793	Provisional
ZCCHC4	Zinc finger, CCHC domain containing, 4	XM_376310	XP_376310	NCBI genome annotation
ANAPC4	Anaphase promoting complex subunit 4	NM_013367	NP_037499	Reviewed
SLC34A2	Solute carrier family 34 (sodium phosphate) member 2	NM_006424	NP_006415	Provisional
RBPSUH	Recombining binding protein suppressor of hairless (drosophila)	NM_005349	NP_005340	Provisional
CCKAR	Cholecystokinin A receptor	NM_000730	NP_000721	Provisional

Table 5-5: Known genes in Minimal Region Two (MR2) (February 2004). MR2 is the region of chromosome 4p defined by the overlapping disease associated haplotypes of families 22, 48 and 50. The known genes are those annotated on the University of California, Santa Cruz genome browser (genome.csi.ucsc.edu). mRNA and protein accession numbers and the National Centre for Biotechnology Information (NCBI) status were obtained from the NCBI locuslink browser (www.ncbi.nln.nih.gov/locuslink).

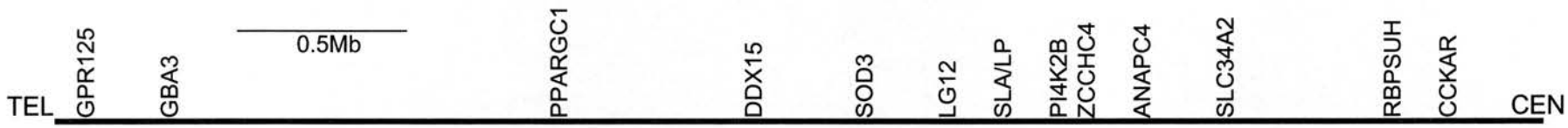


Figure 5-5: Position of the 13 known genes in Minimal Region Two (MR2), a region of chromosome 4p defined by the overlapping disease associated haplotypes of families 22, 48 and 50. Black bar: MR2. TEL: telomeric. CEN: centromeric. See Table 5-5 for full gene names. Mb: Megabase.

5.4.2. Evaluating the Known Genes

CpG islands can be a means of locating the start of a transcript. Antequera and Bird (1993) reported that 56% of genes are associated with CpG islands. Furthermore, it has been shown that all housekeeping genes are associated with CpG islands, compared with 40% of tissue restricted genes (Larsen *et al*, 1992). The completeness of a transcript can be assessed at the 3' end of a gene by the identification of a poly-adenylation signal. The two most common signals are AATAAA or ATTAAA, occurring within 10-30 nucleotides of the end of transcription. It has been estimated that 90% of genes possess the AATAAA poly-adenylation signal, and that the remaining 10% is accounted for by the ATTAAA poly-adenylation signal (Wahle and Keller, 1996). However, there are a small number of alternative poly-adenylation signals that account for some of the variation. Beaudoing *et al* (2000) identified ten such putative variant poly-adenylation signals by identifying sequences significantly overrepresented in the 3' end of mRNA.

Characterisation of the known genes for the presence of a CpG island and a poly-adenylation signal can be seen in Table 5-6. In accordance with what has been shown for other genomic regions, the majority of the genes have a CpG island on or near exon one. Although, in the SOD3 and SLC2A9 genes, they occur in exons 3 and 2 respectively. In SOD3 this may be because the protein coding sequence begins in exon 3. The most common poly-adenylation signal is AATAAA, observed in eleven genes (55%), and the second most common, ATTAAA, is observed in four genes (20%). This also follows the findings from other genomic regions. Five of the genes did not have either of these common poly-adenylation signals.

Gene	MR	CpG island	AATAAA	ATTAAA
GPR78	1	Y	N	N
CpZ	1	N	N	Y
DRD5	1	Y	Y	N
SLC2A9	1	Y (ex2)	N	N
WDR1	1	Y	N	Y
MIST	1	N	N	N
HS3ST1	1	Y	Y	N
GPR125	2	Y	Y	N
GBA3	2	N	Y	N
PPARGC1	2	N	Y	N
DDX15	2	Y	Y	N
SOD3	2	Y (ex3)	N	Y
LG12	2	Y	Y	N
SLA/LP	2	Y	N	N
PI4K2B	2	Y	Y	N
ZCCHC4	2	N	N	Y
ANAPC4	2	Y	Y	N
SLC34A2	2	N	Y	N
RBPSUH	2	Y	Y	N
CCKAR	2	N	N	N

Table 5-6: Characterisation of the known genes in Minimal Region One (MR1) and Minimal Region Two (MR2). Genes are characterised by the presence or absence of CpG islands and poly-adenylation signals (Y: yes, N: no). See Tables 5-4 and 5-5 for full gene names and accession numbers.

5.5. Identifying Novel Genes

5.5.1. cDNA Library Screening

5.5.1.1. Primer Design

I designed primers to putative exonic sequence in order to identify novel genes by cDNA library screening. Using the contigs of the two minimal regions (Tables 5-2 and 5-3), I used ACeDB to identify regions with a build up of transcript evidence. The best evidence for a novel transcript is from multiple splicing ESTs from different libraries, and/or homology to proteins (identified from the protein database SwissProt). However, this was fairly rare in the two regions. As mentioned previously, gene prediction programmes have a certain degree of inaccuracy and therefore I did not consider exon prediction programmes unless they were supported by expression evidence (ESTs, mRNA or protein homology), and never where they were located in a repeat. SwissProt protein matches are good evidence of a gene protein motif, however, to avoid pseudogenes it is imperative to check for a continuous ORF. ESTs are also good evidence for a transcript, however in some instances an EST can be the result of genomic contamination in the cDNA library from which it was generated. Since ESTs are generated from cDNA libraries using dT primers that target the poly-A tail on a mature mRNA, they can also be primed from a contaminating genomic poly-A tract. Therefore, I did not consider single ESTs or ESTs that aligned to a genomic poly-A. I designed 21 STSs in MR1 and 13 STSs in MR2.

Primer pairs were tested to establish optimal PCR conditions on three CEPH control DNAs before cDNA library screening. It is worth noting that at this time (January-March 2002), STS design was made difficult by the unfinished nature of the BACs which made it impossible to see a whole putative gene structure since the order of clone fragments was unknown.

5.5.1.2. cDNA Libraries

My strategy was to screen the STSs on 22 cDNA libraries from various adult and foetal tissues (Table 2-2). Previous evidence from the Sanger Institute (Dr G Howell, *pers comm*) suggested that 80-90% of STSs are expressed in nine of these libraries. Therefore, STSs were first screened on a primary plate of these nine libraries. A secondary plate of nine libraries and a tertiary plate of four libraries were subsequently screened if no positives were identified from the primary plate. Each cDNA library was represented on the plate as five superpools of approximately 100,000 clones each.

5.5.1.3. STS Screening

An example of an agarose gel showing the results of an STS screened on the cDNA library primary plate can be seen in Figure 5-6. Twenty-nine of the 34 STSs were positive for multiple cDNA library superpools. Twenty-five of the 29 STSs were positive on a suitable number of library superpools from the primary plate and the remaining four STSs required screening on the secondary plate. Of the five STSs not positive on the primary plate, two STSs were also screened on the tertiary plate. Neither was positive in any library and therefore are either not part of genes or are not represented in these cDNA libraries. Three STSs were positive for every single superpool and the 'no template' control on the primary plate, although at a lower intensity. Therefore either the reactions were contaminated or, in addition to contamination, the STSs were not specific to chromosome 4. They were not pursued further because the sequence fragment to which they were designed in ACeDB had subsequently been removed from this region of the genome. Table 5-7 details the results for each STS.

5.5.1.4. Vectorette Screening

Three positive superpools, preferably from different libraries in order to maximise access to independent clones, were chosen for each of the 29 STSs. For each STS, six

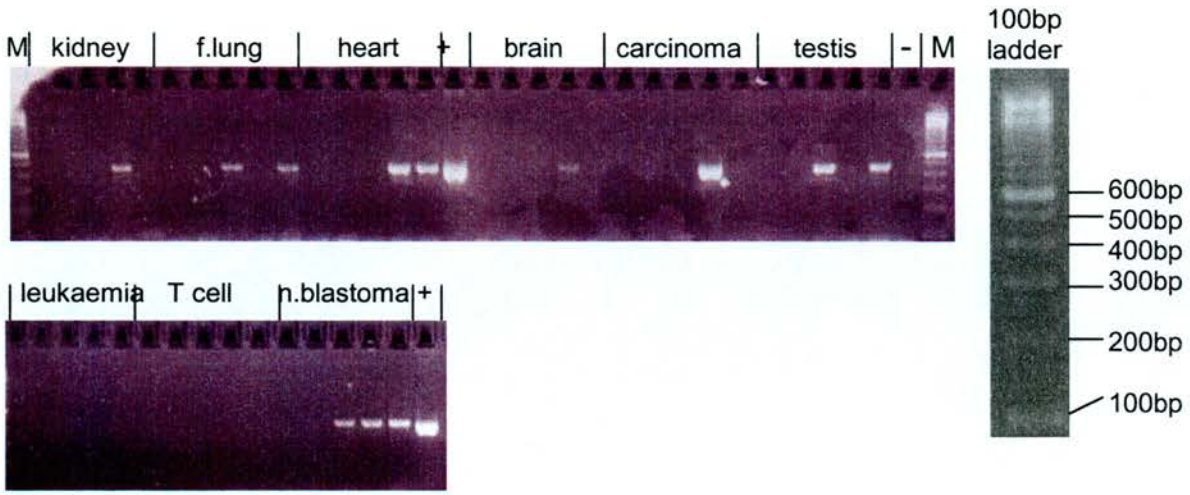


Figure 5-6: Example of an STS screened on the cDNA library primary plate. The primary plate contains five superpools of each of nine cDNA libraries (F.lung: fetal lung; n.blastoma: neuroblastoma), two genomic DNA controls (+), and one 'no template' control (-). Each cDNA library (~500,000 clones) has been divided into five superpools of ~100,000 clones each. M = Ready Load™ 100 bp DNA ladder (Invitrogen). In this example, the STS is positive for seven libraries, but not every superpool in each library.

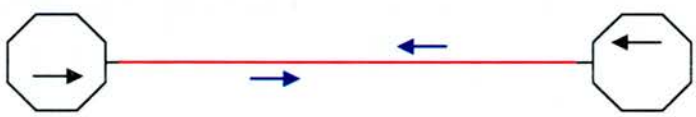


Figure 5-7: Vectorette schematic. Each cDNA library has a vectorette (hexagon) ligated onto the end of the cDNA insert which acts as a specificity regulator. A vectorette is two strands of identical, non-complimentary sequence which cannot anneal to one another. In addition, the vectorette primer (224) sequence is identical to and cannot anneal to the vectorette sequence. When the specific primer (blue arrow) anneals to the cDNA (red line) and amplifies into the vectorette (hexagon) it creates the necessary reverse complimented sequence for the 224 primer (black arrow) to anneal and amplify back.

reactions were performed: each primer was paired with the vectorette primer (224) for each of the three superpools. Only one vectorette primer is required for both ends as the vectorette ligated onto the end of each insert is a bubble of non-basepairing sequence. The 224 primer cannot anneal to the vectorette sequence until the specific primer has annealed to the cDNA insert and extended into the vectorette (Figure 5-7). This acts as a specificity enhancer.

PCRs were run out on agarose gels. Large, well separated bands were cut out for sequencing. In addition, a 1:100 dilution was made of the vectorette PCRs for all 29 STSs. Twenty-five out of 29 STSs had bands cut out in this way from the vectorette PCR. There was not necessarily a suitable band to cut out for every superpool, but this was not due to a lack of amplification, rather that the bands were not well separated from one another. A further round of PCR was performed on the bands before cleaning and sequencing. A successful second round PCR was performed for 20 out of 25 STSs. Therefore, a total of 40 clean bands from 20 STSs were sequenced. Table 5-7 details the results.

Table 5-7 (continued overleaf): cDNA library screening results. cDNA library screening consisted of an initial screen with the specific primers to identify the positive superpools and then a vectorette PCR with each of the specific primers and the vector primer to amplify the full length cDNA. Twenty-two cDNA libraries were arranged on three plates (primary, secondary and tertiary), and each library is represented by five superpools (a-e). The table shows the positive superpools identified from the primary and secondary plates and the libraries used for the vectorette PCR. The results of the vectorette PCR were run on agarose gels and bands were cut out for re-amplification and sequencing. The table shows which libraries had bands cut out and the bands that amplified for sequencing. AK (adult kidney), AH (adult heart), HeLa (cervical carcinoma), T (testis), HBP (T cell), SK (neuroblastoma), FLU (foetal lung), U (leukaemia), AB (adult brain), HSI (small intestine), HL (peripheral blood), ALU (adult lung), BM (bone marrow), FL (foetal liver). Whether the F (forward) or R (reverse) primer of the STS produced the band is noted.

STS	Primary Plate Positives	Secondary Plate	Vectorette PCR	Bands	Second Round PCR
St448G15pt2	AK a b d e, AH a b c d e, HeLa a e, T c e, HPB b, SK c d e.		AK b, HeLa e, Tc.	1 from AK b F, two T c F.	T c F.
St448G15pt3	AK b d, AH a b c d e, T c, SK c.	HSI a b c d e, HL c, ALU e, BM d.	HL c, ALU e, BM d.	1 from HL c F, ALU e F and BM d F.	BM d F
St448G15pt4	AH b d e, T c e, SK c e.		AH b, T c, SK c.	1 from T c R, 1 from SK c R.	All
St26P5pt12	AH e, T e, SK e.	FL e.	FL e, T e, SK e.	1 from FL e R, SK e F and SK e R.	FL e R
St26P5pt13	AH d e, T c e, SK c d e.		AH b, T c, SK c.	1 band from T c R, SK c F and SK c R.	SK c F
St287J14pt15	None	None			
St287J14pt16	None	None			
St287J14pt20	AK c, AH d e, AB b, T e, SK c d e.		AH e, AB b, SK e.	1 from AH e R, 1 from AB b F, 1 from SK e F.	None
St287J14pt21	T c e, SK e, but very faint. T e.	FL c d.	T e, FL c, FL d.	1 from T e F.	None
St287J14pt22	AK a b c, FLU b d, AH b c e, AB a b, T b d e, U d, HPB b c, SK e.		AB b, T b, HPB c.	2 from T b R. 1 from AB b F, T b R and HPB c R.	All
St473M13pt23	AH d e, T c e, SK c d e.		AH d, T c, SK d.	1 from AH d F and T c F.	T c F
St473M13pt26	AK c, FLU b e, AH a c e, AB d e,		AK c, HeLa d, U b.	None	

	HeLa a d e, U a b, HPB a, SK a e.				
St473M13pt27	AH d e, T c e, SK d e.		AH b, T c, SK d.	2 from T c F, 1 from T c R, 1 form SK d F.	All
St437G1pt28	AK a c e, FLU a d, AH c e, AB a b c d, HeLa e, T b c e, U d e, HPB a d e, SK e.		FLU d, HeLa e, SK e.	1 from FLU d R, HeLa e R, SK e F and R.	FLU d R and HeLa e R
St352E6pt30	AH d e, T c e, SK c d e.		AH d, T c, SK c.	1 from AH d F, 1 from AH d R, 1 from SK c F.	All
St301J10pt31	AK a b c d e, FLU a d e, AH c e, AB a b c d, HeLa a b c d e, T d, U c, HPB c d, SK e.		T d, U c, HPB d.	1 from T d F, 1 from T d R, 1 from U c R, 1 from HPB d F, 1 from HPB d R.	T d R, U c R, HPB d F, HPB d R.
St751L19pt32	HeLe a d e, SK c d e, AH d e.		HeLa a, AH d, SK c.	1 from AH d R and SK c R.	None
St74M11pt33	HeLa b, T b e, U a b c, HPB d, SK e.		T b, HPB d, SK e.	1 from T b F and R.	All
St74M11pt34	T b d.	Fl e	T b, T d, FL e.	1 from T b R, T d F and T d R.	T d F and T d R.
St494H11pt35	AH a c e, HeLa c, T e, HPB c, SK c e.		AH a, HeLa c, T e.	None	
St494H11pt36	Contamination/non-specific.				
St17E2pt37	AH a c, SK b.		AH a, AH c, SK b.	1 from AH a R and SK b R.	None
St17E2pt38	AK d, FLU c e, AH d e, AB d, HeLa d, T c e, SK c d e.		HeLa d, AK d, T c.	1 from HeLa d F and T c R.	None
St362I16pt39	AH b d, HeLa b e, AB b, T c d e, U e, HPB a b c d e, SK a b d e.		HeLa e, AH b, T c.	1 from T c R.	All
St106M4pt40	AK d, AH b e, AB a b c e.		AK d, AH b, AB e.	None	
St106M4pt41b	AK a b c d e, FLU b c, AH a b c d e, AB a b c d e, HeLa d, T b d, U a b c d, HPB c d e, SK b c e.		T b, AK e, AB e.	1 from AK e R, AB e F and R.	All
St401G6pt42	AK a c e, AH b e, AB c		AK c, AH b, AB c. 1 from AK c F and AH b F and R.	None	
St310G15pt43	AK a b c e, FLU b c d e, AH a c e, AB a b c d e, HeLa a c d, T b c e, U b e, HPB a c d e, SK b d e.		U e, AK e, AH c.	None	
St470D11pt44	AH d e, AB b e, T c, SK a c e.		SK c, AH e, AB b.	1 from SK c F.	All

St302F12pt45	AK c e, FLU d e, AH c e, AB d, HeLa c d, T e, SK a e.		AB d, AK c, AH c.	1 from AB d R and AH c F.	All
St1004L1pt46	AH d e, T c d, SK c d e		AH d, T d, SK c.	1 from AH d F and T d F.	All
St401G6pt47	AK a c e, AH b.		AH b, AK e, AK c.	1 from AH b F and R, AK e F and AK c F.	All
St470D11pt48	Contamination/non-specific.				
St470D11pt49	Contamination/non-specific.				

5.5.1.5. Vectorette PCR Problems

The product produced after the vectorette PCR had many more bands than was expected (Figure 5-8: A). It was possible that this was due to a proportion of the bands being amplified from genomic regions other than the targeted region. These could be produced by either the 224 primer, or the specific primer, annealing randomly to the cDNA. Previous experiments by others at the Sanger Institute have found that product is produced if a superpool is amplified with 224 only. Figure 5-8 B, shows an example where the forward primer produces clean bands but the reverse primer produces many bands. This discrepancy could be due to the non-specificity of the reverse primer.

However, it is possible that the results could be due to experimenter error. I performed a number of experiments to address this. Firstly, the vectorette PCR was performed on four STSs (st473M13pt26, st751L19pt32, st401G6pt42 and st301J10pt31) by Jackie Bye, by whom this work was routine. The results showed that, whilst there was some variation between reactions, in general the level of specificity was similar between the two experiments (Figure 5-9: A). Secondly, I performed the vectorette PCR using an STS (SG158641) known from previous experiments by others at the Sanger Institute to give clean results with two particular superpools. The results showed that the STS gave clean results as expected (Figure 5-9: B).

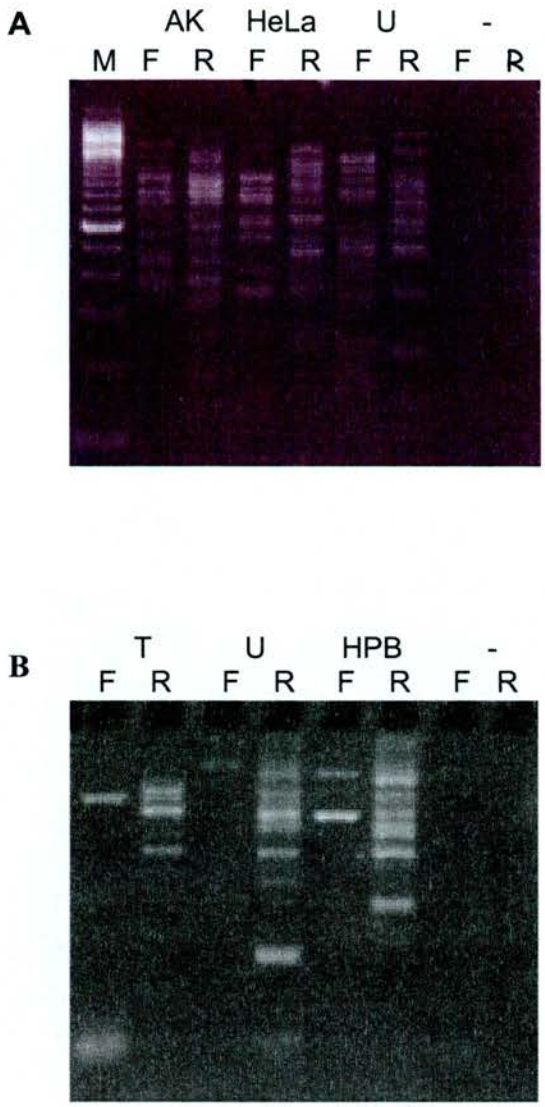


Figure 5-8: Two examples of non-specific amplification from the cDNA libraries. The vectorette PCRs, PCRs on a cDNA library with one specific primer and the vector primer in order to obtain the full length cDNA, produced more bands than was expected. Each primer is amplified on three cDNA superpools and a 'no template' control (-). M = Ready Load™ 100 bp DNA ladder (Invitrogen). **A**. The forward (F) and reverse (R) primers, amplified from adult kidney (AK), cervical carcinoma (HeLa) and leukaemia (U) cDNA libraries, show multiple banding. **B.** The forward (F) primer amplified from testis (T), leukaemia (U) and T cell (HPB) cDNA libraries shows less banding than the reverse (R) primer.

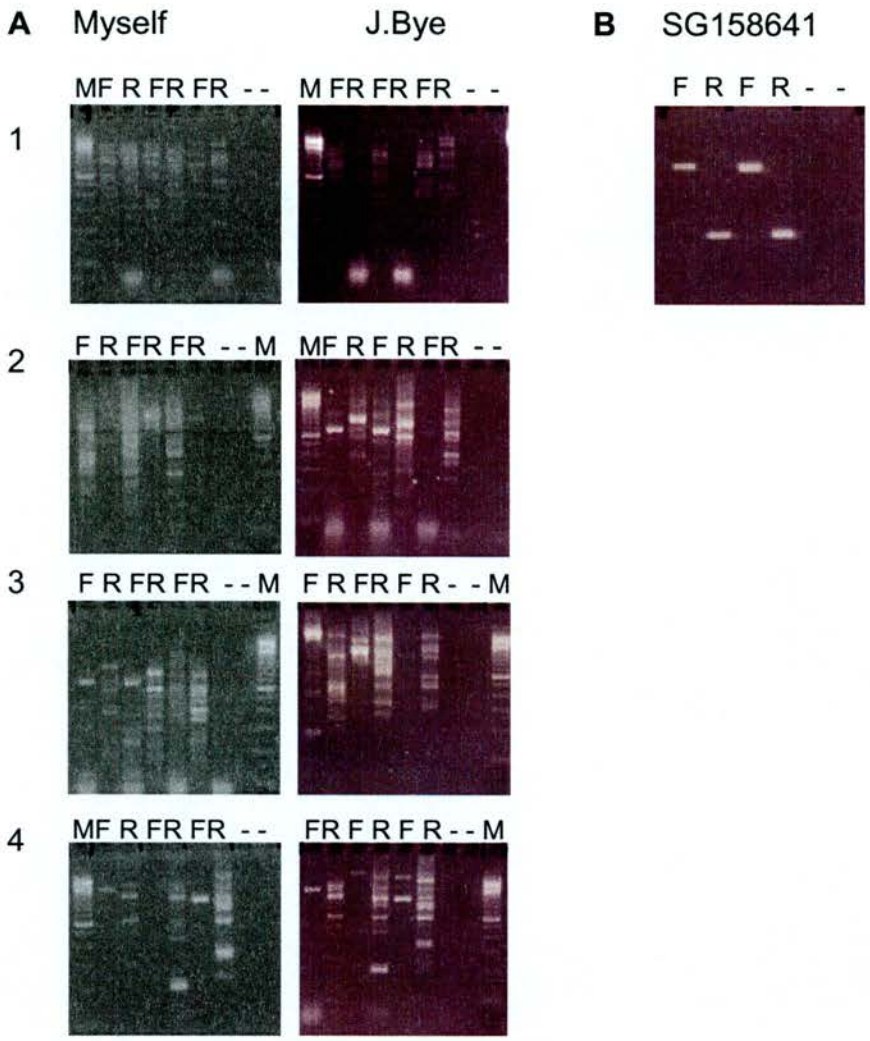


Figure 5-9: Results of the experiments to test the effect of experimenter error. The vectorette PCR is performed with the vector primer (224) and each of the forward (F) and reverse primer (R) of the specific STS in order to obtain a full length cDNA. A 'no template' control (-) is included. M = Ready Load™ 100 bp DNA ladder (Invitrogen). **A:** The vectorette PCR was performed by myself and a second independent person (J. Bye) **A1:** STS st473M13pt26 **A2:** STS st751L19pt32 **A3:** STS st401G6pt42 **A4:** STS st301J10pt31. **B:** The vectorette PCR was performed by me on two superpool cDNA libraries using STS SG158641.

In order to address the lack of specificity, nested primers were designed to four STSs: st352E6pt30, st301J10pt31, st74M11pt34 and st106M4pt41b. A PCR was performed using the nested primer and 224 on a 1:100 dilution of the bands cut out from the vectorette PCR and a 1:100 dilution of the vectorette PCR itself.

Of the ten bands cut out for these four STSs, only one amplified after PCR with the nested primer. However, this failure may have been because the template was too dilute rather than because the band was not from the specific locus. Using the vectorette PCR as a template for the nested PCR revealed that the nested PCR was significantly cleaner than the vectorette PCR. This suggested that some of the product in the original vectorette PCR was non-specific. Figure 5-10 shows an example of the vectorette PCR and the corresponding nested PCR. It can be seen that the nested PCR is cleaner than the vectorette PCR. In addition, the band cut out from the vectorette PCR did not amplify with a nested primer, suggesting it was not specific to chromosome 4. Two bands amplified from the nested PCR were sequenced and found to be from the correct locus on chromosome 4.

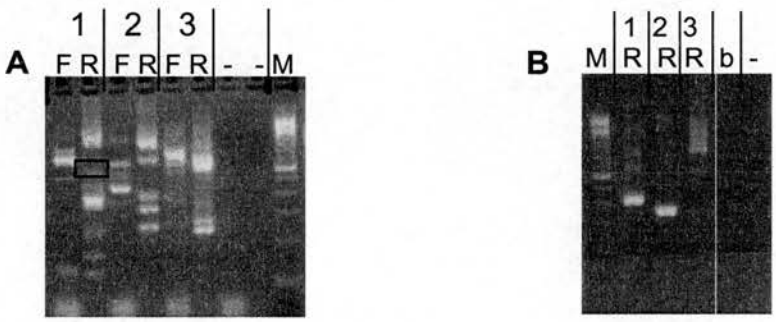


Figure 5-10: Examples of the results obtained from a nested experiment. **A:** PCRs were performed with the forward (F) and reverse (R) primer of st74M11pt34, each with the vector primer (224) on three cDNA libraries (1-3) and a no-template control (-). **B:** A nested primer was designed to the reverse primer. PCR was performed using this nested primer (R) and 224 using a 1:100 dilution of the vectorette PCR as a template. In addition, one band (b) cut out from cDNA library 1 (boxed in A) was used as a template for the nested primer and 224. M = Ready LoadTM 100 bp DNA ladder (Invitrogen).

To further address the problem of specificity, I sequenced eight bands cut out from the vectorette PCR. The sequencing reaction failed for the four STSs. Two of the sequenced products were multiple templates and two successfully sequenced. A BLASTn similarity search of the two successfully sequenced bands revealed that they did not originate from the appropriate locus. The failure of the other sequencing reactions could also have been the result of non-specificity.

Finally, I performed a PCR with only the 224 primer (data not shown) of the bands cut out from the vectorette PCR. The results showed amplification from most bands. However, eleven bands were different from the expected size. I sequenced six of these (the remaining five were too small to be more than vector sequence). Three of the bands gave product non-specific to chromosome 4, two failed to sequence and one was specific to chromosome 4. In conclusion, it was clear that there were non-specific bands in the vectorette PCRs and that the specific bands might be targeted by nested PCR.

5.5.1.6. Nested PCR Results

I designed nested primers to each of the STSs for which I had already performed a vectorette PCR (from Table 5-7). I then used a 1:100 dilution of the vectorette PCR as a template to perform a nested PCR, in which primers were designed to nest the chromosome 4 specific primer. Nested PCR conditions were identical to the original vectorette PCR (see Section 2.5.2.3). I then compared this to a PCR with the 224 primer only and a PCR with the specific primer only, in order to distinguish the bands that were specific to the nested PCR (Figure 5-11). It is possible to see in Figure 5-11, that a band in cDNA libraries 1 and 3 from the nested PCR (N) is not observed in either the specific (S) or the 224 (V) PCR. These bands are likely to be specific to the chromosome 4 locus. In contrast, the bands observed in cDNA library 2 are identical in each PCR and are therefore likely to be non-specific product.

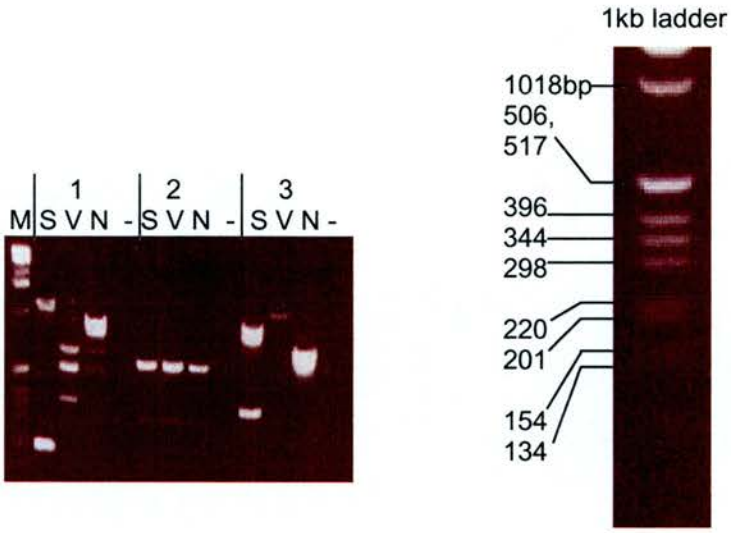


Figure 5-11: A vectorette PCR was performed on three cDNA libraries (1-3) using the vector (224) and the specific primer in order to obtain a full length cDNA. A 1:100 dilution of this vectorette PCR was then used to perform three PCRs with different combinations of primers: the nested primer and the vectorette primer (N), the specific primer only (S) and the vectorette primer only (V). A no-template control (-) was included in each PCR. M = Ready Load™ 1 kb DNA ladder (Invitrogen).

STS	Specific bands
St448G15pt2n	F (1) R (2)
St301J10pt31n	R (2)
St751L19pt32n	F (1)
St74M11pt33n	F (2)
St74M11pt34n	F (2) R (1)
St17E2pt38n	R (1)
St362I16pt39n	F (1) R (1)
St106M4pt40n	F (2) R (2)
St106M4pt41bn	F (3)
St1004L1pt46n	F (2)
St401G6pt47n	F (1)

Table 5-8: Nested vectorette PCR results. A vectorette PCR was performed on cDNA libraries using the vector primer and a chromosome 4 specific primer in order to obtain full length cDNA product. The vectorette PCR was then used as a template for a PCR using the vector primer and a primer designed to nest the specific chromosome 4 primer. The results of this were compared to PCRs using the same template and only the vector primer or only the specific primer. Bands obtained that were present only in the PCR with both the nested and vector primer were sequenced. The table shows the specific primer, either the forward (F) or the reverse (R), that had a band sequenced that was specific to chromosome 4 (number of bands successfully sequenced in brackets).

I obtained and sequenced 91 bands in this way (see Table 5-8 for a summary of the results). The sequencing reaction failed repeatedly for 34 of these. Of the remaining 57, a BLASTn similarity search of the sequences revealed that 30 were not from the intended locus. Twenty-four bands were amplified from the correct locus and gave genic product. Three products were the same size or smaller than the vector sequence and therefore would not contain cDNA sequence. Therefore, despite trying to target specific bands in the vectorette PCR by the above methods, only 26% of bands were actually specific. In addition, only 63% of these bands sequenced successfully. However, sequencing failure could also be due to non-specificity of the bands to chromosome 4 since every sequencing reaction was performed with a pGEM template control. These results are discussed on a gene by gene basis in Section 5.6.

5.5.2. RT-PCR

5.5.2.1. Primer Design

In addition to the cDNA library screening described in the previous section, I also performed RT-PCR in order to identify novel transcripts. Furthermore, an additional source of bioinformatic evidence was available subsequent to cDNA library screening. The NCBI LocusLink database (www.ncbi.nlm.nih.gov/ LocusLink/) provides an integrated site for gene descriptions. Every annotated known gene and hypothetical protein is provided with a unique locus ID (LOC) number. In addition, as part of the NCBI's genome annotation project, locus link also generates its own hypothetical gene predictions based on the results of gene prediction programmes and EST and/or mRNA alignment. Since these hypothetical genes are unique to the NCBI database, their only identifier is the LOC number. S. Morris linked this resource into ACeDB so that all LOCs were overlaid onto the sequence. However, the LOC genes do require individual inspection because many are apparently only based on evidence from gene prediction programmes. For example, there are a number that are contained entirely within repeat sequence (Figure 5-12).

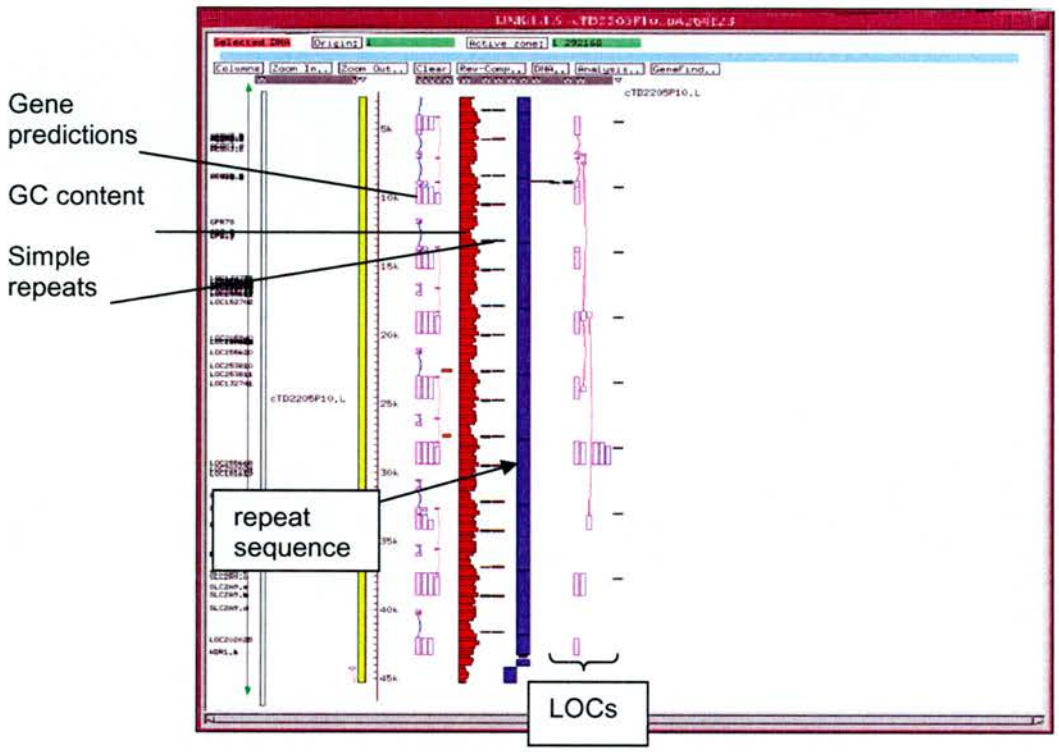


Figure 5-12: A screen shot from ACeDB, the inhouse database of annotated chromosome 4p sequence, showing several hypothetical (LOC) genes from the National Centre for Biotechnology Information (NCBI) LocusLink sequence database. It can be seen that these LOC genes are located entirely within repeat sequence (blue bar).

I classified LOC genes into four groups:

1. *No evidence*: exon prediction programmes either alone and/or in a simple repeat.
2. *Pseudogene:* Swissprot match across intron and/or the match to Swissprot protein is disrupted for at least one exon. No other convincing evidence to other exons.
3. *Putative gene:*

3a. EST match to only one exon or a few ESTs with no splicing, plus evidence a, b or c from category four, to one exon.

3b. Evidence d, e or f from category 4 to one or more exons.

3c. Multiple gene predictions to multiple exons plus one of a-f from category 4 to one exon.

4. *Gene:*

a, b or c of the following plus at least one other:

a. Splicing human/mouse ESTs
b. Multiple ESTs non-splicing
c. Swissprot match with ORF
d. CpG island
e. Mouse/rat homology
f. Multiple exon predictions to more than one exon.

There were 63 LOCs in MR1 and MR2. Fifty-nine novel STSs were designed (Appendix I) to 20 LOC genes for RT-PCR (Table 5-9). Some of the LOCs were part of a known gene already identified from cDNA library screening. All category 4 LOCs were studied, but only a selection of LOCs were chosen from category 2 and 3 in order to test whether LOCs in these categories were genes or not. None of the LOCs in category 1 were studied since it was almost certain that these were not genes. STSs that were located in the repetitive region of MR1 (described in Chapter 3) were designed to lie completely within an exon and then these were tested for specificity to chromosome 4 by a PCR on a MCHP.

RT-PCR was performed on human lymphoblastoid cDNA. Lymphoblastoid cell lines are known to undergo altered gene expression where some gene expression is switched off. Therefore, if an RT-PCR was negative on an agarose gel, a second RT-PCR was then performed using human universal cDNA (Clontech). This comprises cDNA from a selection of human tissues and is more likely to represent a wider number of genes. If this RT-PCR was negative, a small amount of this was used as a template for a further round of PCR since the universal cDNA was used at a relatively dilute concentration.

5.5.2.2. RT-PCR Results

In summary, no evidence from RT-PCR was obtained for eight of the putative genes studied in Table 5-10. RT-PCR produced bands that were not specific to chromosome 4 for two gene structures, and bands that were not supported by sequence data for five gene structures. Specific bands with supporting sequence evidence were obtained for seven structures. These results, and the results from the cDNA library screening, are discussed in more detail in Section 5.6.

Classification	Number in MR1&2	Number Studied
1	29	0
2	10	6
3a	6	2
3b	4	2
3c	5	1
4	9	9

Table 5-9: The classification of LOC genes in Minimal Region One (MR1) and Minimal Region Two (MR2). LOC genes were classified from 1 to 4, according to the amount and type of supporting evidence from other sources to indicate that the LOC was or was not likely to be a real gene (e.g. ESTs, homology to known protein domains, cross species homology, gene prediction programmes), I being a very poor level of supporting evidence and 4 being good level of supporting evidence. I studied a proportion of these for expression evidence in cDNA.

Locus (LOC)	Position (RP11-)	Minimal Region	Classification	All or part of a putative (P) or known (K) gene	Result
285479	264E23	1	4	P	NB
202024	180A12	1	4	P	B
255668	180A12	1	3a	P	NB
285544	180A12	1	4	P	B
202025	448G15	1	4	K*	S
133258	494H11	1	2	P	B
133259	494H11	1	2	P	NB
133260	494H11	1	2	P	S
202015	136I13	1	4	P	B
257414	136I13	1	2	P	B
166522	26P5	1	3b	K	S
166450	473M13	1	3b	P	NS
74M11	74M11	1	-	K*	S
166647	362I16	2	4	K*	S
152715	362I16	2	4	K*	NB
132895	380P13	2	3a	P	S
133214	324H7	2	2	P	NB
133218	750A13	2	4	P	NS
S27	10G12	2	-	P	NB
206041	617A17	2	3c	P	NB
91050	106M4	2	4	K*	S
133225	106M4	2	2	P	NB

Table 5-10: The results of RT-PCR performed on a selection of LOC genes in Minimal Regions One (MR1) and Two (MR2). Table notes the LOC name, the clone it is located on and the minimal region it is in. LOC were classified from 1-4 based on the supporting evidence for it (e.g. ESTs, protein homology, cross species homology), 1 being a very poor level of supporting evidence and 4 being a good level of supporting evidence. The classification of each LOC studied is noted, and whether it forms a completely novel gene (P), part of an existing gene structure identified from cDNA library screening (K*) carried out in the previous sections, or part of a previously known gene (K) annotated in the genome browser at the University of California, Santa Cruz (genome.csi.ucsc.edu). RT-PCR was also performed on two structures that were not LOCs (74M11 and S27). The RT-PCRs were run out on agarose gel and bands were cut out for sequencing. RT_PCR results: NB = no band obtained. B = band obtained but not successfully sequenced. NS = band obtained, successfully sequenced, but sequence comes from a genomic region other than chromosome 4p. S = band obtained, successfully sequenced and sequence is specific to chromosome 4p locus.

5.6. Results from cDNA Library Screening and RT-PCR

5.6.1. Putative Genes Without Evidence

As mentioned above, there were four LOCs for which I obtained bands by RT-PCR on universal cDNA, but for which the sequencing failed (Table 5-11). I concluded that the bands were not specific to the correct locus on chromosome four, but rather the result of spurious amplification of non-specific product after one or two rounds of RT-PCR. Performing a second round of RT-PCR using the first as a template, as was the case for three of the four LOCs, would encourage such non-specific amplification.

LOC	RT-PCR Round	RT-PCR	Sequencing	Conclusion
133258	1	Smear	-	Non-specific
202024	2	Clean band	> 1 sequence	Non-specific
202015	2	Nesting eliminated band	-	Non-specific
285544	2	Second round RT-PCR failed	-	Non-specific

Table 5-11: Troubleshooting four RT-PCR results. If a first round of RT-PCR did not produce a band, a second round was performed using an aliquot of the first round as a template. The results of either one or two rounds of RT-PCRs were suggestive of non-specificity to the intended chromosome 4p locus based either on the band pattern of the RT-PCR on an agarose gel, or the results of sequencing the band.

5.6.2. Novel genes identified

5.6.2.1. 74M11 Gene

Two STS, st74M11pt33 and st74M11pt34, were designed to EST evidence in ACeDB and screened on the cDNA libraries. Both amplified from the testes cDNA library, but st74M11pt34 also amplified from the foetal lung and st74M11pt33 also

amplified weakly from the neuroblastoma and T-cell cDNA libraries. The vectorette PCR revealed two bands for st74M11pt34 from the foetal lung cDNA library and two bands for st74M11pt33 from the testis cDNA library that were successfully sequenced. This revealed a gene structure of five exons, and revealed the 5'-3' orientation of the gene due to the presence of standard splice site consensus sequences. RT-PCR between st74M11pt33 and st74M11pt34 revealed that the gene was not expressed in lymphoblastoid cDNA. This, coupled with the restricted expression of the two STSs in the cDNA libraries, suggests that it is a relatively rare transcript. Two additional STSs, st74M11pt50 and st74M11pt51, were designed from EST evidence for RT-PCR on universal cDNA. RT-PCR was performed on universal cDNA in all combinations for the four STSs. The RT-PCR was positive between st74M11pt33 and st74M11pt51, but not for the other combinations. This was sequenced successfully. The resulting gene structure can be seen in Figure 5-13. The failure of some of the RT-PCRs may have been due to the low level of expression of the gene and the dilute nature of the cDNA.

The gene, positioned at the centromeric end of MR1, has conventional splice sites which orient it 5' to 3', but it does not have continuous ORF. The longest protein within the seven exons obtained runs from exon iv to v and is 88 amino acids long. This is quite short for a protein, but not unprecedented, and untranslated exons have been identified for a number of genes in the human genome (Zhang, 1998).

A BLASTn of the exon sequence, a BLASTp of the protein sequence, and tBLASTx of every possible ORF of the sequence to every possible ORF in the genome, reveal no similarities in any species. The intriguing possibility is that either this gene gives rise to a protein that is a member of a previously unidentified protein family, or that it is a non-coding RNA gene. RNA genes do not code for protein, but have been shown to act by multiple mechanisms and are involved in multiple cellular processes (Szymanski *et al*, 2003). The absence of a CpG island, splicing ESTs or homology to other proteins or genes meant that it was not possible to determine how much of the gene had not been identified.

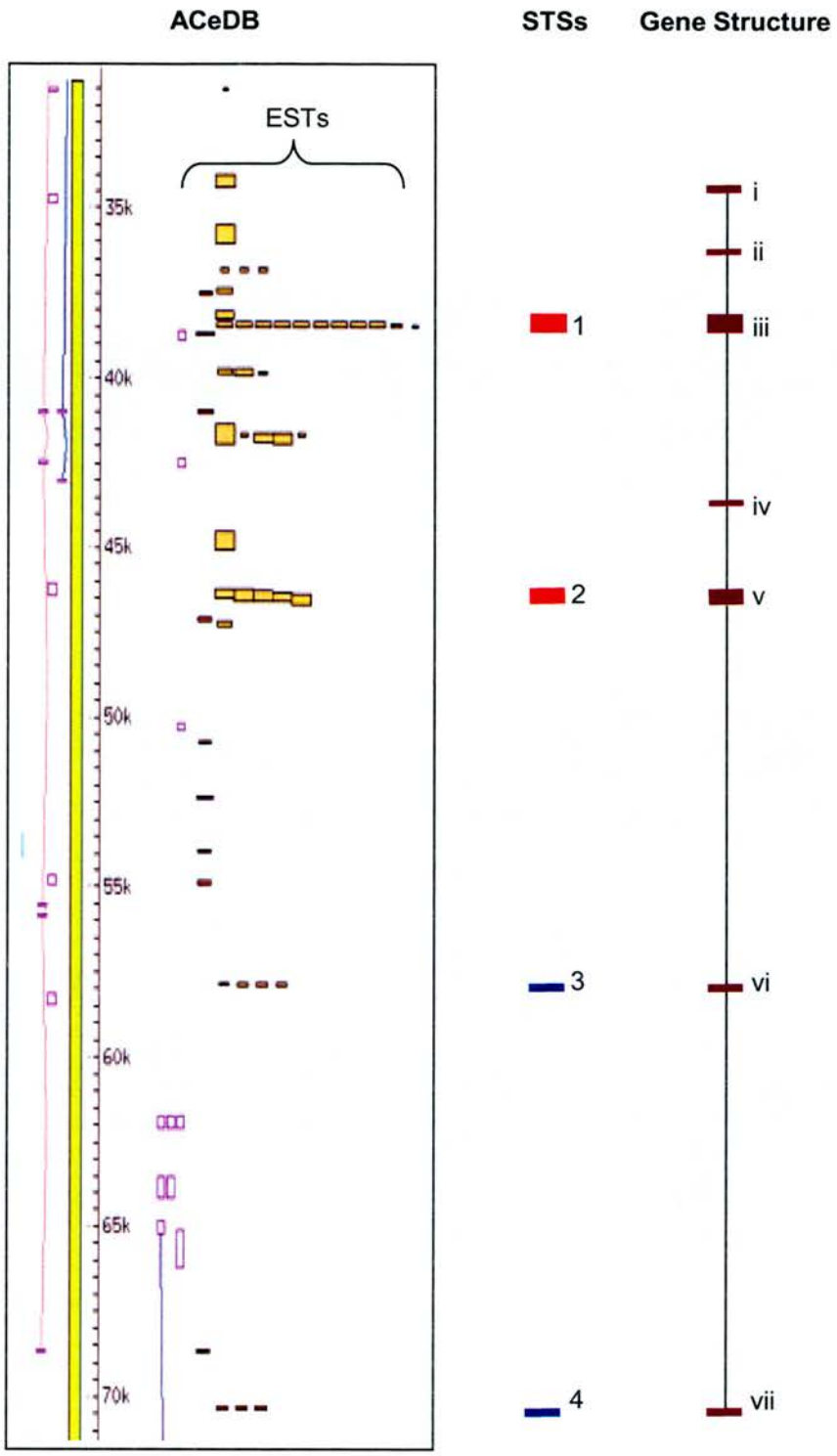


Figure 5-13: A novel gene was identified (74M11) on clone RP11-74M11. Four STSs (1: st74M11pt34, 2: st74M11pt33, 3: st74M11pt50, 4: st74M11pt51) were designed to EST evidence in ACeDB and screened on cDNA libraries (red) and cDNA (blue). Sequencing the bands obtained revealed the seven exon gene structure (purple).

