



Terms and Conditions of Use of Digitised Theses from Trinity College Library Dublin

Copyright statement

All material supplied by Trinity College Library is protected by copyright (under the Copyright and Related Rights Act, 2000 as amended) and other relevant Intellectual Property Rights. By accessing and using a Digitised Thesis from Trinity College Library you acknowledge that all Intellectual Property Rights in any Works supplied are the sole and exclusive property of the copyright and/or other IPR holder. Specific copyright holders may not be explicitly identified. Use of materials from other sources within a thesis should not be construed as a claim over them.

A non-exclusive, non-transferable licence is hereby granted to those using or reproducing, in whole or in part, the material for valid purposes, providing the copyright owners are acknowledged using the normal conventions. Where specific permission to use material is required, this is identified and such permission must be sought from the copyright holder or agency cited.

Liability statement

By using a Digitised Thesis, I accept that Trinity College Dublin bears no legal responsibility for the accuracy, legality or comprehensiveness of materials contained within the thesis, and that Trinity College Dublin accepts no liability for indirect, consequential, or incidental, damages or losses arising from use of the thesis for whatever reason. Information located in a thesis may be subject to specific use constraints, details of which may not be explicitly described. It is the responsibility of potential and actual users to be aware of such constraints and to abide by them. By making use of material from a digitised thesis, you accept these copyright and disclaimer provisions. Where it is brought to the attention of Trinity College Library that there may be a breach of copyright or other restraint, it is the policy to withdraw or take down access to a thesis while the issue is being resolved.

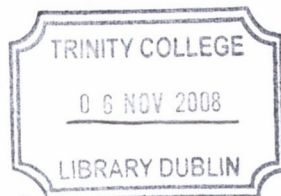
Access Agreement

By using a Digitised Thesis from Trinity College Library you are bound by the following Terms & Conditions. Please read them carefully.

I have read and I understand the following statement: All material supplied via a Digitised Thesis from Trinity College Library is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of a thesis is not permitted, except that material may be duplicated by you for your research use or for educational purposes in electronic or print form providing the copyright owners are acknowledged using the normal conventions. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone. This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

**Investigation of putative candidate genes
involved in the pathogenesis of
Schizophrenia using a large Irish case-
control sample**

Kevin McGhee



A thesis submitted to the University of Dublin
for the degree of Doctor of Philosophy

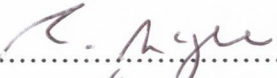
Department of Psychiatry
University of Dublin
Trinity College – August 2008

THESIS
8614

Declaration

Declaration 1:

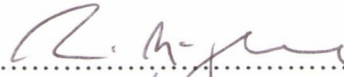
This work has not been previously accepted in substance for any degree and is not being concurrently submitted in candidature for any degree at this or any other university.

Signed.....

Date.....11/8/08

Declaration 2:

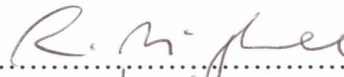
This thesis is the result of my own investigations except where otherwise stated in the text. Other sources are acknowledged in the text. A bibliography is appended.

Signed.....

Date.....11/8/08

Declaration 3:

I hereby give consent for this thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organizations.

Signed.....

Date.....11/8/08

Summary

This research thesis involved a candidate gene search for schizophrenia in the Irish population. It sought to identify candidate genes based on positional and functional information contained in previously published material. The pathogenesis of schizophrenia is currently unknown. However, two prominent theories are that of pathology during neurodevelopment and more recently the involvement of oxidative stress. Eight genes putatively involved in schizophrenia pathogenesis were investigated.

Firstly, Dr. Aiden Corvin and I equally investigated the six genes in the Apolipoprotein-L (APOL) family. I looked at APOL-4, -5 and -6 while Dr. Corvin looked at APOL -1, -2 and -3. This family of genes were both functional and positional candidates as APOL-1, -2 and -4 were found to be significantly up-regulated in post-mortem schizophrenic brains (Mimmack et al 2002), located in a region previously identified in linkage studies and are located on chromosome 22q and some 15 Mb away from the Velo-cardio facial syndrome deletion region. Using public information from NCBI's dbSNP database, 143 SNPs spanning the six genes were selected from 187 SNPs and initially genotyped in a DNA pool sample to confirm heterozygosity, of which 36% were. The 51 SNPs were genotyped in case and control DNA pools. Using a novel method for correcting pooled allele frequencies (McGhee et al 2005) 2 SNPs were identified for further analysis. After correction with true heterozygote frequencies, neither SNP showed association with schizophrenia.

The second investigation was a replication of the initial positive G72/G30 and DAAO finding by Chumakov et al (2002). Five SNPs were selected based on findings from other replication studies and four SNPs were genotyped in G72/G30 and DAAO respectively. Two markers from each gene showed association with schizophrenia. In addition, work by Dr. Aiden Corvin and Dr. Derek Morris showed statistical interaction between G72/G30 and DAAO. This study represents the first association

of G72/G30 and DAAO in the Irish population. Replication in a larger Irish sample (i.e. Resource for Psychosis Genomics Ireland, RPGI), LD analysis of this locus and functional studies are warranted.

The third investigation looked at two genes involved in oxidative stress and mitochondrial dysfunction: Heat shock protein 8 (HSPA8) and Glutathione S-transferase (GSTM3). Both of these genes were positional and functional candidate genes. They were both found to be significantly down-regulated in post mortem schizophrenic brains (Prabakaran et al 2004) and are located within significant meta-analysis linkage bins (Lewis et al 2004). Re-sequencing and mutation detection (DHPLC) was initially carried out. Identified SNPs were commercially genotyped in an Irish Reference Panel (n=92) to determine LD structure. Tag SNPs (tSNPs) were identified and genotyped in the full association sample. No association was found with both genes and schizophrenia.

As no association was found with GSTM3, I then investigated one of its basic leucine transcription factors, Nuclear erythroid factor 2 (NRF2). NRF2 is located on 2q31.2. Although it is not a positional candidate, it was a functional one as it is involved in GSTM3 regulation. LD analysis was carried out in the Irish Reference Panel and tSNPs identified. The tSNPs were commercially genotyped in the full association sample. No association of NRF2 was found with schizophrenia.

In addition to the association study of NRF2, I took the opportunity to compare the then completed Phase I data of the HapMap project with its applicability to future association studies in our sample. By comparing LD structure of CEPH HapMap data with the Irish sample (D' and r^2). I concluded that the applicability of tSNPs identified from HapMap using Phase I data would have been high. However, the density of markers at that time would have led to a loss of power. As I reached the end of my PhD, Phase II HapMap data became available addressing the density issue. Therefore, HapMap will serve as a useful tool in future Irish association studies in schizophrenia.

Acknowledgements

I would like to thank Professor Michael Gill and Dr. Aiden Corvin for the opportunity to undertake these studies. Their support, patience and mentoring has been invaluable and will not be forgotten. I hope that I will uphold their reputation in the field of psychiatric genetics and continue to collaborate with them. I owe all my laboratory training to Dr. Derek Morris. Without his constant advice and patience, this thesis would not have been possible. I admire his approach, not just in genetics, and aspire to follow in his footsteps. I also have to thank Dr. Ziarah Hawi for friendship and support required throughout a student's life. He is a fantastic teacher, with a wealth of knowledge and has always been there for Suzanne, Conor and I.

Many friends have now passed through the group and will be fondly remembered. To Dr's Judith Conroy, Naomi Lowe, David Lambert and Ricardo Segurado, thanks for everything and for keeping me sane (just). I wish all of you well in your future careers. To Kevin Murphy and Niamh Kenny, you were a great laugh. Good luck with your PhD's (and in a few years time don't say I told you so). Thanks to Sarah Clarke who I had great conversations with and always cheered me up. I wish you the very best of luck with the DCLin.Psych.

Thanks also go to Dr. Gary Donohoe for the 'Donegal catch' joke. It still brings a smile to my face. To the remaining team, including Dr. Louise Gallagher, who was the first person I met when I started (and I think you sent me home), Karen and Lynne, good luck with the research. I would like to mention everyone else, but the list is too long. So I won't.

The project of course, would not have been possible without the donation of DNA from patients and controls alike. All of you have played a part in shaping my life. I hope I can now reciprocate.

Lastly, and most importantly, I must thank my family for years of support. To my wife, Suzanne and our son Conor, I am sorry for the bad times during the PhD and thank you for sticking by me. I hope we can now spend every evening and weekend

together and enjoy life as a family. To my parents, Peter and Kathleen, on reflection of my life so far, words cannot express my gratitude for the sacrifices you have made in your own lives for me to get this far. Thank you. As this is the final stage of my PhD, I now owe my Grampa Stokes that meal (but just wish Gran could have been there too).

Statement of Work

This work was the product of the psychosis genetics group, Department of Psychiatry, Trinity College Dublin. The clinical recruitment and collection of samples for DNA were performed by Dr. Aiden Corvin, Siobhan Schwaiger (research nurse) and Dr. Jeanne-Marie Nangle.

DNA extraction and quantification was performed by the author, Dr. Derek Morris and Dr. Aiden Corvin. All genotyping and analysis for APOL-4, 5, 6, HSPA8, GSTM3 and NRF2 was performed by the author or commercially by K-Bioscience Genotyping Services. The novel methodology associated with DNA pooling was jointly by Dr. Derek Morris and the author. Genotyping and analysis of APOL-1, 2, 3 was carried out by Dr. Aiden Corvin. All genotyping of G72/DAAO was jointly carried out by Mr. Kevin Murphy and the author. The G72/DAAO analysis was carried out by the author unless otherwise stated in Chapter 4.

Publications

Investigation of the apolipoprotein-L (APOL) gene family and schizophrenia using a novel DNA pooling strategy for public database SNPs. **Kevin A. McGhee**, Derek W. Morris, Siobhan Schwaiger, Jeanne-Marie Nangle, Gary Donohoe, Sarah Clarke, David Meagher, John Quinn, Paul Scully, John L. Waddington, Michael Gill, Aiden Corvin. (2005) *Schizophrenia Research* (76/2-3) pp 231-238

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. Aiden P. Corvin, Derek W. Morris, **Kevin McGhee**, Siobhan Schwaiger, Paul Scully, John Quinn, David Meagher, David St. Clair, John L. Waddington, Michael Gill. (2004) *Molecular Psychiatry* Feb; 9(2):208-13.

Confirming RGS4 as a susceptibility gene for schizophrenia. Derek W. Morris, Alana Rodgers, **Kevin A. McGhee**, Siobhan Schwaiger, Paul Scully, John Quinn, David Meagher, John L. Waddington, Michael Gill, Aiden P. Corvin. (2004) *American Journal of Medical Genetics (Neuropsychiatric Genetics)* Feb. 15;125B(1):50-3.

No evidence for association of the dysbindin gene [DTNBP1] with schizophrenia in an Irish population-based study. Morris DW, **McGhee KA**, Schwaiger S, Scully P, Quinn J, Meagher D, Waddington JL, Gill M, Corvin AP. (2003) *Schizophrenia Research* Apr 1;60(2-3):167-72.

Abbreviations

5-HT	Serotonin
APOL	Apolipoprotein L
ARE	Anti-oxidant response element
BAC	Bacterial artificial chromosome
BP	Bipolar
CI	Confidence Intervals
cM	Centimorgans
CSF	Cerebral Spinal Fluid
CT	Computer tomography
D'	D prime
DA	Dopamine
DAAO	D-Amino Acid Oxidase
dbSNP	NCBI SNP database
ddH ₂ O	Ultra pure water
ddNTPs	Dideoxy nucleotide triphosphate
dHPLC	Denaturing high performance liquid chromatography
DNA	Dioxyribonucleic acid
dNTPs	Deoxy nucleotide triphosphate
DRD3	Dopamine receptor D3
dsDNA	Double stranded DNA
DSM-III	Diagnostic and Statistical Manual 3 rd Edition
DSM-IV	Diagnostic and Statistical Manual 4 th Edition
DTNBP1	Dysbindin
DZ	Dizygotic Twins
EPUFA	Essential polyunsaturated fatty acids
ExoI	Exonuclease I
fMRI	functional magnetic resonance imaging
GASP	Genetic association study of psychosis
GSMA	Genome Scan Meta-analysis
GSTM3	Glutathione S transferase M3
H ² _b	Heritability
H ₂ O	Water
HDL	High density lipoprotein
HSP70	Heat Shock protein 70
HSPA1	Heat shock protein 1
HSPA1L	Heat shock protein 1L
HSPA8	Heat shock protein 8
ICD-10	International Classification of Diseases 10 th Edition
Kb	Kilobases
KCl	Potassium Chloride
LCD	Lysergic acid diethylamide

LD	Linkage disequilibrium
LOD	Log of the odds
M	Molar
MAF	Minor allele frequency
Mb	Megabase
MgCl ₂	Magnesium Chloride
MIM	Mendelian Inheritance in Man
mM	Millimolar
MRI	Magnetic resonance imaging
mRNA	Messenger Ribonucleic Acid
MZ	Monozygotic Twins
NCBI	National Center for Biotechnology Information
NMDA-R	NMDA receptor
NRF2	Nuclear factor erythroid 2-like 2
NRG1	Neuregulin 1
OC	Obstetric complications
OD	Optical Density
OMIM	Online Mendelian Inheritance in Man
OR	Odds ratio
OSP	Offspring of schizophrenic parents
PCP	Phencyclidine
PCR	Polymerase chain reaction
PM	Post Mortem
PPR	Putative promoter region
PRODH	Proline dehydrogenase
RNA	Ribonucleic Acid
ROI	Region of interest
ROS	Reactive Oxidative Species
RPGI	Resource for psychosis genomics Ireland
SAP	Shrimp Alkaline Phosphatase
SCID	Structured clinical interview for DSM
SNP	Single Nucleotide Polymorphism
SZ	Schizophrenia
Taq	Taq polymerase
TCD	Trinity College Dublin
TDT	Transmission Disequilibrium Test
TEAA	Triethylamine acetate buffer
tSNPs	Tag SNPs
VCFS	Velo-cardio facial syndrome
λ_s	Relative Risk
χ^2	Chi square

Index of Tables

Chapter 1

Table 1.1 DSM-IV-TR Diagnostic Criteria for Schizophrenia	3-4
Table 1.2: Common formats of adoption studies.	16
Table 1.3 Comparison of genes investigated in this study with the meta-analysis results from Lewis et al (2004)	25

Chapter 2

Table 2.1 – SNP coverage density of all six APOL genes	55
Table 2.2 Composition of Agarose Gel loading dye	57
Table 2.3 – The four ddNTPs used in SNaPshot with respective fluorescence colour and dye name	59
Table 2.4 Composition of formamide loading dye	62
Table 2.5 Protein expression is significantly altered in the prefrontal cortex of schizophrenia (taken from Prabakaran et al 2004)	73
Table 2.6: Comparison of the chromosomal location of oxidative stress genes	75
Table 2.7: PCR primers designed to amplify 14 fragments spanning the gene HSPA8.	82-84
Table 2.8 Multiplex of SNPs and Fragments	90
Table 2.9 Reaction mixtures for PCR products multiplexed during the SAP and ExoI stage.	91
Table 2.10 SNaPshot multiplex reactions	91
Table 2.11 Extension Primers	92
Table 2.12: PCR primer details for the 14 fragments designed for DHPLC Analysis	98
Table 2.13 DHPLC conditions for the 14 fragments spanning GSTM3	101
Table 2.14: Details of HapMap and dbSNP SNPs that were used in SNP selection for this study	106

Chapter 3

Table 3.1: Results of validation experiments comparing pooled genotyping with individual genotyping using three SNPs.	111
Table 3.2 – Results of all DNA pooling association results for APOL6 and APOL5	113
Table 3.2 continued – Results of all DNA pooling association results for APOL3 and APOL4	114
Table 3.2 continued – Results of all DNA pooling association results for APOL2 and APOL1	115
Table 3.3 – SNPs that were subjected to stage 2 analysis	117

Chapter 4

Table 4.1 Results of single marker association analyses for G72	128
Table 4.2 D' and r ² values for the five markers genotyped in G72	128
Table 4.3 Haplotype analyses of G72 for 2 marker, 3 marker, 4 marker and 5 marker combinations	129
Table 4.4 Results of single marker association analyses of DAAO	131
Table 4.5 D' and r ² values for the four markers genotyped in DAAO	131
Table 4.6 Haplotype analyses of DAAO for 2 marker, 3 marker and 4 marker combinations	132
Table 4.7 SNPs that have information captured by the genotyped SNPs	137
Table 4.8 Comparison of results from published genetic association studies of G72/G30 and schizophrenia	147
Table 4.9: Comparison of results at the DAAO locus and schizophrenia	147

Chapter 5

Table 5.1 Extension primers and associated NCBI rs number	159
Table 5.2 Results of the association study on HSPA8-1, 3, 4, 5, 7 and 11 in the full sample	163
Table 5.3 Individual marker association results	170

Chapter 6

Table 6.1 MAF comparison of CEPH and Irish samples for 11 SNPs	179
Table 6.2 Alleles captured ($r^2 > 0.8$) in the Irish sample (16 SNPs)	186
Table 6.3 Individual marker association results	187
Table 6.4 Haplotype analysis of 2-, 3- and -4 markers using 10,000 iterations to provide an empirical simulated p value	187

Chapter 7

Table 7.1 Power calculations for the two sample sizes used in these studies	203
---	-----

Appendix

Table A.1 APOL 6 table PCR fragment information	255
Table A.2 APOL5 table PCR information	256
Table A.3 APOL4 table PCR information	257
Table A.4 APOL6 extension primers	258
Table A.5 APOL5 extension primers	259
Table A.6 APOL extension primers	260
Table A.7 PCR program cycles for T; Q and R program	263
Table A.8 PCR program cycles for T; Q and R programs cont/d.	264

Index of Figures

Chapter 1

Figure 1.1: Average risk (%) for the development of SZ in varying classes of relative.	15
Figure 1.2: Liability-threshold model.	20
Figure 1.3: LD around an ancestral mutation	29
Figure 1.4: Decay of LD by recombination.	32
Figure 1.5: Genetic Drift.	38
Figure 1.6 Admixture	41
Figure 1.7 Population Bottleneck	42

Chapter 2

Figure 2.1 – Overview of the SNaPshot primer extension procedure (taken from SNaPshot leaflet, Applied Biosystems)	57
Figure 2.2 Example of sample data for a SNP genotyped in the case DNA pool and control DNA pool using the SNaPshot™ reaction.	67
Figure 2.3 The results for each pool, and the respective difference between them, are plotted against the range of simulated k values.	68
Figure 2.4: HSPA8 fragments designed for sequencing	81
Figure 2.5 A schematic representation of how an electropherogram is created from a sequencing reaction	86
Figure 2.6 This is an example of an electropherogram of a sequence reaction	87
Figure 2.7 A schematic representing the discovery of a poly A tail adjacent to exon 8 that hindered resequencing of fragment 11	88
Figure 2.8 Schematic of GSTM3 showing two known transcripts one with 8 exons the other with 9 exons	94
Figure 2.9: Schematic of regions showing homology with chromosome 19	96
Figure 2.10 Schematic of the GSTM3 fragments designed for DHPLC	97
Figure 2.11 HapMap dump showing high LD structure around NRF2 on chromosome 2q31.2	105

Chapter 4

Figure 4.1 Haploview r^2 output using CEPH HapMap data of a 200Kb region on chromosome 13q containing the G72 locus	134
Figure 4.2 Haploview output of LD region 3	135
Figure 4.3 Haploview output of LD region 4.	136
Figure 4.4 Haploview output of DAAO region.	138
Figure 4.5 CEPH HapMap data of a 200Kb region on chromosome 13q containing the G72 locus.	143
Figure 4.6: Susceptibility and phenotype	145

Chapter 5

Figure 5.1 Output from SeqScape v2.1	158
Figure 5.2 Output from Seqscape of individual heterozygotes for HSPA-8	160
Figure 5.3 Haploview Output	161
Figure 5.4 Haploview Output	162
Figure 5.5 DHPLC analysis of GSTM3 fragment 7 at 60°C	168
Figure 5.6 Electropherogram of the forward primer reaction of fragment 7 amplified in patient 10	169

Chapter 6

Figure 6.1 Mode of action NRF2	175
Figure 6.2 Scatterplot of the MAF of SNPs in the CEPH and Irish samples	179
Figure 6.3 D' comparison between the CEPH sample and the Irish sample	180
Figure 6.4 Scatterplot of D' comparison between the CEPH and the Irish samples	181
Figure 6.5 r^2 comparisons of CEPH and Irish samples	182
Figure 6.6 Scatterplot of r^2 comparisons between the Irish and CEPH sample	183
Figure 6.7 r^2 output from Haploview of all 16 SNPs (11 HapMap and 5 dbSNP) that span NRF2	185
Figure 6.8 Upstream dissociation pathway of NRF2 and KEAP1	190

Figure 6.9 Diagrams taken from two studies which compared LD structure in local populations with HapMap CEPH data	196
--	-----

Appendix

Figure A1 APOL family	244
Figure A2 APOL VI and V all SNPs	245
Figure A3 APOL VI and V Heterozygous SNPs	246
Figure A5 APOL I and III Heterozygous SNPs	247
Figure A4 APOL I and III all SNPs	248
Figure A6 APOL I and III	249
Figure A7 APOL IV all SNPs	250
Figure A8 APOL IV SNPs for pooling	251
Figure A9 APOL IV Heterozygous SNPs	252
Figure A10 APOL II All SNPs	253
Figure A11 APOL II Heterozygous SNPs	254

Contents

Declaration	i
Summary	iii
Acknowledgements	v
Statement of work	vii
Publications	ix
Abbreviations	xi
Index of Tables	xiii
Index of Figures	xvii
Contents	xxi
Chapter 1 – Introduction	1
1.1 Clinical phenotype	1
1.1.1 Psychotic symptoms	5
1.1.2 Negative Symptoms	6
1.1.3 Cognitive impairment	6
1.2 Aetiology	7
1.2.1 Biological Aetiology	7
1.2.2 Environmental aetiology	12
1.3 Genetic factors	13
1.3.1 Genetic epidemiology	13
1.3.2 Mode of inheritance	18
1.3.3 Defining the phenotype	20
1.3.4 Cytogenetic abnormalities	21
1.3.5 Linkage Studies	23
1.3.6 Association Study	26
1.3.6.1 Population Stratification	27
1.3.7 Linkage Disequilibrium	29
1.3.7.1 Defining LD	29
1.3.7.2 Explaining LD	30

1.4 Recombination and Mutation	31
1.4.1 Variable mutation rates.	34
1.5 LD and power	34
1.5.1 Empirical Studies	35
1.5.2 Additional factors affecting LD	36
1.5.2.1 Genetic drift	36
1.5.2.2 Admixture	39
1.5.2.3 Population Bottlenecks, Inbreeding and Assortative Mating	42
1.6 Measures of linkage disequilibrium	43
1.7 Hypotheses	45
Chapter 2 – Materials and Methods	47
2.1 Description of sample	47
2.2 GASP study	48
2.3 DNA extraction from blood	48
2.4 DNA quantification	51
2.4.1 Spectrometry	51
2.4.2 Fluorimetry	52
2.5 DNA storage	54
2.6 Materials and Methods for the Apolipoprotein-L family	54
2.6.1 Identification of SNPs	54
2.6.2 Primer Design	55
2.6.3 PCR optimisation	55
2.6.4 PCR buffer protocol	56
2.6.5 Agarose Gels	56
2.6.6 Primer extension using SNaPshot	57
2.6.7 PCR	58
2.6.8 PCR Cleanup – SAP and ExOI	58
2.6.9 SNaPshot	60
2.6.10 SAP 2 stage	61
2.6.11 Run Samples	61

2.6.12 Sample Analysis	62
2.6.13 Individual genotyping by SNaPshot™ primer extension	63
2.6.14 DNA pooling	64
2.6.14.1 DNA pool preparation	64
2.6.14.2 DNA pooling method for testing association	65
2.6.15 Statistics	69
2.7 Materials and Methods for G72 and DAAO	70
2.7.1 Sample Size	70
2.7.2 SNP genotyping	70
2.7.3 Statistical Analysis	71
2.8 Materials and Methods - HSPA8	72
2.8.1 Identification of candidate genes	72
2.8.2 HSPA1A	72
2.8.3 HSPA1L	74
2.8.4 HSPA8	76
2.8.5 GSTM3	77
2.8.6. Mutation detection using resequencing	78
2.8.6.1 Sample Size	78
2.8.6.2 Identification of exonic and intronic structures	78
2.8.6.3 Primer Design	79
2.8.6.4 PCR optimisation	80
2.8.6.5 Resequencing	80
2.8.6.6 Comparing resequenced SNPs with NCBI's dbSNP	88
2.8.7 Multiplex genotyping of HSPA8 using SNaPshot™	89
2.8.8 Determining LD structure	92
2.8.9 Determining SNPs for genotyping in entire sample	92
2.9 Materials and Methods - GSTM3	93
2.9.1 Identification of exonic and intronic structures	93
2.9.2 Design of Fragments for DHPLC	95
2.9.3 PCR optimisation of fragments for DHPLC	98
2.9.4 Materials and Methods for Denaturing High Performance	

Liquid Chromatography (DHPLC).	99
2.9.5 Resequencing of putative mutations on GSTM3	102
2.9.6 Genotyping of SNPs for GSTM3	103
2.10 Materials and Methods for NRF2	103
2.10.1 Irish Reference Panel	107
2.10.2 Individual Genotyping of entire sample	107
Chapter 3 - Apolipoprotein L (APOL) 1- 6	108
3.1 Introduction	108
3.2 Results	110
3.3 Discussion	117
Chapter 4 - Positive Replication of G72/DAAO	121
4.1 Introduction	121
4.2 Results	126
4.2.1 Association analyses of G72/G30	127
4.2.2 Haplotype Analyses of G72/G30	127
4.2.3. Association Analyses of DAAO	130
4.2.4 Haplotype Analysis of DAAO	130
4.2.5 Retrospectively assessing the amount of variation captured in the Irish study	134
4.2.5.1 G72	134
4.2.5.2 DAAO	139
4.3 Discussion	141
Chapter 5 – Oxidative stress and schizophrenia: HSPA8 and GSTM3	148
5.1 Introduction	148
5.1.1 Plausibility of HSPA8 and GSTM3	155
5.2 Overview HSPA8	156
5.3 Results of HSPA8	157
5.3.1 Results – Resequencing of HSPA8	157
5.3.2 Results – Multiplex Genotypes using SNaPshot	160

5.3.3 Results - LD structure and analysis	161
5.3.4 Association Study in the full sample	162
5.3.5 Results - Determining SNPs for genotyping in entire sample	163
5.4 Discussion	164
5.5 Overview of GSTM3	165
5.6 Results for GSTM3	167
5.6.1 Results of DHPLC for GSTM3	167
5.6.2 Results – Resequencing of GSTM3	169
5.6.3 Results – Association Study of GSTM3	170
5.7 Discussion	171
Chapter 6 – NRF2	174
6.1 Introduction	174
6.2 Results	176
6.2.1 LD comparison of Irish and CEPH samples	177
6.2.2 Efficiency of CEPH tag SNPs in capturing additional SNPs at NRF2	184
6.2.3 Association study of NRF2 in the Irish SZ case-control sample	186
6.4 Discussion	188
Chapter 7 – General Discussion	197
7.1 Review of main findings – Association Studies	197
7.1.1 Apolipoprotein-L 1-6	197
7.1.2 G72 and DAAO	198
7.1.3 HSPA8 and GSTM3	199
7.1.4 NRF2	199
7.2 Strengths of the studies	200
7.3 Limitations of the studies	203
7.4 Future Directions	207
References	213
Appendix	243

Chapter 1 – Introduction

1.1 Clinical phenotype

Schizophrenia (SZ) is a common, severe, disabling disorder that in the majority of instances requires long-term medical and social care (Owen et al 2002). The morbid risk in the general population of developing SZ is 1% (Gottesman 1991). If one includes SZ spectrum disorders such as schizoaffective disorder, schizotypal personality disorder and paranoid personality disorder in this analysis, the morbid risk increases. The incidence of SZ is surprisingly uniform across populations. The annual incidence of SZ is 0.2-0.4 per 1000. The lifetime prevalence (risk) is about 1% (McGuffin et al 1995). Most sufferers of SZ have an onset in early adulthood. Although the incidence of SZ is the same in both sexes, women tend to have a later age of onset than men (Mueser and McGurk 2004). It is thought that women have a higher attainment of social functioning, due to the later onset, which in turn confers a better outcome with less hospital admissions and a more benign course of illness (Mueser and McGurk 2004). Affected individuals may have long periods of illness, are often unable to work and may have difficulties sustaining interpersonal relationships. SZ also has an enormous impact at a societal level through the economic costs of treatment and the social cost: the morbidity and mortality (~5% of individuals commit suicide) (Wyatt et al 1995).

The first comprehensive descriptions of SZ date from the 18th century. The modern definition of SZ was first formalised by Emil Kraepelin, a German psychiatrist. He noted syndromes of catatonia and hebephrenia and distinguished between ‘manic depressive insanity’ and ‘dementia praecox’. He viewed SZ as a single process occurring in youth with dementia as an outcome. However, not all individuals would follow this path of

illness. In 1911, Eugene Bleuler, a Swiss psychiatrist used the term, schizophrenia (SZ). He felt that SZ was more psychological than pathological with the core symptoms characterised “by a ‘loosening of associations’ in language, deficits in volition and attention and by incongruity of affect, ambivalence and autism” (Stefan et al 2002). The symptoms of psychosis, delusions and hallucinations, were, in his view, secondary effects.

However, Bleuler’s criteria were hard to define reliably. In comparison, Kraepelin’s criteria were more consistent because psychosis was more readily defined. In 1959, Kurt Schneider published a list of symptoms that led to a more reliable diagnosis of SZ (known as the first rank symptoms). This was highly influential in determining diagnostic practise. Methods of diagnosis have evolved since then to the present day. Psychiatrists now use two diagnostic systems for SZ: the International Classification of Diseases, tenth edition (ICD-10) and the Diagnostic and Statistical Manual, fourth edition (DSM-IV). These diagnostic systems are highly complementary and have improved the reliability of diagnosis for SZ. See Table 1.1 for DSM-IV TR criteria for SZ.

The essential elements to SZ are a mixture of characteristic signs and symptoms that have been present for a significant proportion of time during a 1-month period, with some signs of the disorder persisting for at least six months. They involve a range of

cognitive and emotional dysfunctions in domains that include perception, inferential thinking, language and communication, behavioural monitoring, affect, fluency and productivity of thought and speech, volition, drive and attention. A group of these signs and symptoms confer diagnosis but individuals may differ substantially in their symptomatology. Impaired role functioning or change in personal behaviour are also

Table 1.1 DSM-IV-TR Diagnostic Criteria for Schizophrenia.

- A. *Characteristic Symptoms:* Two (or more) of the following, each present for a significant portion of time during a 1-month period (or less if successfully treated):
- (1) delusions
 - (2) hallucinations
 - (3) disorganised speech (e.g. frequent derailment or incoherence)
 - (4) grossly disorganised or catatonic behaviour
 - (5) negative symptoms, i.e., affective flattening, alogia, or avolition

Note: 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.

- 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 sign of the disturbance persists for at least six months. This six-month period must include at least one 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. During these prodromal or residual periods, the signs of the disturbance may be manifested by only negative symptoms or two or more symptoms listed in criterion A present in attenuated form (e.g. odd beliefs, unusual perceptual experiences).

Table 1.1 Continued on next page

- D. *Schizoaffective and mood disorder exclusion:* Schizoaffective 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 relative to the duration of the active and residual periods.
- 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:* Is there a history of Autistic Disorder or another pervasive developmental disorder, the additional diagnosis of SZ is made only if prominent delusions or hallucinations are also present for at least a month (or less if successfully treated).

Classification of longitudinal course (can be applied only after at least 1 year has elapsed since the initial onset of active-phase symptoms):

Episodic with interepisode residual symptoms (episodes are defined by the re-emergence of prominent psychotic symptoms) also specify if: **with prominent negative symptoms episodic with no interepisode residual symptoms**

Continuous (prominent psychotic symptoms are present throughout the period of observation) also specify if: **With prominent negative symptoms**

Single episode in partial remission; also specify if: **With prominent negative symptoms**

Single episode in full remission

Other or unspecified pattern.

Adapted from DSM-IV TR 2003 (American Psychiatric Association), p.312-313

Diagnostic systems such as the DSM-IV have improved reliability of diagnosis. Two or more of the signs and symptoms listed here must be present for a substantial proportion of time during a one month period before a diagnosis of schizophrenia can be reached.

included in the diagnostic criteria: reduced ability to work; attend school; parent; have close relationships; take care of oneself; enjoy leisure time; all of these can emerge several years before psychotic symptoms (Mueser and McGurk 2004).

SZ is characterised by the three overlapping forms of deficit termed positive symptoms, negative symptoms and cognitive impairment. Positive symptoms include distortions in thought content (delusions), perception (hallucinations), language and thought processes (disorganised speech) and self-monitoring of behaviour (grossly disorganised or catatonic behaviour). Negative symptoms include restrictions in the range and intensity of emotional expression (affective flattening), in the fluency and productivity of thought and speech (alogia), and in the initiation of goal-directed behaviour (avolition). Cognitive deficits in performance of tasks involving attention, working memory and executive function are also a feature of SZ.

1.1.1 Psychotic symptoms

Psychotic symptoms involve the loss of contact with reality, including false beliefs (delusions), perceptual experiences not shared by others (hallucinations) or bizarre behaviours. Hallucinations occur in many ways: auditory (the most common); visual; olfactory; gustatory or tactile hallucinations. The most common delusions in SZ include persecutory, control (e.g. the belief that others can interfere with their thoughts), grandiose (e.g. the belief that they are Jesus Christ) and somatic (e.g. the belief that their brain is rotting away). The severity and presence of psychotic symptoms seems to be episodic over time (Mueser and McGurk 2004).

1.1.2 Negative Symptoms

Negative symptoms are deficit states in which basic emotional and behavioural processes are diminished or absent influencing affective response. This includes blunted affect with immobile facial expression or a monotonous voice tone, a lack of pleasure or emotion known as anhedonia, diminished ability to initiate and follow through plans, termed avolition or apathy and reduced quantity or content of speech, alogia. Negative symptoms are more pervasive, tend to be stable over time and are strongly associated with poor psychosocial functioning (Mueser and McGurk 2004).

1.1.3 Cognitive impairment

Cognitive impairment in SZ includes problems in attention and concentration, psychomotor speed, learning and memory and executive functioning (e.g. abstract thinking, problem solving). A decline in cognitive performance in patients compared to their healthy state is present in most patients with SZ. This impairment remains stable over time (Mueser and McGurk 2004).

As there are many possible signs and symptoms, SZ is effectively a syndrome (Owen et al 2002). At present, researchers are ignorant of the pathogenesis of SZ and as yet, no biological markers to aid in diagnosis have been discovered. However, current diagnostic criteria may well include a number of heterogenous disease processes. If these processes were delineated it may aid in identifying genes and other aetiological factors quicker (Owen et al 2004). It must be stressed however, that current diagnostic systems (DSM-IV and ICD-10) give reliable and reproducible diagnosis.

Results from family, twin and adoption studies have shown that the phenotype of SZ may include a broad spectrum of other disorders such as schizophreniform disorder, schizoaffective disorder and schizotypal personality disorder. There is also uncertainty over the relationship of SZ and other affective psychoses and nonpsychotic affective disorders (Owen et al 2004).

1.2 Aetiology

1.2.1 Biological Aetiology

The first biological theories on the pathophysiology of SZ came from neuropharmacological and psychopharmacological studies that suggested abnormalities in monoamine neurotransmission, in particular the dopaminergic and serotonergic systems. For more than three decades the prominent hypothesis was that excessive dopaminergic neurotransmission caused SZ (Carlsson 1988). Antipsychotic drugs block dopamine receptors to a degree that is proportionate to their efficacy. Secondly, amphetamine, which is a dopaminergic drug, produces psychomimetic effects. However, dopaminergic pathophysiology has never been established as a primary cause of SZ. It is therefore possible that these neurochemical abnormalities are due to compensatory mechanisms, downstream pathology or perhaps environmental influences.

In recent years newer atypical anti-psychotic drugs have been developed which target other neurotransmission systems such as serotonin (5-HT), GABA and glutamate (Bray and Owen 2001). The 5-HT hypothesis stems from several observations. First, lysergic acid diethylamide (LSD) binds to 5-HT receptors producing psychomimetic and psychogenic symptoms. Second, there is strong evidence to suggest that 5-HT_{2A} receptors are altered in

SZ brains. Third that atypical antipsychotics such as clozapine, bind to 5-HT_{1A} and 5-HT_{2A} receptors (Miyamoto et al 2003). Therefore there is strong evidence to also suggest 5-HT involvement in SZ pathogenesis.

Phencyclidine (PCP) and Ketamine induce SZ-like symptoms in healthy individuals. Both of these drugs are non-competitive antagonists of the NMDA glutamate receptor NMDA-R (Miyamoto et al 2003). Therefore it is thought that decreased NMDA-R functioning may be involved in the pathogenesis of SZ. In addition, these antagonists appear to produce frontal cognitive deficits and post mortem studies have highlighted abnormalities in pre- and postsynaptic glutamatergic pathways. Furthermore, partial deletion of the NMDA-R1 subreceptor in mice produces the same behavioural pattern as PCP induced psychosis in humans (Miyamoto et al 2003). The exact mode of action of these drugs is still unknown. Atypical antipsychotic drugs are still very much in development and greater understanding of the pathophysiology of SZ will aid research in this area. For a fuller understanding of this development, please refer to Miyamoto et al (2003).

More recently, it has been proposed that SZ is a neurodevelopmental disorder that leads to synaptic connectivity abnormalities (Weinberger 1995, Owen et al 2004). Hundreds of published MRI and CT studies have now conclusively shown that patients with SZ differ in various structural brain measurements compared with control patients (Weinberger 1995). The most frequently reported difference is cerebral ventricular enlargement in both chronic treatment patients and first diagnosis patients before chronic treatment. Although there is considerable variation in normal ventricular size, which in turn makes the effect size of this finding small, MZ twin studies, which control for normal variance, have shown that there is a subtle increase in ventricular size for the affected twin (Weinberger 1995). In addition, many studies have shown that there is a statistically significant reduction in cortical

volume in the temporal, prefrontal and parietal lobes (Weinberger 1995). Furthermore, studies on cortical organisation during the second trimester have shown that neurons seem to be going in the opposite direction from the periventricular zone to the pial cortical surface. A cytoarchitectural defect in this “inside out” organisation of cortical layers suggests a defect in the process of neuronal migration during the neurodevelopmental stage (Weinberger 1995). However, there are methodological limitations, negative studies, and the positive studies disagree as to the nature of the alterations (Harrison and Weinberger 2005).

It would be correct to assume that if abnormal neurodevelopment has taken place we should see subtle signs of this in early life. In general, patients who developed SZ in adult life performed less well on neuropsychological tests and at school achievements (Weinberger 1995). Several prospective studies have been carried out in offspring of schizophrenic parents (OSPs). All of these studies showed a higher prevalence rate in SZ and schizoaffective disorder in OSPs than in normal and other psychiatric controls (reviewed by Erlenmeyer-Kimling 2000). Neurological and motor functioning deficits in infancy and mid-childhood have been reported in OSPs. Deficits in cognitive ability have been reported in OSPs, in particular, measures of attention and verbal short term memory. However, there is conflicting evidence of a relationship between obstetric complications and neurological and motor functioning (Erlenmeyer-Kimling 2000) suggesting that there are no definitive liability indicators as yet.

Another avenue of research was that enlarged CSF spaces were indicative of neurodegeneration. However, this line of thought had been dismissed as cross sectional studies of SZ patients scanned ten years apart found no consistent changes in CSF volume (Weinberger 1995). Replication of enlarged CSF spaces has been problematic (Harrison

1999) and the argument for a degenerative process continues (DeLisi 1997; Lieberman 1999). Subsequent studies in childhood onset schizophrenia (Rapoport et al 1997) found neurodegeneration in certain brain structures of 7 – 12 year olds. However, this degeneration appeared to cease at puberty. In addition, some studies followed up adult SZ patients and suggest the possibility of a subgroup of patients in which neurodegeneration occurs (Pearlson and Marsh 1999). Therefore it would still be premature to dismiss entirely a neurodegenerative process in SZ (Lieberman 1999).

Interest in necropsy evidence for developmental pathology resurged after MRI and CT scans. Although cerebral ventricular enlargement (enlarged CSF spaces) indicates neurodegeneration, the most consistent finding is the absence of gliosis. Glial reactions are involved in neurodegeneration. However, neuropathological events in neurodevelopment do not involve gliosis which presents further circumstantial evidence that SZ is neurodevelopmental (Weinberger 1995). There are also at least four studies implicating a defect in the formation of the cortical plate (Weinberger 1995). In addition, current literature suggests macroscopic neuropathology in SZ. If neurodegeneration did occur, one may expect to see Alzheimer's type symptoms such as impaired cognition. However histological findings suggest this is not the case. Therefore it appears that the reported impaired cognition seen in SZ patients is not due to neurodegeneration.

Essentially histology investigations point to dysfunction in synaptic connectivity. Histological findings have reported abnormally clustered neurons in the lamina II of the entorhinal cortex and neocortical white matter. In addition, reduced density of pyramidal neurons in the hippocampus and neocortex as well as being smaller with reduced dendritic spines is indicative of neurodevelopmental abnormality (Harrison and Weinberger 2005). Some studies have reported volumetric and reduced NAA expression in first episodic and

medication free subjects lending support that these findings are not the result of merely illness or treatment. However, the literature remains conflicting and may be due to small sample sizes and the low number of studies performed and treatment artefact as causes (Harrison and Weinberger 2005).

Finally there have been reports of adverse environmental events in utero, which may be involved in neurodevelopmental aetiology. Examples include the Helsinki influenza A2 epidemic of 1957 where foetuses exposed in the second trimester had a significant increase in hospitalisation for SZ compared with first and third trimester foetus' and births for subsequent years. A Dutch study also reported a clear correlation between increased risk and hospitalisation for SZ and exposure to severe famine in early gestation during the Nazi blockade of the Netherlands (Susser et al 1992).

More recently a role for oxidative stress in SZ pathophysiology has been suggested. A study by Sabine Bahn's team in Cambridge (Prabakaran et al 2004) found many significantly altered mitochondrial and oxidative stress mRNA and proteins in SZ post mortem brains compared to controls. The role of cellular oxidative stress in pathology has been investigated in detail elsewhere (e.g cardiovascular and oncology). However, investigations of oxidative stress in psychiatric disorders are relatively new.

Essentially, normal cell physiology produces by-products termed reactive oxidative species (ROS). These ROS are detrimental to cellular sustainability and are captured and destroyed through a complex process of Phase II and anti-oxidant response element (ARE) genes. However, when this mechanism fails, the cell goes into oxidative stress. Neuronal cells are particularly susceptible to this type of stress. Therefore it is hypothesised that a disruption in one or many genes involved in cellular anti-oxidant response harbours a functional

mutation leading to increased neuronal cellular oxidative stress. This results in abnormal cell development or function and hence development of the disorder.

1.2.2 Environmental aetiology

Environmental factors have a role in the aetiology of SZ. They have been shown to marginally increase the risk of developing SZ (Cannon et al 2002). Increased risk has been associated with prenatal and perinatal events including maternal influenza, rubella, malnutrition, diabetes mellitus, smoking during pregnancy and obstetric complications (Diabetes in pregnancy, placental abruption, birth weight < 2Kg e.g. premature baby and caesarean section to name a few) (Cannon et al 2005, Cannon et al 2002). However, none of these events are predictive of developing SZ. Obstetric complications associated with hypoxia, are particularly related to increased risk. It is thought that excitotoxic effects of hypoxia on the foetal neonatal brain might interact with genetic vulnerability to increase the risk of developing SZ. (Maier and Schwab 1998)

Epidemiological studies have identified putative environmental risk factors associated with an increased rate of SZ. This includes complications at birth and during pregnancy (Hultman 1999), delayed developmental milestones (Jones 1994), maternal viral exposure such as influenza during pregnancy, low IQ score or winter births (David et al 1997), personal characteristics concerned with social relations (Malmberg et 1998), urban upbringing (Mortensen et al 1999), immigration (Hutchinson et al 1996), smoking and diabetes mellitus (Meuser and McGurk 2004) and the misuse of illegal drugs such as cannabis (Andreasson et al 1987, Dean et al 2001). Obstetric complications involving hypoxia are at increased risk due to the excitotoxic effects of hypoxia on the foetal

neonatal brain (Meuser and McGurk 2004: Cannon et al 2002). However, most obstetric complications do not lead to SZ and it may be that this environmental effect interacts with genetic susceptibility (Meuser and McGurk 2004).

There are also sociodemographic factors associated with increased risk with SZ. Poverty and lower social class have long been linked to higher rates of SZ. There are currently two hypotheses for this statement. First, social causation (stressful environment conditions increase the risk of SZ) and secondly, downward social drift. This means that SZ reduces social and occupational functioning. (Maier and Schwab 1998)]

Another phenomenon is that individuals born in urban areas are more likely to develop SZ than those in rural areas. Although the incidence of SZ in ethnic groups is similar it has been found that increased rates have been found in second generation Afro-Caribbean people in the UK, Dutch Antillean and Surinamese immigrants in Holland and African-American people. This may reflect the stressful effects of being an ethnic minority in a social environment, which may increase vulnerability in genetically susceptible individuals (Maier and Schwab 1998).

1.3 Genetic factors

1.3.1 Genetic epidemiology

There is an extensive literature on the genetic epidemiology of SZ. Gottesman and Shields (1972) conclusively demonstrated that the risk of SZ is increased in relatives of probands with the disorder, based upon data from some 40 European family and twin studies strating

from 1921 (See Figure 1.1, Gottesman 1991). As mentioned earlier the lifetime risk for SZ is roughly 1%. In siblings of probands this increases to 9% and in offspring of probands to 13%. As the pathology of SZ makes it increasingly difficult to have interpersonal relationships, therefore reducing opportunities to reproduce, it is not surprising that the risk to parents of probands is lower at 6%. In these cases, it is thought that the onset of illness has occurred after child bearing age or it is thought that the parents are 'selected for health' in that they had a predisposition to SZ that never manifested (McGuffin et al 1995). Dual matings, where two schizophrenics mate, produce offspring with a 46% risk of developing this disorder (Gottesman and Bertelsen 1989). It can be argued that many of the studies included in the analysis of Gottesman and Shields (1972) predated contemporary methods. However, in a review by Kendler (2000) of 11 controlled studies using blinded, structured interviews, the findings of a roughly 10% increase in risk to siblings and offspring of probands remained. Thus, the facts mentioned above show that there is evidence to suggest that SZ has a familial component. However, it cannot be assumed that this is genetic.