5.6.2.2. G-protein-coupled receptor 125 (GPR125)

Two STSs, st362I16pt39 and st17E2pt38, were designed for cDNA library screening from EST evidence. Both STSs were expressed in seven out of the nine cDNA libraries on the primary plate, evidence for widespread expression of the gene. Three bands were successfully sequenced from the vectorette nested PCR; one for pt17E2pt38 and two for st362I16pt39. Subsequent to this, two LOC genes and an image clone were annotated on to the genomic sequence. The two bands sequenced for st362I16pt39 revealed that two LOC structures were in fact one gene (Figure 5-14). The fragment of genic sequence obtained for st17E2pt38 was located some way downstream of this gene. I postulated that the two structures may be connected, but RT-PCR did not support this. The RT-PCRs for st362I16pt39 did however show that the gene is expressed in lymphoblastoid cDNA, which together with the cDNA library expression evidence again suggested that it is an abundant and widely expressed gene. The ample EST evidence in ACeDB also supported this.

I performed seven RT-PCR reactions in order to extend the sequence I had obtained from the cDNA library screen. Two of the STSs gave a positive result from RT-PCR of the expected size, but failed to sequence and five failed to give a positive result. However, this may be because the assays did not work as the genomic DNA product was too large to test (except 166647.3 which did work on genomic DNA).

Subsequent to these findings, a gene corresponding to the two LOC structures was published as an orphan g-protein-coupled receptor GPR125 (Fredriksson *et al*, 2003). They identified three splice variants from over 50 ESTs and mRNA sequences. At this time, exactly how much of the gene they had identified was unclear. The paper suggested that GPR125 (AY181243) has 17 exons. However, mRNA AY181243 referred to a partial coding sequence at NCBI of 13 exons, and mRNA XM_291111, with 14 exons, was suggested to represent the full length coding sequence. It appeared therefore that neither Fredriksson *et al*, nor UCSC had identified the full gene structure. I predicted five exons 5' to the existing mRNA, based on homology

(97-98% identity) to a LOC gene on chromosome 18p. I had evidence from RT-PCR for the first five exons, but this was not supported by sequence data and I had not linked this to the remaining GPR125 exons. Therefore, I performed RT-PCR to link the putative 5' start of the gene to the existing exons, but was unable to confirm the link. This could have been due to the primers not working together in the RT-PCR assay. Subsequent to these experiments, the full gene structure, from the CpG island and including the five exons 5' to the existing mRNA, was published (still accession XM_291111).

The GPR125 gene, located right at the telomeric end of MR2, encodes an orphan G-protein-coupled receptor of unknown function. Structurally, it belongs to a group of LN-TM7 G-protein-coupled receptors that have long serine/threonine rich N-terminal domains that are suggested to be involved in cell-cell communication, functioning as ligands for other proteins (Fredriksson *et al*, 2003).

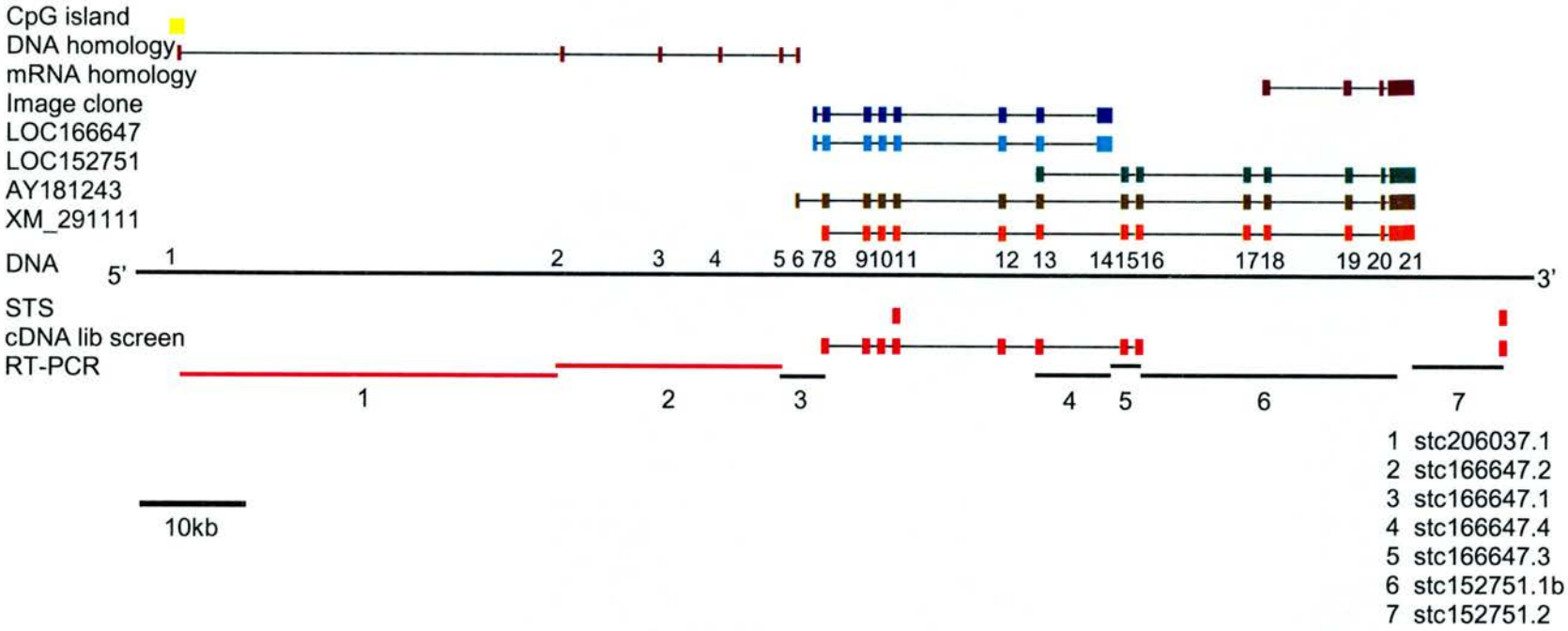


Figure 5-14: Orphan G-protein-coupled receptor 125 (GPR125). Two STSs (red) were designed for cDNA library screening. At this time, the CpG island (yellow), a region of homology to chromosome 18 (light purple), mRNA homology (dark purple) and EST evidence (not shown) were annotated on the sequence in ACeDB. cDNA library screening of the two STSs produced sequence extending over eight exons and sequence 3' to this (red). Two LOC genes (light blue and green) and an image clone (dark blue) were annotated onto the sequence in ACeDB subsequent to the cDNA library screen. Subsequent to this annotation, GPR125 mRNA was published (Fredriksson *et al*, 2003) and annotated in the public sequence databases (olive and orange). The homology to a LOC gene on chromosome 18p and the CpG island suggested the five 5' exons were part of GPR125. Seven RT-PCRs were performed (red and black bars) to extend the sequence obtained form the cDNA library screen. Sequence was obtained for two RT-PCR reactions (red bars). The current GPR125 mRNA (Dec 2003), still accession XM_291111, includes these extra five exons.

5.6.2.3. LOC91050

Four STSs, st106M4pt40, st106M4pt41b, st401G6pt42 and st401G6pt47, were designed to screen the cDNA libraries based on EST evidence adjacent to the SOD3 gene in MR2 (Figure 5-15). The nested vectorette PCR produced bands that were successfully sequenced for STSs st106M4pt40 and st106M4pt41b. This produced sequence which provided no evidence for splicing. Subsequent to this, the sequence was annotated with LOC91050 and the hypothetical protein DKFZp761B107. LOC91050 followed the structure of hypothetical protein DKFZp76B107, except without exon one. RT-PCR between st401G6pt42 and st401G6pt47 on lymphoblastoid cDNA gave a band that was successfully sequenced and followed the LOC structure. I designed four further STSs for RT-PCR and obtained a positive result and sequence for stc91050.1, stc91050.2b and stc91050.5. Therefore, I had extended the hypothetical protein DKFZp761B107 at its 3' end to include more EST evidence. It could be that there are two alternative splice variants, one extending the length of DKFZp761B107 and the other extending to include further 3' exons.

Construction and a BLASTp similarity search of the ORF revealed that the gene has a chromosome segregation ATPase protein domain, contained within both the long and the short version of the protein, and could therefore be involved in cell division and chromosome partitioning. This is a fundamental cellular process without which a cell cannot survive and this gene is therefore not a good candidate for psychiatric illness. Therefore I did not investigate the gene structure further.

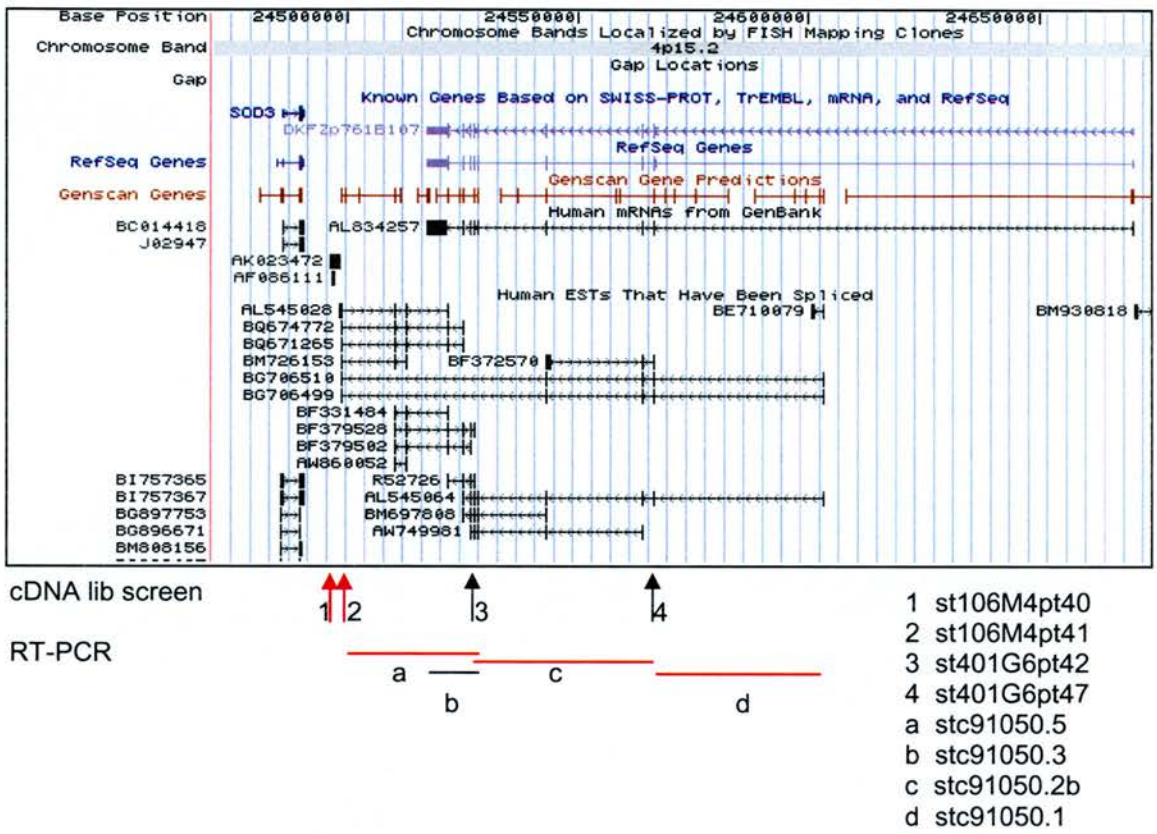


Figure 5-15: Novel gene LOC91050. Figure shows a screen shot from the genome browser at the University of California, Santa Cruz (genome.csi.ucsc.edu). The NCBI (www.ncbi.nlm.nih.gov) predicted gene LOC91050 is not shown on the genome browser but has an identical structure to the annotated mRNA AL834257 and hypothetical protein DKFZp761B107. Four STSs were designed, based on the mRNA and EST evidence, to screen on the cDNA libraries (arrows). Sequence was obtained for STSs 1 and 2 (red), but not for STSs 3 or 4 (black). Three of the four subsequent RT-PCRs (red and black lines) identified the remaining exons (red). RT-PCR was not carried out to confirm the presence of the furthest 5' exon (right of the figure).

5.6.2.4. LOC202025

Two STSs, st448G15pt4 and st448G15pt2, were designed to screen the cDNA libraries (Figure 5-16). LOC202025 was annotated subsequent to the cDNA library screen, and follows the structure of the AL713660 mRNA and the hypothetical protein DKFZp667E0512. One STS gave a band from the nested experiment that was successfully sequenced and exactly matches one LOC exon then splices to near the 3'UTR of the WDR1 gene in MR1. This was interesting because it suggests that the LOC is an alternative 3' UTR of WDR1 (splice sites suggest that the direction of the two genes are the same). It also suggests that the LOC, even if it is separate to WDR1, is itself alternatively spliced since or has been incorrectly annotated, as its second exon was absent from my results.

RT-PCR performed on lymphoblastoid and universal cDNA failed to extend the transcript further. Bands a and b were obtained (Figure 5-16). However, the signal was weak and a second round of RT-PCR, to increase the amount of template for sequencing, did not work. An attempt to sequence the low level of product also failed.

A BLASTp protein similarity search of the hypothetical protein DKFZp667E0512 does not reveal any conserved protein domains. The most significant match reveals 31% amino acid identity to a hypothetical protein in the mycobacterium bovis bacteria, a hypothetical protein in the mycobacterium tuberculosis bacteria and a transcriptional regulator in the pseudomonas putida bacteria.

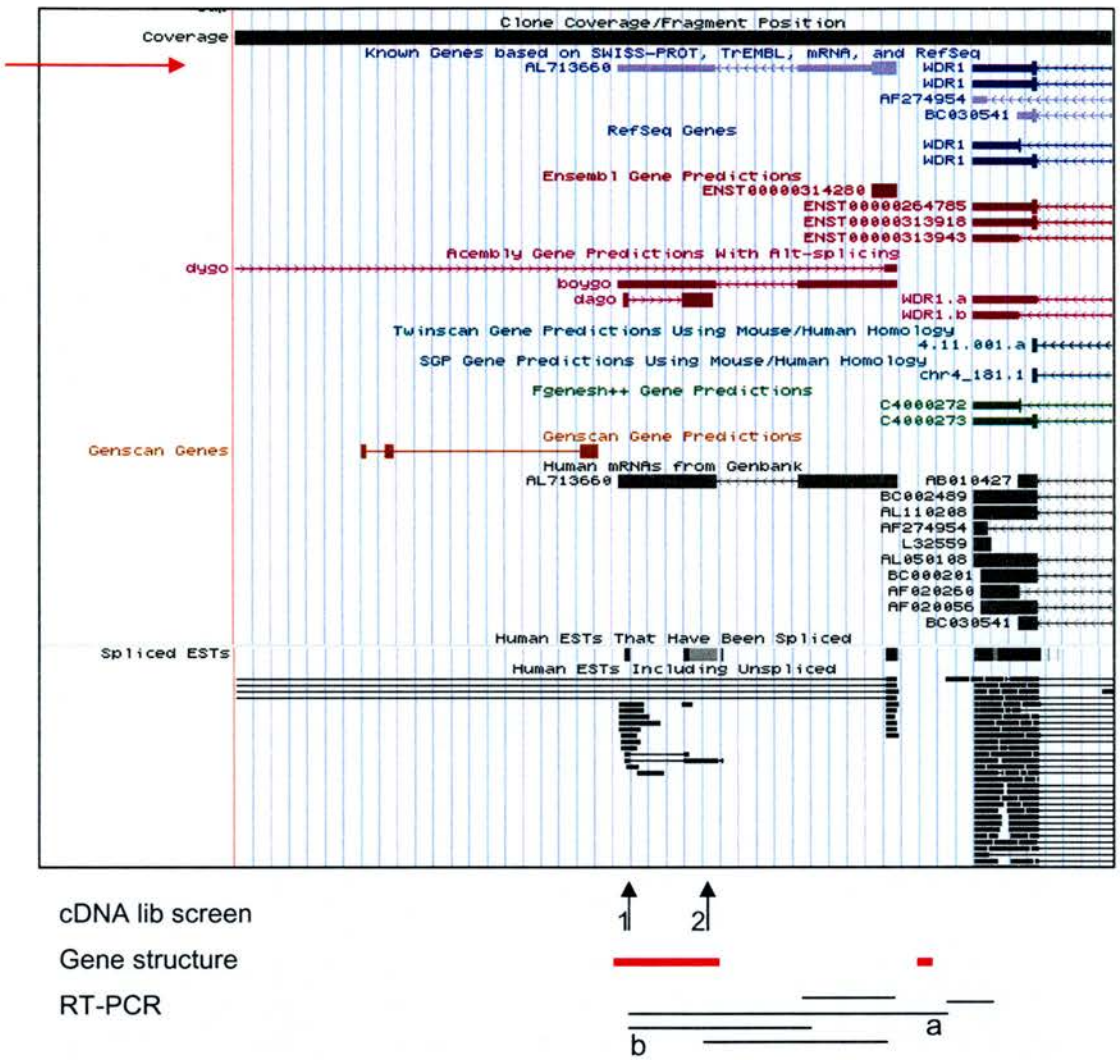


Figure 5-16: Novel gene LOC202025. Figure shows a screen shot from the July 2003 genome browser at the University of California, Santa Cruz (genome.csi.ucsc.edu). The NCBI (www.ncbi.nlm.nih.gov) predicted gene LOC91050 is not shown on the genome browser but the mRNA AL713660 (red arrow), at the 3' end of the gene WD repeat domain 1 (WDR1), follows the structure of LOC202025. Two STSs (black arrows) were used in cDNA library screening and sequence was obtained (red). The five RT-PCR reactions performed to extend the transcript (black lines) failed. Bands were obtained for 'a' and 'b', but they failed to sequence.

5.6.2.5. LOC132895

RT-PCR between two exons of LOC132895 (Figure 5-17) in MR2 revealed sequence specific to chromosome 4, although RT-PCR between the ESTs annotated on Figure 5-17 did not. The sequence did not splice between the exons and did not contain a continuous ORF. Therefore, I suspected that the band obtained was the result of genomic DNA contamination in the cDNA, especially since the expression and prediction evidence supporting the LOC gene structure was not very convincing. Therefore I concluded that the LOC was not a gene.

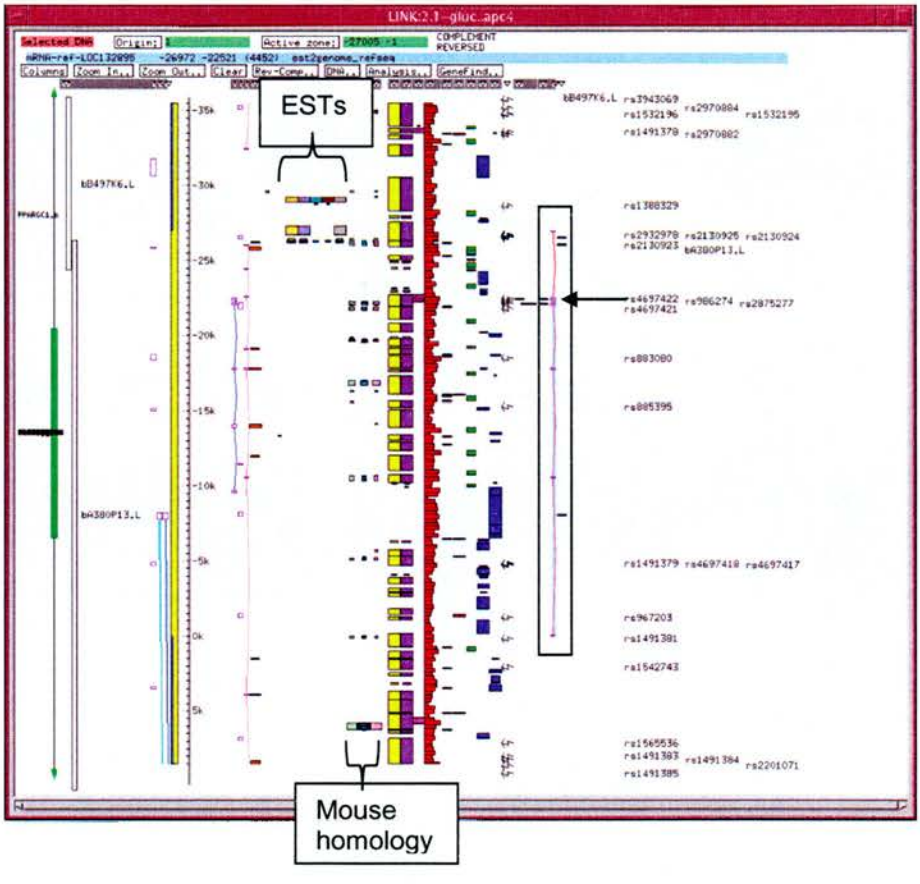


Figure 5-17: A LOC gene, LOC132895 (boxed), from the National Centre for Biotechnology Information (www.ncbi.nlm.nih.gov) annotated in ACeDB. The exons were partially supported by mouse homology but the EST evidence did not overlap with any of the LOC exons. Due to the lack of supporting evidence, only one STS was designed for RT-PCR (arrow).

5.7. Pseudogenes

It is important to characterise the pseudogenes in the minimal regions in order to eliminate them from further study. Pseudogenes are the result of gene duplication that can occur by retrotransposition, where a processed mRNA is inserted into the genomic sequence, or by genome duplication. Pseudogenes are non-functional genes either due to a failure of transcription or translation or by the production of a protein that does not have the same functional repertoire as the protein encoded by the original gene. Evidence of retrotransposition can be found by the presence of a poly-A tail incorporated into the genomic sequence and the absence of intronic and untranslated sequence. The genes in a duplicated genomic region can either remain functionally equivalent, take on new adaptive functions or become neutralised as a pseudogene by the accumulation of mutations. Most pseudogenes are no longer expressed, although it is possible that the insertion of a processed mRNA into a promoter, or genome duplication, may result in continued expression of the gene. However, a novel promoter may alter the pattern of transcription (see Mighell *et al*, 2000 for a review).

There were ten LOCs in MR1 and MR2 that had exons with pseudogene characteristics, although the LOC exon structure and the pseudogene structure did not correspond in five of the LOCs. I studied five LOCs for expression in cDNA. I obtained no product from lymphoblastoid or universal cDNA for any except LOC133260. This was subsequently identified as a chromosome 4 pseudogene in the NCBI public database.

I predict that there are seven pseudogenes in MR1 and four in MR2 (Table 5-12). Nine of the pseudogenes are very similar in structure. They have no poly-A tail and multiple matches to protein domains in the Swissprot database, but none of the matches are exact as evidenced by a disrupted ORF. Some have more than one 'exon' if the protein homology splices. No corresponding expressed gene can be identified in the genome and the multiple protein domain homologies are the only

expression evidence (Figure 5-18). This suggests that they are very old pseudogenes that have been fragmented over time and lost their structural similarity to the original gene.

Pseudogene locus (RP11-)	Minimal Region	>1 Exon	Disrupted ORF	Genomic poly-A	Corresponding spliced gene
2205P10 a	1	Yes	Yes	No	No
2205P10 b	1	Yes	Yes	No	No
2205P10 c	1	No	Yes	No	No
264E23	1	No	Yes	No	No
494H11 a	1	Yes	Yes	No	No
494H11 b	1	Yes	Yes	No	No
287J14	1	Yes	No	Yes	RNPS1 (16p) - 8 exons
10G12 a	2	No	No	Yes	S27 (1q) - 4 exons
302F12 a	2	Yes	Yes	No	No
302F12	2	Yes	Yes	No	No
401G6	2	Yes	Yes	No	No

Table 5-12: Pseudogenes in Minimal Region One (MR1) and Minimal Region Two (MR2). Likely pseudogenes were characterised by the number of exons, the presence of a disrupted open reading frame (ORF), the presence of a poly-A tract in the genomic DNA sequence immediately after the terminal exon and whether a corresponding transcribed gene could be located elsewhere in the genome.

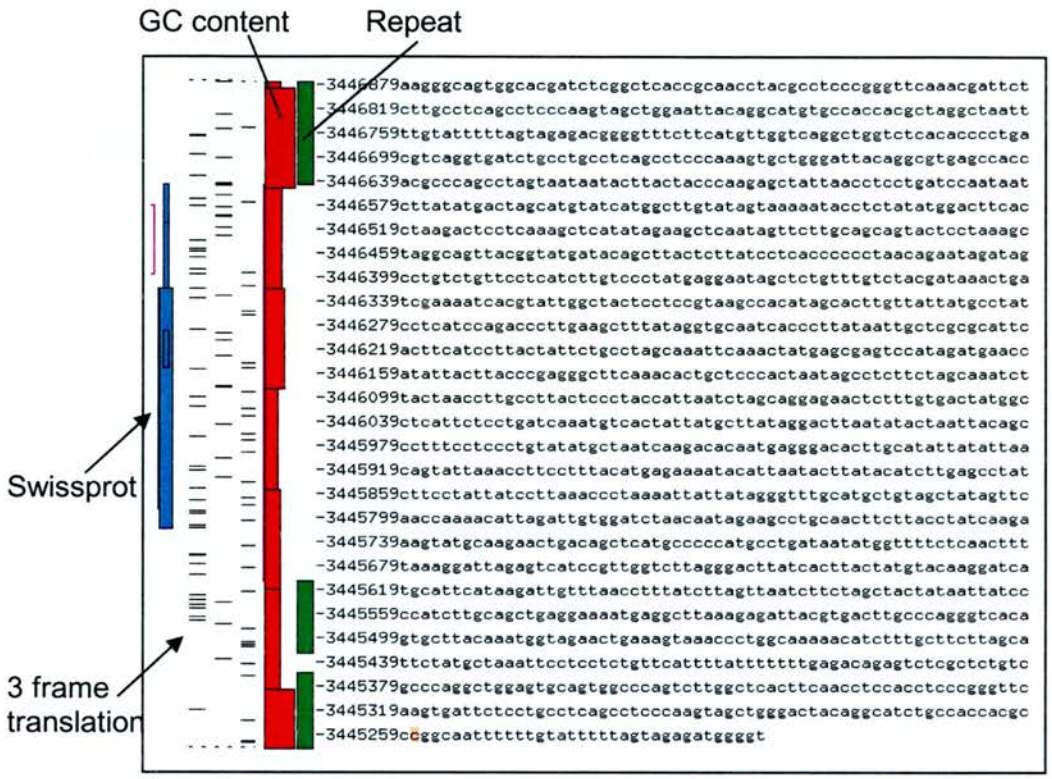


Figure 5-18: A putative pseudogene on chromosome 4p annotated in ACeDB. The region shows homology to an oxidoreductase protein domain (blue bar: the wider the bar, the greater the homology) identified in the protein database Swissprot. There is no genomic poly-A tract and no open reading frame. Three of the possible six translation frames are annotated on the figure. The horizontal lines mark the position of a stop codon.

The pseudogene on clone RP11-10G12 in MR2 shows 100% identity to the 83 amino acid protein for S27 on chromosome 1q (Figure 5-19). The gene on chromosome 1q has four exons and the gene on chromosome 4 has one. The sequence of this exon on chromosome 4 is exactly identical to the four exons of the S27 gene, but with the intronic sequence spliced out and a genomic poly-A tail of 12 nucleotides beginning immediately after exon 4. The pseudogene does not contain the same UTR sequence as the gene. However, the pseudogene has a continual ORF and therefore may still be expressed if it has been inserted in a region under the promoter of another gene. However, I could not detect any expression from RT-PCR of the sequence in either lymphoblastoid or universal cDNA (Section 5.5.2.2).

A second gene, Similar to RNPS1 in MR1, is also likely to be a pseudogene. The pseudogene has two exons and possesses a poly-A tail in the genomic sequence immediately after the second exon. The RNPS1 gene is on chromosome 16p and has eight exons, but the first exon is non-coding. The gene on chromosome 4 codes for the same protein as the RNPS1 gene on chromosome 16, contained entirely within its large second exon, with a complete ORF. As with the S27 pseudogene, the retention of the complete ORF means that it may be expressed, although this is unlikely. RT-PCR was not performed to confirm this.

Figure 5-19: Pseudogene S27 on chromosome 4p. **A:** The genomic sequence containing the S27 gene from chromosome 1. The exons encoding the protein (blue) and the untranslated regions (red) are highlighted. The poly-adenylation signal is boxed. **B:** The S27 pseudogene on chromosome 4p codes for the same protein and shows the same exon sequence. The sequence contained in four exons on the chromosome 1 gene are contained within one exon on chromosome 4p. This lack of introns and the genomic poly-A tract observed immediately after the exon (underlined) suggest that it is a pseudogene.

5.8. Discussion

The characterisation of two susceptibility regions to psychiatric illness on chromosome 4p is described. A combination of bioinformatics analysis, cDNA library screening and RT-PCR has allowed the identification of four novel genes and two processed pseudogenes, and allowed the assessment of these and the known genes with respect to location and Giesma staining. Problems with non-specificity of the cDNA library screening hampered the identification of genes and the majority of STSs tested by RT-PCR were negative. However, it is impossible to conclude anything from a negative result.

Historically, the G-banding technique helped identify the human chromosomes and chromosomal regions and characterise GC content (Bickmore and Sumner 1989). Low GC content is also thought to represent gene poor regions. This is upheld in these two candidate regions, with the majority of genes positioned in a light band. In accordance with what has been shown for other genomic regions, the majority of the genes have a CpG island on or near exon one and the most common poly-adenylation signal is AATAAA. Four of the genes did not have either of the two most common poly-adenylation signals and this therefore lends support to the findings of Beaudoing *et al* (2000) that rarer signals are likely to exist.

It is difficult to predict how many genes there are left to identify in MR1 and MR2. The 23 known genes in MR1 and MR2, totalling approximately 8Mb, gives a gene density of approximately 2.8 genes per Mb. The estimated genome wide gene density has been found to vary between the chromosomes, from five genes per Mb on chromosome 13 to 23 genes per Mb on chromosome 19, although these estimates of gene number must exclude some unknown genes. Furthermore, regional differences in gene density make predicting gene density in the relatively small minimal regions problematic. It seems reasonable to predict however that not every gene has been identified.

Genes are associated with functional elements that help predict their structure. However, exon and gene prediction programmes are not perfect. For example, the accuracy of exon prediction is dependent on length, where very short (less than 70bp) and very long (longer than 200bp) exons are not well predicted (Rogic *et al*, 2001). In the region of 20-30% of Genscan exons are predicted incorrectly (Guigo *et al*, 2000; Burge and Karlin, 1997) with the first and last exons most likely to be inaccurate (Rogic *et al*, 2001). However, the training and testing methods of these programmes have questionable applicability to the prediction of real gene structure. For example, Rogic *et al* (2001) tested seven different gene prediction programmes with GenBank gene sequences submitted after the programmes had been developed, in order to subject them to novel sequences, but removed genes from the test set that had unconventional splice sites. It is likely that they did this to remove incorrect GenBank submissions, but it probably also biased the accuracy of gene prediction programmes to well characterised gene structures. Furthermore, they measured the accuracy of the prediction on the plus strand, the same strand as the GenBank submission, but ignored any predictions made to the minus strand. Therefore, predicted accuracy is again inflated since the correct absence of a prediction is not accounted for.

A consensus between different prediction methods is desirable. Bioinformatic evidence coupled with real transcript evidence, ESTs for example, is better evidence for a novel gene than gene prediction programmes alone. However, there will be many more for which there is currently no known transcript evidence. Low expression levels of the protein or constrained expression times during development will ensure that there are genes for which no expression data is obtained. In addition, genes with an unusual structure will be missed by prediction programmes. Therefore, there will be genes in MR1 and MR2 that have not yet been identified.

Annotation of full length coding features is only the first step in the characterisation process. Regulation, tissue distribution and function pose greater challenges. A BLAST similarity search can give some idea of a gene function. For example, the

novel gene LOC91050 contained a characterised protein domain that strongly suggests a function for the protein. However, this is not always the case. For example, the LOC202025 and 74M11 genes do not possess any known protein motifs and therefore their function remains uncertain. Furthermore, the lack of an ORF in the 74M11 gene raises the interesting question of whether it might be a non-coding RNA gene. As more and more protein coding genes are identified in the human genome, attention will shift towards the elements of the genome largely ignored to date, such as characterising families of non-coding RNA genes.

It is also important to define pseudogenes in the minimal regions. Pseudogenes are generally defined by a set of characteristics including a disrupted ORF, a genomic poly-A tail, the lack of introns and promoter sequence or the presence of evidence for genome duplication of the original locus (Mighell *et al*, 2000). However, whether the pseudogene is still functionally expressed and therefore can really be classified as a pseudogene is harder to define. For example, the pseudogene of S27 retains the complete original ORF. It lacks the original promoter and has a genomic poly-A tail and has therefore arisen by retrotransposition. However, it is impossible to determine from this evidence alone whether it is expressed and if it is expressed whether it is functional in a complimentary or an altered way to the original protein. RT-PCR was required to determine that it is not expressed, although it is still possible that the pseudogene has a specific spatio-temporal expression that has not been captured in the cDNA screened.

The identification of every gene in MR1 and MR2 is an important aim for the group in order to direct SNP detection and association analysis. Since the aim is to identify the susceptibility SNP in the families, directing SNP detection towards coding regions is desirable. These results have gone some way towards completing the transcript map of these two candidate regions for psychiatric illness.

Chapter Six

Genetic Analysis of the Superoxide Dismutase 3 (SOD3) Gene

Genetic Analysis of the Superoxide Dismutase 3 (SOD3) Gene

6.1. Introduction

Here I describe sequence analysis and association studies of the Superoxide Dismutase 3 (SOD3) gene. As mentioned previously, linkage analysis has identified large regions on chromosome 4p that are inherited with psychiatric illness in four families. I have described in Chapters 3 and 4 how these regions have been compared and that the overlap observed permits the definition of smaller candidate regions. However, these sub-regions are still large and contain a number of genes, as discussed in Chapter 5. SOD3 was positioned on chromosome 4pter-q21 by Hendrickson, *et al,* (1990). Today, its exact genomic position can be ascertained from the human genome sequencing project (HGP) and it is located in MR2. MR2, the susceptibility region of F22, F50 and F48, covers ~4.3Mb of sequence and contains 12 other known genes.

In humans, SOD3 is expressed in the heart, brain, placenta, lung, liver, muscle, kidney and pancreas (Folz and Crapo, 1994). In the mouse brain it has a wide distribution, and has been observed in the cortex, the subiculum, the entorhinal cortex, the striatum, the nucleus accumbens, the ventral pallidum, the hippocampus, the diencephalons and hypothalamus, and a small number of scattered cells in the midbrain, hindbrain, cerebellum and brainstem (Oury *et al*, 1999). This localisation of SOD3 to discrete areas suggests that it plays a specialised role in the brain. In most tissues it is anchored to the extra cellular matrix (ECM). However, in the brain, SOD3 is observed in the cytoplasm of neurons and hypothalamic tanycytes and shows a granular (possibly vesicular) distribution.

The function of SOD3 in the brain remains unknown. Mice with SOD3 knocked-out or over-expressed have been shown to exhibit impairments on tasks involving

hippocampal spatial learning and memory (Levin *et al*, 1998). This memory and spatial learning deficit is affected by motivational state. Food restriction increased their ability to learn but did not alter the learning rate of control mice (Levin *et al*, 2000). The actions of SOD3 have also been suggested to be linked to nitric oxide (NO) signalling. SOD3 controls levels of extracellular O^{-}_{2} (Oury *et al*, 1996), and NO is inactivated by O^{-}_{2}. Like SOD3, nitric oxide synthase is found in the neurons in the hippocampus, and studies have found that NO is involved in behaviour, and learning and memory functions (Hawkins, 1996).

SOD proteins are known for their antioxidant properties. Oxygen free radicals are toxic compounds produced by cells during normal utilization of oxygen. SOD proteins form one type of cellular protein that detoxifies oxygen free radicals by catalysing the conversion of superoxide anion radicals to hydrogen peroxide and molecular oxygen. It is unknown whether it is this action that lies behind SOD3's role in learning and memory.

The SOD3 protein is encoded by a three exon gene. The protein is contained entirely within exon three; exons one and two are non-coding (Figure 6-1). Non-coding exons are quite common, estimated to occur in greater than 35% of human genes and are distinguished from untranslated regions (UTRs) by the presence of introns (Zhang, 1998). The gene transcribes a 240 amino acid protein. The first 18 amino acids form the signal peptide. Amino acids 96-193 show ~50% sequence homology to the final two thirds of all known eukaryotic SOD1s, including the amino acid responsible for catalytic activity. Amino acids 194-222 are hydrophilic and contain nine positively charged amino acids, which probably confers affinity to heparin (Hjalmarsson *et al,* 1987). This heparin binding site may anchor SOD3 to the ECM. Proteolysis of the domain abolishes heparin affinity and results in the clearance of SOD3 from the tissue into the serum (Hjalmarsson *et al,* 1987).

Only one functional variant in SOD3 has been described in the literature. Sandstrom *et al* (1994) identified an arginine to glycine substitution at amino acid position 213,

in the centre of the C-terminal cluster of positively charged residues that defines the heparin binding domain. Arginine is positively charged and glycine is uncharged. This variant was observed in 2.2% of 504 random blood donors and results in a 10 fold increase in plasma SOD3 caused by a reduction of affinity for heparin and thus a loss of SOD3 from the ECM to the plasma. The enzymatic activity of SOD3 was not affected. No obvious phenotype was observed with this mutation, although all individuals tested were heterozygote.

It is prudent to study such candidate genes for SNP variation in the linked families. Firstly, this may identify variation that alters the amino acid sequence, and potentially the function, of the protein. Secondly, it contributes to defining the disease associated haplotype in each family. Whilst this is interesting in itself, it also enables the analysis of haplotype sharing between the families. A shared disease haplotype across a gene would provide support for a common ancestral allele and point to the possible role of the gene in the susceptibility to psychiatric illness. The number of control chromosomes available is limited by family size, but this is generally not a problem since two of the families are large. However, there is a maximum of only four disease chromosomes in which to detect allele sharing. Despite this, it is still possible to obtain a statistical measure of the difference in the frequency of control and disease alleles and determine whether allele sharing is expected above chance.

An association study offers a complimentary approach to studying the families. It involves comparing the variation observed on the chromosomes of a large population of unrelated cases versus a large population of unrelated controls. The advantage of studying many different disease chromosomes means that many different recombination events surround the susceptibility locus. Consequently, if a positive association is observed it will only extend over a relatively small genomic region. Furthermore, association analysis has the power to detect variants with small effect sizes. Therefore, performing association analysis within the linked regions in the four families will provide a way to study these large regions in more detail.

In this chapter I describe the results of family and association analyses of the SOD3 gene. I have identified the disease associated SNP haplotypes of the four families across the SOD3 gene, and describe the use of these and other SNPs in an association analysis in a group of unrelated case and controls individuals.

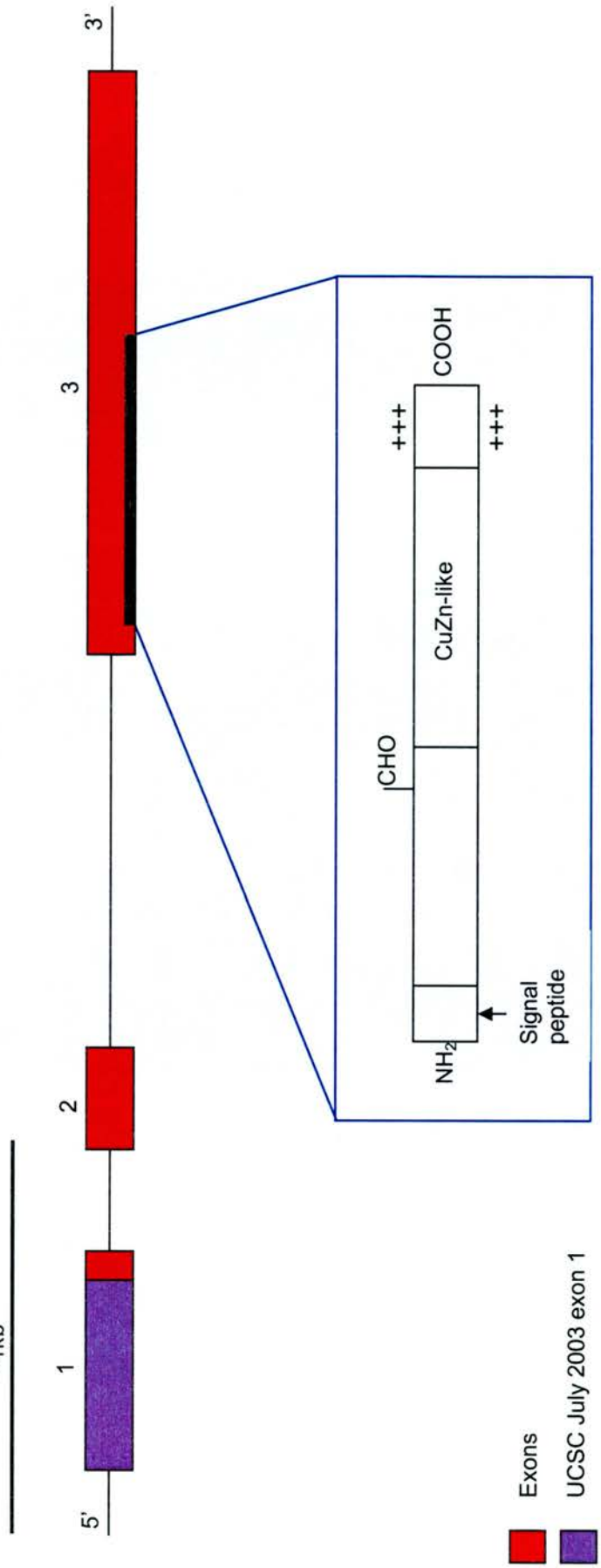


Figure 6-1: The structure of the Superoxide Dismutase 3 (SOD3) gene and peptide. The original publication (Folz and Crapo, 1994) of the SOD3 gene comprised three exons. Two sequence databases report a different exon structure. The University of California, Santa Cruz (UCSC) sequence release extends the size of exon 1 and Ensembl annotates only one exon. The peptide is contained within the third exon. The protein has four domains. The signal peptide is shown by the arrow, followed by a Glycosylated (CHO) amino terminal peptide domain, a region of homology to SOD1 and the C-terminal domain with multiple charged basic residues (+) critical for binding heparin glycosaminoglycans (adapted from Folz & Crapo, 1994).

6.2. SNP Identification

6.2.1. Sample

In order to increase the chance of identifying the SNP responsible for susceptibility to psychiatric illness in the families, SNPs were identified by sequencing family members on the AS DNA panel (Figure 2-5). The panel consists of 46 indivduals with a total of 45 independent chromosomes: four disease chromosomes and 41 control chromosomes. This provides a larger sample from which to identify SNPs and also enables the assessment of allele sharing between the four families to be carried out across the gene.

6.2.2. STS Design and SNP Detection

STSs were designed to cover the coding regions and splice sites of the gene and part of the 3' and 5' UTRs. The aim was to capture SNPs that alter the amino acid sequence of the protein and SNPs involed in the function of the promoter or splicing mechanisms. STSs were designed to an optimum length of ~500bp, to overlap by at least 50bp, and to extend at least 50bp into an intron. A PCR was optimised for each STS on three CEPH control DNAs. The amplified product from each indivdual was sequenced with the reverse primer and run on an ABI 377 or 3730 Genetic Analyser. Sequence chromatograms were aligned using the phredPhrap software and visualised with the Consed programme. Sequences were then manually for polymorphisms. SNPs were positively identified if at least two heterozygotes were observed. However, SNP ih170 was positively identified in only one heterozygote, although the sequence quality was exceptional and therefore the SNP was scored.

6.2.3. SNPs Identified

Sequencing the SOD3 gene in individuals from the AS panel identified eight SNPs. Four of these were in dbSNP (prefixed 'rs') and four were novel (prefixed 'ih') (Figure 6-2).

A number of SNPs from other sources were chosen to provide an even distribution of SNPs across the gene for association analysis. There are a number of SNPs (September 2003) across the SOD3 gene for which an assay has been developed and is available from Applied Biosystems. This 'Assay on Demand' service not only provides a ready made SNP assay, but also provides frequency data from a population of 45 caucasians. One SNP (C_2668721_10) with a caucasian frequency of greater than 0.1 was chosen. The database dbSNP (www.ncbi.nih.nlm.gov/SNP) is a database in which independent researchers deposit the position of SNPs that they have identified from a population. Therefore a SNP can be submitted a number of times. Five SNPs that had been confirmed, at least twice, by independent research groups (rs800399, rs800414, rs800447, rs800448, rs2324580) were chosen, on the premise that more than one independent submission would increase the liklihood of the SNP being present in our population. Figure 6-2 details the position of these six additional SNPs.