In order to determine whether SZ is genetic rather than just familial from environmental effects we need to look at twin and adoption studies. Adoption studies use three study designs (adoptive studies, cross-fostering studies and adoptive family studies) (see Table 1.2). The evidence from these three types of study provide corroborative evidence that there is an increased risk of SZ in biologically first degree relatives of probands and not with non-biologically related adopted or adoptive family members who share the same environment as probands (Owen et al 2002: Heston 1966, Rosenthal et al 1971, Kety et al 1994, Wender et al 1974). The first adoption study in SZ by Heston (1966) looked at 47 adoptees that had been separated within three days from birth from their schizophrenic parent. Five developed SZ compared to zero for 50 control adoptees. This suggested that shared genes rather than shared environment increased susceptibility of developing SZ.

Lifetime Risk of developing SZ

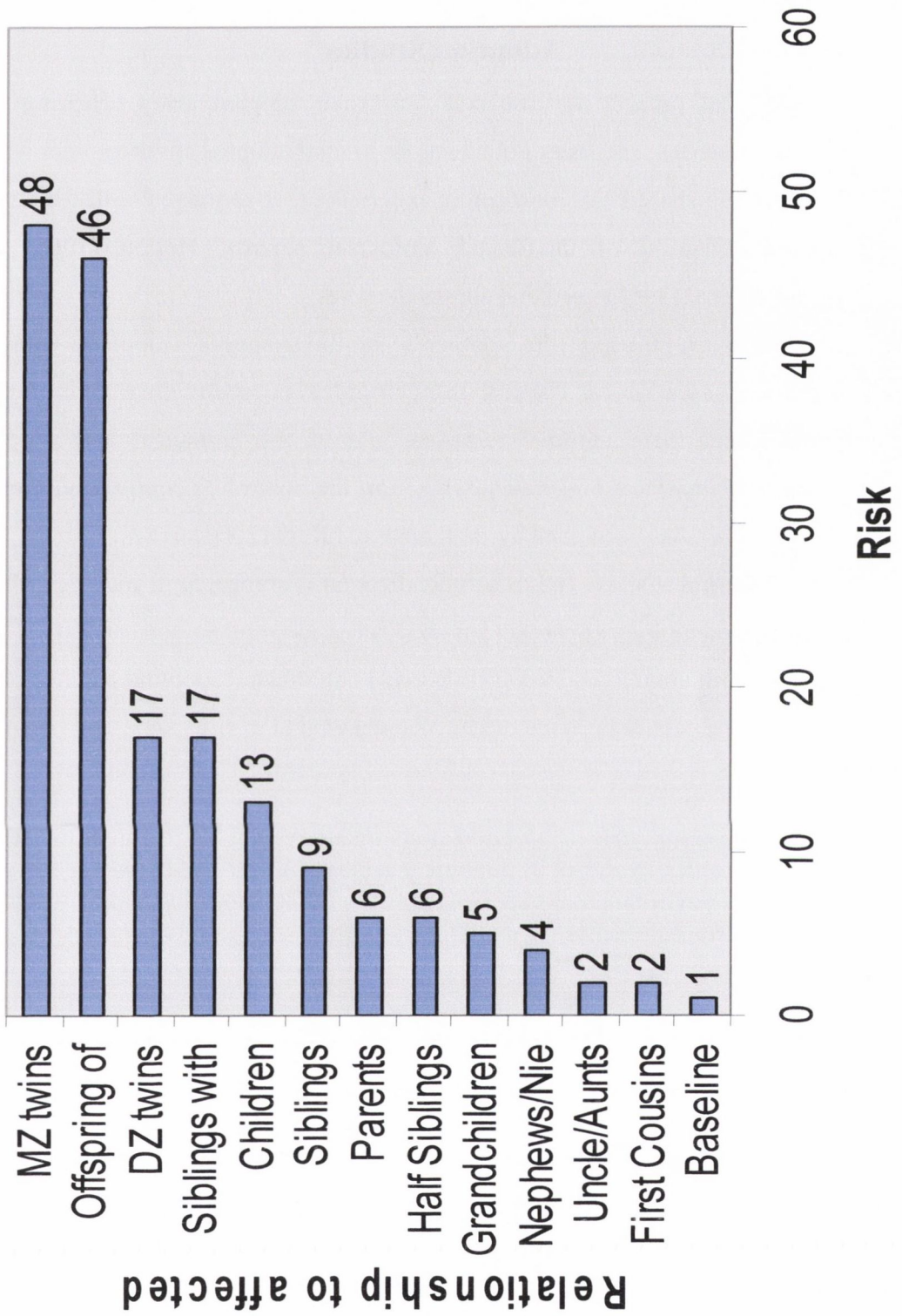


Figure 1.1: Average risk (%) for the development of SZ in varying classes of relative. This figure shows different classes of relative differing in their degree of genetic relatedness and their corresponding average risk of developing SZ related psychosis. The figure shows risk increasing as the number of shared genes also increase. (Adapted from McGuffin et al 1995).

Table 1.2: Common formats of adoption studies.

Adoption Studies

- 1 Ill biological parents as probands who have adopted away offspring (the adoptee strategy). The risk of the disorder in their adopted children, with whom they have no shared environment or experiences, is compared with the risk in adopted offspring of unaffected biological parents. Heston (1966) and Rosenthal et al (1968) used this approach in SZ.
- 2 Ill adoptee as proband (the adoptee's family strategy). Adoptees with the disorder are ascertained. The risk of the disorder is compared in their biological relatives and their adoptive relatives, and in the biological and adoptive relatives of unaffected adoptees (who form the control or comparison group). This approach has been used by Ingraham and Kety (2000) in SZ.
- 3 Cross-fostering studies. The risk of the disorder is compared in adoptees who have affected biological parents but unaffected adopting parents, and in adoptees with unaffected biological parents but affected adopting parents. Difficult to obtain but carried out by Wendel et al (1974) in SZ.

Comparison of the three types of adoption study undertaken in SZ research. In each case, if the risk of the disorder is higher in biological relatives of probands (where the genes are shared but not the environmental experiences) than in the control group, it would suggest that a genetic contribution exists.

More recently, a Danish adoption study (Ingraham and Kety 2000) was re-analysed using DSM-III criteria. This concurred that there was a major genetic influence with SZ. In addition, a Finnish study (Tienari et al 2000) found a higher morbid risk for DSM-III-R SZ in the adopted-away offspring of mothers with SZ compared with control offspring (8.1% vs. 2.3%).

Adoption studies therefore suggest that shared genes rather than shared environment increase the risk in relatives of probands (Owen et al 2002, McGuffin et al 1995). To endorse this view further we turn to twin studies. Monozygotic (MZ) twins share 100% of their genes whereas dizygotic (DZ) twins share on average 50% of their genes (McGuffin

et al 1995). If we assume that both types of twins share the same environmental influences to roughly the same extent, then a significantly higher concordance rate for SZ in MZ twins than DZ twins should indicate a genetic aetiology (McGuffin et al 1995). Recently, five systematically ascertained twin studies showed that the probandwise concordance rate for SZ in MZ twins is 41-65% compared with 0-28% for dizygotic twins. This corresponds to heritability estimates of approximately 80-85% (Cardno and Gottesman 2000). Broad sense heritability (h^2_b) is defined as the proportion of total phenotypic variance in liability that is accounted for by total genetic effects (both additive and non-additive). The broad sense of heritability is the proportion of total phenotypic variance (V_p , or variance in liability) accounted for by total genetic variance (i.e. additive and non-additive genetic effects, V_g) (Cardno and McGuffin 2002)

$$h^2 = V_g/V_p$$

Although there are biases in ascertaining twin samples which might artificially inflate a genetic effect, such as including only the most memorable pairs of twins concordant for SZ, methods are in place to address this issue. One such method is to systematically sample twin registers, such as the hospital-based register at the Maudsley Hospital, London, UK or by matching twins recorded on national registers with those twins receiving psychiatric treatment (McGuffin et al 1995). Another potential assumption in twin studies is the assumption of equal shared environmental experiences. One study addressed this issue (reviewed by McGuffin et al 1995) where 12 sets of MZ twins were raised apart. The concordance rate for SZ was reported as 58%.

Genetic epidemiology therefore points to clearly genetic aetiological mechanisms. Moreover the relative risks associated with close genetic relatedness to an affected subject

are considerably greater than those conferred by the putative environmental risk factors we have discussed above.

1.3.2 Mode of inheritance

Although studies have shown that there is a large genetic contribution to SZ, it is apparent from MZ twin studies that what is inherited is a predisposition to SZ and not the certainty of disease. This is supported by the fact that offspring of discordant MZ pairs show comparable risks in the offspring of both affected and non-affected co-twins (Owen et al 2002; Gottesman and Bertelsen 1989, Kringlen and Cramer 1989). This suggests that specific genetic factors conferring susceptibility to SZ are transmitted, but that they need not be expressed.

The recurrence risk for relatives of probands decreases too rapidly with increasing genetic distance from the proband for SZ to be a single-gene disorder, or collection of single gene disorders, even when incomplete penetrance is taken into account (Risch 1990). It is felt rather, that the mode of transmission is like that of other common disorders is complex and non-Mendelian (Owen 2000). The commonest mode of transmission is probably oligogenic, polygenic or a mixture of the two with a threshold effect (McGuffin et al 1995). However, the number of loci, the disease risk conferred by each locus and the degree of interaction among loci all remain unknown. Risch (1990) has calculated that the data for recurrence risk in the relatives of probands with SZ are incompatible with the existence of a single locus conferring a relative risk (λ_s) of > 3 and unless extreme epistasis exists, models with two or three loci of $\lambda_s \leq 2$ are more plausible. It should be emphasised that these calculations are based upon a model of homogeneity. It is possible that alleles of

larger effect operate in families with a high density of SZ. However, McGue and Gottesman (1989) have demonstrated that such families would be expected to occur even under polygenic inheritance and their existence does not conclusively prove that alleles of large effect exist.

A correlation in liability can be calculated statistically using data on the population risk for SZ. The idea of a liability-threshold model was first proposed by Falconer (1967) and applied to SZ by Gottesman and Shields (1967). It assumes that the liability to develop a disorder is normally distributed in the population and that this distribution reflects the additive effects of several different genes and environmental factors (see figure 1.3). Individuals below a particular threshold are unaffected whereas individuals who at some time exceed this liability-threshold develop the disorder (McGuffin 1995). The position of this threshold is determined by the lifetime morbid risk of the disorder in the population (see figure 1.2). The liability distribution is set to have a mean = 0 and variance = 1 (Cardno and McGuffin 2002). The distance between the mean and the threshold are measured in standard deviations (z-scores) with the threshold designated as the origin, zero (Cardno and McGuffin 2002). Graph (a) shows the population liability distribution for an illness with a 3% morbid risk, with X1 representing the distance between threshold and mean population liability. Graph (b) shows a shift to the right for relatives of affected individuals, with the distance between threshold and mean population liability, X2, less than X1. This gives an increased percentage of affected individuals under the normal distribution graph.

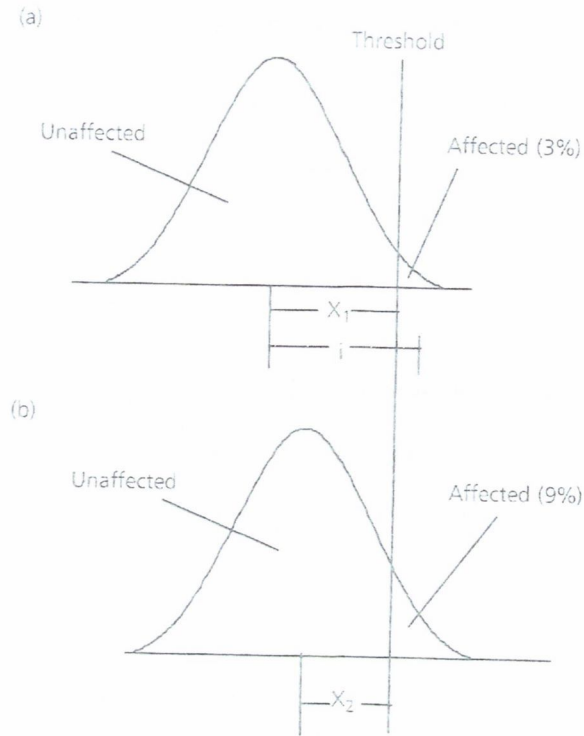


Figure 1.2: Liability-threshold model. (a) Population liability distribution for an illness with a 3% morbid risk. (b) liability distribution for relatives of affected individuals (9% recurrence risk). X_1 is the distance between the threshold and the mean population liability. X_2 is the distance between the threshold and mean liability of relatives. (Adapted from Sham and McGuffin 2002)

1.3.3 Defining the phenotype

The use of semi-structured and structured interviews together with explicit operational diagnostic criteria means that it is possible to achieve a high degree of diagnostic reliability. In addition, SZ has been shown to have a high heritability. Therefore it should be possible to subject this syndrome to genetic analysis. Yet our attempts at identifying susceptibility genes would be greatly facilitated if we could delineate the putative heterogeneous disease processes involved in SZ aetiology.

Further to this problem family, twin and adoption studies have shown that genetic liability extends beyond the core phenotype of SZ to include spectrum disorders such as schizoaffective disorders and schizotypal personality disorder. The limits of these spectrum

disorders to other psychoses are as yet uncertain (Owen 2004; Kendler et al 1998; Teinari et al 2000; Cardno et al 2002).

1.3.4 Cytogenetic abnormalities

Cytogenetic investigations of affected individuals have also been another approach used by researchers to locate susceptibility genes. The main abnormalities involved are translocations and deletions. These pathogenic processes may cause direct disruption of a gene or genes, the formation of a new gene comprised of a fusion of two genes that are normally spatially separated, by indirect disruption of gene dosage in the case of deletions, duplications, and unbalanced translocations. It is possible for a cytogenetic abnormality to be linked to a susceptibility variant in a particular family (Owen et al 2002). However, a single incidence of a cytogenetic abnormality is insufficient to suggest causality applicable to the general population due to the relatively high prevalence of SZ. In order for a cytogenetic abnormality to be of use in determining the pathophysiology of SZ in the greater affected population it should be shown to exist in greater frequency in affected individuals, to disrupt a region already implicated in genetic analyses, or to show co-segregation with the condition in affected families.

There have been numerous reports of association between SZ and chromosomal abnormalities (Basset et al 2000; Baron 2001). However, many have not provided convincing evidence to support a specific locus conferring susceptibility to SZ (Owen et al 2002). Only two chromosomal abnormalities have as yet provided such evidence. The balanced translocation (1;11)(q42;q14.3) is found to co-segregate with SZ and other psychiatric disorders in a large Scottish family (Blackwood et al 2001) giving a LOD score

of 3.6 (SZ phenotype only); 4.5 (affective disorders); and 7.1 (all affected relatives with SZ, bipolar and other affective disorders). Millar et al (2000) reported that the translocation directly disrupts three genes of unknown function on chromosome 1. They identified two of these genes as DISC1 and DISC2. Interestingly, these two genes are located close to genetic markers that were implicated in two Finnish linkage studies (Hovatta et al 2000; Ekelund et al 1999). DISC2 contains no open reading frame and may regulate DISC1 expression via anti-sense RNA. However, until we discover the biological function or mutations at this locus, we cannot understand the mechanism by which this translocation confers risk to SZ. Evidence from functional studies suggests that the function of DISC1 might be relevant to SZ. For example, the truncated product in the translocation family might contribute to SZ by affecting neuronal functions dependent upon intact cytoskeletal regulation such as neuronal migration, neurite architecture, and intracellular transport (Miyoshi et al 2003, Ozeki et al 2003). However, it may be possible that translocations exert effects on genes other than those directly disrupted. For example, there are several mechanisms by which a translocation can influence the expression of neighbouring genes. Thus, in order to unequivocally implicate DISC1 and/or 2 in the pathogenesis of SZ, it is necessary to identify mutations or polymorphisms that are associated with SZ in another population and are not in linkage disequilibrium with neighbouring genes (Craddock et al 2005). To date there have been two negative association findings and two positive association findings published. Further studies in different populations are still required in order to unequivocally implicate DISC1 in the pathogenesis of SZ.

The other chromosomal abnormality associated with SZ is the small interstitial deletion of chromosome 22q11. This deletion causes velo-cardio facial syndrome (VCFS) which is also known as DiGeorge syndrome. Although the phenotype of VCFS is variable, there is strong evidence that individuals with VCFS have a 30 fold increased risk of psychosis, in

particular, SZ (Murphy et al 1999; Pulver et al 1994). The prevalence of VCFS is 1:4000 live births. Therefore VCFS is responsible for less than 1% of SZ cases (Karayiorgou et al 1995). However, the higher rate in prevalence of SZ in VCFS patients suggests that a gene or genes in the deleted region confers increased risk. Further evidence in support of this approach comes from linkage findings that suggest susceptibility loci in the 22q region. Currently, studies are looking at genes in the 22q11 region in SZ patients who do not possess the deletion (Arinami et al 2001). Although the published linkage findings positions do not exactly match the VCFS region, linkage mapping in complex disease is imprecise and modest evidence for linkage (Blouin et al 1998; Lasseter et al 1995; Shaw et al 1998) and for linkage disequilibrium (Owen et al 2002) within the VCFS region has also been observed. Indeed, SNP based haplotype and LD analysis of 22q11 in SZ implicates this region (Li et al 2000). However, there have been reports of negative findings in this region, with the gene proline dehydrogenase (PRODH) that maps to 22q11 and is also an excellent functional candidate. However, a recent study by Williams et al (2003) found no association between 9 SNPs (8 exonic) at this locus and SZ in a large sample (n=368 cases, controls).

1.3.5 Linkage Studies

Linkage analysis is based on the finding that when chromosomal DNA recombines during meiosis, genetic loci that are close to each other are more likely to be co-inherited than more distant loci (Tsuang et al 2004). Linkage studies examine whether a disease is transmitted together with a gene or genetic marker in families with multiply affected individuals (Maier and Schwab 1998). Evidence for linkage occurs if the transmission of alleles occurs more often than would be expected by chance. This is called co-segregation.

This is possible by using a few hundred markers, evenly spaced that cover the entire genome, without requiring a priori knowledge of the disease aetiology (Owen et al 2004). Linkage is particularly useful for detecting genes of large effect in Mendelian disorders. However, the power of linkage decreases when trying to detect genes of moderate to small effect which are likely to be involved in complex genetic disorders such as SZ. In the case of complex disorders, where individual effects are modest it also requires the recruitment of a large number of families in order to achieve reasonable power (Owen et al 2004). Despite these caveats, linkage studies of families with SZ have led to chromosomal regions that contain genes later implicated in SZ by association analyses such as Dysbindin (DTNBP1) and Neuregulin (NRG1). LOD scores are the measure of the strength of linkage. Values of 1.9 – 3.3 are suggestive of linkage (Lander and Kruglyak 1995). When the mode of inheritance is uncertain, a LOD score of 3 no longer implies evidence of strong linkage. Lander and Kruglyak (1995) suggested LOD scores above 3.3 being evidence for linkage, which reduces the probability of a significant finding occurring by chance to 0.05 per genome scan. A cumulative LOD score of -2 should be regarded as strong evidence against linkage (Morton 1955). The affected sib-pair analysis is the most robust procedure if the mode of transmission is unknown.

Early independent linkage findings were reproduced on chromosomes 5q, 6p, 8p, 13q and 22q (Maier and Schwab 1998). However these studies did not provide genome-wide significance. There have also been problems with replication. Prior to a meta-analysis of linkage studies, only three chromosomal regions had reached genome-wide significance (Lewis et al 2003). These are 1q21-22, 13q32-34 and 6p24-22 (Owen et al 2004). To explore this further, Lewis et al (2003) performed a meta-analysis of 20 published genome scans using the rank-based Genome Scan Meta-Analysis (GSMA) method (Wise et al. 1999). Although their study met criteria for linkage in two of the previously reported

genome-wide significant regions (1q and 6p) it did not find any evidence for linkage on 13q. Only a chromosome 2q region was identified as reaching genome-wide significance in the Lewis et al (2003) study. In addition, they identified 19 significant bins of chromosomal regions as providing significantly more evidence for linkage than would be expected by chance. Three of these bins (see Table 1.3) contained some of the genes involved in this study (regions 22pter-q12.3, 11q22.3-q24.1 and 1p13.3-q23.3).

Table 1.3 Comparison of genes investigated in this study with the meta-analysis results from Lewis et al (2004)

Chromosome Bin ^a	Cytogenetic location	Both p values at <.05 ^b	Genes involved in this study
2.5	2p12-q22.1	yes	
5.5	5q23.2-q34	yes	
3.2	3p25.3-p22.1	yes	
11.5	11q22.3-q24.1	yes	HSPA8
6.1	6pter-p22.3	yes	
2.6	2q22.1-q23.3	yes	
1.6	1p13.3-q23.3	yes	GSTM3
22.1	22pter-q12.3	yes	Apolipoprotein-L 1-6
8.2	8p22-p21.1	yes	
6.2	6p22.3-p21.1	yes	
20.2	20p12.3-p11	yes	
14.1	14pter-q13.1	yes	
16.2	16p13-q12.2		
18.4	18q22.1-qter		
10.1	10pter-p14		
1.7	1q23.3-q31.1		
15.3	15q21.3-q26.1		
6.4	6q15-q23.2		
17.3	17q21.33-q24.3		

^aThe chromosomal bin refers to a location assigned by Lewis et al (2004). Three genes investigated have locations within these significant bins. ^b A weighted and unweighted analysis was performed to calculate p-values. Yes signifies both methods reached p < 0.05

1.3.6 Association Study

Allelic association refers to the co-occurrence of an allele at a particular locus and a disease, above the level be expected by chance (Sham and McGuffin 2002: Edwards 1965). Association between two variables suggests, but does not necessarily define causation. Once we start to look for genes of smaller effect than $\lambda_s = 1.5$, the number of family numbers in linkage studies becomes prohibitively large. Therefore association studies provide a more powerful alternative to identifying such genes in realistically sized samples (Owen 2000). At present, researchers select specific genes or loci to be tested: functional candidates, positional candidates or a combination of both. Functional candidates are involved in a biological process that is postulated relevant to the disorder. Positional candidates are chosen because they are located within a locus that has been implicated from linkage or cytogenetic studies (Jurewicz et al 2001).

However, there are some potential problems with association studies. There are a large number of candidate genes within the human genome, $\sim 30,000$. Many genes have the potential to be involved in the pathogenesis of SZ because of the limited information to the aetiology. Therefore the probability of a selected candidate gene being involved in the aetiology is very low, resulting in an extremely low prior probability (e.g. the candidate gene has a $1/30,000$ chance of being associated with SZ). Secondly, due to genetic population sub-structure, many researchers believe that association studies have higher potential to generate false positive results due to population stratification (The concept of stratification is discussed later in section 1.3.6.1). Family-based association studies help to address this problem by using non-transmitted alleles as internal controls (Jurewicz et al 2001). However, there is the issue that you require heterozygotes for TDT analysis and this reduction in sample reduces power. In addition, obtaining large numbers of these families

in SZ are difficult because of the disrupted effects of mental illness on family relations (Owen 2000).

Failures to replicate initial findings of a study are often interpreted as false-positives with the original work. However, it may be that the replication study has false-negatives. Most association studies to date have had insufficient power to detect small genetic effects. It is therefore difficult to conclude which study showed a true negative or positive finding if indeed any (Jurewicz et al 2001). Therefore positive findings should be interpreted cautiously until replicated several times and negative results using small sample sizes should contain appropriate power calculations (Owen 2000). One paper by Lohmeuller et al (2003) attempted to address the question of whether false positive studies were responsible for the inconsistencies. They analysed 25 different reported associations for different disorders in 301 published studies. They found that a large proportion of replications were true positives and that many of the negative reports were due to underpowered studies. Thus, future association studies should be of sufficient size and high power before accepting a negative study as a true finding.

Another caveat of association studies is multiple testing (Owen 2000). A correction for the number of markers investigated must be carried out in order to keep $\alpha = 0.05$. However, as some SNPs will be in complete LD with each other some correction tests such as Bonferroni may be too conservative.

1.3.6.1 Population Stratification

Population stratification means that the sample population consists of several genetic subgroups, each containing genetic variation in allele frequency due to ancestry rather than

association of genes with disease. Therefore the frequency of certain alleles may be higher in some subgroups than others. If this is not accounted for in an association study, then a 'spurious' association may occur, generating a serious false positive (Owen et al 2002; Freedman et al 2004). The inverse is also possible generating a 'Simpson paradox' (Sham and McGuffin 2002). It is unknown how many published association studies can be attributed to population stratification and it has been argued that the effects of stratification can be eliminated simply by carefully matching cases and controls according to geographical location and self-reported ancestry (Freedman et al 2004). There are currently two methods available for testing population stratification. Firstly there is the case-control method of Pritchard and Rosenberg (1999) and secondly testing for genomic control (Devlin & Roeder 1999, Reich & Goldstein 2001).

Genomic control uses a set of anonymous markers to test for differences between allele frequencies in cases and controls that are due to systematic differences in ancestry rather than association of genes with disease (Freedman et al 2004). It examines the distribution of association statistics (chi-squares) between unlinked genetic variants typed in cases and controls (Marchini et al 2004). The statistic at the candidate allele being tested for association can then be compared with the genome-wide distribution of statistics for markers that are probably unrelated to disease to assess whether the candidate allele stands out. In the absence of stratification, association between unlinked genetic variants and disease should follow a chi-squared distribution with 1 degree of freedom. In the presence of stratification, the distribution of association statistics should be inflated by a value termed λ , which becomes larger with increasing sample size.

1.3.7 Linkage Disequilibrium

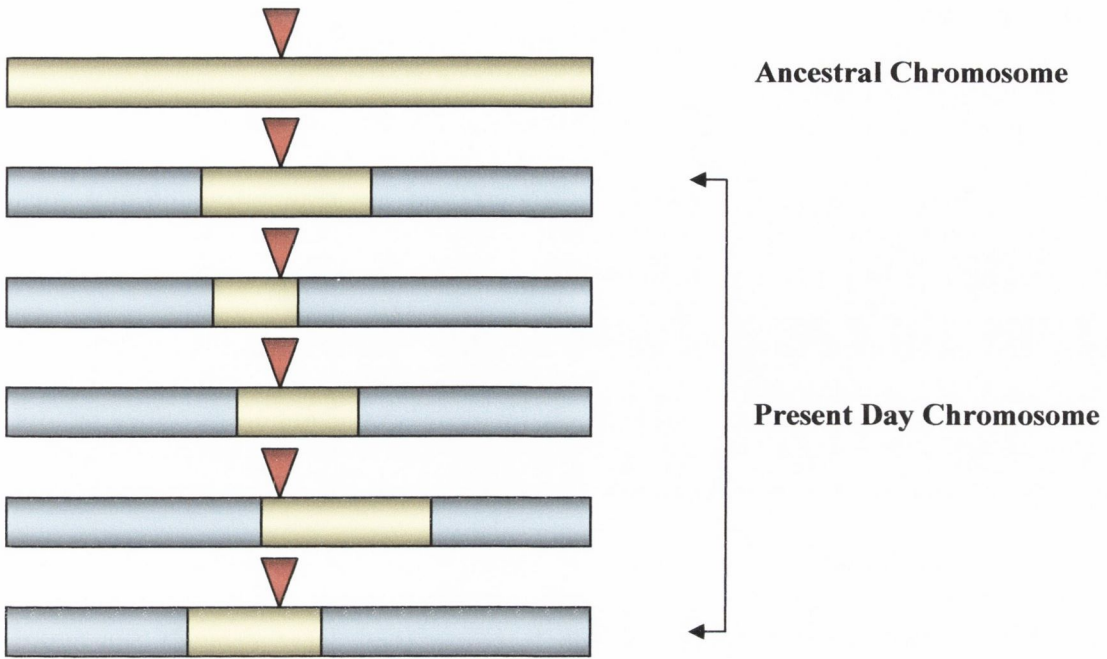


Figure 1.3: LD around an ancestral mutation. The mutation is indicated by a red triangle. Chromosomal stretches derived from the common ancestor of all mutant chromosomes are shown in yellow, and new stretches are shown in blue. Markers that are physically close (that is, in the yellow regions of present day chromosomes) tend to remain associated with the ancestral mutation as recombination limits the extent of the region of association over time.

1.3.7.1 Defining LD

The formal definition of LD is the non-random association of alleles at adjacent loci (Ardlie et al 2002). If a set of markers co-occurs on the same haplotype more often than is expected by chance, the alleles are said to be in LD (Wall and Pritchard 2003). That is when a particular allele at one locus is found together on the same chromosome with a specific allele at a second locus – more often than expected if the loci were segregating independently in a population – the loci are in disequilibrium (Ardlie et al 2002). So how does LD arise and how can it help geneticists in positional cloning?

1.3.7.2 Explaining LD

LD refers to the correlations among neighbouring alleles reflecting haplotypes (a combination of alleles at neighbouring loci) descended from single ancestral chromosomes (Reich et al 2001). When a new mutation arises, it does so, on a specific chromosomal haplotype. The association between each mutant allele and its ancestral haplotype is disrupted mainly by mutation and recombination in successive generations (other factors affecting LD are discussed later in section 1.4). Thus, it should be possible to track each variant allele in the population by identifying (through the use of anonymous genetic markers) the particular ancestral segment on which it arose (Gabriel et al 2002) (see figure 1.3). For example, if most individuals affected in a population share the same mutant allele at a causative locus, it is possible to detect the narrow genetic interval around the disease locus by detecting disequilibrium between nearby markers and the disease locus (Ardlie et al 2002). This approach makes use of the many crossover recombination events since the first appearance of the mutation.

Each pair of chromosomes undergo on average two recombination events per meiosis. Therefore genes or loci may be ‘reshuffled’ during meiosis which tends to reduce the level of LD between all pairs of loci from one generation to the next (Sham and McGuffin 2002). Consequently, markers that are closely physically located tend to remain together over many generations – analogous to playing cards sticking together until disrupted by repeated shuffling (Corvin and Gill 2003). Therefore these alleles will have a slow decay of LD over generations (Sham and McGuffin 2002). However, other factors that influence the extent of LD include such events as admixture, genetic drift, multiple mutations, natural selection, the size and age of a population, together with the mutational and recombinational histories of the loci make establishing LD difficult (Jorde 2000).

Variation in the Human Genome sequence plays a powerful but poorly understood role in the aetiology of common medical conditions (Gabriel et al 2002). Identifying the genetic variants that predispose to common human diseases such as psychiatric disorders is the major goal of the human geneticist. The traditional and primary strategy is to test for linkage in families with 300 – 400 microsatellite markers. However, this approach relies heavily on crossover recombination events and only identifies chromosomal regions of 10 – 20 Mb in length. Fine scale mapping of these large regions is then required in order to identify candidate genes. This was prohibitively costly with the technology available prior to the publication of the Draft Human Genome (Boehnke 2000). However, technology, as well as our knowledge of LD patterns in the human genome, has now advanced to a level where mass automation and more efficient genotyping has reduced costs substantially.

1.4 Recombination and Mutation

The larger the number of recombination events between two loci, the less likely it is that LD will be preserved between them. The number of recombination events depends on two factors: the recombination fraction between the loci and the number of doubly heterozygous crossing over events that have occurred between the current population and the initial mutant chromosome. If we assume random mating, the number of meioses is strongly related to the number of generations separating the founder chromosomes and the sampled population. Greater LD is therefore generally expected between markers that are close (low recombination fraction) and between markers in populations that began with a small number of founders relatively recently (few meiosis). The practical implication for LD mapping is that when used in studies, recently founded populations may be very useful

because they may present LD over large regions. However, mapping will be of poor resolution. In contrast, older populations may only exhibit LD over short regions but will provide more accurate positional data for fine mapping of disease loci (Jorde 2000). Figure 1.4 represents an example of how mutation and recombination affect LD.

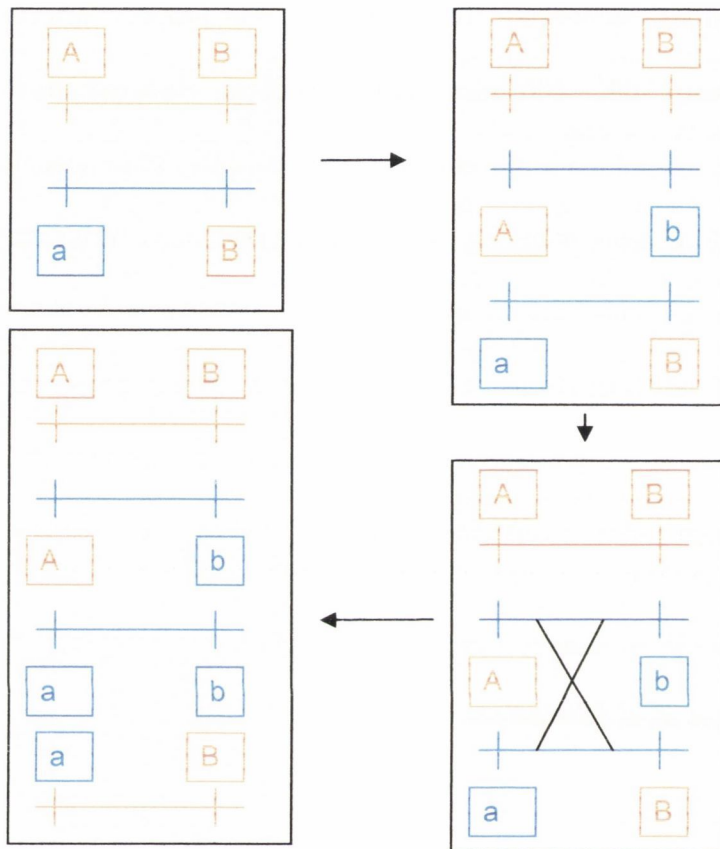


Figure 1.4: Decay of LD by recombination. At the start, there is a polymorphic locus with alleles A and a. When a mutation occurs at a nearby locus, changing allele B to b, this occurs on a single chromosome bearing either allele A or a at the first locus (A in this case). Therefore, early in the life of the mutation, only three out of four possible haplotypes will be observed in the population. The b allele will always be found on a chromosome with the A allele at the adjacent locus. The association between alleles at the two loci will gradually be disrupted by recombination between the loci. This will result in the creation of the fourth possible haplotype and eventual decline in LD among the markers in the population as the recombinant chromosome (a, b) increases in frequency.

The extent of LD is inversely proportional to the local recombination rate. From recent empirical studies it has been proposed that the genome is arranged into discrete haplotypic blocks of low recombination frequency separated by recombination ‘hot spots’ (Daly et al 2001, Gabriel et al 2002, Dawson et al 2002). Underlying haplotypic structure in a region is useful in understanding the LD relationship between genetic markers and in the

identification of high-risk haplotypes within that region. Due to the limited variation in haplotype blocks, genotyping of a single SNP on the haplotype may be almost as powerful in detection of LD between haplotypes and disease carrying haplotypes (Johnson et al. 2001). This method of haplotype-tagging is more cost effective due to a reduction in the amount of genotyping required. Haplotypes are likely to vary in their size in differing populations. It has been shown that haplotypes in older populations are smaller due to recombination events over a greater number of years. In contrast, younger populations such as European and Asian, have larger haplotypes (Gabriel et al. 2002) because there is less time for recombination events to have occurred. The extent of LD in the Irish population is unknown and there is no published data concerning underlying haplotypic structure. However, Kendler et al. (1999) have shown that for the regions of the genome they studied with dense marker maps they were able to identify LD at the 5% level of significance in 96% of marker pairs within 0.5cM distance from each other, and 67% of pairs 0.5-1cM apart, 35% of pairs 1-2cM apart, 15% of pairs 2-4cM apart and 8% of pairs 5-10cM apart.

Inversion polymorphisms can also influence recombination rates. These occur when segments of chromosomes become inverted. It has been suggested that heterozygous individuals for the inversion have little or no recombination between the normal and the inverted chromosomal regions (Martin et al 1999). Therefore it is possible that these regions will display higher LD. For example if a disease mutation within the inverted region causes disease, and the region has no recombination, extensive LD structure will build up between the disease mutation and other markers in this inverted section. Inversion polymorphisms have been identified as large as 3MB (Pritchard and Przeworski 2001) which could lead to strong LD over large distances.

Gene conversion also results in chromosomal exchange of genetic material between homologous chromosomes (Andolfatto and Nordberg 1998). It is the nonreciprocal exchange of short tracts of genetic material between homologous chromosomes. The average length of gene conversion is thought to be between 350 – 1000 base pairs (Ardlie et al 2002). If one of the markers that contribute to LD is contained within the short tract, this will result in disruption of LD over shorter distances. Therefore, by definition, gene conversion, should not impact on LD over long distances. Ardlie et al (2002) noted that a significant proportion of tightly linked marker pairs were not in complete LD and suggested that gene conversion could be responsible for this phenomenon. Therefore gene conversion could be an important factor in fine LD mapping.

1.4.1 Variable mutation rates.

Mutation at a locus will also affect LD. If two alleles are associated at two loci, and a mutation event occurs at one locus, then it is possible the two alleles may no longer be associated. Some SNPs, particularly those in CpG islands have high mutation rates and are therefore not in LD with surrounding markers.

1.5 LD and power

Another reason for LD studies is that linkage studies have less power to detect genes or variants with small effect size. As the vast majority of heterozygosity is attributable to common variants and because the evolutionary history of common human diseases (which determined the allele spectrum for causal alleles) is not yet known, one promising approach is to comprehensively test common genetic variation for association to medical conditions (Gabriel et al 2002). With the development of a dense map of biallelic SNPs to

facilitate this process using either case control studies or simpler families using TDT (Boehnke 2000), LD offers a shortcut to genome wide association studies (Ardlie et al 2002). LD can be used to detect association with a disease susceptibility locus indirectly using a nearby marker. No prior hypothesis of putative variants is required (HapMap 2003). If disequilibrium were extensive the number of markers for a genome wide association study could be substantially reduced, increasing efficiency and decreasing genotyping costs (Ardlie et al 2002).

1.5.1 Empirical Studies

In order to perform genome LD mapping it is essential that we understand the behaviour of LD in the human genome (Ardlie et al 2002). Initially this was carried out using computer simulations. More recently, thanks to the vast number of SNPs identified, empirical studies have shown LD to be complex and variable throughout the genome (Reich et al 2001; Gabriel et al 2002; (Daly et al 2001; , Dawson et al 2002).

Prior to empirical studies, it was theoretically thought that the strength of LD would be strong for tightly linked loci and weak with increasing genetic distance (Sham and McGuffin 2002). However, recent studies have shown that there is great variability in LD structure throughout the genome (Abecasis et al 2001). One study by Daly et al 2001, found that haplotype blocks span up to 100 kb containing at least 5 common SNPs (with a frequency > 5%). They found a maximum of 4 haplotypes per block that account for > 90% of haplotypes in cases in their sample. Their results and other (Gabriel et al 2002, Dawson et al 2002) suggest that there are discrete haplotype blocks (tens to hundreds of kilobases), punctuated by sites of apparent recombination, throughout the genome. However, other studies have shown that this simplistic picture may not be the case. Jeffries

and May (2005) noted disruption of LD at short intervals in human sperm. They suggest that recombination events can occur at short distances to each other but are halted through some 'DNA checking method', resulting instead in gene conversion. This phenomenon is when a short stretch of one copy of a chromosome is transferred to another during meiosis without reciprocation. This in turn breaks down LD at short distances. Again another study showed LD structure in certain genes with a single haplotype frequency over 50% whereas one third of the genes examined did not have a single predominant haplotype (Stephens et al 2001) indicating that there was no or little LD over a short range. Therefore at this time further studies into genome wide LD patterns are required (Ardlie et al 2002).

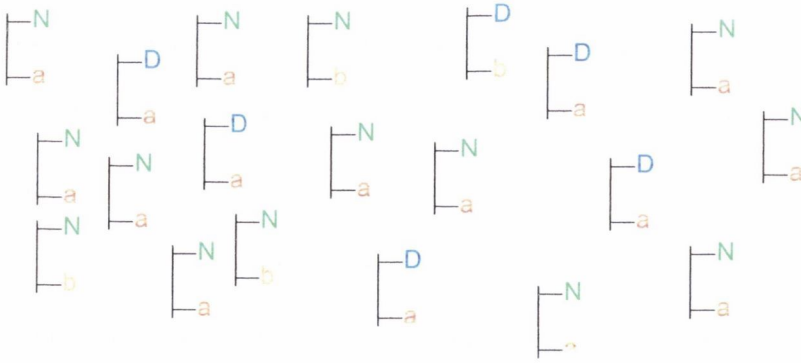
1.5.2 Additional factors affecting LD

1.5.2.1 Genetic drift

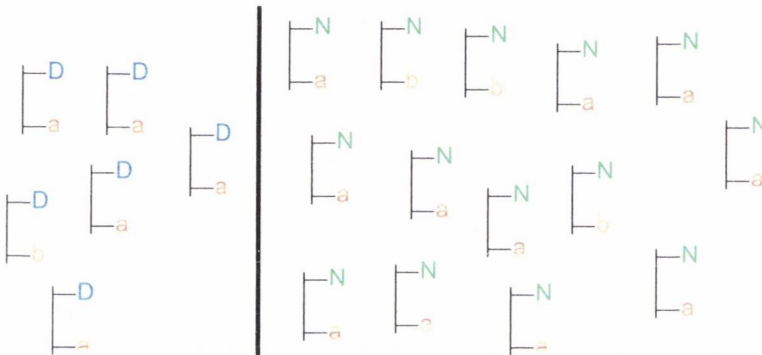
Changes in gene or haplotype frequencies between generations due to random sampling of gametes are called genetic drift. In small populations, this effect is more pronounced as increasing genetic drift will increase the LD as some haplotypes are lost. Ireland's position on the western edge of Europe suggests that the genetics of its population should have been relatively undisturbed by the demographic movements that have shaped variation on the mainland of the UK and Europe. The influence of genetic drift has been demonstrated in the Irish population based on geographical location, although it is not pronounced (Hill et al 2000).

Genetic drift has the potential to create LD between closely linked loci. For example, consider two loci, one for disease and one marker locus. The disease locus has a disease allele and a normal allele. The marker locus is biallelic. As they are physically very close

to each other the recombination rate is very close, if not equal to zero. The allele frequencies at the marker locus will fluctuate from one generation to the next because of the chance assortment of chromosomes at meiosis. However, as there is no recombination between the disease and marker locus, the biallelic marker frequencies will fluctuate independently. The degree of fluctuation depends on the size of the population: the smaller the population, the greater the variation in frequency; the larger the population the more likely that allele frequencies will tend to equal out due to the random nature of chance assortment. However, we would expect the allele frequency of the disease allele to be lower in the population than the normal allele. This means that the subpopulation of chromosomes carrying the disease allele will be smaller than the subpopulation of chromosomes carrying the normal allele. Therefore the fluctuation of allele frequencies at the marker locus on the disease allele chromosomes would be greater (and independent of) the fluctuation of allele frequencies at the marker locus on normal allele chromosomes. This is illustrated in the following figure, Figure 1.5.



Population of chromosomes. Two linked loci; disease locus with allele D (pathogenic) and allele N (non-pathogenic) and biallelic markers a and b.



Separation of chromosomes into subpopulations of disease and normal chromosomes.

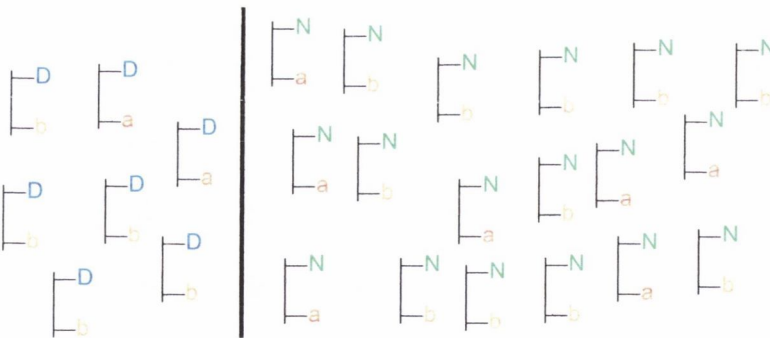
D chromosomes; $f(a) = 0.667$

$f(b) = 0.333$

N chromosomes; $f(a) = 0.667$

$f(b) = 0.333$

Disease and marker loci are not in LD



Chance assortment of chromosomes at meiosis results in fluctuation of allele frequencies at marker locus. Fluctuations within each subpopulation are independent. Greater fluctuation in smaller population of chromosomes carrying the disease allele.

D chromosomes; $f(a) = 0.333$

$f(b) = 0.667$

N chromosomes; $f(a) = 0.611$

$f(b) = 0.389$

Disease and marker loci are in LD

Figure 1.5: Genetic Drift. The above schematic represents a small isolated population, showing the effects genetic drift has on the creation of LD. (Reproduced with permission from Dr. Derek Morris).

The greater fluctuation of allele frequencies at the marker locus in the subpopulation of chromosomes carrying the disease allele creates LD between the marker locus and the disease locus. However, because of the random nature of genetic drift, it is possible that the allele frequencies at the marker locus will be equal in both the disease and normal subpopulations of chromosomes and LD will not be detected between the two loci.

Genetic drift is most likely to create LD in a small population that has not expanded in size over time. As the overall population expands, so will the subpopulation of disease chromosomes. As the disease subpopulation gets bigger, the fluctuation in allele frequencies at the marker locus from one generation to the next will lessen. As it is these fluctuations that are responsible for generating LD the expanding disease subpopulation will consequently not be enriched with LD to the same extent as the one of constant size (Terwilliger et al 1998). The idea of drift mapping has been proposed (Terwilliger et al 1998). It is proposed that LD mapping in small populations of constant size will maximise the use of LD that has been created by genetic drift.

1.5.2.2 Admixture

Admixture is the interbreeding of two populations. If these two populations have different allele frequencies at each of two loci, admixture will produce LD between the loci in the resultant admixed population (Chakraborty and Weiss, 1998). Such a demographic event could result from migration. The amount of LD generated between pairs of loci depends on the difference between the allele frequencies in the two founding populations and the degree of admixture (Stephens et al 1994).

Figure 1.6, illustrates an example of how LD is generated by admixture. If we have two populations, X and Y and two polymorphic loci A (A1 and A2) and B (B1 and B2), we can show that by interbreeding both populations to form a new population Z we produce LD.

In population X, the allele frequencies of A1, $f(A1) = 1.0$ and B1 $f(B1) = 1.0$. In population Y, the allele frequencies of A2, $f(A2) = 1.0$ and B2, $f(B2) = 1.0$. If X and Y populations interbreed to form population Z and both founder populations contribute equally to Z, the allele frequencies at the two loci for the first generation of Z will be as follows: $f(A1) = 0.5$, $f(A2) = 0.5$, $f(B1) = 0.5$ and $f(B2) = 0.5$. Given these allele frequencies, one would expect the haplotype frequencies under linkage equilibrium to be as stated in figure (b). However, because the (A1 B2) and (A2 B1) haplotypes did not exist in either of the founder populations, the real haplotype frequencies are very different: $f(A1 B1) = 0.5$ and $f(A2 B2) = 0.5$. Therefore these haplotypes show that A and B are in complete LD in the Z population. This LD will of course decay over several generations due to recombination.

(a) Under linkage equilibrium one would expect:

P1 A1 B1 x A2 B2

Frequency 0.5 0.5 0.5 0.5

Gametes A1 B1 A2 B2

F1 A1 B1 A1 B2 A2 B1 A2 B2

Frequency 0.25 0.25 0.25 0.25

(b) However, because A1 B2 and A2 B1 did not exist in either population X or Y it would not exist in population Z. Therefore the haplotype frequencies would be:

F1 A1 B1 A2 B2

Frequency 0.5 0.5

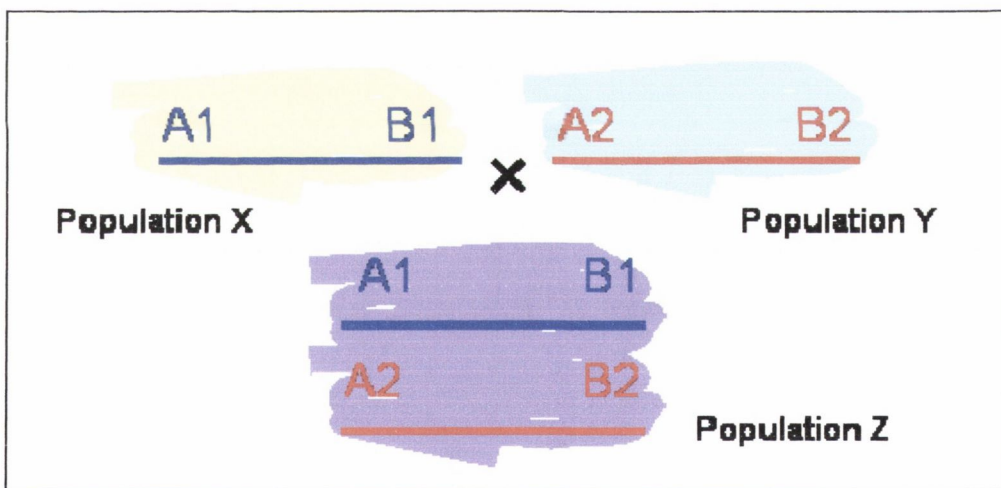


Figure 1.6 Admixture. (a) Allele frequencies one would expect under linkage equilibrium. (b) Haplotype frequencies of population Z showing complete LD between alleles (A1 B1) and (A2 B2) each with a haplotype frequency of 50%.

1.5.2.3 Population Bottlenecks, Inbreeding and Assortative Mating

Other population factors that affect LD are population bottlenecks (see Figure 1.7), inbreeding and assortative mating. Bottlenecks result from dramatic temporary reductions in population size. If the population is significantly reduced, the number of haplotypes may also be reduced because not all haplotypes are represented in proportion to that expected by the frequencies of each allele. If the bottleneck is recent (in terms of generations), there may be extensive LD in the present population because of the reduced opportunity for recombination to decimate haplotypes. There may also be reduced locus and allelic heterogeneity because of the reduced number of disease haplotypes. Inbreeding and assortative mating can lead to increased levels of LD by reducing the number of doubly heterozygous recombination events to disrupt haplotypes.

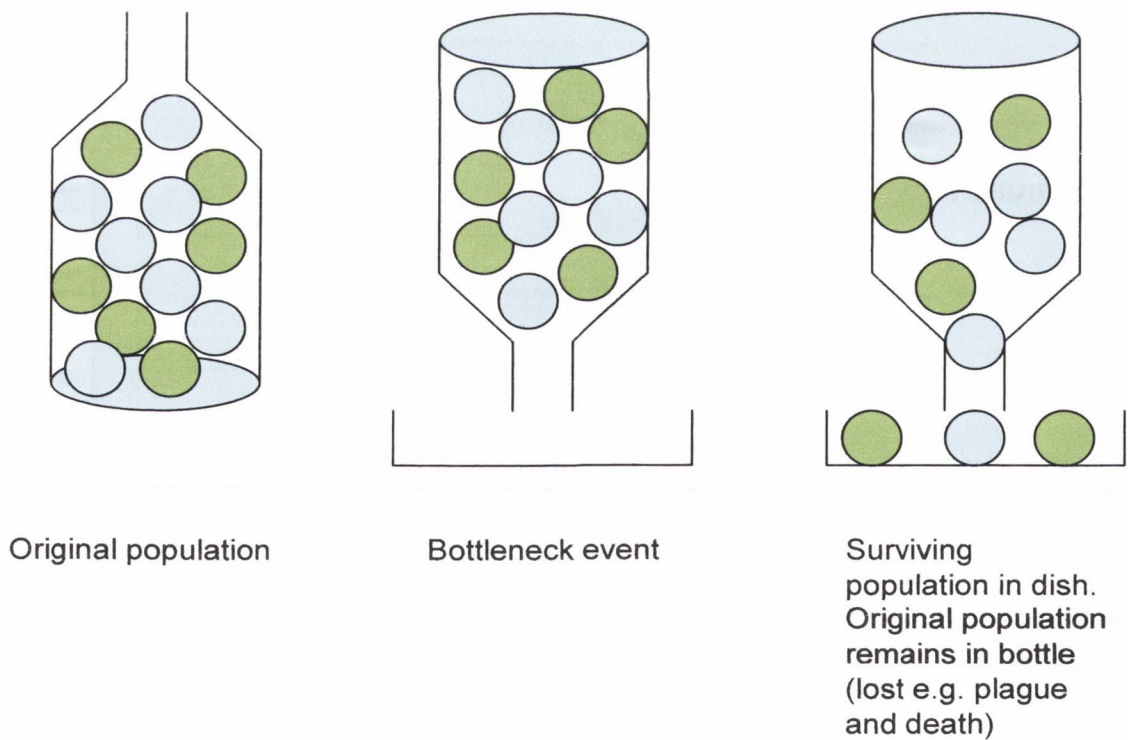


Figure 1.7 Population Bottleneck. This figure represents how a ‘bottleneck effect’ can reduce the size and amount of variation seen in a population which may lead to fewer haplotypes and increased LD between markers.

1.6 Measures of linkage disequilibrium

Lewontin's D (1964) quantified linkage disequilibrium as the difference between the observed frequency of a two locus haplotype and the frequency it would be expected to show if the alleles were segregating at random.

The observed frequency of the haplotype that consists of alleles A and B is represented by P_{AB} . Assuming the independent assortment of alleles at the two loci, the expected haplotype frequency is calculated as the product of the allele frequency of each of the two alleles, or $P_a \times P_b$, where P_a is the frequency of allele A at the first locus and P_b is the frequency of allele B at the second locus. So one of the simplest measures of LD is

$$D = P_{ab} - P_a \times P_b$$

The two most common measures are the absolute value of D' and r^2 . The absolute value of D' is determined by dividing D by its maximum possible value, given the allele frequencies at the two loci. This means that if $D'=1$, if and only if, two SNPs have not been separated by recombination (or recurrent mutation or gene conversion) during the history of the sample. In this case, at most three out of the four possible two-locus haplotypes are observed in the sample. This is known as complete LD. Values of D' less than 1 indicate complete ancestral LD has been disrupted. However, values less than 1 have no clear interpretation.

Although D' is less dependant on allele frequencies (Jorde 2000) values of D' are strongly inflated in small sample sizes, even for SNPs with common alleles, but especially for SNPs with rare alleles. So high D' values can be obtained even though markers may be in

linkage equilibrium. As values of D' are dependant on sample size, it is difficult to make sample comparisons. Values close to a $D' = 1$ suggest minimal historical recombination, but intermediate values should not be used for comparisons of the strength of LD between studies, or to measure the extent of LD.

The measure r^2 is in some ways complimentary to D' . It has recently emerged as the measure of choice for quantifying and comparing LD in the context of mapping (Pritchard and Przeworski 2001) (Weiss and Clark 2002). It is the correlation of alleles at the two sites, and is formed by dividing D^2 by the product of the four alleles frequencies at the two loci.

$$r^2 = D^2 / (P_{AB} + P_{Ab} + P_{aB} + P_{ab})$$

$r^2 = 1$ if, and only if, the markers have not been separated by recombination and have the same allele frequency. In this case, exactly two out of the four possible two-locus haplotypes are observed in the sample. The case of $r^2 = 1$ is known as perfect LD. In this case the observations at one marker provide complete information about the other marker, making the two redundant. Intermediate values of r^2 are easily interpretable.

The value of r^2 is related to the amount of information provided by one locus about the other. To have the same power to detect the association between the disease and the marker locus, the sample size must be increased by roughly $1/r^2$ when compared with the sample size for detecting association with the susceptibility locus itself (Kruglyak 1999, Pritchard and Przeworski 2001). Lastly, r^2 shows much less inflation in small sample sizes than D' .

The power of LD is affected by the methods of measurement used, the number of available samples, inheritance mode, recombination patterns and mutations in a region, the age of the mutations, the degree of allelic and locus heterogeneity, the type of markers assayed and many aspects of population history (Jorde 2000).

1.7 Hypotheses

The studies contained within this thesis investigated eight genes putatively involved in SZ pathogenesis that had previously been identified as potential candidate genes using linkage data and/or gene expression studies. Specifically, genes were selected to test two current hypotheses on the pathogenesis of SZ: that of abnormal neurodevelopment and more recently oxidative stress. The Apolipoprotein-L (APOL) genes (4, 5 and 6) were selected as functional candidate genes based on their known role in neurodevelopment and evidence for their abnormal expression in post-mortem SZ samples; and positional candidates as the gene family localize to a known SZ locus. This combination of positional cloning methods and functional studies has been termed convergent functional genomics (Chapter 3). In addition, a positive replication study of G72/G30 and DAAO putatively involved in neurodevelopment was also investigated (Chapter 4).

To test the hypothesis of oxidative stress and SZ a further three genes involved in cellular oxidative stress response were also investigated: HSPA8, GSTM3 and NRF2 (Chapters 5 and 6). Two of these genes (HSPA8 and GSTM3) were selected based on positional evidence from a linkage meta-analysis scan (Lewis et al 2004) and functional evidence from a differential expression study (Prabakaran et al 2004). In addition, NRF2 was selected as a functional candidate because it is a known transcription factor for GSTM3.

A second goal of the study was to determine how feasible it was at the time of this project to use data from the HapMap project on haplotype structure in populations of European origin as a proxy for the haplotype structure in the Irish population. I compared LD structure and tagging SNPs (tSNPs) between the CEPH sample and the Irish sample using Phase I data available on NRF2. By comparing D' and r^2 values from both samples I could investigate the similarity in LD between the CEPH and Irish populations and identify the amount of common variation in the Irish sample which was captured by tSNPs from the European HapMap sample. Results of this are presented in Chapter 6.

2 Materials and Methods

2.1 Description of sample

The Irish schizophrenia association sample consisted of 219 cases and 231 controls from the Republic of Ireland. Ethics Committee approval for the study was obtained from all participating hospitals and centres. Most patients were approached on wards and at day clinics, with their respective consultant's permission. The hospitals involved in the study included St James Hospital, St. Patricks Hospital in Dublin, St. Davnets in Monaghan and St. Vincent's in Fairview. All cases gave written informed consent and were interviewed by a psychiatrist or psychiatric nurse trained to use the Structured Clinical Interview for DSM (SCID). Diagnosis was based on DSM-III-R criteria using all available information (interview, family or staff report and chart review). All cases were over 18 years of age, of Irish origin and had been screened to exclude substance-induced psychotic disorder or psychosis due to a general medical condition. Cases included in this investigation met criteria for SZ or schizoaffective disorder. The control sample, drawn from Irish blood donors, was not specifically screened for psychiatric illness; however, donors were not taking regular prescribed medication as such individuals are excluded from blood donation in Ireland. The blood samples were collected in 2 x 6ml tubes containing EDTA and one non-EDTA tube for the serum sample. The serum sample was obtained by centrifuging the blood to separate the serum from plasma. The serum was then retained for any future protein or chemical analysis. The two blood samples were frozen at -80°C until DNA extraction.

2.2 GASP study

Within the Neuropsychiatric Genetics group there were several projects looking at other disorders such as bipolar disorder under the direction of Professor Michael Gill. It was essential to identify this sample set from the other projects. Therefore the sample set was called the 'Genetic Association Study of Psychosis' (GASP). Each patient was given a unique lab number preceded by the acronym GASP.

2.3 DNA extraction from blood

DNA extraction from whole blood was performed by a modification of the standard phenol/chloroform method (Sambrook et al 1982). The procedure was a two day process. On average, 16 samples were run in one batch. The samples were thawed out on ice for two hours prior to extraction. The 6mls of blood was transferred to a 50ml falcon tube. A few millilitres of sterile water were added to the 50ml falcon tube in order to transfer any remaining residual blood in the EDTA tube. Sterile water was then added to the blood in the falcon tube and made up to 12.5mls. A further 12.5mls of 1x Lysis buffer was added giving a total volume of 25mls. Lysis buffer is a salt solution that is slightly osmolar to the contents of erythrocytes (red blood cells). It is used to lyse erythrocytes from the blood sample while maintaining the integrity of the lymphocytes (white blood cells) that contain the DNA. Erythrocytes absorb water from the solution through osmosis which then swells and ruptures the erythrocytes. The falcon tube was gently inverted twice and then placed on a rocker for 30 minutes. The mixture was then centrifuged at 3500 rpm for 15 minutes to form a pellet. The supernatant was carefully discarded until ~4mls remained. The falcon was then refilled to 25mls with 1x lysis buffer and tapped once at the bottom of the tube to release the pellet. The tube was then placed on ice and gently mixed on a rocker for 10

minutes to allow the pellet to dissolve into solution. Again the tube was centrifuged at 3500 rpm for 15 minutes. The supernatant was poured off until the lymphocyte (white blood cell) layer began to move. The secured pellet was resuspended in 1ml of 1x Suspension buffer. Suspension buffer contains a high salt concentration, Tris and EDTA. Tris interacts with the lipopolysaccharides present on the outer membrane of lymphocytes which helps to permeabilize the membrane. This effect is enhanced with the addition of EDTA. Using a sterile pastette the red pellet was transferred to a sterile 15ml falcon tube. Residue in the 50ml falcon tube was rinsed with a further 1ml of 1x suspension buffer and then added to the 15ml falcon. Finally, 150 μ l of 10% SDS and 60 μ l of Proteinase K was added to the 15ml falcon tube and placed on a rocker overnight at 37°C. SDS is a detergent with a negative charge that has a high affinity to proteins, promoting proteolysis. Proteinase K is a powerful proteolytic enzyme that ensures the degradation of nucleic proteins. This overnight step promotes the denaturation and digestion of the cell and nuclear membranes, enzymes and other proteins that remain in solution after the two lysis buffer steps.

On day two, 2mls of phenol was added to the 15ml falcon tube. Phenol is an organic neurotoxin. Therefore this part of the procedure was carried out in a fume hood. The solution was now cloudy and was shaken and then rocked for ten minutes. The tube was centrifuged at 3800 rpm for 10 minutes. The supernatant was removed using sterile pastettes and placed in a new sterile 15ml falcon tube. 1ml of chloroform (24:1 chloroform:iso-amyl-alcohol) and 1ml of phenol were added to the solution and again the tube was placed on a rocker for ten minutes and then centrifuged at 3800 rpm for 10 minutes. Phenol and chloroform are organic solvents, which capture the lysed cell debris that is mostly protein. The DNA, which is insoluble in the organic phase, is present in the non-miscible aqueous phase (supernatant). Again the supernatant was removed and

transferred to a sterile 15ml falcon tube. 50µl of 3M sodium acetate was added followed by ~6mls of absolute alcohol at -20°C. The ethanol depletes the hydration shell from DNA, exposing negatively charged phosphates. The sodium ions in Sodium Acetate are monovalent cations, which bind to the phosphate groups reducing the repulsive forces between the chains, rapidly precipitating DNA from the aqueous solution. The tube was gently inverted until a pellet was formed in the tube. The pellet was then removed using a sterile 1ml pipette tip and transferred to a sterile eppendorf tube. 1ml of 70% alcohol at -20°C was added to the eppendorf and spun at 13000 rpm for five minutes in a microfuge. This removes residual salt and moisture allowing the DNA to be pelleted. The alcohol was removed using a sterile 1ml and 200µl pipette. This procedure was repeated again. The extracted DNA was allowed to dry on the bench for 3 hours. Once dry, the DNA pellet was resuspended in 250µl of TE buffer. The TE Buffer reduces base hydrolysis by chelating divalent cations from the DNA solution with EDTA and by resisting pH changes with the Tris buffer. Both alkalinity and free cations have been implicated in catalysing the hydrolysis of DNA. The eppendorf was then stored and gently spun for 5 days at 4°C until the DNA had fully dissolved. The final product was then stored at -20°C until quantification.

Some of the extracted samples had brown specs in the pellet that indicated protein contamination. Therefore these samples had to undergo an additional cleanup step before they could be accepted. The following was added to these samples before incubating at 37°C for 3-4 hours: 60µl of proteinase K; 150 µl of SDS and 490µl of ddH₂O. As previously mentioned, SDS has a high affinity for proteins and Proteinase K is a powerful proteolytic enzyme. After the incubation period the contents of the eppendorf were transferred to a 15ml falcon tube and 1ml of H₂O added. These samples were then taken through day two extraction procedures.

Prior to the GASP study, DNA samples from 52 patients had been extracted by an MSc student, Steve Smith, under the supervision of Dr. Ziarh Hawi. I obtained the spectrometry readings for these samples and discovered that the calculations had been out by a factor of 10. Therefore, once I accounted for the discrepancy I could confidently calculate the correct concentrations of DNA extracted. These additional samples were then added to our GASP collection.

2.4 DNA quantification

DNA quantification was carried out initially using spectrometry to obtain an indication on the quality and amount of DNA extracted from frozen blood samples. Secondly DNA concentrations were measured using fluorimetry to obtain a more accurate measurement of concentration which is essential to prevent wastage of DNA and important when designing DNA pools. We used this strategy initially to roughly quantify the extracted DNA so that 80ng/μl solutions could be made from raw stocks. We then used fluorimetry to accurately quantify the 80 ng/μl stocks which then allowed us to produce 8ng/μl stocks for use in individual and pooled analysis.

2.4.1 Spectrometry

DNA absorbs light at 260nm whereas RNA absorbs light at 280nm. Comparing the ratio of these two values gives an indication of the quality of the DNA. If the ratio was <1.8 it indicated that the extracted sample contained proteins and aromatic hydrocarbons. In these cases, the sample was re-cleaned with SDS and proteinase K before undergoing the phenol/chloroform steps again (see section2.3). If the ratio was >2 it indicated that the

sample contained a substantial quantity of RNA. No sample extracted contained excess RNA.

The DNA concentration was calculated using the following formula:

$$\text{DNA concentration} = \frac{\text{Optical Density (OD)}_{260\text{nm}} \times 50}{5} \mu\text{g}/\mu\text{l}$$

However, we wanted to know the DNA concentration in ng/ μl . Therefore, in the example of GASP4001, the $\text{OD}_{260\text{nm}}$ was 0.53 $\mu\text{g}/\mu\text{l}$ which is 530ng/ μl . As the DNA is dissolved in 250 μl of 1x TE, it equates to $250 \times 530 = 132500$ ng of DNA.

Samples were quantified as follows. Firstly, the spectrometer was blanked with 1xTE buffer contained in 1ml cuvettes. I then added 5 μl of each sample and made up to 1000 μl with 1x TE. This 1:200 dilution of the raw stock was necessary to facilitate the range of the spectrometer. These 1ml preparations of the stock solution were transferred to a fresh cuvette and readings at 260 and 280nm were recorded. By applying the equation above we could determine the raw stock concentrations, which would allow us to dilute the samples appropriately.

2.4.2 Fluorimetry

DNA quantification was carried out using the PicoGreen dsDNA quantification kit (Molecular probes, Eugene, Oregon, USA) on a Wallac Victor 2, 1420 Multilabel counter (Perkin Elmer) in the department of microbiology at TCD. Before quantification can occur, a standard curve must be obtained for each run of samples. This was achieved by using a set of standard DNA concentrations supplied with the kit (ng): 750; 500; 250; 100; 50 and 0 these standards are run concurrently with the samples. The DNA standards were used to

draw a standard curve which produced an equation, $y=mx$, which represented the relationship between DNA concentration and fluorimetric reading. This equation was then used to determine the DNA concentration of case samples from the fluorimetric readings.

To obtain these readings the following was carried out. The raw stock was initially quantified using the spectrometry method (listed above) and then diluted to 80 ng/ μ l stocks. An aliquot of 5 μ l was taken from the 80ng/ μ l stocks and diluted with 95 μ l of ddH₂O, to make a final volume of 100 μ l. The purpose of the dilution was for accurate quantitation. As the standard curve increases, the fluorescence eventually reaches a ceiling. This means that if the sample had a higher concentration than that of the standard curve, we could not detect it due to the ceiling of fluorescence. In addition, it also saved on sample being used for fluorimetry.

The 100 μ l was added to the fluorimetry plate using an electronic pipette. In addition, a 1:200 dilution (0.5 μ l of picogreen diluted using 99.5 μ l TE, making 100 μ l of 1x) of picogreen mixture was made for each sample being analysed. Dilutions were made up in a light protective tube, thereby retaining the fluorescence of picogreen. 100 μ l of picogreen (1x) was added to each sample on the fluorimetry plate using an electronic pipette and covering each column with tin foil (to protect picogreen from fluorescing) as I moved along the plate.

The fluorimetry plate was automatically mixed inside the fluorimeter by gentle mechanical shaking for two minutes. The samples were then read on the fluorimeter. Once fluorimetry results were obtained, 8ng/ μ l working stocks were designed. The design of DNA pools required fluorimetry and this is detailed in section 3.2.14.

2.5 DNA storage

DNA stock solutions, DNA pools and DNA working solutions were stored at -20°C. Approximately 600µl of DNA at 8ng/µl were stored in 96 deep well boxes for individual genotyping. When not in use, these boxes were stored at -20°C. However, when individual genotyping was performed these boxes were stored at 4°C to reduce the effects of freeze-thaw damage. When DNA pools were used they were constantly stored at -20°C.

2.6 Materials and Methods for the Apolipoprotein-L family

2.6.1 Identification of SNPs

The six genes (APOL1-6) span 617 KB. Information on exonic structure and transcript details are available at the AceView website (<http://www.ncbi.nlm.nih.gov/IEB/Research/Acembly/>). The SNPs genotyped in this study are detailed in the results section Chapter 3 (Table 3.1) and were all identified from the dbSNP database (<http://www.ncbi.nlm.nih.gov/SNP/>). From a total of 187 SNPs we selected 143 for investigation (see Appendix Figures A.1 – A.12) concentrating on SNPs close to the putative promotor region (PPR) and exonic regions of the six genes. We then screened these SNPs for heterozygosity in a pooled DNA sample of 100 individuals using SNaPshotTM (Applied Biosystems). Of 143 potential markers, 51 (36%) were found to be polymorphic. The average and median inter-marker distances for each gene is contained in Table 2.1.

Table 2.1 – SNP coverage density of all six APOL genes

Gene	Number of Markers	Average Inter-marker Distance (bp)	Median Inter-marker Distance (bp)
<i>APOL6</i>	7	5598.6	1845.5
<i>APOL5</i>	8	3911.6	4137
<i>APOL4</i>	10	1549.7	736
<i>APOL3</i>	5	10421.6	7651
<i>APOL2</i>	14	2335.5	1713
<i>APOL1</i>	7	2382.6	1495

2.6.2 Primer Design

Primers for all fragments were designed using Primer3 (http://www.genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi). Primer pairs were obtained from Invitrogen (Paisley, Scotland). All PCR reactions used 8ng of genomic DNA, 5pmol of each primer and 1µM dNTPs. The Buffer consisted of 1.5 mM MgCl₂, 50mM KCl and 10mM Tris HCl (pH 8.3). One unit of Taq polymerase (Qiagen) was used in reactions. All PCR reactions were carried out in 10µl volumes. All PCR reactions were carried out on DNA engine TetradTM thermal cyclers (MJ Research)(For a more comprehensive list of PCR and extension primers please see the Appendix Table A.1 – A.6).

2.6.3 PCR optimisation

A range of PCR programs were used (for optimisation see Appendix Table A.7 and A.8). In a standard PCR mix I used 1.5mM MgCl₂. However, some reactions required additional MgCl₂ to increase the specificity of the reaction. I used a concentration gradient of 1.5 (normal); 2 and 2.5mM. To achieve this I had a 100mM stock of MgCl₂ which I diluted to

10mM MgCl₂. Adding 0.75µl and 1.5µl of the 10mM stock to the 10µl PCR reaction gave concentrations of 2 and 2.5mM respectively.

2.6.4 PCR buffer protocol

I prepared the PCR buffer 'on the bench'. This is composed of 100mM Tris HCl pH 8.9; 15mM MgCl₂; 500 mM KCl; 1% Triton x100 (Sigma). In some preparations 1% gelatin is also added. I did not use gelatin in this buffer. To obtain the above concentration the following volumes were aliquoted from the stock buffers: 0.5ml 1M Tris HCl pH 8.9; 0.75ml 100mM MgCl₂; 1.25 ml 2M KCl; 50µl Triton x100. The solution was spun at 13000 rpm in a microfuge for one minute to remove bubbles. It was then autoclaved before use. All PCR buffer was stored at -20°C until required.

2.6.5 Agarose Gels

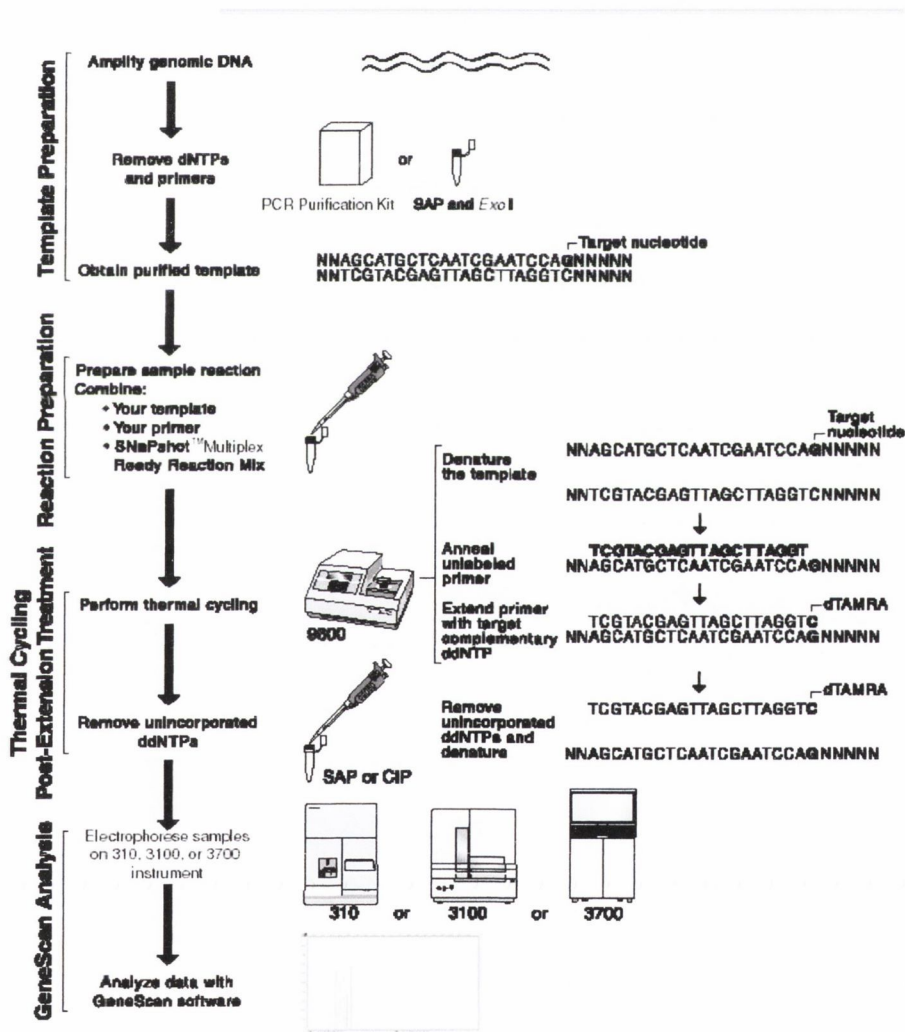
All PCR fragments were run out on 1.5% agarose gels. 1.5g of agarose was weighed out and transferred to a 250ml conical flask. 100mls of 1xTAE buffer was added and microwaved until the agarose had dissolved (roughly 1.5 mins). The flask was then allowed to cool on the bench. 5µl of Ethidium Bromide was added to the solution, mixed and then poured into an appropriate gel mould to set. Once the gel was set, it was placed in a gel rig with 250mls of 1xTAE buffer. PCR samples were loaded into the wells of agarose gels using 4µl of PCR product and 2µl of loading dye (Table 2.2). Samples were run at 80V for 30 minutes.

Table 2.2 Composition of Agarose Gel loading dye

Agarose Gel Loading Dye	
Bromophenol Blue	12.5mg
Xylene Cyanol	12.5mg
Glycerol	15ml
H ₂ O	up to 50ml

2.6.6 Primer extension using SNaPshot

Overview of the Procedure



9

Figure 2.1 – Overview of the SNaPshot primer extension procedure (taken from SNaPshot leaflet, Applied Biosystems)

The SNaPshot™ method of primer extension (Figure 2.1) employs the principle of extending an unlabelled oligonucleotide primer in the presence of fluorescently labelled ddNTPs. Each of the four ddNTPs is tagged by a different fluorescent dye (Table 2.3). Hence, when the primer extension products are run on an ABI DNA genetic analyser, the specific allele products can be differentiated from each other on the basis of which fluorescent dye they carry. This method can be used for individual and pooled genotyping (Pastinen et al 1996, Norton et al 2002). The pooled genotyping method will be outlined in detail later (section 2.6.14).

2.6.7 PCR

In order for primer extension to occur, we must amplify a fragment that contains the polymorphism we are interested in. Therefore the initial step in the SNaPshot reactions is PCR. The procedures for PCR are listed in section 2.6.3.

2.6.8 PCR Cleanup – SAP and ExOI

Shrimp Alkaline Phosphatase (SAP) (USB Corp., Cleveland, Ohio, USA) and Exonuclease I (ExoI) (New England Biolabs, Beverly, MA, USA) were used to cleanup any remaining dNTPs. These steps were important because any unincorporated dNTPs may interfere with the SNaPshot™ primer extension step. The following volumes of reagents were used for post-PCR cleanup: PCR Product 5µl; SAP (1U/µl) 0.5µl; SAP dilution buffer 0.5µl; ddH₂O 1µl; Exo I (1U/µl) 0.05µl. Total reaction volume 7.05µl. The PCR mixture was placed on a thermocycler under the following conditions (37°C for 30 min, 85°C for 15 min).

2.6.9 SNaPshot

The 10 μ l primer extension reaction is then set up. The ABI SNaPshot™ ddNTP primer extension kit (Perkin Elmer Applied Biosystems) provides the SNaPshot™ ready reaction mix which contains AmpliTaq™ DNA polymerase, fluorescently labelled ddNTPs (Table 2.3) and reaction buffer. The SNaPshot™ reaction mixture contained a total reaction volume of 10 μ l. This consisted of: 1 μ l SNaPshot™ (Standard Concentration); 1.5 μ l SNaPshot™ buffer; 2 μ l PCR product (post - SAP+ExoI stage); 0.4 μ l Extension Primer (5pmol/ μ l); 5.1 μ l ddH₂O. The thermocycler conditions for the extension reaction were 94°C for 2 min, followed by 25 cycles of 94°C for 5 sec, 43°C for 5 sec and 60°C for 5 sec.

Table 2.3 – The four ddNTPs used in SNaPshot with respective fluorescence colour and dye name

ddNTP	Dye Label	Colour of Analysed Data
A	dR6G	Green
C	dTAMRA	Black
G	dR110	Blue
T	dROX	Red

The only element in the SNaPshot™ primer extension reaction that is altered for a different polymorphism is the amount of extension primer added to the reaction. This is of particular importance when genotyping DNA pools for accurate estimation of allele frequencies, peak heights (which are a direct count of the fluorescence emitted) should be below a count of 6000. Genescan analysis software is unable to accurately quantify fluorescent signals above this level.

Norton and colleagues in Cardiff developed the method of genotyping DNA pools by SNaPshot™ primer extension. An extension primer was designed for each sequence variant that was to be genotyped in the pools. This included the 20 bases upstream of the variant with the base at the 3' end of the primer immediately adjacent to the variant. A homozygous individual, a heterozygous individual and a negative control (ddH₂O) were PCR amplified. Where a homozygous individual for the second allele was identified, it too would have been amplified. Following amplification, the PCR products were treated with Shrimp Alkaline Phosphatase and Exonuclease I as detailed in section 3.2.8 and the extension reaction was performed. For all sequence variants, the initial extension primer concentration used was 0.5pmol/μl. Following the extension reaction, the samples were again treated with Shrimp Alkaline Phosphatase (section 2.6.10) and run on the 377 Genetic Analyser (section 3.2.11).

Analysis of the heterozygote and the negative control determined if the reaction had worked and guarded against problems with contamination and self-priming (where the extension primer annealed to itself and extended by one base, thus creating a spurious allele). Analysis of the peak height of the homozygotes established what the appropriate concentration of extension primer to use. If the peak heights were below 6000, the extension primer concentration would remain at 0.5pmol/μl for genotyping the variant in the pools. If the peak height was above 6000, I reduced the primer concentration by a factor of 5 to 1pmol/μl and repeated the extension reactions. The extension reactions for all variants genotyped in the pools worked between 1pmol/μl and 5pmol/μl primer concentrations.

2.6.10 SAP 2 stage

If the SNaPshot™ reaction was left untreated, the unincorporated fluorescent ddNTPs would co-migrate with fragments of interest rendering the genotypes void. Removal of the 5' phosphoryl groups by phosphatase treatment alters the migration of the unincorporated fluorescent ddNTPs and thus prevents interference. Therefore it was important to degrade any of these ddNTPs before loading onto the genetic analyser. The following reaction was used for Post-SNaPshot™ reaction: SNaPshot™ product 5µl; SAP (1U/µl) 0.5µl; SAP dilution buffer 0.5µl; ddH₂O 1µl. Total reaction volume 7µl. The thermocycler reactions are identical to those stated for the post-PCR clean-up stage (section 3.2.8).

2.6.11 Run Samples

Initially we used 8% polyacrylamide gels to run SNaPshot™ products. This consisted of 11g of Urea; 3mls 10x TBE; 14mls Sigma Water; 5mls Long Ranger; 21µl TEMED; 150µl 10% AMPS. However, the SNaPshot™ products were running through the gel too fast, which affected quality. Therefore we increased the polyacrylamide gel to 10%: 11g Urea; 13 mls Sigma Water; 3 mls TBEx10; 6 mls Long Ranger; 150 µl 10 % APS; 21 µl TEMED.

All reactions were genotyped on an ABI 377 Genetic Analyser (Perker Elmin Applied Biosystems). In order to prepare samples for this, 1µl of treated extension reaction product was added to 0.5µl size standard (in house, Dr. Morris) and 2µl of Hi-Di Formamide (Table 2.4) (Perker Elmin Applied Biosystems). This mix was then denatured for 2 minutes then immediately placed onto ice. The rapid cooling process and added formamide

allows the single strands of DNA to remain denatured therefore exposing the ddNTPs to the laser capture.

Samples were loaded onto the ABI 377 according to manufacturer's instructions.

Table 2.4 Composition of formamide loading dye

Formamide Loading Dye	
Formamide	10ml
Xylene Cyanol	10mg
Bromophenol Blue	10mg
0.5M EDTA (pH 8.0)	200µl

2.6.12 Sample Analysis

Data collected by the ABI 377 Genetic Analyser was analysed using Genescan Analysis v3.1.2 and samples were analysed using Genotyper 2.5 (both, Perker Elmer Applied Biosystems).

The ABI Genescan v3.1.2 and Genotyper software suites performed allele calling for the SNaPshot™ reaction. Genescan v3.1.2 processes the resulting gel image. It detects the lanes on the gel and then extracts fluorescence intensity data over time for the four dyes used. A size curve is fitted to each set of three size standard peaks. A local average size is then estimated for each fragment peak.

The information for each lane of a gel is stored in a unique sample file, which can then be imported into the Genotyper software. Genotyper allows the alleles for each marker to be

inspected manually and labelled with a size in base pairs. Finally the genotypes for each marker can be outputted in table format listing the sample identifier, marker name and genotype. Samples for which no genotyped was assigned have "-1" entered into their genotype fields in the table.

2.6.13 Individual genotyping by SNaPshot™ primer extension

The methodology for genotyping sequence variants in individuals is the same as that used for pools. However in order to save both time and money, it is best to try and genotype a number of variants for an individual simultaneously. This is done by designing different sized primers for different variants. When the fluorescently tagged primer extension products representing all the alleles of a number of different variants are separated by electrophoresis each allele can be identified on the basis of both its size and colour. Therefore, where two or more variants share one or both extension ddNTPs, it is necessary to design their respective extension primers to be of different lengths.

There were three SNPs that were genotyped individually. They are listed in the results section, Chapter 3, Table 3.1.

2.6.14 DNA pooling

2.6.14.1 DNA pool preparation

DNA pools of 216 cases and 221 controls were created from fluorimetry readings. For the purposes of the study contained within this report, Dr. Derek Morris prepared the pools. The procedure initially involved re-quantifying the 8ng/μl stocks that were contained in the 96 deep well plates. The reason for this was that repeated uses of these boxes could expose the samples to evaporation and therefore have an impact on the concentration of DNA contained. Again, a 1:20 dilution of each sample was made before fluorometry (section 2.4.2). Once readings were available, equal concentrations of individual DNA samples were added to the pool eppendorf. However, as all samples are not equal starting volumes, different volumes must be taken to acquire the individual desired concentration. Therefore when pooling a case sample and a control sample, the final concentrations of each pool will differ due to the difference in volumes added. In order to have equal concentrations of case and control pool before PCR experiments, a working concentration of pools must be decided and higher concentration pools increased accordingly.

The next stage is to then quantify the pools using fluorimetry using the procedure in section 2.4.2. Once the required concentration was confirmed, the case and control pools were genotyped four times concurrently, along with three known heterozygotes using two SNP markers that had been previously individually genotyped. The results of these validation experiments are not shown as they were carried out by Dr. Morris.

2.6.14.2 DNA pooling method for testing association

We adopted a novel three-stage DNA pooling method to identify evidence of association. This technique is a modification of the DNA pooling method described by Norton et al, (2002). The Norton et al (2002) method used the SNaPshot™ genotyping assay (Applied Biosystems) to determine peak heights for a pair of SNP alleles in a pool of case DNA samples and in a pool of control DNA samples, and to determine the peak heights for one individual known to be heterozygous for this SNP. The ratio of the peak heights in the heterozygous sample (k) is used to correct for the unequal representation of alleles in a DNA pool. In practice, the frequency f in the pool of allele A is calculated as $f(a)=A/(A+kB)$ where A and B are the peak heights of the SNaPshot™ products representing alleles A and B. Where SNPs are taken from web-based databases, as in the case of this study, it is a time-consuming undertaking to genotype them in a panel of samples to identify a heterozygous individual. To circumvent this task we applied simulated heterozygote ratios (range of 0.1 to 10) to the analysis of each SNP. Each simulated k results in an estimate of the allele frequencies in the case and control pools (Figure 2.2). Computing allele frequencies with a range of these k values allows the identification of the maximum *potential* difference in allele frequencies between the case and control pools (Figure 2.3). Where this maximum potential difference in SNP allele frequency reaches nominal significance levels, further investigation is warranted – this involves identification of a heterozygous individual to determine the true k value for the SNP and hence accurate allele frequency estimation in the case and control pools. However, where this maximum potential difference in SNP allele frequency *does not* reach nominal significance levels, it can be categorized as showing no evidence of association in the case-control sample and thus excluded from further analysis. It is possible that the maximum potential difference in SNP allele frequency will not be identified within the

range of simulated k values stated (0.1 to 10). In such cases the maximum difference will have been found for $k = 0.1$ or $k = 10$. In these cases the range of k values should be extended (e.g. 0.1 to 0.01 , or 10 to 100) until the maximum potential difference has been found.

Therefore, stage 1 of this study involved genotyping 51 SNPs in case and control pooled DNA samples. Each SNP was assayed 5 times for each pool. Simulated k values were applied to each replicate and allele frequency estimates were an average of the 5 assay results. At the point of maximum potential difference in allele frequencies between the pools, allele frequencies were computed to allele counts and SNPs were tested for association using a 2 x 2 contingency table to calculate χ^2 statistics and p values. We set a conservative cut off value of $p < 0.1$ for transfer of any SNP to stage 2 to allow for experimental error. Each SNP that reached stage 2 was genotyped in a panel of individuals to identify a heterozygous individual. This provided the true allele peak height ratio (k) and was used to determine the estimated allele frequencies in the case and control pools. These frequencies were computed to allele counts and SNPs were tested for association using a 2 x 2 contingency table to calculate χ^2 statistics and p values. Any SNP that still demonstrated evidence of association at $p < 0.1$ was taken to stage 3 for individual genotyping of all cases and controls in the total sample to confirm the pooling result.

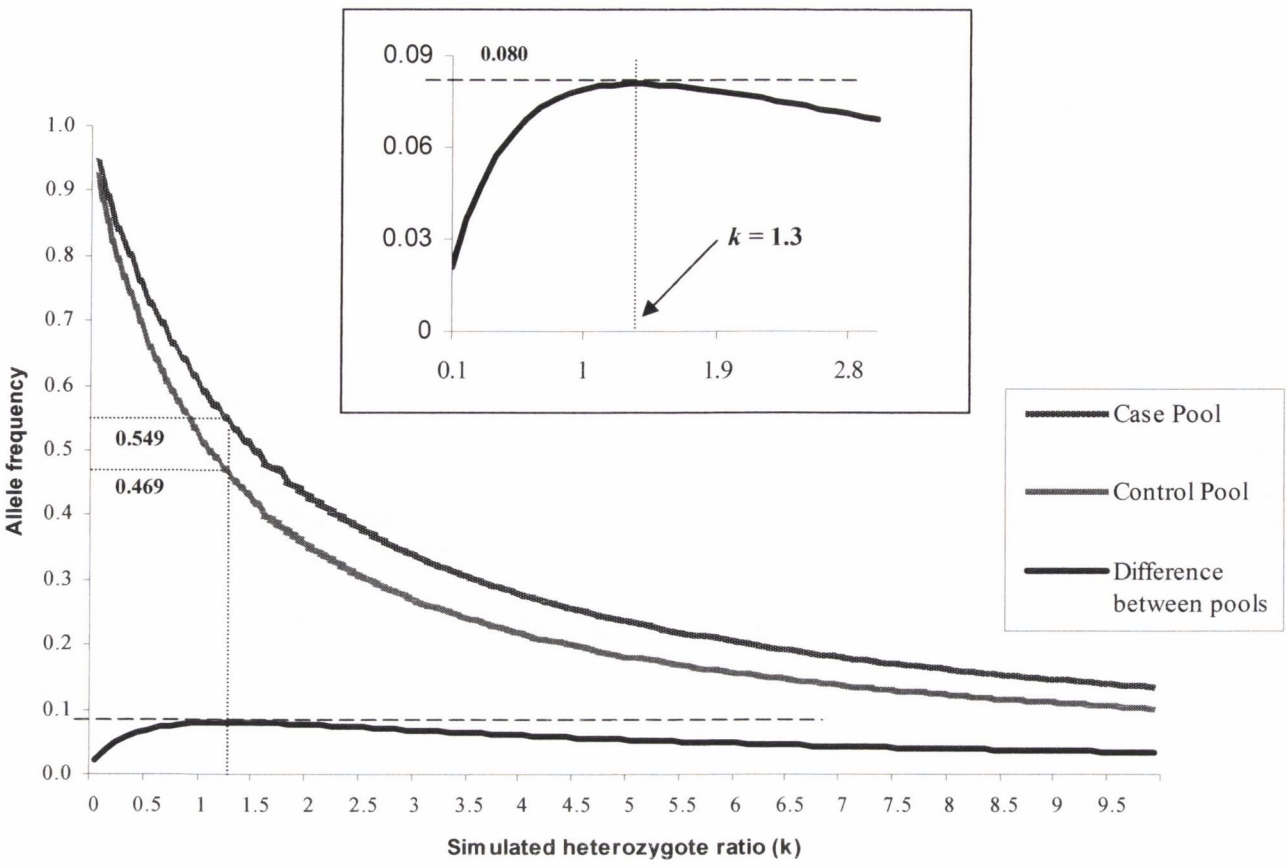
Case Pool	
Allele	Peak Height
A	1798
B	1135

Control Pool	
Allele	Peak Height
A	1500
B	1309

$$f(a) = A/(A + kB)$$

<i>k</i>	Case Pool Data	Control Pool Data	Difference
0.1	1798/(1798 + 0.1*1135) = 0.941	1500/(1500 + 0.1*1309) = 0.920	0.021
0.2	1798/(1798 + 0.2*1135) = 0.888	1500/(1500 + 0.2*1309) = 0.851	0.037
⋮	⋮	⋮	⋮
1	1798/(1798 + 1*1135) = 0.613	1500/(1500 + 1*1309) = 0.534	0.079
⋮	⋮	⋮	⋮
3.6	1798/(1798 + 3.6*1135) = 0.306	1500/(1500 + 3.6*1309) = 0.242	0.064
⋮	⋮	⋮	⋮
10	1798/(1798 + 10*1135) = 0.137	1500/(1500 + 10*1309) = 0.103	0.034

Figure 2.2 Example of sample data for a SNP genotyped in the case DNA pool and control DNA pool using the SNaPshot™ reaction. The raw data for each pool is fed into the formula $f(a) = A/(A + kB)$ where *k* is in the range 0.1 to 10. The difference between the results for each pool is listed for each simulated value of *k*.



<i>k</i>	Case Pool Data	Control Pool Data	Difference
1.3	$1798 / (1798 + 1.3 * 1135) = 0.549$	$1500 / (1500 + 1.3 * 1309) = 0.469$	0.080

Figure 2.3 The results for each pool, and the respective difference between them, are plotted against the range of simulated *k* values. The line plotting the difference between the pools, peaks at a value of 0.080 on the y-axis. This maximum potential difference between the pools occurs at *k* = 1.3 on the x-axis. This can be seen more clearly in the magnified section of the graph displayed in the box above the graph itself. At *k* = 1.3, the frequency of allele A is estimated at 0.549 in the case pool and 0.469 in the control pool. These estimates can be computed to allele counts and tested for evidence of association using a 2 x 2 contingency table to calculate a χ^2 statistic.