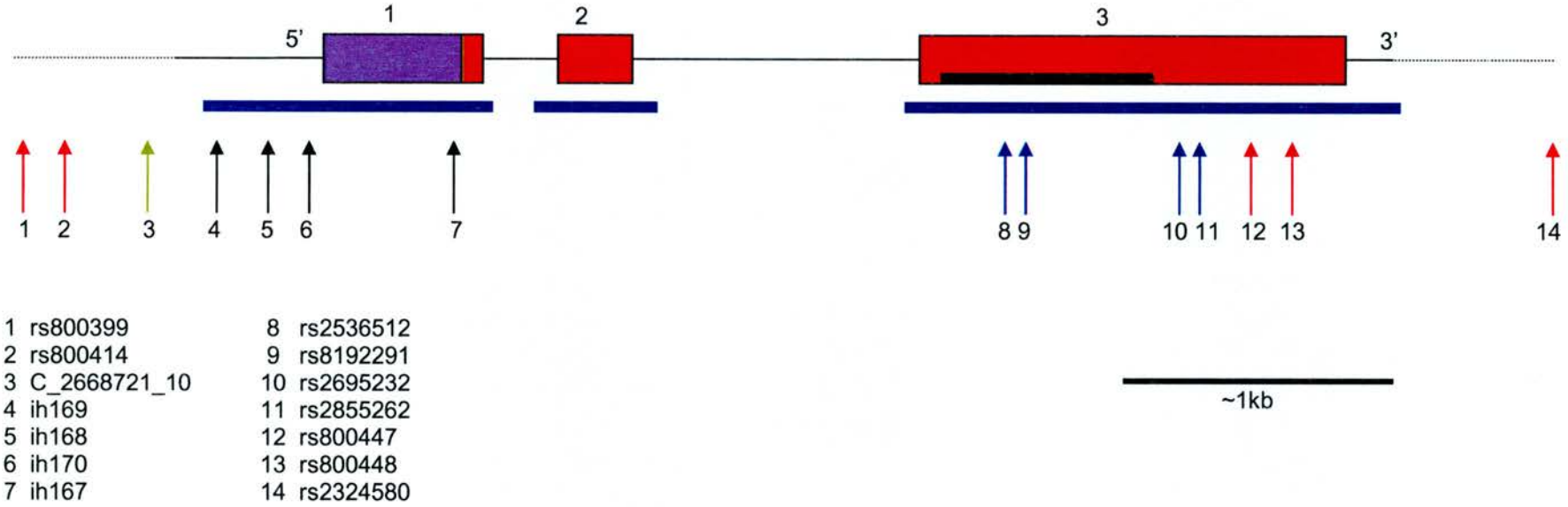


Figure 6-2: Single nucleotide polymorphism (SNP) identification in the Superoxide Dismutase 3 (SOD3) gene. The regions of SOD3 that were screened for SNPs on the Allele Sharing (AS) DNA panel, a DNA panel of 46 individuals, by sequencing are shown by the blue bars. SNPs identified in only the AS panel are marked with black arrows. SNPs identified in the AS DNA panel that are also in the public SNP database dbSNP are marked with blue arrows. The SNPs used for association analysis from dbSNP are marked with red arrows. The SNP available as a genotypng assay from Applied Biosystems is marked with a green arrow. SNPs names are provided. The dotted line is not to scale (SNP 1, 2 and 3 are ~16, ~12.4 and ~2.4kb from the 5' end of exon 1 respectively. SNP 14 is ~5.4kb from the 3' end of exon 3).

6.3. SNP Analysis

6.3.1. Amino Acid Changes

Of the eight SNPs identified from sequencing the AS panel, two are located in the peptide sequence: rs2536512 and rs8192291 (Figure 6-2). SNP rs8192291 is a c>t change of the first nucleotide of codon 'ctg' at amino acid position 71. Both alternative codons code for leucine. SNP rs2536512 is an a>g change of the first nucleotide of codon 'acg'. This codes for threonine at amino acid position 58, and is changed to an alanine if the first nucleotide is a guanine. The frequency of this variant could provide some idea of its functional importance. As has been seen, the SNP was identified by sequencing chromosomes from the AS DNA panel. The AS panel gives a total of 45 independent chromosomes. The variant was observed on 18 chromosomes, which translates into a frequency of 40%. Therefore, this is a common variant in this sample. This suggests that an alanine substitution would not have a significant impact on protein function. However, the Common Disease/Common Variant (CD/CV) hypothesis (Discussed in detail in Section 1.5.5) suggests that common variants will underly the susceptibility to common complex diseases such as psychiatric illness. Therefore, a susceptibility variant would be expected to be common in the control chromosomes. Despite this, it is possible that assortative mating has made the SNP is more common in the families compared to the general population. The mutation was observed on the disease chromosome of the Celtic families 22, 59 and 50, but not the Jewish family 48. It forms part of the most common eight SNP haplotype observed from 32 chromosomes (Section 6.3.2).

The properties of each amino acid give some idea of whether a substitution would be tolerated or not. Bordo and Argos (1991) identified the ajmmino acid substitutions that are least likely to disturb protein structure. According to their results, threonine and alanine can be substituted with 95% confidence. Despite this, alanine and threonine are placed in different positions on four different hydrophobicity scales (Janin, 1979; Wolfenden *et al*, 1981; Kyte and Doolittle, 1982; Rose *et al*, 1985).

Alanine is positioned near the top in each scale, and is consistantly considered to be more hydrophobic than threonine which is placed near the middle in each scale. Therefore, this difference could alter the secondary structure of the protein because a more hydrophobic residue may be more likely to be buried inside the protein than a less hydrophobic residue.

Figure 6-3 displays the patial results of a Protein Identification of Unknown Sequences (PIX) analysis at the MRC Rosalind Franklin Centre for Genomics Research (MRC-RFCGR, www.hgmp.mrc.ac.uk). A PIX analysis runs twelve different types of analysis programme on a peptide sequence. I therefore ran the analysis on the two alternative proteins to assess what impact an amino acid substitution might have on predicted protein structure and functional domains. The results of most interest displayed in the figure are the secondary structure prediction programmes DSC and Simpa96, the protein domain predictor programmes Pfam and BLOCKS, and the antigenicity predictor programme Antigenic. The protein with the alanine variant (6-3: B) at position 58, due to its hydrophobicity, would be predicted to be buried. However, it is predicted to be on the surface by Antigenic. It is also likely to be part of a fairly continuous coil structure. The protein with the threonine variant (6-3: A) has an inconclusive coil/beta strand structure or an alpha helix structure for the five or six amino acids immediately 5' to it and the threonine itself is predicted to be in a non-antigenic region, suggesting that it is not exposed on the protein surface. This is not consistant with the findings of the four scales of hydrophobicity mentioned above in which alanine is more hydrophobic than threonine. The Pfam Copper Zinc superoxide dismutase protein domain prediction is unchanged by the alternative amino acid, but a BLOCKS protein domain prediction is created. The BLOCKS programme predicts a MADS-box domain from amino acids 46 to 59. A MADS-box domain is a DNA-binding domain that has been identified mostly in plants but also in several vertebrates, including humans. A protein with a DNA binding domain might be involved in the regulation of gene transcription. However, the domain prediction is identical at only seven out of fourteen amino acids and is therefore not likely to be real.

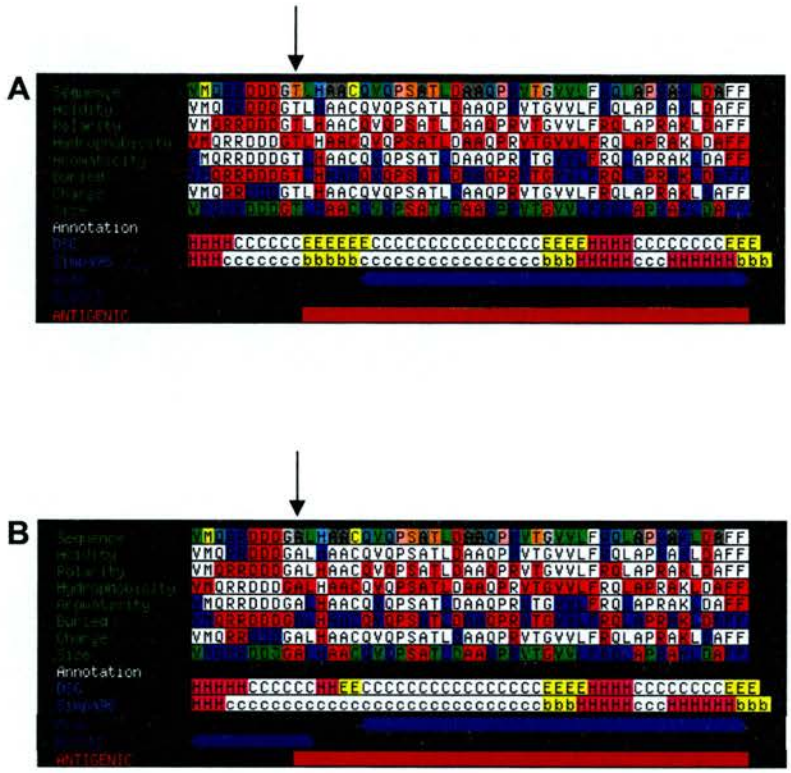


Figure 6-3: The results of a Protein Identification of Unknown Sequences (PIX) analysis (carried out at the MRC-RFCGR: www.rfcgr.mrc.ac.uk) for the part of the Superoxide Dismutase 3 (SOD3) protein that contains SNP rs2536512. The first line follows the RasMol colouring scheme (The molecular graphics programmes to visualise the 3D structure of proteins, developed by Roger Sale, University of Edinburgh, where each amino acid is a different colour). The next seven lines show the individual properties of each amino acid. White is the default colour. Red = acidic, blue = basic. Red = polar. Red = hydrophobic. Red = aromatic, blue = aliphatic. Red = surface, blue = buried. Red = positive, blue = negative. Red = tiny, green = small, blue = large. The sequence is annotated with two secondary structure prediction programmes DSC and Simpa96, two protein domain prediction programmes Pfam and BLOCKs, and the antigenicity predictor programme Antigenic. The secondary structure programmes predict three structures: helix (H), coil coil (C/c) and beta strand (E/b). The two amino acid variants created by SNP rs253652 are shown (arrow). **A:** Threonine variant , **B:** Alanine variant.

The poor antigenicity of the alanine variant, compared with the surface availability of the threonine, could have an impact on protein function if the alanine is predicted to create or abolish a functional motif. PROSITE (Bairoch *et al*, 1997) (www.expasy.org/prosite) is a database of protein families and domains. It searches a protein sequence for biologically relevant sites, patterns and profiles. Therefore I used this to determine whether the two alternative proteins would differ at a functional motif. The predictions for the normal SOD3 protein are shown in Table 6-1. As can be seen, the threonine at position 58 (bold) is predicted to be part of an N-myristoylation site. The alanine variant does not alter this functional motif.

Position	Sequence	Motif
107-110	NSSS	N-glycosylation site
232-235	PRES	cAMP- and cGMP-dependent protein kinase phosphorylation site
109-111 226-228	SSR SER	Protein kinase C phosphorylation site
20-23	TGED	Casein kinase II phosphorylation site
14-19 **57-62** 124-129 162-167 190-195 191-196	GASDAW **GTLHAA** GCESTG GLAASL GGNQAS GNQASV	N-myristoylation site
201-204	AGRR	Amidation site
112-122	AIHVHQFGDLS	Copper/zinc superoxide dismutase signature 1
199-210	GNAGRRLACCVV	Copper/zinc superoxide dismutase signature 2

Table 6-1: Results of a PROSITE database scan (www.expasy.org/prosite) for functional motifs in the Superoxide Dismutase 3 (SOD3) protein. The position refers to the amino acid position in the SOD3 protein. The sequence shows the amino acid sequence that identifies the functional motif.

The phosphorylation of a residue in a protein is determined by its local environment. The NetPhos version 2.0 server (www.cbs.dtu.dk/services/NetPhos/) uses a neural network method to predict the phosphorylation of individual serine, threonine and tyrosine residues. The results (Table 6-2) show that the threonine at position 58 (bold) is not predicted to be phosphorylated since the score of 0.240 does not reach the threshold of 0.50.

Therefore, a functional variant at amino acid position 58 of the SOD3 protein that replaces a threonine with an alanine is predicted to alter the secondary structure of the amino acids immediately adjacent to it and it is predicted to be hidden from the protein surface but leaves the phosphorylation status unchanged, does not create or abolish a known predicted functional motif, and is a common variant in a population of 46 individuals from four families. Therefore I conclude that it is not likely to alter the function of the protein significantly. Functional *in vitro* studies would be required to test this.

Position	Context	Score	Prediction
4	SXDTMLAL	0.081	-
24	SDAWTGEDS	0.146	-
47	TAKVTEIWQ	0.043	-
62	**DDDGTLHAA**	**0.240**	**-**
74	QPSATLDAA	0.187	-
83	QPRVTGVVL	0.731	*T*
108	EGFPTEPNS	0.076	-
132	GCESTGPHY	0.042	-

Table 6-2: The prediction of the phosphorylation status of the Superoxide Dismutase 3 (SOD3) protein. The phosphorylation of Individual threonine residues are predicted using the NetPhos 2.0 server. The position refers to the amino acid position in the SOD3 protein. The context shows the nine residue sequence centred on the threonine residue being analysed. The score ranges from 0.0-1.0. Residues for which the value exceeds the threshold value of 0.50 are predicted to be phosphorylated.

6.3.2. Allele Sharing

As discussed previously, the genotypes for eight SNPs were obtained by sequencing individuals from the AS DNA panel which consists of members from the four linked families. Genotyping family members allows for the unambiguous identification of haplotypes, including that of the disease chromosome, by following the inheritance of alleles from parent to offspring. Figure 6-4 and Table 6-3 detail the results obtained.

The AS panel defines 45 independent chromosomes. I identified 32 chromosomes for which the data was complete and unambiguous for the entire eight SNP haplotype. Eleven different haplotypes were observed in total. One haplotype is by far the most common, accounting for 18, or 56%, of the 32 chromosomes. The remaining ten haplotypes did not account for more than one or two chromosomes each (3 or 6% respectively). Therefore ten of the eleven haplotypes could be considered rare.

The disease associated haplotypes for families 22, 50 and 59 were the most common haplotype. Definition of the disease associated haplotype in F48 is incomplete, as it was not possible to unambiguously determine the segregation of SNP ih167. Therefore, there are two possible eight SNP haplotypes in F48. One of the possible haplotypes is observed only once out of the 32 chromosomes, the other possible haplotype would be a novel twelth haplotype. Therefore, the disease assoicated hapotypes for three of the four families are identical, but different to the disease associated haplotype of F48. It is possible that the common disease haplotypes of the Celtic families are due to a more recent common ancestor. The fact that the shared disease haplotype is the most common haplotype observed in this study does not support this hypothesis. However, this may be biased by the predominantly Scottish chromosomes on the AS panel.

A Fishers exact statistical test is a non-parametric test that is similar to Chi-square but is suitable for counts less than five. A Fishers exact test was performed using the Analyse-it® (version 1.71) software for microsoft excel (available from

www.analyze-it.com/). Each SNP was analysed individually for allele sharing. A two-tailed Fishers exact test on each of the eight SNPs shows that there is no significant difference between the allele frequency of the four disease and the non-disease chromosomes (Table 6-4). Since the disease haplotype of family 59 does not extend into MR2, the test was also performed without the F59 data. However, there was still no difference in allele frequency between the disease and non-disease chromosomes. In conclusion, there is no obvious pattern of excess allele sharing between the families that would suggest a common susceptibility variant to psychiatric illness.

Figure 6-4 (continued overleaf): Genotyping results from the Allele Sharing (AS) panel, a DNA panel compised of members from families 22, 48, 50, 59. Eight single nucleotide polymorphisms (SNPs) were identified by sequencing exons of the Superoxide Dismutase gene on the AS panel. SNPs are labelled down the lefthand side of the figure. The family ID and chromosome number (relatedness between family members means that 45 different chromosomes are represented on the AS panel) are shown along the top of the figure (d = haplotype inherited with the disease). A representative of each chromosome is highlighted (blue and purple). Ambiguous results, where allele segregation cannot be determined, are boxed. \ = missing data. 32 chromosomes had complete eight SNP haplotypes (purple).

Family	F50														F59									
Member ID	1		2		3		4		7		12		15		2078		2079		2080		3688		3784	
Chromosome	d	1	2	3	D	2	4	5	1	3	d	5	2	5	d	1	2	3	d	2	d	5	4	5
1 ih169	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	A	G
2 ih168	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	A	G	G	A	\	\	\	\
3 ih170	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
4 ih167	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	A	C
5 rs2536512	A	A	G	A	A	G	\	G	A	A	A	G	G	G	A	G	A	A	A	A	A	A	G	A
6 ih166	C	C	T	C	C	T	\	C	C	C	C	C	T	C	C	T	C	C	C	C	C	C	C	C
7 rs2695232	T	T	T	C	T	T	T	C	T	C	T	C	T	C	T	C	T	T	T	T	T	T	C	T
8 rs2855262	C	C	C	T	C	C	C	T	C	T	C	T	C	T	C	T	C	C	C	C	C	C	T	C

Family	F22																							
Member ID	16		47		48		49		111		17		52		53		112		19		54		55	
Chromosome	d	1	D	3	4	5	1	2	2	4	6	7	8	9	d	7	d	8	10	11	12	13	d	10
1 ih169	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	A	G	G	G	G	A
2 ih168	G	G	G	G	G	G	\	\	A	G	G	A	G	G	G	A	G	A	A	G	G	G	G	A
3 ih170	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
4 ih167	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
5 rs2536512	A	A	A	A	G	A	A	A	\	\	G	G	G	A	A	G	A	G	G	A	G	A	A	G
6 ih166	C	C	C	C	C	C	C	C	C	C	T	T	C	C	C	T	C	C	C	C	C	C	C	C
7 rs2695232	T	T	T	T	T	T	T	T	T	T	C	C	C	T	T	C	T	C	C	T	T	T	T	C
8 rs2855262	C	C	C	C	C	C	C	C	C	C	T	T	T	C	C	T	C	T	T	C	C	C	C	T

Family	F22																							
Member ID	114		63		22		62		23		24		68		70		87		88		89		90	
Chromosome	d	12	1	15	14	15	d	14	d	1	16	17	D	16	d	17	d	18	19	20	18	21	22	23
1 ih169	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	\	\	G	G	G	G	G	G
2 ih168	G	G	G	G	\	G	G	\	G	G	A	G	G	A	G	G	\	\	A	G	G	G	G	G
3 ih170	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	\	\	T	C	C	C	C	C
4 ih167	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	\	\	C	C	C	C	C	C
5 rs2536512	A	G	A	A	A	A	A	A	\	\	A	A	A	A	A	A	\	\	G	G	A	A	A	A
6 ih166	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	\	\	T	T	C	C	C	C
7 rs2695232	T	T	T	T	T	T	T	T	\	\	T	T	T	T	T	T	\	\	C	C	T	T	T	T
8 rs2855262	C	C	C	C	C	C	C	C	\	\	C	C	C	C	C	C	\	\	T	T	C	C	C	C

Family	F22														F48					
Member ID	91		92		132		134		136		93		95		823		5		836	
Chromosome	25	26	d	24	d	19	21	23	24	25	D	28	d	27	d	2	d	3	d	4
1 ih169	G	G	G	G	G	G	G	G	G	G	G	A	G	A	A	A	A	G	A	G
2 ih168	G	G	G	G	G	A	G	G	G	G	G	G	G	G	G	G	G	G	G	G
3 ih170	\	\	C	C	C	T	C	C	C	C	C	C	C	C	C	C	C	C	C	C
4 ih167	C	C	C	C	C	C	C	C	C	C	C	\	C	A	A	C	A	C	A	C
5 rs2536512	A	G	A	A	A	G	A	A	A	A	A	\	A	G	G	G	G	G	G	A
6 ih166	C	T	C	C	C	T	C	C	C	C	C	C	C	C	C	C	C	T	C	C
7 rs2695232	T	C	T	T	T	C	T	T	T	T	T	\	T	C	C	C	C	T	C	T
8 rs2855262	C	T	C	C	C	T	C	C	C	C	C	\	C	T	T	T	T	T	T	T

	SNP	1	2	3	4	5	6	7	8	Frequency
	Distance	-	231	167	657	4744	39	595	27	no. (%)
haplotype	1	g	g	c	c	a	c	t	c	18 (56)
	2	g	a	t	c	g	t	c	t	2 (6)
	3	g	g	c	c	g	c	c	t	2 (6)
	4	g	g	c	c	g	t	c	t	2 (6)
	5	g	a	c	c	a	c	t	c	2 (6)
	6	a	g	c	a	g	c	c	t	1 (3)
	7	a	a	c	c	g	c	c	t	1 (3)
	8	g	g	c	c	g	t	t	c	1 (3)
	9	g	g	c	c	g	c	t	c	1 (3)
	10	g	g	c	c	a	c	c	t	1 (3)
	11	g	a	c	c	g	t	c	t	1 (3)

Table 6-3: Haplotypes identified from the Allele Sharing (AS) panel, a DNA panel compised of members from families 22, 48, 50, 59. Eight single nucleotide polymorphisms (SNPs) (labelled 1-8 in the table) were identified by sequencing exons of the Superoxide Dismutase (SOD3) gene on the AS panel. Relatedness between family members means that 45 different chromosomes are represented on the AS panel. Thirty-two of the 45 chromosomes had complete data for all eight SNPs. Eleven eight-SNP haplotypes in SOD3 are observed from these 32 chromosomes. Haplotype 1 predominates over the others, accounting for 56% of the variation. SNP number corresponds to the number in Figure 6-4. Distance is the distance in base pairs between it and the preceding SNP. Frequency refers to the number (and percentage) of the 32 chromosomes upon which the observations are based.

SNP	Chromosome	Allele 1 inc. F59	Allele 2 inc. F59	p-value inc. F59	p-value exc. F59
1	D	(G)3	(A)1	1.0	0.83
	ND	35	6		
2	D	(G)4	(A)0	1.0	1.0
	ND	32	5		
3	D	(C)4	(T)0	1.0	1.0
	ND	39	1		
4	D	(C)3	(A)0	1.0	1.0
	ND	35	2		
5	D	(A)3	(G)1	0.9766	1.0
	ND	22	15		
6	D	(C)4	(T)0	0.8678	1.0
	ND	32	8		
7	D	(T)3	(C)1	1.0	1.0
	ND	26	14		
8	D	(C)3	(T)1	1.0	1.0
	ND	25	15		

Table 6-4: The results of a Fishers exact test on each of the eight single nucleotide polymorphisms (SNPs) in the Superoxide Dismutase 3 (SOD3) gene (labelled 1-8 in the table). SNPs were identifed by sequencing SOD3 exons on the Allele Sharing (AS) panel, a DNA panel compised of members from families 22, 48, 50, 59. Relatedness between family members means that 45 different chromosomes are represented on the AS panel. The number of chromosomes successfully genotyped for each allele is included in the table (with the corresponding allele in brakets). A two-tailed hypothesis for a difference in allele frequency between 'diease' and 'non-disease' alleles has been used. D = allele inherited with the disease. ND = allele not inherited with the disease. The results have been calculated including and excluding the data from Family 59.

6.4. Association Study on Pooled DNA

As discussed in the Introduction, association analysis can provide a suitable way to narrow down a large linkage region. However, association studies typically require large samples numbers, especially when studying heterogeneous disorders with small effect sizes. DNA pooling provides one way to reduce the cost and time of genotyping in large scale association studies, and has been found to be a reliable and quantifiable method for genotyping (Daniels *et al*, 1998; Breen *et al*, 1999; Sham *et al*, 2002). Here I describe an association study on three SNPs in SOD3 using the DNA pooling technique.

6.4.1. Sample

The DNA was pooled by S.Le Hellard (see: Le Hellard *et al*, 2002, for details on the methodology). Four DNA pools were constructed according to diagnosis: 383 control (CTL) individuals, 271 bipolar affective disorder (BPAD) patients, 192 schizophrenic (SCZ) patients and 87 recurrent major depressive (RMD) patients. A Pools plate was constructed (Figure 2-5) which contained four replicates of each of the four DNA pools and 16 individual control DNAs.

6.4.2. SNPs

Three SNPs, rs2536512, rs2695232 and rs2855262, that were identified in the individuals on the AS panel (Section 6.2.3), were used in the study. At the time of this study, the remaining five SNPs had not yet been identified from the AS panel.

6.4.3. Genotyping

Previous studies have identified the SNapShot^{TM} method of genotyping as suitable for estimating allele frequency from DNA pools (Le Hellard *et al,* 2002). The

SNapShot™ reaction was run on the ABI PRISM® 3100 Genetic Analyser and analysed using the GeneScan version 3.0 software.

6.4.4. Analysis

The genotype from a SNapShot™ reaction, when analysed on the ABI PRISM® platform, is read as a peak of fluorescence. The allele frequency is determined by measuring the peak heights of each allele and translating this into a frequency (Figure 6-5). However, when the SNaPshot™ genotyping reaction is performed on an individual, unequal allele amplification is observed. This pattern is reproducible in individuals, but not between individuals, and therefore has to be accounted for in the DNA pool. Therefore, a set of 16 individual control DNAs are run with each pool to ensure that approximately five heterozygotes are obtained for each SNP. The *K* ratio measures the ratio between the allele peak heights of an individual heterozygote. This is then subtracted from the peak heights of the DNA pools before further analysis. Approximately five heterozygotes are needed to calculate an average measure of the between-individual variation in the *K* ratio. Too much variation makes the assay unreliable, and therefore, a standard error of the mean (SEM) of greater than 10% was not tolerated.

Variation in the amplification of the individual DNAs in a pool will also be observed. Therefore, each pool was replicated four times and the average variation between replicate peak height ratios was calculated. A SEM of greater than 0.01 (1% variation around the mean) was deemed unreliable. A SEM calculation that exceeds the required cut off can be reduced by increasing the number of replicates of either the heterozygotes or the pools.

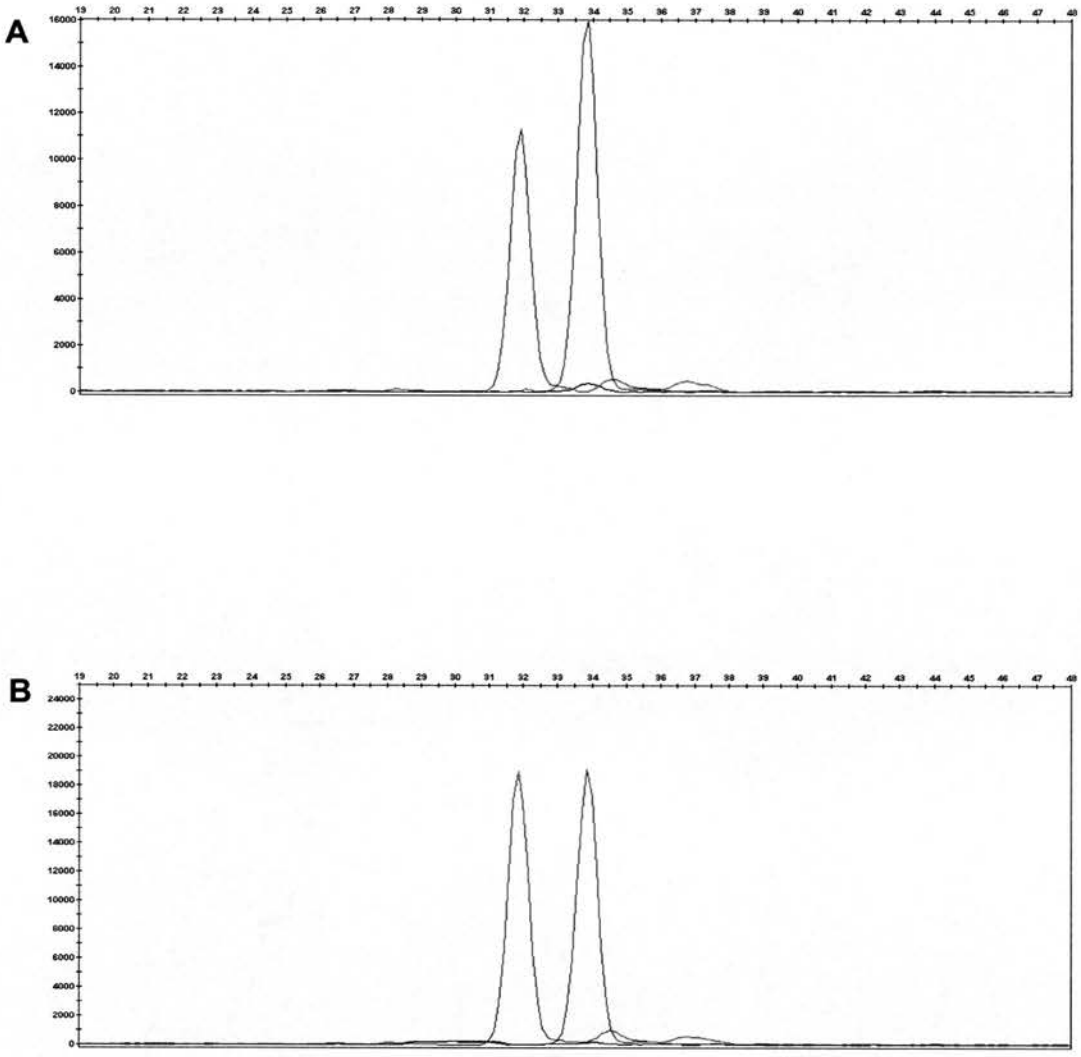


Figure 6-5: Example of the fluorograms for two DNA pools for the single nucleotide polymorphism (SNP) rs2536512. Genotyping was performed by the SNaPshot™ method, run on an ABI PRISM® 3730 genetic analyser and signal intensity was visualised as a fluorogram using the GeneMapper (version 3.0) software. The coloured peaks represent each allele of the SNP. Allele frequency in a DNA pool is assessed by comparing the signal intensity (peak height) of each allele. **A:** Control DNA pool. **B:** Recurrent major depression (RMD) DNA pool. In this example, the blue allele is over-represented in the RMD DNA pool compared to controls.

Significance was measured by the non-parametric Chi-square statistic which measures the difference between expected and observed frequencies. In a test for association, the expected frequency for each allele is the frequency observed in the control group. The main problem of using the Chi-square statistic for pooled DNA is that one of the requirements of the test is that real frequencies are used. To meet these requirements, the peak heights are first translated into real allele frequencies before inputting the figures into the equation. In addition, an adjustment factor was included into the p-value to account for the error involved in this translation. The *K* ratio, the calculation of allele frequency from peak heights and the p-value correction factor were developed by Peter Visscher (published in Le Hellard *et al*, 2002).

6.4.5. Results

The results are detailed in Table 6-5 (see Appendix II for the raw data). The difference in allele frequency of SNP rs2536512 between the RMD patient group and the CTL group approaches significance. However, the individual heterozygote variation (*K* ratio) and the RMD pool SEM are a little high. The differences in allele frequency of SNP rs2695232 between the control and patient groups was not significant. This SNP also displayed too much variation between pool replicates in each of the four pools. Subsequent efforts to increase the number of replicates, in order to reduce the SEM, failed.

The difference in allele frequency between the control and the RMD DNA pools for SNP rs2855262 reveals a p-value of <0.003. This is a highly significant difference. The allele frequency of the major allele in the control group is 65% compared to 37% in the RMD patient group (Appendix II). Therefore, the minor allele is overrepresented in the unipolar affective disorder group. However, it is impossible from this to predict whether it might be the under representation of the major allele or the over representation of the minor allele that is responsible for the association and the predicted change in protein function. Unfortunately, this could be a false positive since the unipolar group is the smallest of the four pools, containing only 87

individuals, and, due to a genotyping failure, the result was based on only two pool replicates (Appendix II).

6.4.6. Problems

Subsequent analysis of DNA profiles revealed that the pools contained a number of historical duplicate samples, meaning that the results were not reliable. This inflexibility is an obvious disadvantage of pooling DNA. The effort required to construct accurate pools and a decrease in genotyping cost led to the decision to carry out association studies on individual samples.

		SNP		
		rs2536512	**rs2695232**	**Rs2855262**
SEM	**Het**	0.12	0.01	0.02
	CTL	0.004	0.02	0.002
	BPAD	0.005	0.02	0.005
	SCZ	0.007	0.03	0.01
	RMD	0.015	0.07	0.002
p-value	**BPAD**	0.37	0.82	0.63
	SCZ	0.83	0.4	0.42
	RMD	**0.08**	0.65	**0.003**

Table 6-5: The results of a Chi-Square test for association on pooled DNA. Three single nucleotide polymorphisms (SNPs) in the Superoxide Dismutase 3 gene were tested. The allele frequency of each SNP was measured in DNA pools of control individuals (CTL), bipolar affective disorder patients (BPAD), schizophrenic patients (SCZ) and recurrent major depressive patients (RMD). Allele frequency comparisons between the CTL group and each of the case groups are given as a p-value. Each DNA pool, and five individual heterozygotes (Het) are genotyped a number of times and the standard error of the mean (SEM) measures the variation between these genotyping replications.

6.5. Phase I Association Study on Individuals

The association study was to be carried out in two phases. The Phase I sample consisted of a small number of individuals and was used to identify the linkage disequilibrium (LD) between SNPs and to identify SNPs with a trend towards a positive association to carry through to a Phase II study on a larger group.

6.5.1. Sample

The Phase I sample consisted of 95 controls, 93 BPAD patients and 95 schizophrenic patients.

6.5.2. SNPs

Seven of the eight SNPs that had been identified by sequencing the SOD3 gene in individuals from the AS DNA panel were tested in Phase I. SNP ih167 was not included in the study because the sequence quality was poor and I was not absolutely sure that the SNP was real. The seven SNPs were clustered in exon one and exon three and therefore a further six SNPs were included to ensure a more even coverage of the gene to hopefully cover all LD blocks. One SNP was chosen from Applied Biosystems as an 'Assay on Demand' TaqMan® assay. The Caucasian minor allele frequency was 0.41. Five SNPs from dbSNP were also included. These were chosen on the basis that there had been more than one independent submission to dbSNP. The 13 SNPs included in Phase I gives a coverage of one SNP every ~1.75kb. This aimed to account for potential assay failures and the unknown nature of the LD block structure across the region.

6.5.3. Genotyping and Analysis

Twelve SNPs were genotyped at the Sanger Institute by the Sequenom® MASSARRAY™ primer extension method. The results of this analysis revealed a

number of gaps due to assay failure. Therefore, I re-genotyped rs2536512 and rs2324580, by SNaPshot™ and sequencing respectively, in order to provide a better coverage of the gene. The SNP that was provided as an 'Assay on Demand' by Applied Biosystems was genotyped by the TaqMan® method by Alison Condie.

Association was measured using Chi-square and LD was measured using D' and r^2. Association, LD and Hardy-Weinberg equilibrium were calculated using the UNPHASED software, version 2.403, developed by Frank Dudbridge (www.hgmp.mrc.ac.uk/~fdudbrid/software/unphased/) (Dudbridge, 2003), by Naomi Wray.

6.5.4. Results

Table 6-6 details the results of the association study for each SNP. Eleven out of twelve SNPs sent to the Sanger Institute for genotyping by the Sequenom® MASSARRAY™ method failed. The reasons for failure were as follows. The assay failed for four of the SNPs, the negative control was contaminated for one SNP, and three SNPs were not polymorphic. There are possible explanations for three of the SNPs not being polymorphic in the DNA panels used for the association study. SNP ih170 was observed only once on the AS panel and was therefore very rare, SNP ih168 was identified from bad quality sequence and therefore may not be real, and SNP rs800448 selected from dbSNP and therefore was not known to be polymorphic in this population. Therefore, eight SNPs were discarded at this stage. Out of the remaining four SNPs, all four were found to be in Hardy Weinberg equilibrium. However, an insufficient number of chromosomes were successfully genotyped for three of the SNPs. This left one SNP for which the assay worked on a sufficient number of chromosomes to test for association. In order to attain better coverage of the gene and to calculate LD, I retyped rs2324580 and rs2536512 inhouse. Together with SNP C_2668721_10 from Applied Biosystems, this produced four SNPs with sufficiently high quality data for further analysis.

In a cost-effective approach, the case-control study was undertaken in two phases. In Phase I, all SNPs were genotyped on 95 cases and 95 controls. All SNPs with association p-values of <0.2 were to be chosen for the Phase II association study, to be performed on a larger sample. This threshold was selected based on power calculations carried out by Dr N Wray (using an in house programme developed by N. Wray based on standard formulas, e.g. Schaid and Rowland, 1998). Under a multiplicative genotype relative risk (RR) model, a sample size of 95 cases and 95 controls, using a liberal type 1 error of 0.20, has over 80% power to detect an association with a heterozygote genotype RR of 2, even when the frequency of the associated allele is low at 0.10 and over 90% power to detect the same RR when the allele frequency is 0.40-0.60. Therefore, despite the small sample size in Phase I, it is very unlikely that any SNP that is truly associated will be eliminated from the list of SNPs tested in Phase II. Moreover, the power for the Phase II sample was similar to the power of the Phase I sample when different corrections for multiple testing based on the full set *vs* a selected subset of SNPs was taken into account. In this way, a high cost-effectiveness is achieved with little loss in overall power.

SNP haplotypes were not analysed since the case-control data precluded definitive haplotype construction. Instead, each SNP was tested independently for association. A bonferroni correction for multiple testing, to account for the number of SNPs tested, was not applied to the p-value cut-off at which significance was declared because multiple tests within the one gene are not considered to represent independent tests if LD is maintained between them (Nyholt, 2001). A correction for multiple testing would also be appropriate if the sample group was to be split, for example, by gender, for further comparison. However, this was not done here.

The results of the association analysis showed that none of the SNPs showed a significant difference in allele frequency between the control group and either the schizophrenic or the BPAD groups or both case groups combined.

Linkage disequilibrium is a measure of how likely two loci (in this case SNPs) will be separated by a recombination event. Therefore, the Phase I set of individuals was to provide a measure of LD as well as the results of association. Rather than exhibiting a gradual decay over increasing genomic distance, LD has been suggested to occur in a block-like structure with clustered recombination 'hotspots' (Daly *et al*, 2001; Jeffreys and May, 2004). The extent to which these blocks exist and whether recombination hotspots are the mechanism by which they occur is still controversial (Oord and Neale, 2003). However, an association study ought to be able to pick a minimal set of SNPs to represent all or most of the haplotype diversity and therefore provide an even coverage of the gene with respect to LD. LD can be measured in a number of different ways, of which D' and r^2 are two of the most frequently used examples. D' is a widely accepted method, however, r^2, unlike D', is more sensitive to allele frequency. For example, with D', high LD can be observed between two SNPs if one has a rare minor allele if the rare allele is only ever observed in combination with one allele of the second SNP. The r^2 measure is a pairwise correlation and therefore is not so influenced by allele frequency.

The results of the LD calculations show high LD between SNP 3 and 7, suggesting that it would not be necessary to type further SNPs in between these. The D' and r^2 values are very similar, reflecting the fact that the allele frequency of both SNPs is the same. However, there is quite low LD, by both D' and r^2, between SNP 7 and 10, and SNP 10 and 13. Therefore, these three SNPs do not adequately cover the LD landscape across this region. This level of LD would suggest that, in contrast to the premise used to design the study, the tests for association performed with each SNP are independent and, therefore, criterion for declaring significance should be adjusted accordingly. However, since no association was found, further correction for multiple testing would be unnecessary. In addition, the lack of LD between the SNPs means that the lack of association to psychiatric illness observed in the region of SOD3 covered by SNPs 7, 10 and 13 does not necessarily apply to the intervening regions. Therefore, more SNPs need to be tested.

SNP name	Distance (bp)	MAF ch (%)	Method	Assay	HW	No ch's [ctl]	MAF [ctl]	BPAD p- value	SCZ p-value	SCZ+BP p-value	D'	r^2
Rs800399	-	-	S	Failed								
Rs800414	3644	-	S	Failed twice								
C_2668721_10	9948	- [0.41]	T		0.69	190	31	0.754	0.978	0.869	-	-
Ih169	1908	5 [11.4]	S		0.73	44	7					
Ih168	231	4 [9]	S	Monomorphic								
Ih170	167	1 [2.3]	S	Monomorphic								
Rs2536512	5404	17 [39]	Sn		0.34	186	31	0.668	0.688	0.631	0.78	0.75
Rs8192291	39	8 [18]	S		0.69	16	13					
Rs2695232	595	15 [34]	S	Contaminated								
Rs2855262	27	15 [34]	S		0.73	172	34	0.32	0.79	0.70	0.50	0.34
Rs800447	229	-	S	Failed								
Rs800448	118	-	S	Monomorphic								
Rs2324580	5668	-	Se		0.30	190	10	0.689	0.716	0.661	0.09	0.07

Table 6-6: The Phase I association study results. Thirteen single nucleotide polymorphisms (SNPs) in the Superoxide Dismutase 3 gene were genotyped in 95 controls (CTL), 95 schizophrenic (SCZ) and 93 bipolar affective disorder (BPAD) patients. The table details SNP name, the distance between a SNP and the preceding SNP, the minor allele frequency (MAF) determined from genotyping the Allele Sharing DNA panel or supplied by Applied Biosystems, the genotyping method (S = Sequenom®, T = TaqMan®, Sn = SNaPshot™, Se = sequencing), the reason for assay failure, the Hardy-Weinberg (HW) calculation expressed as a p-value ($p<0.05$ not tolerated), the number of control chromosomes genotyped successfully, the MAF in the CTL individuals, the Chi-square p-value for BPAD, SCZ and both combined and the linkage disequilibrium (LD) calculations (D' and r^2) between it and the preceding SNP (red = high LD, pink = moderate LD, white = low LD; SNPs highlighted blue were not included in the LD analysis).

6.6. Phase II Association Study on Individuals

As discussed in the previous section, the Phase I association study was performed on a restricted set of individuals and a high p-value threshold of <0.2 was set. This aimed to capture all possible true association, but would also include a number of false positive results.

6.6.1. Sample

The Phase II sample consisted of 388 controls, 331 BPAD patients and 254 schizophrenic patients. The individuals in this group did not overlap with the individuals in Phase I. Therefore, this represented a replication study with an independent sample, but also enabled the results from both groups to be analysed together in one larger group.

6.6.2. SNPs and Results

None of the SNPs in SOD3 passed the threshold p-value of <0.2. Therefore none of the SNPs were taken forward to Phase II.

6.7. Discussion

Here I describe the genetic analysis of the SOD3 gene in MR2. MR2 is a candidate region for the susceptibility to psychiatric illness in three families that show linkage to the region, and SOD3 is a good functional candidate gene within this region. It is expressed in the brain, is believed to play a role in protection against damage induced by oxygen toxicity and may be involved in learning and memory.

I identified eight SNPs by sequencing the coding and regulatory regions of the SOD3 gene in members of the four linked families. One SNP was found to alter amino acid 58 of the protein from a threonine to an alanine. Bioinformatic and frequency analysis showed that this substitution is unlikely to significantly alter the function of the protein.

The haplotypes of the four families were ascertained and assessed. The eight SNP haplotype that occurs on the disease chromosome of families 22, 59 and 50 is the same. This is the most common haplotype in the population; it is observed on 56% of 32 chromosomes. Ten other haplotypes were observed, each accounting for between 3% and 6% of the total. The eight-SNP haplotype on the disease chromosome of family 48 was rare in the individuals sampled. However, due to the uninformative nature of one of the SNPs it was not possible to determine whether it was unique or was the same as one of the ten rare haplotypes observed. The haplotype sharing of families 22, 50 and 59 might reflect the common ancestry of the three Celtic families. In addition, the frequency of the common haplotype might be an artefact of the predominantly Celtic chromosomes on the AS panel: 41 out of the 45 chromosomes on the panel are from families 22, 59 or 50. Chromosomes from other ethnic groups would need to be genotyped to test this.

There are certain limitations to analysing haplotype sharing using the AS panel. There are only four disease chromosomes to compare and the control chromosomes are family members which could be biased by assortative mating. However, having

closely related control chromosomes is advantageous, and studying families enables the unambiguous determination of haplotypes.

Three of the SNPs that had been identified in the families were tested for association in a set of pooled DNA samples. The results revealed a significant positive association between SNP rs2855262 and the RMD patient group compared to controls. However, the results were deemed unreliable because the pools contained replicate DNAs. Therefore, comparison of pooled and individual association results is not possible. The inflexibility of DNA pooling is an obvious disadvantage. Once an individual is classified into a particular group he/she cannot be removed and therefore precludes the correction of sample errors. In addition, psychiatric illnesses are clinically and genetically heterogeneous, and diagnoses can change over a persons lifetime. Therefore, samples that have not been pooled can be more appropriately analysed at a later stage.

Seven of the SNPs that had been identified from the four families, and an additional six SNPs identified from the public databases, were tested for association in a population of unrelated individuals. This preliminary association study consisted of 95 controls, 95 schizophrenic patients and 93 BPAD patients. This preliminary screen intended to identify SNPs which showed a trend towards a significant positive association before testing these on a larger sample. None of the SNPs showed a significant positive association in this preliminary screen and so none were tested on the larger population.

There are a number of advantages to using association studies in this way. It provides a good way to narrow down a large linkage region. In addition, using this data set it was possible to create an LD map of the gene. This allows the assessment of the coverage obtained in Phase I of the study and ensures that there is no unnecessary duplication in Phase II. Due to genotyping failures, only four SNPs were included in the LD calculations. This revealed that whilst SNP 3 and 7 appear to be in high LD, the remaining SNPs do not. Therefore, future studies are required to analyse a higher

density of SNPs across SOD3 in order to be able to rule out association with psychiatric illness completely.

Caution is required when interpreting association studies. Psychiatric illness has a heterogeneous genetic cause. In a family, the genetic cause is likely to be the same for all individuals. However, studying unrelated individuals significantly reduces this likelihood and more than one genetic locus will be responsible for a clinical diagnosis. Consequently the sample size required to detect the loci will be increased. The effect size of the loci that contributes to susceptibility to psychiatric illness in the four families is unknown, so the required association study sample size is unknown. A reasonable guess would be the kinds of effect sizes shown in other studies. For example, the results of association and haplotype analysis in the NRG1 gene and schizophrenia have observed effect sizes in the region of 2 (Stefansson *et al*, 2002). Power calculations by Dr. Naomi Wray suggest that an association with this effect size would be detectable in the Phase I population if a p-value of <0.2 is considered.

In summary, the results of genetic analysis of SOD3 in the four families did not reveal a mutation in SOD3 that would alter protein function and contribute to psychiatric illness. Association studies in pooled DNA revealed a significant association for one SNP in RMD patients. However, the sample proved unreliable. Association studies on individuals did not reveal a significant association to schizophrenia, BPAD or RMD. LD across the gene has not been covered and therefore further SNPs should be tested in SOD3

Chapter Seven

Genetic Analysis of the G-Protein-Coupled Receptor 78 (GPR78) Gene

Genetic Analysis of the G-Protein-Coupled Receptor 78 (GPR78) Gene

7.1. Introduction

Here I describe sequence analysis and association studies of the orphan G-protein-coupled receptor 78 (GPR78) gene. Previously I have described how the comparison of linkage regions in four families has prioritised two regions: MR1 and MR2. MR1 currently (February 2004) extends over ~3.8Mb of sequence and contains seven known genes. GPR78 is positioned in MR1, and it is therefore a candidate gene for psychiatric illness in F22, F50 and F59.

In the previous chapter, I explained the rationale of studying individual genes within these regions for SNP variation in the families. As in the previous chapter, I have chosen a candidate gene from MR1 to identify the haplotypes that are observed on the family disease chromosomes, identify SNPs that potentially alter the function of the protein and perform association studies of these SNPs in a group of unrelated cases and controls.

GPR78 is a good candidate gene to study within MR1 because it is a member of the G-protein-coupled receptor family. G-protein-coupled receptors are involved in cellular signalling pathways and are therefore well positioned to mediate the subtle changes in cellular function that may be involved in psychiatric illness. GPR78 was identified by virtue of its homology to GPR26, which itself was identified in human and rat by Lee *et al* (2000). A human EST encoding a fragment of the G-protein-coupled receptor GPR26 was used to screen cDNA libraries. This identified the full length rat cDNA of GPR26 and a partial genomic DNA fragment of human GPR26. Rat GPR26 cDNA encoded a 317 amino acid protein distantly related to the serotonin 5-HT5α and gastrin releasing hormone BB2 receptors. Expression of the mRNA was identified in numerous brain regions.