2.6.15 Statistics

The most commonly used test statistic for genetic association studies is the chi-square test. The chi-square distribution is one of the most widely used theoretical probability distributions in statistical significance tests. This is because under reasonable assumptions easily calculated quantities such as allele and genotype counts are proven to have distributions approximate to the chi-square distribution if the null hypothesis is true. Pearson's chi-square test (χ^2) is one such chi-square test calculated by finding the difference between each observed and theoretical frequency for each possible outcome, squaring them, dividing each by the theoretical frequency, and taking the sum of the results:

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$$

Where O_i = an observed frequency; E_i = an expected (theoretical) frequency, asserted by the null hypothesis. It makes two types of comparison: it tests goodness of fit and it tests independence. A test of goodness of fit establishes whether or not an observed frequency distribution differs from a theoretical distribution. A test of independence assesses whether paired observations on two variables, expressed in a contingency table, are independent of each other. Consultation of the chi-square distribution for 1 degree of freedom shows the probability of observing this difference when looking at a bi-allelic marker. In practice, the following websites were used in calculating the chi-square and p values for allele frequencies determined from the pooled DNA.

- Web Chi-squared calculator

<http://www.fourmilab.ch/rpkp/experiments/analysis/chiCalc.html>

- p value

http://www.georgetown.edu/faculty/ballc/webtools/web_chi.html

2.7 Materials and Methods for G72 and DAAO

2.7.1 Sample Size

At this point in the project, additional samples had been ascertained. The Irish sample now consisted of 299 cases and 645 controls. Details of the collection criteria for the additional cases are identical to that stated in section 2.1 (Materials and Methods). The additional control sample came from anonymous plasma donors via the Irish Blood Transfusion Service. Participating individuals gave written informed consent, met the same ethnicity criteria as cases, but were not specifically screened for psychiatric illness. Individuals taking regular prescribed medication are excluded from blood donation in Ireland and donors are not financially remunerated, making it unlikely that patient or socially disadvantaged groups (which may have higher rates of psychosis) were over-represented among controls. In the case sample 234 individuals met criteria for schizophrenia/schizophreniform disorder and 65 met criteria for schizoaffective disorder.

2.7.2 SNP genotyping

SNPs were selected for analysis based on a PubMed and Web of Science review of the schizophrenia G72/G30 and DAAO association literature and data presented at the XI and XII World Congress on Psychiatric Genetics. As many posters and a review presentation at the conferences showed, four of the original SNPs in G72 and four of the original SNPs in DAAO from the Chumakov et al study (2002) were statistically significant in several populations. This is highlighted in Chapter 4 Table 4.8 and Table 4.9. The selected SNPs for this study are detailed in Chapter 4 (Table 4.1 and Table 4.4). All SNPs were genotyped using the Taqman® assay (Applied Biosystems, Warrington, UK) on an ABI PRISM 7900HT Sequence Detection System. I carried out this genotyping of these SNPs.

However, in order to train a research assistant in our laboratory (Mr. Kevin Murphy) in Taqman® genotyping, he assisted me in setting up several of the plates for one assay.

2.7.3 Statistical Analysis

SNPs were tested for association with schizophrenia using a 2x2 contingency table to calculate a χ^2 statistic. Odds ratios were calculated for individual marker association using Epicalc 2000 v1.02 (Gilman and Myatt, 1998). Because of linkage disequilibrium (LD) individual SNPs may not be independent so applying a traditional Bonferroni correction (based on the assumption that individual tests are independent) would increase type II error under nominal significance levels ($p=0.05$). I applied the correction for multiple testing of SNPs based on spectral decomposition of matrices of pair-wise LD between SNPs described by Nyholt (2004). This correction, implemented in the programme SNPSpD (<http://genepi.qimr.edu.au/general/daledN/SNPSpD/>), allows for the potential non-independence of SNPs in estimating the total number of tests carried out when multiple markers are genotyped at a locus.

Haplotype association analysis was performed using FASTEHPLUS, which employs the gene-counting algorithm, a form of the EM algorithm, to calculate haplotype frequencies (Zhao and Sham, 2002). Significance levels were determined empirically using simulations with 10,000 iterations. The program GENE-COUNTING was used to output individual haplotype frequencies in cases and control samples (Zhao et al 2002). Inter-marker linkage disequilibrium (LD) was measured using D' and r^2 using Haploview (Barrett et al 2005; detailed in section 2.10).

2.8 Materials and Methods - HSPA8

2.8.1 Identification of candidate genes

The study by Prabakaran et al 2004 identified a large number of plausible SZ candidate genes from their functional studies (See Table 2.5). However, there are a number of reasons (see Chapter 5), for differences in gene expression in SZ post mortem studies, which are not related directly to disease pathogenesis. In selecting candidate genes for investigation for this study, I screened functional candidate genes from the Prabakaran study for additional positional evidence of involvement in SZ susceptibility.

I reviewed positional evidence from linkage studies in the Irish and other populations and from a recent SZ linkage meta-analysis by Lewis et al (2003). The meta-analysis was carried out because replication of significant genetic linkage has been inconsistent. By applying a rank based genome scan meta-analysis (GSMA) to 20 published linkage studies and assigning each marker to one of 120 x 30cM bins, regions of significant linkage were identified. In order to identify a smaller number of candidate genes I compared the chromosomal location of the oxidative stress genes identified by Prabakaran et al (2004) (n=22) with the significant chromosomal bins identified by Lewis et al (2003) (see Table 2.6). By using this approach I identified four candidate genes based on positional and functional evidence.

2.8.2 HSPA1A

Oxidative stress induces molecular chaperones (Macario and Macario 2005) such as Heat shock 70kDa (HSP70), which preserve cellular integrity (Fekete et al 2005) by binding to

Table 2.5 Protein expression is significantly altered in the prefrontal cortex of schizophrenia (taken from Prabakaran et al 2004).

Protein	White matter		Gray matter		Accession number (SwissProt/TrEMBL)	Mascot score	Sequence coverage (%)
	Fold change	P-value	Fold change	P-value			
Pyruvate kinase, muscle (PKM1)	-1.29 to -1.58 (4)	0.00045 - 0.0074	-1.3	0.039	P14618	27 8-907	14-48
Pyruvate kinase, muscle (PKM2)					P14786	278-907	14-48
Acetate kinase, mitochondrial (ACO2)	-1.25 to -1.29 (2)	0.036 - 0.043	-1.37 to -1.41 (3)	0.000011 - 0.0012	Q99798	85-723	5-33
Phosphoglycerate dehydrogenase (PHGDH)			-1.3	0.004	Q43175	75	4
EH-domain containing protein 2 (EHD2)	1.29	0.048			Q9N2N4	89-141	4
EH-domain containing protein 3 (EHD3)					Q9N2N3	89-141	4
Triosephosphate isomerase 1 (TPI1)			-1.47	0.0096	P00938	72	8
Hexokinase 1 (HK1)	-1.31	0.035	1.71	0.01	P19367	168	8
Tu translation elongation factor, mitochondrial (TUFM)			-1.4	0.0043	P49411	430	23
Ubiquinol-cytochrome c reductase core protein 1 (UQCRC1)	-1.44	0.016	-1.35	0.0023	P31990	143	11
Glucose-regulated protein, 58 kDa (GRP58)			-1.27	0.0068	P30101	247	20
Moesin (MSN)			1.19	0.029	P26038	115	4
Gelsolin (amyloidosis, Finnish type) (GSN)	-1.42	0.0028	-1.36	0.025	P06396	332-459	19-20
Malate dehydrogenase 1, NAD (soluble) (MDH1)	-1.29	0.00064	-1.55	0.00027	P40925	79	6
Peroxisomal protein 1 (PRDX1)			-1.6	0.024	Q05839	89	14
Peroxisomal protein 2 (PRDX2)	-1.28	0.02	-1.38	0.00032	P32119	124	22
Heat-shock 70 kDa protein 1 (HSPA1A)					P06107	520	24
Heat-shock-related 70 kDa protein 2 (HSPA2)					P34832	465-554	23-28
Heat-shock 70 kDa protein-like 1 (HSPA1L)	-1.31 to -1.56 (3)	0.012 - 0.039	-1.23 to -1.41 (4)	0.00086 - 0.032	P34931	520	24
Heat-shock 70 kDa protein 8 (HSPA8)					P11142	507	28
Heat-shock 70 kDa protein 5 (HSPA5)					P11021	59	6
Aldolase A, fructose-bisphosphate (ALDOA)	-1.25	0.027	-1.24 to -1.47 (5)	0.000085 - 0.035	P04075	387-611	39-56
Aldolase C, fructose-bisphosphate (ALDOC)					P06972	119-509	15-48
Glyceroldehyde-3-phosphate dehydrogenase (GAPD)			-1.32 to -1.5 (5)	0.000095 - 0.016	P00354	162-279	23-29
Pyruvate dehydrogenase E1 component, alpha 1 (PDHA1)			-1.48	0.0051	P04859	387	22
NADH dehydrogenase (ubiquinone) Fe-S protein 1 (NDUFS1)	-1.36 to -1.39 (2)	0.0094 - 0.024	-1.35 to -1.52 (2)	0.0013 - 0.0022	Q8H1C4	54-75	2
Tubulin, alpha 2 (TUBA2)					Q13748	90	13
Tubulin, alpha 6 (TUBA6)	-1.24 to -1.61 (2)	0.034 - 0.0035	-1.29 to -1.53 (5)	0.0031 - 0.043	Q890E3	90-283	9-20
Tubulin, alpha 1 (testis specific) (TUBA1)					P06209	109-283	11-20
Tubulin, beta 5 (TUBB5)					P06218	248-294	18-25
Glutathione-S-transferase M3 (brain) (GSTM3)			-1.19	0.044	P21295	143	14
Glutathione transferase omega (GSTL2)					P78417	94	14
Enolase 2, (gamma, neuronal) (ENO2)	-1.29	0.042	-1.18 to -1.41 (6)	0.00065 - 0.03	P09104	441	26
Enolase 1, (alpha) (ENO1)					P06793	167-610	11-37
Leucine aminopeptidase 3 (LAP3)			-1.25	0.025	P28838	201	11
Aldehyde dehydrogenase 1, family member A1 (ALDH1A1)	-1.28	0.014			P00052	294	14
Fascin homolog 1, actin binding protein (FSCN1)	-1.26	0.039			Q16628	388	20
N-ethylmaleimide-sensitive factor (NSF)			1.62	0.011	P46459	141-193	8
CDC10 cell division cycle 10 homolog (S. cerevisiae) (CDC10)	-1.22 to -1.36 (2)	0.028 - 0.038	-1.52	0.0017	Q16181	123-214	11-18
Glutamate-aminonucleoside ligase (glutamine synthase) (GLUL)					P15104	183	10
Actin, alpha 2, smooth muscle, aorta (ACTA2)					P03996	222-369	16-29
Actin, beta (ACTB)					P02570	193-369	15-26
Actin, alpha cardiac muscle (ACTC)	-1.15 to -1.55 (5)	0.011 - 0.048	-1.18	0.025	P04270	222-316	16-29
Actin, gamma 1 (ACTG1)					P02571	259-318	18-42
Actin, alpha 1, skeletal muscle (ACTA1)					P02568	222-316	16-29
Spectrin, alpha, nonerythrocytic 1 (alpha-fodrin) (SPTAN1)	-1.41	0.012	-1.29 (2)	0.037 - 0.048	Q13813	104-612	2-9
Creatine kinase, brain (CKB)			-1.24 to -1.3 (3)	0.0014 - 0.0023	P12277	95-193	9-12
Actinin, alpha 4 (ACTN4)	-1.3 to -1.61 (2)	0.0055 - 0.0074	1.58	0.026	Q43707	155-234	4-8
Carbonyl reductase 3 (CBR3)					Q75829	103	11
Carbonyl reductase 1 (CBR1)			-1.23 to -1.41 (2)	0.0029 - 0.03	P16152	151-387	26-39
Quinoid dihydropteridine reductase (QDPR)	-1.36	0.021			P08417	154	12
Phosphoglycerate mutase 1 (brain) (PGAM1)			-1.18	0.025	P18669	95	15
Phosphoglycerate mutase 2 (muscle) (PGAM2)					P15259	95	15
Phosphotyrosine phosphatase inorganic pyrophosphate phosphatase			-1.25	0.0067			
Ubiquitin carboxyl-terminal esterase L1 (UCHL1)			-1.41	0.0022	Q9H008	102	12
Esterase Dfomylglutathione hydrolase (ESD)	-1.35	0.02			P10988	245	32
Tyrosyl-tRNA synthetase (YARS)	-1.3	0.021			P54577	340	17
Glycolipid transfer protein (GLTP)			-1.34	0.0042	Q9N2D2	189	9-14
Actin-related protein 2/3 complex, subunit 1A, 41 kDa (ARPC1A)			-1.62	0.0017	Q92747	60	4
Actin-related protein 2/3 complex, subunit 1B, 41 kDa (ARPC1B)					Q15143	60	4
Dynamin 1 (DNM1)					Q05193	150-186	5-6
Dynamin 2 (DNM2)	-1.17	0.033	1.48 to 1.82 (2)	0.01 - 0.034	P06570	150-186	5-6
Transferrin (TF)	1.39	0.007	-1.17 to -1.22 (2)	0.0094 - 0.02	P02787	63-126	7-9
ATPase, H ⁺ transporting, (vacuolar proton pump) (ATP9A1E1)			-1.33	0.0069	P36543	336	33
2,3'-cyclic nucleotide 3' phosphodiesterase (CNP)	-1.28 to -1.63 (7)	0.0024 - 0.032	-1.41	0.0069	P09543	54-853	2-51
Brain abundant, membrane attached signal protein 1 (BASP1)	2.03 to 2.4 (2)	0.0089 - 0.035			P08073	177-204	33-47
Dihydropyrimidinase-like 2 (DPYSL2)					Q16333	106-486	16-33
Dihydropyrimidinase-like 5 (DPYSL5)	-1.33 to -1.53 (4)	0.0001 - 0.034	-1.33 to -1.46 (3)	0.0031 - 0.041	Q9BPU6	85	5
Dihydropyrimidinase-like 4 (DPYSL4)					Q14531	52	2
Collapsin response mediator protein 1 (CRMP1)					Q14194	321	16
Septin 3 (SEPT3)	-1.2	0.022	-1.19	0.0063	Q9LH03	53	6
SH3-domain GRB2-like 2 (Endophilin 1) (SH3GL2)			-1.32	0.014	Q99902	68	6
Albumin (ALB)	-1.69	0.0063			P02768	240-261	14

misfolded or unfolded proteins promoting either correct assembling or proteolytic degradation (Mahadik et al 2001; Fekete et al 2005). Heat shock 70kD protein 1A (HSPA1A) (MIM 140550) is part of the HSP70 family and is located on chromosome 6p21.3. It spans 2382 base pairs containing one exon and no introns. It is located 530 base pairs downstream of HSPA1L. Milner and Campbell (1990) determined that the HSPA1A gene encodes a predicted 641-amino acid protein. By Northern blot analysis of HeLa cell RNA, they detected an approximately 2.4-kb HSPA1A transcript that was constitutively expressed at low levels and was induced following heat shock.

An NCBI nucleotide search, identified a transcript of mRNA NM_005345.4 whose sequence was found on the contig NT_007592 (gi:29804415). By identifying exonic and intronic structures (section 5.3.3.2) it was apparent that this gene had pseudogenes with almost 100% homology. This made designing primers very difficult because one could not be sure that you were amplifying the correct gene. Although it is sometimes possible to amplify a highly homologous sequence using nested PCR, the process is laborious and may still yield unsuccessful results. A fellow PhD student has tried for over one year to amplify another pseudogene with no success. After advice from my peers and senior post-docs in the lab, this gene was dropped from further analysis.

2.8.3 HSPA1L

The Heat shock 70kDa protein 1-like (HSPA1L) also known as HSP70-HOM (MIM 140559) is also part of the HSP70 (see HSPA1A) family and is located on chromosome 6p21.3. The literature on this gene is conflicting. Sargent et al. (1989) identified the HSP70-HOM gene as a region with similarity to HSP70-1 (HSPA1A; 140550) that was located approximately 4 kb telomeric to HSP70-1 in the class III region of the major

Table 2.6: Comparison of the chromosomal location of oxidative stress genes identified by Prabakaran et al (2004) with the significant chromosomal bins identified by Lewis et al (2003). Matches are highlighted in yellow.

Gene	Clone Number	Chromosome	Lewis et al SZ MetaAnalysis (Y/N)	PubMed search with SZ	Pubmed ID
PRDX1	NM_002574	1p34.1	No	Nil	
PRDX2	NM_181738	19p13.2	No	Nil	
HSPA1A	NM_005345	6p21.3	Yes	Nil	
HSPA2	NM_021979	14q24.1	No	Nil	
HSPA1L	NM_005527	6p21.3	Yes	Nil	
HSPA8	NM_153201	11q24.1	Yes	Nil	
HSPA5	NM_005347	9q33-q34.1	No	1	10693161
ALDOA	NM_184043	16q22-q24	No	Nil	
ALDOC	NM_005165	17cen-q12	No	Nil	
GAPD	NM_002046	12p13	No	1	10392541
PDHA1	NM_000284	Xp22.2-p22.1	No	Nil	
NDUFS1	NM_005006	2q33-q34	No	Nil	
TUBA2	NM_006001	13q11	No	Nil	
TUBA6	NM_032704	12q12-q14	No	Nil	
TUBA1	NM_006000	2q36.1	No	Nil	
TUBB5	NM_006087	19p13.3	No	Nil	
GSTM3	NM_000849	1p13.3	Yes	Nil	
GSTTLp28	NM_004832	10q25.1	No	Nil	
ENO2	NM_001975	12p13	No	Nil	
ENO1	NM_001428	1p36.3-p36.2	No	Nil	
LAP3	NM_015907	4p15.33	No	Nil	
ALDH1A	NM_000689	9q21.13	No	Nil	
FSCN1	NM_003088	7p22	No	Nil	
NSF	XXX	XXX		5	12892855, 11555028, 11403936, 10641584, 7178849
CDC10	BC025987	???		Nil	
GLUL	BC018992	???		Nil	

histocompatibility complex on 6p21.3. However, NCBI positions HSPA1L only 530 base pairs apart from HSPA1A. The predicted 641-amino acid HSP70-HOM protein is 84% identical to HSPA8 (600816) protein. Using NCBI's nucleotide search, I identified a transcript of mRNA NM_005527.2 whose sequence could be found on the contig NT_007592. (gi:29804415). By identifying exonic and intronic structures (section 5.3.3.2) it was apparent that this gene had pseudogenes with almost 100% homology. Again, this made designing primers very difficult because there was no unique sequence for annealing primers to amplify this gene and as mentioned in section 5.2.1 the consensus of colleagues was to drop this gene from further analysis.

2.8.4 HSPA8

Heat shock 70kDa protein 8 (HSPA8) (MIM:600816) is located on chromosomal region 11q24.1. One of the significant linkage bins in the meta-analysis by Lewis et al (2003) spanned this region (11q22.3-q24.1). By using the search facility of NCBI's nucleotide database I identified an mRNA NM_153201.1 whose sequence was found on the contig NT_033899. (gi:37540935). Further research using NCBI and ACEVIEW identified two suggested isoforms of HSPA8: NM_006597 mRNA coding for NP_006588 protein and NM_153201 mRNA coding for NP_694881 protein.

The product encoded by HSPA8 belongs to the heat shock protein 70 family which is involved in cellular stress response, such as heat shock and oxidative stress. They are an evolutionarily highly conserved family of proteins that act as molecular chaperones ensuring that correct protein folding occurs. Intracellular accumulation of abnormally folded proteins has a significant effect on the cell by activating the heat shock/stress response. Interestingly, one sign of ageing is the accumulation of abnormally folded

proteins, which may suggest that HSP70 have less ability to activate the stress response with increasing age (Welch et al 1998). The HSPA8 protein binds to nascent polypeptides to facilitate correct folding. It also functions as an ATPase in the disassembly of clathrin-coated vesicles during the transport of membrane components through the cell

Two alternatively spliced variants have been characterized to date. Variant 2 uses an alternate in-frame splice site in the 3' coding region, unlike variant 1, resulting in a shorter protein (isoform 2).

2.8.5 GSTM3

Glutathione S-transferase M3 (brain) (GSTM3) (MIM:138390) is located on chromosomal region 1p13.3. One of the significant linkage bins in the meta-analysis by Lewis et al (2003) spanned this region (1p13.3-q23.3). By using NCBI's search facility of the nucleotide database I identified an mRNA NM_000849.2 whose sequence was found on the contig NT_019273. (gi:37539512)

According to ACEVIEW, cytosolic and membrane-bound forms of GSTM3 are encoded by two distinct supergene families. At present, eight distinct classes of the soluble cytoplasmic mammalian glutathione S-transferases have been identified: alpha, kappa, mu, omega, pi, sigma, theta and zeta. This gene encodes a glutathione S-transferase that belongs to the mu class. The mu class of enzymes functions in the detoxification of electrophilic compounds, including carcinogens, therapeutic drugs, environmental toxins and products of oxidative stress, by conjugation with glutathione. The genes encoding the mu class of enzymes are organized in a gene cluster on chromosome 1p13.3 and are known to be highly polymorphic. These genetic variations can change an individual's

susceptibility to carcinogens and toxins as well as affect the toxicity and efficacy of certain drugs. Mutations of this class mu gene have been linked with a slight increase in a number of cancers, likely due to exposure with environmental toxins. According to LocusLink, GSTM3 is involved in the establishment of the blood/nerve barrier, metabolism and glutathione transferase activity.

2.8.6. Mutation detection using resequencing

2.8.6.1 Sample Size

At this point in my PhD, additional SZ and control samples had been ascertained in order to increase the power of our sample. The collection criteria for the additional SZ samples is identical to that stated in section 2.1. The additional control samples came from anonymous plasma donors via the Irish Blood Transfusion Service. The new sample size consisted of SZ patients (n=299) and controls (n=645). A reference panel from the control sample (n=92) was sent to Ellipsis, Canada, for commercial genotyping in order to determine LD structure for the genes under study.

2.8.6.2 Identification of exonic and intronic structures

In order to perform candidate gene analysis, the intron/exon boundaries of HSPA8 had to be determined. Using NCBI's nucleotide database, a search was performed using each individual gene as the query. This gave a number of hits and mRNA from *Homo sapiens* was identified (the exact mRNA's identified for the four candidate genes are contained in the 'Identification of candidate genes' section 2.8.1). If there was alternative splicing identified in any of the genes, each alternative exon was also resequenced. The primers were designed to ensure that both alternatively spliced exons were resequenced. Each

identified mRNA was then blasted against the human genome database to identify homology between the exonic structures of the individual gene and the entire genome. Any matching regions indicated areas of homology where primers should not be designed. This is important in large exonic structures over 400 base pairs as complete or near complete homology with an exon in another genomic region can make amplification of this exon by PCR problematic. This also helps identify possible pseudogenes.

Blasting the mRNA against the entire genome also served the purpose of identifying the full genomic region that contained the gene. Once identified, the entire gene (exons and introns) were blasted against the genome in order to determine possible intronic homology with the remainder of the human genome. If introns contain high homology with other genomic regions, amplifying PCR fragments spanning the gene becomes difficult. As previously mentioned, two of the four identified candidate genes for this study (HSPA1A and HSPA11), were removed from the study due to the presence of pseudogenes.

2.8.6.3 Primer Design

Primers were designed using the Primer3 website, described in (section 2.6.2). When resequencing a gene it is important to have overlap between PCR fragments of between 20 – 40 base pairs. This is because the first 10 - 20 base pairs of a fragment are difficult to read in a sequencing reaction and a SNP in this region may be difficult to identify. By overlapping fragments, you ensure that you capture all sequence information. Once the PCR fragments were designed, they were blasted against the human genome in order to make sure that they only showed homology to the region of interest. Figure 2.4 represents a schematic of the fragments designed to span HSPA8 and Table 2.7 contains details of the

primer design. Fourteen fragments were designed to span the gene. In addition, 1.5 Kb was sequenced upstream of the gene, and 500 bp was sequenced downstream of the gene.

2.8.6.4 PCR optimisation

All fragments were optimised initially using PCR programs 54Q and 57Q (see appendix Table A.8). During PCR optimisation, fragments 2 and 3 did not amplify well at normal annealing temperatures. This was likely to have been due to the high GC content in the 5' region of the gene. Therefore, three new sets of PCR primers were designed: Fragment-2a, Fragment-2b and Fragment-3 with annealing temperatures of 72°C. This higher annealing temperature allows use of Deep Vent (exo-) DNA polymerase (New England Biolabs) and programme DV72 (see appendix). Deep Vent (exo-) DNA polymerase is more stable at high temperatures (72°C) than normal DNA polymerases and is used when the fragment being amplified has a high GC nucleotide content.

2.8.6.5 Resequencing

Resequencing was carried out in 15 randomly selected SZ individuals (30 chromosomes) in order to identify novel mutations. This gives 95% power to detect SNPs with a minor allele frequency (MAF) > 0.1 and 80% power to detect SNPs with MAF > 0.05. However in order to obtain 95% power for detecting MAF < 0.1, prohibitively large numbers of patients would require resequencing. Therefore one caveat of this method is that rare variants would probably not be detected. Following PCR, a cleanup step was performed to remove residual PCR primers and unlabelled dNTPs to ensure that they did not interfere with the subsequent sequencing reaction. The cleanup process was performed using the QIAquick PCR Purification Kit (Qiagen Ltd, Crawley, West Sussex, UK).

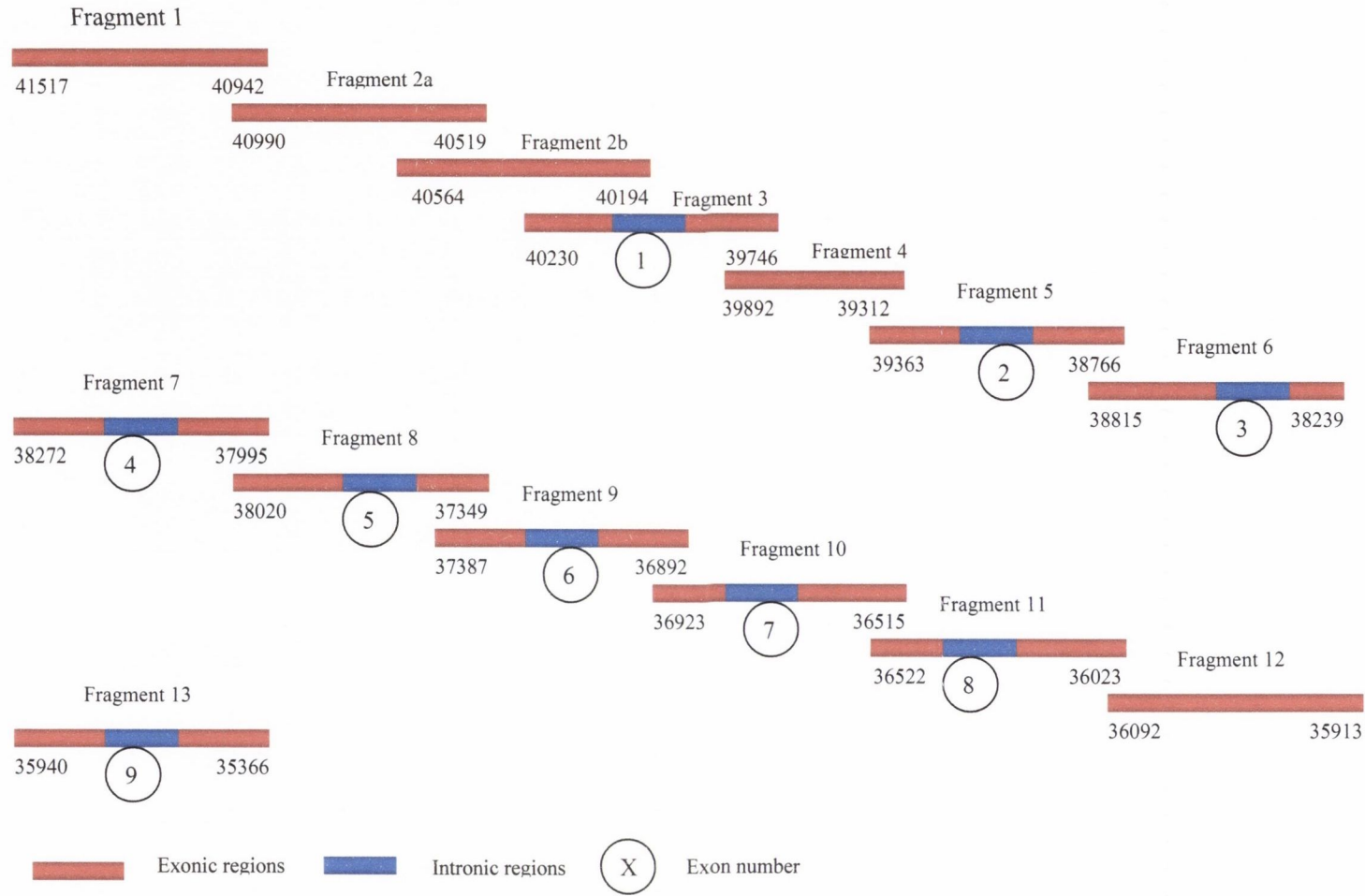


Figure 2.4: HSPA8 fragments designed for sequencing. Each fragment is represented by a bar with the fragment name on top. Exons are highlighted in blue and are identified by a number in a circle. The numbers below the fragments refer to the base pairs covered using BAC AP000926 (gi#:31790747). Details on primer sequence and PCR protocols for this gene given in Table 5.1.

Table 2.7: PCR primers designed to amplify 14 fragments spanning the gene HSPA8.

HSPA8 FRAGMENTS FOR SEQUENCING							
Fragment Number		PCR primer sequence	Primer	Start ^a	Stop ^a	Size of Fragment (bp)	PCR protocol
1	F	CTTTTCAGAACCGCCTCA	Left	40942	40959	575	57Q
	R	AAATCTTTTCCCCCATCG	Right	41517	41500		
2a	F	TCCTCGGGCCTCGCCCCCTTC	Left	40194	40214	370	DV72
	R	CCGGCTCCCGCGCCCAAAC	Right	40564	40546		
2b	F	GGAATCGCGCGGGGCCTGAAG	Left	40519	40539	471	DV72
	R	TCCGCGGAGGGATCCAGACACGA	Right	40990	40968		
3	F	GCCTCCCCTGGGGCCACTGC	Left	39746	39765	484	DV72
	R	CGGTGGGGTGGGTGCGGAAG	Right	40230	40211		
4	F	TTAACCAGGAAAAACGTATGG	Left	39312	39332	580	54Q
	R	GGGTTCTGAGAATCTCGT	Right	39892	39874		

^a The base pair reference for the start and stop region of the designed PCR fragments refers to the BAC AP000926 (gi#31790747).

Table 2.7 continued.

HSPA8 FRAGMENTS FOR SEQUENCING							
Fragment Number		PCR primer sequence	Primer	Start ^a	Stop ^a	Size of Fragment (bp)	PCR protocol
5	F	CAAGTCCATGATTACTCAAATACC	Left	38766	38789	597	54Q
	R	GGTGATGGGCACTATTACCT	Right	39363	39344		
6	F	CAGCAAATGGATTAATGCTG	Left	38239	38258	576	57Q
	R	TGAATTCACCTACCTATGAACTGT	Right	38815	38792		
7	F	TAATCCGAACTTGCATCACA	Left	37995	38014	277	57Q
	R	AGCTGTTCTTTCACCAGCAT	Right	38272	38253		
8	F	CCTGCCTTTAGGGTTAATTG	Left	37349	37368	671	57Q
	R	ACCATTTGTGATGCAAGTTC	Right	38020	38001		
9	F	CCATCATCATAGCGAGTCAG	Left	36892	36911	495	57Q
	R	GCAGGTAACAATGGTATCTCAA	Right	37387	37366		

^a The base pair reference for the start and stop region of the designed PCR fragments refers to the BAC AP000926 (gi#31790747).

Table 2.7 continued

HSPA8 FRAGMENTS FOR SEQUENCING							
Fragment Number		PCR primer sequence	Primer	Start^a	Stop^a	Size of Fragment (bp)	PCR protocol
10	F	TCAAGACCAGATGACAGTGC	Left	36515	36534	408	57Q
	R	GGGAAGTCTTGACTGACTCG	Right	36923	36904		
11	F	CTTGAATTCTGGTGGAAACC	Left	36023	36042	499	57Q
	R	GGTCTTGACAGGGATAATGG	Right	36522	36503		
12	F	AGTCTGGCGCAAACCTTAC	Left	35913	35932	179	57Q
	R	AGCTGGATTCAGTGTAGGG	Right	36092	36073		
13	F	CTGGCCTGGGAGATTATTTA	Left	35366	35385	574	57Q
	R	GCCCTCTGTAAGAGTTTGC	Right	35940	35921		

^a The base pair reference for the start and stop region of the designed PCR fragments refers to the BAC AP000926 (gi#31790747).

Sequencing of the amplified DNA was performed using the ABI BigDye Terminator (v3.0) Cycle Sequencing kit (Applied Biosystems, Foster City, CA, USA). The process is similar in nature to that of a PCR except fluorescently labelled ddNTPs are present in the reaction mix that randomly terminate the amplification process and the mix contains proof reading polymerase. With the exception of fragment 7 and 12, the reaction mix contained 4µl BigDye Terminator reaction mix, 10µl PCR product (post cleanup) and 2.5pmol primer. The mix was made up to a final volume of 20 µl with sterile H₂O. For fragments 7 and 12, the PCR product, Big Dye Terminator reaction mix and primer were reduced by half, compensating with sterile H₂O to a final volume of 20µl. The cycling conditions for the sequencing reaction were an initial 94°C for 30 sec, 25 cycles of 96°C for 10 sec, 50°C for 10 sec and 60°C for 4 min and a final 4 °C for 10 min. The run time was 2hr 35 min. In addition, a positive control supplied by the manufacturer was run with every set of reactions. This consisted of 1µl pGem (DNA template), 2µl M13 (primer) 4µl BigDye Terminator reaction mix and 13µl of water. All sequencing reactions were performed in the forward and reverse directions to confirm results.

Following the sequencing reaction another cleanup step was performed using the DyeEx 2.0 Spin Kit (Qiagen Ltd, Crawley, West Sussex, UK). This process removes the excess unincorporated fluorescently labelled ddNTPs, which would otherwise interfere with the analysis of the sequencing products.

Sequencing reactions were run on the ABI 3100 DNA Sequencer using a 36cm array and the results were visualised and analysed using SeqScape (v.2) and Sequencing Analysis 5.1 (Applied Biosystems, Foster City, CA, USA). Figure 2.5 and 2.6 is a schematic representing the methodology behind resequencing with fluorescently labelled ddNTPs.

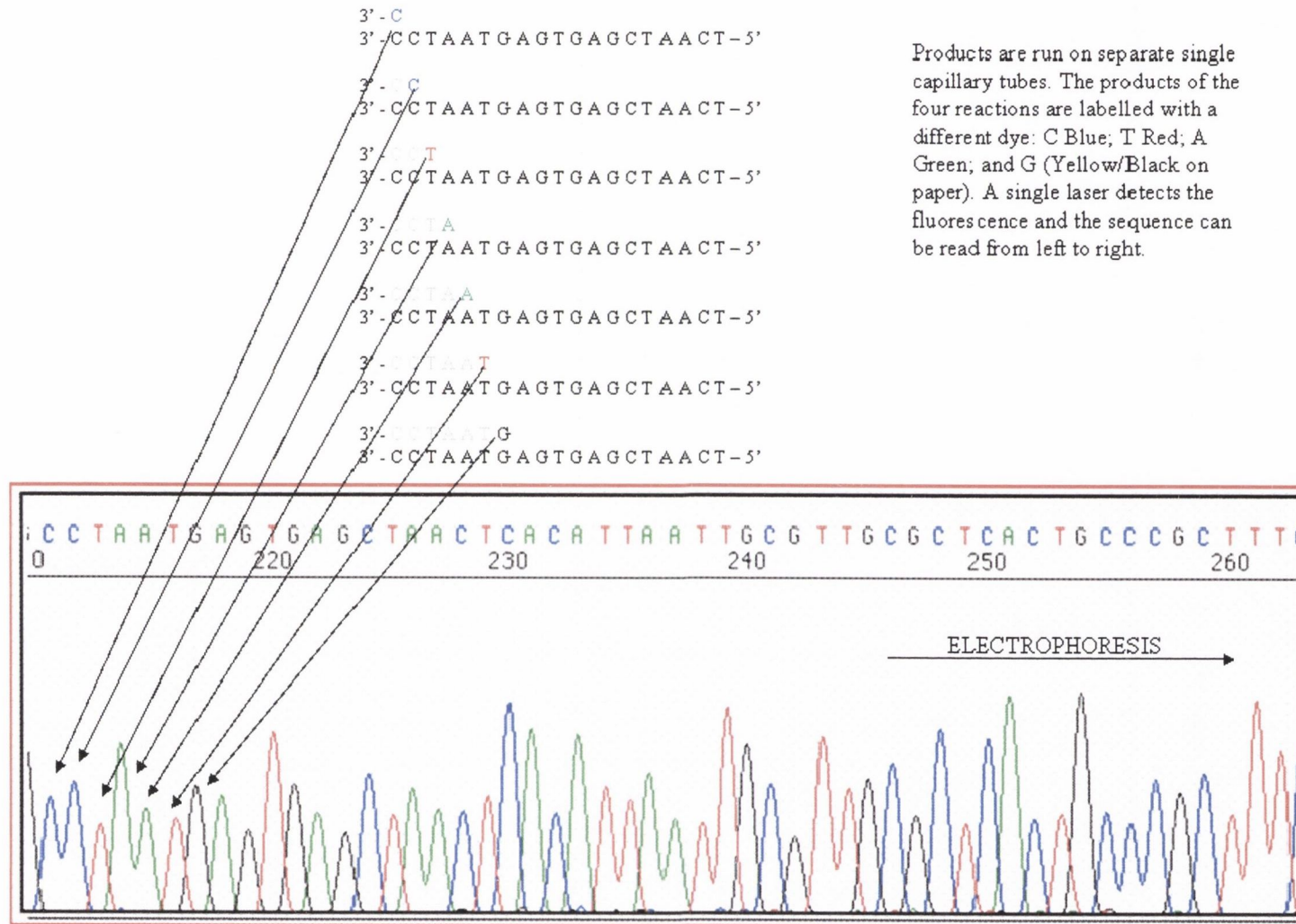


Figure 2.5 A schematic representation of how an electropherogram is created from a sequencing reaction. Contained within the Big Dye mix are ddNTPs each labelled with a different fluorescence. As the template DNA is read, the corresponding ddNTP anneals and terminates further extension.

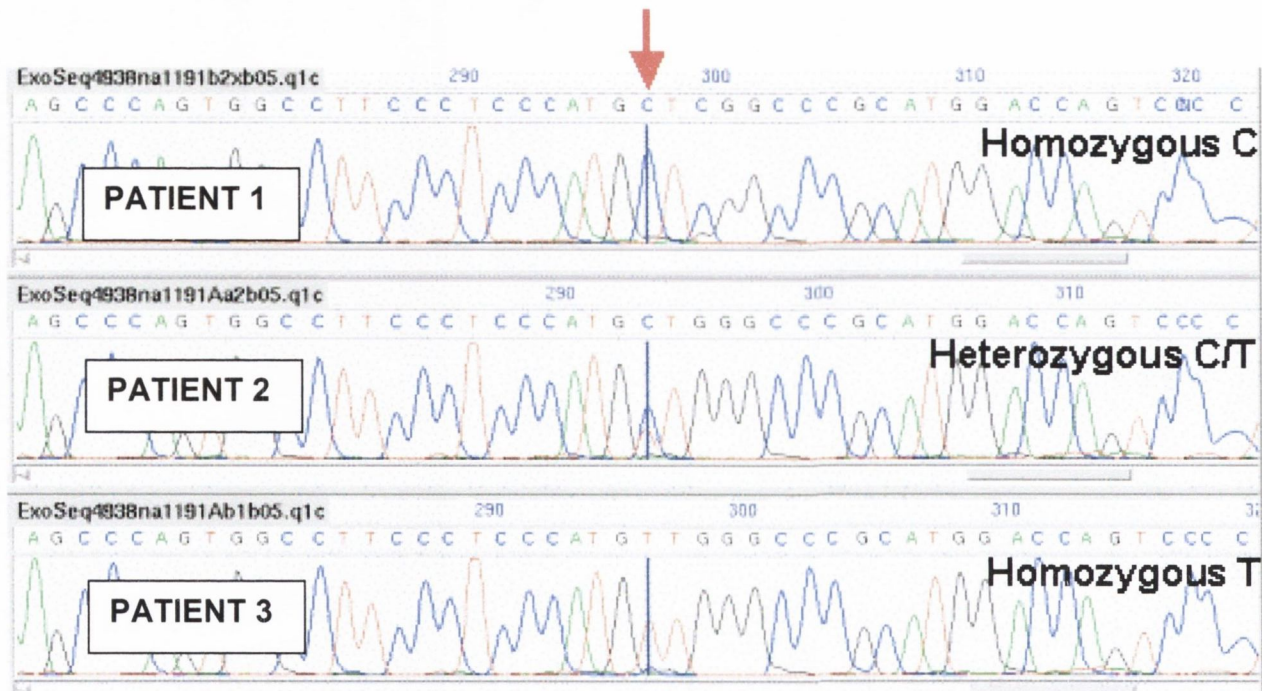


Figure 2.6 This is an example of an electropherogram of a sequence reaction. The DNA sequence of three patients is shown. A SNP has been located with a C/T polymorphism (indicated by a red arrow). Patient one is homozygous for the C allele. Patient two is heterozygous (C and T alleles). The third patient is homozygous for the other allele (T).

During resequencing of fragment 11, a poly A tail located after exon 8 interfered with the ddNTP labelling process resulting in poor sequencing results for that fragment. A new second forward primer was designed in order to allow resequencing of exon 8 (see figure 2.7) with the resultant loss of information for 75 base pairs.

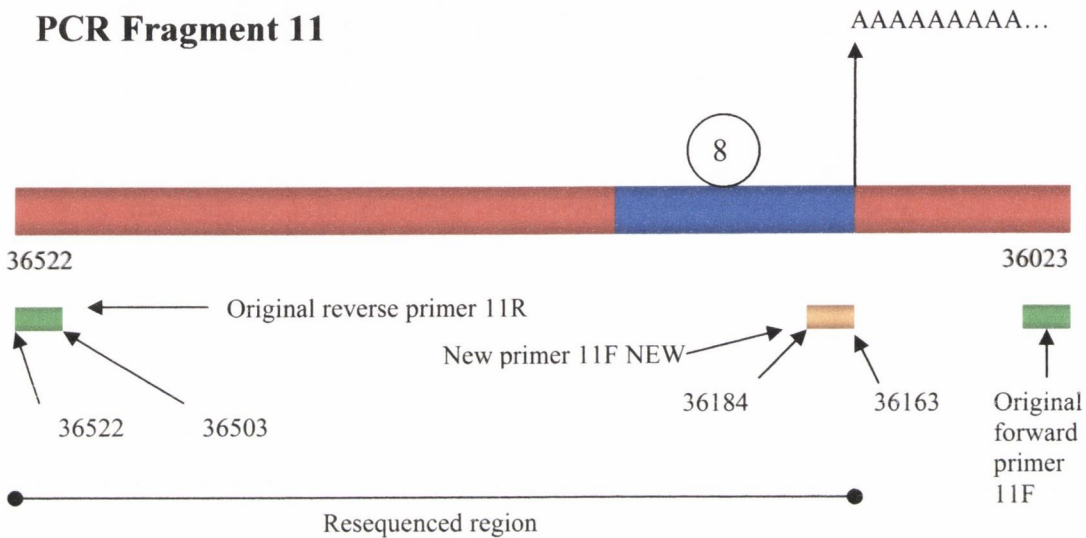


Figure 2.7 A schematic representing the discovery of a poly A tail adjacent to exon 8 that hindered resequencing of fragment 11. Both original primers (green bars) were used to amplify fragment 11 at PCR stage and then the original reverse primer (green LHS) and a new forward primer (orange bar) were used to resequence exon 8 (blue) and intron 8 (red bar to left of exon 8). The remaining 75 base pairs were not resequenced. Numbers represent base pairs according to the contig used to design primers.

2.8.6.6 Comparing resequenced SNPs with NCBI's dbSNP

To identify SNPs, the sequence analysis program SeqScape v.2.1 (Applied Biosystems) requires a region of interest (ROI) as e-reference sequence in order to identify potential SNPs. In this case a BAC was used for the ROI in HSPA8 (AP000926 *Homo sapiens* genomic DNA, chromosome 11 clone:RP11-762B21[gi#31790747] (this BAC was also used to design PCR primers)). The dbSNP database for HSPA8 is based on contig (NT_033899). Therefore by BLASTing the dbSNP contig NT_033899 against the BAC AP000926 I compared the identified 13 SNPs in the 15 samples with the SNPs in dbSNP

2.8.7 Multiplex genotyping of HSPA8 using SNaPshot™

The 13 SNPs identified by the resequencing were genotyped by SNaPshot in four multiplex reactions (see Table 2.8). The details of the SNaPshot method of genotyping are described elsewhere (see section 2.6.9.). Multiplexing is a method of analysing several different SNPs in the one reaction, thereby reducing costs and labour. By designing extension primers of varying lengths, specific SNPs can be identified from each other within one reaction. The goal was to genotype the reference panel (n=92) in order to determine LD structure across the gene. The SNPs were multiplexed together based upon their location within PCR fragments and their specific alleles. Extension primers (Table 2.11) were designed to be 16, 20 or 24 base pairs long in order to distinguish individual SNPs. For example in fragment 13 there were 2 SNPs (HSPA8-1, -2) with A/G and A/T alleles respectively. As both SNPs carried the A allele, the extension primer for HSPA8-1 was designed to be 16 base pairs long and the extension primer for HSPA8-2 was designed to be 20 base pairs long. This difference in product size allowed identification of the respective A alleles and successful genotype calls for both SNPs. Each multiplex reaction was optimised using homozygote DNA samples (identified from resequencing), heterozygote DNA samples (identified from resequencing) and H₂O samples (negative control). PCR was carried out on relevant fragments using the protocols specified in the appendix. Prior to the SAP and ExoI stage (section 2.6.8) PCR products were multiplexed for each of

Table 2.8 Multiplex of SNPs and Fragments

SNP Name (HSPA8-)	Fragment	Optimised Multiplex Reaction
1	13	A
2		
3	10+6+5	B
4		
5		
6		
7	4	C
8		
9		
10	2b+3	D
11		
12		
13		

In order to genotype the 13 SNPs identified from the resequencing more efficiently, PCR fragments containing the relevant SNPs were multiplexed together. Once optimised, four multiplex reaction mixes were used (version A-D).

the four multiplex reactions (details of which are contained in Table 2.9). SNaPshot (section 2.6.9) was carried out on the four multiplex reactions (Table 2.10). SAP2 cleanup step (section 2.6.10) was carried out on each of the multiplex reactions before samples were loaded onto an ABI3100 for genotyping.