GPR78 was identified by the same group (Lee *et al*, 2001) by using partial GPR26 sequence as a probe to search the EST and high throughput genomic sequences (HTGS) databases. GPR78 mRNA is expressed in the pituitary and the placenta, but not in the brain. This places GPR78 in a good position to play a role in the hypothalamic-pituitary-adrenal (HPA) axis. The HPA axis is involved in hormone and stress regulation and has been found to be dysfunctional in affective disorder patients (Pariante & Miller, 2001), and schizophrenic patients (Altamura *et al*, 1999).

The closest relation to GPR78 is GPR26. Both are orphan receptors with unknown ligands, they share 65% identity in their transmembrane domain (TMD) regions and have a similar protein structure: they lack an asparagine-linked extra cellular glycosylation site, have a short amino terminus, have a similar intron-exon structure, and share cationic arginine and lysine residues in transmembrane domains V and VI respectively (Lee *et al*, 2001). These residues have previously been found to play a role in purinergic binding in P2Y receptors; G-protein-coupled receptors for the nucleotides ADP, ATP, UDP, and UDP-glucose. Therefore GPR78 belongs to subgroup 1-C of the group 1 family of G-protein-coupled purigenic and adenosine receptors that use cAMP as a second messenger.

The GPR78 gene has three exons and gives rise to the classic seven transmembrane receptor structure of G-protein-coupled receptors, with an extracellular N-terminus and an intracellular C terminus (Figure 7-1). Within MR1, the gene is positioned on BAC RP11-301J10 and lies within the recombination breakpoint interval at the telomeric end of MR1. In chapter 4, I described refinement of this recombination interval, but unfortunately I could not definitively rule GPR78 in or out of MR1. However, even if in the future it is ruled out of MR1 it could still be under promoter control from within MR1, and it will still remain on the disease chromosome of F22. Here I describe the results of family and association analyses for this gene.

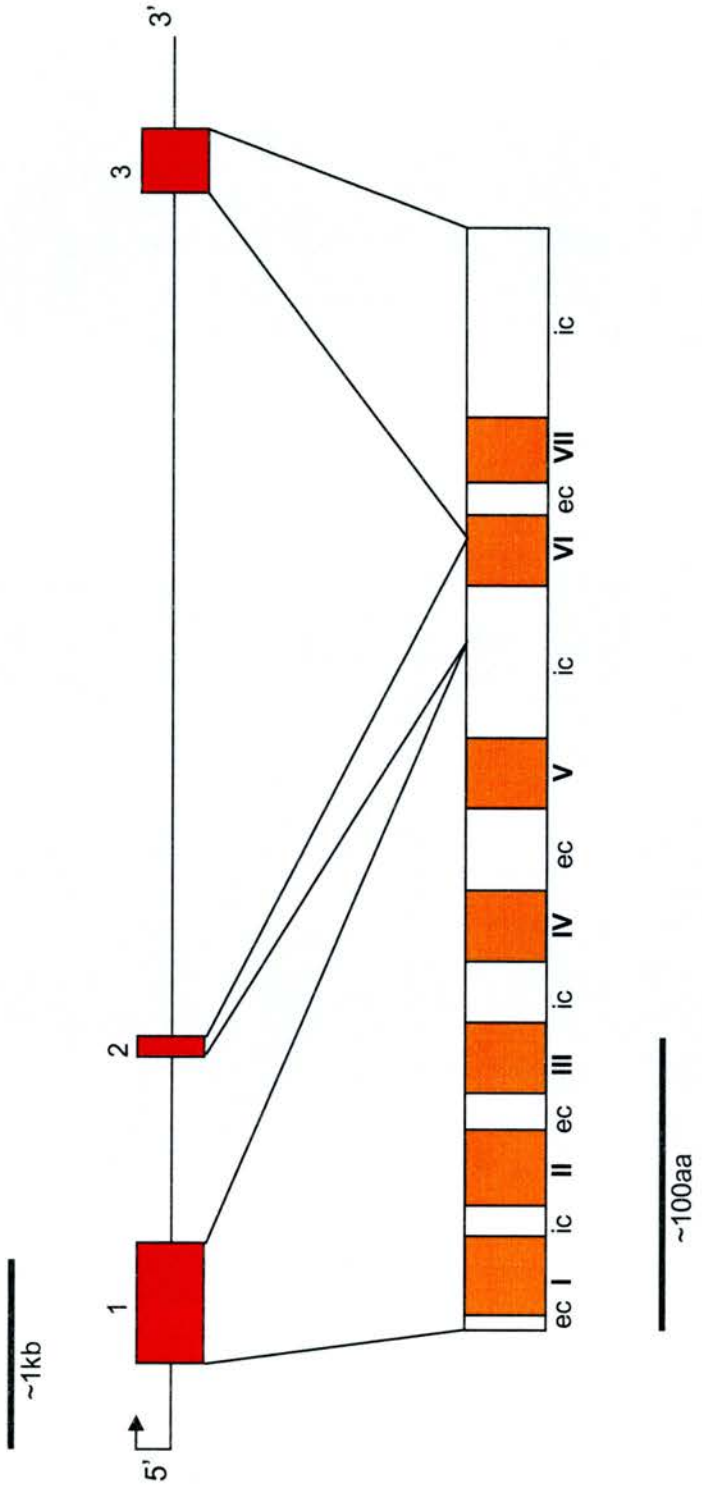


Figure 7-1: The structure of the orphan G-protein-coupled receptor 78 (GPR78) gene and peptide. The gene has three exons (red). The relationship of the exon structure to the protein domains is shown by the lines. The protein has seven transmembrane domains (orange), and the intervening sequence is marked as either intracellular (ic) or extracellular (ec). aa = amino acids. kb = kilobase.

7.2. SNP Identification

7.2.1. Sample

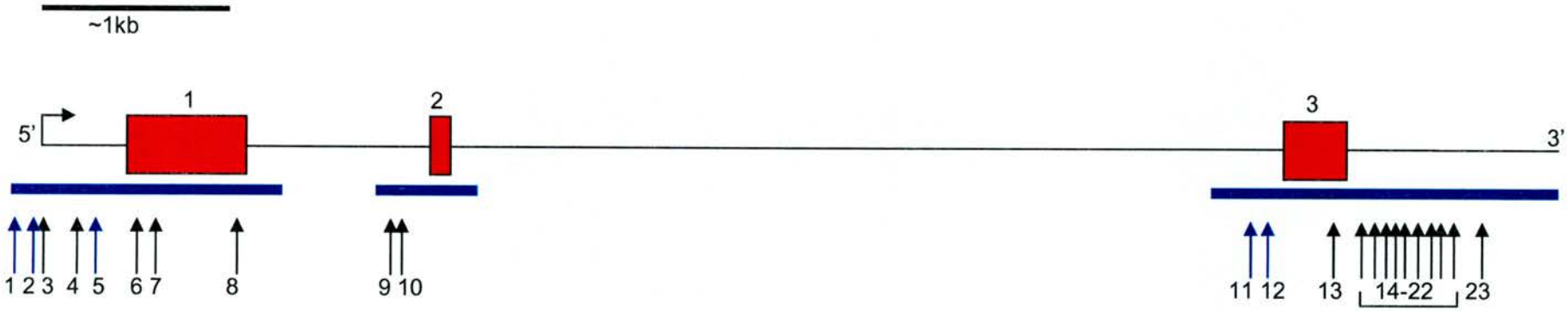
As in Chapter 6, the allele sharing (AS) DNA panel was used to identify SNPs (Figure 2-5).

7.2.2. STS Design and SNP Detection

STSs were designed, PCR's were optimised and sequenced and SNPs were identified as in Section 6.2.2. Every STS was sequenced with both the forward and reverse primer to confirm the genotype obtained.

7.2.3. SNPs identified

Twenty-three SNPs were identifed from the AS DNA panel (Figure 7-2). None of these SNPs were in the public database dbSNP at the time (February 2002). Today (February 2004), five of these SNPs have since been submitted to dbSNP. Two of the four exonic SNPs (ih31 and ih34) change the amino acid sequence of the protein.



1	rs6447884	13	ih34
2	rs4385041	14	ih44
3	ih162	15	ih43
4	ih165	16	ih42
5	rs4235270	17	ih41
6	ih33	18	ih40
7	ih32	19	ih36
8	ih31	20	ih37
9	ih48	21	ih39
10	ih47	22	ih38
11	rs3115388	23	ih35
12	rs3115387		

Figure 7-2: Single nucleotide polymorphism (SNP) identification in the orphan G-protein-coupled receptor 78 (GPR78) gene. The regions of GPR78 that were screened for SNPs on theAllele Sharing (AS) DNA panel by sequencing are shown by the blue bars. SNPs (numbered arrows) identified in only the AS panel are marked with black arrows. SNPs identified in the AS DNA panel that are also in the public SNP database dbSNP are marked with blue arrows. SNP names are provided.

7.3. SNP Analysis

7.3.1. Amino Acid Changes

7.3.1.1. ih31

SNP ih31 in exon one changes amino acid 201 from an arginine to a serine. The SNP is an a>c variant, the third nucleotide of codon 'aga'. The mutation occurs in the intracellular loop between TMD V and VI, just after TMD V. It changes a charged residue to an uncharged residue in a string of charges. It could be hypothesised to play a role in TMD anchoring, as this is thought to be under the control of such a string of charged residues. The frequency of this variant could provide some idea of its functional importance. The SNP was identified by sequencing the 45 independent chromosomes from the AS DNA panel. This SNP was observed on 10 chromosomes, which translates into a frequency of 23%. This is fairly common, and might suggest that the alternative protein would not have a significant impact on protein function. However, as discussed in Chapter 6, the CDCV hypothesis suggests that common variants will underly the susceptibility to common diseases, or assortative mating may have made the SNP more common in the families. Despite this, the variant was only observed on the disease chromosome of F50, and not families 22, 59 or 48.

The properties of each amino acid should give some idea of whether a substitution would be tolerated or not. According to Bordo and Argos (1991) these two amino acids cannot be substituted. Four different scales of hydrophobicity place arginine at or near the bottom and therefore is consistently considered to be more hydrophilic than serine, which is placed near the middle (Janin, 79; Wolfenden *et al*, 81; Kyte and Doolittle, 82; Rose *et al*, 85). Therefore, this could alter the secondary structure of the protein because the more hydrophilic residues will tend to be exposed on the protein surface.

PIX analysis was described in Section 6.3.1. The results of a PIX analysis of the two GPR78 variant proteins (Figure 7-3) shows that the serine variant (B) has extended the region of predicted antigenicity by four residues, and slightly altered the secondary structure of two adjacent amino acids from an alpha helix to a coil structure in the DSC secondary structure prediction programme. However, these are still surrounded by an extensive region of a predicted alpha helical structure. The transmembrane domain predictions, by the programmes TMHMM, TMPRED, TMAP and DAS, are the same in both protein variants. Therefore, structurally, the serine variant does not appear to have a significant impact on the protein.

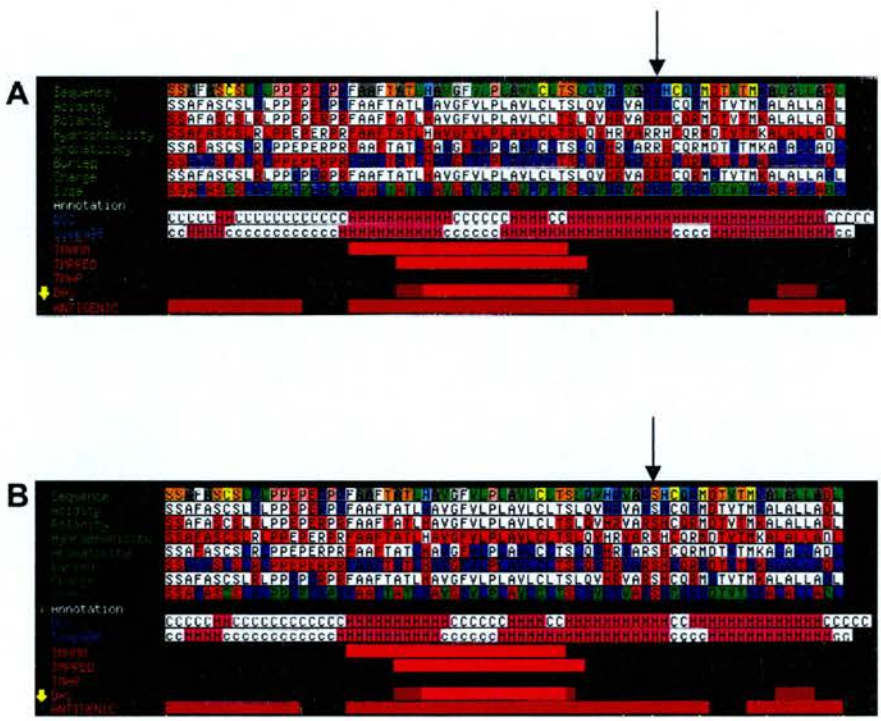


Figure 7-3: The results of a Protein Identification of Unknown Sequences (PIX) analysis (MRC-RFCGR) for part of the orphan G-protein-coupled receptor 78 (GPR78) protein containing single nucleotide polymorphism (SNP) ih31 at amino acid position 201 (arrow). The two amino acid variants created by SNP ih31 are shown (arrow). **A:** Arginine variant. **B:** Serine variant. See Figure 6-3 for an explanation of the analysis programmes.

If the variant amino acid is predicted to alter a functional motif it could be postulated to impact upon protein function. The PROSITE database, described in Section 6.3.1, searches for such functional motifs. The results of a PROSITE scan in the normal GPR78 protein can be seen in Table 7-1. The arginine-serine variant at position 201 falls between two predicted protein kinase C phosphorylation sites, but does not fall directly within any existing motif. A PROSITE search with the variant serine at position 201 does not create any additional motifs.

Position	Sequence	Motif
47-50	NLSL	N-glycosylation site
348-351	NGSV	
355-358	NDSC	
157-159	SLR	Protein kinase C phosphorylation site
210-212	TMK	
223-225	SVR	
240-242	TRK	
316-318	TPR	
340-342	TPR	
321-324	STHD	Casein kinase II phosphorylation site
325-328	SSLD	
353-356	TEND	
286-294	KAVADPFTY	Tyrosine Kinase phosphorylation site
66-71	GVMRGR	N-myristoylation site
124-129	GLLLGC	
128-133	GCAWGQ	
132-137	GQSLAF	
139-144	GAALGC	
244-249	GIAIAT	
349-354	GSVDTE	

Table 7-1: Results of a PROSITE scan for functional motifs in the orphan G-protein-coupled receptor 78 (GPR78) protein. The position refers to the amino acid position in the GPR78 protein. The sequence shows the amino acid sequence that identifies the functional motif.

The NetPhos 2.0 server, decribed in Section 6.3, predicts phosphorylation of individual serine, threonine and tyrosine residues in a protein. Phosphorylation of proteins can be an important part of cellular signalling cascades. The results of the predicted phosphorylation status of serine residues in GPR78 are shown in Table 7-2. As can be seen, the predicted phosphorylation status of the serine in the alternative protein (position 201 in the table) does not exceed the threshold value of 0.5 and is therefore not predicted to be phosphorylated.

Position	Context	Score	Prediction
22	VALLSNALV	0.002	-
33	CCAYSAELR	0.007	-
41	RTRASGVLL	0.398	-
49	LVNLSLGHL	0.008	-
74	GRTPSAPGA	0.875	*S*
91	TFLASNAAL	0.009	-
96	NAALSVAAL	0.571	*S*
101	VAALSADQW	0.173	-
134	AWGQSLAFS	0.003	-
138	SLAFSGAAL	0.039	-
145	ALGCSWLGY	0.007	-
150	WLGYSSAFA	0.007	-
151	LGYSSAFAS	0.011	-
155	SAFASCSLR	0.006	-
157	FASCSLRLP	0.015	-
192	LCLTSLQVH	0.006	-
201	**RVARSHCQR**	**0.059**	**-**
223	DLHPSVRHG	0.620	*S*
279	WGILSKCLT	0.005	-
285	CLTYSKAVA	0.027	-
295	PFTYSLLRR	0.010	-
321	PRPASTHDS	0.998	*S*
325	STHDSSLDV	0.655	*S*
326	THDSSLDVA	0.428	-
345	PRPASTHNG	0.996	*S*
350	THNGSVDTE	0.045	-
357	TENDSCLQQ	0.203	-

Table 7-2: The prediction of the phosphorylation status of the orphan G-protein-coupled receptor 78 (GPR78) protein. The phosphorylation of individual serine residues are predicted using the NetPhos 2.0 server. The position refers to the amino acid position in the GPR78 protein. The context shows the nine residue sequence centred on the serine residue being analysed. The score ranges from 0.0-1.0. Residues for which the value exceeds the threshold value of 0.50 are predicted to be phosphorylated.

Therefore, a functional variant at amino acid position 201 that changes the residue from an arginine to a serine is unlikely to greatly alter the function of the GPR78 protein, if at all. The variant is relatively common on control chromsomes and it does not effect the predicted phosphorylation state of the protein or the protein structure.

7.3.1.2. SNP ih34

SNP ih34 in exon three changes amino acid 342 from an arginine to a histidine. The SNP is variant g>a , the second nucleotide of codon 'cgc'. The variant occurs in the protein's C-terminal, and is the 23rd amino acid from the end. The C-terminal region is situated intracellularly and is therefore in a position to be involved in intracellular signalling cascades. This variant was observed on five chromosomes from the AS panel; a frequency of 11.4%. This minor allele frequency is quite low and therefore this variant is quite rare. It was observed on the disease chromosome of family 22, but not families 59, 50 or 48.

According to Bordo and Argos (1991) these two amino acids cannot be interchanged. Using four different scales of hydrophobicity, arginine is consistently considered to be more hydrophilic than histidine, being positioned at or near the bottom whilst histidine is positioned near the middle (Janin (1979); Wolfenden *et al*, 1981; Kyte and Doolittle, 1982; and Rose *et al*, 1985).

The results of a PIX analysis at the MRC-RFCGR is shown in Figure 7-4. The secondary structure of this region of the protein is not altered by the histidine variant. However, the level of antigenicity is increased. This does not follow with the observation that histidine is more hydrophobic that arginine. However, if histidine is more exposed on the protein surface, this could alter the availability of functional motifs to intracellular signalling molecules and enzymes.

The PROSITE database predictions for functional motifs in the GPR78 protein have already been shown in Table 7-1. Performing the same search with the variant

histidine at amino acid 342 changes the motif 'TPR' to 'TPH' at position 340-342. This would abolish a predicted protein kinase C phosphorylation site. Since the variant histidine occurs in the intracellular C-terminal region it is well positioned for interactions with components of intracellular signalling cascades.

Therefore, the consequence of this variant is unclear. It is possible that it disrupts an intracellular signalling cascade. However, the programmes used are predictions only and *in vitro* functional studies would be required to determine this unequivocally. It is a rare polymorphism which therefore could indicate that it is either relatively new or that it is under negative selection pressure. The CDCV hypothesis would not postulate a rare polymorphism to be involved in a common illness. Furthermore, if it does alter the function of the protein it does not play a role in the disease pathology of families 50, 59 or 48 because the mutation occurs only on the F22 disease chromosome. However, it is an interesting variant, and if a positive association were found between ih34 or other GPR78 markers in LD with ih34, it might justify further *in vitro* work being carried out.

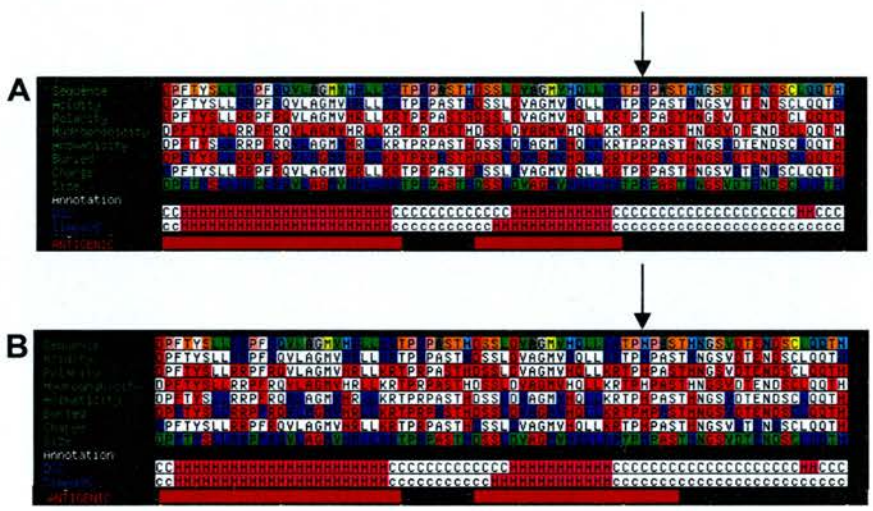


Figure 7-4: The results of a Protein Identification of Unknwon Sequences (PIX) analysis (MRC-RFCGR) for part of the orphan G-protein-coupled receptor 78 (GPR78) protein that contains single nucleotide polymorphism (SNP) ih34. The two amino acid variants created by SNP ih34 are shown (arrow). **A:** Arginine variant. **B:** Histidine variant. See Figure 6-3 for an explanation of the analysis programmes.

7.3.2. Allele Sharing

As mentioned above, the genotypes for 23 SNPs were obtained by sequencing individuals from the AS DNA panel which consists of members from the four linked families. This allows for the unambiguous identification of haplotypes, including that of the disease chromosome, by following the inheritance of alleles from parent to offspring.

Figure 7-5 details the genotypes of the 46 individuals from the AS panel. SNPs ih47 and ih48 were removed from the analysis because the genotyping was unreliable since the results did not segregate. I suspected that this was because the minor allele amplified inefficiently in the sequencing reaction and that therefore heterozygote genotypes were mistaken for homozygote genotypes. Thirty-six unambiguous 21-SNP haplotypes were identified from the 45 chromosomes. These represented twelve different haplotypes (Table 7-3). Two of the twelve haplotypes are common and differed at only one SNP. They accounted for 10 chromosomes (28%) each, therefore representing 56% of the observed variation. The third most common haplotype accounts for five chromosomes (14%). Therefore, 70% of the haplotype diversity of these 23 SNPs is represented by three haplotypes, and this can be captured by sampling just three SNPs (for example, SNPs 1, 5 and 11). The remaining nine haplotypes are relatively rare, accounting for between one (2.7%) and two (5.5%) chromosomes each.

The disease associated haplotype for family 59 was one of the most common and for family 50 it was the third most common. The disease associated haplotype for family 22 was unique and the disease associated haplotype for family 48 could not be unambiguously determined. Therefore, the disease assoicated hapotypes of the three families are all different. Family 50, 48 and 59 share a common 12 SNP haplotype from SNPs 11-23, excluding 22 (Table 7-3). There are only two haplotypes that are observed in this region. One is composed entirely of the common alleles, and the other is composed entirely of the rare alleles, and family 22 has the rare haplotype.

The frequency of the rare 12 SNP haplotype on the AS panel is 11.4%. In the gene they cover a distance of just over 1kb, from intron two to the 3' UTR. Therefore, this is a small region with a lot of SNPs that are observed infrequently in the population and that are all in LD with one another in this sample. I have already described that SNP ih34 within this haplotype alters the amino acid sequence of the protein, but that its effect on the function of the protein is unclear. The fact that this rare haplotype is observed on the F22 disease haplotype, and that ih34, part of this haplotype, may alter the function of the protein, is interesting. As mentioned previously, a positive association in this region would justify further study of these SNPs.

A Fishers exact test was performed using the Analyse-it® (version 1.71) software for microsoft excel. Since the disease haplotype of F48 does not extend into MR1, the position of GPR78, the analysis was carried out with and without F48. Each SNP was analysed individually for allele sharing. A two-tailed Fishers exact test on each of the 21 SNPs shows that there is no significant difference between the allele frequency of disease and non-disease chromosomes with or without F48 included in the analysis (Table 7-4). Therefore, in conclusion, there is no obvious pattern of excess haplotype sharing between the four families that would suggest the presence of a common susceptibility variant to psychiatric illness.

Figure 7-5 (continued over four pages): Genotyping results from the Allele Sharing (AS) panel, a DNA panel compised of members from families 22, 48, 50, 59. Twenty-three single nucleotide polymorphisms (SNPs) were identified by sequencing exons of the orphan G-protein-coupled receptor 78 gene on the AS panel. SNPs are labelled down the lefthand side of the figure. The family ID and chromosome number (relatedness between family members means that 45 different chromosomes are represented on the AS panel) are shown along the top of the figure (d = haplotype inherited with the disease). A representative of each chromosome is highlighted (blue and purple). Ambiguous results, where allele segregation cannot be determined, are boxed. Ih47 and ih48 genotypes were unrelaible and therefore not used. \ = missing data. Thirty-six chromosomes had complete haplotypes (purple).

		F50														F59									
		1		2		3		4		7		12		15		2078		2079		2080		3688		3784	
		d	1	2	3	d	2	4	5	1	3	D	5	2	5	D	1	2	3	d	2	d	5	4	5
1	rs6447884	A	G	A	G	A	A	G	A	\	\	A	G	A	G	G	G	G	A	G	G	G	A	G	A
2	rs4385041	C	G	C	G	C	C	C	G	\	\	C	G	C	G	G	G	G	C	G	G	G	C	G	C
3	ih162	A	G	A	G	A	A	A	G	\	\	A	G	A	G	G	G	G	G	G	G	G	A	G	A
4	ih165	C	C	C	C	C	C	C	C	\	\	C	C	C	C	C	C	C	C	C	C	C	C	C	C
5	rs4235270	A	C	A	C	A	A	A	A	\	\	A	A	A	A	C	A	A	C	C	A	C	A	A	A
6	ih33	A	G	A	G	A	A	A	G	\	\	A	G	A	G	G	G	G	G	G	G	G	A	G	A
7	ih32	C	A	C	A	C	C	C	A	\	\	C	A	C	A	A	A	A	A	A	A	A	C	A	C
8	ih31	G	T	G	T	G	G	G	T	\	\	G	T	G	T	T	T	T	T	T	T	T	G	T	G
9	ih48	G	C	G	C	G	G	G	C	C	C	C	C	C	C	C	C	C	C	\	\	C	C	G	C
10	ih47	A	G	A	G	A	A	A	G	G	G	G	G	G	G	G	G	G	G	G	G	G	A	G	A
11	rs3115388	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
12	rs3115387	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
13	ih34	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
14	ih44	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G
15	ih43	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T
16	ih42	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G
17	ih41	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
18	ih40	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
19	ih36	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G
20	ih37	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T
21	ih39	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T
22	ih38	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	T
23	ih35	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C

F22
16 47 48 49 111 17 52 53 112 54 55
d 1 d 3 4 5 1 2 2 4 6 7 8 9 D 7 d 8 10 11 12 13 d 10
1 rs6447884 G A G A G G A A A G G A G G G A G G G G G G A G G
2 rs4385041 G C G C G G C C C G G C G G G C G G G G C G G G
3 ih162 G A G A G G A A A G G A G G G A G G G G A G G G
4 ih165 C C C C C C C C C C C C C C C C C C C C C C C C
5 rs4235270 A A A A A A A A A A A A A A A A A A A A A A A C
6 ih33 G A G A G G A A A G G A G G G A G G G G A G G G
7 ih32 A C A C A A C C C A A C A A A A C A A A C A A A
8 ih31 \ \ T G T T G G G T T G T T G T T T C T T T
9 ih48 C C C C G G C C C C C G C C C C C C C C C C C C
10 ih47 G G G G A A G G A G G G G G A G G G G G G G G
11 rs3115388 T C T C C C C C C C C C C C T C T C C C C C T T
12 rs3115387 G A T G A A A A A A A A A A G A G A A A G G G
13 ih34 T C T C C C C C C C C C C C T C T C C C T T T
14 ih44 A G A G G G G G G G G G G G A G A G G G A A A
15 ih43 C T C T T T T T T T T T T T C T C T T T C C C
16 ih42 A G A G G G G G G G G G G G A G A G G G A A A
17 ih41 C A C A A A A A A A A A A A C A C A A A C C C
18 ih40 G C G C C C C C C C C C C C G C G C C C G G G
19 ih36 C G C G G G G G G G G G G G C G C G G G C C C
20 ih37 G T G T T T T T T T T T T T G T G T T T G G G
21 ih39 C T C T T T T T T T T T T T C T C T T T C C C
22 ih38 G G G G G G G G G G G G G G G G G G G G G G G
23 ih35 T C T C C C C C C C C C C C T C T C C C T T T

F22
114 63 22 62 24 68 70 87 88 90
d 12 1 15 14 15 D 14 d 1 16 17 d 16 D 17 d 18 19 20 18 21 22 23
1 Rs6447884
2 Rs4385041
3 Ih162
4 Ih165
5 Rs4235270
6 Ih33
7 Ih32
8 Ih31
9 Ih48
10 ih47
11 rs3115388
12 rs3115387
13 ih34
14 ih44
15 ih43
16 ih42
17 ih41
18 ih40
19 ih36
20 ih37
21 ih39
22 ih38
23 ih35

F22
149 150 99 147 153 103 143 F48 823 836 5
25 26 d 24 D 19 21 23 24 25 d 28 D 27 d 2 d 3 d 4
1 rs6447884 G G G G G G G G G G G A G A G A
2 rs4385041 G G G G G G G G G G G G G G C G C
3 ih162 G G G G G G G G G G G G G G A G A
4 ih165 C C C C C C C C C C C C C C C C C
5 rs4235270 C A C A A A C C C A C A C C C A C
6 ih33 G G G G G G G G G G G G G G A G A
7 ih32 A A A A A A A A A A A A A A C A C
8 ih31 T T T T T T T T T \ T T T T G T G
9 ih48 C C C C C G C C C G C G G G G G G
10 ih47 G G G G G G G G G G A G A A A A A
11 rs3115388 C T C T C C T C C T C T C C C C C
12 rs3115387 A G A G A A G A A G A G A A A A C
13 ih34 C T C T C C T C C T C T C C C C C
14 ih44 G A G A G G A G G G A G G G G G G
15 ih43 T C T C T C T T T C T C T T T T T
16 ih42 G A G A G G A G G G A G G G G G G
17 ih41 A C A C A A C A A A C A A A A A A
18 ih40 C G C G C C G C C G C G C C C C C
19 ih36 G C G C G G C G G C G T G G G G G
20 ih37 T G T G T T G T T T G T G T T T T
21 ih39 T C T C T T C T T T C T C T T T T
22 ih38 G G G G G G G G \ G G G G G G G
23 ih35 C T C T C C T C C T C T C C C C C

	SNP	1	2	3	4	5	6	7	8	11	12	13	14	15	16	17	18	19	20	21	22	23	Freq.
	Distance	-	111	28	173	106	229	114	456	5059	56	338	132	51	35	22	19	66	68	20	41	237	no. (%)
HAPLOTYPE	1	G	G	G	C	C	G	A	T	C	A	C	G	T	G	A	C	G	T	T	G	C	10 (28)
	2	G	G	G	C	A	G	A	T	C	A	C	G	T	G	A	C	G	T	T	G	C	10 (28)
	3	A	C	A	C	A	A	C	G	C	A	C	G	T	G	A	C	G	T	T	G	C	5 (14)
	4	A	C	G	C	C	G	A	T	C	A	C	G	T	G	A	C	G	T	T	G	C	2 (5.5)
	5	A	C	A	C	A	A	C	G	C	A	C	G	T	G	A	C	G	T	T	T	C	2 (5.5)
	6	A	C	A	C	A	A	C	T	C	A	C	G	T	G	A	C	G	T	T	G	C	1 (2.7)
	7	G	G	G	C	A	G	A	T	C	A	C	G	T	G	A	C	G	T	T	T	C	1 (2.7)
	8	G	C	A	C	A	A	C	G	C	A	C	G	T	G	A	C	G	T	T	G	C	1 (2.7)
	9	A	G	G	C	A	G	A	T	C	A	C	G	T	G	A	C	G	T	T	G	C	1 (2.7)
	10	A	C	A	T	A	A	C	G	C	A	C	G	T	G	A	C	G	T	T	G	C	1 (2.7)
	11	G	G	G	C	A	G	A	T	T	G	T	A	C	A	C	G	C	G	C	G	T	1 (2.7)
	12	G	G	G	C	C	G	A	T	T	G	T	A	C	A	C	G	C	G	C	G	T	1 (2.7)

Table 7-3: Haplotypes identified from the Allele Sharing (AS) panel; a DNA panel comprised of members from families 22, 48, 50 and 59. Twenty-one single nucleotide polymorphisms (SNPs) were identified by sequencing exons of the orphan G-protein-coupled receptor (GPR78) gene on the AS panel. Relatedness between family members means that 45 different chromosomes are represented on the AS panel. Thirty-six of the 45 chromosomes had complete data for all 21 SNPs. Twelve 21-SNP haplotypes are observed from these 36 chromosomes. SNP number corresponds to the number in Figure 7-4. Distance is the distance in base pairs between it and the preceding SNP. Frequency refers to the number (and percentage) of the 36 chromosomes upon which the observations are based.

SNP	Chromosome	Allele 1 inc. F48	Allele 2 inc. F48	p-value inc. F48	p-value exc. F48
1	D	(G)2	(A)1	1.00	-
	ND	25	13		
2	D	(G)3	(C)1	1.00	1.0
	ND	27	14		
3	D	(G)3	(A)1	1.00	1.0
	ND	29	12		
4	D	(C)4	(T)0	1.00	1.0
	ND	40	1		
5	D	(A)2	(C)2	0.8447	1.0
	ND	25	12		
6	D	(G)3	(A)1	1.00	1.0
	ND	29	12		
7	D	(A)3	(C)1	1.00	1.0
	ND	29	12		
8	D	(T)3	(G)1	1.00	1.0
	ND	29	11		
11	D	(C)3	(T)1	0.3756	0.2923
	ND	36	1		
12	D	(A)3	(G)1	0.3756	0.2923
	ND	36	1		
13	D	(C)3	(T)1	0.3756	0.2923
	ND	36	1		
14	D	(G)3	(A)1	0.3756	0.2923
	ND	36	1		
15	D	(T)3	(C)1	0.3756	0.2923
	ND	36	1		
16	D	(G)3	(A)1	0.3756	0.2923
	ND	36	1		
17	D	(A)3	(C)1	0.3756	0.2923
	ND	36	1		
18	D	(C)3	(G)1	0.3756	0.2923
	ND	36	1		
19	D	(G)3	(C)1	0.3756	0.2923
	ND	36	1		
20	D	(T)3	(G)1	0.3756	0.2923
	ND	36	1		
21	D	(T)3	(C)1	0.3756	0.2923
	ND	36	1		
22	D	(G)4	(T)0	1.00	1.0
	ND	39	2		
23	D	(C)3	(T)1	0.3756	0.2923
	ND	36	1		

Table 7-4: The results of a Fishers exact test on 21 single nucleotide polymorphisms identified by sequencing exons of the orphan G-protein-coupled receptor 78 gene in the Allele Sharing DNA panel (46 members of families 22, 48, 50 and 59, totalling 45 different chromosomes). The number of chromosomes successfully genotyped for each allele is included in the table (with the corresponding allele in brakets). A two-tailed hypothesis for a difference in allele frequency between 'diease' and 'non-disease' alleles has been used. D = allele inherited with the disease. ND = allele not inherited with the disease. The results have been calculated including and excluding the data from Family 48.

7.4. Association Study on Pooled DNA

As discussed in the previous Chapter, association analysis can provide a suitable way to narrow down a large linkage region, and DNA pooling provides a way to reduce the cost and time of genotyping.

7.4.1. Sample

Pools were constructed as Section 6.4.1. The pools plate (Figure 2-5) contained six replicates of each of the four DNA pools and 24 individual DNAs.

7.4.2. SNPs

At the time of the study on pooled DNA, 18 of the 23 SNPs (SNPs 6-23 in Figure 7-2) had been identified from a subset of 11 chromosomes from the AS panel. Linkage disequilibrium (LD) was calculated from these 11 chromosomes in order to determine a smaller set to study for association. Taking each SNP in turn, I calculated the occurrence of each of the two alleles with each allele of every other SNP. An example of complete LD therefore would be where allele A of SNP 1 only ever occurs with allele A of SNP 2. Based on such small numbers of chromosomes, it would only be feasible to calculate an LD block as instances of complete LD and type one member from this block.

Three groups of LD were identified (Figure 7-6). Group one consisted of SNPs 6 and 7, group two consisted of SNPs 9 and 10 and group three consisted of SNPs 11-13, excluding SNP 22. Therefore, six SNPs were chosen for the association study: SNP 7 from group one, SNP 9 from group two, SNPs 13 and 16 from group three, and SNPs 8 and 22. Despite the lack of segregation of SNPs 9 and 10 in the larger AS panel, at this time, the genotypes did segregate in these eleven chromosomes, and therefore they were included in the study.

S		6		7		8		9		10		11		12		13		14		15	
	Al.	G	A	A	C	T	G	C	G	G	A	C	T	A	G	C	T	G	A	T	C
6	G	7	0																		
	A	0	4																		
7	A	7	0	7	0																
	C	0	4	0	4																
8	T	6	1	6	1	7	0														
	G	1	3	1	3	0	4														
9	C	6	2	6	2	6	2	8	0												
	G	1	2	1	2	1	2	0	3												
10	G	6	2	6	2	6	2	8	0	8	0										
	A	1	2	1	2	1	2	0	3	0	3										
11	C	6	4	6	4	6	4	7	3	7	3	10	0								
	T	1	0	1	0	1	0	1	0	1	0	0	1								
12	A	6	4	6	4	6	4	7	3	7	3	10	0	10	0						
	G	1	0	1	0	1	0	1	0	1	0	0	1	0	1						
13	C	6	4	6	4	6	4	7	3	7	3	10	0	10	0	10	0				
	T	1	0	1	0	1	0	1	0	1	0	0	1	0	1	0	1				
14	G	6	4	6	4	6	4	7	3	7	3	10	0	10	0	10	0	10	0		
	A	1	0	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1		
15	T	6	4	6	4	6	4	7	3	7	3	10	0	10	0	10	0	10	0	10	0
	C	1	0	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1	0	1
16	G	6	4	6	4	6	4	7	3	7	3	10	0	10	0	10	0	10	0	10	0
	A	1	0	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1	0	1
17	A	6	4	6	4	6	4	7	3	7	3	10	0	10	0	10	0	10	0	10	0
	C	1	0	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1	0	1
18	C	6	4	6	4	6	4	7	3	7	3	10	0	10	0	10	0	10	0	10	0
	G	1	0	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1	0	1
19	G	6	4	6	4	6	4	7	3	7	3	10	0	10	0	10	0	10	0	10	0
	C	1	0	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1	0	1
20	T	6	4	6	4	6	4	7	3	7	3	10	0	10	0	10	0	10	0	10	0
	G	1	0	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1	0	1
21	T	6	4	6	4	6	4	7	3	7	3	10	0	10	0	10	0	10	0	10	0
	C	1	0	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1	0	1
22	G	6	4	6	4	6	4	7	3	7	3	9	1	9	1	9	1	9	1	9	1
	T	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0
23	C	6	4	6	4	6	4	7	3	7	3	10	0	10	0	10	0	10	0	10	0
	T	1	0	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	1	0	1

Figure 7-6 (continued overleaf): Linkage disequilibrium (LD) calculations of 18 single nucleotide polymorphisms (SNPs) in the orphan G-protein-coupled receptor 78 (GPR78) gene. Eleven chromosomes were used to calculate the occurrence of each allele of each SNP with each allele of every other SNP. Three groups of complete LD (where two of the four possible genotypes are seen) are shown (yellow, blue and purple). S=SNP, Al.=allele.

16		17		18		19		20		21		22		23	
G	A	A	C	C	G	G	C	T	G	T	C	G	T	C	T
10	0														
0	1														
10	0	10	0												
0	1	0	1												
10	0	10	0	10	0										
0	1	0	1	0	1										
10	0	10	0	10	0	10	0								
0	1	0	1	0	1	0	1								
10	0	10	0	10	0	10	0	10	0						
0	1	0	1	0	1	0	1	0	1						
10	0	10	0	10	0	10	0	10	0	10	0				
0	1	0	1	0	1	0	1	0	1	0	1				
9	1	9	1	9	1	9	1	9	1	9	1	10	0		
1	0	1	0	1	0	1	0	1	0	1	0	0	1		
10	0	10	0	10	0	10	0	10	0	10	0	9	1	10	1
0	1	0	1	0	1	0	1	0	1	0	1	1	0	0	1

7.4.3. Genotyping

As described in the previous chapter, the SNaPshot™ method of genotyping was used to esitmate the allele frequency from DNA pools. The SNaPshot™ reaction was run on the ABI PRISM® 3100 DNA sequencer and analysed using the GeneScan version 3.0 software.

7.4.4. Results

Section 6.4.4 describes in detail how the allele frequency of each DNA pool is determined and how Chi-square is calculated. The results can be seen in Table 7-5. The minor allele for SNP 22 was not visible. This was thought to be because it is too rare in the population. This was supported by the fact that the minor allele frequency of the SNP from the AS DNA panel was only 4.4% (or two chromosomes). The assay for SNP 8 failed. The SEM for SNP 9 is a little high for the control, BPAD and RMD groups, so that the significant association observed in the SCZ group is not reliable. A significant association was also found for ih34 between the control and RMD groups. This is interesting in the light of the previous observations about the amino acid change in the protein caused by this SNP, and its occurrence on the disease chromosome of the family that segregates affective disorder (F22). An almost significant association was also found in SNP 16 between the BPAD and control groups. The different results obtained for SNPs 13 and 16 do not follow with the fact that, based on 11 chromosomes from the AS panel, they are in complete LD with each other. Upon genotyping the entire AS panel, the two SNPs were still in complete LD (Figure 7-5). Since the association results suggest that the two SNPs are not in LD in that population, it would appear that 45 or less chromosomes is not enough from which to calculate LD reliably. An alternative explanation is that the genotyping assay is unreliable.

7.4.5. Problems

AS mentioned in chapter 6, a subsequent discovery that the pools contained duplicated samples meant that work on pools had to be stopped. Therefore, I did not attempt to further reduce the SEM for SNP 9 or get the assay to work for SNP 8 and unfortunately the positive associations observed were not reliable. I therefore moved on to readdress the question of whether any of the SNPs showed significant association by undertaking a study on individuals.

		SNP				
		7	**8**	**9**	**13**	**16**
SEM	**Het**	0.09	-	0.09	0.03	0.06
	CTL	0.009	-	0.04	0.005	0.005
	BPAD	0.009	-	0.06	0.003	0.004
	SCZ	0.007	-	0.01	0.004	0.004
	RMD	0.01	-	0.03	0.002	0.01
p-value	**BPAD**	0.6	-	0.9	0.3	**0.06**
	SCZ	0.7	-	**0.03**	0.95	0.7
	RMD	0.6	-	0.3	**0.05**	0.1

Table 7-5: The results of a Chi-Square test for association on pooled DNA. Five single nucleotide polymorphisms (SNPs) in the G-protein-coupled receptor gene were tested. The allele frequency of each SNP was measured in DNA pools of control individuals (CTL), bipolar affective disorder patients (BPAD), schizophrenic patients (SCZ) and recurrent major depressive patients (RMD). Allele frequency comparisons between the CTL group and each of the case groups are given as a p-value. Each DNA pool, and five individual heterozygotes (Het) are genotyped a number of times and the standard error of the mean (SEM) measures the variation between these genotyping replications.

7.5. Phase I Association Study on Individuals

7.5.1. Sample

The Phase I sample consisted of 95 controls, 93 BPAD patients and 95 schizophrenic patients.

7.5.2. SNPs

Twenty-two of the 23 SNPs that had been identified by sequencing individuals from the AS DNA panel were tested in Phase I. Three additional SNPs were chosen from Applied Biosystems as an 'Assay on Demand' TaqMan® assay. SNPs were chosen that had a causacian minor allele frequency greater than 0.1. SNPs C_1221917_10 and C11352300_10 are 8.8kb and 2.9kb from the 5' end of exon one respectively, and SNP C_1221895_10 is 3.6kb from the 3' of exon 3 (not shown on Figure 7-2). Therefore, a total of 25 SNPs across GPR78 were included. The 25 SNPs included in Phase 1 gave one SNP every ~870bp. The structure of LD was unknown since it was with these SNPs that LD was to be determined. Therefore, a SNP every 800bp, accounting for potential assay failures, was predicted to cover LD across the region.

7.5.3. Genotyping and Analysis

The 22 SNPs identified from the AS DNA panel were sent to the Sanger Institute for genotyping by the Sequenom® MASSARRAY™ primer extension method. The three SNPs that were provided as an 'Assay on Demand' by Applied Biosystems were genotyped by the TaqMan® method by Alison Condie on an ABI PRISM® 3730 Genetic Analyser.

As Chapter 6, association was measured using Chi-square method and LD was measured using D' and r^2 using the UNPHASED software (Section 6.5.3).

7.5.4. Results

Table 7-6 details the results of the association analyses. Fifteen SNPs could not be used in the association analysis for the following reasons. Three SNPs failed the primer design stage of the assay, the assay failed for five of the SNPs and one SNP was not polymorphic in this population. Six SNPs were not in Hardy-Weinberg equilibrium. One of these five also had a minor allele frequency that was too low and another did not have enough chromosomes genotyped. The fifteenth SNP had a minor allele frequency that was too low. Therefore, nine SNPs were suitable for LD calculation. Section 6.5.4 describes how and why LD was measured. The results show that the D' measurements between each SNP show moderate to high LD (>0.6) except between SNPs ih31 and oh34, where LD is low (D'=0.37). The r^2 measurements are more variable, but show high LD between ih33 and 31 and high to moderate LD between ih40 and ih36. Based on the D' measurements, if an LD block is defined by a continuous stretch of pairwise LD exceeding the threshold of 0.60, the nine SNPs form two LD blocks (SNPs c_1221917_10 to ih31 and ih34 to C_1221895_10). The D' values within the blocks (> 0.64) exceed the D' value between the blocks (0.37). Therefore, this suggests that the intervening region between SNPs ih31 and ih34 has not been covered in terms of the LD observed, and further association analysis might want to be undertaken. Furthermore, it also suggests that the Phase I association analysis constituted two independent tests for association (Nyholt, 2001).

Section 6.5.4 also describes how SNPs with a p-value <0.2 were to be chosen for the Phase II association study. The results of the Chi-square test for association on the Phase I sample revealed that one SNP, ih31, is significantly associated with schizophrenia and BPAD and schizophrenia combined. Therefore, only one SNP from the two LD blocks was chosen for study on the Phase II sample.

Table 7-6 (continued on next page): The Phase I association study results. Twenty-five single nucleotide polymorphisms (SNPs) in the orphan G-protein-coupled receptor 78 gene were genotyped in 95 controls (CTL), 95 schizophrenic (SCZ) and 93 bipolar affective disorder (BPAD) patients. The table details SNP name, the distance between a SNP and the preceding SNP, the minor allele frequency (MAF) determined from genotyping the Allele Sharing DNA panel or supplied by Applied Biosystems, the genotyping method (S = Sequenom®, T = TaqMan®), the reason for assay failure, the Hardy-Weinberg (HW) calculation expressed as a p-value ($p<0.05$ not tolerated), the number of control chromosomes genotyped successfully, the MAF in the CTL individuals, the Chi-square p-value in BPAD, SCZ and both groups combined and the linkage disequilibrium (LD) calculations (D' and r^2) between it and the preceding SNP (red = high LD, pink = moderate LD, white = low LD; SNPs highlighted blue were not included in the LD analysis)

SNP Name	Distance (bp)	MAF ch (%)	Method	Assay	HW	No. ch's [CTL]	MAF [CTL]	BPAD p-value	SCZ p-value	SCZ + BPAD p-value	D'	r^2
C_1221917_10	-	[23]	T		0.9	190	31	0.49	0.99	0.69		
C_11352300_30	5811	[27]	T		0.97	190	18	0.78	0.37	0.74	1	0.31
rs6447884)	2337	15 [34]	S		0.005	16	38					
rs4385041	111	14 [32]	S		0.5	188	39	0.56	0.88	0.68	0.81	0.47
ih162	28	13 [29]	S		0.00	162	15					
rs4235270	106	14 [32]	S	Failed primer design								
ih33	229	13 [29]	S		0.4	172	37	0.93	0.46	0.63	0.92	0.86
ih32	114	13 [29]	S		0.00	116	45					
ih31	456	10 [23]	S		0.2	122	45	0.46	0.08	0.15	1	1
ih48	777	-	S		0.0025	150	38					
ih47	6	-	S	Failed								
rs3115388	4276	5 [11.4]	S	Failed								

rs3115387	56	5 [11.4]	S	Failed								
ih34	338	5 [11.4]	S		0.5	164	87	0.67	0.97	0.79	0.37	0.11
ih44	132	5 [11.4]	S		0.00	144	16					
ih43	51	5 [11.4]	S	Monomorphic								
ih42	35	5 [11.4]	S	Failed primer design								
ih41	22	5 [11.4]	S	Failed primer design								
ih40	19	5 [11.4]	S		0.08	172	12	0.78	0.61	0.65	0.64	0.63
ih36	66	5 [11.4]	S		0.07	174	9	0.61	0.90	0.82	1	0.83
ih37	68	5 [11.4]	S		0.00	144	4					
ih39	20	5 [11.4]	S	Failed								
ih38	41	2 [4.4]	S		0.08	170	2					
ih35	237	5 [11.4]	S	Failed								
C_1221895_10		[32]	T		0.58	190	29	0.93	0.54	0.69	0.88	0.41

7.6. Phase II Association Study on Individuals

As discussed in the previous section, the Phase I association study was performed on a restricted set of individuals and a high p-value threshold of <0.2 was set. Power calculations suggested that this would capture all true association.

7.6.1. Sample

As discussed in Section 6.6.1., the Phase II sample consisted of 388 controls, 331 BPAD patients and 254 schizophrenic patients.

7.6.2. SNPs and Genotyping

One SNP, ih31, showed a significant association to schizophrenia ($p < 0.08$) and schizophrenia and BPAD combined ($p < 0.15$) (Table 7.6) and was therefore tested on this larger sample. Genotyping was performed the Sequenom® method, carried out by the Sanger Institute.

7.6.4. Results

The results showed that SNP ih31 was not associated with either schizophrenia, BPAD in the Phase II sample (Table 7-7).

				p-value	
SNP	**HW**	**MAF.**	**No. Ch's (CTL)**	**BPAD**	**SCZ**
Ih31	0.41	35	766	0.67	0.46

Table 7-7: The Phase II association study results. SNP ih31 in the orphan G-protein-coupled receptor 78 gene was genotyped in 388 controls (CTL), 331 bipolar affective disorder (BPAD) and 254 schizophrenic (SCZ) patients. The table details the test for Hardy-Weinberg (HW) equilibrium ($p<0.05$ not tolerated), the minor allele frequency (MAF) in the CTL group, the number of CTL chromosomes genotyped and the results (p-value) of the Chi-square test for a difference in allele frequency between the CTL and the case groups (SCZ and BPAD).

7.7. Discussion

Here I describe the genetic analysis of the GPR78 gene in MR1. MR1 is a candidate region for the susceptibility to psychiatric illness in three families that show linkage to the region. GPR78 is expressed in the placenta and the pituitary and is therefore positioned to play potential roles in the HPA axis and during pregnancy. The HPA axis plays an important role in stress regulation and has been shown to be dysfunctional in patients suffering from schizophrenia and affective disorders (Pariante and Miller, 2001; Altamura, 1999). The effect of maternal factors, such as stress hormones, on an unborn child has also been suggested to be involved in psychiatric pathology (Weinstock, 1997).

I identified twenty-three SNPs by sequencing the gene in members of the four linked families. Two SNPs alter the amino acid sequence of the protein. SNP 8 (ih31) changes amino acid 201 from an arginine to a serine, and SNP 12 (ih34) changes amino acid 342 from an arginine to a histidine. Bioinformatic and frequency analysis showed that these substitutions are unlikely to significantly alter the function of the protein.

Despite this, predicting the relationship between protein sequence, structure and function is inexact. Many tools (e.g. Pfam) rely on homology to known protein sequences and functional domains whilst others use algorithms extrapolated from past *in vitro* studies (e.g. NetPhos). Experimentally determined structures are only available for a limited number of proteins, and computational methods with varying degrees of accuracy are required to model the structures of the rest (Kopp and Schwede, 2004). Biological and biochemical data is still needed to reliably go from a sequence to function. For example, a cysteine to arginine amino acid mutation at two residues of the apolipoprotein E protein (residues 112 and 158) creates three common isoforms: E2 (cys/cys), E3 (cys/arg) and E4 (arg/arg) (Weisgraber *et al*, 1981). The E4 allele confers a susceptibility risk to late onset Alzheimer's disease (LOAD) (Corder *et al*, 1993) and is associated with an increase in the amount in beta-amyloid

deposition in the post-mortem brains of LOAD patients (Lambert *et al*, 2001). Subsequent *in vitro* work has shown that the E4 allele binds with significantly higher affinity to beta amyloid than the E3 allele (Sanan *et al*, 1994). However, it is unknown how an arginine at residue 112 should influence this, and without biochemical analysis it might not have been predicted. Therefore, prediction of the likely effect of an amino acid variant on protein function when very little is known about the protein, as is the case for GPR78, is very difficult.

The haplotypes of the four families were ascertained and assessed. The twenty-three SNP haplotype that occurs on the disease chromosomes of the families are different to each other. Thirteen haplotypes were observed in total, with three that are common. The disease haplotype of families 48, 50 and 59 are common and the disease haplotype of F22 is rare. A Fishers exact test revealed that there was no significant allele sharing between the four families and therefore a susceptibility haplotype could not be determined. As discussed in the previous chapter, there are certain limitations concerning the number of disease chromosomes and the potential for assortative mating. However, a Fishers Exact test is more suited to low numbers and efforts have been made to rule out assortative mating is unlikely since there is a disease associated risk haplotype inherited from one side in each family.

Six of the SNPs that had been identified in the families were tested for association in a set of pooled DNA samples. The results revealed a significant positive association between SNP 9 and the schizophrenic patient group, SNP 16 and the BPAD patient group, and SNP 13 and the RMD patient group compared to controls. However, the results were deemed unreliable because the pools contained replicate DNA's. For this reason, comparison of pooled and individual association results is not possible.

Twenty-two of the SNPs that had been identified from the four families and an additional three SNPs identified from the public databases were tested for association in a population of unrelated individuals. This preliminary association population consisted of 95 controls, 95 schizophrenic patients and 93 BPAD patients. A relaxed

p-value cut off of <0.2 was chosen to pick SNPs for a second round of association analysis on a larger population. SNP 8 (ih31) showed a significant positive association in this preliminary screen ($p<0.08$) and so was tested on the larger population. In addition, this data was used to measure LD across the region. It is not possible to do this with the AS DNA panel because the number of chromosomes is too few to be reliable. This was supported by the fact that the complete LD seen between some of the SNPs on the AS panel was not observed in the Phase I control population (inferred from the differences in allele frequency of the SNPs). Pairwise LD between the nine SNPs suggested the presence of two blocks (SNPs c_1221917_10 to ih31 and ih34 to C_1221895_10), in which D' values within the blocks (> 0.64) exceeded the D' value between the blocks (0.37). Therefore, this suggests that the intervening region between SNPs ih31 and ih34 has not been covered in terms of the LD observed, and further association analysis might want to be undertaken.

The results of the Chi-square test of SNP 8 on the Phase II sample did not find an association between this SNP and psychiatric illness. Using a high p-value cut off ($p<0.2$) for Phase I means that a large number of false positives would be expected (20%). The lack of association in SNPs apparently in LD with SNP 8 and the lack of replication of the Phase I finding in the larger Phase II population, suggests that the positive association of SNP 8 to schizophrenia was a false positive. A 20% false positive rate for nine SNPs would mean that 1.8 positive associations would be expected by chance. Therefore, one positive association cannot be considered unexpected.

An interesting haplotype was observed in family 22 between SNPs 11 and 23. The minor allele of SNP 22 was extremely rare (4.4%) and the minor alleles of the remaining 12 were moderately rare (11.4%) in the AS panel. There are two interesting things about this region. Firstly, it is a highly polymorphic region, with 12 SNPs in just over 1 kb. Secondly, the minor allele was observed only in F22, and only two haplotypes were observed. One haplotype is composed of the rare alleles

and the second haplotype is composed of the common alleles. The small distance and the presence of only two haplotypes suggests that they are so tightly in LD that they are never separated by a recombination event, but also that they occurred together in one founding individual. However, it might simply be that the AS panel does not contain enough chromosomes to sample all possible haplotypes. Evidence for this comes from the association study results. As mentioned previously, the fact that the Phase I sample consists of unrelated individuals means that haplotypes can only be inferred. However, the finding that the minor allele frequency of a subset of SNPs from the 12-SNP haplotype are different, suggests that more than two haplotypes exist in this larger population. Since a positive association was not identified between any SNPs in this haplotype and schizophrenia or BPAD in Phase I, the haplotype was not studied further.

In summary, the results of genetic analysis of GPR78 in the four families did not reveal a mutation that would unambiguously alter the function of the protein and contribute to psychiatric illness. However, the amino acid change caused by SNP ih34 is interesting because it is rare, located on the disease haplotype of F22, and abolishes a predicted intracellular protein kinase C phosphorylation site. However, the lack of a positive association between SNPs in the gene and psychiatric illness suggest that this is not worth pursuing.

Chapter Eight

Discussion

Discussion

8.1. Summary

This thesis describes the genetic analysis of candidate regions implicated in the susceptibility to psychiatric illness. Linkage in four families has previously identified a region of chromosome 4p linked to psychiatric illness. The focus of this thesis has been on MR1 and MR2, two candidate sub-regions within the linkage region. I have successfully refined the recombination breakpoint intervals of MR1. The refinement of the telomeric end of MR1 was aided by the positioning of four BAC clones in this region. I have also carried out analysis to identify novel genes within MR1 and MR2, and analysed two candidate genes for association to psychiatric illness.

It has become accepted that the identification of disease genes in complex disorders such as psychiatric illness will require the implementation of a number of complimentary approaches (Evans *et al*, 2001_b). This reflects the approach taken here.

8.2. Families

The results of linkage analysis in four families have been used to guide the work done here. There are a number of advantages of studying large families to identify genes in complex disease. A single gene of major effect is more likely to underlie the cause and should be common to the majority of affected members (Blackwood *et al*, 2001). The advantage of studying multiple linked families is that they allow the determination of a common shared haplotype between the families. However, reduced pentrance means that the disease susceptibility variant will also be present in family members without the illness.

A possible caveat of studying large families with relatively high disease penetrance is that the causative variant could be rare and may not generalise to the population of patients. However, even if this is the case, the identification of susceptibility genes

may highlight cellular pathways involved in the pathophysiology of psychiatric illness, and thus identify further candidate genes. Another caveat is that there may be more than a single gene operating in single families via mechanisms of genetic heterogeneity or assortative mating, where multiple genes are inherited from both parents (Blackwood *et al*, 2001). In these four families, the presence of only one disease associated haplotype, inherited from only one side of the family, makes assortative mating unlikely. Furthermore, genetic heterogeneity, at least in F22, ought to be limited. The genetic effect of the disease associated haplotype in F22 is large. Using the data from Blackwood *et al* (1996), there are 95 individuals in F22 for which the haplotype data (ascertained or inferred) and diagnostic status are known. In these 95 family members, the disease associated haplotype occurs in 100% of BPAD patients, 87.5% of RMD patients and 22% of the remaining non-affected family members. Therefore most of the genetic liability in this family is likely to come from this one genetic locus. It is, however, more difficult to make this kind of assumption about the smaller families.

8.3. Minimal Region One

8.3.1. Minimal Region One Contig

In order to study the linked regions in the future for genes and association, it is necessary to have the continuous sequence of the region. The production of high quality human genome sequence is of considerable value to the mapping of complex disease genes. The complete sequence not only lends itself to gene identification, but has also led to large scale projects to facilitate the identification of the variants underlying complex disease. For example, the efforts of dbSNP and the SNP consortium to identify SNPs in the human genome, and of the HapMap, which aims to characterise patterns of LD between SNPs and identify 'tagSNPs' that capture haplotype diversity, aim to provide the tools necessary for researchers to design and implement the appropriate study.

The complete human sequence at both Ensembl and the UCSC Golden Path were published in July 2003. The Sanger Institute, the curators of the Ensembl genome database, reports that 93.22% of the genome has been sequenced, of which 88.99% is finished, whilst the international human genome sequencing committee (IHGSC), curators of the UCSC Golden Path genome database, reports that 99% of gene containing regions have been sequenced. However, the quality of the draft sequence across the genome has been variable during its production, and even with the announcement of the finished product, gaps and inconsistencies are still evident. This was observed with the STS st175378snp (Section 3.3.2). Whilst it amplified a region of chromosome 15, it did not produce a match to any chromosome 15 clone after a BLASTn similarity search.

One of the gaps in the genome sequence lies within MR1, the minimal region formed by the overlapping disease haplotypes of families 22, 59 and 50. A contig gap within the recombination breakpoint interval of MR1 is flanked on one side by a region of low representation in genomic libraries, and on the other side by a region of highly repetitive sequence. The contig flanking the gap on the repetitive side has been highly unstable in previous releases of the UCSC Golden Path. It is preferable for our purposes to have unequivocal evidence for clone position. Bioinformatic alignment alone does not prove that two clones overlap, since it is not possible to rule out sequencing or clone labelling errors, or the alignment of repetitive clones from different chromosomal regions. The HGP is a large scale project that does not carry out detailed manual curation of the data, therefore making such errors more likely than the small scale work undertaken here. The colony PCR results described in Chapter three provided an extra layer of analysis to the HGP.

Positioning novel clones in the gap, and eventually closing the gap is a continuing aim of the project. It is important to understand the genomic landscape in this region, since it is within or around just such repetitive regions that chromosomal duplication, deletions or rearrangements are predicted to occur and that might affect gene expression or function (Cheung *et al*, 2003). The region will also be a source of novel SNP and microsatellite markers to refine the telomeric end of MR1. The

refinement of the recombination breakpoint interval is important in order to rule out genes from MR1, and novel markers are required to do this. Furthermore, the region may be a source of novel genes.

The predicted repetitive nature of the intervening sequence in the gap might be one reason that has precluded its identification, or its inclusion into a BAC in the genomic libraries, and contributed to the variability of the HGP contigs. The evidence for low copy number of the clones immediately telomeric to the gap might be further evidence of this. The possibility of using the chromosome 4 specific repeat sequence to probe BAC libraries has not been investigated and might prove worthwhile. However, several copies of the tandem repeat has also been identified on chromosome 8p (Gondo *et al*, 1998).

Using the Golden Path, it is possible to obtain clones for which the sequence is not yet finished and/or which cannot be placed with certainty at a specific position on the chromosome (http://genome.ucsc.edu/goldenPath/hg16/chromosomes/). It is likely that the reason they have not been positioned is because they are repetitive or that they are under represented in libraries. A future experiment may be to inspect these clones on an individual basis to design specific STSs to test whether any of them map to this region of MR1.

8.3.2. Minimal Region One Recombination Breakpoints

Typing SNP and microsatellite markers in the families enabled the successful refinement of both recombination breakpoint intervals of MR1 (Chapter 4). However, the resolution was not sufficient to rule in or out two candidate genes from the telomeric end of MR1. Since candidate genes cannot be identified by function alone, and association analysis is required to analyse MR1 further, ruling genes out of MR1 is desirable to reduce the future work load. Therefore, future work should concentrate on identifying novel markers to further refine the recombination breakpoint intervals.

Refinement of the telomeric recombination breakpoint of MR1 was precluded by an extensive region of homozygosity in the transmitting parent F50-3. This highlights a problem with this type of experiment; the fact that you are relying on chance to identify heterozygosity in parents transmitting the recombination breakpoint. Furthermore, recombination breakpoint mapping relies on random recombination events around the disease gene and since recombination only occurs at meiosis and once or rarely twice per chromosome per generation, the disease locus identified by recombinants is likely to be large.

Combining data from four families effectively increases the number of meioses and the chance of recombination events defining a smaller disease gene region. However, the fact that the four families show linkage to the same chromosomal region is encouraging but is not evidence that they share the same ancestry or disease gene. If the same gene and the same variant underlie the susceptibility in the families, the variant might have arisen in a single, relatively recent, common ancestor, (e.g. with the Celtic families) meaning that some or all of the families will share a common haplotype around this variant. Alternatively, the hypothetical common ancestor might be sufficiently far back in evolutionary time that recombination has degraded the surrounding haplotype (they will be the same within the LD block that the variant is in but this will be a substantially smaller region). A further scenario would be where the same variant arose independently in different ancestors (convergent evolution) or different variants arose in the same gene in different ancestors (allelic heterogeneity). Convergent evolution or allelic heterogeneity could have arisen on the same background haplotype (if this background haplotype is older than the variants), or different background haplotypes.

The highly significant Lod score obtained in F22 defines a region that is extremely likely to contain a susceptibility variant for psychiatric illness. The Lod score in F48 is also significant and, used in conjunction with the region defined by F22, defines a smaller overlapping linkage region. However, allelic heterogeneity might still operate. The Lod scores obtained in the two small families, F50 and F59, are much more likely to be spurious, but, if they are real, allelic heterogeneity is still a

possibility. Therefore, the use of families 48, 50 and 59 should be viewed as a way of prioritising the search for a functional susceptibility variant in F22, rather than searching exclusively for a common susceptibility variant in the four families.

A project such as this cannot rely on using linkage and recombination breakpoint mapping alone to identify the candidate gene and the susceptibility variant to a complex disorder. It simply does not produce the desired resolution and other complimentary approaches are needed. The refinement of linkage regions is really just the beginning of the process for the identification of the disease gene.

8.4. Allele Sharing

Analysis of the family haplotypes in the genic regions provides a complimentary approach for narrowing down this large linkage region and SNPs were identified from members of the four families with this aim. The AS panel was constructed to have enough control chromosomes to compare the frequency of allele sharing between the families compared to controls. A region of increased haplotype sharing would support the common ancestor hypothesis. If this concurred with the results of association analysis then this would provide compelling evidence for the identification of the susceptibility region. However, the marker haplotypes are currently biased by the focus of SNP discovery in coding and regulatory regions and by the natural uneven spread of those SNPs that are identified within these regions. Within a gene this is not so much of a problem, but between all genes within MR1 and MR2 this results in vast regions of 'empty' space.

Allele sharing was not observed in GPR78 or SOD3, as measured by the Fishers exact test on a marker-by-marker basis (Sections 6.3.2 and 7.3.2). An alternative application of the Fishers exact test would be to study the allele frequency of a sliding window of two or more marker haplotypes since this may be more informative than marker alleles alone. For example, it has been seen that the association observed between a haplotype in the NRG1 gene and schizophrenia is not observed for each of the individual alleles that comprise the haplotype (Stefansson *et*

al, 2002). The analysis of haplotype sharing across the whole extent of MR1 and MR2 is a worthwhile future task and highlights the advantage of this project by being able to study several families that show linkage to the same region.

8.5. Identification of Genes

Without the HGP, the task of identifying the disease gene operating in these families would be significantly harder. It is vital to have a map of all the genes in the linked region since any one of them might harbour the disease variant and the production of finished human sequence is a critical component to the identification of genes. The current release of Ensembl (version 19.34b.2) reports the identification of 31,609 gene transcripts. It is unknown what the final gene count will be, but the early estimates of 100,000 have been curbed dramatically to the order of 30,000 or less since the completion of the human genome (Claverie, 2001; Pennisi, 2003).

It is an important aim of the group to have a catalogue of all the genes within the linked region. Abundantly expressed genes are readily identified due to the presence of vast numbers of ESTs and other transcript evidence. The majority of currently known genes constitute such abundant transcripts, for example, GPR125 in MR2. However, it is more problematic to identify genes with low expression levels and/or that are expressed at specific developmental time points or where expression is restricted to specific spatial regions. An example of this is the identification of the 74M11 gene in MR1 (Section 5.6.2.1). This gene had limited EST evidence supporting it and to only some of the exons. The 74M11 ESTs were not remarkably different to others in MR1 for which RT-PCR or cDNA library screening has not yet been performed. It is therefore entirely possible that other low level expressed genes have been missed. Furthermore, there is no reason to suppose that these rare or restricted genes are not of relevance for psychiatric illness.

To further complicate gene identification, some expression evidence is spurious. For this reason it is necessary to be cautious when performing RT-PCR or library screening based on a single line of evidence. A consensus of more than one line of

evidence is preferable. The more genes that are identified, the more it will be possible to learn about the characteristics that predict genes. However, this highlights the circularity of gene prediction, since prediction methods can only be based upon the known characteristics of confirmed genes.

Most of the results from RT-PCR and library screening were negative (Sections 5.5.1.6 and 5.5.2.2). This could be because the transcript evidence is spurious, but it could also be due to assay failure. It is difficult to draw definite conclusions from a negative result. A significant proportion of the RT-PCR assays were optimised before screening and therefore negative results are more likely to reflect an absence of expression. However, the library screening assays were also optimised before screening, but this did not preclude vast numbers of non-specific products being amplified. Despite this, the importance of prior optimisation can be seen with the RT-PCR attempted in order to extend the GPR125 transcript (Section 5.6.2.2). RT-PCR failed to link the first five exons with the last 14 exons, despite the fact that others subsequently published confirmation of the link in the UCSC Golden Path.

Despite an optimal assay, a negative result can always be attributed to the rarity of the transcript, as mentioned above. Attempts to address this were made by the use of universal cDNA (Clontech), a mixture of 37 human tissues to provide a wide range of cDNAs, in addition to the lymphoblastoid cDNA extracted inhouse.

A number of alternative techniques might be employed for future transcript mapping. 5' RACE (rapid amplification of cDNA ends), and other cDNA based techniques, would be subject to the same limitations as RT-PCR and cDNA library screening. Northern blotting would provide an alternative method for transcript verification and would also immediately identify the size of the transcript and give an idea of the existence of alternative transcripts. However, this would be relatively time consuming and costly to apply to large regions such as these. Real time PCR would be a good method for detecting genes with low expression levels. However, this would depend on the availability of human tissue, or the identification of the homologue in other species. An interesting transcript identification technique

described by Li *et al* (2004_a) involved the construction of a microarray using short segments of genomic sequence that represented the chromosome 4q22-24 region. The genomic sequence is then used to identify novel genes from mRNA. This would provide a possible way to identify rare transcripts or genes with a structure that would not normally be predicted by prediction programmes.

For the purposes of this study, it is vital to have a complete transcript map of the candidate region in order to be able to direct and interpret future association analysis. There are advantages in the level of individual inspection of the transcript evidence that is possible from working as an individual experimenter. However, with the advent of the finished genome sequence, there are a number of large scale gene identification projects underway, for example Ota *et al* (2004) that may provide valuable data for this and other projects. Such large scale approaches might also dissuade the individual from carrying out the work his- or herself. However, in the same way that the production of the genome sequence, in the absence of detailed examination of the sequence assemblies, leads to errors, not every gene will be captured in a high throughput project and individual inspection of the region might still be productive.

The identification of genes is not trivial and will remain a challenge in the future. However, it is an essential endeavour for the aim of the project and for other groups attempting to identify susceptibility variants in complex disorders. It is interesting to wonder, from the point of view of the HGP, whether one can ever be sure that every human gene has been identified.

8.6. Identification of Association

Association studies were used here as another approach to narrow down the large regions of linkage in the families. SNPs were identified in the families to ensure that they were relevant to the population, and were identified around the coding and regulatory regions of known genes. Based on position and function, two candidate genes were studied for association.

DNA pooling has been suggested to be a powerful tool to reduce the time and cost involved with the large scale genotyping required to detect variants with small to moderate effect size in common complex diseases (Shaw *et al*, 1998; Daniels *et al*, 1998; Breen *et al*, 1999; Sham *et al*, 2002). Allele frequency estimation of SNPs is not significantly affected by genotyping technique (SNaPshot™ or primer extension measured with dHPLC or mass spectrometry), DNA pool size and the natural unequal amplification that occurs for a SNP is not affected by the allele frequency in a pool (LeHellard *et al*, 2002). However, the biggest disadvantage, not discussed in most papers, is their lack of flexibility. This was seen in this study when it was identified that the pools contained a number of sample ID errors (Section 6.4.6). Simple human error, where samples of the wrong diagnosis were included in the pool, could have been avoided by better methods of sample labelling, for example barcoding. However, the pools also contained individuals where the diagnoses changed over time. This type of error cannot be planned for, and is unfortunately not unusual in psychiatric illness. A second limitation of DNA pooling is the inability to construct haplotypes, although there are a number of programmes available that can estimate them. The decreasing price of individual genotyping meant that it was considered more reliable to perform association studies on individuals.

Association analysis on individuals was divided into two phases. The smaller Phase I population aimed to calculate association but also to identify LD between SNPs. Based on this, a non-redundant set of SNPs that are associated with psychiatric illness and that represent LD blocks can be chosen for Phase II. A relaxed threshold for significance aimed to capture all true association in the small sample, accepting that a greater number of false positives will also be taken through. The importance of LD in association analysis is two fold. Firstly, it is a way of reducing marker redundancy and secondly, it ensures that the gene has been adequately covered. The rationale of the two phases was based on a set of power calculations performed by Dr. Naomi Wray. However, these power calculations assume that the causal SNP is being studied, and therefore, the power reduces if it is not.

Consequently, splitting association analysis into two phases saves time and money, an important consideration in large projects such as this. Association analysis of two genes, SOD3 and GPR78, involving a total of 38 markers, involved a great deal of time and effort. Genotyping this entire set on the Phase I and Phase II sample combined would be costly, especially if every gene known in MR1 and MR2 (currently 23) is to be tested in a similar manner. In the future, a several tiered approach seems sensible, where regions within MR1 and MR2 that show positive association are followed up with a higher density of markers to cover LD more adequately and confirm results.

The results of the phase I and II association studies on two genes in the minimal regions, SOD3 (Section 6.5.4) and GPR78 (Sections 7.5.4 and 7.6.4), found that there was no positive association between SNPs in the genes and psychiatric illness. However, the LD calculations suggested that neither gene fully covered all LD blocks and future studies may want to test more SNPs in these genes.

8.7. Identifying the Causal Variant

A positive association between markers within or near a gene and psychiatric illness is really only the first stage in the process of identifying the causal variant. A number of factors could complicate the implementation and interpretation of association analysis in the future for the group. This is mainly because an associated SNP is not likely to be the causative SNP itself, making the identification of the disease variant more difficult. Alternative splicing, the mechanism by which protein diversity is increased, is a highly complex and poorly understood process (Caceres and Kornblihtt, 2002; Cartegni *et al*, 2002). Alteration in splicing machinery has been shown to operate in human disease, such as in cancer and cystic fibrosis, resulting in changes in the relative levels of alternative spliced isoforms (Nissim-Rafinia and Kerem, 2002). These types of mutation are harder to identify by sequence analysis. The search for disease genes has traditionally centred on identifying differences in protein sequence. Here I analysed the potential effects of the SNPs in SOD3 (Section 6.3.1) and GPR78 (Section 7.3.1) which resulted in an altered protein sequence.

However, the causative variant in psychiatric illness is likely to cause subtle alterations in protein function, as exhibited by amino acid substitutions that have non-significant impact on protein function as assessed by conventional programmes, or by subtle effects on splicing or promoter machinery.

Epigenetic factors are heritable changes in the methylation status of genes and the structure and methylation, phosphorylation, acetylation, ubiquitination and ADP-ribosylation status of chromatin that affect gene expression, but do not involve a change in DNA sequence (Nakao, 2001). The differential expression of paternal and maternal genes by the mechanism of CpG methylation is thought to be important for the regulation of reproduction, placentation, energy homeostasis, lactation and behaviour (Surani, 2001). In psychiatric illness, there is some evidence from linkage and association studies for epigenetic control of inheritance, such as paternal or maternal inheritance or evidence for the effects of imprinting on gene expression (Petronis, 2000). Such effects would mean that an association might be missed. For example, it might be necessary to characterise differential expression of imprinted alleles when interpreting the results of allelic association studies. Sex differences were not analysed here and might be worth considering for future analysis if no association is found in the minimal regions. This is another reason why performing association analysis on individuals rather than pools is perhaps preferable. At the beginning of a study, it is not necessarily possible to account for all the ways a case sample is to be divided up and analysed in the future.

The premise of this work assumes that the causal variant lies within or near a gene in this region of chromosome 4p and that the variant affects this genes functioning. However, long distance gene regulation has been identified in vertebrates (Kluppel *et al*, 1997) and, therefore, it is possible that a causal variant a considerable distance away in a different part of the genome regulates a gene in the chromosome 4p region, or alternatively, that a variant in the chromosome 4p region regulates a gene a considerable distance away.

8.8. Future Work

Association studies and haplotype analysis are currently being performed on all the known genes in MR1 and MR2, and MR1 and MR2 wide haplotype and LD maps are being constructed as a result. However, SNP discovery is focused on genes and the genes are unevenly spaced and therefore there will be an uneven coverage of markers across these regions. It might be that a variant affecting gene expression or function operates a large distance away from the gene. In this situation, analysis of the intergenic regions would be desirable.

The public HapMap project (www.hapmap.org/) has recently published the results of its measurements of LD across the genome (March 2004). Using the SNPs that make up the HapMap across MR1 and MR2, 442 SNPs would be required to capture all haplotypes with a frequency of greater than 10%, or 271 SNPs would be required to represent one SNP per haplotype block (S. Le Hellard, personal communication). In order to minimise the cost of sequencing intergenic regions in the families, the HapMap will be useful in the future in order to close LD gaps formed after SNP discovery in the families and genotyping in the phase I population.

Biochemical approaches will be the next step in testing for altered gene function in affected versus non-affected family members. As mentioned previously, association studies in themselves will not necessarily identify the causative variant and the causative variant will not necessarily be obvious. Furthermore, an area of association may not be identified at all. Therefore, complimentary methods for identifying the causative polymorphism form an important adjunct to association analysis.

Protein and RNA expression profiling by microarray is becoming a popular technique in complex diseases (Bunney *et al*, 2003), and could be considered in the future for this project. The microarray assay is able to reflect the complexity of the disease process by identifying multiple differences in a hypothesis free manner. This technology could be utilised to detect a difference in the protein expression profiles of disease haplotype and non-disease haplotype carriers in the four families. The

ultimate aim would be to identify differential expression of a protein that resides in one of the minimal regions, or that functions in the same cellular pathway as a protein in one of the minimal regions. If the results of protein expression profiling point to the same genomic region as the results of association and/or haplotype analysis, this would provide compelling evidence for the location of the susceptibility gene.

Other protein expression profiling techniques that could be used include mass spectrometry using SELDI-TOF or MALDI-TOF. In a similar way to microarrays, they identify differences in protein expression on the basis of protein mass rather than mRNA or protein sequence, in a hypothesis free manner.

The ultimate aim of the association analyses and/or protein or mRNA expression profiling will be to highlight sub-regions or individual genes within MR1 and MR2 that are associated with psychiatric illness. Hypothesis driven biochemical analysis of particular candidate genes could then provide information about altered protein function in affected and non-affecteds. For example, quantitative RT-PCR in patients versus controls would provide information on relative protein expression levels, whereas quantitative RT-PCR in mouse tissue at different developmental time points would provide information on developmental expression. *In situ* hybridisation or immunohistichemistry in mouse or human post-mortem samples could highlight the regional distribution of the mRNA or protein. Therefore, a whole array of future experiments will become possible as association and haplotype analysis proceed.

8.9. Findings Genes in Psychiatric Illness

Identifying genes in psychiatric illness is proving quite difficult. This is due to the problems discussed concerning the application of linkage and association analyses to complex disorders and the problem of extracting the real data from the false positives and false negatives (Merikangas and Risch, 2003).

This group is well positioned to identify the causal variant or variants on chromosome 4p predisposing to psychiatric illness in these families. The linkage findings are convincing and have been replicated. However, considerable effort is required in the future to continue with haplotype refinement, continue the search for allele sharing between the families and implement association and functional analyses across the entire region in order to pinpoint the gene involved. This is not trivial, and puts other research groups without the advantages of strong linkage families and the availability of large association populations at a distinct disadvantage. The emerging picture from this research is the importance of the use of complimentary techniques and the convergence of data and evidence. An exciting future prospect is the convergence of association and functional evidence to a gene or genes within one of the minimal linkage regions.

There is much work left to be done, but these are exciting times for psychiatric genetics. The accumulation of linkage, association and haplotype data is beginning to appear for certain candidate genes, for example NRG1 on chromosome 8 (Collier and Li, 2003). It won't be long before this group is in a similar position. As candidate genes become well established and confirmed and even susceptibility variants are identified, it is possible to envisage an explosion of findings in the field with the consistent implication of biochemical pathways and as methodologies become refined.

8.10. Conclusions

In conclusion, the work carried out here has contributed to the refinement of linkage regions for psychiatric illness on chromosome 4p, contributed to the identification of novel genes in these regions and analysed two positional and functional candidate genes for allele sharing in the families and association in a population of unrelated cases and controls. Whilst no positive association was identified, this has contributed to the analysis of the entire minimal regions, which contain a number of other positional candidate genes to study in the future.

References

Aita, V.M., Liu, J., Knowles, J.A., Terwilliger, J.D., Baltazar, R., Grunn, A., Loth, J.E., Kanyas, K., Lerer, B., Endicott, J., Wang, Z., Penchaszadeh, G., Gilliam, T.C. and Baron, M. (1999). "A comprehensive linkage analysis of chromosome 21q22 supports prior evidence for a putative bipolar affective disorder locus." Am J Hum Genet **64**(1): 210-7.

Akey, J., Jin, L. and Xiong, M. (2001). "Haplotypes vs single marker linkage disequilibrium tests: what do we gain?" Eur J Hum Genet **9**: 291-300.

Allen, M.G. (1976). "Twin studies of affective illness." Arch Gen Psychiatry **33**(12): 1476-8.

Allen, M.G., Cohen, S., Pollin, W. and Greenspan, S.I. (1974). "Affective illness in veteran twins: a diagnostic review." Am J Psychiatry **131**(11): 1234-9.

Als, T.D., Dahl, H.A., Flint, T.J., Wang, A.G., Vang, M., Mors, O., Kruse, T.A. and Ewald, H. (2004). "Possible evidence for a common risk locus for bipolar affective disorder and schizophrenia on chromosome 4p16 in patients from the Faroe Islands." Mol Psychiatry **9**(1): 93-8.

Altamura, A.C., Boin, F. and Maes, M. (1999). "HPA axis and cytokines dysregulation in schizophrenia: potential implications for the antipsychotic treatment." Eur Neuropsychopharmacol **10**(1): 1-4.

Altshuler, D., Pollara, V.J., Cowles, C.R., Van Etten, W.J., Baldwin, J., Linton, L. and Lander, E.S. (2000). "An SNP map of the human genome generated by reduced representation shotgun sequencing." Nature **407**(6803): 513-6.

American Psychiatric Association (2000) Diagnostic and Statistical Manual of Mental Disorders DSM-IV-RT. 4th edition, text revision.

Angst, J. (1995). "The epidemiology of depressive disorders." Eur Neuropsychopharmacol **5 Suppl**: 95-8.

Angst, J. and Sellaro, R. (2000). "Historical perspectives and natural history of bipolar disorder." Biol Psychiatry **48**(6): 445-57.

Antequera, F. and Bird, A. (1993). "Number of CpG islands and genes in human and mouse." Proc Natl Acad Sci U S A **90**(24): 11995-9.

Araki, H., Suemara, K., Gomita, Y. (2002). "Neuronal nicotinic receptor and psychiatric disorders: functional and behavioural effects of nicotine." Jpn J Pharmacol **88**(2): 133-138.

Arango, V., Underwood, M.D. and Mann, J.J. (1996). "Fewer pigmented locus coeruleus neurons in suicide victims: preliminary results." Biol Psychiatry **39**(2): 112-20.

Asherson, P., Mant, R., Williams, N., Cardno, A., Jones, L., Murphy, K., Collier, D.A., Nanko, S., Craddock, N., Morris, S., Muir, W., Blackwood, B., McGuffin, P. and Owen, M.J. (1998). "A study of chromosome 4p markers and dopamine D5 receptor gene in schizophrenia and bipolar disorder." Mol Psychiatry **3**(4): 310-20.

Badner, J.A. and Gershon, E.S. (2004). Meta-analysis of whole-genome linkage scansof bipolar disorder and schizophrenia." Mol Psychiatry **7**: 405-411

Banks, R.E., Dunn, M.J., Hochstrasser, D.F., Sanchez, J.C., Blackstock, W., Pappin, D.J. and Selby, P.J. (2000). "Proteomics: new perspectives, new biomedical opportunities." Lancet **356**(9243): 1749-56.

Baron, M. (1997). "Genetic linkage and bipolar affective disorder: progress and pitfalls." Mol Psychiatry **2**(3): 200-10.

Baron, M. (2001). "The search for complex disease genes: fault by linkage or fault by association?" Mol Psychiatry **6**(2): 143-9.

Batzer, M.A. and Deininger, P.L. (2002). "Alu repeats and human genomic diversity." Nat Rev Genet **3**(5): 370-9.

Baumann, B. and Bogerts, B. (2001). "Neuroanatomical studies on bipolar disorder." Br J Psychiatry Suppl **41**: s142-7.

Beaudoing, E., Freier, S., Wyatt, J.R., Claverie, J.M. and Gautheret, D. (2000). "Patterns of variant polyadenylation signal usage in human genes." Genome Res **10**(7): 1001-10.

Benazzi, F. $(2003)_a$. "Bipolar II disorder and major depressive disorder: continuity or discontinuity?" World J Biol Psychiatry **4**(4): 166-71.

Benazzi, F. $(2003)_b$. "Diagnosis of bipolar II disorder: a comparison of structured versus semistructured interviews." Prog Neuropsychopharmacol Biol Psychiatry **27**(6): 985-91.

Bernstein, H.G., Stanarius, A., Baumann, B., Henning, H., Krell, D., Danos, P., Falkai, P. and Bogerts, B. (1998). "Nitric oxide synthase-containing neurons in the human hypothalamus: reduced number of immunoreactive cells in the paraventricular nucleus of depressive patients and schizophrenics." Neuroscience **83**(3): 867-75.

Berrettini, W.H. $(2000)_a$. "Are schizophrenic and bipolar disorders related? A review of family and molecular studies." Biol Psychiatry **48**(6): 531-8.

Berrettini, W.H. $(2000)_b$. "Susceptibility loci for bipolar disorder: overlap with inherited vulnerability to schizophrenia." Biol Psychiatry **47**(3): 245-51.

Berridge, M.J., Downes, C.P. and Hanley, M.R. (1989). "Neural and developmental actions of lithium: a unifying hypothesis." Cell **59**(3): 411-9.

Berry, N., Jobanputra, V. and Pal, H. (2003). "Molecular genetics of schizophrenia: a critical review." J Psychiatry Neurosci **28**(6): 415-29.

Bertelsen, A. (2002). "Schizophrenia and related disorders: experience with current diagnostic systems." Psychopathology **35**(2-3): 89-93.

Bertelsen, A., Harvald, B. and Hauge, M. (1977). "A Danish twin study of manic-depressive disorders." Br J Psychiatry **130**: 330-51.

Bickmore, W.A. and Sumner, A.T. (1989). "Mammalian chromosome banding--an expression of genome organization." Trends Genet **5**(5): 144-8.

Blackwood, D.H., He, L., Morris, S.W., McLean, A., Whitton, C., Thomson, M., Walker, M.T., Woodburn, K., Sharp, C.M., Wright, A.F., Shibasaki, Y., St Clair, D.M., Porteous, D.J. and Muir, W.J. (1996). "A locus for bipolar affective disorder on chromosome 4p." Nat Genet **12**(4): 427-30.

Blackwood, D.H., Visscher, P.M. and Muir, W.J. (2001). "Genetic studies of bipolar affective disorder in large families." Br J Psychiatry **178**(Suppl 41): S134-6.

Bordo, D. and Argos, P. (1991). "Suggestions for "safe" residue substitutions in site-directed mutagenesis." J Mol Biol **217**(4): 721-9.

Botstein, D. and Risch, N. (2003). "Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease." Nat Genet **33 Suppl**: 228-37.

Breen, G., Sham, P., Li, T., Shaw, D., Collier, D.A. and St Clair, D. (1999). "Accuracy and sensitivity of DNA pooling with microsatellite repeats using capillary electrophoresis." Mol Cell Probes **13**(5): 359-65.

Breier, A., Su, T.P., Saunders, R., Carson, R.E., Kolachana, B.S., de Bartolomeis, A., Weinberger, D.R., Weisenfeld, N., Malhotra, A.K., Eckelman, W.C. and Pickar, D. (1997). "Schizophrenia is associated with elevated amphetamine-induced synaptic dopamine concentrations: evidence from a novel positron emission tomography method." Proc Natl Acad Sci U S A **94**(6): 2569-74.

Brett, D., Pospisil, H., Valcarcel, J., Reich, J. and Bork, P. (2002). "Alternative splicing and genome complexity." Nat Genet **30**(1): 29-30.

Bromet, E.J. and Fennig, S. (1999). "Epidemiology and natural history of schizophrenia." Biol Psychiatry **46**(7): 871-81.

Brown, G.W. and Birley, J.L. (1968). "Crises and life changes and the onset of schizophrenia." J Health Soc Behav **9**(3): 203-14.

Bunney, W.E., Bunney, B.G., Vawter, M.P., Tomita, H., Li, J., Evans, S.J., Choudary, P.V., Myers, R.M., Jones, E.G., Watson, S.J. and Akil, H. (2003). "Microarray technology: a review of new strategies to discover candidate vulnerability genes in psychiatric disorders." Am J Psychiatry **160**(4): 657-66.

Burge, C. and Karlin, S. (1997). "Prediction of complete gene structures in human genomic DNA." J Mol Biol **268**(1): 78-94.

Caceres, J.F. and Kornblihtt, A.R. (2002). "Alternative splicing: multiple control mechanisms and involvement in human disease." Trends Genet **18**(4): 186-93.

Cardno, A.G., Marshall, E.J., Coid, B., Macdonald, A.M., Ribchester, T.R., Davies, N.J., Venturi, P., Jones, L.A., Lewis, S.W., Sham, P.C., Gottesman, II, Farmer, A.E.,

McGuffin, P., Reveley, A.M. and Murray, R.M. (1999). "Heritability estimates for psychotic disorders: the Maudsley twin psychosis series." Arch Gen Psychiatry **56**(2): 162-8.

Cardon, L.R. and Bell, J.I. (2001). "Association study designs for complex diseases." Nat Rev Genet **2**(2): 91-9.

Carpenter, W.T., Strauss, J.S., Muleh, S. (1973). "Are there pathognomonic symptoms in schizophrenia? An empirical investigation of Schneider's first-rank symptoms." Arch Gen Psychiatry **28**:847-852.

Cartegni, L., Chew, S.L. and Krainer, A.R. (2002). "Listening to silence and understanding nonsense: exonic mutations that affect splicing." Nat Rev Genet **3**(4): 285-98.

Cheung, J., Estivill, X., Khaja, R., MacDonald, J.R., Lau, K., Tsui, L.C. and Scherer, S.W. (2003). "Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence." Genome Biol **4**(4): R25.

Clarke, A.S. and Schneider, M.L. (1993). "Prenatal stress has long-term effects on behavioral responses to stress in juvenile rhesus monkeys." Dev Psychobiol **26**(5): 293-304.

Clements, A.D. (1992). "The incidence of attention deficit-hyperactivity disorder in children whose mothers experienced extreme psychological stress." Georgia Ed Res **91**:1-14.

Claverie, J.M. (2001). "Gene number. What if there are only 30,000 human genes?" Science **291**(5507): 1255-7.

Collier, D.A. and Li, T. (2003). "The genetics of schizophrenia: glutamate not dopamine?" Eur J Pharmacol **480**(1-3): 177-84.

Cooper, J.R., Bloom, F.E. and Roth, R.H. (1996). "The biochemical basis of neuropharmacology" 7th edition. Oxford Uni Press.

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

Corvin, A.P., Morris, D.W., McGhee, K., Schwaiger, S., Scully, P., Quinn, J., Meagher, D., Clair, D.S., Waddington, J.L. and Gill, M. (2004). "Confirmation and refinement of an 'at-risk' haplotype for schizophrenia suggests the EST cluster, Hs.97362, as a potential susceptibility gene at the Neuregulin-1 locus." Mol Psychiatry **9**(2): 208-13.

Coyle, J.T. and Manji, H.K. (2002). "Getting balance: drugs for bipolar disorder share target." Nat Med **8**(6): 557-8.

Craddock, N. and Jones, I. (1999). "Genetics of bipolar disorder." J Med Genet **36**(8): 585-94.

Daly, M.J., Rioux, J.D., Schaffner, S.F., Hudson, T.J. and Lander, E.S. (2001). "High-resolution haplotype structure in the human genome." Nat Genet **29**(2): 229-32.

Daniels, J., Holmans, P., Williams, N., Turic, D., McGuffin, P., Plomin, R. and Owen, M.J. (1998). "A simple method for analyzing microsatellite allele image patterns generated from DNA pools and its application to allelic association studies." Am J Hum Genet **62**(5): 1189-97.

De Haan, L. and Bakker, J.M. (2004). "Overview of neuropathological theories of schizophrenia: from degeneration to progressive developmental disorder." Psychopathology **37**(1): 1-7.

Detera-Wadleigh, S.D., Badner, J.A., Berrettini, W.H., Yoshikawa, T., Goldin, L.R., Turner, G., Rollins, D.Y., Moses, T., Sanders, A.R., Karkera, J.D., Esterling, L.E., Zeng, J., Ferraro, T.N., Guroff, J.J., Kazuba, D., Maxwell, M.E., Nurnberger, J.I., Jr. and Gershon, E.S. (1999). "A high-density genome scan detects evidence for a bipolar-disorder susceptibility locus on 13q32 and other potential loci on 1q32 and 18p11.2." Proc Natl Acad Sci U S A **96**(10): 5604-9.