During the course of optimisation for each multiplex reaction alterations were made in the amount of PCR product and primer extension primer used in order that all SNPs within each multiplex could be genotyped together. Unfortunately (HSPA8-7) could not be optimised for genotyping in the multiplex reactions (see table 2.10). Therefore, this SNP was genotyped individually in the reference panel using SNaPshot. The remaining optimised multiplex reactions were carried out on the reference panel in order to determine LD structure across HSPA8.

Table 2.9 Reaction mixtures for PCR products multiplexed during the SAP and ExoI stage.

Multiplexing of PCR fragments for SAP + ExoI stage								
13			10+6+5		4		2b+3	
	Version A		Version B		Version C		Version D	
DNA	5	Fragment 13	3	Fragment 10	5	Fragment 4	3	Fragment 2b
DNA	-		3	Fragment 6	-		3	Fragment 3
DNA	-		3	Fragment 5	-		-	
SAP	0.5		1		0.5		1	
SAP dilution buffer	0.5		-		0.5		-	
ExoI	0.05		0.1		0.05		0.1	
Water	1						-	
Total	7.05		10.1		1		7.1	

Table 2.10 SNaPshot multiplex reactions

Optimised SnaPshot multiplex reactions (µl)								
	Version A		Version B		Version C		Version D	
SSHOT	1.5		2		1.5		2	
Buffer	1		0.5		1		0.5	
PCR	5		5		4		5	
PeX	1	PeX 1	4	PeX 3	1.5	PeX 7	1	PeX 9
PeX	0.5	PeX 2	0.5	PeX 4	0.25	PeX 8	1	PeX 10
PeX	-		1	PeX 5	-		4	PeX 11
PeX	-		1	PeX 6	-		1	PeX 12
Water	1		-		1.75		-	
Total	10		14		10		14.5	

Specific details of the SNaPshot multiplex reactions (versions A-D). Note extension primer 7 was subsequently removed from version C and genotyped individually (see text).

Table 2.11 Extension Primers

SNP Name ^a	Polymorphism	Extension Primer Size	Extension Primer
HSPA8- 1	A/G	16	CAAATTCTTCCTTCTC
HSPA8- 2	A/T	20	GCAGTAATTCCTTTTTCTCT
HSPA8- 3	G/T	16	CAGCAGAGACATTGAG
HSPA8- 4	G/A	20	TAAGGAAGAATGGTCGCTCA
HSPA8- 5	T/C	20	GAGCTGAGCCCCATCTGTTC
HSPA8- 6	T/C	24	TTCAAACCTTCAACCTCCTACGTTA
HSPA8- 7	C/A	16	GAGTCAACGGGCTTTT
HSPA8- 8	G/A	20	ACTGGAAGCACGCCAAGAAC
HSPA8- 9	G/T	24	CACGGTGGGACTGGACTAAGCAGG
HSPA8- 10	C/T	24	TTCGTTATTGGAGCCAGGCCTACA
HSPA8- 11	G/A	20	GGCCTGGCTCCAATAACGAA
HSPA8- 12	T/C	16	AGGAAGCCACAAAAAA
HSPA8- 13	G/A	24	CTGTATTTCATAGCGTGGCTTTGG

^a SNP name as listed in table 2.8

2.8.8 Determining LD structure

LD structure for HSPA8 was determined using Haploview version 2.05 (Barrett et al 2005). Detailed methodology on how to use Haploview is contained in section 2.10.

2.8.9 Determining SNPs for genotyping in entire sample

Haploview analysis revealed relatively little strong LD spanning HSPA8 (see Chapter 5 figure 5.3 and 5.4). Using the data generated from resequencing and SNP genotyping in the Irish reference panel, I needed to identify SNPs for genotyping in the full sample. From 13 SNPs originally identified, two turned out to be non-polymorphic in the full sample (highlighting possibly a chance rare variant in the sample panel of 15). In addition, five SNPs genotyped in the Irish reference panel had a MAF <0.1. Therefore the full sample would have had prohibitively low power to detect any association signal from these SNPs. Hence, six SNPs (HSPA8-1, HSPA8-3, HSPA8-4, HSPA8-5, HSPA8-7 and HSPA8-11) were selected based on location within the gene and their MAF's (>0.1) in the Irish

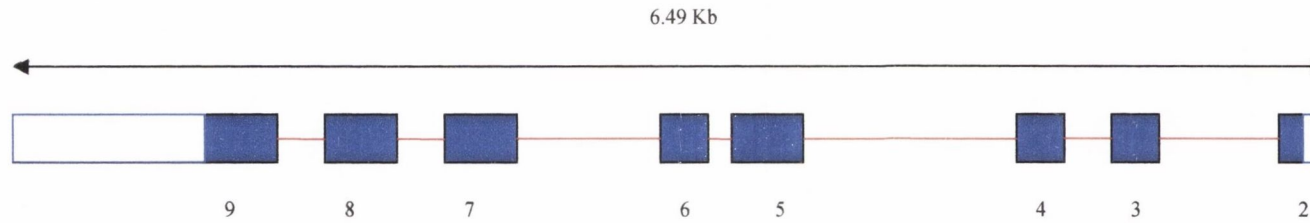
reference panel. This gave an average marker density of 2.2Kb. Those SNPs chosen for analysis in our full case-control sample were genotyped by Kbiosciences (Hoddesdon, UK; <http://www.kbioscience.co.uk>). FASTA files were created for each SNP containing the relevant sequence data required by KBiosciences for genotyping. KBiosciences uses both its own novel form of competitive allele specific PCR system (KASPar) and Taqman™ (Applied Biosystems) chemistries for genotyping.

2.9 Materials and Methods - GSTM3

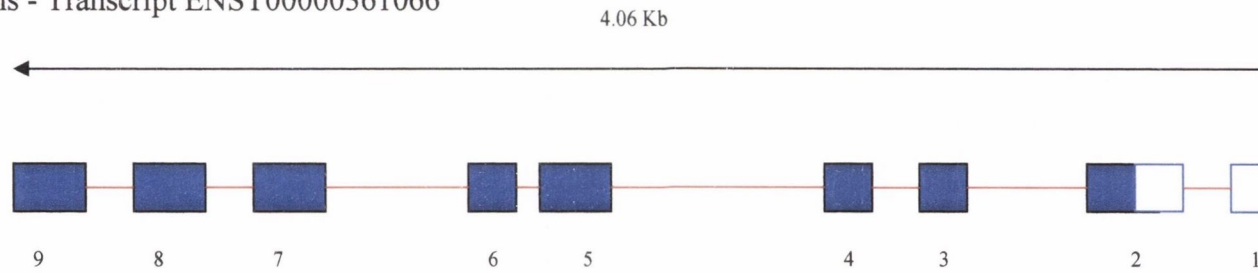
2.9.1 Identification of exonic and intronic structures

In order to perform candidate gene analysis, the intron/exon boundaries of GSTM3 had to be determined. The methodology for this has previously been described for HSPA8 (section 2.8). Regions of GSTM3 were found to contain areas of homology with chromosome 19. PCR primers were not designed in these homologous regions (this is represented schematically in Figure 2.9). The Ensembl website (<http://www.ensembl.org>) identified two major isoforms of GSTM3, one with eight exons (Transcript ENST00000256594) and one with nine exons (Transcript ENST00000361066; see figure 2.8). In order to include all alternatively spliced exons, PCR fragments were designed to include all nine exons.

8 exons - Transcript ENST00000256594



9 exons - Transcript ENST00000361066



Schematic of the two transcripts suggested for GSTM3 according to Ensembl (www.ensembl.org). When designing PCR primers for amplifying fragments consideration was taken to include all nine exons.

Figure 2.8 Schematic of GSTM3 showing two known transcripts one with 8 exons the other with 9 exons

2.9.2 Design of Fragments for DHPLC

Primers were designed using the Primer3 website, described in (section 2.6.2). As previously described in section (2.8.6.3) it is important to have a 20 – 40 base pair overlap between fragments. In total, 14 fragments were designed to capture all nine possible exons, including 1.6 Kb upstream of the gene. A schematic of the PCR fragments designed can be found in Figure 2.10. Due to homologous regions, regions of high GC content and regions contain poly A sequence there are small gaps between certain fragments (see figure 2.9).

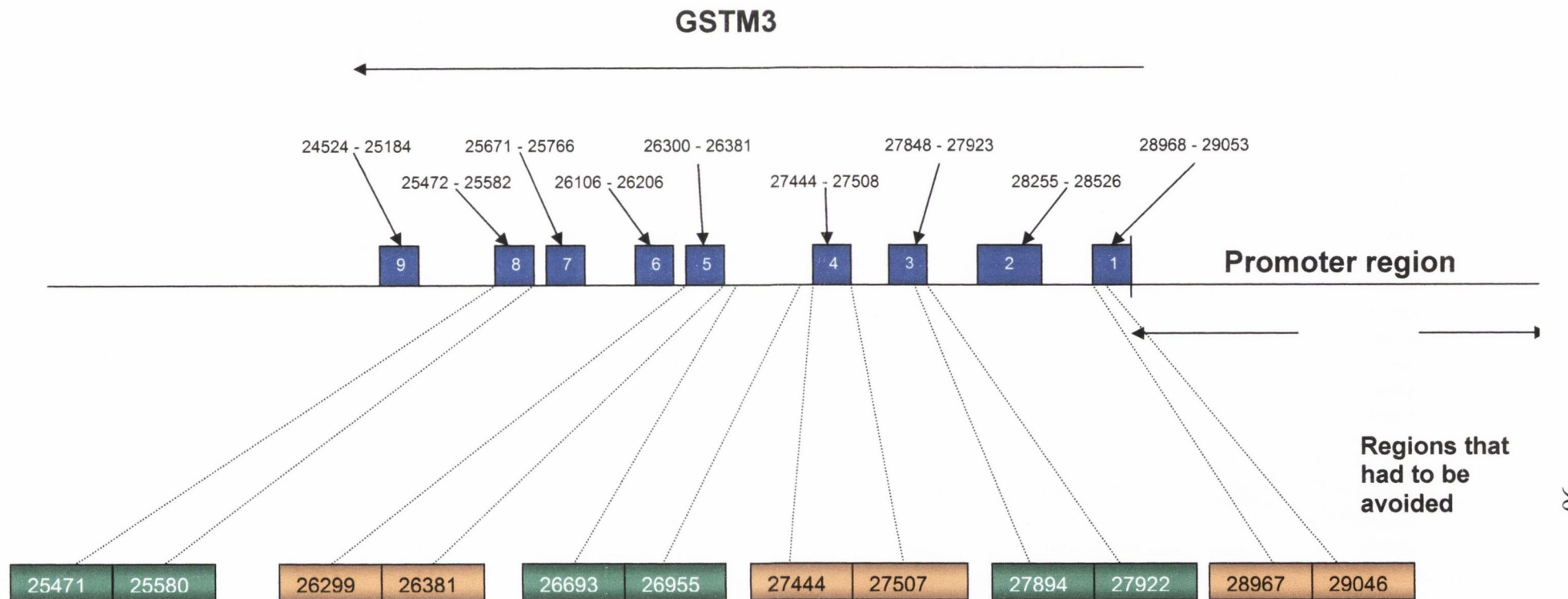
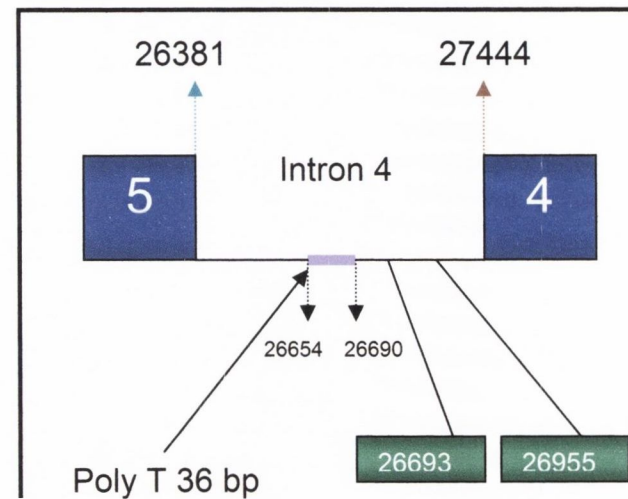


Figure 2.9: Schematic of regions showing homology with chromosome 19. These regions had to be avoided when designing primers because they could anneal to the homologous sequence on a different chromosome and amplify the wrong fragment.



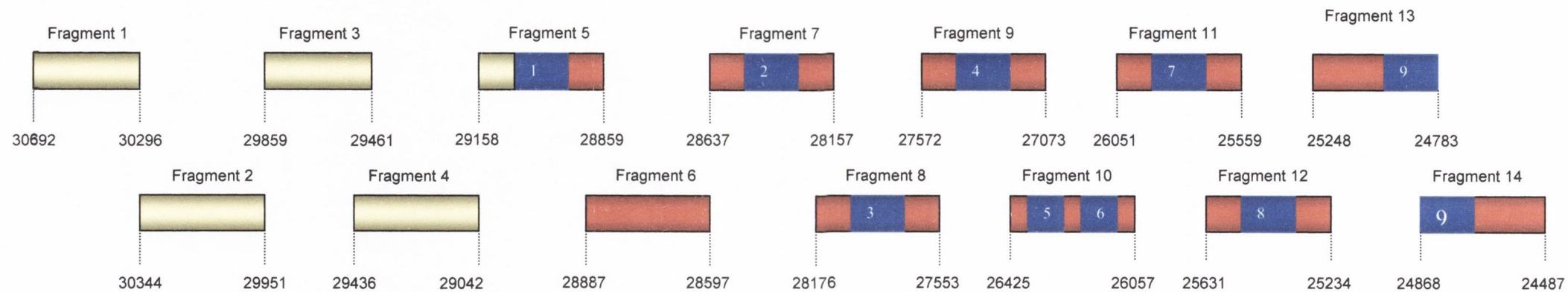


Figure 2.10 Schematic of the GSTM3 fragments designed for DHPLC. Base pair numbers refer to contig gi#15131467

2.9.3 PCR optimisation of fragments for DHPLC

Table 2.12: PCR primer details for the 14 fragments designed for DHPLC analysis

GSTM3 Fragments for DHPLC						
Fragment number	PCR primer sequence	Primer	Start	Stop	Size of fragment	PCR protocol
1	GGTCAACTCGGGAGACATAG	Left	30296	30315	396	57Q
	TGGAATAAATTAGCTAACATACGTAAA	Right	30692	30666		
2	TCTAAGGGTTTGAGGTGCTC	Left	29951	29970	393	57Q
	CTGTCACCTCCTCCACATTCA	Right	30344	30325		
3	GAGAGATAGGGGTCTCAGG	Left	29461	29480	398	57Q
	TTTCCTGGAGTTCTCTCTGC	Right	29859	29840		
4	TCAGAACCAACCTCTATCCTGG	Left	29042	29063	394	57Q
	GGATGTGTATCACCACCTGA	Right	29436	29417		
5	CTCAGCCCGTCACAACTT	Left	28859	28876	299	57Q
	GGGATTGGAGGAGGAAATTA	Right	29158	29139		
6	TTAATGGTGCCAGTGAAGAC	Left	28597	28618	290	57Q
	GCTGGCTTATGAAGTTGTGA	Right	28887	28868		
7	CGTAGAGCTGCTCCTGCT	Left	28157	28176	480	57Q
	GGCTAGTGGTGCTTTTCG	Right	28637	28618		
8	ACCCAACCTCCTTAATACCC	Left	27533	27552	643	57Q
	AGCAGGAGCAGCTCTACG	Right	28176	28157		
9	CTGCAGTGAGCTGTGATTG	Left	27073	27091	499	T61
	TCTTTTGCTCTTGTCAGGTG	Right	27572	27553		
10	CCCCTTTTTCTGATACTCCA	Left	26057	26076	368	57Q
	GAGAGAGGCAGGTTGTTCAT	Right	26425	26405		
11	GGTGAGAAAATCCACAAAGG	Left	25559	25578	492	61Q
	AGGAACCCAAGGGTTATCTC	Right	26051	26032		
12	AGATGCATACCAGAACAGCA	Left	25234	25253	397	57Q
	TGCAGGCTACTACTCCTCAG	Right	25631	25612		
13	TAACACACCTGCTCTCTCCA	Left	24783	24802	465	57Q
	TTCTGGTATGCATCTGTTGG	Right	25248	25229		
14	CGTGGTCACAAAAATTCTTACA	Left	24487	24510	381	57Q
	TTCAAGGGCTGTAGGTTTCAT	Right	24868	24849		

All fragments (except fragments 5 and 7) were optimised using PCR programs 54Q and 57Q (see appendix A7 and A8). Fragments 5 and 7 did not amplify properly using either of these PCR programmes and were subjected to other programmes and temperatures. After further optimisation, fragment 5 amplified correctly using programme T61 (see appendix A7) and fragment 7 amplified correctly using programme 61Q (see appendix A8). Full details of PCR primer sequence and fragment size can be found in Table 2.12.

2.9.4 Materials and Methods for Denaturing High Performance Liquid

Chromatography (DHPLC).

DHPLC was carried out using the same 15 SZ individuals as in the HSPA8 study and therefore gave the same power to detect sequence variants as previously indicated (section 2.8.6.5). DHPLC is a highly sensitive and specific method for mutation detection (O'Donovan et al 1998). The methodology has been described in detail by Underhill et al (1997). In short, the process uses ion-pair reverse-phase high-performance liquid chromatography under partially denaturing conditions to differentially retain double stranded heteroduplex and homoduplex molecules. The PCR products are heated to 95°C for 5 mins and then cooled at a rate of 0.5°C per minute to a temperature of 40°C to allow for the formation of heteroduplexes. The eluted molecules are detected using a UV detector set at 260 nm. DHPLC was performed on a WAVE DNA Fragment Analysis System (Transgenomic Inc, Omaha, NE, USA) containing a DNASep column. Column temperature for DHPLC analysis of each PCR fragment was determined using the DHPLC Melt software (<http://insertion.stanford.edu/melt.html>) (Jones et al 1999). To ensure maximum sensitivity, in addition to the temperature suggested by the software (n°C), each fragment was also run at n+2°C. Empirical studies gave sensitivity for mutation detection greater than 95% (Jones et al 1999). Column temperatures for each fragment are listed in table 2.13 at the end of this section. As some fragments have multiple melt domains, they were run at more than two temperatures.

PCR products were eluted from the column using a linear acetonitrile gradient in a 0.1M triethylamine acetate buffer (TEAA), pH 7 at a constant flow rate of 0.9 ml/min. The gradient was achieved by the mixing of Buffer A (0.1 M TEAA (pH 7), 0.1 mM Na₄EDTA) and Buffer B (25% acetonitrile in 0.1M TEAA (pH 7)). The analytical gradient

for each PCR fragment was determined by the same software that suggested column temperature (n°C) for each fragment. For any fragment the software suggests three concentrations of eluent B (B1, B2 and B3) for use at the recommended column temperature (n°C). Empirically, colleagues in the Cardiff lab (DW Morris, personal communication) found that the initial %B to be used for the gradient is B3 - 2%. The gradient then runs from this %B to (B3 - 2%) + 10% over 5 minutes. For example, if the software suggested a B3 of 67% B, then the gradient used would be 65% to 75% over 5 minutes. The analytical gradients for each fragment (%B) are listed in parenthesis in table 5.12. At the end of each analytical run the column was washed with a solution of 50% acetonitrile in 0.1M TEAA (pH 7). Results were automatically recorded in the form of chromatogram images, by the WAVEMAKER software (Transgenomic Inc, Omaha, NE, USA), which were then visually analysed to determine difference between the traces indicative of the presence of a sequence variation. Individual samples with differing traces for the same amplified region were sequenced using both the forward and reverse primers of the original PCR to identify the sequence variation.

Table 2.13 DHPLC conditions for the 14 fragments spanning GSTM3

Fragment Name	Base Pairs	Temperature 1 (°C) (DHPLC gradient (%B))	Temperature 2 (°C) (DHPLC gradient (%B))	Temperature 3 (°C) (DHPLC gradient (%B))	Temperature 4 (°C) (DHPLC gradient (%B))
1	396	56 (63 – 73)	60 (58 – 68)	62 (57 – 67)	na
2	393	59 (59 – 69)	61 (57 – 67)	na	na
3	398	54 (66 – 76)	60 (60 – 70)	62 (58 – 68)	na
4	394	53 (68 – 78)	58 (61 – 71)	60 (59 – 69)	na
5	280	50 (68 – 78)	55 (63 – 73)	60 (56 – 66)	62 (54 – 64)
6	270	57 (62 – 72)	62 (55 – 65)	64 (53 – 63)	na
7	467	60 (65 – 75)	65 (58 – 68)	67 (56 – 66)	na
8	625	59 (71 – 81)	64 (65 – 75)	66 (63 – 73)	na
9	499	60 (61 – 71)	62 (59 – 69)	na	na
10	348	60 (57 – 67)	62 (53 – 63)	na	na
11	474	59 (61 – 71)	61 (59 – 69)	na	na
12	386	60 (59 – 69)	62 (57 – 67)	na	na
13	442	60 (60 – 70)	62 (58 – 68)	na	na
14	362	54 (65 – 75)	59 (58 – 68)	61 (56 – 66)	na

2.9.5 Resequencing of putative mutations on GSTM3

Where DHPLC analysis of a fragment suggested that a patient was heterozygous, the PCR product of that individual, along with the PCR product of a patient who appeared homozygous, was resequenced to determine the nature of the polymorphism. A maximum of two heterozygotes and two homozygotes were selected for each fragment. The details of the resequencing method have been described in detail elsewhere (section 2.8.6.5). In order to identify a polymorphism, the forward and reverse sequences for the heterozygous and homozygous samples were compared. Ideally, the polymorphism would be seen in both heterozygous sequencing reactions. This was not always the case and in some instances, a polymorphism was only identified in one of the reactions. This may occur for a number of reasons. If a polymorphism is located within 20 base pairs of the sequencing primer, it may be difficult to identify as in some cases the first 30 – 40 base pairs of a sequencing reaction are of poor quality. Secondly, in fragments over 500 base pairs, the quality of the sequencing can begin to deteriorate toward the end of the reaction. So if the polymorphism is located over 500 base pairs downstream it may be hard to detect. Thirdly, the sequencing of a fragment can be disrupted by a run of A's or T's in the sequence. For example, if there is a run of A's 100 bases downstream of the forward primer and a polymorphism exists 150 bases downstream of the forward primer, the sequence reaction may only be readable as far as the run of A's. Hence, the polymorphism may not be identified in the forward reaction. However, the polymorphism should be evident in the reverse reaction, as it will be upstream of the run of T's in the reverse sequence.

On occasion, samples that were thought to be heterozygous for the DHPLC chromatograms were found not to be so when sequenced. In this instance, two heterozygous samples, along with two homozygous samples, were PCR amplified again. If

the subsequent chromatograms again suggested the presence of a polymorphism, the fragments were resequenced again. If the sequence again failed to identify a polymorphism, despite the quality of the sequence being satisfactory, no further investigation was carried out. It should be noted that this only happened in instances where chromatograms for a fragment were only vaguely suggestive of the presence of a polymorphism. The fragment was being resequenced to rule out the possibility of a sequence variant being present. Figure 5.6 (Chapter 5) shows the electropherogram that identified the sequence variant found in fragment 3.

2.9.6 Genotyping of SNPs for GSTM3

All SNPs were sent for commercial genotyping using KBiosciences, UK. The company used a variety of methods for genotyping including Taqman and their own method KASPar. KASPar assays are a proprietary in-house system. They developed this to replace the previously used Amplifluor system. Typically they are able to convert approximately >80% of database SNPs (NCBI) to assays.

2.10 Materials and Methods for NRF2

NRF2 is 34 Kb in length located between base pairs 178331546 – 178297622 on chromosome 2q31.2 (NCBI Build 35, October 2004). Data from the CEPH trios in HapMap indicate that there are blocks of high LD in the region of this gene (see figure 2.11). Across the exonic region of the gene and extending into the 5' putative promoter region there are 11 SNPs that are informative in the CEPH trios. The availability of this small but relatively dense CEPH dataset offered me the opportunity to investigate the

similarities in LD structure between the CEPH sample and our sample from Ireland. To undertake this I genotyped the same sets of SNPs in an Irish reference panel and compared LD measurements with the CEPH data. In addition, investigation of the NRF2 gene in dbSNP indicated that there were a further 5 SNPs with MAF > 10% in at least one European population. These 5 SNPs were also genotyped in my Irish reference panel. I then examined how efficiently tag SNPs, selected on the basis of LD analysis of data from the 11 HapMap SNPs, captured the additional 5 SNPs from dbSNP across the NRF2 locus.

Details of the 16 SNPs examined in this study of NRF2 are given in Table 2.14. These 16 SNPs were genotyped by Ellipsis (Canada) in the Irish reference panel (n = 92 controls) (see section 2.10.1). SNPs selected for analysis in the full case-control sample were genotyped by Kbiosciences (UK) (see section 2.10.2). Data from the 11 HapMap SNPs genotyped in CEPH sample were downloaded from the HapMap website. Inter-marker linkage disequilibrium (LD) was measured using D' and r^2 using Haploview (Barrett et al, 2005). Tag SNPs were chosen from reference panel data using Tagger (de Bakker et al, 2005; <http://www.broad.mit.edu/mpg/tagger/>). Selection of tags is based on r^2 values between alleles of variable sites. Tagger employs both pairwise and effective haplotype predictors to capture alleles of interest. I used an r^2 threshold of 0.8, LOD of 3.0 and the 2- and 3- marker haplotype aggressive tagging option for tag SNP selection.

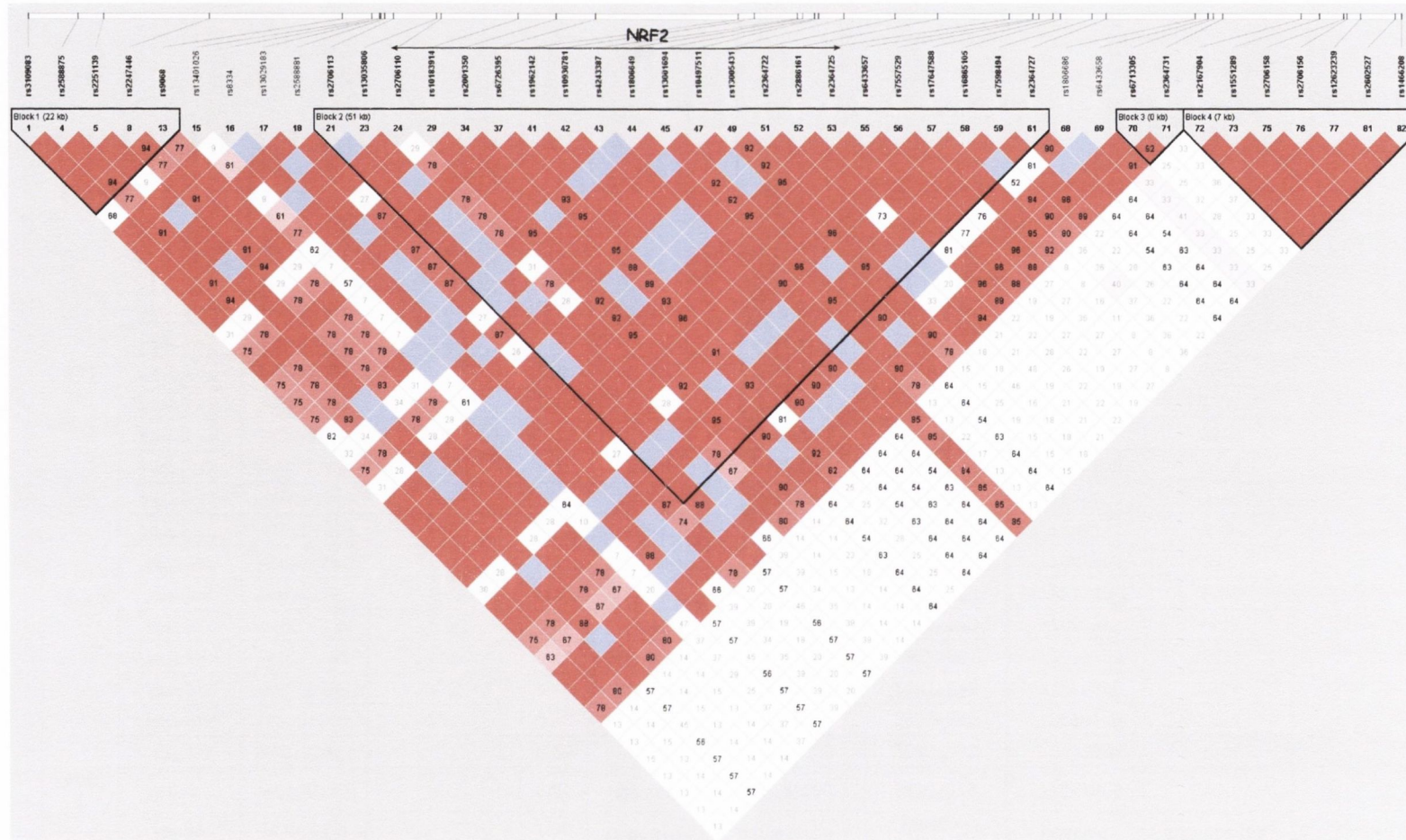


Figure 2.11 HapMap dump showing high LD structure around NRF2 on chromosome 2q31.2. The entire length shown is 100 Kb. Block 2 (defined by the Gabriel Method) spans over half this genomic region (51 Kb).

Table 2.14: Details of HapMap and dbSNP SNPs that were used in SNP selection for this study

	SNP		Polymorphism	MAF	HAPMAP or dbSNP	Chromosome 2 position	dbSNP contig position	Ensembl position	Inter SNP distance (bp)
	Name	rs number							
3' end	NRF2-1	rs2706110	A/G	0.21	HAPMAP	178294706	28301579	177917883	0
NRF2	NRF2-3	rs10183914	C/T	0.25	dbSNP	178300210	28307083	177923387	5504
	NRF2-4	rs2001350	A/G	0.10	HAPMAP	178302969	28309842	177926146	2759
	NRF2-5	rs6726395	A/G	0.47	HAPMAP	178305773	28312646	177928950	2804
	NRF2-6	rs1962142	C/T	0.10	HAPMAP	178316028	28322901	177939205	10255
	NRF2-7	rs10930781	G/A	0.20	dbSNP	178317176	28324049	177940353	1148
	NRF2-8	rs4243387	C/T	0.10	HAPMAP	178320309	28327182	177943486	3133
	NRF2-9	rs1806649	C/T	0.31	HAPMAP	178320696	28327569	177943873	387
	NRF2-10	rs13001694	A/G	0.28	dbSNP	178321534	28328407	177944711	838
	NRF2-11	rs10497511	T/C	0.17	dbSNP	178321840	28328713	177945017	306
	NRF2-12	rs13005431	T/C	0.27	dbSNP	178323656	28330529	177946833	1816
5' end	NRF2-13	rs2364722	A/G	0.29	HAPMAP	178327331	28334204	177950508	3675
	NRF2-15	rs2364725	G/T	0.47	HAPMAP	178335532	28342405	177958709	8201
	NRF2-16	rs2364727	C/T	0.10	HAPMAP	178342431	28349304	177965608	6899
	NRF2-17	rs4465800	C/T	0.10	HAPMAP	178348813	28355686	177971990	6382
	NRF2-18	rs2364731	A/G	0.49	HAPMAP	178350748	28357621	177973925	1935

2.10.1 Irish Reference Panel

95 controls were randomly selected from the entire sample to be used for LD analysis in the Irish population. Aliquots were sent to Ellipsis (Canada) for storage and subsequent genotyping. The company used Taqman assays to perform genotyping.

2.10.2 Individual Genotyping of entire sample

All SNPs for individual genotyping were sent commercially genotyped using KBiosciences, UK. Details of their method are contained in section 2.9.6.

Chapter 3 – Apolipoprotein L (APOL) 1- 6

3.1 Introduction

A recent meta-analysis of SZ linkage studies provided statistically significant evidence that at least 10 chromosomal loci are likely to contain susceptibility genes (Lewis et al, 2003). Investigation at two of these loci has established firm statistical evidence implicating dystrobrevin binding protein 1 (dysbindin) at chromosome 6p (Straub et al, 2002; Schwab et al, 2003; Williams et al, 2004) and neuregulin-1 at chromosome 8p (Stefansson et al, 2002; Stefansson et al, 2003; Williams et al, 2003; Corvin et al, 2004) in SZ susceptibility. There are a number of lines of evidence suggesting that genes on chromosome 22 are also involved in the pathobiology of SZ. Several studies have suggested linkage to chromosome 22q11-13 in SZ (Gill et al, 1996; Blouin et al, 1998; DeLisi et al, 2002) and the recent meta-analysis provides additional statistical support for SZ susceptibility genes on chromosome 22 (Lewis et al, 2003). A recent family-based association study by Takahashi et al (2003), also reported evidence of association between the microsatellite marker D22S683, located on 22q12.3, and SZ in two small samples of European American and Chinese origin respectively. This microsatellite marker is situated 27.4 Kb from the 5' end of the Apolipoprotein-3 gene (APOL3) and 33.85Kb from the 3' end of the Apolipoprotein-5 gene (APOL5). Another interesting phenomenon is that individuals with chromosomal deletions involving this chromosome 22q region (in velo-cardio-facial syndrome (VCFS) (MIM 192430)) have a 30-fold increased risk of SZ and related psychotic disorders (Murphy 2002). The Apolipoprotein L gene family lie some 15Mb apart from the VCFS region.

In this study Dr. Aiden Corvin and I looked at a family of six genes (Apolipoprotein L1-6) that lie in the putative SZ susceptibility locus region 22q11-13 (within bin 22.1 of the meta-analysis). Dr. Corvin investigated APOL 1, 2 and 3 while I investigated APOL 4, 5 and 6. The APOL gene families are functional and positional candidate genes for SZ. The APOL family transport high-density lipoprotein (HDL) in cell membranes and have an important role in the development and maintenance of cell membrane structure and function (Page et al, 2002; Sutcliffe 2003). SZ is increasingly considered a neurodevelopmental disorder resulting from abnormalities in the formation and maintenance of cell membranes (Tkachev et al, 2003). This gene cluster has arisen recently in evolutionary terms that is, it is shared only with other primates, which is compatible with the theory of SZ is a disorder of evolution (Crow 2003, Monajemi et al, 2002). Most intriguingly, gene expression studies in two independent SZ samples data indicate that APOL1, 2 and 4 are up regulated in SZ. Initially, 10 SZ and 10 control post mortem (PM) brains from New Zealand and Japan were screened with a custom made cDNA array of more than 300 genes. The most significant finding was a 2.6-fold up-regulation of APOL1, which was confirmed by real-time PCR and replicated in an independent sample from the Stanley Foundation (which included 15 SZ samples). Further investigation of the APOL gene family using real-time PCR demonstrated >2.4-fold up-regulation of APOL2 in both samples and also up-regulation of APOL4 in the Stanley Foundation sample only (Mimmack et al, 2002).

The use of genetic association studies to identify susceptibility genes for complex disorders such as SZ requires the investigation of large numbers of SNPs in large samples. The task of SNP identification in candidate genes has been greatly aided by the recent explosion in the numbers of SNPs catalogued on public databases on the web. Additional advances in genotyping methodologies have allowed cheaper and faster assaying of these SNPs (for

review see Syvanen 2001). One such methodology is DNA pooling which accurately and reliably measures SNP allele frequencies in pooled DNA samples (Hoogendoorn et al, 2000, Norton et al, 2002). The DNA pooling technique of Norton and colleagues requires the identification of a heterozygous individual to determine accurate SNP allele frequencies and hence test for evidence of association in pooled samples of cases and controls. Where SNPs are taken from public databases, it is a time-consuming undertaking to genotype them in a panel of samples to identify a heterozygous individual. To circumvent this task in this study we have applied a range of simulated heterozygote data to the calculation of allele frequencies of each SNP in the DNA pools.

This modified DNA pooling strategy is more time and cost efficient than existing strategies for analysis of public database SNPs. We have used this new strategy to investigate the APOL gene family in an Irish SZ case-control sample. The six genes (APOL 1-6) covered a region of 617 KB. A total of 187 SNPs across this gene family were identified using the dbSNP database (<http://www.ncbi.nlm.nih.gov/SNP/>), from which we selected a total of 143 markers for analysis. The modified DNA pooling methodology is detailed in full in section 2.6.14.

3.2 Results

Initially, it was important to determine that our DNA pools would provide accurate estimates of allele frequencies. Therefore we randomly selected three SNPs, APOL 1-29, APOL 2-2 and APOL 2-20, from our study and genotyped them individually and in the pools to confirm the accuracy of the pools. The results of the individual and pooled genotyping gave a mean error in estimating the difference between cases and controls of only 2 alleles per thousand for the three SNPs when using pooled genotyping rather than

individual genotyping. These results show that our DNA pools are able to give accurate estimations of absolute allele frequencies as well as accurate estimations of the difference in allele frequency between different pooled DNA samples (Table 3.1).

Table 3.1: Results of validation experiments comparing pooled genotyping with individual genotyping using three SNPs.

	Allele Frequencies in Cases			Allele frequencies in Controls		
	Individual	Pooled	Difference	Individual	Pooled	Difference
Allele A	0.166	0.202	0.037	0.152	0.187	0.035
Allele G	0.834	0.798		0.848	0.813	

Apol 1-29

	Allele Frequencies in Cases			Allele frequencies in Controls		
	Individual	Pooled	Difference	Individual	Pooled	Difference
Allele G	0.828	0.693	0.135	0.843	0.710	0.133
Allele T	0.172	0.307		0.157	0.290	

Apol 2-2

	Allele Frequencies in Cases			Allele frequencies in Controls		
	Individual	Pooled	Difference	Individual	Pooled	Difference
Allele G	0.773	0.758	0.015	0.807	0.794	0.012
Allele T	0.227	0.242		0.193	0.206	

Apol 2-20

In order to validate the pooled allele frequency results with individual allele frequency results, three SNPs were randomly selected and genotyped individually and using DNA pools. The table shows that over the three SNPs only 2 alleles out of 1000 will be incorrectly called.

Fifty-one SNPs were tested for evidence of association in our case-control sample using our 3-stage method of DNA pooling association analysis. Of the 51 SNPs in stage 1, two reached our criteria ($p < 0.1$) for advancement to stage 2: rs132734 (APOL4-49) and rs2017689 (APOL2-6) (see Table 3.2). Following identification of a heterozygous individual for each SNP, accurate estimates of allele frequencies in the case and control pools were calculated for each SNP. Allele frequencies were computed to allele counts and both SNPs were tested for association. However, neither of the 2 SNPs reached significance levels required for advancement to stage 3 ($p < 0.1$) (Table 3.3). It should be noted that these p-values were not corrected for multiple testing.

Table 3.2 – Results of all DNA pooling association results for APOL6 and APOL5

Gene	dbSNP ID	Chromosomal position ^a	Position within gene ^b	Alternative Marker Name	Polymorphism	Allele Ratio in Cases (Pools) ^{c,d}	Frequency in Cases using simulated k value ^{d,e}	Allele Ratio in Controls (Pools) ^{c,d}	Frequency in Controls using simulated k value ^{d,e}	<i>p</i> value ^f
Apol6	dbSNP:926754	34301481	-22805	Apol6-3	G/C	0.843	0.519	0.832	0.497	0.538
	dbSNP:2103768	34304784	-19502	Apol6-4	G/C	0.613	0.442	0.600	0.429	0.429
	dbSNP:762910	34304919	-19367	Apol6-5	T/C	0.585	0.413	0.607	0.436	0.464
	dbSNP:1883986	34305307	-18979	Apol6-6	T/C	0.810	0.516	0.797	0.495	0.541
	dbSNP:1894469	34309889	-14397	Apol6-7	A/G	0.686	0.522	0.683	0.519	0.920
	dbSNP:2413359	34310099	-14187	Apol6-9	C/T	0.682	0.517	0.677	0.512	0.876
	dbSNP:2010168	34322402	-1884	Apol6-12	A/C	0.907	0.661	0.901	0.645	0.590
Apol5	dbSNP:2413367	34373674	-12058	Apol5-2	A/G	0.549	0.549	0.548	0.548	0.992
	dbSNP:1997883	34375522	-10210	Apol5-5	T/C	0.586	0.586	0.573	0.573	0.690
	dbSNP:1970777	34382149	-3583	Apol5-9	G/A	0.729	0.473	0.730	0.474	0.965
	dbSNP:1540297	34382278	-3454	Apol5-10	T/C	0.570	0.570	0.588	0.588	0.582
	dbSNP:2413369	34385614	-118	Apol5-11	T/C	0.395	0.566	0.388	0.559	0.813
	dbSNP:2009168	34389667	3935	Apol5-13	G/A	0.689	0.526	0.698	0.536	0.774
	dbSNP:2009169	34389705	3973	Apol5-14	A/G	0.527	0.527	0.527	0.527	0.965
	dbSNP:2073198	34395126	9394	Apol5-17	A/G	0.573	0.573	0.617	0.617	0.179

^a Position on chromosome 22 determined from April 2003 freeze of Human Genome Assembly using the ensembl browser (<http://www.ensembl.org>)

^b Position of SNP within gene in relation to ATG site (base pairs)

^c Ratio for alleles a and b in pooled sample. Formula = $A/(A+B)$, where A and B represent the peak heights of alleles a and b.

^d Number represents frequency for first allele listed

^e This is the SNP allele frequency estimate based on the simulated k value that produces the maximum potential difference between case and control pools. Formula is $f(a) = A/(A+kB)$. This value would only be a true estimate of the SNP allele frequency if the true heterozygote ratio (k, unknown) were equal the simulated k value.

^f *p* value determined from allele frequencies estimated from simulated k value that produces the maximum potential difference between case and control pools.

Table 3.2 continued – Results of all DNA pooling association results for APOL3 and APOL4

Gene	dbSNP ID	Chromosomal position ^a	Position within gene ^b	Alternative Marker Name	Polymorphism	Allele Ratio in Cases (Pools) ^{c,d}	Frequency in Cases using simulated k value ^{d,e}	Allele Ratio in Controls (Pools) ^{c,d}	Frequency in Controls using simulated k value ^{d,e}	p value ^f
Apol3	dbSNP:2413381	34792045	-16397	APOL3-2	C/T	0.393	0.393	0.406	0.406	0.712
	dbSNP:80576	34811617	3175	APOL3-11	A/G	0.749	0.499	0.744	0.493	0.899
	dbSNP:1807498	34812887	4445	APOL3-17	A/G	0.382	0.553	0.401	0.573	0.566
	dbSNP:2105915	34824152	15710	APOL3-23	C/T	0.554	0.554	0.542	0.542	0.730
	dbSNP:132653	34828636	20194	APOL3-26	G/T	0.145	0.459	0.139	0.446	0.694
Apol4	dbSNP:2227169	34859765	-9948	APOL4-8	C/T	0.713	0.554	0.689	0.526	0.267
	dbSNP:2097466	34859925	-9788	APOL4-9	G/A	0.592	0.592	0.573	0.573	0.543
	dbSNP:132704	34863087	-6626	APOL4-17	T/C	0.912	0.674	0.898	0.638	0.268
	dbSNP:2007468	34863193	-6520	APOL4-18	G/A	0.457	0.457	0.405	0.405	0.115
	dbSNP:132712	34867057	-2656	APOL4-27	G/A	0.133	0.434	0.146	0.461	0.409
	dbSNP:132719	34868829	-884	APOL4-33	G/T	0.340	0.507	0.314	0.478	0.391
	dbSNP:132720	34868911	-802	APOL4-34	G/A	0.298	0.563	0.265	0.522	0.226
	dbSNP:132721	34868954	-759	APOL4-35	G/A	0.805	0.508	0.819	0.531	0.480
	dbSNP:132734	34869690	-23	APOL4-49	G/A	0.720	0.462	0.792	0.559	0.004
	dbSNP:1807673	34870580	867	APOL4-55	A/G	0.202	0.502	0.175	0.459	0.191

^a Position on chromosome 22 determined from April 2003 freeze of Human Genome Assembly using the ensembl browser (<http://www.ensembl.org>)

^b Position of SNP within gene in relation to ATG site (base pairs)

^c Ratio for alleles a and b in pooled sample. Formula = $A/(A+B)$, where A and B represent the peak heights of alleles a and b.

^d Number represents frequency for first allele listed

^e This is the SNP allele frequency estimate based on the simulated k value that produces the maximum potential difference between case and control pools. Formula is $f(a) = A/(A+kB)$. This value would only be a true estimate of the SNP allele frequency if the true heterozygote ratio (k, unknown) were equal the simulated k value.

^f p value determined from allele frequencies estimated from simulated k value that produces the maximum potential difference between case and control pools.

Table 3.2 continued – Results of all DNA pooling association results for APOL2 and APOL1

Gene	dbSNP ID	Chromosomal position ^a	Position within gene ^b	Alternative Marker Name	Polymorphism	Allele Ratio in Cases (Pools) ^{c,d}	Frequency in Cases using simulated k value ^{d,e}	Allele Ratio in Controls (Pools) ^{c,d}	Frequency in Controls using simulated k value ^{d,e}	p value ^f
Apol2	dbSNP:1315	34887948	-19855	APOL2-1	A/C	0.788	0.482	0.798	0.497	0.628
	dbSNP:1317	34887976	-19827	APOL2-2	G/T	0.705	0.544	0.692	0.529	0.645
	dbSNP:763086	34890870	-16933	APOL2-4	A/G	0.794	0.491	0.806	0.510	0.548
	dbSNP:132757	34893522	-14281	APOL2-5	T/C	0.655	0.486	0.659	0.491	0.879
	dbSNP:2017689	34894533	-13270	APOL2-6	A/G	0.764	0.447	0.802	0.503	0.099
	dbSNP:129607	34894719	-13084	APOL2-7	T/C	0.661	0.494	0.696	0.534	0.236
	dbSNP:132759	34894961	-12842	APOL2-8	T/C	0.909	0.666	0.922	0.703	0.232
	dbSNP:2005998	34897317	-10486	APOL2-10	C/T	0.351	0.520	0.391	0.562	0.226
	dbSNP:2010499	34901279	-6524	APOL2-15	A/T	0.346	0.514	0.391	0.562	0.157
	dbSNP:2157249	34902852	-4951	APOL2-18	T/C	0.858	0.548	0.830	0.494	0.101
	dbSNP:2157251	34905240	-2563	APOL2-20	G/T	0.758	0.440	0.794	0.491	0.126
	dbSNP:1557535	34905265	-2538	APOL2-21	A/C	0.304	0.570	0.261	0.517	0.122
	dbSNP:136144	34907257	-546	APOL2-23	A/G	0.779	0.540	0.770	0.527	0.698
	dbSNP:713791	34914183	6380	APOL1-1	A/G	0.575	0.575	0.562	0.562	0.653
Apol1	dbSNP:136147	34924702	1887	APOL1-3	G/T	0.418	0.589	0.396	0.568	0.503
	dbSNP:136148	34924744	1929	APOL1-4	C/T	0.907	0.661	0.906	0.657	0.898
	dbSNP:136151	34926825	4010	APOL1-8	A/G	0.715	0.455	0.716	0.456	0.998
	dbSNP:713929	34929409	6594	APOL1-18	A/G	0.806	0.509	0.823	0.538	0.404
	dbSNP:713753	34930347	7532	APOL1-21	C/T	0.711	0.450	0.741	0.488	0.234
	dbSNP:136171	34932874	10059	APOL1-29	A/G	0.750	0.501	0.770	0.528	0.397
	dbSNP:136174	34933349	10534	APOL1-33	A/C	0.807	0.491	0.828	0.456	0.303

^a Position on chromosome 22 determined from April 2003 freeze of Human Genome Assembly using the ensembl browser (<http://www.ensembl.org>)

^b Position of SNP within gene in relation to ATG site (base pairs)

^c Ratio for alleles a and b in pooled sample. Formula = A/(A+B), where A and B represent the peak heights of alleles a and b.