Don, R.H., Cox, P.T., Wainwright, B.J., Baker, K. and Mattick, J.S. (1991). "'Touchdown' PCR to circumvent spurious priming during gene amplification." Nucleic Acids Res **19**(14): 4008.

Drevets, W.C., Price, J.L., Simpson, J.R., Jr., Todd, R.D., Reich, T., Vannier, M. and Raichle, M.E. (1997). "Subgenual prefrontal cortex abnormalities in mood disorders." Nature **386**(6627): 824-7.

Dubertret, C., Gorwood, P., Ades, J., Feingold, J., Schwartz, J. and Sokoloff, P. (1998). "Meta-analysis of DRD3 gene and scizophrenia." Am J Med Genet **81**:318-322.

Dudbridge, F. (2003). "Pedigree disequilibrium tests for multilocus haplotypes." Genet Epidemiol **25**(2): 115-21.

Duman, R.S., Heninger, G.R. and Nestler, E.J. (1997). "A molecular and cellular theory of depression." Arch Gen Psychiatry **54**(7): 597-606.

Eaves, I.A., Merriman, T.R., Barber, R.A., Nutland, S., Tuomilehto-Wolf, E., Tuomilehto, J., Cucca, F. and Todd, J.A. (2000). "The genetically isolated populations of Finland and sardinia may not be a panacea for linkage disequilibrium mapping of common disease genes." Nat Genet **25**(3): 320-3.

Egeland, J.A., Gerhard, D.S., Pauls, D.L., Sussex, J.N., Kidd, K.K., Allen, C.R., Hostetter, A.M. and Housman, D.E. (1987). "Bipolar affective disorders linked to DNA markers on chromosome 11." Nature **325**(6107): 783-7.

Elston, R.C. (1998)."Methods of linkage analysis - and the assumptions underlying them." Am J Hum Genet **63**: 931-934.

Evans, K.L., Le Hellard, S., Morris, S.W., Lawson, D., Whitton, C., Semple, C.A., Fantes, J.A., Torrance, H.S., Malloy, M.P., Maule, J.C., Humphray, S.J., Ross, M.T., Bentley, D.R., Muir, W.J., Blackwood, D.H. and Porteous, D.J. $(2001)_a$. "A 6.9-Mb

high-resolution BAC/PAC contig of human 4p15.3-p16.1, a candidate region for bipolar affective disorder." Genomics **71**(3): 315-23.

Evans, K.L., Muir, W.J., Blackwood, D.H. and Porteous, D.J. $(2001)_b$. "Nuts and bolts of psychiatric genetics: building on the Human Genome Project." Trends Genet **17**(1): 35-40.

Ewald, H., Degn, B., Mors, O. and Kruse, T.A. (1998). "Support for the possible locus on chromosome 4p16 for bipolar affective disorder." Mol Psychiatry **3**(5): 442-8.

Fischer, M., Harvald, B. and Hauge, M. (1969). "A Danish twin study of schizophrenia." Br J Psychiatry **115**(526): 981-90.

Folz, R.J. and Crapo, J.D. (1994). "Extracellular superoxide dismutase (SOD3): tissue-specific expression, genomic characterization, and computer-assisted sequence analysis of the human EC SOD gene." Genomics **22**(1): 162-71.

Fowles, D.C. (1992). "Schizophrenia: diathesis-stress revisited." Annu Rev Psychol **43**: 303-36.

Frangou, S. and Murray, R. (1997). Schizophrenia, Martin Dunitz.

Fredriksson, R., Gloriam, D.E., Hoglund, P.J., Lagerstrom, M.C. and Schioth, H.B. (2003). "There exist at least 30 human G-protein-coupled receptors with long Ser/Thr-rich N-termini." Biochem Biophys Res Commun **301**(3): 725-34.

Freedman, R., Adler, L.E., Waldo, M.C., Pachtman, E. and Franks, R.D. (1983). "Neurophysiological evidence for a defect in inhibitory pathways in schizophrenia: comparison of medicated and drug-free patients." Biol Psychiatry **18**(5): 537-51.

Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A. and Faggart, M. (2002) "The structure of haplotype blocks in the human genome." Science. **296**: 2225-2229.

Gershon, E.S., DeLisi, L.E., Hamovit, J., Nurnberger, J.I., Jr., Maxwell, M.E., Schreiber, J., Dauphinais, D., Dingman, C.W., 2nd and Guroff, J.J. (1988). "A controlled family study of chronic psychoses. Schizophrenia and schizoaffective disorder." Arch Gen Psychiatry **45**(4): 328-36.

Ginns, E.I., St Jean, P., Philibert, R.A., Galdzicka, M., Damschroder-Williams, P., Thiel, B., Long, R.T., Ingraham, L.J., Dalwaldi, H., Murray, M.A., Ehlert, M., Paul, S., Remortel, B.G., Patel, A.P., Anderson, M.C., Shaio, C., Lau, E., Dymarskaia, I., Martin, B.M., Stubblefield, B., Falls, K.M., Carulli, J.P., Keith, T.P., Fann, C.S., Paul, S.M. and et al. (1998). "A genome-wide search for chromosomal loci linked to mental health wellness in relatives at high risk for bipolar affective disorder among the Old Order Amish." Proc Natl Acad Sci U S A **95**(26): 15531-6.

Glatt, C.E. and Freimer, N.B. (2002). "Association analysis of candidate genes for neuropsychiatric disease: the perpetual campaign." Trends Genet **18**(6): 307-12.

Green, P.M, Saad, S., Lewis, C.M. and Giannelli, F. (1999). "Mutation rates in humans:

overall and sex-specific rates obtained form a population study of hemophilia B." Am J Hum Genet **65**:1572-1579.

Gondo, Y., Okada, T., Matsuyama, N., Saitoh, Y., Yanagisawa, Y. and Ikeda, J.E. (1998). "Human megasatellite DNA RS447: copy-number polymorphisms and interspecies conservation." Genomics **54**(1): 39-49.

Gordon D., Finch S.J., Nothnagel M., and Ott J. (2002) Power and sample size calculations for case-control genetic association tests when errors present: application to single nucleotide polymorphisms. Human Heredity 54:22-33

Guigo, R., Agarwal, P., Abril, J.F., Burset, M. and Fickett, J.W. (2000). "An assessment of gene prediction accuracy in large DNA sequences." Genome Res **10**(10): 1631-42.

Halliday, G.M. (2001). "A review of the neuropathology of schizophrenia." Clin Exp Pharmacol Physiol **28**(1-2): 64-5.

Harrison, P.J. (1999). "The neuropathology of schizophrenia. A critical review of the data and their interpretation." Brain **122 (Pt 4)**: 593-624.

Harrison, P.J. (2002). "The neuropathology of primary mood disorder." Brain **125**(Pt 7): 1428-49.

Hashimoto, R., Straub, R.E., Weickert, C.S., Hyde, T.M., Kleinman, J.E. and Weinberger, D.R. (2003). "Expression analysis of neuregulin-1 in the dorsolateral prefrontal cortex in schizophrenia." Mol Psychiatry.

Hawkins, R.D. (1996). "NO honey, I don't remember." Neuron **16**(3): 465-7.

Hendrickson, D.J., Fisher, J.H., Jones, C. and Ho, Y.S. (1990). "Regional localization of human extracellular superoxide dismutase gene to 4pter-q21." Genomics **8**(4): 736-8.

Herman, J.P., Prewitt, C.M. and Cullinan, W.E. (1996). "Neuronal circuit regulation of the hypothalamo-pituitary-adrenocortical stress axis." Crit Rev Neurobiol **10**(3-4): 371-94.

Hjalmarsson, K., Marklund, S.L., Engstrom, A. and Edlund, T. (1987). "Isolation and sequence of complementary DNA encoding human extracellular superoxide dismutase." Proc Natl Acad Sci U S A **84**(18): 6340-4.

Holmes, T.H. and Rahe, R.H. (1967). "The Social Readjustment Rating Scale." J Psychosom Res **11**(2): 213-8.

Hugot, J.P., Chamaillard, M., Zouali, H., Lesage, S., Cezard, J.P., Belaiche, J., Almer, S., Tysk, C., O'Morain, C.A., Gassull, M., Binder, V., Finkel, Y., Cortot, A., Modigliani, R., Laurent-Puig, P., Gower-Rousseau, C., Marcey, J., Colombel, J.F., Sahbatou, M. and Thomas, G (2001). "Association of the NOD2 leucine rich repeat variants with susceptibility to Crohn's disease." Nature **411**: 599-603.

Huttunen, M.O. and Niskanen, P. (1978). "Prenatal loss of father and psychiatric disorders." Arch Gen Psychiatry **35**(4): 429-31.

Ikonomov, O.C. and Manji, H.K. (1999). "Molecular mechanisms underlying mood stabilization in manic-depressive illness: the phenotype challenge." Am J Psychiatry **156**(10): 1506-14.

The International HapMap Consortium (2003). "The International HapMap Project." Nature **426**(6968): 789-96.

Iwata, N., Suzuki, T., Ikeda, M., Kitajima, T., Yamanouchi, Y., Inada, T. and Ozaki, N. (2004). "No association with the neuregulin 1 haplotype to Japanese schizophrenia." Mol Psychiatry **9**(2): 126-7.

Janca, A. (2001). "Reliability of DSM-IV axis V scales." Am J Psychiatry **158**(11): 1935-7.

Janin, J. (1979). "Surface and inside volumes in globular proteins." Nature **277**(5696): 491-2.

Jeffreys, A.J., Kauppi, L. and Neumann, R. (2001). "Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex." Nat Genet **29**(2): 217-22.

Jeffreys, A.J. and May, C.A. (2004). "Intense and highly localized gene conversion activity in human meiotic crossover hot spots." Nat Genet **36**(2): 151-6.

Johnson, G.C., Esposito, L., Barratt, B.J., Smith, A.N., Heward, J., Di Genova, G., Ueda, H., Cordell, H.J., Eaves, I.A., Dudbridge, F., Twells, R.C., Payne, F., Hughes, W., Nutland, S., Stevens, H., Carr, P., Tuomilehto-Wolf, E., Tuomilehto, J., Gough, S.C., Clayton, D.G. and Todd, J.A. (2001). "Haplotype tagging for the identification of common disease genes." Nat Genet **29**(2): 233-7.

Jope, R.S. $(1999)_a$. "Anti-bipolar therapy: mechanism of action of lithium." Mol Psychiatry **4**(2): 117-28.

Jope, R.S. $(1999)_b$. "A bimodal model of the mechanism of action of lithium." Mol Psychiatry **4**(1): 21-5.

Kendler, K.S. and Diehl, S.R. (1993). "The genetics of schizophrenia: a current, genetic-epidemiologic perspective." Schizophr Bull **19**(2): 261-85.

Kety, S.S. (1988). "Schizophrenic illness in the families of schizophrenic adoptees: findings from the Danish national sample." Schizophr Bull **14**(2): 217-22.

Kety, S.S., Rosenthal, D., Wender, P.H. and Schulsinger, F. (1971). "Mental illness in the biological and adoptive families of adpoted schizophrenics." Am J Psychiatry **128**(3): 302-6.

Kitayama, I., Yaga, T., Kayahara, T., Nakano, K., Murase, S., Otani, M. and Nomura, J. (1997). "Long-term stress degenerates, but imipramine regenerates, noradrenergic axons in the rat cerebral cortex." Biol Psychiatry **42**(8): 687-96.

Kluppel, M., Nagle, D.L., Bucan, M. and Bernstein, A. (1997). "Long-range genomic rearrangements upstream of Kit dysregulate the development pattern of Kit expression in W57 and Wbanded mice and interfere with distinct steps in melanocyte development." Development **124**(1): 65-77.

Kogi, M., Fukushige, S., Lefevre, C., Hadano, S. and Ikeda, J.E. (1997). "A novel tandem repeat sequence located on human chromosome 4p: isolation and characterization." Genomics **42**(2): 278-83.

Kopp, J. and Schwede, T. (2004). "Automated protein structure homology modelling: a progress report." Pharmacogenomics **5**(4): 405-416

Kruglyak, L. (1999)$_a$. "Genetic isolates: separate but equal?" Proc Natl Acad Sci U S A **96**(4): 1170-2.

Kruglyak, L. (1999)$_b$. "Prospects for whole-genome linkage disequilibrium mapping of common disease genes." Nat Genet **22**(2): 139-44.

Kyte, J. and Doolittle, R.F. (1982). "A simple method for displaying the hydropathic character of a protein." J Mol Biol **157**(1): 105-32.

Lambert, J.C., Mann, D., Goumidi, L., Harris, J., Amouyel, P., Iwatsubo, T., Lendeon, C. and Chartier-Harlin, M.C. (2001). "Effect of the APOE promoter polymorphisms on cerebral amyloid peptide deposition in Alzheimer's disease." Lancet **357**: 608-609.

Lander, E. and Kruglyak, L. (1995). "Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results." Nat Genet **11**(3): 241-7.

Lander, E.S.Linton, L.M.Birren, B.Nusbaum, C.Zody, M.C.Baldwin, J.Devon, K.Dewar, K.Doyle, M.FitzHugh, W.Funke, R.Gage, D.Harris, K.Heaford, A.Howland, J.Kann, L.Lehoczky, J.LeVine, R.McEwan, P.McKernan, K.Meldrim, J.Mesirov, J.P.Miranda, C.Morris, W.Naylor, J.Raymond, C.Rosetti, M.Santos, R.Sheridan, A.Sougnez, C.Stange-Thomann, N.Stojanovic, N.Subramanian, A.Wyman, D.Rogers, J.Sulston, J.Ainscough, R.Beck, S.Bentley, D.Burton, J.Clee, C.Carter, N.Coulson, A.Deadman, R.Deloukas, P.Dunham, A.Dunham, I.Durbin, R.French, L.Grafham, D.Gregory, S.Hubbard, T.Humphray, S.Hunt, A.Jones, M.Lloyd, C.McMurray, A.Matthews, L.Mercer, S.Milne, S.Mullikin, J.C.Mungall, A.Plumb, R.Ross, M.Shownkeen, R.Sims, S.Waterston, R.H.Wilson, R.K.Hillier, L.W.McPherson, J.D.Marra, M.A.Mardis, E.R.Fulton, L.A.Chinwalla, A.T.Pepin, K.H.Gish, W.R.Chissoe, S.L.Wendl, M.C.Delehaunty, K.D.Miner, T.L.Delehaunty, A.Kramer, J.B.Cook, L.L.Fulton, R.S.Johnson, D.L.Minx, P.J.Clifton, S.W.Hawkins, T.Branscomb, E.Predki, P.Richardson, P.Wenning, S.Slezak, T.Doggett, N.Cheng, J.F.Olsen, A.Lucas, S.Elkin, C.Uberbacher, E.Frazier, M.Gibbs, R.A.Muzny, D.M.Scherer, S.E.Bouck, J.B.Sodergren, E.J.Worley, K.C.Rives, C.M.Gorrell, J.H.Metzker, M.L.Naylor, S.L.Kucherlapati, R.S.Nelson, D.L.Weinstock, G.M.Sakaki, Y.Fujiyama, A.Hattori, M.Yada, T.Toyoda, A.Itoh, T.Kawagoe, C.Watanabe, H.Totoki, Y.Taylor, T.Weissenbach, J.Heilig, R.Saurin, W.Artiguenave, F.Brottier, P.Bruls, T.Pelletier, E.Robert, C.Wincker, P.Smith, D.R.Doucette-Stamm, L.Rubenfield, M.Weinstock, K.Lee, H.M.Dubois, J.Rosenthal, A.Platzer, M.Nyakatura, G.Taudien, S.Rump, A.Yang, H.Yu, J.Wang, J.Huang, G.Gu, J.Hood, L.Rowen, L.Madan, A.Qin, S.Davis, R.W.Federspiel,

N.A.Abola, A.P.Proctor, M.J.Myers, R.M.Schmutz, J.Dickson, M.Grimwood, J.Cox, D.R.Olson, M.V.Kaul, R.Shimizu, N.Kawasaki, K.Minoshima, S.Evans, G.A.Athanasiou, M.Schultz, R.Roe, B.A.Chen, F.Pan, H.Ramser, J.Lehrach, H.Reinhardt, R.McCombie, W.R.de la Bastide, M.Dedhia, N.Blocker, H.Hornischer, K.Nordsiek, G.Agarwala, R.Aravind, L.Bailey, J.A.Bateman, A.Batzoglou, S.Birney, E.Bork, P.Brown, D.G.Burge, C.B.Cerutti, L.Chen, H.C.Church, D.Clamp, M.Copley, R.R.Doerks, T.Eddy, S.R.Eichler, E.E.Furey, T.S.Galagan, J.Gilbert, J.G.Harmon, C.Hayashizaki, Y.Haussler, D.Hermjakob, H.Hokamp, K.Jang, W.Johnson, L.S.Jones, T.A.Kasif, S.Kaspryzk, A.Kennedy, S.Kent, W.J.Kitts, P.Koonin, E.V.Korf, I.Kulp, D.Lancet, D.Lowe, T.M.McLysaght, A.Mikkelsen, T.Moran, J.V.Mulder, N.Pollara, V.J.Ponting, C.P.Schuler, G.Schultz, J.Slater, G.Smit, A.F.Stupka, E.Szustakowski, J.Thierry-Mieg, D.Thierry-Mieg, J.Wagner, L.Wallis, J.Wheeler, R.Williams, A.Wolf, Y.I.Wolfe, K.H.Yang, S.P.Yeh, R.F.Collins, F.Guyer, M.S.Peterson, J.Felsenfeld, A.Wetterstrand, K.A.Patrinos, A.Morgan, M.J.Szustakowki, J.de Jong, P.Catanese, J.J.Osoegawa, K.Shizuya, H.Choi, S.Chen, Y.J. (2001). "Initial sequencing and analysis of the human genome." Nature **409**(6822): 860-921.

Larsen, F., Gundersen, G., Lopez, R. and Prydz, H. (1992). "CpG islands as gene markers in the human genome." Genomics **13**(4): 1095-107.

Le Hellard, S., Ballereau, S.J., Visscher, P.M., Torrance, H.S., Pinson, J., Morris, S.W., Thomson, M.L., Semple, C.A., Muir, W.J., Blackwood, D.H., Porteous, D.J. and Evans, K.L. (2002). "SNP genotyping on pooled DNAs: comparison of genotyping technologies and a semi automated method for data storage and analysis." Nucleic Acids Res **30**(15): e74.

Lee, D.K., Lynch, K.R., Nguyen, T., Im, D.S., Cheng, R., Saldivia, V.R., Liu, Y., Liu, I.S., Heng, H.H., Seeman, P., George, S.R., O'Dowd, B.F. and Marchese, A. (2000). "Cloning and characterization of additional members of the G protein-coupled receptor family." Biochim Biophys Acta **1490**(3): 311-23.

Lee, D.K., Nguyen, T., Lynch, K.R., Cheng, R., Vanti, W.B., Arkhitko, O., Lewis, T., Evans, J.F., George, S.R. and O'Dowd, B.F. (2001). "Discovery and mapping of ten novel G protein-coupled receptor genes." Gene **275**(1): 83-91.

Leckman, J.F., Weissman, M.M., Prusoff, B.A., Caruso, K.A., Merikangas, K.R., Pauls, D.L. and Kidd, K.K. (1984). "Subtypes of depression: family study perspective." Arch Gen Psychiatry **41**: 833-838.

Lenox, R.H. and Hahn, C.G. (2000). "Overview of the mechanism of action of lithium in the brain: fifty-year update." J Clin Psychiatry **61 Suppl 9**: 5-15.

Lerer, B., Segman, R.H., Hamdan, A., Kanyas, K., Karni, O., Kohn, Y., Korner, M., Lanktree, M., Kaadan, M., Turetsky, N., Yakir, A., Kerem, B. and Macciardi, F. (2003). "Genome scan of Arab Israeli families maps a schizophrenia susceptibility gene to chromosome 6q23 and supports a locus at chromosome 10q24." Mol Psychiatry **8**(5): 488-98.

Levin, E.D., Brady, T.C., Hochrein, E.C., Oury, T.D., Jonsson, L.M., Marklund, S.L. and Crapo, J.D. (1998). "Molecular manipulations of extracellular superoxide dismutase: functional importance for learning." Behav Genet **28**(5): 381-90.

Levin, E.D., Brucato, F.H. and Crapo, J.D. (2000). "Molecular overexpression of extracellular superoxide dismutase increases the dependency of learning and memory performance on motivational state." Behav Genet **30**(2): 95-100.

Lewy, A.J., Kern, H.A., Rosenthal, N.E. and Wehr, T.A. (1982). "Bright artificial light treatment of a manic-depressive patient with a seasonal mood cycle." Am J Psychiatry **139**(11): 1496-8.

Li, L.H., Li, J.C., Lin, Y.F., Lin, C.Y., Chen, C.Y. and Tsai, S.F. $(2004)_a$. "Genomic shotgun array: a procedure linking large-scale DNA sequencing with regional transcript mapping." Nucleic Acids Res **32**(3): e27.

Li, T., Stefansson, H., Gudfinnsson, E., Cai, G., Liu, X., Murray, R.M., Steinthorsdottir, V., Januel, D., Gudnadottir, V.G., Petursson, H., Ingason, A., Gulcher, J.R., Stefansson, K. and Collier, D.A. $(2004)_b$. "Identification of a novel neuregulin 1 at-risk haplotype in Han schizophrenia Chinese patients, but no association with the Icelandic/Scottish risk haplotype." Mol Psychiatry.

Lieberman, J.A., Sheitman, B.B. and Kinon, B.J. (1997). "Neurochemical sensitization in the pathophysiology of schizophrenia: deficits and dysfunction in neuronal regulation and plasticity." Neuropsychopharmacology **17**(4): 205-29.

Lloyd, A.J., Harrison, C.L., Ferrier, I.N. and Young, A.H. (2003). "The pharmacological treatment of bipolar affective disorder: practice is improving but could still be better." J Psychopharmacol **17**(2): 230-3.

Lohmueller, K.E., Pearce, C.L., Pike, M., Lander, E.S. and Hirschhorn, J.N. (2003). "Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease." Nat Genet **33**(2): 177-82.

Lonjou, C., Collins, A. and Morton, N.E. (1999). "Allelic association between marker loci." Proc Natl Acad Sci U S A **96**(4): 1621-6.

Lopez, J.F., Chalmers, D.T., Little, K.Y. and Watson, S.J. (1998). "A.E. Bennett Research Award. Regulation of serotonin1A, glucocorticoid, and mineralocorticoid receptor in rat and human hippocampus: implications for the neurobiology of depression." Biol Psychiatry **43**(8): 547-73.

Maier, W., Lichtermann, D., Minges, J., Hallmayer, J., Heun, R., Benkert, O. and Levinson, D.F. (1993). "Continuity and discontinuity of affective disorders and schizophrenia. Results of a controlled family study." Arch Gen Psychiatry **50**(11): 871-83.

Malhotra, A.K. and Goldman, D. (1999). "Benefits and pitfalls encountered in psychiatric genetic association studies." Biol Psychiatry **45**: 544-550.

Manji, H.K. and Lenox, R.H. (2000). "Signaling: cellular insights into the pathophysiology of bipolar disorder." Biol Psychiatry **48**(6): 518-30.

Manji, H.K., Moore, G.J., Rajkowska, G. and Chen, G. (2000). "Neuroplasticity and cellular resilience in mood disorders." Mol Psychiatry **5**(6): 578-93.

Manning, J.S., Connor, P.D. and Sahai, A. (1998). "The bipolar spectrum: a review of current concepts and implications for the management of depression in primary care." Arch Fam Med **7**(1): 63-71.

McGue, M., Gottesman, II and Rao, D.C. (1985). "Resolving genetic models for the transmission of schizophrenia." Genet Epidemiol **2**(1): 99-110.

Mendlewicz, J. and Rainer, J.D. (1977). "Adoption study supporting genetic transmission in manic--depressive illness." Nature **268**(5618): 327-9.

Meier. A. (1985). "Child psychiatric sequelae of maternal war stress." Acta Psychiatr Scand **72**: 505-511.

Merikangas, K.R. and Risch, N. (2003). "Will the genomics revolution revolutionize psychiatry?" Am J Psychiatry **160**(4): 625-35.

Mighell, A.J., Smith, N.R., Robinson, P.A. and Markham, A.F. (2000). "Vertebrate pseudogenes." FEBS Lett **468**(2-3): 109-14.

Millar, J.K., Christie, S. and Porteous, D.J. (2003). "Yeast two-hybrid screens implicate DISC1 in brain development and function." Biochem Biophys Res Commun **311**(4): 1019-25.

Millar, J.K., Wilson-Annan, J.C., Anderson, S., Christie, S., Taylor, M.S., Semple, C.A., Devon, R.S., Clair, D.M., Muir, W.J., Blackwood, D.H. and Porteous, D.J. (2000). "Disruption of two novel genes by a translocation co-segregating with schizophrenia." Hum Mol Genet **9**(9): 1415-23.

Moller, H.J. (2003). "Bipolar disorder and schizophrenia: distinct illnesses or a continuum?" J Clin Psychiatry **64**(suppl6): 23-27.

Muir, W.J., Thomson, M.L., McKeon, P., Mynett-Johnson, L., Whitton, C., Evans, K.L., Porteous, D.J. and Blackwood, D.H. (2001). "Markers close to the dopamine D5 receptor gene (DRD5) show significant association with schizophrenia but not bipolar disorder." Am J Med Genet **105**(2): 152-8.

Mynett-Johnson, L., McKeon, P. (1996). "The Molecular Genetics of Affective Disorders: An Overview." Irish Journal of Psychological Medicine **13**(4): 155-61.

Nakao, M. (2001). "Epigenetics: interaction of DNA methylation and chromatin." Gene **278**(1-2): 25-31.

Nissim-Rafinia, M. and Kerem, B. (2002). "Splicing regulation as a potential genetic modifier." Trends Genet **18**(3): 123-7.

Nordborg, M. and Tavare, S. (2002). "Linkage disequilibrium: what history has to tell us." Trends Genet **18**(2): 83-90.

Norman, R.M.G.and Malla, A.K. (1993). "Stressful Life Events and Schizophrenia I: A Review of the Research." British Journal of Psychiatry **162**: 161-166.

Nyholt, D.R. (2001). "Genetic case-control association studies - correction for multiple testing." Hum Genet **109**: 564-565.

O'Donovan, M.C. and Owen, M.J. (1999). "Candidate-gene association studies of schizophrenia." Am J Hum Genet **65**(3): 587-92.

Ogura, Y., Bonen, D.K., Inohara, N., Nicolae, D.L., Chen, F.F., Ramos, R., Britton, H., Moran, T., Karaliuskas, R., Duerr, R.H., Achkar, J.P., Brant, S.R., Bayless, T.M., Kirschner, B.S., Hanauer, S.B., Nunex, G. and Cho, J.H. (2001). "A frameshift mutation in NOD2 is associated with susceptibility to Crohn's disease." Nature **411**: 603-606.

Ota, T., Suzuki, Y., Nishikawa, T., Otsuki, T., Sugiyama, T., Irie, R., Wakamatsu, A., Hayashi, K., Sato, H., Nagai, K., Kimura, K., Makita, H., Sekine, M., Obayashi, M., Nishi, T., Shibahara, T., Tanaka, T., Ishii, S., Yamamoto, J., Saito, K., Kawai, Y., Isono, Y., Nakamura, Y., Nagahari, K., Murakami, K., Yasuda, T., Iwayanagi, T., Wagatsuma, M., Shiratori, A., Sudo, H., Hosoiri, T., Kaku, Y., Kodaira, H., Kondo, H., Sugawara, M., Takahashi, M., Kanda, K., Yokoi, T., Furuya, T., Kikkawa, E., Omura, Y., Abe, K., Kamihara, K., Katsuta, N., Sato, K., Tanikawa, M., Yamazaki, M., Ninomiya, K., Ishibashi, T., Yamashita, H., Murakawa, K., Fujimori, K., Tanai, H., Kimata, M., Watanabe, M., Hiraoka, S., Chiba, Y., Ishida, S., Ono, Y., Takiguchi, S., Watanabe, S., Yosida, M., Hotuta, T., Kusano, J., Kanehori, K., Takahashi-Fujii, A., Hara, H., Tanase, T.O., Nomura, Y., Togiya, S., Komai, F., Hara, R., Takeuchi, K., Arita, M., Imose, N., Musashino, K., Yuuki, H., Oshima, A., Sasaki, N., Aotsuka, S., Yoshikawa, Y., Matsunawa, H., Ichihara, T., Shiohata, N., Sano, S., Moriya, S., Momiyama, H., Satoh, N., Takami, S., Terashima, Y., Suzuki, O., Nakagawa, S., Senoh, A., Mizoguchi, H., Goto, Y., Shimizu, F., Wakebe, H., Hishigaki, H., Watanabe, T., Sugiyama, A., Takemoto, M., Kawakami, B., Watanabe, K., Kumagai, A., Itakura, S., Fukuzumi, Y., Fujimori, Y., Komiyama, M., Tashiro, H., Tanigami, A., Fujiwara, T., Ono, T., Yamada, K., Fujii, Y., Ozaki, K., Hirao, M., Ohmori, Y., Kawabata, A., Hikiji, T., Kobatake, N., Inagaki, H., Ikema, Y., Okamoto, S., Okitani, R., Kawakami, T., Noguchi, S., Itoh, T., Shigeta, K., Senba, T., Matsumura, K., Nakajima, Y., Mizuno, T., Morinaga, M., Sasaki, M., Togashi, T., Oyama, M., Hata, H., Komatsu, T., Mizushima-Sugano, J., Satoh, T., Shirai, Y., Takahashi, Y., Nakagawa, K., Okumura, K., Nagase, T., Nomura, N., Kikuchi, H., Masuho, Y., Yamashita, R., Nakai, K., Yada, T., Ohara, O., Isogai, T. and Sugano, S. (2004). "Complete sequencing and characterization of 21,243 full-length human cDNAs." Nat Genet **36**(1): 40-5.

Ott, J. (2000). "Predicting the range of linkage disequilibrium." Proc Natl Acad Sci U S A **97**(1): 2-3.

Oury, T.D., Card, J.P. and Klann, E. (1999). "Localization of extracellular superoxide dismutase in adult mouse brain." Brain Res **850**(1-2): 96-103.

Oury, T.D., Day, B.J. and Crapo, J.D. (1996). "Extracellular superoxide dismutase: a regulator of nitric oxide bioavailability." Lab Invest **75**(5): 617-36.

Owen, M.J., Cardno, A.G. and O'Donovan, M.C. (2000). "Psychiatric genetics: back to the future." Mol Psychiatry **5**(1): 22-31.

Papageorgiou, C., Grapsa, E., Christodoulou, N.G., Zerefos, N., Stamatelopoulos, S. and Christodoulou, G.N. (2001). "Association of serum nitric oxide levels with depressive symptoms: a study with end-stage renal failure patients." Psychother Psychosom **70**(4): 216-20.

Pariante, C.M. and Miller, A.H. (2001). "Glucocorticoid receptors in major depression: relevance to pathophysiology and treatment." Biol Psychiatry **49**(5): 391-404.

Patil, N., Berno, A.J., Hinds, D.A., Barrett, W.A., Doshi, J.M., Hacker, C.R., Kautzer, C.R., Lee, D.H., Marjoribanks, C., McDonough, D.P., Nguyen, B.T., Norris, M.C., Sheehan, J.B., Shen, N., Stern, D., Stokowski, R.P., Thomas, D.J., Trulson, M.O., Vyas, K.R., Frazer, K.A., Fodor, S.P. and Cox, D.R. (2001). "Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21." Science **294**(5547): 1719-23.

Pennisi, E. (2003). "Human genome. A low number wins the GeneSweep Pool." Science **300**(5625): 1484.

Petronis, A. (2000). "The genes for major psychosis: aberrant sequence or regulation?" Neuropsychopharmacology **23**(1): 1-12.

Petty, F. (1995). "GABA and mood disorders: a brief review and hypothesis." J Affect Disord **34**(4): 275-81.

Phillips, M.S., Lawrence, R., Sachidanandam, R., Morris, A.P., Balding, D.J., Donaldson, M.A., Studebaker, J.F., Ankener, W.M., Alfisi, S.V., Kuo, F.S., Camisa, A.L., Pazorov, V., Scott, K.E., Carey, B.J., Faith, J., Katari, G., Bhatti, H.A., Cyr, J.M., Derohannessian, V., Elosua, C., Forman, A.M., Grecco, N.M., Hock, C.R., Kuebler, J.M., Lathrop, J.A., Mockler, M.A., Nachtman, E.P., Restine, S.L., Varde, S.A., Hozza, M.J., Gelfand, C.A., Broxholme, J., Abecasis, G.R., Boyce-Jacino, M.T. and Cardon, L.R. (2003). "Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots." Nat Genet **33**(3): 382-7.

Pope, H.G., Jr., Jonas, J.M., Cohen, B.M. and Lipinski, J.F. (1982). "Failure to find evidence of schizophrenia in first-degree relatives of schizophrenic probands." Am J Psychiatry **139**(6): 826-8.

Potash, J.B., Willour, V.L., Chiu, Y.F., Simpson, S.G., Mackinnon, D.F., Pearlson, G.D., DePaulo, J.R. and Mcinnis, M.G. (2001). "The familial aggregation of psychotic symptoms in bipolar pedigrees." Am J Psychiatry **158**: 1258-1264.

Potash, J.B., Zandi, P.P., Willour, V.L., Lan, T.H., Huo, Y., Avramopoulos, D., Shugart, Y.Y., Mackinnon, D.F., Simpson, S.G., McMahon, F.J., DePaulo, J.R. and McInnis, M.G. (2003). "Suggestive linkage to chromosomal regions 13q31 and 22q12 in families with psychotic bipolar disorder." Am J Psychiatry **160**: 680-686.

Pritchard, J.K. and Cox, N.J. (2002) "The allelic architecture of human disease genes:

common disease-common variant...or not?" Hum Mol Gen **11**(20): 2417-2423.

Reich, D.E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P.C., Richter, D.J., Lavery, T., Kouyoumjian, R., Farhadian, S.F., Ward, R. and Lander, E.S. (2001). "Linkage disequilibrium in the human genome." Nature **411**(6834): 199-204.

Reich, D.E., and Lander, E.S. (2001) "On the allelic spectrum of human disease." Trends in Genet. **17**(9): 502-510.

Ressler, K.J. and Nemeroff, C.B. (1999). "Role of norepinephrine in the pathophysiology and treatment of mood disorders." Biol Psychiatry **46**(9): 1219-33.

Riemann, D., Berger, M. and Voderholzer, U. (2001). "Sleep and depression--results from psychobiological studies: an overview." Biol Psychol **57**(1-3): 67-103.

Riley, J.H., Butler, R., Ogilvie, D., Finniear, R., Jenner, D., Powell, S., Anand, R., Smith, J.C. and Markham, A.F. (1990). "A novel, rapid method for the isolation of terminal sequences from yeast artificial chromosome (YAC) clones."Nucleic Acids Res **18**:2887-2890.

Rioux, J.D., Daly, M.J., Silverberg, M.S., Lindblad, K., Steinhart, H., Cohen, Z., Delmonte, T., Kocher, K., Miller, K., Guschwan, S., Kulbokas, E.J., O'Leary, S., Winchester, E., Dewar, K., Green, T., Stone, V., Chow, C., Cohen, A., Langelier, D., Lapointe, G., Gaudet, D., Faith, J., Branco, N., Bull, S.B., McLeod, R.S., Griffiths, A.M., Bitton, A., Greenberg, G.R., Lander, E.S., Siminovitch, K.A. and Hudson, T.J. (2001). "Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease." Nat Genet **29**(2): 223-8.

Risch, N.J. (1990). "Genetic linkage and complex diseases, with special reference to psychiatric disorders." Genet Epidemiol **7**(1): 3-16; discussion 17-45.

Risch, N.J. (2000). "Searching for genetic determinants in the new millennium." Nature **405**(6788): 847-56.

Risch, N.J. and Merikangas, K. (1996). "The future of genetic studies of complex human diseases." Science **273**(5281): 1516-7.

Rogic, S., Mackworth, A.K. and Ouellette, F.B. (2001). "Evaluation of gene-finding programs on mammalian sequences." Genome Res **11**(5): 817-32.

Rose, G.D., Geselowitz, A.R., Lesser G.J., Lee R.H. and Zehfus M.H. (1985). "Hydrophobicity of amino acid residues in globular proteins." Science **229**(4716): 834-8.

Rozen, S. and Skaletsky, H. (2000). "Primer3 on the WWW for general users and for biologist programmers." Methods Mol Biol **132**: 365-86.

Saitoh, Y., Miyamoto, N., Okada, T., Gondo, Y., Showguchi-Miyata, J., Hadano, S. and Ikeda, J.E. (2000). "The RS447 human megasatellite tandem repetitive sequence encodes a novel deubiquitinating enzyme with a functional promoter." Genomics **67**(3): 291-300.

Samonte, R.V. and Eichler, E.E. (2002). "Segmental duplications and the evolution of the primate genome." Nat Rev Genet **3**(1): 65-72.

Sanan, D.A., Weisgraber, K.H., Russell, S.J., Mahley, R.W., Huang, D., Saunders, A., Schmechel, D., Wisniewski, T., Frangione, B., Roses, A.D. and Strittmatter, W.J. (1994(. "Apolipoprotein E associated with beta-amyloid peptide of Alzheimer's disease to form novel monofibrils: isoform apoE4 associates more efficiently that apoE3." J Clin Invest **94**: 860-869.

Sanacora, G., Rothman, D.L., Mason, G. and Krystal, J.H. (2003). "Clinical studies implementing glutamate neurotransmission in mood disorders." Ann N Y Acad Sci **1003**: 292-308.

Sandstrom, J., Nilsson, P., Karlsson, K. and Marklund, S.L. (1994). "10-fold increase in human plasma extracellular superoxide dismutase content caused by a mutation in heparin-binding domain." J Biol Chem **269**(29): 19163-6.

Saykin, A.J., Shtasel, D.L., Gur, R.E., Kester, D.B., Mozley, L.H., Stafiniak, P. and Gur, R.C. (1994). "Neuropsychological deficits in neuroleptic naive patients with first-episode schizophrenia." Arch Gen Psychiatry **51**(2): 124-31.

Schlager, D.S. (1994). "Early-morning administration of short-acting beta blockers for treatment of winter depression." Am J Psychiatry **151**(9): 1383-5.

Schneider, M.L. (1992). "Prenatal stress exposure alters postnatal behavioral expression under conditions of novelty challenge in rhesus monkey infants." Dev Psychobiol **25**(7): 529-40.

Selemon, L.D. and Goldman-Rakic, P.S. (1999). "The reduced neuropil hypothesis: a circuit based model of schizophrenia." Biol Psychiatry **45**(1): 17-25.

Sham, P., Bader, J.S., Craig, I., O'Donovan, M. and Owen, M. (2002). "DNA Pooling: a tool for large-scale association studies." Nat Rev Genet **3**(11): 862-71.

Shan, J. and Krukoff, T.L. (2001). "Intracerebroventricular adrenomedullin stimulates the hypothalamic-pituitary-adrenal axis, the sympathetic nervous system and production of hypothalamic nitric oxide." J Neuroendocrinol **13**(11): 975-84.

Shaw, S.H., Carrasquillo, M.M., Kashuk, C., Puffenberger, E.G. and Chakravarti, A. (1998). "Allele frequency distributions in pooled DNA samples: applications to mapping complex disease genes." Genome Res **8**(2): 111-23.

Shean (2004). Understanding and treating schizophrenia. Contemporary Research Theory and Practise. New York, The Haworth Clinical Practice Press.

Sher, L. (2003). "Aetiology and pathogenesis of mood disorders." Qjm **96**(4): 309-13.

Shifman S., Bronstein M., Sternfeld M., Pisante A., Weizman A., Reznik I., Spivak B., Grisaru N., Karp L., Schiffer R., Kotler M., Strous R.D., Swartz-Vanetik M., Knobler H.Y., Shinar E., Yakir B., Zak N.B. and Darvasi A. (2004). "COMT: a

common susceptibility gene in bipolar disorder and schizophrenia." Am J Med Genet **128**(1): 61-4.

Skolnick, P., Popik, P., Janowsky, A., Beer, B. and Lippa, A.S. (2003). ""Broad spectrum" antidepressants: is more better for the treatment of depression?" Life Sci **73**(25): 3175-9.

Smit A.F. and Riggs A.D. (1996). "Tiggers and DNA transposon fossils in the human genome." Proc Natl Acad Sci U S A. **93**(4): 1443-8.

Smith, D.J. and Lusis, A.J. (2002) "The allelic architecture of common disease." Hum Mol Gen **11**(20): 2455-2461.

Soderlund, C. and Dunham, I. (1995). "SAM: a system for iteratively building marker maps." Comput Appl Biosci **11**(6): 645-55.

Spitzer, R.L., Endicott, J. and Robins, E. (1978). "Research diagnostic criteria: rationale and reliability." Arch Gen Psychiatry **35**(6): 773-82.

Spring J. (2002). "Genome duplication strikes back." Nat Genet **31**(2): 128-9.

Stefansson, H., Sarginson, J., Kong, A., Yates, P., Steinthorsdottir, V., Gudfinnsson, E., Gunnarsdottir, S., Walker, N., Petursson, H., Crombie, C., Ingason, A., Gulcher, J.R., Stefansson, K. and St Clair, D. (2003). "Association of neuregulin 1 with schizophrenia confirmed in a Scottish population." Am J Hum Genet **72**(1): 83-7.

Stefansson, H., Sigurdsson, E., Steinthorsdottir, V., Bjornsdottir, S., Sigmundsson, T., Ghosh, S., Brynjolfsson, J., Gunnarsdottir, S., Ivarsson, O., Chou, T.T., Hjaltason, O., Birgisdottir, B., Jonsson, H., Gudnadottir, V.G., Gudmundsdottir, E., Bjornsson, A., Ingvarsson, B., Ingason, A., Sigfusson, S., Hardardottir, H., Harvey, R.P., Lai, D., Zhou, M., Brunner, D., Mutel, V., Gonzalo, A., Lemke, G., Sainz, J., Johannesson, G., Andresson, T., Gudbjartsson, D., Manolescu, A., Frigge, M.L., Gurney, M.E., Kong, A., Gulcher, J.R., Petursson, H. and Stefansson, K. (2002). "Neuregulin 1 and susceptibility to schizophrenia." Am J Hum Genet **71**(4): 877-92.

Stott, D.H. (1973). "Follow-up study from birth of the effects of prenatal stresses." Dev Med Child Neurol **5**: 770-787.

Sullivan, P.F., Neale, M.C. and Kendler, K.S. (2000). "Genetic epidemiology of major depression: review and meta-analysis." Am J Psychiatry **157**(10): 1552-62.

Surani, M.A. (2001). "Reprogramming of genome function through epigenetic inheritance." Nature **414**(6859): 122-8.

Suzuki, E., Yagi, G., Nakaki, T., Kanba, S. and Asai, M. (2001). "Elevated plasma nitrate levels in depressive states." J Affect Disord **63**(1-3): 221-4.

Syvanen, A.C. (2001). "Accessing genetic variation: genotyping single nucleotide polymorphisms." Nat Rev Genet **2**(12): 930-42.

Szymanski, M., Barciszewska, M.Z., Zywicki, M. and Barciszewski, J. (2003). "Noncoding RNA transcripts." J Appl Genet **44**(1): 1-19.

Taillon-Miller, P., Bauer-Sardina, I., Saccone, N.L., Putzel, J., Laitinen, T., Cao, A., Kere, J., Pilia, G., Rice, J.P. and Kwok, P.Y. (2000). "Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28." Nat Genet **25**(3): 324-8.

Tamminga, C.A. (1998). "Serotonin and schizophrenia." Biol Psychiatry **44**(11): 1079-80.

Terwilliger, J.D., Haghighi, F., Hiekkalinna, T.S. and Goring, H.H. (2002). "A bias-ed assessment of the use of SNPs in human complex traits." Curr Opin Genet Dev **12**(6): 726-34.

Thaker, G.K. and Carpenter, W.T., Jr. (2001). "Advances in schizophrenia." Nat Med **7**(6): 667-71.

Thakore, J.H. (1998). "Stabilisation of the Hypothalamic-Pituitary-Adrenal Axis as a Treatment Modality for Mood Disorders." Human Psychopharmacology **13**: 77-81.

Thase, M.E. and Sachs, G.S. (2000). "Bipolar depression: pharmacotherapy and related therapeutic strategies." Biol Psychiatry **48**(6): 558-72.

Thompson, W.R. (1957). "Influence of prenatal maternal anxiety on emotionality in young rats." Science **125**(3250): 698-9.

Tienari, P. (1991). "Interaction between genetic vulnerability and family environment: the Finnish adoptive family study of schizophrenia." Acta Psychiatr Scand **84**(5): 460-5.

Tienari, P., Wynne, L.C., Moring, J., Lahti, I., Naarala, M., Sorri, A., Wahlberg, K.E., Saarento, O., Seitamaa, M., Kaleva, M. and et al. (1994). "The Finnish adoptive family study of schizophrenia. Implications for family research." Br J Psychiatry Suppl(23): 20-6.

Tishkoff, S.A. and Verrelli, B.C. (2003). "Role of evolutionary history on haplotype block structure in the human genome: implications for disease mapping." Curr Opin Genet Dev **13**(6): 569-75.

Tkachev, D., Mimmack, M.L., Ryan, M.M., Wayland, M., Freeman, T., Jones, P.B., Starkey, M., Webster, M.J., Yolken, R.H. and Bahn, S. (2003). "Oligodendrocyte dysfunction in schizophrenia and bipolar disorder." The Lancet **362**: 798-805.

Tomassi, P. (2004)."Tall stories: the devilish detail of genetic association studies." Clin Endocrin **60**: 394-396.

Torrey, E.F. (1999). "Epidemiological comparison of schizophrenia and bipolar disorder" Schizophrenia Res **39**: 101-106.

Tsuang, M.T., Taylor, L. and Faraone, S.V. (2004). "An overview of the genetics of psychotic mood disorders." J Psychiatr Res **38**(1): 3-15.

Van den Oord, E.J.C.G., Neale B.M. (2004). "Will haplotype maps be useful for finding genes?" Molecualr Psychiatry **9**: 227-236.

Vawter, M.P., Freed, W.J. and Kleinman, J.E. (2000). "Neuropathology of bipolar disorder." Biol Psychiatry **48**(6): 486-504.

Venter, J.C.Adams, M.D.Myers, E.W.Li, P.W.Mural, R.J.Sutton, G.G.Smith, H.O.Yandell, M.Evans, C.A.Holt, R.A.Gocayne, J.D.Amanatides, P.Ballew, R.M.Huson, D.H.Wortman, J.R.Zhang, Q.Kodira, C.D.Zheng, X.H.Chen, L.Skupski, M.Subramanian, G.Thomas, P.D.Zhang, J.Gabor Miklos, G.L.Nelson, C.Broder, S.Clark, A.G.Nadeau, J.McKusick, V.A.Zinder, N.Levine, A.J.Roberts, R.J.Simon, M.Slayman, C.Hunkapiller, M.Bolanos, R.Delcher, A.Dew, I.Fasulo, D.Flanigan, M.Florea, L.Halpern, A.Hannenhalli, S.Kravitz, S.Levy, S.Mobarry, C.Reinert, K.Remington, K.Abu-Threideh, J.Beasley, E.Biddick, K.Bonazzi, V.Brandon, R.Cargill, M.Chandramouliswaran, I.Charlab, R.Chaturvedi, K.Deng, Z.Di Francesco, V.Dunn, P.Eilbeck, K.Evangelista, C.Gabrielian, A.E.Gan, W.Ge, W.Gong, F.Gu, Z.Guan, P.Heiman, T.J.Higgins, M.E.Ji, R.R.Ke, Z.Ketchum, K.A.Lai, Z.Lei, Y.Li, Z.Li, J.Liang, Y.Lin, X.Lu, F.Merkulov, G.V.Milshina, N.Moore, H.M.Naik, A.K.Narayan, V.A.Neelam, B.Nusskern, D.Rusch, D.B.Salzberg, S.Shao, W.Shue, B.Sun, J.Wang, Z.Wang, A.Wang, X.Wang, J.Wei, M.Wides, R.Xiao, C.Yan, C.Yao, A.Ye, J.Zhan, M.Zhang, W.Zhang, H.Zhao, Q.Zheng, L.Zhong, F.Zhong, W.Zhu, S.Zhao, S.Gilbert, D.Baumhueter, S.Spier, G.Carter, C.Cravchik, A.Woodage, T.Ali, F.An, H.Awe, A.Baldwin, D.Baden, H.Barnstead, M.Barrow, I.Beeson, K.Busam, D.Carver, A.Center, A.Cheng, M.L.Curry, L.Danaher, S.Davenport, L.Desilets, R.Dietz, S.Dodson, K.Doup, L.Ferriera, S.Garg, N.Gluecksmann, A.Hart, B.Haynes, J.Haynes, C.Heiner, C.Hladun, S.Hostin, D.Houck, J.Howland, T.Ibegwam, C.Johnson, J.Kalush, F.Kline, L.Koduru, S.Love, A.Mann, F.May, D.McCawley, S.McIntosh, T.McMullen, I.Moy, M.Moy, L.Murphy, B.Nelson, K.Pfannkoch, C.Pratts, E.Puri, V.Qureshi, H.Reardon, M.Rodriguez, R.Rogers, Y.H.Romblad, D.Ruhfel, B.Scott, R.Sitter, C.Smallwood, M.Stewart, E.Strong, R.Suh, E.Thomas, R.Tint, N.N.Tse, S.Vech, C.Wang, G.Wetter, J.Williams, S.Williams, M.Windsor, S.Winn-Deen, E.Wolfe, K.Zaveri, J.Zaveri, K.Abril, J.F.Guigo, R.Campbell, M.J.Sjolander, K.V.Karlak, B.Kejariwal, A.Mi, H.Lazareva, B.Hatton, T.Narechania, A.Diemer, K.Muruganujan, A.Guo, N.Sato, S.Bafna, V.Istrail, S.Lippert, R.Schwartz, R.Walenz, B.Yooseph, S.Allen, D.Basu, A.Baxendale, J.Blick, L.Caminha, M.Carnes-Stine, J.Caulk, P.Chiang, Y.H.Coyne, M.Dahlke, C.Mays, A.Dombroski, M.Donnelly, M.Ely, D.Esparham, S.Fosler, C.Gire, H.Glanowski, S.Glasser, K.Glodek, A.Gorokhov, M.Graham, K.Gropman, B.Harris, M.Heil, J.Henderson, S.Hoover, J.Jennings, D.Jordan, C.Jordan, J.Kasha, J.Kagan, L.Kraft, C.Levitsky, A.Lewis, M.Liu, X.Lopez, J.Ma, D.Majoros, W.McDaniel, J.Murphy, S.Newman, M.Nguyen, T.Nguyen, N.Nodell, M.Pan, S.Peck, J.Peterson, M.Rowe, W.Sanders, R.Scott, J.Simpson, M.Smith, T.Sprague, A.Stockwell, T.Turner, R.Venter, E.Wang, M.Wen, M.Wu, D.Wu, M.Xia, A.Zandieh, A.Zhu, X. (2001). "The sequence of the human genome." Science **291**(5507): 1304-51.

Visscher, P.M., Haley, C.S., Heath, S.C., Muir, W.J. and Blackwood, D.H. (1999). "Detecting QTLs for uni- and bipolar disorder using a variance component method." Psychiatr Genet **9**(2): 75-84.

Wahle, E. and Keller, W. (1996). "The biochemistry of polyadenylation." Trends Biochem Sci **21**(7): 247-50.

Wakshlak, A. and Weinstock, M. (1990). "Neonatal handling reverses behavioral abnormalities induced in rats by prenatal stress." Physiol Behav **48**(2): 289-92.

Walker, E.F. and Diforio, D. (1997). "Schizophrenia: a neural diathesis-stress model." Psychol Rev **104**(4): 667-85.

Wall, J.D., Pritchard, J.K. (2003). "Haplotype Block Structure and Linkage Disequilibrium in the Human Genome." Nature Reviews Genetics **4**: 587-597.

Wehr, T.A., Sack, D., Rosenthal, N., Duncan, W. and Gillin, J.C. (1983). "Circadian rhythm disturbances in manic-depressive illness." Fed Proc **42**(11): 2809-14.

Weinberger, D.R. and Lipska, B.K. (1995). "Cortical maldevelopment, anti-psychotic drugs, and schizophrenia: a search for common ground." Schizophr Res **16**(2): 87-110.

Weinstock, M. (1997). "Does prenatal stress impair coping and regulation of hypothalamic-pituitary-adrenal axis?" Neurosci Biobehav Rev **21**(1): 1-10.

Weisgraber, K.H., Rall, S.C. and Mahley, R.W. (1981). "Human E apoprotein heterogeneity: cysteine-arginine interchanges in the amino acid sequence of the apo-E isoforms." J Biol Chem **256**: 9077-9083.

Weiss, K.M. and Clark, A.G. (2002). "Linkage disequilibrium and the mapping of complex human traits." Trends Genet **18**(1): 19-24.

Weissman (1991). Affective Disorders. In: Psychiatric Disorders in America. The Epidemiological Catchment Area Study., Free Press, New York.

Williams, N.M., Rees, M.I., Holmans, P., Norton, N., Cardno, A.G., Jones, L.A., Murphy, K.C., Sanders, R.D., McCarthy, G., Gray, M.Y., Fenton, I., McGuffin, P. and Owen, M.J. (1999). "A two-stage genome scan for schizophrenia susceptibility genes in 196 affected sibling pairs." Hum Mol Genet **8**(9): 1729-39.

Williams, R.S., Cheng, L., Mudge, A.W. and Harwood, A.J. (2002). "A common mechanism of action for three mood-stabilizing drugs." Nature **417**(6886): 292-5.

Wolfenden, R., Andersson, L., Cullis, P.M. and Southgate, C.C. (1981). "Affinities of amino acid side chains for solvent water." Biochemistry **20**(4): 849-55.

World Health Organisation (1992) International Statistical Classification of Diseases and Related Health Problems. 10th revision.

Wright, A.F. and Hastie, N.D. (2001) "Complex genetic diseases: controversy over the Croesus code." Genome Biology **2**(8): 2007.1-2007.8.

Yang, J.Z., Si, T.M., Ruan, Y., Ling, Y.S., Han, Y.H., Wang, X.L., Zhou, M., Zhang, H.Y., Kong, Q.M., Liu, C., Zhang, D.R., Yu, Y.Q., Liu, S.Z., Ju, G.Z., Shu, L., Ma, D.L.

and Zhang, D. (2003). "Association study of neuregulin 1 gene with schizophrenia." Mol Psychiatry **8**(7): 706-9.

Zhang, M.Q. (1998). "Statistical features of human exons and their flanking regions." Hum Mol Genet **7**(5): 919-32.

Zhang, K., Deng, M., Chen, T., Waterman, M.S. and Sun, F. (2002) "A dynamic programming algorithm for haplotype block partitioning" Proc Natl Acad Sci USA. **99**: 7335-7339

Zondervan, K.T. and Cardon, L.R. (2004). "The complex interplay amount factors that influence allelic association." Nat Rev Genet **5**:89-100.

Zuker, M. (2003). "Mfold web server for nucleic acid folding and hybridization prediction." Nucleic Acids Res **31**(13): 3406-15.

Appendix I: Primer Sequences

The following tables detail the STSs used in each chapter with their primer sequences, the resultant STS size, the optimal PCR cycling annealing temperature, the optimal PCR buffer, the additives included in the PCR and any further comments. PCR cycling that utilised the touch down technique is prefixed 'TD'. PCR cycling of microsatellite markers that utilised a 50 minute final extension step are suffixed 'EXT'.

Chapter 3 Contig Mapping

STS	Primers	Size (bp)	PCR	Buffer	Comment
St751L19.m1	GCCTTGGGACTCTAAGCA AGAGAACTAGGACCCTGA	244	TD60	Gibco BRL	-
St180A12.m1	CTTCTGGGTGACTCCAATGCA AGCTAATGAGGGTCATGTGAC	550	TD60	Perkin Elmer	-
St180A12.m2	GCAGGGCAGGATTCCATCTC CACCTCCACAGACACATGGC	765	TD60	Perkin Elmer	-
St26424.m1	CCCTGAGATGGTCCAATAC AGGCACATTCGTGAGGCCT	330	TD60	Perkin Elmer	-
St26424.m2	CAAGGACACTGAGAACAG AGCTGCAGGTAGCCACAT	238	TD60	Perkin Elmer	Questionable specificity to chromosome 4
St175378snp	CTTCAAGAACGTTCCAAGAACC CAGATGGCATGGTCTGAGC	244	TD60	Gibco BRL	Amplifies chromosomes 3,4,11,15
St175378.m1	TCCAGATGGCATGGTCTG GTTATTCACGTCTTGCTA	288	TD60	Invitrogen	Questionable specificity to chromosome 4
St175378.m2	TCCAGATGGCATGGTCTG CAGCCTATCTTTTCTGAAGAAC	184	TD57	Perkin Elmer	-

STS	Primers	Size (bp)	Label	PCR	Buffer	Additive	Comment
Stb74M11.p1	CCCACCATTAACTGCCTCAT ACCCACCTATGGGACAGTAAGG	336	hex	TD48	Roche	-	-
Stb1J7.p1	GGGTGTTTGGGTAGCAACTT GCCATGTCCTTCTTCAGAACT	222	fam	TD55	Roche	-	-
Stb4E12.p1	AGGCTGAAAGCCTCTGAGTG AGGATGAGGCTGATGCTACC	150	fam	TD55	Roche	-	-
Stb352E6.p1	GAATGGAATGAAAGCTTGTGC TGGGAACATCTCACCACAAA	375	tamra	TD55	Roche	-	-
Stb426F15.p1	CACTGCTGAGTCTCTCATGTCA CATAGAGTTGGATCCATTGTTGG	262	hex	TD55	Roche	-	-
Stb426F15.p3	GTCAGTAGCGCACATACATTCC CCCATCTTTTATTCAGACTCTTACAGG	341	tet	TD55	Roche	-	-
Stb22A3.p1	CCATAGCATCCTACAAAAAGCA TGCATGGTAGTGATTGAATGC	293	fam	TD50	Roche	-	-
Stb3M2.p1b	GGAGGGAACATGAACACAAGG GAGGATTTAGGGGAGTTAGTGC	241	fam	TD55	Roche	-	-
Stb11C13.p1b	GCGCCATTAATCTTTCTATTCC AAGAACTCAGAGGTGAGCTTGG	254	hex	TD55	Roche	-	-
St585142.snp	TACTCACCTGCGCTGTTATTCC AGAAACCCTTGTCATCTTCACC	169	-	TD55	Roche	-	-
St585128.snp	GCAAGTGTGAATAAAGGTGTTGG TGCACTGGTAACAATCTCTTCC	410	-	TD55	Roche	-	-
St352E6.p2	GCCTCCTACTCCGCTATGG CATTTAACATTGTCTGCCAACC	210	Fam	TD55	Perkin Elmer	-	-
stD4S2906	CAGTCTAGATTCAAAGGAATTAGAC AATTAGAGATGCCCGTGAAA	164	hex	TD55	Perkin Elmer	-	-
St426F15.p2	CCAGCATGACCTAATAAAATGC CTGTCTTACGAGACAACAAGC	172	Tet	TD55	Roche	-	-
St426F15.p5	TTGAGGGACTACAATTTTGC	316	Fam	TD55	Perkin Elmer	-	-

	GAATGAGGGCACTGTCAGG						
St426F15.p4	CATCTATTTTGTCCTGGAAGTGC CAATGGAGGATTAAGCAAAAGC	253	hex	TD55	Perkin Elmer	-	-
stSG4961.ca	CAGCAACACAGACAAAGTCAGG TGCCCTATGAAAATCCTTAGC	230	fam	TD55	Perkin Elmer	-	-
Stb264E23.p1b	CATTTTTAGGAAGCCCCA TTTCGGGGTGGATTTCCA	433	hex	TD62ext	Perkin Elmer	DMSO	-
stCeGAx54.p1	TTCTTGGACCTTCTGGCATC ACAAAAACCGGAACTGTTGG	194	fam	TD55ext	Sanger Salts	-	-
St689P11.p1	TGAATGGTGAGCTCCTGTTTGC GGGCTGCCATCACAAAATGC	511	Fam	TD72ext	Perkin Elmer	-	-
St689P11.p2b	TGGAATCTCGCTCTGTTGC GGAAGGTGGAAAGACAGG	420	Hex	TD60ext	Gibco	-	-
St2205P10.p1b	GCAGTGATATAGAGTAGGAAAAATA CTAAACTAATTTCCTCATTTCAGG	350	Hex	TD62ext	Roche	-	-
St2205P10.p2	TGAGGTGAAATAAACTGAAACTTGC CAGGGTCGGCTATTTAAACATCTCCCA	421	Fam	TD70	Perkin Elmer	-	Amplifies other chromosomes
St2205P10.p3	GTCTCAGTGTGTAAAAAGTACT GCATGCAAGTTTCAGTTTATTTCA	345	Fam	TD63ext	Perkin Elmer	-	-
St180A12.p1	ATACCCATTCTTTTGCATCC ATTATGAGGGACTTAGGTAGCA	391	Fam	TD55ext	Roche	DMSO	-
St180A12.p2	CTTTCTCTGACAGCTCAGGG GGAGAAAAACAGAGACTCCCT	510	Fam	TD55ext	Roche	DMSO	Amplifies other chromosomes

Chapter 5 cDNA library screening

224 cgaatcgtaaccgttcgtacgagaatcgct

STS	Primers	Size (bp)	PCR
St448G15pt2	GAACATTTTTGTGGTGACAAGG TCTGTGAGTCTTCCTCATTTGC	260	TD55
St448G15pt3	GAACACAGGACATGAAATAAGTATCC GAACATTTTTGTGGTGACAAGG	624	TD55
st448G15pt4	CAGCATGACGACAAAAGATGC AGAATGGGAGCCAAGAGTGC	279	TD50
st26P5pt12	TTTTTCCTTCCTCTCTATTTCC CACAGACCTCCAAATGTCAGC	105	TD50
st26P5pt13	CCCTCTTCCACTTCTCTCTCC GGACTGCATACCACACTCAAGG	209	TD50
st287J14pt15	AAAAATCCGAACGTCAAAGG ATTAACCCTTGGGTGGAAGC	553	TD55
st287J14pt16	ATTAACCCTTGGGTGGAAGC TTCCCACAATGTTATCATGAGC	441	TD50
st287J14pt20	GACCTTGAAAGAAACTGGAACC CAGATACCATGCAGGATTGTAGG	499	TD55
st287J14pt21	ATGTTTGACTTAAACCCTGACTCC CAGATACCATGCAGGATTGTAGG	582	TD55
st287J14pt22	GACAAAGGAAAATGCTATGC GCTGGAATGAATGTGTGACC	197	TD55
st473M13pt23	CGGGAAGACACAAAGTAATATGG CAGAGAACACATTGACATCTCG	558	TD55
st473M13pt26	TGGATTCATAAAACCTGTGACC CTGGCTAGAGAAAGACTCACAGC	97	TD50
st473M13pt27	GTTGTTGATGCCAGAGAAAGC TGATGAGGACAAAGGAAGTGC	96	TD50
st437G1pt28	CGAATATCAACAGAGAAGAACAGC CAAGTAACTTCTGCATTTCTGAGG	114	TD55
st352E6pt30	TGAATCTGGGAGAAGAAGATGG CAAACATTTCCTGACCACTAAGG	163	TD55
St301J10pt31	CCCCAGGAGACATTGTAAGG GGAAGCGAAACAGAGACTGC	463	TD55
St751L19pt32	AGGAAGCCACAGAAACATGG ATGTGCCTGGTCACAGTGC	336	TD55
St74M11pt33	GGTCCCAGTAAGGAGAAAGTAGC AATGAGGTCCTGGTTCTTGC	261	TD55
St74M11pt34	TGAGTCACTTGCCCTGTAGC AGGCCTGCTCATCTGACTCC	108	TD50
St494H11pt35	GGAAAGCCTGGTCCTATTGC CCACCAAGTGGATTGTCTCC	404	TD55
St494H11pt36	GCTTCAAAGTCACCGTCACC GCGTTTCGTTATTGGCTTCC	564	TD60
St17E2pt37	CCACTGCTCCATTCTTAGGG AGCGAGAGGAAAAGGAAAGG	254	TD55
St17E2pt38	CACACACACAGTCTCTTTTCTCC GCCATGAGAAGCTCCTAAGC	458	TD55
st362I16pt39	TTTCCTCCTTGGTAAATCTTCC CCCCTCAATCTTACCAATGC	142	TD55
St106M4pt40	CTTGGCTGTGTTTTCCTTCG	543	TD50

	GGTAGCCTCTCAGTGAAAATGC		
St106M4pt41	GTTTGCAACAGGATCAGTGC GAAAACCTGAAGGGGAGAGG	154	TD55
St401G6pt42	GTCAATGATGCGGTTCTCG TCTGTGGACGAGCTTCAGG	117	TD55
St310G15pt43	CTTTCCTTTTGCCCTTTTGG CCTTTTGGTAAAACAGGTTAGC	269	TD55
St470D11pt44	TTCTCTGATGTTTGGTGAGAGC AATGTGGGAAAACCTTGTGG	471	TD55
St302F12pt45	GGCAATTGTGATGGATAGTGC CCCGGTCCCAGAATTAGG	183	TD55
St1004L1pt46	AGATGACTCACCTCTGGATGG GAAGTCTGGGATAGGGAATGG	119	TD55
st401G6pt47	TGGCCATGAGTTTGTACTGG CCTGGTGTGTAAGAGGAAGC	107	TD50
st470D11pt48	GGGTAAAAACACCAGAATACCG CGGCGCAAAGAGTACTGG	144	TD55
st470D11pt49	GCTTCTTTCGACGGATGACC GGATGACAGCGCTAAATCG	234	TD55

Nested PCR

STS	Primers	PCR
St448G15pt2n	GCAGTGGCAGCTTTAGAATCC CATAAAATCAGGGATGGTTGG	60
St448G15pt3n	CAAGATTTTCCAGGCAGTGG TCTGTGAGTCTTCCTCATTTGC	60
st448G15pt4n	GGACAGTGAGGGTTTCATGG AACCCAACAGGAGCCATAGG	60
st26P5pt12n	ACACTGTGGCAAGAAAGTGC TGCCATGCACTTTCTTGC	60
st26P5pt13n	TGGGAACAGAGTTTGGATGG TTGACCCTGGAGACAGAACC	60
st287J14pt15n	ATTAACCCTTGGGTGGAAGC CCTCTTCTCCTCCCTCAGC	60
st287J14pt20n	CAGCCCAGAGATTTTGTTGG CATTAAATCACGATACCCAAGC	60
st287J14pt21n	GACCTTGAAAGAAACTGGAACC CAGATACCATGCAGGATTGTAGG	60
st287J14pt22n	TCTCTCCCCACATATAATCATCC CTGAAACACCACATAAAGTCAAGG	60
st473M13pt23n	CTGACAGTCACCAGCACACC CATTAGTGGAGGCAGACAGG	60
st473M13pt26n	TTTCAAGAAGTGGACTTACATCC TTTGAAGCTTATCAACAGC	60
st473M13pt27n	GCCTGCTGCTGTAGTATTTGC GCAAATACTACAGCAGCAGGC	60
st437G1pt28n	GTGGATAAGGCAGAGTTTGG TCCCAAACTCTGCCTTATCC	60
st352E6pt30n	TCTTGACAGCCACCTAGA TTAGAACATTGCCTTCATGC	60
St301J10pt31n	CACGGACTGCTGCCCGTTT CATCTGAAACAAGCCAGGCTGTC	60
St751L19pt32n	CATCATGTTGGCCTTCAGC	60

	CATGATGGCTGAATGATATAGAGG	
St74M11pt33n	GATGTCCCCAAAGCAAGACC TTTGCCCTCAGTCTCAGTTCC□	60
St74M11pt34n	CCTGTAGCTGCCACTGTT TCATCTGACTCCAAGGACTTC	60
St494H11pt35n	TGTGATGGAGGCTATCTAGGG CTGATGTGCAAGGTCAGTGG	60
St17E2pt38n	TGAACAGAAGACTAGTGAGATTTGG GGTCATAACTGGGCTACGG	60
st362I16pt39n	CAGAAAAGTTGGCTGCTTCC GACAGTTACTGGCTTACACTGTGG	60
St106M4pt40n	GACATTTGCAGGCAGTTAAGG CTGGACACGAGGGGTCAGC	60
St106M4pt41n	TTCCGATGCTTCATTCTTGA AGAGGGATCTGACACAATGAC	60
St401G6pt42n	TGAGCCTCTCCACTTTGTCC CAAAGTGGAGAGGCTCAACC	60
St310G15pt43n	TGTGCGCATGTAAGATGG GACATCAGCTTCATATTCTCATGG	60
St470D11pt44n	AGCTTCAGAGAACACACACAGG CACTACATTTATAGGGTTTGTCTCC□	60
St302F12pt45n	GAGCCCAAGTATTTGAAACTCC GCATTTTTCCCCTCTACTCC□	60
St1004L1pt46n	GCTCTCTGCTTTAACAATGTGG CCACATTGTTAAAGCAGAGAGC	60
st401G6pt47n	GCTCCTTGGAGAGGATCAGC GACACCTGTCAACAGGAAAGG	60

RT-PCR

STS	Primers	DNA size (bp)	cDNA size (bp)	PCR	Buffer	Additive
St133218.1	GTTGGAAAATTTGGGAATGG TAGGCAACAGTCTTGCATGG	615	252	TD60	Roche	
stc91050.1	TGGCCATGAGTTTGTACTGG AGGAGGAGGCCATGAACG	36286	168	-	-	-
stc91050.3	GTTTGCAACAGGATCAGTGC CCCATCCCTGTATCATGTCC	1111	1111	TD60	Perkin Elmer	
stc166647.1	TGATTATCTTGCGTCATTACGG ACCACAACACTTGCATGTCC	4754	415	-	-	-
stc166647.2	GAAGAATGGCTCATTTTCTGG AAAACAAATTCCCCGAAAGG	26321	207	-	-	-
stc206037.1	CCCAGATACTCTGCCCAACC GAAAATGAGCCATTCTTCAGC	41764	72	-	-	-
stc166647.3	GGAAATTCCGGTGAATATGG TGGAAAAAGCTTTCCATTGC	926	547	TD55	Gibco	
stc166647.4	CCCAATATTGCTCTGGAAGC GCAGAAATTTCCTTGAGACTAGG	6755	477	-	-	-
stc133214.1	CCAAGAACAAGTGTCATTCAAGG CTGGCTGAGACCAGATTGC	132	132	TD60	Perkin Elmer	-
Stc133259.1	TCGGCAAACTTTGTTTCTTCC GTACAAGGCCTTCCTGCACTGG	60	60	TD55	Perkin Elmer	-
Stc202025.1	GAACTGCAAGTTCCGTGAGC TCCCACAGATTCACAGTTGG	383	383	TD65	Roche	-
Stc132895.1	AGACAAGCTTTCCCCTGAGC GCATTTGCTCCAAGATCTCC	439	272	TD55	Perkin Elmer	-
Stc132895.2	TTTATTCCGGGAGAATCTGG TTTTCCTCGAAACCATCTCC	145	145	TD60	Perkin Elmer	-
stc132895.3	AAAGAGGAGCCTGTTGTCTCC GCTGCAAAGTAATTAATATCAGTCG	64	64	TD55	Perkin Elmer	DMSO
Stc133260.1	AGAGGGGTCAACATCTACGC TTTTTCAGAGGCCTTTTTGG	1005	249	TD60	Perkin Elmer	
Stc133225.1	ATTTTGAGCAATGGGGAAGG TTCAATCACTTCGGTTTTCC	313	298	TD55	Perkin Elmer	

Stc202024.1	GTCACACTCTCCGAGAGT GTGGTGACAACTGTAACA	683	614	TD60	Perkin Elmer	
stc202024.2	GCCTGGTCACATGGGACA CACCACCTTCTCTCAAGA	736	667	TD60	Perkin Elmer	
Stc91050.5	TGGAGAGACGGAAAAACTCG CTGTGGAAGTTGATGCTTGG	3935	639	-	-	-
Stc91050.2b	TCTGACCTCTTCCAGTTCTGC TCTGTGGACGAGCTTCAGG	28789	806	-	-	-
Stc152751.1b	GGATTTCGATTTGCTGAACG AGCTCATATTTGCGCTCAGG	24486	891	-	-	-
Stc133205.1	CTGGACCTCCTTGAGAAAGC TGAAGGGTATCACTGGAAGC	577	256	TD65	Perkin Elmer	
Stc255668.1	ATGAGACCATATTCAGAGCTTTCG GTCTGCGTGGTTAAGGAAGG	5306	201	-	-	-
Stc133258.1	GGTGGACTCCAAGACAGTGC TAGACCCCCTGTCAGTCAGC	684	257	TD60	Perkin Elmer	
Stc91050.6	ACTGGATTTGGGGATTTGC GCCAGTCGCTCTCAGTCC	466	466	Failed	Failed	
Stc202015.1	GGGAGACTATCAGGCTGTGG AATGCAGTACATGGTCTTATTCTCC	624	301	TD60	Perkin Elmer	-
Stc202015.2	GCCATAATTCATTTTCCAAGC TCTATAGCCTCTTTTACAGTGTTGC	925	210	TD55	Perkin Elmer	-
Stc257414.1	GGCAATATGGTCTTCTTCTTTCC AGGTCCACATTTAGCCATGC	591	212	TD60	Perkin ELmer	-
Stc257414.2	GCGGGACCCGGTACAAGG AGCGAGGTCCACATTTAGCC	102	102	TD55	Perkin Elmer	-
Stc257414.3	TTGTGAAACTCATGGCAGTCC CTTCCTGCCTGTAGCCTTGG	118	118	TD60	Perkin Elmer	-
StcKIAA1729.1	TGTCAATAGCAAATTTAAGGTAAGG TGGCAACAGCTTTGACTGG	302	302	TD55	Perkin Elmer	-
Stc166522.1	AAAGAAACTTTGCTGCAGTCC CATATTCAGATTCCTTTATAGGC	138	138	TD55	Perkin Elmer	-
Stc166522.2	CATCGAACACTACAAGAATTTTCC CAGAGGCAAGAGGTGTCTGG	121	121	TD55	Perkin Elmer	-

Stc166450.1	TAAATCCCTCCCAAAAGTGG CCCATTGGTGTTTTAGTCTGG	167	167	TD60	Perkin Elmer	-
Stc166450.2	CAGATGAAAAGGATGGTATGAGC GGGGAGTTGTCTCTGAAAGG	381	381	TD60	Perkin Elmner	-
Stc166450.3	TCTTTTGAGCATCTTTTCAGTGC GGGAAGGAAGCACGAAGC	201	201	TD60	Perkin Elmer	-
stc202025.2	GGATTCGATTCCACCATATCC GCACATCAGCTTCATTCTGC	503	503	TD60	Perkin Elmer	-
stc202025.3	GGATTCGATTCCACCATATCC TCAAATTCTTAGCAGCCTGTATAGC	1679	?	-	-	-
stc202025.4	TTTGAGAAGTAGATGGGATGC TGGAATTGTGGGTATCTTTCC	761	?	-	-	-
stc202024.3	GAAGTATCCTGTGTGTGTGAAGC GCACATCAGCTTCATTCTGC	3768	906	-	-	-
Stc202024.4	CGCCGTCAGTCAAGTATGC GCACATCAGCTTCATTCTGC	3742	880	-	-	-
Stc285544.1	GCTGTGACCGTCTACGACAAA CGACCGCTCCGTGGCTGGA	883	144	TD60	Invitrogen	-
Stc285544.2	TTCATGAAGCTGGGCAGCACA AGCCTGCTCCCAGCATCCT	846	106	TD60	Invitrogen	-
Stc285544.3	GACCTCTGCAGTACTCCTTAG ATTCTGGCTCCGAAAATAGCAA	238	238	TD60	Invitrogen	-
Stc284455.4	TGCGTTTGGTCTCCAGGTGC TTCCAAATCAGGAACCACAAAG	335	335	TD60	Invitrogen	-
Stc202015.1n	TGCCTTTATTGTTGAAGTGAGC GGTCCACTGGTTCACTTTGC	414	91	TD60	Perkin Elmer	
St74M11pt50	CACTCAACATCAATGAAGTAACAGC TTTATTGCATCCGAATCTTCAGC	82	82	TD60	Perkin Elmer	DMSO
St74M11pt51	ATTTTTCCTCTGGAAATGAGC ACTGTGGTGACCTCAAAATGG	91	91	TD60	Perkin Elmer	DMSO
Stc133218.1b	TTGCTGAGTTCTGATGGAGAGC TGCATGGAACATCAAGCAAACC	676	311	TD60	Invitrogen	-
Stc133218.2	CAGCCCTGGTTATTGAGAAACTCC CCAAGTTGTGCTGTTTTGATTGC	10554	?	-	-	-

Stc285479.1	CTCGTGGCTTACAGCCTGGA AGAGGATCCAGCTGTGTTGA	172	172	TD65	Perkin Elmer	
Stc285479.2	TGGCTGCCACGCCCGGTTT CCCACACCCCAAAGGGCG	143	143	TD65	Perkin Elmer	
Stc285479.3	CTAAGCTGGCGGTGCTGCTT GCCTGCTAAGGGCCCGCA	575	575	TD60	Perkin Elmer	
Stc285479.4	AGTGAGGGGCCGCCTGGTT ACGCGGCCCGGGTCCGA	358	358	TD60	Invitrogen	
Stc285544.5	GCTTTTCTCAAACTTTCCTGAC CTCTCATCCCATCGGAGCTT	90	90	TD55	Invitrogen	
Stc285544.6	GCTGTGACCGTCTACGACAAA CAGACGAGAGAAGGAAGGTTG	1893	?	-	-	-
Stc285544.7	GCTGTGACCGTCTACGACAAA AGAGAAGGAAGGTTGTGTCCG	1887	?	-	-	-
Stc285544.8	GCTGTGACCGTCTACGACAAA CAGTTCCTCTGAGTCCTCAAG	5909	?	-	-	-
Stc285544.9	GCTGTTTGCCTGTGTGTGCCA ATTCTGGCTCCGAAAATAGCAA	2774	?	-	-	-
Stc285544.10	ATGAAGAAAACGGCCTGGTCC ATTCTGGCTCCGAAAATAGCAA	1706	?	-	-	-
Stc202024.5	GACCTCTGCAGTACTCCTTAG TCCCAGTCCAGCTGGGCCAA	3378	?	-	-	-
Stc202024.6	GACCTCTGCAGTACTCCTTAG CGTGGGAGATGATGGCCGTA	3023	?	-	-	-
Stc206041.1	CCTGGTCACAGAGCTGACATCG CTCTGGGGTGGCCAATCG	537	?	TD60	Perkin Elmer	
St10G12pt52	CCTTCATCCCTCTCCAGAAC GGAATAACCTTCTGTAAGCCTC	214	214	TD55	Perkin Elmer	
Stc133218.3	TGCTCCAGACAGCACACTGC TACCCCAAGGCCATGTCAGC	153	153	TD55	Roche	
stc133218.4	TCTCCATTTCCATGATTCAGTAGC GGAGCCGTTCTAACACCTTTGC	138	138	TD55	Roche	
stc133218.5	AAATACGGCAGAGCTGGAAACC AAATTTTCTCTTCCCAAGTTGTGC	103	103	TD60	Perkin Elmer	

Chapter 6 Sequencing Primers

STS	Primers	Size (bp)	PCR	Buffer	Additive
stSOD3ex1	GGGAAGTCTCCCTCTTATCTCG CTGTAAAATGGGGGTGTGAGG	711	TD60	Perkin Elmer	-
stSOD3ex2	AGGCGAAGCAATTCTACAACC TCCTTCCCAAGAAGATGATCC	329	TD60	Perkin Elmer	-
stSOD3ex3a	TTTCTTCAGATTGGGTCTCACC CTCAGGTCCCCGAACTGG	607	TD60	Perkin Elmer	-
stSOD3ex3a.2	TGTGACTCAGATCCCTGTGG GCCAGCGGGTTGTAGTGG	670	TD55	Gibco	PCR Enhancer
stSOD3ex3b	AGCTCGACGCCTTCTTCG GGGTGTTTCGGTACAAATGG	700	TD60	Perkin Elmer	-
stSOD3ex3b.4	CACTCAGAGCGCAAGAAGC GAAGCATGTTTGCCACTCC	777	TD55	Gibco	PCR Enhancer
stSOD3ex3c	CACTCTGAGGTCTCACCTTCG GGCAGAGGATAAGGAGAAGTCC	676	TD60	Perkin Elmer	-
stSOD3.3	TGTTTGTTTTAGGCTTCTCTCC ACAGCTAGGAAGTGAGGCTACC	726	TD60	Perkin Elmer	DMSO
stSOD3.5c	GGAAGAGGGAATGAATGTTGC GTTTCGGTGGCAGGAAAGC	635	TD57	Invitrogen	DMSO
stSOD3.5b	CGAGAAAAAGGCAGATTTCC AGCGACGAAGAATGAACAGG	613	TD60	Perkin Elmer	-

SNaPshot™ Genotyping Oligonucleotides

SNP	Primer	Orientation	Genotype	Size (bp)	oligo
rs2536512	CGGCGGGACGACGACGGC	F	G/A	19	2μM
rs2855262	CACTCTGAGGTCTCACCTTCGCCT	F	T/C	25	2μM
rs2695232	TCTGCTCCAACAGACACC	F	C/T	19	2μM

Chapter 7 Sequencing Primers

STS	Primers	Size	PCR	Buffer	Additive
stGpr26lex1a	GAAGGCGCTGCTGTAGCC CTCTGCACCTCCCAGAAGC	579	TD60	Gibco BRL	PCR Enhancer
stGpr26lex1b.3	GCTTCCCACTGCGCTACG GCACCAATGAACCAACAGG	513	TD57	Invitrogen	DMSO
stGpr26lex2	CCAGATATCTTGCTGGGAAACC CTCTGTGTGGAGAAGGATGAGC	467	TD60	Perkin Elmer	DMSO
stGpr26lex3	CAGCCTTAGCTCAAAGTCAAGC AGTGCTGCCACCTACTGAGC	691	TD60	Perkin Elmer	DMSO
stGpr26l.2b	GGAATGAATTGGCAAGAAGC CTTCATTTCTGTCATGTTGTGG	561	TD55	Perkin Elmer	DMSO
stGPR26l.3b	GTTCTCATGGCTCTGCTTCG ACCTTGTTTCCACTTCTTGTCG	729	TD55	Gibco BRL	PCR enhancer
stGpr26l.4	GAGAACAAATCGGAACAATGC ATGACAGCTCTCTGGATGTGG	587	TD55	Perkin Elmer	DMSO
stGpr26l.5	GTGATGCACACCACTACTCAGG TAAGTGACTGCTGTTCATGTGG	622	TD55	Perkin Elmer	DMSO

SNaPshot™ Genotyping Oligonucleotides

SNP	Primer	Orientation	Genotype	Size (bp)	Oligo
lh32	CCGCCAGCAGCAGGTGGCCCAG	R	G/T	23	2μM
lh31	CGGTGTCCATGCGCTGGCAGTG	R	C/A	23	2μM
lh48	TCACCTACAAAAAAGCAA	F	G/C	19	8μM
lh34	TGCTGAAGAGAACCCCGC	F	T/C	19	2μM
lh42	CTGTCCACACCCACATCC	F	A/G	19	2μM
lh38	GTGGGTGAGGTGGTCACTCTTG	R	G/C	23	2μM

Appendix II: DNA Pool Peak Heights of SOD3 SNPs

(see Le Hellard *et al*, 2002, for details on statistical analysis)

Table 1: pool peak heights and peak height ratio, mean (x), standard deviation (SD) and standard error of the mean (SEM).

Table 2: individual heterozygote peak heights and peak height ratios.

SNP rs2536512

	CTL			BP			SCZ			UP		
	allele 1	allele 2	Ratio	allele 1	allele 2	Ratio	allele 1	allele 2	ratio	allele 1	allele 2	ratio
Rep 1	3501	1700	2.059	2491	1090	2.285	3043	1650	1.839	2891	803	3.6
Rep 2	2807	1437	1.953	3476	1591	2.185	3838	1907	2.013	4480	1546	2.898
Rep 3	3516	1806	1.947	4470	2188	2.043	3109	1516	2.051	3661	987	3.709
Rep 4	3203	1734	1.847	5597	2432	2.301	3763	1732	2.173	5019	1872	2.681
X	3256.75	1669.25	1.952	4008.5	1825.25	2.204	3463	1701.25	2.019	4012.75	1302	3.222
SD	332.682	161.016	0.087	1332	604.2	0.119	423.11	163.527	0.138	933.258	494.187	0.509
SEM	132.817	65.735	0.035	543.787	246.667	0.049	172.734	66.759	0.056	381.001	201.751	0.208

Table 1

	Heterozygote		
	allele 1	allele 2	Ratio
1	2791	2421	1.153
2	3617	3300	1.096
3	2104	1848	1.139
4	1022	588	1.738
5	2206	1905	1.158
X	2237.25	1910.25	1.257
SD	1064.468	1108.099	0.270
SEM	434.567	452.380	0.121

Table 2

SNP rs2855262

	CTL			BP			SCZ			UP		
	allele 1	allele 2	Ratio	allele 1	allele 2	ratio	allele 1	allele 2	ratio	allele 1	allele 2	ratio
Rep 1	1954	2860	0.683	2040	2584	0.789	687	917	0.749	-	-	-
Rep 2	1465	2067	0.709	1423	1970	0.722	1438	1651	0.871	-	-	-
Rep 3	1244	1763	0.706	832	1164	0.715	1498	2092	0.716	2981	2497	1.194
Rep 4	1952	2830	0.690	-	-	-	1574	1774	0.887	803	652	1.232
X	1653.75	2380	0.697	1432.667	1906	0.742	1299.25	1608.5	0.806	1892	1574.5	1.213
SD	357.13	551.228	0.012	604.047	712.160	0.041	411.943	497.038	0.086	1540.079	1304.612	0.027
SEM	145.798	225.038	0.005	246.601	290.738	0.017	168.175	202.915	0.035	628.734	532.606	0.011

Table 1

	Heterozygote		
	allele 1	allele 2	Ratio
1	1876	4334	0.433
2	2055	3795	0.542
3	1002	2414	0.415
4	594	1352	0.439
5	1488	3364	0.442
X	1403	3051.8	0.454
SD	606.494	1182.155	0.05
SEM	247.6	482.613	0.022

Table 2

SNP 2695232

	CTL			BP			SCZ			UP		
	allele 1	allele 2	Ratio	allele 1	allele 2	ratio	allele 1	allele 2	ratio	allele 1	allele 2	ratio
Rep 1	636	3621	0.176	570	3765	0.151	78	1482	0.053	643	4301	0.150
Rep 2	408	2750	0.148	452	2932	0.154	330	3882	0.085	58	1763	0.033
Rep 3	248	2424	0.102	173	1714	0.101	440	2965	0.148	-	-	-
Rep 4	663	4075	0.163	-	-	-	346	3176	0.109	-	-	-
X	488.75	3217.5	0.147	398.333	2803.667	0.135	298.5	2876.25	0.099	350.5	3032	0.091
SD	197.084	762.958	0.032	203.868	1031.505	0.030	154.802	1008.827	0.040	413.675	1794.637	0.082
SEM	80.459	311.477	0.013	83.229	421.11	0.012	63.198	411.852	0.016	168.875	732.657	0.034

Table 1

	Heterozygote		
	allele 1	allele 2	Ratio
1	969	3279	0.296
2	1049	3753	0.280
3	662	2072	0.319
4	346	1229	0.282
5	749	2682	0.279
X	755	2603	0.291
SD	277.578	994.429	0.017
SEM	113.321	405.974	0.008

Table 2

Appendix III: DNA Pool Peak Heights of GPR78 SNPs

(see Le Hellard *et al*, 2002, for details on statistical analysis)

Table 1: pool peak heights and peak height ratio, mean (x), standard deviation (SD) and standard error of the mean (SEM).

Table 2: individual heterozygote peak heights and peak height ratios.

SNP ih31

	CTL			BP			SCZ			UP		
	allele 1	allele 2	Ratio	allele 1	allele 2	Ratio	allele 1	allele 2	ratio	allele 1	allele 2	ratio
Rep 1	2342	2024	1.157	1044	1025	1.019	1785	2564	0.696	1273	1071	1.189
Rep 2	799	759	1.053	592	742	0.798	1288	1690	0.762	1372	1460	0.940
Rep 3	2191	1495	1.466	1385	1595	0.868	716	1074	0.667	1819	1755	1.036
Rep 4	1911	1050	1.820	838	1207	0.694	1826	2445	0.747	861	808	1.066
X	1810.750	1332	1.374	964.750	1142.250	0.845	1403.750	1943.250	0.718	1331.250	1273.500	1.058
SD	697.736	551.751	0.345	335.608	357.362	0.136	519.630	696.854	0.044	393.314	418.059	0.103
SEM	284.849	225.251	0.141	137.012	145.892	0.056	212.138	284.489	0.018	160.570	170.672	0.042

Table 1

	Heterozygote		
	allele 1	allele 2	Ratio
1	1554	3512	0.442
2	1059	2015	9.526
3	662	782	0.847
4	3188	3575	0.892
5	1245	1504	0.828
X	1541.600	2277.600	0.707
SD	1126.227	1184.138	0.207
SEM	459.780	483.422	0.092

Table 2

SNP ih32

	CTL			BP			SCZ			UP		
	allele 1	allele 2	Ratio	allele 1	allele 2	ratio	allele 1	allele 2	ratio	allele 1	allele 2	ratio
Rep 1	3724	5234	0.712	2607	3851	0.677	6443	10549	0.611	4958	9248	0.536
Rep 2	6060	8282	0.732	1608	2533	0.635	8475	12092	0.701	3327	5075	0.656
Rep 3	4071	6325	0.644	1717	3034	0.566	6520	9980	0.653	4031	7027	0.574
Rep 4	3382	4519	0.748	2684	4118	0.652	5356	8891	0.602	4104	5850	0.702
Rep 5	3889	5927	0.656	913	1612	0.566	5449	8399	0.649	4559	7125	0.640
Rep 6	3481	5992	0.581	1973	2868	0.688	-	-	-	4022	5846	0.688
X	4101	6047	0.679	1917	3003	0.631	6449	9982	0.643	4167	6695	0.632
SD	992.675	1271.67	0.063	664.985	909.648	0.053	1255.43	1455.14	0.039	552.890	1474.595	0.065
SEM	405.258	519.156	0.026	271.479	371.362	0.022	561.445	650.760	0.018	225.716	602.001	0.027

Table 1

	Heterozygote		
	allele 1	allele 2	Ratio
1	3416	5812	0.588
2	2066	6134	0.337
3	2658	5632	0.472
4	2373	5971	0.397
5	3206	7065	0.454
6	5023	8351	0.601
X	3124	6494	0.475
SD	1058.179	1037.862	0.104
SEM	432.0	423.705	0.043

Table 2

SNP ih48

	CTL			BP			SCZ			UP		
	allele 1	allele 2	Ratio	allele 1	allele 2	ratio	allele 1	allele 2	ratio	allele 1	allele 2	ratio
Rep 1	3664	708	5.175	2079	126	16.5	3447	1027	3.356	3577	336	10.646
Rep 2	7587	722	10.508	2218	94	23.596	3992	914	4.368	1365	132	10.341
Rep 3	5033	512	9.830	1167	77	15.156	2883	737	3.912	2086	324	6.438
Rep 4	3949	228	17.320	2112	281	7.516	3074	707	4.348	4500	539	8.349
Rep 5	2739	390	7.023	6716	602	11.156	806	167	4.826	4316	388	11.124
Rep 6	-	-	-	3651	1009	3.618	1936	431	4.492	3577	111	32.225
X	4594	512	9.971	2991	365	12.924	2690	664	4.217	3237	305	13.187
SD	1862.305	210.983	4.637	1991.990	371.878	7.080	1146.379	317.095	0.514	1250.583	161.575	9.489
SEM	832.848	94.355	2.074	813.226	151.819	2.890	468.007	129.453	0.210	510.549	65.963	3.874

Table 1

	Heterozygote		
	allele 1	allele 2	Ratio
1	1179	426	2.768
2	2723	833	3.269
3	1542	348	4.431
4	1779	963	1.847
5	2643	1217	2.172
6	3540	1143	3.097
X	2234	822	2.931
SD	884.415	363.472	0.913
SEM	361.061	148.387	0.373

Table 2

SNP ih34

	CTL			BP			SCZ			UP		
	allele 1	allele 2	Ratio	allele 1	allele 2	ratio	allele 1	allele 2	ratio	allele 1	allele 2	ratio
Rep 1	5101	364	14.014	4583	333	13.763	77294	494	14.765	8566	721	11.881
Rep 2	7585	440	17.239	2033	156	13.032	9288	616	15.078	5508	461	11.948
Rep 3	10017	562	17.824	2277	153	14.882	9461	514	18.407	6733	535	12.585
Rep 4	14761	893	16.530	5621	418	13.447	8338	469	17.778	11770	965	12.197
Rep 5	4297	261	16.464	9700	682	14.223	6978	416	16.774	9920	775	12.8
Rep 6	3438	190	18.095	-	-	-	7219	445	16.222	-	-	-
X	7533	452	16.694	4843	348	13.869	8096	492	16.504	8499	691	12.282
SD	4279.539	252.821	1.470	3112.225	218.843	0.714	1095.893	69.825	1.446	2489.378	200.167	0.40
SEM	1747.115	103.214	0.600	1391.830	97.870	0.320	447.397	28.506	0.590	1113.284	89.517	0.179

Table 1

	Heterozygote		
	allele 1	allele 2	Ratio
1	2861	1179	2.427
2	2941	1042	2.822
3	8536	3344	2.553
4	-	-	-
5	-	-	-
X	4779	1855	2.601
SD	3253.615	1291.330	0.202
SEM	1878.475	745.550	0.117

Table 2

SNP ih42

	CTL			BP			SCZ			UP		
	allele 1	allele 2	Ratio	allele 1	allele 2	ratio	allele 1	allele 2	ratio	allele 1	allele 2	ratio
Rep 1	7773	1500	5.182	5823	1743	3.341	9583	2190	4.376	4580	1198	3.823
Rep 2	11752	2795	4.205	9073	2247	3.708	4655	1219	3.819	9372	2513	3.729
Rep 3	12500	2531	4.939	10927	3223	3.390	6990	1560	4.481	12142	3710	3.273
Rep 4	9006	1995	4.514	7062	2194	3.219	2046	460	4.448	8092	3014	2.685
X	10258	2205	4.710	8221	2402	3.414	5819	1357	4.281	8547	2609	3.378
SD	2236.167	576.050	0.435	2246.629	620.110	0.208	3221.250	720.821	0.311	3138.405	1060.874	0.521
SEM	118.084	228.025	0.218	1123.314	310.055	0.104	1610.625	360.411	0.156	1569.203	530.437	0.260

Table 1

	Heterozygote		
	allele 1	allele 2	Ratio
1	5829	7812	0.746
2	7663	12861	0.596
3	4172	8763	0.476
X	5888	9812	0.606
SD	1746.248	2682.984	0.135
SEM	1008.197	1549.021	0.078

Table 2

Appendix IV: Raw Data of the Phase I and Phase II Association Study

SOD3

Raw data for four SNPs tested in Phase I (1: C-2669721_10, 2: 2536512, 3: rs2855262, 4: 2324580). Table shows sex (0: unknown, 1: male, 2: female), diagnosis (0: control, 1: BPAD, 2: SCZ) and genotype (1: A, 2: C, 3: G, 4: T) of each chromosome (a or b).