^d Number represents frequency for first allele listed

^e This is the SNP allele frequency estimate based on the simulated k value that produces the maximum potential difference between case and

control pools. Formula is $f(a) = A/(A+kB)$. This value would only be a true estimate of the SNP allele frequency if the true heterozygote ratio (k, unknown) were equal the simulated k value.

^f p value determined from allele frequencies estimated from simulated k value that produces the maximum potential difference between case and control pools.

Results of all 51 SNPs genotyped in the Stage 1 analysis. Two SNPs reached $p < 0.1$ for advancement to stage 2 of the study (these p-values had not been corrected for multiple testing).

Table 3.3 – SNPs that were subjected to stage 2 analysis

dbSNP ID	Gene	Alternative Marker Name	Simulated k value	HZ ratio value k (true)	Frequency in Cases (Pools) ^{a,b}	Frequency in Controls (Pools) ^a	True p value
rs132734	4	APOL4-49	3	0.24	0.720	0.792	0.128
rs2017689	2	APOL2-6	4	6.17	0.764	0.802	0.109

^a Ratio for alleles a and b in pooled sample. Formula = $A/(A+B)$, where A and B represent the peak heights of alleles a and b.

^b Number represents frequency for first allele listed

Corrected p value of two SNPs identified as possibly showing significance in our pooled sample using the correction factor k .

3.3 Discussion

We have performed an extensive genetic association study of the six known APOL genes and SZ using a modified method of DNA pooling. The APOL genes are both positional and functional candidate genes for SZ. This gene family maps to chromosome 22q12.3, a region implicated by SZ linkage studies as likely to contain one or more SZ susceptibility genes. A recent gene expression study demonstrated up-regulation of APOL1, 2 and 4 in post-mortem brain samples from SZ patients compared to controls in two independent samples. To test for genetic association we investigated 143 SNPs from dbSNP that are located within these genes. Of these, 51 (36%) were polymorphic in our Irish sample and were genotyped using a three-stage DNA pooling approach. We found no evidence to support the hypothesis that genetic variation at these genes contributes to SZ susceptibility in our sample.

In completing this study we developed a three-stage DNA pooling strategy that increases efficiency for studies that employ public database SNPs in large-scale association studies. This strategy represents a modification of the DNA pooling method proposed by Norton et al (2002). That method required the identification of an individual heterozygous for the

SNP under analysis in order to accurately test for association at that SNP. By simulating the information likely to be provided by the heterozygous individual, our strategy circumvents this requirement, resulting in a more economical but no less accurate method of association analysis. In this study 49 of 51 SNPs assayed could be categorized as not showing evidence for association without the need to identify a heterozygous sample. Although this increased efficiency of the study, we could not determine LD structure due to our methodology.

There are several possible interpretations of these findings. Firstly, it may well be that up regulation of APOL genes is of relevance to the pathobiology of SZ, but that this is influenced by genetic variation at cis or trans acting elements or transcription factors on other chromosomes. Secondly, APOL expression abnormalities in post-mortem SZ samples may not be directly related to the genetic aetiology of SZ. For example, expression differences may represent secondary effects caused by differences in factors such as drug treatment, substance abuse or smoking behaviour between case and control groups (Harrison and Weinberger 2005). It may be that the gene expression study (Mimmack et al 2002) findings were due to chance (type I error). Or, there could be genetic heterogeneity between the Irish sample and the gene expression samples. Thirdly, genetic variation in the APOL gene family is of aetiological importance in SZ but because the effect size is unknown, it is possible that our study lacked statistical power to detect association. Power calculations indicate that for the markers tested, our sample has 80% power to detect an effect of odds ratio (OR) >2 with 95% confidence. However, as for other susceptibility genes identified for SZ the effect sizes are smaller (OR < 1.5) such as 5-HT_{2a} (OR = 1.2, C allele) (Williams et al 1997); DRD3 (OR = 1.3, Ser9Gly polymorphism) (Williams et al 1998); Neuregulin-1 (NRG1) (OR = 1.25, core haplotype) (Williams et al 2003). This is consistent with the work by Risch (1990) who calculated that the recurrence risk in the

relatives of probands with SZ are incompatible with the existence of a single locus conferring a relative risk (λ_s) in siblings of >3 and unless extreme epistasis exists, models of two or three loci of $\lambda_s \leq 2$ are more plausible.

We did not test all genetic variation at this gene family for association as I used public database SNPs from NCBI's dbSNP (freeze May 26th 2001 and Dec 11th 2001 gi#: 13449280). Since then there has been one new version of the draft (gi#: 22035660) with 13 new revisions.

In addition, it should be noted that during these experiments no multiple testing had been accounted for. This was partly because my studies were in their infancy and the field of psychiatric genetics had not yet insisted on using this form of correction because no single gene had yet been identified as having an association with SZ and discussions at the World Congress of Psychiatric Genetics as to how best to identify association in the light of low prior probability were continuing. It should be noted if any SNP in this study had been found at stage 3 to have a p value < 0.05 , it would never have survived any form of correction (e.g Bonferroni would have required a p value $< 9.43 \times 10^{-4}$).

Although chromosome 22 has been extensively genotyped to determine LD structure (Dawson et al 2003) there are areas of long range high LD together with regions of little or no LD. The markers involved in our study did not match any of the markers used in the Dawson et al (2003) study. Therefore not all the genetic variation on chromosome 22 has been captured by the Dawson et al (2003) study. An increase in marker density is required to fully understand the majority of genetic variation and LD in this region. At present the HapMap project aims to genotype 1 SNP per 1000 bases. In addition the ENCODE project aims to genotype every possible SNP in selected 500 Kb regions. Unfortunately,

chromosome 22 is not an ENCODE region. In addition, the HapMap project at the time of this study lacked sufficient marker density in order to determine LD structure. Given the positional and functional evidence, the APOL gene family warrants further genetic investigation. This will require a better understanding of the haplotypic structure across this locus and larger sample sizes to definitively exclude smaller genetic effects. A second study should be commissioned now that HapMap Phase II data is available. This should increase the power and coverage of the study while greatly reducing the number of SNPs genotyped.

Chapter 4 – Positive Replication of G72/DAAO

4.1 Introduction

Linkage analysis data from many SZ samples, sampled from different ethnic populations, suggests that a 68Mb region of chromosome 13q is involved in SZ susceptibility (Lin et al 1995; Lin et al 1997; Blouin et al 1998; Shaw et al 1998; Brzustowicz et al 1999; Camp et al 2001; Cardno et al 2001; Faraone et al 2002; Wijsman et al 2003 and Abecasis et al 2004). The first *suggestive* report of linkage to chromosome 13q32 was by Lin et al (1995) using 18 markers and 13 families from Europe and Japan. They found a multipoint LOD score of 2 under a narrow diagnostic model of SZ. In 1997, the same group repeated the scan using the same markers but a different sample. Under the same narrow diagnostic model, they found suggestive linkage with Caucasian but not Oriental samples. A subsequent study by Shaw et al (1998) used a sample of 70 pedigrees consisting of 42 families from the USA, 14 from the UK and Ireland, 12 from Northern Italy, and 2 from Belgium. They produced suggestive evidence for linkage at 13q12-q13 (LOD=1.8, $p=0.01$). In addition a study published almost ten years later by Abecasis et al (2004) showed suggestive linkage of (LOD = 1.88, $p=0.02$) in a study consisted of 143 small families of Afrikaner origin from South Africa.

The study by Blouin et al (1998) was the first to obtain a LOD > 3 in the 13q32 region. Their sample consisted of 54 families. They found a LOD score = 4.18 ($p = 0.00002$). In addition, a second sample of 51 families obtained a LOD score = 2.36 ($p = 0.007$) ~1.5 Mb from the original marker. A different study by Cardno et al (2001) study consisted of 164 SZ patients from the U.K. and Republic of Ireland. None of the linkage results achieved

genome-wide statistical significance, but the peak LOD score showed suggestive linkage (LOD = 1.68) that, although suggestive, coincided with the region showing maximum evidence for linkage in the study by Blouin et al (1998).

Around the same time as the other studies mentioned, a study by Brzustowicz et al (1999) studied 21 Canadian families of Celtic or German descent and obtained a LOD score of 3.92. A later study by Faraone et al (2002) showed suggestive linkage (LOD = 1.43) within 3 cM of the Brzustowicz et al (1999) report but within the same region as the Blouin et al (1998) report. His sample consisted of 166 families; the majority being Northern European-American families and a similar proportion were African-American kindred.

Lastly, the study by Camp et al (2001) used Palauans from Micronesia and found evidence of linkage (LOD = 3.6 (under broad diagnostic classification) and LOD = 3.5 (under spectrum diagnostic classification)). Interestingly, after correction for multiple testing their results still had genome-wide significance.

Then, Badner and Gershon (2002) presented a technique of linkage meta-analysis using published genome scans. They found the strongest evidence for susceptibility loci on 13q ($p < 6 \times 10^{-6}$) for Bipolar disorder and 13q ($p < 7 \times 10^{-5}$) for SZ. However, a more recent meta-analysis using a different methodology, with more data from 20 completed genome scans did not find any evidence for linkage in the 13q region (Lewis et al 2004).

In a fine-mapping study based on positional evidence of a 5Mb sub-section of chromosomal region 13q33, Chumakov et al (2002) identified association with two novel genes, G72 (MIM 607408) and G30 (MIM 607415) in French-Canadian and Russian case-control samples. Initially, they identified SNPs from pools of 100 French individuals

through re-sequencing 500 bp amplicons covering the 5Mb region. 191 SNPs for chromosome 13, 8 SNPs for DAAO (located on chromosome 12q24) and 27 'genome-wide' SNPs (27 SNPs randomly selected throughout the genome for genomic control purposes) were identified and subsequently registered in dbSNP. The two samples consisted of 213 cases and 241 controls of French-Canadian background matching DSM-IV criteria. The second sample consisted of 183 cases (DSM-III R criteria) and 183 controls of Russian background. Genomic control was carried out using 27 SNPs on random chromosomes.

Two regions, one 1400Kb and one 65Kb region showed association with disease. Six markers in the 65 Kb regions of the French Canadian sample were associated in allelic tests (M12, M14, M15, M22, M23 and M24) and four of the six also showed genotypic association with SZ (M12, M22, M23 and M24). In the Russian sample two alleles (M23 and M24) were also associated in allelic tests. Subsequently gene discovery techniques (see Chumakov et al 2002) identified G72 and G30 overlapping each other on opposite strands in this 65Kb region.

G72 is ~ 25Kb in length with 5 exons (UCSC Genome Browser, May 2004; <http://www.ensembl.org>) and encodes a 742 bp mRNA with up to five alternative transcripts (Chumakov et al 2002). G30 is transcribed in the opposite direction and it was suggested (by Chumakov and colleagues) that both transcripts may regulate each other. However, functional analysis *in vitro* (Chumakov et al 2002) has suggested that only the G72 gene is actively translated. In addition, G72 is only expressed in the amygdale, caudate nucleus, spinal cord and testis of primates (Chumakov et al 2002) although the exons have been found in the dog genome but lack intronic sequences (Detera-Wadleigh and McMahon, 2006).

The next stage was to identify what both of these genes transcribed. However, it appeared that only G72 gave a protein product. In order to identify the function of this novel gene, Yeast 2 hybrid experiments were carried out *in vitro* and identified that G72 interacted with DAAO (OMIM 124050). This was a very interesting finding as it was known that DAAO oxidizes D-serine which activates NMDA-type glutamate receptors (Mothet et al 2000). Subsequently they found that increasing G72 levels in the presence of DAAO *in vitro* increased oxidation of D-serine. They concluded that G72 was an activator of DAAO. As such G72 has now been renamed D-amino acid oxidase activator (DAOA) (Craddock et al 2006). However, for the purposes of this study, I will continue to use the term G72.

An association study was carried out on DAAO using eight SNPs in the French Canadian sample. Four of these SNPs were associated with SZ (MDAAO-4, MDAAO-5, MDAAO-6 and MDAAO-7). Finally as *in vitro* studies had shown that G72 regulated DAAO, they investigated for evidence of statistical interaction between markers at the gene, using logistic regression modeling on G72-M22 (OR = 1.89) and MDAAO-6 (OR = 1.04) (each marker showed highest significance in each gene). This gave a significant finding of $p = 0.04$, OR = 5.02.

DAAO is located on chromosome 12q24. It spans 42.11 Kb, contains 17 exons with 4 known alternative transcripts (NCBI 35, August 2004). (ACEVIEW, <http://www.ncbi.nlm.nih.gov/IEB/Research/Acembly/av.cg>). This locus has previously been implicated in linkage studies in bipolar disorder (for a review of these studies, the reader is referred to Sklar, 2002). This is interesting because the pattern of findings emerging from recent genetic susceptibility studies is showing increased evidence of an

overlap between Bipolar and Schizophrenia phenotypes (Craddock et al 2005; Craddock et al 2005), challenging the traditional Kraepelinian dichotomy still in DSM-IV classification today. This issue is discussed further at the end of this chapter. DAAO is a principal flavoenzyme of peroxisomes, and is expressed in the brain, kidney and liver of many mammalian species. It acts on a wide range of D-amino acids however in the human brain it oxidizes D-serine, a potent activator of N-methyl-D-aspartate-type glutamate receptor.

Genetic association and functional data provide some support for this G72/G30-DAAO hypothesis. Independent studies have reported evidence for association at both the G72/G30 (Schumacher et al, 2004; Wang et al, 2004; Hall et al, 2004; Korostishevsky et al, 2004; Zou et al, 2005; Korostishevsky et al, 2005; Addington et al, 2004) and DAAO genes (Yamada et al, 2004; Liu et al, 2004; Schumacher et al, 2004). However, negative studies have also been reported with G72/G30 (Mulle et al, 2005). In addition, the positive studies have differed in the risk variants (see tables 4.8 and 4.9 for comparison of published case-control studies) and haplotypes identified (see Detera-Wadleigh and McMahon, 2006).

Another complicating factor is that genetic association has been reported with specific psychotic symptoms (persecutory delusions) and other psychiatric disorders (bipolar disorder, panic disorder) (Schultze et al, 2005; Hattori et al, 2003; Schumacher et al, 2004). This suggests that the proposed mechanism may contribute to specific aspects of symptomatology or a broader phenotype of psychiatric morbidity.

The few reported functional studies indicate reduced serum D-serine levels *in vivo* in patients and a decrease in the ratio of D-serine/DAAO expression and increased G72

expression in post-mortem schizophrenia brain samples (Hashimoto et al 2004; Toro et al, 2004; Korostishevsky et al, 2004). However, interaction between G72/G30 and DAAO has not been established *in vivo* and association studies in two large Caucasian samples have failed to replicate the epistatic interaction between these genes described in the original report (Schumacher et al, 2004; Williams et al, personal communication).

Association studies of the DAAO and G72 genes conducted to date have investigated different genetic markers in different ethnic populations. The primary aim of this study was to investigate markers that have demonstrated association in previous studies of Caucasian populations in a suitably powered independent case-control sample. A subsequent later aim was to determine how much variation was captured in this study by using Phase II HapMap data to answer this question.

4.2 Results

There are three sets of results presented here. First, the association study of the five single markers genotyped in G72 followed by G72 haplotype analysis (section 4.2.1 and 4.2.2). Second, the association study of the four single markers genotyped in DAAO followed by DAAO haplotype analysis (section 4.2.3 and 4.2.4). Lastly, some time after this replication study was carried out; Phase II data from HapMap became available. Therefore I present retrospective analysis of how much variation was captured with the markers used this study (section 4.2.5).

4.2.1 Association analyses of G72/G30

Five G72/G30 SNPs were genotyped in the case-control sample: rs3916965 (G72-M12); rs2391191 (G72-M15); rs778293 (G72-M22); rs3918342 (G72-M23) and rs1421292 (G72-M24) (see Table 4.1). Application of SNPSPD indicated the effective number of independent SNPs at this locus was 3.81 and a significance threshold of $p=0.013$ was required to keep Type I error at $\alpha = 0.05$. Evidence for single marker association was identified at two SNPs: G72-M12 ($\chi^2=7.87$, $p=0.005$, OR 1.34(95% CI=1.09 – 1.65) and G72-M15 ($\chi^2=6.40$, $p=0.011$, OR 1.31 (95% CI= 1.06-1.61)).

4.2.2 Haplotype Analyses of G72/G30

Haplotype analyses indicated high LD between markers G72- M12/G72-M15 and between G72-M23/G72-M24 (see Table 4.2). Based on the r^2 value ($r^2=0.99$) between G72-M12 and G72-M15, including both markers in haplotype analysis provided no additional information, therefore we included G72-M12 only. Haplotype analysis identified three significant two-marker haplotypes: G72-M12/G72-M22 (sim $p=0.0245$); G72-M12/G72-M23 (sim $p=0.0114$) and G72-M12/G72-M24 (sim $p=0.0101$) (see Table 4.3). Analysis using GENECOUNTING indicated that each of these haplotypic associations resulted from excess haplotypes in the case sample containing the C allele of G72-M12.

Table 4.1 Results of single marker association analyses for G72

SNP marker ^a	SNP rs #	Polymorphism ^b	Distance ^c (kb)	Allele Frequency		p value	Odds Ratio (95% CI)
				Cases ^d	Controls		
G72-M12	rs3916965	C/T	0.000	0.689	0.622	0.005	1.34 (1.09, 1.65)
G72-M15	rs2391191	A/G	16.086	0.317	0.378	0.011	1.31 (1.06, 1.61)
G72-M22	rs778293	A/G	65.839	0.592	0.625	0.137	1.15 (0.96, 1.38)
G72-M23	rs3918342	C/T	82.389	0.490	0.457	0.186	1.14 (0.94, 1.39)
G72-M24	rs1421292	A/T	94.875	0.473	0.506	0.181	1.14 (0.94, 1.39)

^a Name of SNP marker used in the Chumakov et al paper (2002)

^b Alleles called with respect to Taqman assay (Alleles in brackets relate to Chumakov et al 2002 paper)

^c Distance of SNPs in relation to each other using the format of Korostishevsky et al (2004)

^d Frequency shown relates to the first allele listed in the polymorphism column

Table 4.2 D' and r² values for the five markers genotyped in G72

	G72-M12	G72-M15	G72-M22	G72-M23	G72-M24	
G72-M12		0.993	0	0	0	r ²
G72-M15	0.997		0	0	0.001	
G72-M22	0.01	0.03		0.491	0.611	
G72-M23	0.011	0.027	0.831		0.866	
G72-M24	0.025	0.038	0.995	1		
D'						

Table 4.3 Haplotype analyses of G72 for 2 marker, 3 marker, 4 marker and 5 marker combinations

Marker 1	Marker 2	Marker 3	Marker 4	p value ^a
G72-12	G72-22			0.0245
G72-12	G72-23			0.0114
G72-12	G72-24			0.0101
G72-22	G72-23			0.6578
G72-22	G72-24			0.5082
G72-23	G72-24			0.5969
G72-12	G72-22	G72-23		0.1442
G72-12	G72-22	G72-24		0.1261
G72-12	G72-23	G72-24		0.0985
G72-22	G72-23	G72-24		0.8741
G72-12	G72-22	G72-23	G72-24	0.6216

^a Simulated p value determined empirically from simulations with 10,000 iterations. Significant p values highlighted in bold.

4.2.3. Association Analyses of DAAO

Four markers were genotyped at the DAAO locus: DAAO-M4 (rs2111902), DAAO-M5 (rs3918346), DAAO-M6 (rs3741775) and DAAO-M9 (rs888531) (see Table 4.4). Application of SNPSpD indicated the effective number of independent SNPs at this locus was 3.5 and a significance threshold of $p=0.014$ was required to keep Type I error at $\alpha=0.05$. Significant evidence for single marker association was identified at DAAO-M5 ($\chi^2=8.55$, p value = 0.0034, OR = 1.43 (1.12 – 1.84) and marginal evidence for association at DAAO-M4 ($\chi^2=5.64$, p value = 0.018, OR = 1.29 (1.04-1.61).

4.2.4 Haplotype Analysis of DAAO

LD analysis using Haploview revealed complete LD between markers DAAO-M5 and DAAO-M6 ($D'=1$). However, r^2 analysis revealed a low correlation between markers DAAO-M5 and DAAO-M6 ($r^2=0.21$) (see Table 4.5). Therefore, all four markers were included in haplotype analysis. Using FASTEPLUS I identified five 2-marker haplotypes: DAAO-M4/DAAO-M5 (sim p value = 0.0313); DAAO-M4/DAAO-M6 (sim p value = 0.0199); DAAO-M4/DAAO-M9 (sim p value = 0.0168); DAAO-M5/DAAO-M6 (sim p value = 0.0313); DAAO-M5/DAAO-M9 (sim p value = 0.0031) and three 3-marker haplotypes DAAO-M4/DAAO-M5/DAAO-M9 (sim p value 0.0313); DAAO-M4/DAAO-M6/DAAO-M9 (sim p value = 0.0199); DAAO-M5/DAAO-M6/DAAO-M9 (sim p value = 0.0313) (see Table 4.6). Analysis using GENECOUNTING indicated that each of these haplotypic associations resulted from excess haplotypes in the case sample containing the G allele of DAAO-M4 and the C allele of DAAO-M5.

Table 4.4 Results of single marker association analyses of DAAO

SNP marker ^a	SNP rs #	Polymorphism ^b	Allele Frequency		p value	Odds Ratio (95% CI)
			Cases ^d	Controls		
DAAO-M4	rs2111902	G/T	0.266	0.320	0.018	1.30 (1.05, 1.61)
DAAO-M5	rs3918346	C/T	0.812	0.751	0.003	1.43 (1.12, 1.82)
DAAO-M6	rs3741775	T/G	0.517	0.554	0.136	1.16 (0.95, 1.41)
DAAO-M9	rs888531	T/C	0.708	0.738	0.185	1.16 (0.93, 1.45)

^a Name of SNP marker used in the Chumakov et al paper (2002)

^b Alleles called with respect to Taqman assay (Alleles in brackets relate to Chumakov et al 2002 paper)

^c Distance of SNPs in relation to each other using the format of Korostishevsky et al (2004)

^d Frequency shown relates to the first allele listed in the polymorphism column

Table 4.5 D' and r² values for the four markers genotyped in DAAO

	DAAO-M4	DAAO-M5	DAAO-M6	DAAO-M9	
DAAO-M4		0.6	0.04	<0.01	r ²
DAAO-M5	0.97		0.21	<0.01	
DAAO-M6	0.34	1		<0.01	
DAAO-M9	0.15	0.14	0.02		
D'					

Table 4.6 Haplotype analyses of DAAO for 2 marker, 3 marker and 4 marker combinations

Marker 1	Marker 2	Marker 3	Marker 4	p value ^a
DAAO-M4	DAAO-M5			0.0313
DAAO-M4	DAAO-M6			0.0199
DAAO-M4	DAAO-M9			0.0168
DAAO-M5	DAAO-M6			0.0313
DAAO-M5	DAAO-M9			0.0031
DAAO-M6	DAAO-M9			0.1365
DAAO-M4	DAAO-M5	DAAO-M6		0.1448
DAAO-M4	DAAO-M5	DAAO-M9		0.0313
DAAO-M4	DAAO-M6	DAAO-M9		0.0199
DAAO-M5	DAAO-M6	DAAO-M9		0.0313
DAAO-M4	DAAO-M5	DAAO-M6	DAAO-M9	0.1448

^a Simulated p value determined empirically from simulations with 10,000 iterations. Significant p values highlighted in bold.

4.2.5 Retrospectively assessing the amount of variation captured in the Irish study

The HapMap Project was in its infancy when the genes G72 and DAAO were investigated in the Irish population. Many studies had been performed in different populations and using different markers. Although many studies reported significant association, it was unclear as to the amount of variation (ie SNP information across the genes) that had been captured from the SNPs selected for this study. Therefore when the HapMap phase II data became available some time after this association study I was able to address this question.

4.2.5.1 G72

To determine how much common variation ($MAF > 0.05$, HapMap) was captured in this study, I used the computer programme Tagger implemented in Haploview (de Bakker, 2005) and the current Phase II data (Oct 05) from HapMap (NCBI b34, dbSNP b134). Parameters were set to pairwise tagging only, r^2 threshold of 0.8 and $LOD = 3$. I force included markers G72-M15, G72-M22, G72-M23 and G72-M24 (G72-M12 was not genotyped in the HapMap sample). G72-M15 captured 16 SNPs in LD region 3 (12.4%). 34% of common variation was captured in LD region 4: G72-M22 captured seven markers (6.8%); G72-M23 captured 12 markers (11.7%) and G72-M24 captured 16 markers (15.5%). The r^2 output of LD structure (Figure 4.1) is very similar to the block structure (below). However, as I used Tagger to determine how much variation I had captured with four G72 markers, I based my results on the r^2 output above. Each dark triangle is referred to as a LD region. Marker G72-15 is located in LD region 3. Markers G72-M22, G72-M23 and G72-M24 are located in LD region 4 respectively.

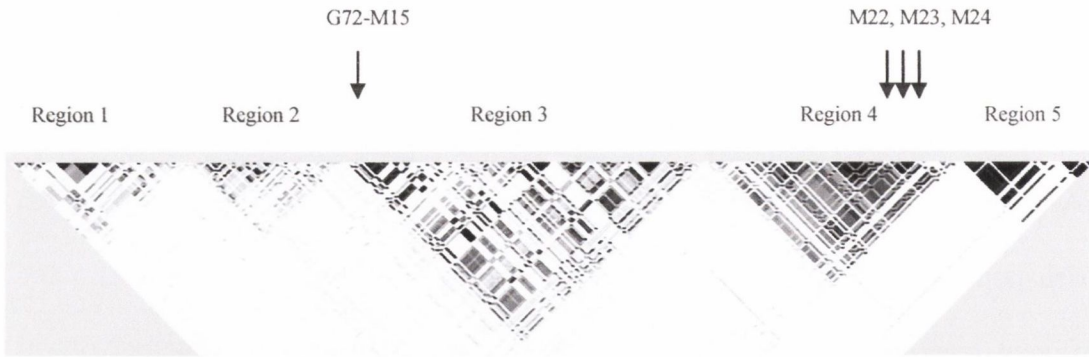


Figure 4.1 Haploview r^2 output using CEPH HapMap data of a 200Kb region on chromosome 13q containing the G72 locus.

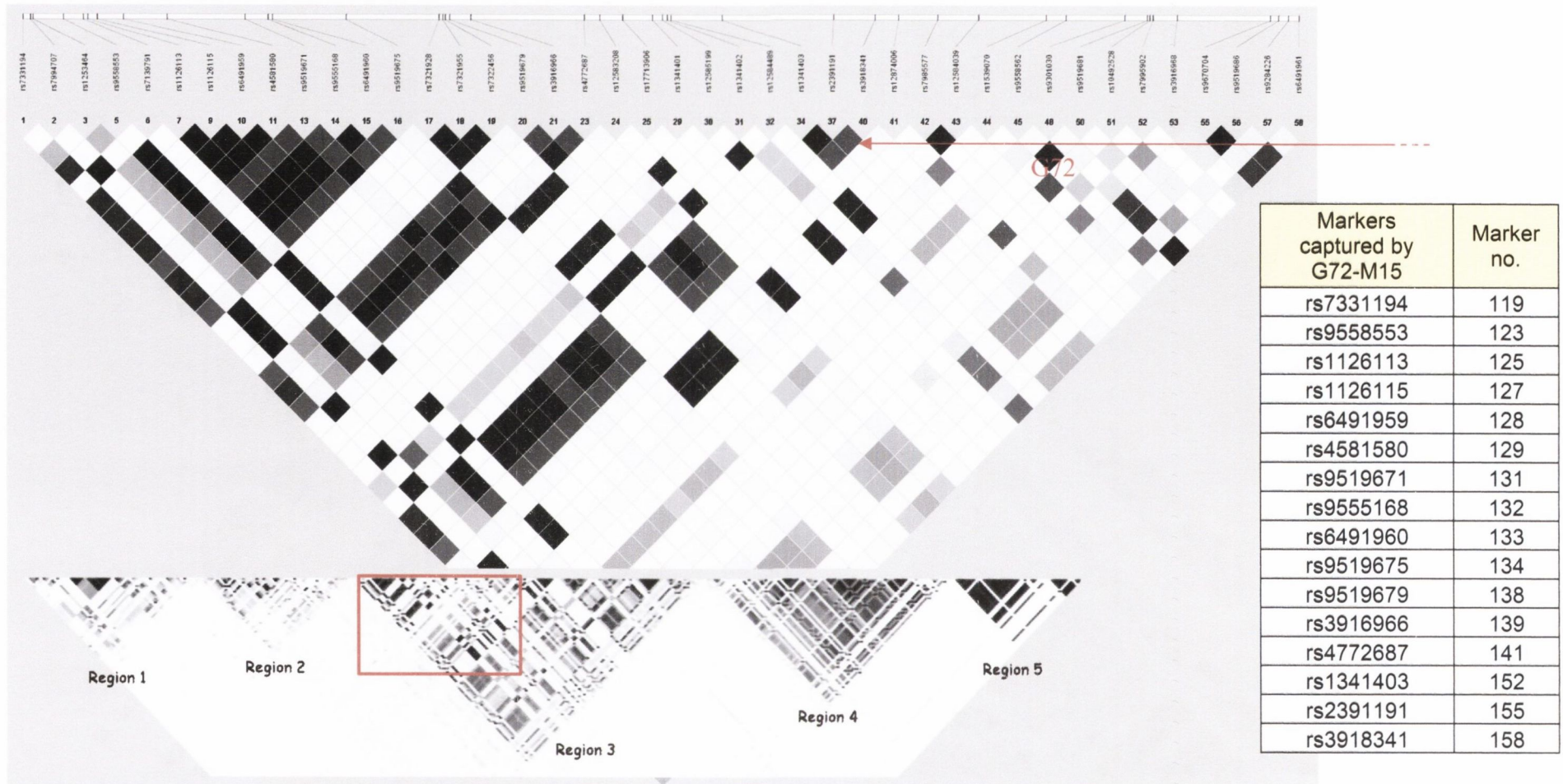


Figure 4.2 Haploview output of LD region 3. Marker G72-M15 captures 16 SNPs ($r^2 = 0.8$, LOD = 3). rs2391191 is the HapMap SNP nearest the 3' end of G72. The nearest HapMap SNP to the start of G72 is rs778294 located 5' of the partial region shown. Marker no. refers to the marker number assigned by Haploview using HapMap (Phase II Oct 05 on NCBI B34, dbSNP b124) when 200 Kb Region spanning G72 was downloaded (April 4th 2006)

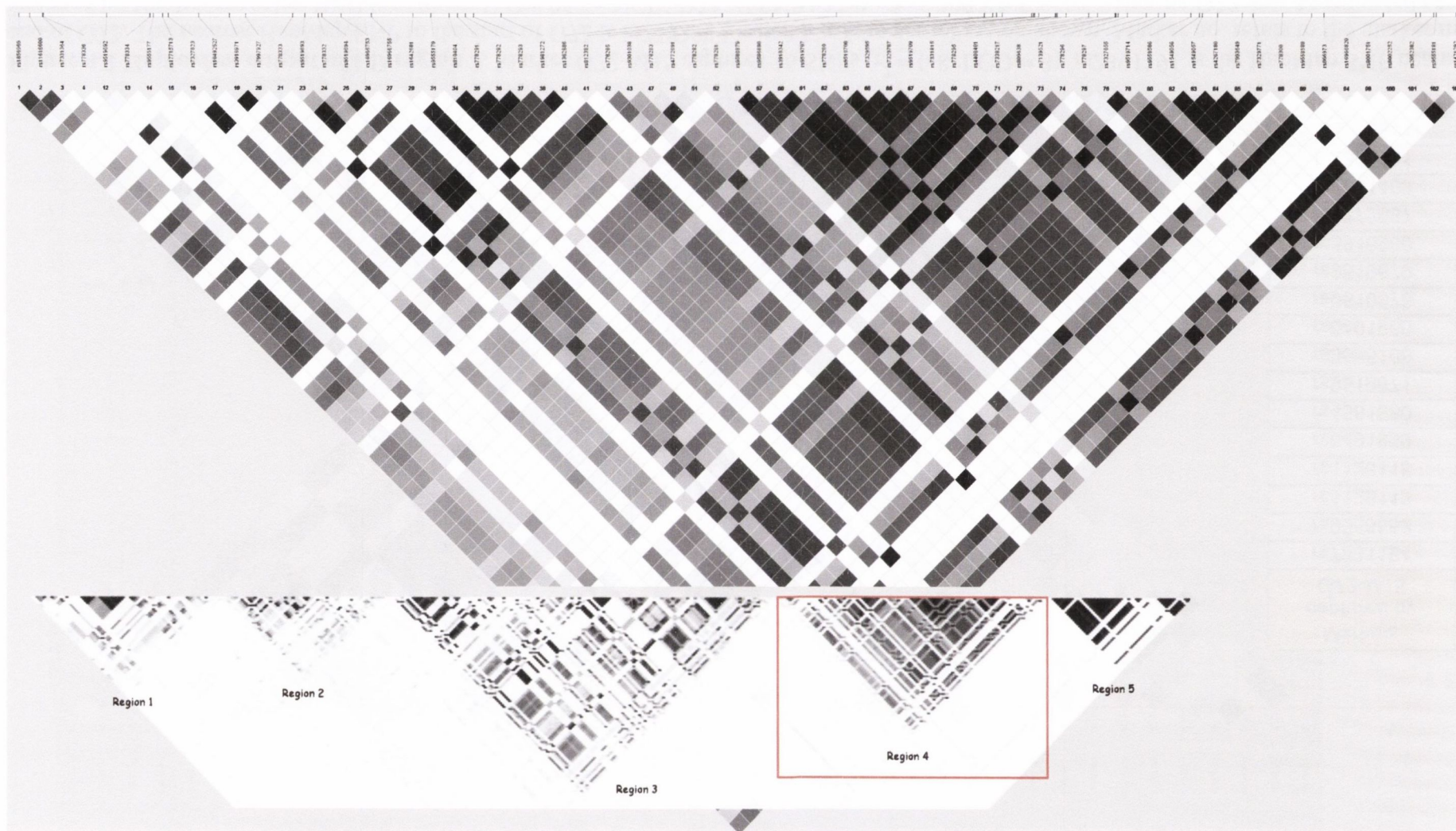


Figure 4.3 Haploview output of LD region 4. Marker G72-M22 captured 6 SNPs, G72-M23 captured 12 markers and G72-M24 captured 13 SNPs in region. Details of the SNPs captured are contained in Table 4.7.

Table 4.7 SNPs that have information captured by the genotyped SNPs

Markers captured by G72-M22	Marker no. ^a	Markers captured by G72-M23	Marker no.	Markers captured by G72-M24	Marker no.
rs7331364	258	rs9586879	308	rs9519693	278
rs1642681	284	rs9586880	312	rs9519694	280
rs810404	289	rs3918342	315	rs16966753	281
rs778291	290	rs9519707	316	rs1981272	293
rs778292	291	rs9519708	318	rs1362886	295
rs778293	292	rs7329595	320	rs9301038	298
rs778285	297	rs7329787	321	rs7327655	331
		rs9519709	322	rs9519714	333
		rs7331815	323	rs1549056	337
		rs4408401	325	rs1549057	338
		rs7338217	326	rs4267180	339
		rs7139521	328	rs7993649	340
				rs7329771	341
				rs866973	345
				rs2052382	356
				rs1421292	359
^a Refers to the marker number assigned by Haploview using HapMap (Phase II Oct 05 on NCBI B34, dbSNP b124) when 200 Kb Region spanning G72 was downloaded (April 4th 2006)					

Details of three markers analysed in the Irish study located within LD region 4 and the additional SNPs that they tag using CEPH Phase II Hapmap data.

4.2.5.2 DAAO

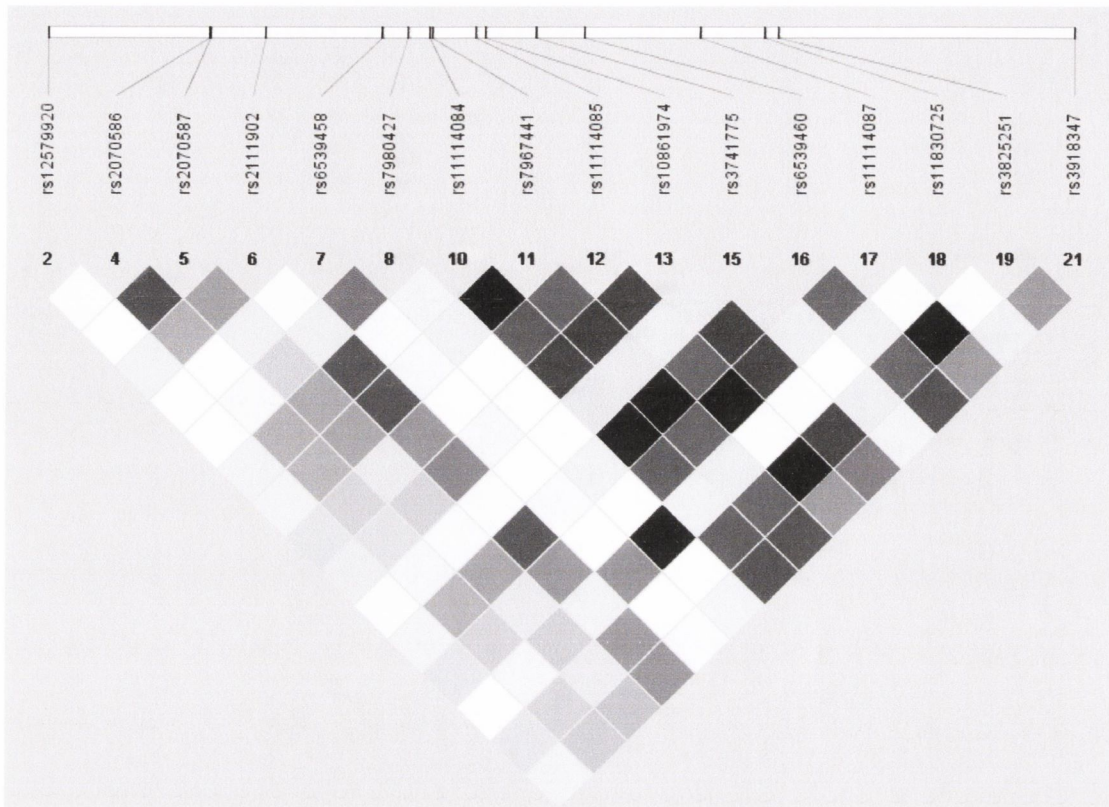


Figure 4.4 Haploview output of DAAO region. Markers DAAO-M4 (rs2111902) and DAAO-M6 (rs3741775) are number 6 and number 15 respectively.

In addition to testing how much variation was captured by the SNPs in G72 I also used HapMap data to assess the level of variation captured by the four SNPs used in DAAO. Two of the SNPs (DAAO-M5 and DAAO-M9) were not genotyped in HapMap. Therefore analysis was carried out on the two remaining SNPs: DAAO-M4 and DAAO-M6. Using Tagger with parameters set to pairwise analysis, $r^2 = 0.8$, $LOD = 3$, I ‘force included’ these markers. ‘Force included’ simply means that the program must include markers I have selected in addition to any markers the program chooses. Analysis suggested that the markers I ‘force included’ only captured information for their own marker and did not infer information on any other HapMap marker.

4.3 Discussion

The results presented in this chapter contain the first investigation of the G72/G30 and DAAO loci in an Irish SZ population. After correction for multiple testing using a novel method designed by Nyholt (2004) individual genotyping of markers G72-M12 and G72-M15 both met criteria for association with SZ at a significance level of $p=0.05$. Further analysis revealed that there was high LD between both of these markers ($D'=0.997$ and $r^2=0.993$). This suggests that either marker may truly be associated with SZ, or that both markers are in high LD with another susceptibility variant. Subsequently G72-M15 was dropped from haplotype analysis, as it provided no additional information. Furthermore, individual genotyping of markers DAAO-M4 and DAAO-M5 also met statistical significance after correction for multiple testing. Again, LD analysis revealed high correlation ($D'=0.97$ and $r^2=0.6$) suggesting that either marker may be the true susceptibility variant or in high LD with another untyped marker. These results are interesting first of all because they are a direct replication of two of the six significant G72/G30 markers initially found by Chumakov et al (2002). Although two other case-control published studies found significance with markers G72-M12 and G72-M15 as well (Schumacher et al 2004; Wang et al 2004), the signal was in the other direction (lower frequency in cases compared to controls) (see Table 4.8). These findings cannot be simply due to genetic heterogeneity as the German sample is obviously a Caucasian population. Instead it highlights the complications that currently challenge the field.

Two-marker haplotype analysis of the G72/G30 locus also found three significant associations with SZ. Subsequent analysis using GENECOUNTER revealed that the C allele of single marker G72-M12 accounted for this finding. Therefore no additional information was found.

Two markers from DAAO also showed association with SZ in the Irish population (DAAO-M4 and DAAO-M5). In comparison with three other published studies (plus a study from Cardiff, in preparation) (see Table 4.9), it is interesting to note the similarity in allele frequencies across the different populations (with the exception of the Chinese, Liu et al (2004)). However, when compared with the Irish, UK and German samples the alleles in the original French Canadian study (Chumakov et al 2002) are in the opposite direction (cases = 0.36 and controls = 0.28).

In addition, haplotype analysis of the DAAO locus found significant association with SZ for five 2-marker haplotypes and three 3-marker haplotypes. Again the association was found to be coming from the G allele of DAAO-M4 and C allele of DAAO-M5. Again this provided no additional information.

Chumakov et al (2002) also presented data on statistical interaction between the two markers most significantly associated in G72 and DAAO. Other work by our group suggests interaction between markers G72-M12 (rs3916965) and DAAO-M5 (rs3918346) in our sample (OR=9.3, (CI: 1.4-60.5), p=0.008). In addition, testing for interaction between the markers originally reported by Chumakov et al (2002) found no evidence of epistasis. However, this result should be interpreted with caution because as highlighted by Cordell (2002), definitions of epistasis can be conflicting and statistical tests of interaction are limited to testing specific hypotheses concerning precisely defined quantities. Therefore ideally, statistical interaction should be backed up by biological evidence.

In comparing the results of this study with that of other published data, the sample size of these case-control studies raises another issue (see table 4.8). The largest sample (537

cases and 538 controls) was the Chinese study (Wang et al 2004); the smallest sample was 60 cases and 130 controls (Korostishevsky et al 2004). Therefore given the odd ratios expected for complex diseases ($OR < 1.3$) and the low sample sizes, all studies lacked statistical power. However, most studies report at least one significant association. It is also possible that all of these studies have inflated Type I errors (false positive) but based on the functional evidence by Korostishevsky et al (2004), G72 showed increased transcripts (but not G30) in the Dorsal Lateral Prefrontal Cortex (DLPFC) of 40 PM SZ brains compared to controls. Therefore it seems unlikely that all of these studies can be explained by false positives.

As there are now 11 published studies, seven of which have looked at G72/G30 in SZ (Case-Control: Chumakov et al 2002; Schumacher et al 2004; Korostishevsky et al 2004; Wang et al 2004; Family based TDT: Mulle et al 2005; Zou et al 2005 and Early childhood onset TDT: Addington et al 2004) a meta-analysis was carried out by Detera-Wadleigh and McMahon (2006) based on the argument that genetic association studies should be compared at the gene level (Neale and Sham 2004) and not at the marker by marker level. All studies have used different phenotypes, study design (TDT or case-control) and ethnicity. They noted that the associated alleles across the studies varied, except for G72-M24 (rs1421292). In addition, no study to date has comprehensively studied all common variation at the G72/G30 locus. The meta-analysis revealed three highly significant ($p < 0.001$) associations from published SZ studies.

Detera-Wadleigh and McMahon (2006) also looked at the LD structure in this gene using the same Phase II data from HapMap to look at population specific LD structure between CEPH (see figure 4.5) and Han Chinese. The CEPH data showed G72 and the 5' region is split into 2 large blocks (using D' measure) with each block sub-divided into four smaller

blocks (see figure 4.5) similar in pattern to the r^2 correlations I used in section (4.3.5). The three highly significant markers from the meta-analysis span over 82Kb (between blocks 12 and 19) and are located at least over 50Kb distal to the predicted coding region of G72. In addition, they found that different markers, which have shown association in other studies, were not in LD with surrounding markers (Detera-Wadleigh and McMahon, 2006).

It is interesting that the marker G72-M15 is found in LD block 12 (region 3) in both the CEPH and Han-Chinese samples (Detera-Wadleigh and McMahon, 2006). Four case-control studies have found association with this marker (Chumakov et al 2002, Schumacher et al 2004, Wang et al 2004, Irish study). The major allele (G) in the Chumakov et al (2002) study was in the same direction as the Irish study (higher in cases) whereas the major allele (G) in the German study (Schumacher et al 2004) was lower in SZ cases. In the Han Chinese sample (Wang et al 2004) the G allele was the minor allele and is lower in cases than controls (i.e. the major allele A was in excess in the cases). This apparent difference in associated major allele between samples of European origin and the Han-Chinese is probably due to genetic heterogeneity. Therefore it is highly likely that G72-M15 is in high LD with a functional variant in block 12 (Detera-Wadleigh and McMahon, 2006).