Sample	Sex	Diag	1a	1b	2a	2b	3a	3b	4a	4b
413	0	0	2	2	3	3	2	2	2	2
424	0	0	1	1	1	1	4	4	2	2
428	0	0	1	2	1	3	4	4	2	2
431	0	0	1	1	1	3	4	4	2	2
433	0	0	1	1	1	3	2	4	2	2
444	0	0	1	2	1	3	2	4	2	4
445	0	0	1	2	1	3	4	4	2	2
447	0	0	1	1	1	1	4	4	2	2
450	0	0	1	1	1	1	4	4	2	2
452	0	0	1	2	1	3	4	4	2	2
456	0	0	1	1	1	1	4	4	2	2
457	0	0	1	1	1	1	4	4	2	2
458	0	0	1	1	1	1	4	4	2	2
459	0	0	1	1	1	1	4	4	2	2
460	0	0	1	2	1	3	2	2	2	2
461	0	0	1	2	3	3	2	4	2	2
462	0	0	1	1	1	1	4	4	2	2
464	0	0	1	2	1	3	2	4	2	4
465	0	0	1	1	1	1	4	4	2	2
467	0	0	1	2	1	3	2	4	2	2
472	0	0	2	2	3	3	2	4	2	2
485	0	0	1	1	1	1	4	4	2	2
486	0	0	1	2	1	1	4	4	2	2
487	0	0	1	1	1	1	4	4	2	2
488	0	0	1	2	1	3	4	4	2	2
489	0	0	1	1	1	1	2	4	2	2
490	0	0	1	2	1	3	2	4	2	2
491	0	0	1	2	1	3	4	4	2	2
492	0	0	1	1	1	1	4	4	2	2
493	0	0	1	2	1	3	4	4	2	2
494	0	0	1	2	1	3	4	4	2	2
495	0	0	1	2	1	3	2	4	2	2
496	0	0	1	1	1	1	4	4	2	2
497	0	0	2	2	3	3	4	4	2	2
498	0	0	1	2	1	3	0	0	2	4
500	0	0	1	1	1	1	4	4	2	2

501	0	0	1	2	1	3	4	4	2	2
502	0	0	1	2	3	3	4	4	4	4
503	0	0	1	1	1	1	2	4	2	4
504	0	0	1	2	1	3	4	4	2	4
505	0	0	1	2	1	3	4	4	2	2
506	0	0	1	1	1	1	4	4	2	4
521	0	0	1	1	1	1	4	4	2	2
531	0	0	1	2	1	1	2	4	2	2
532	0	0	2	2	3	3	4	4	2	4
533	0	0	1	2	1	3	0	0	2	2
535	0	0	1	1	1	1	0	0	2	4
536	0	0	2	2	0	0	4	4	2	2
537	0	0	1	1	1	1	4	4	2	2
545	0	0	1	1	1	1	2	4	2	2
554	0	0	1	2	1	1	4	4	2	4
556	0	0	1	2	1	3	2	4	2	2
561	0	0	1	2	1	3	2	4	2	2
570	0	0	1	1	1	3	4	4	2	4
572	0	0	1	1	1	1	4	4	2	4
580	0	0	1	1	1	1	4	4	2	2
587	0	0	2	2	0	0	4	4	2	2
588	0	0	1	1	1	1	4	4	2	2
591	0	0	1	1	1	3	4	4	2	2
592	0	0	1	1	1	1	2	4	2	2
595	0	0	1	2	3	3	2	4	2	2
598	0	0	1	1	1	1	4	4	2	2
599	0	0	1	2	1	3	4	4	2	2
600	0	0	1	1	1	1	4	4	2	2
603	0	0	1	1	1	1	2	4	2	2
604	0	0	1	2	1	3	4	4	2	2
607	0	0	1	1	1	1	4	4	2	2
610	0	0	1	1	1	1	4	4	2	2
613	0	0	1	2	1	3	4	4	2	2
615	0	0	1	1	1	1	4	4	2	2
616	0	0	1	1	1	1	4	4	2	2
618	0	0	1	1	1	1	4	4	2	2
619	0	0	1	1	1	3	2	4	2	2
621	0	0	1	1	1	3	2	4	2	2
622	0	0	1	2	1	3	2	4	2	2
624	0	0	2	2	3	3	2	4	2	4
625	0	0	2	2	3	3	2	4	2	2
627	0	0	1	2	1	1	4	4	2	2
633	0	0	1	2	1	1	0	0	2	2
634	0	0	1	1	1	1	0	0	2	2
635	0	0	2	2	1	3	2	4	2	4
639	0	0	1	2	1	3	0	0	2	2
640	0	0	1	2	1	3	0	0	2	2
641	0	0	1	1	1	1	4	4	2	2
644	0	0	1	1	1	1	4	4	2	2
645	0	0	1	2	1	3	4	4	2	2

648	0	0	1	2	1	1	4	4	2	2
650	0	0	1	2	3	3	4	4	4	4
651	0	0	1	1	1	3	2	4	2	4
653	0	0	1	1	1	1	4	4	2	2
655	0	0	1	1	1	1	4	4	2	2
656	0	0	1	2	1	3	0	0	2	2
657	0	0	1	2	1	1	0	0	2	2
659	0	0	2	2	3	3	4	4	2	4
682	0	0	1	1	1	1	2	4	2	4
243	1	1	1	1	1	1	0	0	2	2
297	1	2	1	1	1	1	4	4	2	2
302	1	1	1	2	1	3	4	4	2	2
319	2	1	1	2	1	3	4	4	2	4
323	2	2	2	2	3	3	4	4	2	2
346	2	1	2	2	3	3	2	4	2	4
358	1	2	1	2	1	3	2	4	2	2
404	1	1	1	2	1	3	4	4	2	2
462	2	2	1	2	1	3	2	4	2	4
774	2	2	1	2	1	3	2	4	2	2
899	2	2	1	2	1	3	4	4	2	2
1080	1	2	1	2	1	3	2	2	2	2
1085	1	2	1	2	1	3	4	4	2	2
1190	2	1	1	1	1	1	4	4	2	2
1325	2	1	1	1	1	1	4	4	2	2
1326	1	1	1	1	0	0	4	4	2	2
1475	1	2	1	2	1	3	4	4	2	2
1548	1	1	1	1	1	1	2	4	2	2
1563	2	1	1	2	1	3	2	4	2	4
1567	1	1	1	1	1	1	4	4	2	2
1645	2	1	1	1	1	3	4	4	2	2
1646	1	2	1	2	0	0	4	4	2	2
1695	1	2	1	1	1	3	0	0	2	2
1705	2	1	1	1	1	1	0	0	2	4
1725	1	1	1	1	1	1	4	4	2	2
1758	2	1	1	2	1	1	4	4	2	4
1767	2	1	1	2	1	3	0	0	2	2
1780	1	2	2	2	3	3	4	4	2	2
1797	1	2	1	2	1	3	4	4	2	2
1961	1	2	1	1	1	3	2	2	2	2
2075	1	2	1	2	1	3	4	4	2	2
2086	1	2	1	1	1	1	4	4	2	2
2092	1	1	2	2	3	3	2	4	2	2
3000	2	1	1	2	1	3	4	4	2	2
3016	2	1	1	2	0	0	4	4	2	2
3042	2	1	1	2	1	3	4	4	2	2
3049	1	1	1	1	1	1	2	4	2	2
3071	1	2	1	2	1	3	4	4	2	2
3109	2	1	1	2	1	3	4	4	2	2
3133	1	1	1	2	1	3	4	4	2	2
3136	1	2	1	2	1	1	4	4	2	2

3141	1	2	1	1	1	3	4	4	2	2
3150	2	1	1	2	1	3	2	4	2	4
3159	2	1	2	2	3	3	4	4	2	2
3162	1	1	2	2	3	3	4	4	2	4
3165	1	1	1	1	1	1	4	4	2	2
3183	2	1	1	1	1	1	4	4	2	2
3193	1	1	1	1	1	1	4	4	2	2
3194	1	2	1	1	1	1	2	4	2	2
3195	1	1	1	1	1	1	2	2	2	4
3196	2	1	1	1	0	0	4	4	2	2
3205	2	2	1	2	1	3	4	4	2	2
3206	1	2	1	1	1	1	4	4	2	2
3207	2	2	1	1	1	1	4	4	2	2
3208	1	2	1	2	1	3	4	4	2	2
3209	2	2	1	1	1	1	4	4	2	2
3223	2	1	1	2	1	1	2	4	2	2
3339	1	2	2	2	3	3	4	4	2	2
3340	2	2	1	2	1	3	2	4	2	2
3347	1	1	1	1	1	1	4	4	2	2
3389	2	1	1	1	1	1	2	4	2	2
3422	1	2	1	2	1	3	0	0	2	4
3435	2	1	2	2	3	3	2	4	2	2
3458	1	1	1	1	1	3	4	4	2	4
3467	2	1	1	2	1	3	4	4	2	4
3477	2	1	1	2	1	3	4	4	2	4
3489	2	1	1	1	1	1	4	4	2	2
3505	2	2	2	2	3	3	4	4	2	2
3537	1	2	2	2	1	3	4	4	2	4
3542	1	1	1	2	3	3	4	4	4	4
3545	1	2	1	2	3	3	4	4	2	2
3559	1	2	1	1	0	0	4	4	2	2
3570	2	1	1	2	0	0	4	4	2	2
3572	1	2	1	2	1	3	4	4	2	2
3581	1	2	1	1	1	1	4	4	2	2
3590	1	1	1	2	0	0	0	0	2	2
3594	1	2	1	1	1	1	2	4	2	2
3596	1	1	1	2	0	0	4	4	2	2
3622	1	2	1	2	1	3	2	4	2	4
3623	2	1	1	1	1	1	4	4	2	2
3639	1	1	1	2	1	3	4	4	2	2
3640	1	2	1	2	1	3	4	4	2	2
3647	1	2	1	2	1	3	4	4	2	2
3652	2	2	2	2	3	3	4	4	2	2
3655	2	1	1	1	1	1	4	4	2	2
3664	2	1	1	1	1	3	4	4	2	2
3684	2	1	1	1	1	1	4	4	2	2
3685	1	2	1	1	1	1	4	4	4	4
3695	2	1	1	1	1	1	4	4	2	2
3705	2	1	1	1	1	1	2	4	2	4
3742	1	2	1	1	1	1	4	4	2	2

3743	1	2	1	2	1	3	4	4	2	2
3756	2	2	1	1	1	1	4	4	2	2
3766	1	2	1	1	1	1	4	4	2	2
3782	2	2	0	0	1	3	2	4	2	2
3783	1	2	1	1	1	1	2	4	2	4
3787	2	2	0	0	0	0	4	4	2	2
3791	2	2	1	1	1	1	4	4	2	2
3793	1	1	1	2	1	1	4	4	2	2
3800	1	2	1	1	1	3	2	4	2	4
3802	2	2	1	2	1	3	4	4	2	2
3805	1	2	1	2	1	3	4	4	2	4
3845	2	2	1	1	1	1	4	4	2	2
3867	1	1	2	2	3	3	4	4	2	2
3880	2	1	1	1	1	1	4	4	2	2
3889	1	2	1	2	1	3	2	4	2	2
3895	1	2	1	2	1	3	4	4	2	2
3900	1	1	1	1	0	0	4	4	2	2
3911	2	2	1	2	1	3	4	4	2	4
3918	2	1	2	2	0	0	2	4	2	2
3919	2	2	1	1	1	1	4	4	2	4
3920	2	1	1	1	1	1	4	4	2	2
3921	1	2	1	2	1	3	4	4	2	2
3938	1	1	1	1	1	1	2	4	2	4
3944	1	2	1	1	1	3	4	4	0	0
3945	2	2	1	2	1	1	4	4	2	4
3953	2	2	1	2	1	3	4	4	2	4
3969	1	2	1	1	1	1	2	4	2	2
4000	2	1	1	1	1	1	4	4	2	2
4008	1	2	1	2	1	3	2	4	2	2
4012	1	1	1	1	1	1	4	4	2	2
4018	2	1	1	1	1	3	4	4	2	4
4020	2	1	2	2	3	3	2	4	2	2
4042	2	2	1	2	1	3	2	4	2	2
4052	2	1	1	1	1	1	2	4	2	2
4058	1	2	1	2	1	3	0	0	2	2
4066	1	1	1	1	1	1	4	4	2	2
4067	1	1	1	2	1	3	4	4	2	2
4068	2	1	1	2	1	3	4	4	2	4
4101	1	2	1	2	1	1	2	4	2	4
4102	1	2	1	1	1	1	4	4	2	2
4106	1	2	1	2	1	1	2	2	2	2
4111	1	2	1	1	1	1	4	4	2	2
4112	2	2	1	1	1	1	4	4	2	2
4157	2	1	1	1	0	0	4	4	2	2
4171	2	1	1	1	1	1	4	4	2	4
4178	2	1	1	1	1	1	4	4	2	2
4179	1	2	1	2	1	3	2	4	2	2
4206	2	1	1	2	3	3	4	4	2	2
4277	2	1	1	2	1	3	4	4	2	2
4344	2	1	1	1	1	1	4	4	2	2

4355	2	2	1	1	1	3	4	4	2	2
4358	2	1	1	2	1	3	2	4	2	2
4387	1	1	1	2	1	3	2	4	2	2
4403	1	1	1	2	1	3	2	4	2	2
4421	2	1	1	1	1	1	4	4	2	2
4424	1	2	1	1	1	1	2	4	2	2
4428	1	1	1	2	3	3	4	4	2	4
4441	1	2	1	2	1	3	4	4	2	2
4443	2	2	1	1	1	1	4	4	2	2
4460	1	2	1	2	1	1	4	4	2	2
4463	1	2	1	2	1	3	4	4	2	2
4512	1	2	1	1	1	1	2	4	2	2
4571	2	1	1	1	1	3	4	4	2	4
4577	2	1	1	2	0	0	4	4	2	2
4620	2	1	1	1	1	3	4	4	2	2
4622	2	1	1	2	1	3	4	4	2	2
4683	1	1	1	1	1	1	2	4	2	2
4694	1	2	1	2	1	3	2	4	2	2
4698	1	2	1	1	1	1	2	4	2	4
4706	1	2	1	1	1	1	4	4	2	2
4707	1	2	1	1	1	1	4	4	2	4
4710	1	2	1	1	1	1	2	4	2	4
4715	1	2	1	1	1	1	4	4	2	4
4716	1	2	1	1	0	0	2	4	2	2
4721	2	1	1	2	1	3	2	4	2	2
4731	2	2	2	2	3	3	2	2	2	2
4748	1	2	2	2	1	1	4	4	2	2
4753	2	1	1	1	0	0	4	4	2	2
4762	2	1	1	1	1	1	2	4	2	2
4775	1	1	2	2	3	3	4	4	2	2
4793	1	1	1	2	1	3	0	0	2	2
5302	2	1	1	2	1	3	4	4	2	2
5315	1	1	1	1	1	1	2	4	2	2
5337	2	2	1	1	1	1	4	4	2	2
5338	1	2	1	1	1	1	4	4	2	2
5350	1	2	1	1	3	3	2	4	2	4
5363	2	1	1	1	1	1	4	4	2	2
5367	1	2	1	1	3	3	4	4	2	4
5376	1	2	1	1	1	1	4	4	2	2
5379	1	2	1	1	1	1	2	4	2	4
5390	1	2	1	2	1	1	2	4	2	2
5423	1	1	1	1	0	0	4	4	2	4
5452	1	2	1	1	0	0	4	4	2	2
5463	1	1	1	2	1	3	4	4	2	2
5484	1	1	1	2	1	3	4	4	2	2
5487	1	2	1	2	0	0	4	4	2	4
6057	2	1	1	2	1	3	4	4	2	4

GPR78

Raw data for nine SNPs tested in Phase I (1: C_1221917_10, 2: C_11352300_30, 3: ih63, 4: ih33, 5: ih31, 6: ih34, 7:ih40, 8: ih36, 9: C_1221895_10). Sa = sample number. Se = sex. Codes are as previous table.

Sa	Se	D	1a	1b	2a	2b	3a	3b	4a	4b	5a	5b	6a	6b	7a	7b	8a	8b	9a	9b
413	0	0	2	2	4	4	3	3	4	4	2	2	3	3	3	3	2	2	4	4
424	0	0	2	4	2	2	2	2	2	2	1	1	0	0	3	3	2	2	2	4
428	0	0	2	2	2	4	2	3	2	4	1	2	0	0	2	3	2	3	4	4
431	0	0	2	2	2	2	2	2	2	2	0	0	3	3	3	3	2	2	4	4
433	0	0	2	2	2	2	2	3	2	2	1	1	3	3	3	3	2	2	2	4
444	0	0	2	2	2	2	2	2	2	2	1	1	3	3	3	3	2	2	2	4
445	0	0	4	4	2	2	2	3	2	4	1	2	3	3	3	3	2	2	2	4
447	0	0	2	4	2	2	2	3	2	4	1	2	3	3	3	3	2	2	4	4
450	0	0	2	4	2	2	2	2	2	2	1	1	3	3	2	3	2	2	2	4
452	0	0	2	2	2	4	2	2	2	2	1	1	3	3	3	3	2	2	4	4
456	0	0	2	2	2	4	2	2	2	2	1	1	3	3	3	3	2	2	4	4
457	0	0	2	4	2	2	2	2	2	2	1	1	3	3	3	3	2	2	2	4
458	0	0	2	2	2	2	2	2	2	2	0	0	3	3	3	3	2	2	4	4
459	0	0	4	4	2	2	2	2	2	2	0	0	3	3	3	3	2	2	2	4
460	0	0	2	4	2	4	2	3	2	4	0	0	3	3	3	3	2	2	4	4
461	0	0	2	4	2	4	3	3	4	4	2	2	3	3	3	3	2	2	4	4
462	0	0	2	2	2	4	2	3	2	4	0	0	3	3	3	3	2	2	4	4
464	0	0	2	2	2	4	2	3	2	4	0	0	3	3	3	3	2	2	4	4
465	0	0	2	2	2	4	3	3	2	4	0	0	1	3	2	3	2	3	2	4
467	0	0	2	4	2	2	2	2	2	2	0	0	1	3	2	3	2	3	2	4
472	0	0	2	2	2	2	2	2	2	2	0	0	1	3	3	3	2	2	2	4
485	0	0	2	4	2	2	2	3	2	4	1	2	3	3	3	3	2	2	2	2
486	0	0	2	2	2	2	2	2	2	2	0	0	3	3	3	3	2	2	4	4
487	0	0	2	4	2	2	2	3	2	4	1	2	3	3	3	3	2	2	4	4
488	0	0	2	4	2	2	2	2	2	2	1	1	3	3	3	3	2	2	2	4
489	0	0	2	2	2	2	3	3	2	4	1	2	3	3	3	3	2	2	2	2
490	0	0	2	2	2	4	3	3	2	4	1	2	3	3	3	3	2	2	2	4
491	0	0	4	4	2	2	3	3	4	4	2	2	3	3	3	3	2	2	2	4
492	0	0	2	2	2	4	2	3	2	4	1	2	3	3	3	3	2	2	4	4
493	0	0	2	2	2	2	2	3	2	2	1	1	3	3	3	3	2	2	4	4
494	0	0	2	4	2	2	2	3	2	4	1	2	3	3	3	3	2	2	4	4
495	0	0	2	2	2	2	2	2	2	2	1	1	3	3	3	3	2	2	4	4
496	0	0	2	4	2	2	2	3	2	4	1	2	1	3	2	3	2	3	2	4
497	0	0	2	2	2	2	2	2	2	2	1	1	3	3	3	3	2	2	2	4
498	0	0	2	4	2	2	2	2	0	0	0	0	0	0	0	0	0	0	2	4
500	0	0	2	2	2	2	2	2	2	2	0	0	3	3	3	3	2	2	4	4
501	0	0	2	2	2	2	2	2	2	2	0	0	3	3	3	3	2	2	4	4
502	0	0	2	2	2	4	2	3	2	4	1	2	3	3	2	3	2	2	2	2
503	0	0	2	4	2	2	2	3	2	4	1	2	3	3	3	3	2	2	4	4
504	0	0	2	4	2	2	2	3	2	4	1	2	1	3	3	3	2	2	4	4
505	0	0	2	4	2	4	2	3	2	4	1	2	1	3	2	2	2	3	2	2
506	0	0	2	2	2	2	2	3	2	2	0	0	0	0	3	3	2	2	4	4
521	0	0	2	4	2	2	2	3	2	4	1	2	1	3	3	3	2	2	4	4

531	0	0	2	4	2	2	2	3	2	4	1	2	3	3	3	3	2	2	4	4
532	0	0	2	4	2	2	2	3	2	4	1	2	0	0	3	3	2	2	4	4
533	0	0	2	4	2	2	2	3	0	0	0	0	0	0	0	0	0	0	4	4
535	0	0	2	4	2	2	2	3	0	0	0	0	0	0	0	0	0	0	4	4
536	0	0	2	4	2	2	2	3	2	4	1	2	3	3	3	3	2	2	4	4
537	0	0	2	2	2	4	2	3	2	4	0	0	3	3	3	3	2	2	4	4
545	0	0	2	2	2	2	2	3	2	2	0	0	1	3	3	3	2	2	4	4
554	0	0	2	2	2	4	2	3	2	4	0	0	3	3	3	3	2	2	4	4
556	0	0	2	2	2	4	3	3	2	4	0	0	1	3	2	3	2	3	2	4
561	0	0	2	4	2	2	2	2	2	2	0	0	1	3	3	3	2	2	2	4
570	0	0	4	4	2	2	3	3	4	4	2	2	3	3	3	3	2	2	2	4
572	0	0	2	2	2	2	2	2	2	2	0	0	1	3	3	3	2	2	2	4
580	0	0	4	4	2	2	2	2	2	2	0	0	3	3	3	3	2	2	2	4
587	0	0	2	2	2	2	2	2	2	2	0	0	1	3	3	3	2	2	4	4
588	0	0	2	4	2	4	2	3	4	4	2	2	3	3	3	3	2	2	4	4
591	0	0	2	2	2	2	2	2	2	2	0	0	3	3	3	3	2	2	2	4
592	0	0	2	4	2	2	2	3	2	4	1	2	3	3	3	3	2	2	2	2
595	0	0	2	2	2	2	2	2	2	2	1	1	3	3	3	3	2	2	4	4
598	0	0	2	2	2	4	2	3	2	4	1	2	3	3	3	3	2	2	4	4
599	0	0	2	2	2	2	2	3	2	4	1	2	3	3	3	3	2	2	4	4
600	0	0	2	4	2	2	2	3	2	4	1	2	1	3	2	3	2	3	2	4
603	0	0	2	4	2	4	2	3	2	4	1	2	3	3	3	3	2	2	2	4
604	0	0	2	4	2	4	2	3	2	4	1	2	1	3	2	3	2	3	2	4
607	0	0	2	4	2	4	2	3	2	4	1	2	3	3	3	3	2	2	2	2
610	0	0	2	2	2	2	2	2	2	2	1	1	1	3	2	3	2	3	2	4
613	0	0	2	4	2	4	2	2	2	4	1	2	3	3	3	3	2	2	4	4
615	0	0	2	4	2	2	2	2	2	2	1	1	1	3	2	3	2	3	2	4
616	0	0	2	2	2	4	2	3	2	4	1	2	3	3	3	3	2	2	4	4
618	0	0	2	4	2	4	2	3	2	4	1	2	3	3	2	3	2	2	2	4
619	0	0	4	4	2	2	3	3	4	4	2	2	3	3	3	3	2	2	2	2
621	0	0	4	4	2	2	2	2	2	2	0	0	1	1	2	2	3	3	2	2
622	0	0	4	4	2	2	3	3	4	4	2	2	3	3	3	3	2	2	4	4
624	0	0	2	4	2	2	2	3	2	4	1	2	3	3	3	3	2	2	4	4
625	0	0	2	4	2	4	2	3	2	4	1	2	3	3	2	3	2	2	2	4
627	0	0	2	2	2	4	2	3	2	4	1	2	3	3	2	3	2	2	2	4
633	0	0	2	4	2	2	2	2	0	0	0	0	0	0	0	0	0	0	4	4
634	0	0	4	4	2	2	2	2	0	0	0	0	0	0	0	0	0	0	2	4
635	0	0	2	2	2	4	2	3	2	4	1	2	1	3	2	3	2	3	2	2
639	0	0	2	2	2	2	2	2	0	0	0	0	0	0	0	0	0	0	2	4
640	0	0	2	2	4	4	3	3	0	0	0	0	0	0	0	0	0	0	2	4
641	0	0	2	4	2	2	2	3	4	4	2	2	3	3	3	3	2	2	4	4
644	0	0	2	2	4	4	3	3	4	4	2	2	3	3	3	3	2	2	2	4
645	0	0	2	4	2	2	2	2	2	2	1	1	3	3	3	3	2	2	4	4
648	0	0	2	4	2	2	2	3	2	4	0	0	3	3	3	3	2	2	4	4
650	0	0	2	2	2	4	2	3	2	4	1	2	1	1	2	2	3	3	2	2
651	0	0	2	2	2	2	2	3	2	4	1	2	3	3	3	3	2	2	4	4
653	0	0	2	4	2	4	3	3	4	4	2	2	3	3	3	3	2	2	2	4
655	0	0	2	4	2	2	2	3	2	4	1	2	3	3	3	3	2	2	4	4
656	0	0	2	2	2	2	0	0	0	0	0	0	0	0	0	0	0	0	2	4
657	0	0	2	2	2	2	2	3	0	0	0	0	0	0	0	0	2	2	2	4

659	0	0	2	2	2	2	2	3	2	4	0	0	3	3	3	3	2	2	4	4
682	0	0	2	2	2	2	2	2	2	2	1	1	3	3	3	3	2	2	2	4
243	1	1	2	4	2	2	2	3	0	0	1	2	3	3	3	3	2	2	4	4
297	1	2	2	2	2	2	2	2	2	2	1	1	3	3	3	3	2	2	2	4
302	1	1	2	4	2	2	2	3	2	4	0	0	3	3	3	3	2	2	4	4
319	2	1	2	2	2	4	2	3	2	4	0	0	3	3	3	3	2	2	2	4
323	2	2	2	2	2	2	2	2	2	2	1	2	3	3	3	3	2	2	2	4
346	2	1	2	2	2	2	2	2	2	2	1	1	3	3	3	3	2	2	2	4
358	1	2	4	4	2	2	2	3	2	4	1	2	3	3	3	3	2	2	4	4
404	1	1	2	4	2	2	2	2	2	2	1	1	3	3	3	3	2	2	2	4
462	2	2	2	2	2	2	2	3	2	4	1	2	3	3	3	3	2	2	2	2
774	2	2	2	4	2	2	3	3	2	4	1	2	3	3	2	3	2	2	2	2
899	2	2	2	4	2	2	3	3	4	4	2	2	3	3	3	3	2	2	2	4
1080	1	2	2	4	2	2	2	2	2	2	1	1	3	3	3	3	2	2	4	4
1085	1	2	2	4	2	2	2	2	2	2	1	1	1	3	2	3	2	3	4	4
1190	2	1	4	4	2	2	2	3	2	4	1	2	3	3	3	3	2	2	2	4
1325	2	1	2	2	2	4	2	3	2	4	1	2	3	3	3	3	2	2	4	4
1326	1	1	2	2	2	2	2	2	2	2	0	0	1	3	2	3	2	3	2	4
1475	1	2	2	4	2	2	2	3	2	2	1	1	0	0	3	3	2	2	2	4
1548	1	1	2	4	2	2	3	3	2	4	1	2	3	3	3	3	2	2	2	4
1563	2	1	2	2	2	4	2	3	2	4	1	2	3	3	3	3	2	2	4	4
1567	1	1	4	4	2	2	2	2	2	2	1	1	3	3	3	3	2	2	2	4
1645	2	1	4	4	2	2	2	3	2	4	1	2	1	3	2	3	2	3	2	2
1646	1	2	4	4	2	2	2	3	2	4	1	2	3	3	3	3	2	2	4	4
1695	1	2	4	4	2	2	2	2	2	2	1	1	1	3	3	3	2	2	4	4
1705	2	1	2	4	2	4	3	3	0	0	0	0	0	0	0	0	0	0	4	4
1725	1	1	2	2	2	4	2	3	2	4	1	2	1	3	2	3	2	3	2	2
1758	2	1	2	2	2	2	2	2	2	2	1	1	1	3	2	3	2	3	2	4
1767	2	1	2	4	2	4	2	3	0	0	0	0	0	0	0	0	0	0	4	4
1780	1	2	2	4	2	2	2	3	2	4	1	2	1	3	2	3	2	3	2	2
1797	1	2	2	4	2	2	3	3	4	4	2	2	3	3	3	3	2	2	4	4
1961	1	2	2	2	2	2	2	2	2	2	1	1	1	3	2	3	2	3	2	4
2075	1	2	0	0	2	4	3	3	4	4	2	2	3	3	3	3	2	2	4	4
2086	1	2	2	2	2	4	2	3	2	4	1	2	3	3	3	3	2	2	2	4
2092	1	1	2	2	2	2	2	2	2	2	0	0	1	3	2	3	2	3	2	2
3000	2	1	2	4	2	2	2	3	2	4	1	2	3	3	3	3	2	2	2	4
3016	2	1	4	4	2	2	2	3	2	4	1	2	3	3	3	3	2	2	4	4
3042	2	1	2	2	2	2	2	2	2	2	1	1	3	3	3	3	2	2	2	4
3049	1	1	2	4	0	0	3	3	2	4	1	2	3	3	3	3	2	2	4	4
3071	1	2	2	2	2	2	2	3	2	2	1	1	3	3	3	3	2	2	2	4
3109	2	1	2	2	2	2	2	2	2	2	1	1	1	3	2	3	2	3	2	2
3133	1	1	2	2	2	2	2	2	2	2	1	1	3	3	3	3	2	2	4	4
3136	1	2	2	2	2	4	2	3	2	4	1	2	3	3	2	3	2	2	2	4
3141	1	2	2	4	2	2	2	3	2	4	1	2	3	3	3	3	2	2	4	4
3150	2	1	2	4	2	4	2	3	2	4	1	2	1	3	2	3	2	3	4	4
3159	2	1	2	2	2	4	2	3	2	4	1	2	3	3	3	3	2	2	4	4
3162	1	1	2	4	2	2	2	2	2	2	1	1	3	3	3	3	2	2	2	4
3165	1	1	4	4	2	2	3	3	4	4	2	2	3	3	3	3	2	2	4	4
3183	2	1	2	4	2	4	2	3	2	4	1	2	3	3	3	3	2	2	4	4
3193	1	1	2	4	2	2	2	2	2	2	1	1	1	3	2	3	2	3	2	4

3194	1	2	2	2	2	2	2	3	2	4	1	2	3	3	3	3	2	2	4	4
3195	1	1	2	2	2	2	2	2	2	2	1	1	1	3	2	3	2	3	2	4
3196	2	1	2	2	4	4	3	3	4	4	2	2	3	3	2	3	2	2	2	4
3205	2	2	2	4	2	2	2	3	2	4	1	2	3	3	3	3	2	2	4	4
3206	1	2	2	2	2	4	2	3	2	4	1	2	1	3	2	3	2	3	2	4
3207	2	2	2	4	2	2	2	2	2	2	0	0	3	3	3	3	2	2	2	4
3208	1	2	2	2	2	2	2	2	2	2	1	1	3	3	3	3	2	2	4	4
3209	2	2	2	2	2	4	2	3	2	4	1	2	3	3	3	3	2	2	4	4
3223	2	1	2	2	2	2	2	2	2	2	1	1	3	3	3	3	2	2	4	4
3339	1	2	4	4	2	2	3	3	4	4	2	2	3	3	3	3	2	2	4	4
3340	2	2	0	0	2	4	2	2	2	2	1	1	3	3	3	3	2	2	4	4
3347	1	1	2	4	2	4	2	3	2	4	1	2	3	3	3	3	2	2	2	4
3389	2	1	2	4	2	2	2	3	2	4	1	2	3	3	3	3	2	2	4	4
3422	1	2	2	4	2	4	2	3	2	4	1	2	3	3	3	3	2	2	4	4
3435	2	1	2	2	2	4	3	3	2	4	1	2	1	3	2	3	0	0	2	4
3458	1	1	2	4	2	2	2	3	2	4	1	2	3	3	3	3	2	2	4	4
3467	2	1	2	4	2	2	2	2	2	2	0	0	1	3	2	3	2	3	2	2
3477	2	1	2	2	2	4	2	3	2	4	0	0	3	3	3	3	2	2	4	4
3489	2	1	2	2	2	2	2	2	2	2	0	0	1	3	2	3	2	3	2	4
3505	2	2	2	2	2	4	2	3	2	4	1	2	0	0	3	3	2	2	4	4
3537	1	2	2	2	2	2	2	3	2	4	1	2	3	3	3	3	2	2	4	4
3542	1	1	4	4	2	2	2	3	2	4	1	2	3	3	3	3	2	2	4	4
3545	1	2	2	2	2	4	0	0	2	4	1	2	0	0	3	3	2	2	4	4
3559	1	2	2	4	2	2	3	3	4	4	2	2	3	3	3	3	2	2	4	4
3570	2	1	2	4	2	4	2	3	2	4	0	0	1	3	3	3	2	2	4	4
3572	1	2	4	4	2	2	3	3	4	4	2	2	3	3	3	3	2	2	4	4
3581	1	2	2	4	2	2	2	3	2	2	1	1	3	3	3	3	2	2	2	4
3590	1	1	2	2	2	4	0	0	0	0	0	0	0	0	0	0	0	0	2	4
3594	1	2	2	2	2	4	2	3	2	4	1	2	3	3	3	3	2	2	2	4
3596	1	1	4	4	2	2	2	3	2	4	1	2	3	3	3	3	2	2	4	4
3622	1	2	2	4	2	2	2	3	2	4	1	2	3	3	3	3	2	2	2	4
3623	2	1	2	4	2	2	2	3	2	4	1	2	3	3	3	3	2	2	4	4
3639	1	1	2	4	2	2	2	3	2	4	1	2	3	3	3	3	2	2	2	4
3640	1	2	2	4	2	2	2	2	2	2	1	1	1	3	2	3	2	3	2	2
3647	1	2	2	2	2	2	2	3	2	4	1	2	1	3	2	3	2	3	2	4
3652	2	2	2	2	2	4	2	3	2	4	1	2	1	3	3	3	2	2	4	4
3655	2	1	2	2	2	4	3	3	4	4	2	2	3	3	3	3	2	2	4	4
3664	2	1	2	4	2	2	2	3	0	0	0	0	3	3	3	3	2	2	2	4
3684	2	1	2	2	4	4	3	3	4	4	2	2	3	3	3	3	2	2	4	4
3685	1	2	2	4	2	2	2	2	2	2	1	1	3	3	3	3	2	2	4	4
3695	2	1	2	2	2	4	2	3	4	4	2	2	3	3	3	3	2	2	4	4
3705	2	1	2	4	2	2	2	3	2	4	1	1	3	3	3	3	2	2	2	4
3742	1	2	4	4	2	2	2	3	2	4	1	2	3	3	3	3	2	2	2	4
3743	1	2	2	4	2	2	2	2	2	2	1	1	1	3	2	3	2	3	2	4
3756	2	2	2	2	2	2	2	3	2	4	1	2	3	3	3	3	2	2	4	4
3766	1	2	2	4	2	2	2	2	2	2	1	1	1	3	3	3	2	2	4	4
3782	2	2	2	4	2	2	2	3	2	4	1	2	3	3	3	3	2	2	4	4
3783	1	2	2	4	2	2	2	2	2	2	1	1	1	3	2	3	2	3	2	2
3787	2	2	0	0	0	0	2	2	2	2	1	1	1	3	3	3	2	2	0	0
3791	2	2	2	2	2	4	2	3	2	4	1	2	3	3	3	3	2	2	4	4

3793	1	1	2	4	2	4	2	3	2	4	1	2	3	3	3	3	2	2	4	4
3800	1	2	2	2	2	2	2	2	2	2	1	1	0	0	3	3	2	2	2	4
3802	2	2	2	2	2	2	2	2	2	2	1	1	1	3	3	3	2	2	4	4
3805	1	2	2	4	2	2	2	2	2	2	1	1	1	3	2	3	2	3	4	4
3845	2	2	2	4	2	4	2	2	2	2	1	1	3	3	3	3	0	0	2	4
3867	1	1	2	4	2	2	2	3	2	4	2	2	1	3	2	3	2	3	2	4
3880	2	1	2	4	2	2	2	2	2	2	1	1	0	0	3	3	2	2	2	4
3889	1	2	2	2	2	2	2	2	2	2	1	1	1	3	3	3	0	0	4	4
3895	1	2	2	2	2	2	3	3	4	4	2	2	3	3	3	3	2	2	2	4
3900	1	1	2	2	2	4	2	3	2	4	1	2	3	3	3	3	2	2	4	4
3911	2	2	4	4	2	2	2	3	2	4	1	2	3	3	3	3	0	0	4	4
3918	2	1	2	4	2	2	3	3	4	4	2	2	3	3	3	3	2	2	4	4
3919	2	2	2	4	2	4	0	0	2	4	1	2	0	0	3	3	2	2	2	4
3920	2	1	2	4	2	2	2	2	2	2	0	0	3	3	3	3	2	2	4	4
3921	1	2	2	4	2	2	2	2	2	2	1	1	1	3	3	3	2	2	4	4
3938	1	1	2	4	2	4	3	3	4	4	2	2	3	3	3	3	2	2	4	4
3944	1	2	2	2	2	2	2	3	2	4	1	2	1	3	2	3	2	3	4	4
3945	2	2	2	2	0	0	2	2	2	2	1	1	3	3	3	3	2	2	2	4
3953	2	2	2	4	2	2	2	2	2	2	0	0	3	3	3	3	2	2	2	2
3969	1	2	2	2	2	4	2	3	2	4	1	2	3	3	3	3	2	2	2	4
4000	2	1	2	4	2	2	2	2	2	2	1	1	0	0	3	3	2	2	2	2
4008	1	2	2	4	2	2	3	3	2	4	1	2	3	3	3	3	2	2	2	4
4012	1	1	2	4	2	2	2	3	2	2	1	1	1	3	2	3	2	3	2	4
4018	2	1	4	4	2	2	2	3	2	4	1	2	3	3	3	3	2	2	2	4
4020	2	1	2	2	2	2	3	3	4	4	2	2	3	3	3	3	2	2	4	4
4042	2	2	2	2	2	2	2	2	2	2	1	1	3	3	3	3	2	2	4	4
4052	2	1	2	2	2	2	2	3	2	4	1	2	3	3	3	3	2	2	4	4
4058	1	2	2	4	2	2	0	0	0	0	0	0	0	0	3	3	0	0	2	4
4066	1	1	2	4	2	2	2	3	2	4	1	2	3	3	3	3	2	2	2	4
4067	1	1	2	4	4	4	3	3	4	4	2	2	3	3	3	3	2	2	4	4
4068	2	1	2	2	2	2	2	2	2	2	1	1	3	3	3	3	2	2	4	4
4101	1	2	2	2	2	4	2	3	2	4	1	2	3	3	2	3	2	2	2	4
4102	1	2	2	2	2	2	2	3	2	2	1	1	3	3	3	3	2	2	4	4
4106	1	2	4	4	2	4	2	2	2	2	1	1	3	3	3	3	2	2	4	4
4111	1	2	2	4	2	4	2	3	2	4	1	2	1	3	2	3	2	3	4	4
4112	2	2	2	4	2	2	2	3	2	4	1	2	3	3	3	3	2	2	2	4
4157	2	1	2	2	2	4	2	2	2	2	1	1	3	3	3	3	2	2	4	4
4171	2	1	4	4	2	2	2	3	2	2	1	1	1	3	2	3	2	3	2	2
4178	2	1	2	4	2	2	2	2	2	2	1	1	0	0	3	3	2	2	2	4
4179	1	2	2	4	2	2	2	2	2	2	1	1	3	3	3	3	2	2	4	4
4206	2	1	2	4	2	2	2	2	2	2	1	1	3	3	3	3	2	2	2	4
4277	2	1	2	2	2	2	0	0	2	2	1	1	0	0	3	3	2	2	2	4
4344	2	1	2	4	2	2	2	2	2	2	1	1	1	3	2	3	2	3	2	4
4355	2	2	2	4	2	2	3	3	2	4	1	2	3	3	3	3	2	2	2	4
4358	2	1	2	2	2	2	2	3	2	4	1	2	3	3	3	3	2	2	4	4
4387	1	1	2	4	2	2	2	3	2	4	1	2	3	3	3	3	2	2	2	4
4403	1	1	2	2	2	2	2	2	2	2	1	1	3	3	3	3	2	2	4	4
4421	2	1	2	4	2	2	2	3	2	2	1	1	0	0	2	3	2	3	2	4
4424	1	2	2	4	2	2	2	3	2	4	1	2	3	3	3	3	2	2	4	4
4428	1	1	2	4	2	4	3	3	4	4	2	2	0	0	3	3	2	2	4	4

4441	1	2	2	4	2	4	3	3	4	4	2	2	3	3	3	3	2	2	4	4
4443	2	2	2	2	2	2	2	2	2	2	1	1	0	0	2	3	2	2	2	4
4460	1	2	2	2	2	4	2	2	2	2	1	1	3	3	3	3	2	2	4	4
4463	1	2	2	2	2	2	2	3	2	4	1	2	1	3	2	3	2	3	2	4
4512	1	2	2	4	2	2	2	2	2	2	1	1	3	3	3	3	2	2	4	4
4571	2	1	2	4	2	2	2	2	2	2	1	1	3	3	3	3	2	2	2	4
4577	2	1	2	4	2	4	2	3	2	4	1	2	0	0	2	3	2	3	2	4
4620	2	1	2	4	2	2	2	2	2	2	1	1	0	0	3	3	2	2	2	4
4622	2	1	2	2	4	4	3	3	4	4	2	2	3	3	3	3	2	2	4	4
4683	1	1	2	2	2	2	2	3	2	2	0	0	3	3	3	3	2	2	2	4
4694	1	2	2	2	2	2	3	3	4	4	2	2	3	3	3	3	2	2	4	4
4698	1	2	2	2	2	2	2	3	2	2	1	1	3	3	3	3	2	2	2	4
4706	1	2	2	4	2	4	3	3	4	4	2	2	3	3	3	3	2	2	2	4
4707	1	2	2	2	2	4	0	0	2	2	1	1	0	0	3	3	2	2	4	4
4710	1	2	2	2	2	2	2	2	2	2	1	1	1	3	2	3	2	3	2	2
4715	1	2	2	4	2	4	3	3	2	4	1	2	0	0	2	3	2	3	4	4
4716	1	2	2	4	2	2	3	3	2	4	1	2	3	3	3	3	2	2	2	4
4721	2	1	2	4	2	2	2	2	2	2	1	1	3	3	3	3	2	2	2	4
4731	2	2	2	4	2	2	2	2	2	2	1	1	3	3	3	3	2	2	2	4
4748	1	2	2	2	2	2	2	3	2	4	1	2	0	0	3	3	2	2	4	4
4753	2	1	2	2	2	4	2	3	2	4	1	2	3	3	3	3	2	2	4	4
4762	2	1	2	4	2	4	3	3	4	4	2	2	3	3	3	3	2	2	4	4
4775	1	1	2	2	2	4	3	3	4	4	2	2	3	3	3	3	2	2	4	4
4793	1	1	2	2	2	4	2	3	0	0	0	0	0	0	0	0	0	0	4	4
5302	2	1	2	4	2	2	2	2	2	2	0	0	0	0	3	3	2	2	2	4
5315	1	1	2	2	2	2	2	2	2	2	0	0	3	3	3	3	2	2	2	4
5337	2	2	2	2	4	4	3	3	4	4	2	2	3	3	3	3	2	2	4	4
5338	1	2	2	2	2	2	2	2	2	2	1	1	3	3	3	3	2	2	4	4
5350	1	2	2	4	2	2	2	3	2	4	1	2	3	3	3	3	2	2	2	2
5363	2	1	2	2	2	2	2	2	2	2	0	0	3	3	3	3	2	2	4	4
5367	1	2	2	2	2	2	2	2	2	2	0	0	0	0	3	3	2	2	4	4
5376	1	2	2	2	2	4	3	3	2	2	1	1	1	3	2	3	2	3	2	2
5379	1	2	2	2	2	4	2	3	2	4	1	2	3	3	2	3	2	2	2	4
5390	1	2	2	2	2	2	2	2	2	2	1	1	3	3	3	3	2	2	4	4
5423	1	1	2	4	2	2	2	3	2	4	1	2	3	3	3	3	2	2	2	4
5452	1	2	2	4	2	2	2	3	2	4	0	0	0	0	3	3	2	2	4	4
5463	1	1	2	2	2	2	2	2	2	2	1	1	3	3	3	3	2	2	2	4
5484	1	1	2	2	2	4	2	3	2	4	1	2	1	3	2	3	2	3	2	4
5487	1	2	2	2	2	2	2	2	2	2	1	1	3	3	3	3	2	2	4	4
6057	2	1	2	4	2	2	3	3	2	4	1	2	3	3	3	3	2	2	4	4