LD analysis of the five G72 SNPs investigated in this study (Table 4.2) showed two distinct regions of LD (G72-M12 and G72-M15) and (G72-M22, G72-M23 and G72-M24). This concurs with the HapMap data showing two distinct LD regions based on r^2 values (regions 3 and 4, see figure 4.1). Retrospective analysis of the Irish study using Tagger showed how much little variation was captured by the selected markers. Using HapMap data for G72-15, a maximum of 12.4% of common variation was captured for region 3. In addition, the three markers genotyped in region 4 collectively captured a

maximum of 34% of common variation. Therefore it is likely that additional association was missed. It is also evident that other published studies on G72 have investigated little of the variation at this locus. Therefore future studies should include a higher density of SNPs in order to fine map the region for in depth association testing.

To date it is still unclear why the associated alleles vary across studies, or whether they are all in high LD with one functional variant or represent multiple functional variants contributing to SZ susceptibility at this locus.

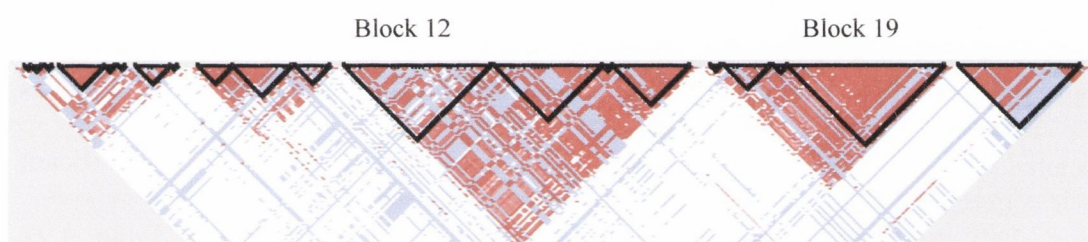


Figure 4.5 CEPH HapMap data of a 200Kb region on chromosome 13q containing the G72 locus. The review by Detera-Wadleigh and McMahon (2006), used the Gabriel et al method of D' block structure to compare LD. Block 12 (29 Kb) contains marker G72-M15 and G72 is located in this region and in block 13. Block 19 (29 Kb) contains G72-M22, -M23 and -M24.

I also compared the 4 case-control studies for DAAO. The most significant finding in the DAAO study was the C allele of DAAO-M5 ($p=0.003$) (Table 4.4) that was higher in cases than controls. The same pattern was seen in the German sample (Schumacher et al 2004). However, in the Chumakov et al (2002) study the C allele was lower in cases than controls (Table 4.9). There have also been negative results for DAAO-M5 in a large Han-Chinese sample and UK sample (see Table 4.9). Subsequent analysis similar to that carried out in G72 was not carried out due to time constraints on the project.

Retrospective analysis on the amount of variation captured by the two DAAO markers using Phase II HapMap data revealed that no additional variation had been captured. Marker DAAO-M4 was significantly associated ($p=0.018$) with SZ in both this study and three other published Caucasian studies (see Table 4.9). Therefore, one could infer that this is a causal variant. However, this study showed high LD between this marker and marker DAAO-M5 ($D'=0.97$, $r^2=0.6$), not genotyped in HapMap. This finding highlights marker density issues suggesting that HapMap should be used as a useful tool but not as a definitive guide in determining the amount of variation captured at a particular locus.

In addition, the second marker (DAAO-M6) investigated using Phase II HapMap data was not associated with SZ in this sample. However, Chumakov et al (2002) found significant association with the same marker in their French-Canadian sample (see Table 4.9). Therefore it is possible that LD structure between the French-Canadian and Irish populations differ slightly with DAAO-M6 being in high LD with a causal variant elsewhere at the locus.

This study has focused on the SZ phenotype. However, further comment is warranted on G72 involvement in Bipolar disorder. To date, it is the best supported locus for Bipolar disorder (Craddock et al 2006). Five studies consisting of two European case-control samples (German and UK), (Schumacher et al 2004; Williams et al, personal communication) and three American family studies (Hattori et al 2003; Chen et al 2004) have all shown association, although with different variants. The most interesting finding to date is that the UK sample found association in Bipolar disorder but not SZ. However when the two samples were divided into a subset of individuals with major mood disorder ($n=818$) association was found (Williams et al, personal communication). This evidence

suggests that the G72/G30 locus is involved in major mood disorder rather than the Kraepelin dichotomy of SZ and BP (Craddock et al 2006). Evidence reviewed by Craddock et al (2006) suggests that other genes recently implicated in SZ pathophysiology (NRG1, DISC1, Dysbindin) may also be involved in an overlap of phenotypes such as schizoaffective, bipolar and mood disorders (see Figure 4.6). Future genetic studies should incorporate the strategy of looking at symptoms dimensionally for psychosis, depression and mania suggested by Craddock et al (2006), which may eventually have implications for clinical diagnosis (Craddock et al 2005).

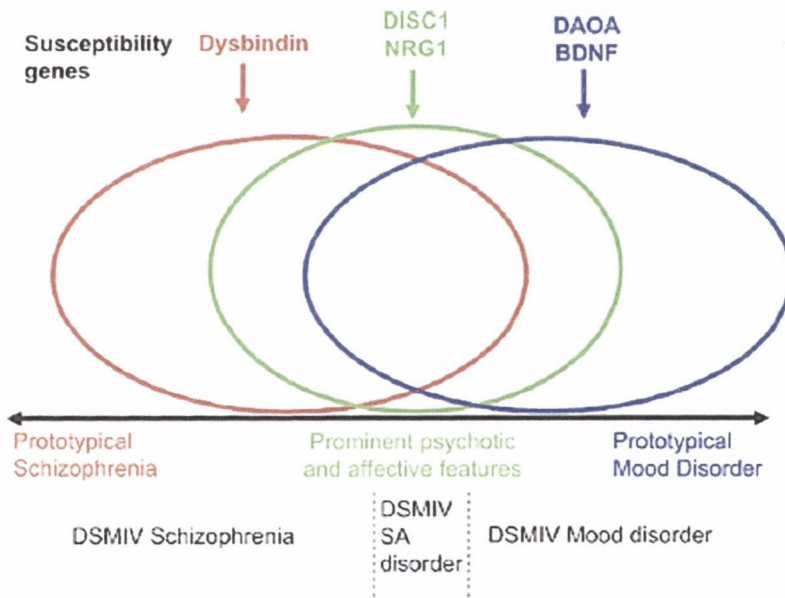


Fig. 1. Simplified hypothesized relationship between specific susceptibility genes (above the black line) and clinical phenotype (below the line) using the model outlined in Craddock and Owen⁸. The overlapping ellipses represent overlapping sets of genes: red influencing susceptibility to phenotypes with prominent schizophrenia-like features, blue to prominent mood features, and green to phenotypes with a prominent mix of both types of feature. These assignments are based on current data and are likely to require revision as more data accumulate.

Figure 4.6: Susceptibility and phenotype Adapted from Nick Craddock, Michael C. O'Donovan and Michael J. Owen *Schizophrenia Bulletin* vol. 32 no. 1 pp. 9–16, 2006 NB: DAOA = G72

The Chumakov et al (2002) study provided the first evidence of genetic association within a linkage region for SZ. It was the first study to show that G72 was expressed in the brain and interacted with DAAO suggesting a link to NMDA signaling (Detera-Wadleigh and McMahon, 2006). This Irish study has shown direct replication of two markers in G72/G30 (G72-M12 and G72-M15) yet opposite replication of two markers in DAAO (DAAO-M4 and DAAO-M5). In addition, this study showed statistical interaction between G72/G30 and DAAO (G72-M12/DAAO-M4) yet failed to replicate the original interaction (Chumakov et al, 2002). This study provides further evidence of association of G72/G30 and DAAO in SZ susceptibility but adds to the confusion over associated marker variability in previously published studies. This lack of consistency across studies may reflect differences in sampling methods, ascertainment, ethnicity, genetic heterogeneity, or false positive findings. In order to access this gene further identification of functional variants, LD structure and biological studies need to be carried out. Therefore although our study encourages the hypothesis of G72/G30 and DAAO involvement in the pathogenesis of schizophrenia, further studies, including functional, are warranted.

Table 4.8 Comparison of results from published genetic association studies of G72/G30 and schizophrenia

Markers	Alleles ^a	Irish Study			Chumakov et al (Canadian)			Chumakov et al (Russian)			Schumacher et al (Germany)			Korostishevsky et al (Ashkenazi Jew)			Wang et al (Chinese)		
		Allele Freq Cases	Allele Freq Controls	p value	Allele Freq Cases	Allele Freq Controls	p value	Allele Freq Cases	Allele Freq Controls	p value	Allele Freq Cases	Allele Freq Controls	p value	Allele Freq Cases	Allele Freq Controls	p value	Allele Freq Cases	Allele Freq Controls	p value
M12	C/T	0.689	0.622	0.005	0.64	0.55	0.007	ns	ns	ns	0.59	0.64	0.048	ns	ns	ns	0.38	0.43	0.019
M15	G/A	0.683	0.622	0.011	0.65	0.58	0.032	na	na	na	0.59	0.64	0.037	ns	ns	ns	0.38	0.45	0.001
M22	A/G	0.592	0.625	ns	0.69	0.60	0.003	ns	ns	ns	na	na	na	ns	ns	ns	ns	ns	ns
M23	T/C	0.510	0.543	ns	0.57	0.49	0.019	0.60	0.51	0.017	0.47	0.54	0.033	62.3	43.3	0.001	0.49	0.46	0.16
M24	T/A	0.527	0.494	ns	0.55	0.47	0.019	ns	ns	ns	0.56	0.5	0.036	na	na	na	na	na	na
Sample Size																			
Cases		299			213			183			299			60			537		
Controls		645			241			183			300			130			538		

^a Frequencies relate to first allele listed

ns = marker was genotyped in other studies but the p value was not significant. na = marker was not genotyped in other studies, not applicable.

-147--

Table 4.9: Comparison of results at the DAAO locus and schizophrenia

Markers	Alleles ^a	Irish Study			Chumakov et al (Canadian)			Yui et al 2004			Schumacher et al (Germany)			Williams et al (Cardiff)		
		Allele Freq Cases	Allele Freq Controls	p value	Allele Freq Cases	Allele Freq Controls	p value	Allele Freq Cases	Allele Freq Controls	p value	Allele Freq Cases	Allele Freq Controls	p value	Allele Freq Cases	Allele Freq Controls	p value
DAAO-M4	G/T	0.266	0.320	0.018	0.36	0.28	0.017	0.5074	0.4758	0.319117	0.28	0.34	0.026	0.28	0.33	0.02
DAAO-M5	C/T	0.812	0.751	0.003	0.72	0.79	0.007	0.5054	0.5093	0.934585	0.78	0.72	0.019	0.75	0.75	0.99
DAAO-M6	T/G	0.517	0.554	0.136	0.59	0.49	0.001	0.7569	0.6623	0.000001	0.53	0.61	0.021	0.56	0.55	0.55
DAAO-M9	T/C	0.708	0.738	0.185	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a

^a Allele Frequency represents first allele listed

Chapter 5 – Oxidative stress and schizophrenia: HSPA8 and GSTM3

5.1 Introduction

There is a substantial amount of literature suggesting that pre-, peri-, or post-natal complications are involved in SZ pathogenesis. A recent meta-analysis of obstetric complications (OC) and SZ sought to categorise whether specific OC increase risk (Cannon et al 2002). The authors found three main categories of complications associated with risk: 1) complications of pregnancy 2) abnormal foetal growth and development (low birth weight, congenital malformations and small head circumference) 3) complications of delivery (Asphyxia, uterine atony and C-section). The common mechanisms appear to be foetal hypoxia or anoxia. The developing foetus is most sensitive to stress and environmental insults during the second trimester. It is at this time that brain regions such as the hippocampus develop (Cannon et al 2005) which have been implicated in SZ pathology (Schulze et al 2003). Indeed the hypothesis that perinatal hypoxia may be etiologically relevant to hippocampal dysfunction in high risk individuals who later developed SZ was first hypothesised over 35 years ago (Mednick 1971).

Aerobic cellular respiration occurs in the mitochondria and under normal circumstances produces reactive oxygen species (ROS). Examples of ROS include $O_2^{\bullet-}$, OH^{\bullet} , OH , NO^{\bullet} and $ONOO^{\bullet}$. If ROS are not eliminated effectively, they can cause oxidative cell damage such as peroxidation of DNA, cell membrane phospholipids and esterified essential polyunsaturated fatty acids (EPUFAs) (Mahadik et al 2001) and denature proteins such as

cellular receptors and enzymes, which disrupts function (Macario and Macario 2005). It is essential for the cell to eliminate these destructive ROS using molecular chaperones such as heat shock protein 70 KDaltons (HSP70) (Macario and Macario 2005), enzymatic (superoxide dismutase, SOD; glutathione peroxidase, GSHPx and catalase, CAT) and non-enzymatic (GSH and uric acid) antioxidant defences and dietary antioxidant defences (vitamins A, E and C, β -carotene, Q-enzyme and flavons) (Mahadik et al 2001). However, if this defence mechanism fails, it leads to increased cellular levels of ROS -termed oxidative stress (Mahadik et al 2001).

Several events are thought to occur after a hypoxic-ischemic insult pre- or peri-natally. During this insult, the immune system initiates an antibody response causing damage to the micro vascular system, which makes the brain more susceptible to infections and environmental damage. The inflammation response also interferes with normal aerobic respiration by overproducing ROS, abnormal cellular Ca^{2+} homeostasis and activation of apoptotic proteins leading to neuronal damage. During reperfusion of the neonate (which means that the brain is starting with a low pH and increased cellular concentration of Ca^{2+}) these conditions promote the generation of ROS, which increases OS and promotes apoptosis and necrosis 24-48 hours post-hypoxic insult. If the antioxidant response genes functioned properly, neurons would be able to protect against this oxidative stress pathology. However, if patients with SZ were genetically predisposed with an impaired antioxidant response, one would expect the brain of a SZ patient to show evidence of low pH (increased iron content), lipid peroxidation and DNA fragmentation of neurons (which may be a primary cause of enlarged ventricles), and a lower gene expression of antioxidants involved in OS protection.

Recently it had been suggested that that the neurotoxic effects of foetal hypoxia leads to an early onset of schizophrenia through premature cortical synaptic pruning (Rosso et al 2000). Post mortem data has shown reduced neuronal size and reduced levels of synaptic proteins in the hippocampus (Harrison 1999). Recent meta-analysis of MRI scans (Wright et al 2001) has shown that there are consistent reports of enlarged ventricles and volumetric decreases of temporal lobe structures in SZ. Subsequent studies have shown that genetic loading together with hypoxic OCs are associated with ventricular enlargement (Falkai et al 2003; Schulze et al 2003), reductions in the medial temporal lobe and reduction of the left hippocampus (Schulze et al 2003) in SZ patients. Oxidative stress injury predominately affects both brain structure and function. There is evidence that peroxidative loss of membrane phospholipids and EPUFAs cause a reduction in neuronal processes and synapses in SZ patients (Mahadik et al 2001). This could explain the secondary effects of reduction in certain brain regions and increased ventricular size. This adds further support to the MRI data that abnormalities in genes affecting early brain development and environmental insults pre- or perinatally cause hippocampal reduction (Schulze et al 2003). There is also evidence that membrane receptor-mediated phospholipid-dependant signal transduction of several neurotransmitters have structural changes which would have happened during neurodevelopment (Mahadik et al 2001). Post mortem studies have reported increased iron content in the brain. Iron is transported using transferrin and remains bound to this transporter unless the pH of the brain is lower than normal (as in the event of mitochondrial dysfunction or hypoxia). In this case, the iron is deposited in the brain. Free iron can then assist in creating hydroxyl radicals in the presence of vitamin C or superoxides (Mahadik et al 2001) increasing cellular oxidative stress.

There is also data from animal studies on oxidative stress. The long-term effects of perinatal hypoxia on brain development has become an area of increasing research interest (Boksa 2004). It has been established (Hallivell 1992; Mahadik and Mukherjee 1996) using animal studies that the brain is the organ most affected in global oxidative stress. This is because the brain has a very high aerobic metabolism and blood perfusion, which generates high levels of ROS and has a relatively lower enzymatic antioxidant defence. In addition, it is enriched in lipids which are susceptible to damage from oxidative stress and neuronal DNA in an adult brain cannot be repaired since there is no DNA replication (Mahadik et al 2001). Furthermore, because of the high blood perfusion, there is an increased susceptibility to environmental stressors such as smoke, pollution and radiation. If there is a high oxidative stress event during pregnancy this may cause abnormal neurodevelopment in the foetus (Mahadik et al 2001). There are three main rat models of perinatal hypoxia: 1) unilateral carotid artery ligation together with exposure to a hypoxic atmosphere in postnatal rats; 2) Exposing rat pups to global anoxia during a caesarian section birth by immersion of the isolated intact uterus in a saline bath; 3) exposure of postnatal rats to anoxic or hypoxic atmospheres (Boksa 2004). The rat brain at birth is less mature than human brain. Therefore using postnatal rats ($n > 2$ days) in hypoxia experiments creates a paradigm more similar to a human brain at birth. Studies of this sort have shown that a single episode of hypoxia initiates little behavioural effect whereas chronic repeated exposure to hypoxic conditions may have a more profound impact on nervous tissue (Boksa 2004).

Therefore, epidemiological, post mortem, MRI and animal studies all suggest that oxidative stress is involved in the pathogenesis of SZ. However, studies have shown that more than 90% of individuals who experience OC do not develop SZ. This suggests that hypoxic OC must act additively or interactively with genetic factors in influencing liability

(Cannon et al 2000). The hypothesis of this study was that normal mitochondrial and cellular oxidative stress responses were dysfunctional in SZ patients because of mutations in genes involved in these two processes. These mutations would lead to a decrease in antioxidant gene expression, thereby increasing ROS and oxidative stress. In order to identify possible candidate genes I turned to gene expression studies of post mortem SZ patients.

Microarray technology has developed at an increasing rate of pace and has become a standard tool in many genomics research laboratories. It has revolutionized biological research because instead of working on a gene-by-gene basis, scientists can now study tens of thousands of genes at once. However, this generates a lot of complex data at any one time and careful investigation and interpretation of data analysis is required (Leung et al 2003). In addition, post mortem studies of schizophrenia can be affected by a host of confounding variables.

Although it is possible to match experimental subjects on the basis of post mortem interval, gender, ethnic origin or time since death, other variables are more difficult to control. These include agonal state, ambient temperature after death and the accurate calculation of the post mortem interval. Furthermore, the assumption that patients with schizophrenia and healthy controls can be matched ignores the higher incidence of alcohol and substance abuse in patients with schizophrenia, the consequences of a lifetime of treatment with neuroleptics and other drugs, the stress of acute exacerbations of illness and admissions to hospital, and the effects of a chronic mental illness on quality of life (Perlman et al 2004).

Other limitations include a decreased sensitivity of the arrays to the detection of genes with low expression levels (low-abundance genes). Microarrays do not measure

posttranslational modifications such as phosphorylation. Another possibility is confounding microarray results through a process of cross-hybridization in which specific components of the arrays will cross-hybridize because of sequence similarity of the probes as defined by Affymetrix. Lastly, tissue heterogeneity continues to be a persistent challenge for microarray studies, especially in brains with multiple densely packed cell types (Bunney et al 2003).

Good experimental design for microarrays should contain at least four elements: 1) a clearly defined biological question and/or hypothesis; 2) treatment, perturbation and observation of the biological materials, as well as the microarray experimental protocols, should be as little affected by systematic and experimental errors as possible; 3) a simple, sensible and statistically sound microarray experimental arrangement that will give the maximal amount of information given the cost structure and complexity of the study; and 4) compliance with the standard of microarray information collection (Minimum Information About a Microarray Experiment (MIAME) (<http://www.mged.org/Workgroups/MIAME/miame.html>) (Leung et al 2003).

Microarray studies have been used successfully (Middleton et al 2002, Hakak et al 2001) to identify myelin-related gene changes in neuropsychiatric disorders. However, inter-individual variance in gene expression can be high which makes detecting true differences difficult. By using this three-way parallel approach Prabakaran et al (2004) provides a more comprehensive picture into the disease processes of SZ by providing insight into disease-specific regulatory mechanisms and metabolic networks, giving a more comprehensive picture of the SZ brain.

A recent study by Sabine Bahn's team in Cambridge, UK proposed that oxidative stress and the ensuing cellular adaptations are linked to the pathogenesis of SZ (Prabakaran et al 2004). Their study involved a parallel transcriptomics (mRNA), proteomics (proteins) and metabolomics (low molecular weight intermediates) set of experiments using human post-mortem (PM) brain tissue (white and grey matter from the prefrontal cortex).

Initially they used 10 SZ and 10 controls PM prefrontal cortex samples for their proteomics and metabolomics experiments. In addition, they carried out a gene expression study (transcriptome) using 48 SZ and 44 control PM prefrontal cortex samples, which included the 20 samples used in the proteomics and metabolomics studies. Isolated total RNA was processed through the Affymetrix preparation protocol (<http://www.affymetrix.com>) and hybridized to 92 HGU133A GeneChips (Affymetrix CA, USA).

The proteomics results indicated 50 significantly altered proteins: 19 associated with mitochondrial function; 16 with oxidative stress and three with peroxisomal function. The remainder were cytoskeletal proteins and proteins associated with protein trafficking/turnover (see Table 2.5). Using the same samples used in the proteomics study, Prabakaran et al (2004) found 10 out of 60 metabolites significantly altered in white matter but not grey matter of the prefrontal cortex.

Results from the gene expression study identified 3406 transcripts that were found to be significantly altered in the SZ sample compared to controls. Genes were localised depending on their cellular localisation and with the largest number of genes assigned to mitochondrial and mitochondria related proteins. Further analysis of the 92 gene chips identified 59 significantly altered genes mainly involved in energy metabolism and mitochondria. The investigation of gene and protein expression combined with measurements of metabolites provides insights into disease-specific regulatory

mechanisms and metabolic networks, generating a more comprehensive and converging picture of the diseased brain. They identified several significantly altered metabolic pathways in SZ brain tissue. Therefore, the results presented by Prabakaran et al (2004) provide evidence of increased oxidative stress in the SZ brain.

Their findings, together with epidemiological, MRI, post-mortem and animal studies strongly suggest that genetic predisposition to impaired cellular anti-oxidative responses combined with chronic pre- or perinatal hypoxic events induces cellular oxidative stress and the production of ROS which is involved in the pathogenesis of SZ. As there are many plausible genes involved in cellular antioxidant response, I took a converging approach using experimental and positional data in order to reduce the number of genes for this study.

5.1.1 Plausibility of HSPA8 and GSTM3

Overall, HSPA8 and GSTM3 seem plausible candidates for SZ pathogenesis as they are both involved in cellular oxidative stress response. HSPA8 is part of the family of HSP70 proteins that are molecular chaperones aiding proteins to fold correctly. Misfolding can cause dysfunction of certain proteins. HSPA8 is also a mitochondrial protein and is involved in active respiration. GSTM3 transfers glutathione, which is the main non-protein antioxidant that protects cells from ROS generated from DA metabolism. If glutathione levels are reduced, membrane peroxidation and microlesions occur around dopaminergic terminals resulting in a loss of connectivity with excitatory glutamatergic neurons (Grima et al 2003). This may be involved in synaptic pruning (Rosso et al 2000) and reduction in neuronal size (Harrison 1999). Glutathione is also important in drug detoxification, gene expression and neurodevelopment (Grima et al 2003). The proteomics, metabolomics and

gene expression study by Prabakaran et al (2004) found a significant decrease in white and grey matter for HSPA8 and a significant decrease in grey matter for GSTM3 in PM SZ brain tissue. Therefore, if both of these genes function at lower levels in nerve cells compared to a normal control, it suggests that the antioxidative stress response cannot cope with the production of ROS. The nerve cells are eventually subjected to chronic oxidative stress which in turn causes e.g. membrane peroxidation and microlesions leading to a loss of DA connections with glutamate, both of which are neurotransmitters implicated in SZ pathophysiology.

5.2 Overview of HSPA8

Using 15 SZ patients I identified 13 SNPS from direct resequencing of HSPA8. By comparing the Irish data with the NCBI database I discovered two of these, HSPA8-6 and HSPA8-9 were novel. The next stage was to take a sub-sample of the control group (n=92), termed an Irish reference panel, and determine the LD structure of HSPA8 in order to identify tagging SNPs that could cover all variation spanning the gene. However, after genotyping and LD analysis, no LD structure was identified. Therefore, six SNPs were genotyped in the full sample based upon SNP location and minor allele frequency (MAF) values in Irish populations. Genotyping was carried out commercially using Kbiosciences.

I will describe my work on HSPA8 in the following sections: materials and methods of sequencing; results of sequencing; materials and methods of SNP genotyping in an Irish reference panel; results of SNP genotyping; materials and methods for LD analysis in an Irish reference panel; results of LD analysis, and materials and methods for SNP genotyping by Kbiosciences in the full SZ case-control sample, and results from full SZ case-control sample.

5.3 Results of HSPA8

5.3.1 Results – Resequencing of HSPA8

Of the 14 fragments resequenced, seven fragments were found to contain a total of 13 SNPs that were polymorphic in the Irish sample. Results of HSPA8-1 (rs3179174) are shown as an example of the sequencing results in Figure 5.1.

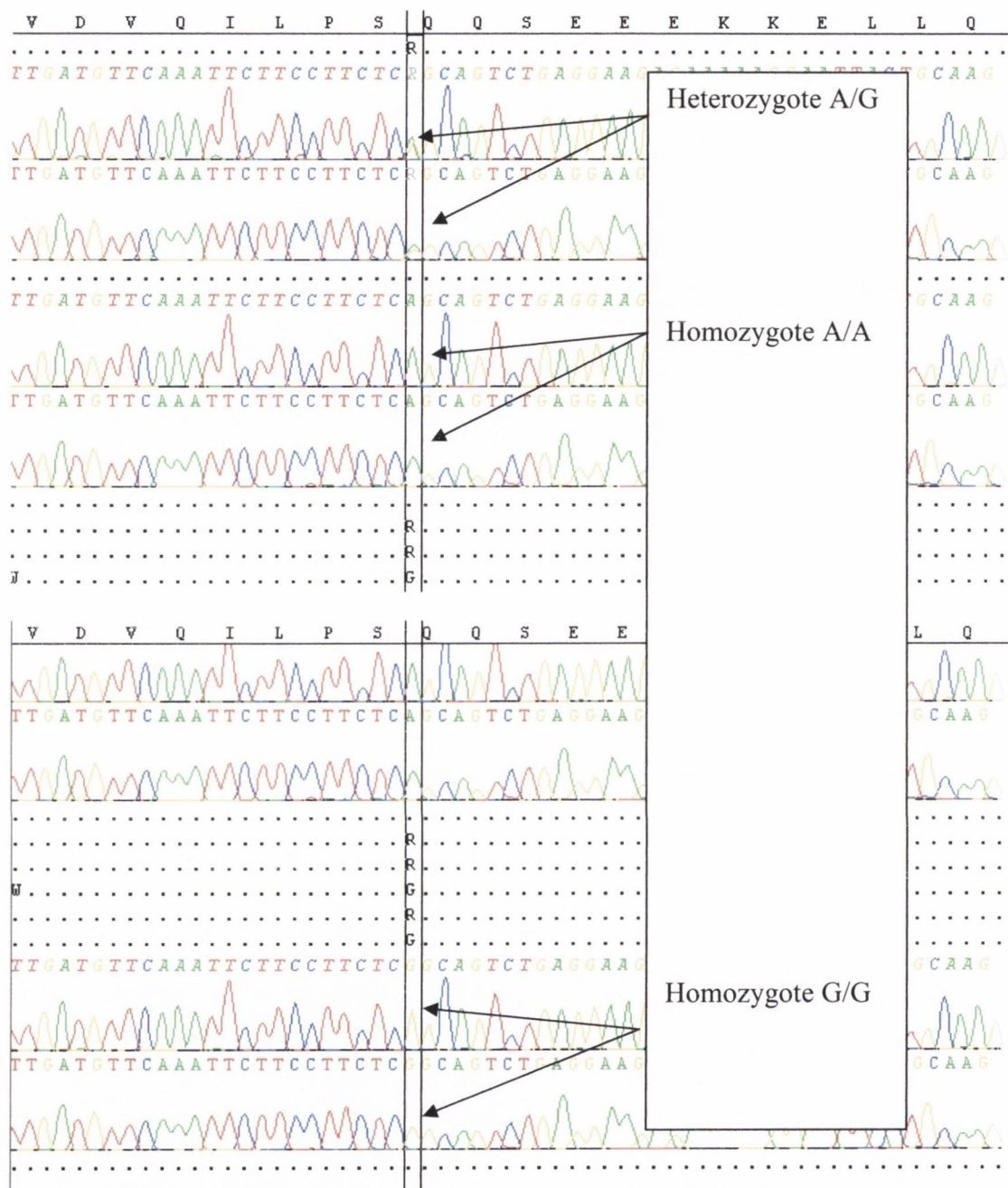


Figure 5.1 Output from SeqScape v2.1 of HSPA8-1 (rs3179174), at 35825 bp on contig AP000926 (ROI) – polymorphism is Nucleic Acid Code R (GA purine).

Next, these SNPs were checked against the NCBI dbSNP database (see Table 5.1) in order to determine if any of these SNPs were already known. Eleven of the SNPs were found to be in the dbSNP database. Two SNPs (HSPA8-6 and HSPA8-9) were identified as novel polymorphisms in the Irish sample.

Table 5.1 Extension Primers and associated NCBI rs number

dbSNP rs#	Fragment number	Alleles	Extension Primer / SNP ^a
rs3179174	13	A/G	1
rs3763897		A/T	2
rs1064585	10	G/T	3
rs1461496		G/A	4
rs1461497	6	T/C	5
HSPA8- 6	5	T/C	6
rs4935825	4	C/A	7
rs11218941		G/A	8
HSPA8-9		G/T	9
rs1136141	3	G/A	10
rs2276075		G/A	11
rs2276077		T/C	12
rs2236660	2B	G/A	13

a = Each SNP identified was assigned a number in numerical order across the gene. The corresponding extension primer had the prefix HSPA8- placed in front of each SNP number for clarity

5.3.2 Results – Multiplex Genotypes using SNaPshot

Fragment HSPA8-4: multiplex of extension primers 8 and 9 (Figure 5.2). Output from Genemapper version 3 showing homozygous A for SNP 8 and homozygous G for SNP 9. Sample is sample 11. Note the size of extension primer 8 (23bp) and extension primer 9 (27bp). These primers were designed as 20 and 24 base pairs respectively. The increase in base pairs is an artefact of sizing in the genotyping process.

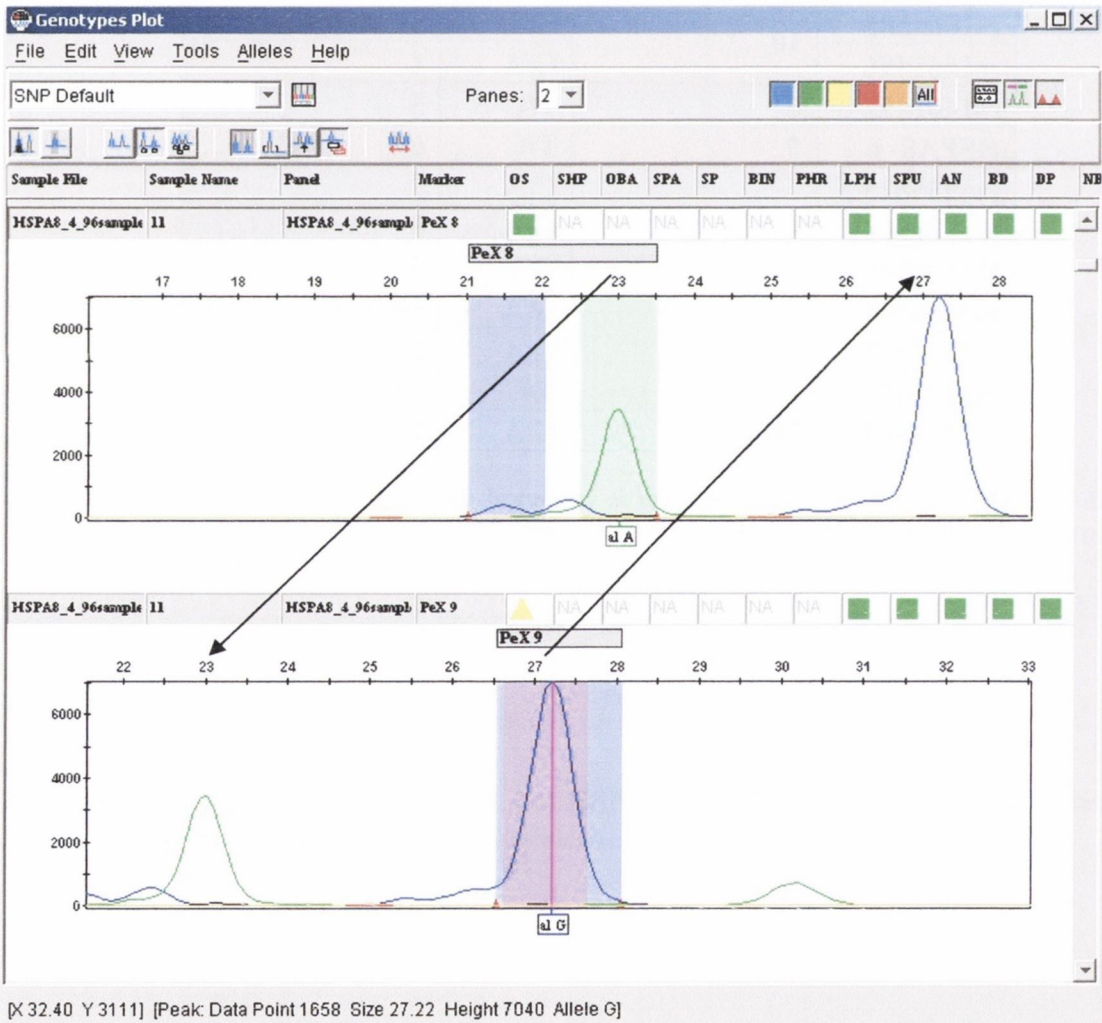


Figure 5.2. Output from Seqscape of individual heterozygous for HSPA8-4 (A/G).

5.3.3 Results - LD structure and analysis

Although HSPA8-6 and HSPA8-9 were detected through re-sequencing, both SNPs were found to be uninformative in the Irish reference panel and these two SNPs were removed from all subsequent analysis. The remaining 11 SNPs underwent LD analysis (see Figure 5.3), but no ‘block-like’ LD structure was detected. The analysis was limited to relatively common SNPs ($MAF < 0.1$), thereby excluding 5 SNPs from the analysis. Again no strong patterns of LD were detected (Figure 5.4).

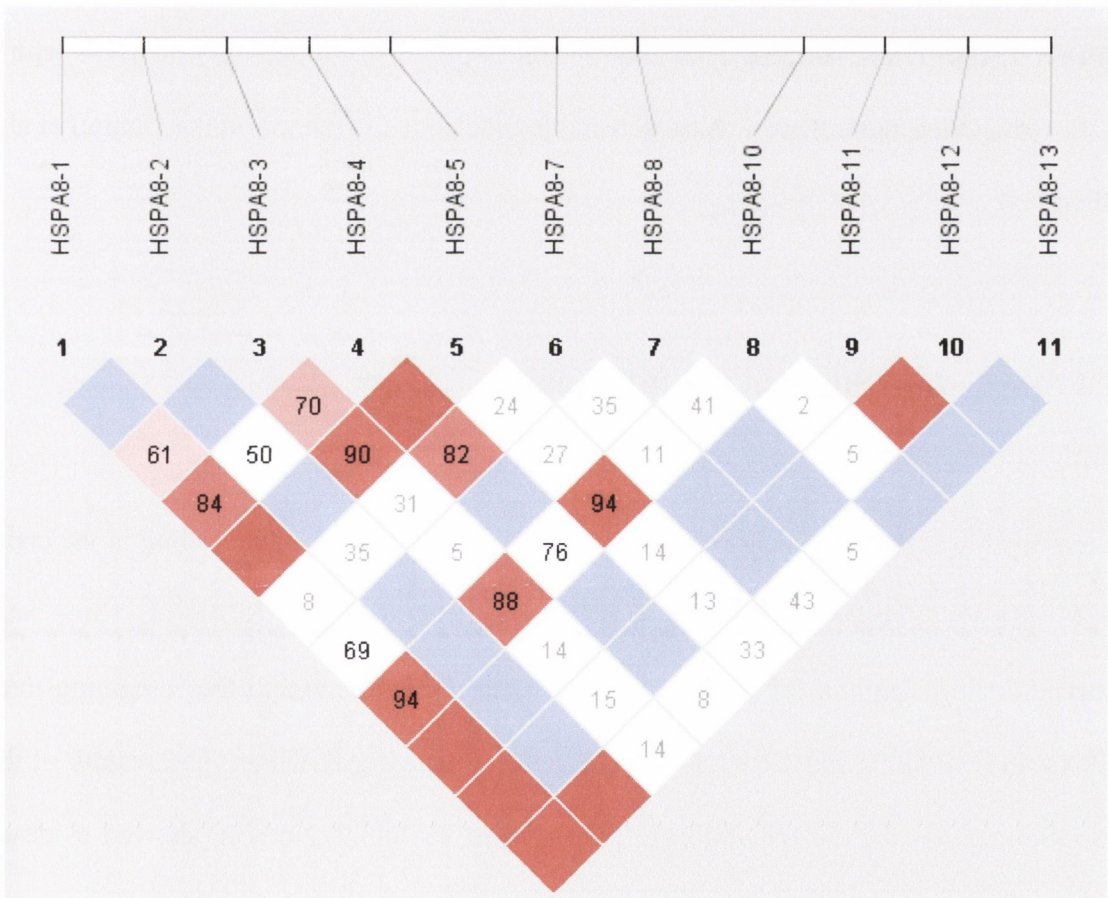


Figure 5.3 Haploview output. No detectable blocks of LD using the Gabriel et al method.

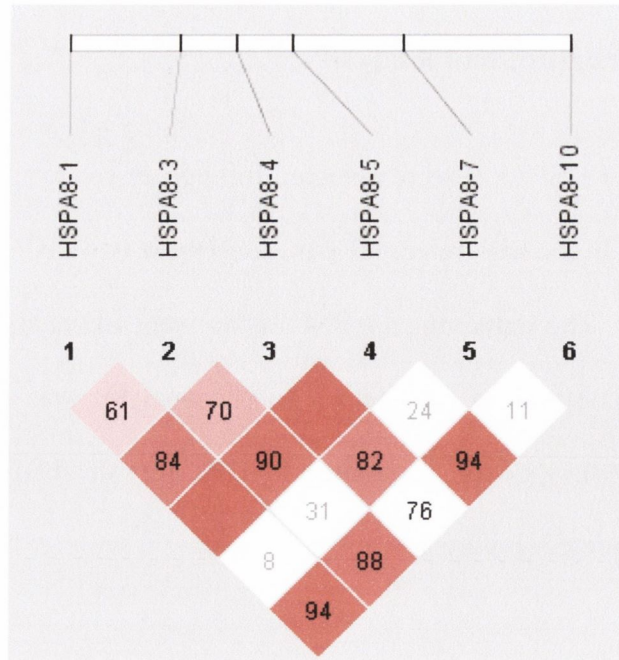


Figure 5.4 Haploview Output. This output from Haploview is based only on SNPs with MAF >10%. Once again there was no detectable blocks of LD (based on the Gabriel et al method).

5.3.4 Association Study in the full sample

For reasons highlighted in section 2.8.9 six SNPs were selected for genotyping in the entire sample HSPA8-1, HSPA8-3, HSPA8-4, HSPA8-5, HSPA8-7 and HSPA8-11. Three were genotyped using Taqman (HSPA8-1, HSPA8-3 and HSPA8-11) and the remaining three (HSPA8-4, HSPA8-5, and HSPA8-7) were genotyped using KASPar. The results of the association studies for the six SNPs are presented in Table 5.2. No SNP showed evidence of association with SZ. To try and capture unknown variants that could be associated at each gene (i.e. variants not analysed in our reference panel), I performed 2- and 3-marker haplotype analyses using FASTEHPPLUS and Genecounting (described in section 2.7.3) using all combinations of SNPs available. No haplotypes were found to be significantly associated with the phenotype.

5.3.5 Results - Determining SNPs for genotyping in entire sample

Table 5.2 Results of the association study on HSPA8-1, 3, 4, 5, 7 and 11 in the full sample

dbSNP ID	Chromosomal position ^a	Position within gene ^b	Alternative marker name	Polymorphism	Frequency of Cases ^c	Frequency of controls ^c	p value
rs3179174	122442204	4302	HSPA8-1	A/G	0.772	0.776	0.846
rs1064585	122441419	3517	HSPA8-3	T/G	0.825	0.840	0.447
rs1461496	122441202	3300	HSPA8-4	A/G	0.356	0.348	0.706
rs1461497	122439613	1711	HSPA8-5	C/T	0.821	0.840	0.274
rs4935825	122438461	559	HSPA8-7	A/C	0.583	0.578	0.802
rs2276075	122438024	122	HSPA8-11	G/A	0.963	0.959	0.704

^a Position on Chromosome 11 determined from according to July 2004, NCBI 35 freeze of Human Genome Assembly using the Ensembl b

^b Position of SNP within gene in relation to ATG site (base pairs)

^c Number represents frequency for first allele listed

All markers were genotyped in the entire sample

5.4 Discussion

This is the first association study of HSPA8 and SZ. It was investigated based upon functional and positional evidence. The gene maps to chromosome 11q24.1 that was within one of the significant bins in the SZ linkage meta-analysis by Lewis et al (2003). The same region was suggestive of linkage in a study by Gurling et al (2001). In addition, Prabakaran et al found that HSPA8 was significantly down regulated in both white and grey matter of PM SZ brain.

My strategy employed re-sequencing this gene to identify common variants in the Irish population. This identified 13 SNPs, 11 of which were registered in dbSNP. Two other SNPs, HSPA8-6 and HSPA8-9 were novel to the Irish population. All 13 SNPs were genotyped to determine LD structure. However, the two novel SNPs were found to be uninformative and were dropped from LD analysis. LD analysis revealed no identifiable patterns of LD between the remaining 11 markers. A further five SNPs had a MAF < 0.1 reducing the power of the full sample to detect any association. Six SNPs were chosen for genotyping in the full sample based on heterozygosity information from dbSNP and the Irish reference panel and also on marker location. None of the SNPs were found to be associated with SZ in the Irish sample. Subsequent 2 and 3 marker haplotype analysis also revealed no evidence of association.

There are several interpretations of this finding. The first is that the study was underpowered to detect variants with OR < 1.3 (e.g. see power calculations in section 3.3). In addition as this was the first association study of HSPA8 the prior probability of this gene being involved in the pathogenesis of SZ was low. As there was no detectable LD for HSPA8 with the markers identified, SNPs genotyped in the full sample were chosen based

on MAF and location. Therefore it is possible that I missed an association with another variant. It is also possible that a causal variant may be located some distance away in a regulatory region that I did not sequence. There may also be issues related to genetic heterogeneity between the sample used by Prabakaran et al (2004) and the Irish sample (discussed in section 3.3). It is also possible that, for example, a transcription factor regulating expression of HSPA8 and other heat shock proteins contains a functional variant associated with SZ pathology, which resulted in the reduced expression of HSPA8 in PM SZ brain samples.

It may also be that the findings by Prabakaran et al (2004) are due to false positives. For example, it is known that HSPA8 decreases in expression as a natural function of age. It is also known that SZ patients have a reduced life span of ~ 10 years. Therefore it may be possible that HSPA8 expression decreases more rapidly in the brains of SZ patients due to long-term drug therapy and that the significant decrease found by Prabakaran and colleagues is simply an artefact of these phenomena. Further biological study to address this issue is warranted. Given the speculation from the literature of oxidative stress involvement in SZ pathogenesis, and the positional and functional evidence presented in this chapter, HSPA8 warrants further investigation in larger sample sizes, other populations and additional biological studies.

5.5 Overview of GSTM3

The objective of my research on GSTM3 was to identify sequence variants across the gene and genotype them in my SZ case-control sample to determine if they were associated with SZ. As resequencing is expensive and as my PhD studies require me to learn as many relevant techniques as possible, I carried out mutation detection analysis using denaturing

high performance liquid chromatography (DHPLC) to identify gene regions likely to contain sequence variants. These regions were then sequenced to confirm and identify the variant. As in the study of HSPA8, I used a sample of 15 SZ cases to detect sequence variants. Only one SNP (rs1332018) was identified as polymorphic across GSTM3. In addition to the one SNP identified by DHPLC and sequencing, a further two SNPs were selected from dbSNP for association analysis based on their known informativity in Caucasian samples. These 3 SNPs were genotyped commercially in the full SZ case-control sample by Kbiosciences. I performed association analysis across GSTM3 using the resultant genotypes.

I will describe my work on GSTM3 in the following sections: materials and methods for design of fragments for DHPLC; materials and methods for PCR optimisation; materials and methods for DHPLC; results from DHPLC analysis of GSTM3; materials and methods sequencing of GSTM3 PCR fragments; results from sequencing of GSTM3 fragments; materials and methods for genotyping of GSTM3 SNPs; results from association study of GSTM3.

5.6 Results for GSTM3

5.6.1 Results of DHPLC for GSTM3

Three fragments showed DHPLC chromatogram traces that suggested sequence variants: fragments 7, 10 and 13. The DHPLC chromatogram result for fragment 7 showed a probable presence of a polymorphism whereas fragments 10 and 13 were only vaguely suggestive of containing polymorphisms. Fragment 7 contains exon 2, fragment 10 contains exons 5 and 6, and fragment 13 contains part of exon 9. Two patient samples (10 and 11) were selected as possible heterozygotes. Both showed chromatogram shifts for fragment 7 at 60°C and 65°C but not at 67°C. Two other samples that did not show shifts in the chromatogram were selected as homozygotes (patients 1 and 9) (see Figure 5.5). Six patients showed chromatogram shifts at 60°C, but not at 62°C. Patient samples 9 and 10 showed chromatogram shifts for fragment 10 at 60°C, but not at 62°C. These were classed as possible heterozygotes. Two other samples that did not show shifts in the chromatogram were selected at random as homozygotes (patients 1 and 8). Patient samples 1 and 2 showed chromatogram shifts for fragment 13 under at 60°C but not at 62°C conditions. These were classed as possible heterozygotes. Two other samples that did not show shifts in the chromatogram were selected at random as homozygotes (patients 3 and 4). Therefore all three fragments underwent resequencing with two chosen heterozygotes and two homozygotes in order to identify or rule out the possibility of a sequence variant being present.

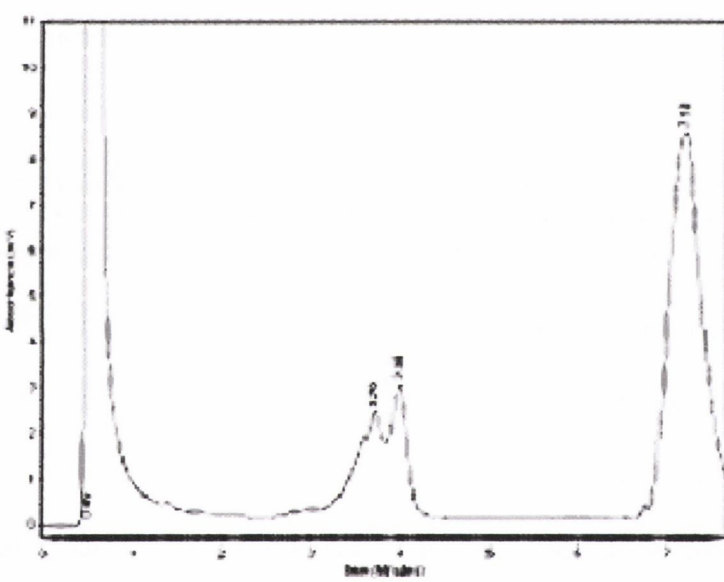
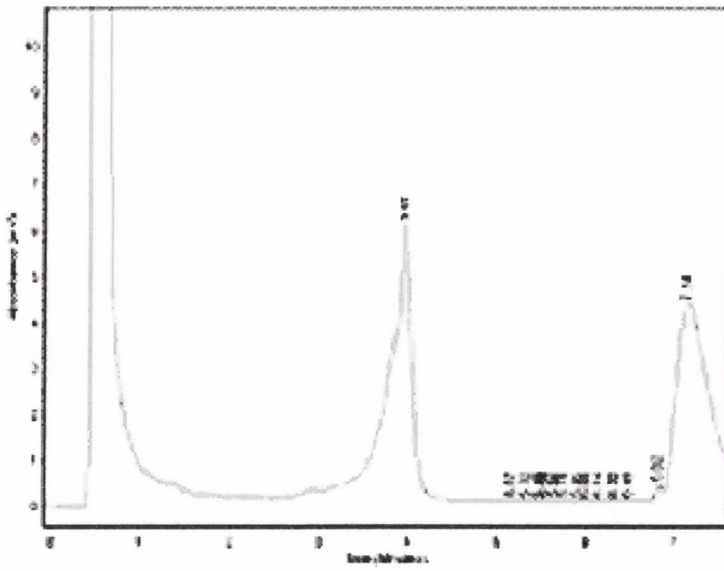


Figure 5.5 DHPLC analysis of GSTM3 fragment 7 at 60°C. Six samples showed only one peak indicative of no polymorphism (Patient 1). A number of other samples showed the presence of two distinct peaks (as in Patient 10 in the lower figure) indicative of a polymorphism in the sequence.

5.6.2 Results – Resequencing of GSTM3

DHPLC analysis suggested the presence of sequence variants in three GSTM3 fragments. DHPLC chromatograms from fragment 7 suggested a probable mutation was present in this region whereas DHPLC chromatograms from fragments 10 and 13 suggested possible mutations were present in these fragments. Two putative heterozygotes and two homozygotes were selected for resequencing. Fragment 7 identified a SNP mutation located at 296 bp, which equated to 28365 bp (on contig gi#15131467) that was used to design PCR fragments. In order to identify if this SNP was already deposited in the dbSNP database I blasted the first 295 bp preceding the SNP against the contig (NT_019273) used to construct information for GSTM3 on dbSNP NCBI. This identified SNP rs1332018 (G/T) at position 6369087 bp (dbSNP contig NT_019273). For the purposes of this experiment, rs1332018 was renamed GSTM3-1. Resequencing analysis of fragments 10 and 13 revealed no detectable variants in either respective forward or reverse reactions. The following figure 5.6 shows an electropherogram of GSTM3-1.

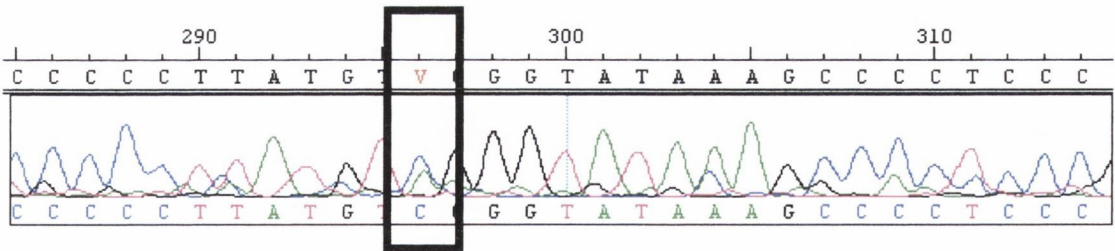


Figure 5.6: Electropherogram of the forward primer reaction of fragment 7 amplified in patient 10. A SNP mutation was identified at base pair 296 (highlighted by arrow). The polymorphism was a C/A. Although the reverse reaction did not work very well, this polymorphism was also evident.

5.6.3 Results – Association Study of GSTM3

Only one SNP (GSTM3-1, rs1332018) was found in the entire gene using DHPLC and resequencing techniques,. There are several possible reasons for this relatively low number of SNPs found and this issue is expanded upon in the Discussion below (section 5.7). Therefore we selected a further three SNPs (GSTM3-2 rs1537236, GSTM3-3 rs4970737 and GSTM3-4 rs7483) based on location and MAF values > 0.1 from European samples deposited in dbSNP to be genotyped in our full association sample. However, they were unable to genotype GSTM3-2 (rs1537236), using either Taqman or their own in-house method. GSTM3-2 lies in the untranslated region of GSTM3 (24365 bp on contig gi#15131467 – 122 bp downstream of fragment 14). This SNP was subsequently dropped from analysis. The remaining three SNPs (GSTM3-1, -3 and -4) were genotyped in the full sample. The results are presented in the next section (Table 5.3).

Table 5.3 Individual marker association results

dbSNP ID	Alternative marker name	Polymorphism	Frequency of Cases ^a	Frequency of controls ^a	p value
rs1332018	GSTM3-1	G/T	0.421	0.427	0.783425
rs4970737	GSTM3-3	C/G	0.729	0.741	0.532369
rs7483	GSTM3-4	A/G	0.313	0.295	0.384267

^a Frequency of first allele listed

All markers were genotyped in the entire sample

The three SNPs (GSTM3-1, GSTM3-3 and GSTM3-4) all showed no association with SZ in the Irish population. In addition, 2 and 3 marker haplotype analysis was performed using FASTEHPPLUS and GENECOUNTING (see section 2.7.3). This also showed no association with SZ in the Irish population. The results of this negative finding are discussed in the next section.

5.7 Discussion

This is the first association study of GSTM3 and SZ in any population. GSTM3 is normally involved in the oxidative stress response but was significantly down regulated in PM SZ brain grey matter (Prabakaran et al 2004). It is also located on chromosome 1p13.3 that is located in a significant linkage bin reported in the recent SZ meta-analysis paper by Lewis et al (2003). Therefore it is both a positional and functional candidate gene.

My original aim was to identify all SNP variation spanning GSTM3 in order to determine LD structure in the Irish reference panel and choose tSNPs for genotyping in the full association sample. As my PhD required that I learn as many skills as possible and I had previously learnt resequencing using HSPA8, I undertook mutation detection using DHPLC to identify fragments harbouring SNPs in the Irish population. This identified one probable fragment and two possible fragments that contained sequence variants. All three fragments were resequenced using the method outlined in section 2.9.5. However, only one SNP in fragment 7 was identified during re-sequencing. As GSTM3 is 4.5Kb long and my region of analysis extended into the promoter region of the gene, one might expect more than one common polymorphism to have been identified. The fact that only one SNP was detected was a worry and raised concerns that method may not have worked in full. At the time of my study other investigators were also experiencing difficulties with the WAVE DNA Fragment Analysis System used for DHPLC analysis and it may not have been functioning properly. With such high user traffic, buffer had to be replenished by different individuals weekly. As the buffer was very sensitive to concentrations, this practice may have introduced systematic errors, resulting in a malfunction. Retrospectively, it may have been prudent to have one technician preparing buffers and running all users plates, thereby limiting the introduction for errors.

Unfortunately, due to time constraints on the project and funding, subsequent repeated attempts or undertaking a full re-sequence of the gene was not attainable. Instead, I used the information from NCBI's dbSNP to select a further three SNPs across the gene for association analysis. However, for reasons mentioned in section 2.9.6 only three SNPs were genotyped. No association between SZ and GSTM3 was found with single marker analysis or haplotype analysis.

The interpretation of the negative finding of HSPA8 discussed in section 5.4 is also relevant here. However, GSTM3 expression does not decrease with age. It is possible that I missed an association by only genotyping a limited number of markers. There were several areas of the gene that were not analysed using DHPLC because they showed homology with chromosome 19 and amplifying these regions would have been problematic.

It is also possible that the gene expression study (Prabakaran et al 2004) may have had a false positive result with GSTM3. There are several caveats to using PM tissue. Individual expression changes are usually modest and variability between individuals is high, which makes it difficult to distinguish true physiological differences from normal human variation (Prabakaran et al 2004). Good quality PM tissue is hard to achieve as events such as apoptosis of cells immediately occurs after death. The time between death and collection of the tissue is critical. In addition, mRNA levels may be affected by chronic medication use (Harrison and Weinberger 2005).

It is also possible that a transcription factor regulating GSTM3 expression contains a functional mutation that interferes with the normal response in oxidative stress. It is now widely known that the basic leucine transcription factor NRF2 is a transcription factor for

GSTM3. Therefore the next step in this project was to investigate NRF2 for its putative role in SZ pathobiology (Chapter 6).

Chapter 6 – NRF2

6.1 Introduction

NFE2L2 or NRF2 (NUCLEAR FACTOR ERYTHROID 2-LIKE 2) is found at chromosome 2q31.2 (OMIM 600492). It is a member of a family of basic leucine transcription factor genes along with NFE2 on chromosome 12 and NFE2L1 on chromosome 17. Although this family of genes are located on different chromosomes, they share high homology suggesting that they probably derived from a single ancestor by chromosomal duplication. Interestingly, other gene families such as the collagen, integrin, and HOX genes also map to the same regions of the 3 chromosomes. Several animal models have shown that NRF2 is expressed in the liver, lung and GI tract and most other tissues. However, until recently it was not known if it was expressed in the brain.

A recent study by Lee et al (2003) showed that NRF2 regulates expression of antioxidant response element (ARE) driven genes in astrocytes. The ARE (5'-TGA[C/T]NNNGC-3') is a cis-acting element governing the regulation of multiple phase 2 genes encoding proteins that protect against oxidative and electrophilic stresses (Kwak et al 2003). There are several studies that show NRF2 deficient mice do not respond well to experimental induced oxidative stress (Itoh et al 1997, Kwak et 2001, Shih et a 2005). A paper by Chanas et al (2002) showed that NRF2 is a transcription factor for glutathione S-transferases (which includes GSTM3, covered in chapter 5). GSTM3 was among many oxidative stress and mitochondrial genes found to be significantly down-regulated in PM SZ white matter (Prabakaran et al 2004). Therefore it is plausible that a functional mutation in NRF2 impedes transcription of genes such as GSTM3, leading to their down-

regulation. My hypothesis is that as NRF2 regulates genes that respond to oxidative stress and Prbakaran et al (2004) have demonstrated reduced expression in SZ, NRF2 may contribute to susceptibility to this disorder (Figure 6.1 (a) and (b)).

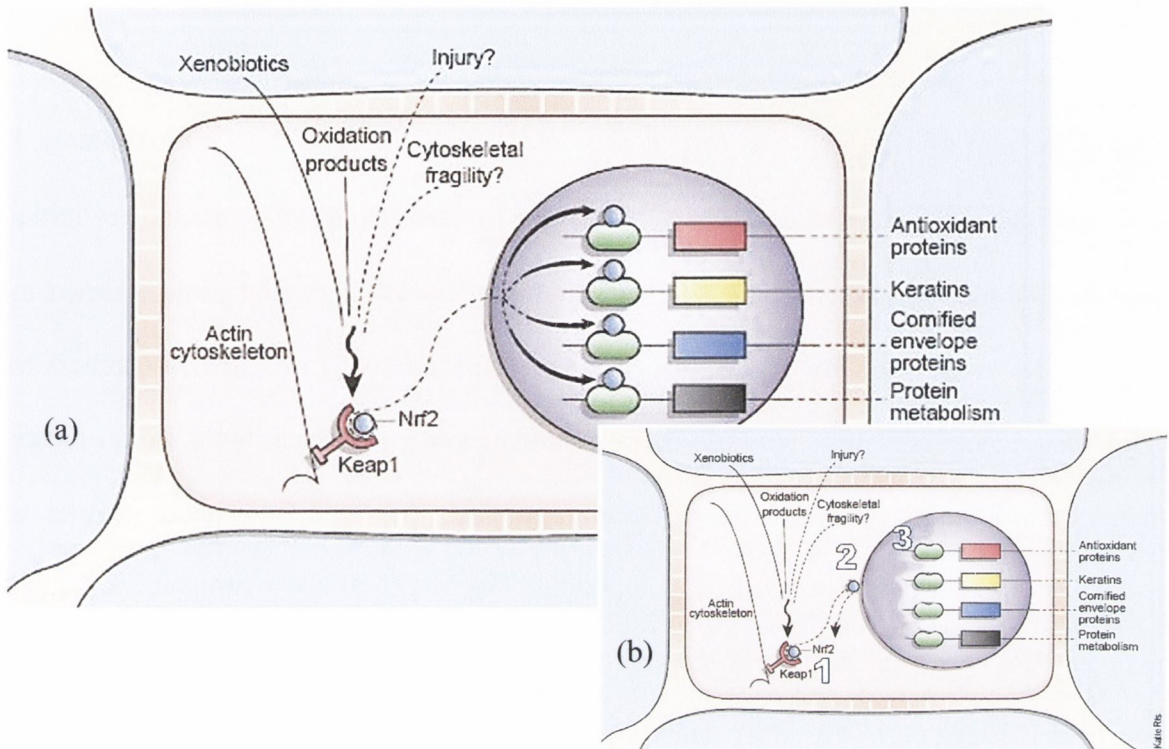


Figure 6.1 Mode of action NRF2. (a) Cells undergo many stresses including oxidative stress. The main diagram above shows the sequence of events involved in releasing NRF2 into the nucleus to initiate transcription. NRF2 is initially bound to KEAP1 which represses its function. When the cell is stressed from injury, xenobiotics or reactive oxidative species, KEAP1 releases NRF2. The NRF2 then freely enters the nucleus and with the aid of a MAF protein initiates transcription of many proteins including antioxidants in order for the cell to survive. **(b) Inset.** If however, NRF2 harbours a functional mutation it is possible that (1) it will not be able to detach from its repressor KEAP1 (see discussion) (2) have the ability to enter the nucleus or (3) may enter the nucleus but is unable to bind with the ARE to initiate transcription. (Adapted from Magin 2003, Nature Genetics)

In addition to investigating NRF2 as a putative susceptibility gene for SZ, I also took advantage of the relatively dense SNP data from the NRF2 locus in HapMap to study patterns of genetic variation between the CEPH sample used in HapMap and the Irish Reference Panel (described in chapter 5). The HapMap project aims to determine the common patterns of DNA sequence variation in the human genome by characterising sequence variants, their frequencies and correlations between them in 269 DNA samples

from African, Asian and European origin populations (HapMap 2003). The goal of the International HapMap project is to help researchers across the world discover genetic factors that contribute to disease susceptibility, protect against illness and impact on drug response by developing HapMap as a research tool.

Initially the aim of HapMap was to genotype 600,000 SNPs spaced at approximately 1 SNP per 5 Kb, with a MAF ≥ 0.05 (HapMap 2003). This would give researchers initial genomic LD and tagging SNPs (tSNPs) information. Subsequently the project aimed to increase SNP marker density to 1 SNP per 1 Kb (HapMap 2005) to allow researchers to carry out indirect association studies on any candidate gene, positional locus from linkage studies or to conduct genome-wide association studies. As one of the main samples is descended from European ancestry, it is hoped that HapMap will allow researchers to pick tagging SNPs based on the closest relevant population and perform association studies in their local European sample, while retaining maximum power for detecting association. At the time of this experiment, HapMap had published Phase 1 data (1 SNP per 5 KB). It was thought that the Phase 1 density of SNP markers might be too low to capture all common SNP variation (i.e. MAF $> 10\%$). My second hypothesis was to test if SNP marker density in the Phase 1 HapMap was currently at a useful level to allow the selection of tSNPs for association studies in the Irish population.

6.2 Results

The results in this chapter will be presented in 3 sections. Section 6.3.1 will detail my investigation of LD structure in the CEPH and Irish samples. Section 6.3.2 will describe how efficiently tag SNPs selected in the CEPH reference panel capture other SNPs in the

region. Section 6.3.3 will details the results of the association study of NRF2 in the Irish SZ case-control sample.

6.2.1 LD comparison of Irish and CEPH samples

I had data on 11 SNPs that were genotyped in both the CEPH sample and the Irish reference panel. Details of the MAF for each SNP in each sample are given in Table 6.1, and graphically represented in figure 6.2. MAF in both samples are similar. The largest difference in MAF is for NRF2-9, rs1806649 where there is a 10% difference in allele frequency. Overall, the MAF's are highly correlated ($r^2 = 0.9354$).

LD was measured using D' and r^2 between all possible pairwise combinations of SNPs in both samples. The results of the comparisons of the D' data from the two samples are displayed in figures 6.3 and 6.4. The schematic at the top of the diagram in figure 6.3 represents the position of each HapMap SNP in relation to the gene. Although the 11 SNPs span the length of the gene, none of them were located in exons. From the two Haploview outputs in figure 6.4 one can see that the Irish sample would appear to have less pairwise LD between SNPs compared to the CEPH sample. Figure 6.4 is a scatterplot of the D' measurements in the two samples. The correlation co-efficient ($r^2 = 0.1092$) indicates that the D' values are not highly correlated between the two samples. When I examined the results for the r^2 comparisons, the results were quite different. The results of the comparison between r^2 data from the two samples are displayed in figures 6.5 and 6.6. The schematic at the top of the diagram in figure 6.6 represents the position of each HapMap SNP in relation to the gene. From the r^2 Haploview output one can see that the Irish and CEPH sample have similar pairwise LD between SNPs. This is confirmed in a scatterplot

of the r^2 measurements in the two samples (Figure 6.6), which has a high correlation coefficient ($r^2 = 0.9369$), indicating that the r^2 values are highly correlated.

In order to further investigate the similarities in the LD structure between the CEPH and the Irish samples, I compared the set of tag SNPs chosen by Tagger in the CEPH data to the set of tag SNPs chosen by Tagger from the Irish data. From the CEPH data, Tagger chose the following 6 tag SNPs, with the additional SNPs captured by each tag in parenthesis; NRF2-1, NRF2-5 (NRF2-15), NRF2-8 (NRF2-4, NRF2-6, NRF2-16, NRF2-17), NRF2-9, NRF2-13, NRF2-18. By the term ‘capture’ I mean the following: A SNP at position 1 that has have a high correlation ($r^2 \geq 0.8$) with another SNP at position 2 will be able to capture the genotypic information of the SNP at position 2, effectively tagging it. From the Irish data, Tagger chose the same 6 tag SNPs along with 1 additional tag SNP; NRF2-4 (NRF2-6). This additional SNP was chosen by Tagger as the r^2 values in the Irish sample for NRF2-8 – NRF2-4 ($r^2 = 0.787$), and NRF2-8 – NRF2-6 ($r^2 = 0.728$) were below the $r^2 = 0.8$ Tagger threshold. Had the r^2 threshold been set at 0.7, Tagger would have chosen the same set of tag SNPs from both reference panels. Overall, this highlights the high similarity in LD structure at this locus between the CEPH and Irish reference panels.

Table 6.1: MAF comparison of CEPH and Irish samples for 11 SNPs

SNP Name	rs number	Alleles	MAF CEPH ^a	MAF Irish ^a	Allele frequency difference
NRF2-1	rs2706110	A/G	0.21	0.25	0.04
NRF2-4	rs2001350	G/A	0.10	0.08	0.02
NRF2-5	rs6726395	A/G	0.47	0.45	0.02
NRF2-6	rs1962142	T/C	0.10	0.08	0.02
NRF2-8	rs4243387	C/T	0.10	0.08	0.02
NRF2-9	rs1806649	T/C	0.31	0.21	0.10
NRF2-13	rs2364722	G/A	0.29	0.33	0.04
NRF2-15	rs2364725	G/T	0.47	0.47	0.00
NRF2-16	rs2364727	T/C	0.10	0.08	0.02
NRF2-17	rs1806686	C/T	0.10	0.08	0.02
NRF2-18	rs2364731	G/A	0.49	0.41	0.08

^a Allele frequency represents the first allele listed

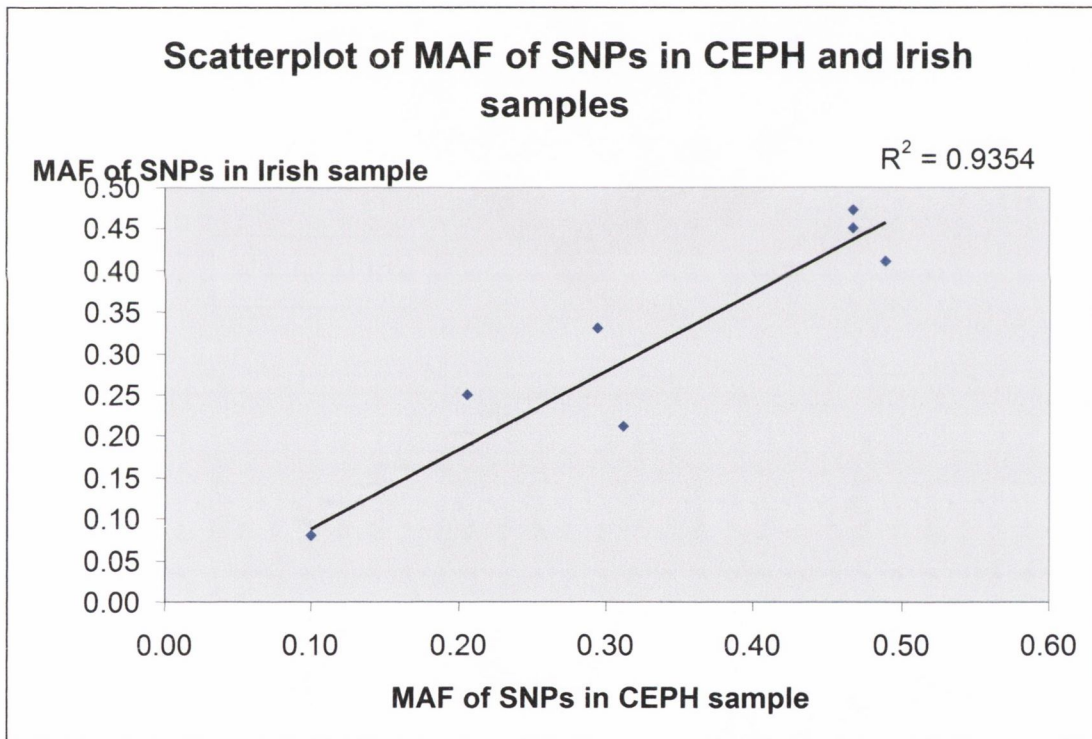


Figure 6.2: Scatterplot of the MAF of SNPs in the CEPH and Irish samples

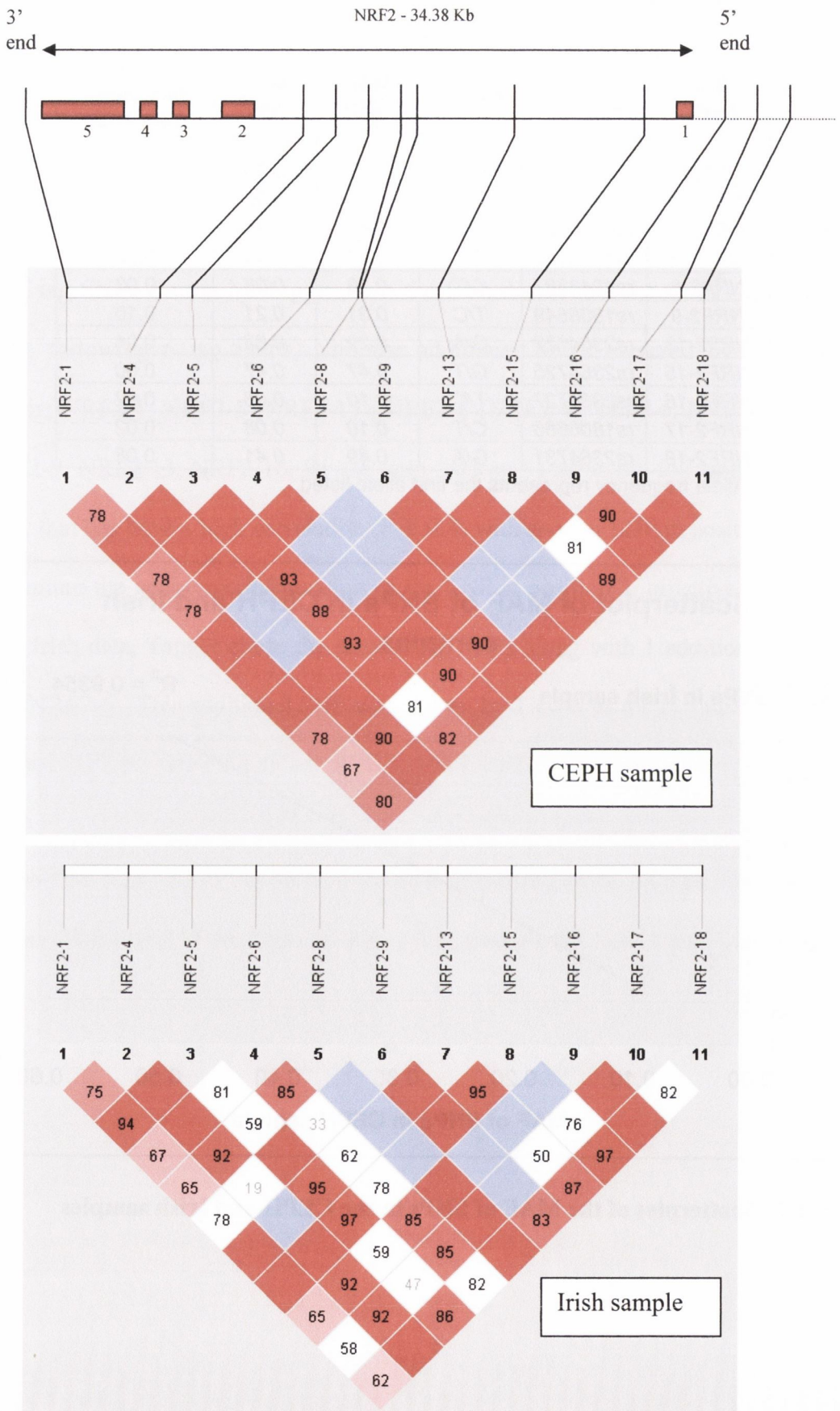


Figure 6.3: D' comparison between the CEPH sample and the Irish sample

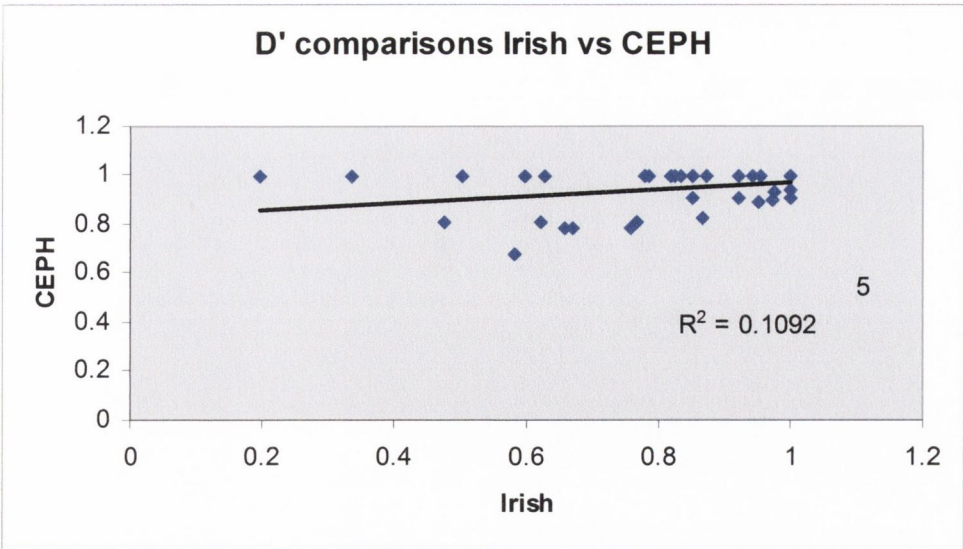
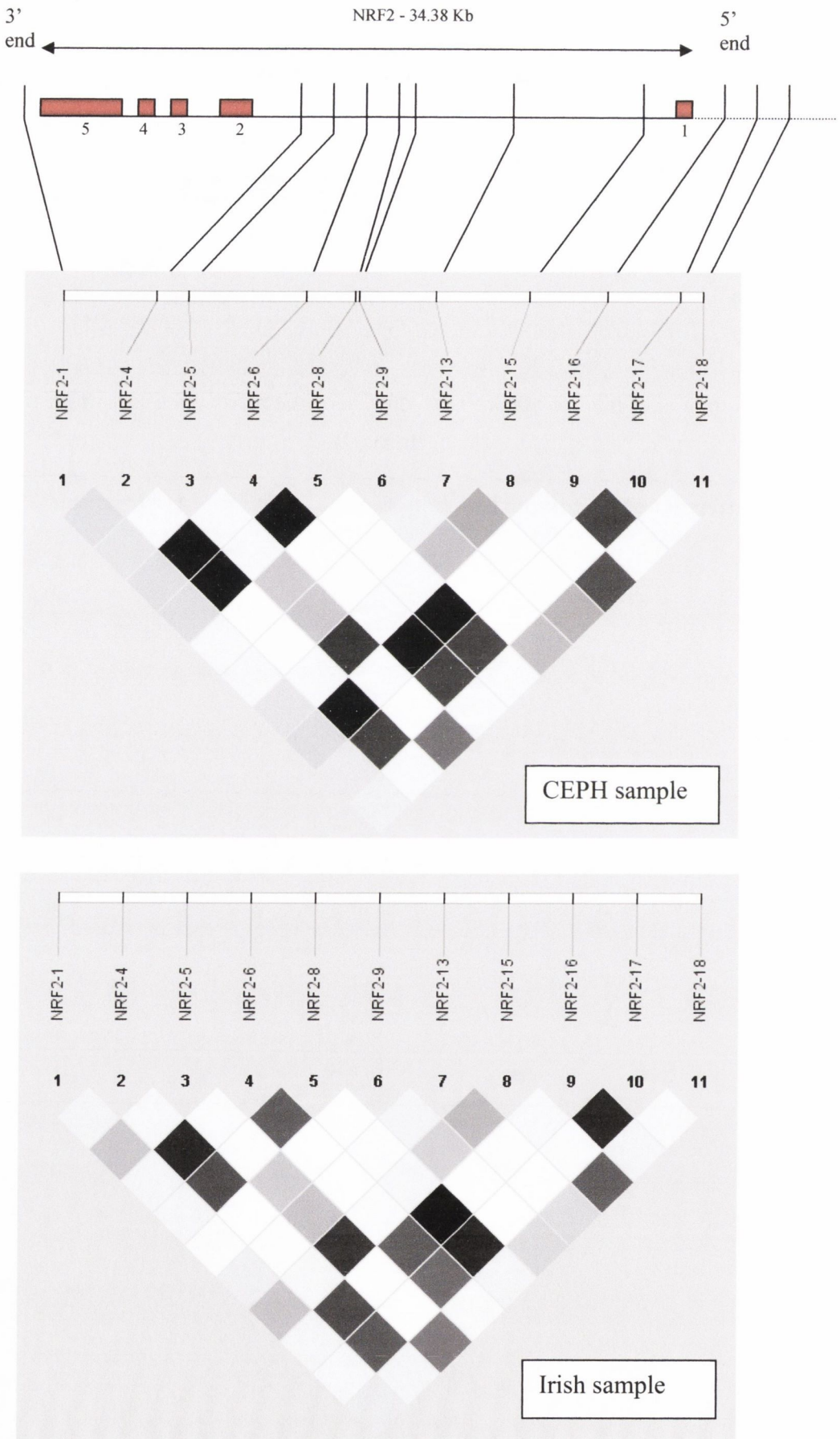


Figure 6.4: Scatterplot of D' comparison between the CEPH and Irish samples



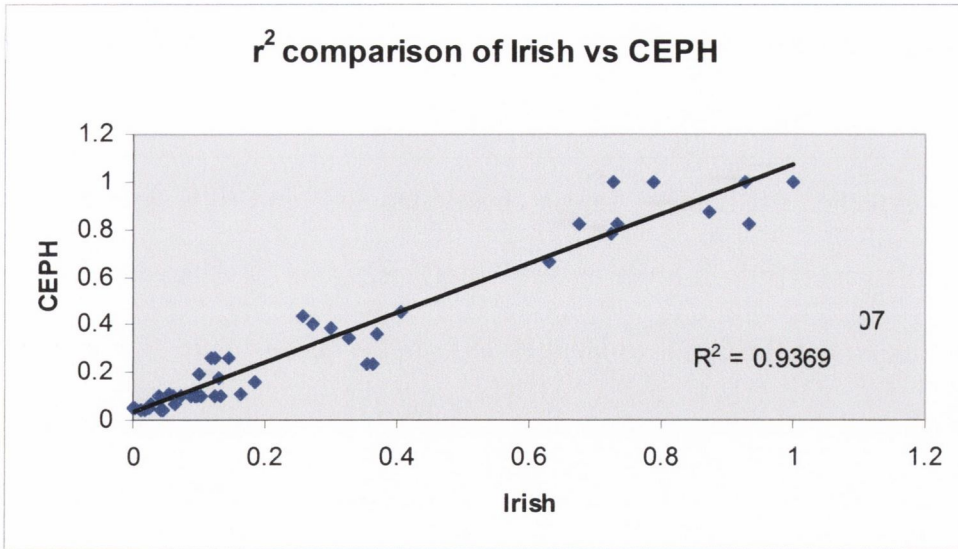


Figure 6.6: Scatterplot of r² comparisons between the Irish and CEPH sample

6.2.2 Efficiency of CEPH tag SNPs in capturing additional SNPs at NRF2

As outlined in section 6.3.1 above, Tagger chose 6 tag SNPs in CEPH samples from the 11 HapMap SNPs genotyped. In order to test the efficiency of these tag SNPs in capturing additional variation at NRF2 (i.e. to identify SNPs in the gene that are highly correlated ($r^2 \geq 0.8$)), I evaluated the performance of these 6 tag SNPs in tagging the additional 5 SNPs at NRF2 chosen from dbSNP. This task was performed using genotype data from the Irish reference panel for the extended set of 16 SNPs across NRF2 (11 HapMap and 5 dbSNP; figure 6.7). LD parameters in Tagger were as before ($r^2 = 0.8$ and $\text{LOD} = 3$). Table 6.2 gives a breakdown of tag SNP selection in the Irish panel. The 11 HapMap SNPs are highlighted in blue, the 5 dbSNP SNPs are highlighted in orange. The original 6 tag SNPs (highlighted in bold in the first column of table 6.2) chosen from the CEPH data were ‘force included’ for analysis, thereby insuring that they were again chosen as tag SNPs in the Irish reference panel. These 6 tag SNPs were not sufficient to capture all 16 SNPs. Only 1 of the 5 new dbSNPs (NRF2-11) was tagged by one of the 6 tag SNPs. Two additional tag SNPs (NRF2-3 and NRF2-7) were required to capture the remaining SNPs. Therefore, if I had used the HapMap data to choose tag SNPs for an association study in the Irish sample, I would not have captured all the variation at NRF2 (i.e. all SNPs, HapMap and dbSNP, used in this study). In total, 8 tag SNPs were required to capture all 16 SNPs genotyped in the Irish reference panel at NRF2.

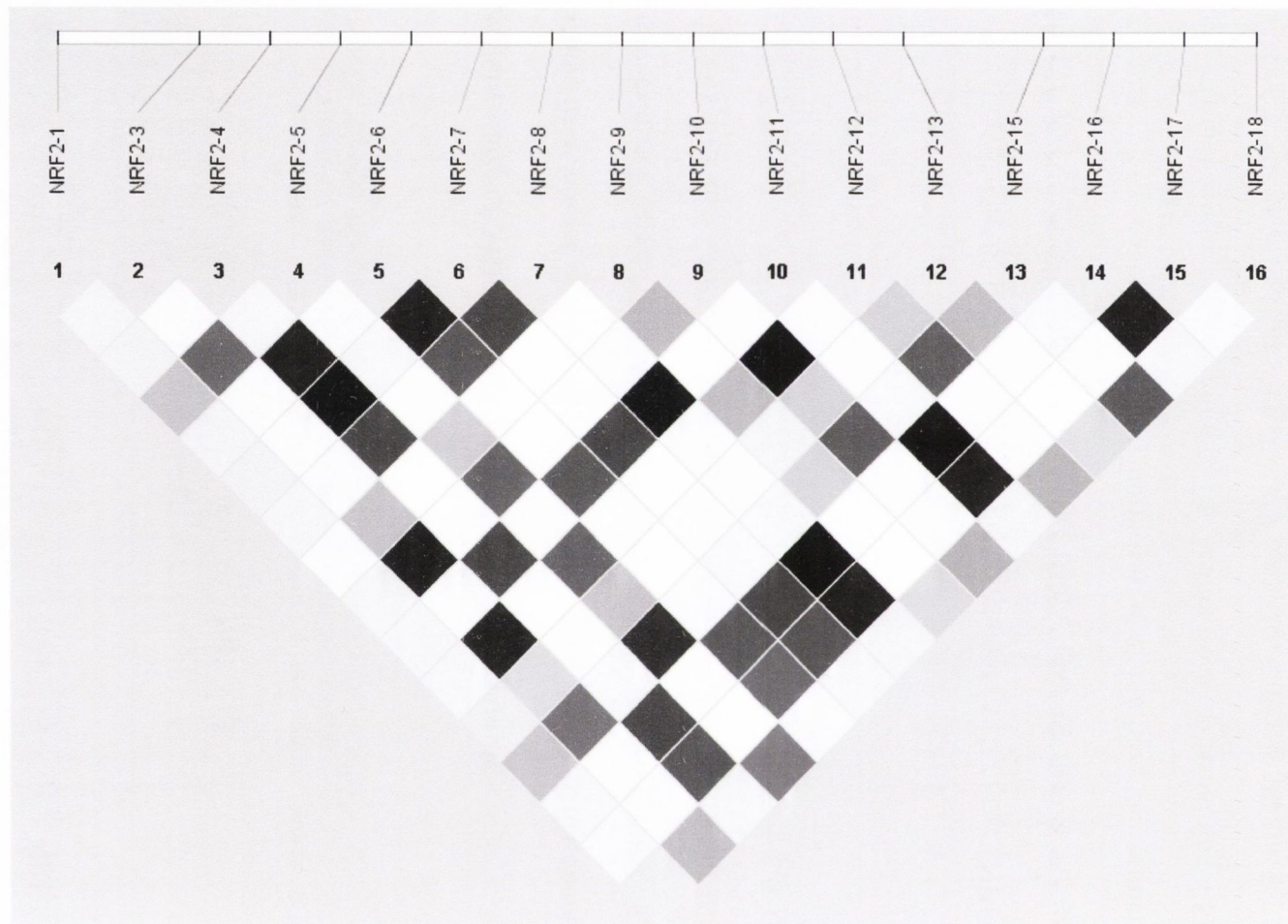


Figure 6.7: r^2 output from Haploview of all 16 SNPs (11 HapMap and 5 dbSNP) that span NRF2

Table 6.2: Alleles captured ($r^2 > 0.8$) in the Irish sample (16 SNPs)

Tag SNP ^a	Alleles Captured in the Irish sample (16 SNPs) ^b			
NRF2-1	NRF2-1			
NRF2-5	NRF2-5	NRF2-15		
NRF2-8	NRF2-8	NRF2-11	NRF2-16	NRF2-17
NRF2-9	NRF2-9			
NRF2-13	NRF2-13			
NRF2-18	NRF2-18			
NRF2-7	NRF2-4	NRF2-6	NRF2-7	
NRF2-10	NRF2-3	NRF2-10	NRF2-12	

^a Dark blue tag SNPs were HapMap CEPH tag SNPs force included in the Irish sample. Orange tag SNPs were additional tag SNPs (from dbSNP) identified by Tagger using the Irish sample.

^b Light blue SNPs are HapMap SNPs. Light orange SNPs are additional SNPs (dbSNPs)

6.3.3 Association study of NRF2 in the Irish SZ case-control sample

Eight tag SNPs (first column of table 6.2) were selected from the 16 SNPs genotyped in the Irish reference panel across NRF2. Of these 8 SNPs, 6 SNPs (NRF2-1, NRF2-5, NRF2-9, NRF2-10, NRF2-13 and NRF2-18) were selected for genotyping in the full Irish case-control sample (n=299 cases, n=645 controls). The other 2 SNPs (NRF2-7 and NRF2-8) were not analysed due to their low MAF (< 0.1).

Of the 6 SNPs genotyped in the full sample, data was unavailable on 2 of the SNPs (NRF2-1 and NRF2-13) due to assay design failure. This failure may be due to the sequence surrounding these SNPs containing a polymorphism that inhibits the primers from annealing properly and continuing the reaction. The results from the remaining 4 SNPs are detailed in table 6.3. No SNP showed evidence of association with the phenotype. In order to try and capture unknown variants that could be associated at each gene (i.e. variants not analysed in the reference panel), I performed 2-, 3- and 4-marker haplotype analyses using all combinations of SNPs available (for haplotype analysis methodology). Haplotype analysis did not identify association between NRF2 and schizophrenia (Table 6.4).

Table 6.3 Individual marker association results

dbSNP ID	Alternative marker name	Polymorphism	Frequency of Cases ^a	Frequency of controls ^a	p value
rs6726395	NRF2-5	A/G	0.401	0.423	0.940
rs1806649	NRF2-9	C/T	0.767	0.754	0.513
rs13001694	NRF2-10	A/G	0.671	0.637	0.110
rs2364731	NRF2-18	A/G	0.586	0.590	0.846

^a Frequency of first allele listed

All markers were genotyped in the entire sample

Analyses of SNP genotyping data from either the Irish reference panel or the full case sample or the full control sample showed that all SNPs were in Hardy-Weinberg Equilibrium ($p > 0.05$). The reference panel of 92 control samples was part of our full case-control sample. Therefore, for the 4 SNPs in table 6.3 above I had genotype data on 92 samples generated by different genotyping technologies at independent sites. This allowed me to cross-validate genotype data to estimate genotyping accuracy. Out of 368 genotypes generated in duplicate there were no discrepancies, giving an estimated genotyping accuracy of 100%.

Table 6.4 Haplotype analysis of 2-, 3- and 4-markers using 10,000 iterations to provide an empirical simulated p value

Marker 1	Marker 2	Marker 3	Marker 4	p value
NRF2-5	NRF2-9			0.329
NRF2-5	NRF2-10			0.385
NRF2-5	NRF2-18			0.351
NRF2-9	NRF2-10			0.076
NRF2-9	NRF2-18			0.862
NRF2-10	NRF2-18			0.120
NRF2-5	NRF2-9	NRF2-10		0.146
NRF2-5	NRF2-10	NRF2-18		0.458
NRF2-5	NRF2-9	NRF2-18		0.369
NRF2-9	NRF2-10	NRF2-18		0.116
NRF2-5	NRF2-9	NRF2-10	NRF2-18	0.125

6.4 Discussion

NRF2 maps to chromosome 2q31.2 a region not previously reported as linked to schizophrenia. However, traditional methods of linkage analysis may lack power to detect the modest genetic effects that individual genes contribute to SZ risk. NRF2 is nonetheless a functional candidate gene for SZ. NRF2 was studied because it is a basic leucine transcription factor playing a pivotal role in cellular activation of antioxidant response elements (ARE), which initiate transcription of anti-oxidative stress genes. Several genes (including those studied in chapter 5) involved in anti-oxidative stress were shown to be significantly decreased in SZ PM brain in a recent gene expression paper (Prabakaran et al 2004). The hypothesis in this experiment was that some functional mutation in the NRF2 gene coded for a dysfunctional NRF2 protein that was unable to bind to the ARE in response to oxidative stress.

I carried out a two-stage design association study in a large Irish case-control SZ sample (n=299, n=645). The first stage involved determining LD structure across NRF2 using Phase 1 data from the HapMap project and from NCBI's dbSNP database. I identified 11 HapMap SNPs informative in the CEPH trios and five additional SNPs from dbSNP with a MAF > 0.1 in European populations. All 16 SNPs were commercially genotyped in the Irish reference panel (n=92 controls) identifying eight tag SNPs for genotyping in the full sample (stage 2). Two of these SNPs (NRF2-7 and NRF2-8) were not analysed due to their low MAF (<0.1). In addition, data is unavailable on two further tag SNPs due to assay design failure (NRF2-1 and NRF2-13). Therefore results are available for the remaining four tag SNPs (NRF2-5, NRF2-9, NRF2-10 and NRF2-18). The genotyping of these four tag SNPs showed no evidence of individual or haplotypic association of NRF2 with SZ in the Irish population.

There are several interpretations of this result. First of all, it could be that this is a true negative finding. The prior probability of any selected candidate gene being involved in SZ pathology is low because we do not understand the aetiology of SZ and that that almost every gene in the human genome is expressed in the brain at some time during development. Although NRF2 is expressed in astrocytes *in vivo* and *in vitro*, this was not one of the genes identified as being differentially expressed in PM SZ brains from the Prabakaran et al (2004) study.

Second, the proposed mechanism of ARE activation could contribute to SZ but involving different genes. The literature shows that NRF2 is only one of many genes involved in ARE activation. From Figure 6.8 (overleaf), you can see that NRF2 is sequestered by KEAP1 and ubiquitin. When NRF2 dissociates from KEAP1 it is in its active form and free to travel to the nucleus and bind with the ARE. In addition, other factors upstream in this complicated pathway stimulate release of NRF2 from KEAP1. Therefore it is possible that any of these factors in the pathway leading to binding of ARE and transcription of anti-oxidant proteins may harbour risk variants (see Figure 6.8) with the result of significantly altered expression of genes involved in oxidative stress response.

The results of this study may represent a false negative report in failing to identify association between NRF2 and SZ. The study had 67% power to detect a risk variant with OR=1.5 and MAF \leq 0.1 at 95% CI. To have 80% power at these parameters and maintaining an $\alpha = 0.05$ I would have required a case sample size of 407 SZ patients and 879 controls. The Irish sample has >98% power to detect a risk variant with an OR \geq 2. However, in reality, putative susceptibility genes identified to date report OR typically < 1.3. Therefore there is the possibility that this study was lacking power. .

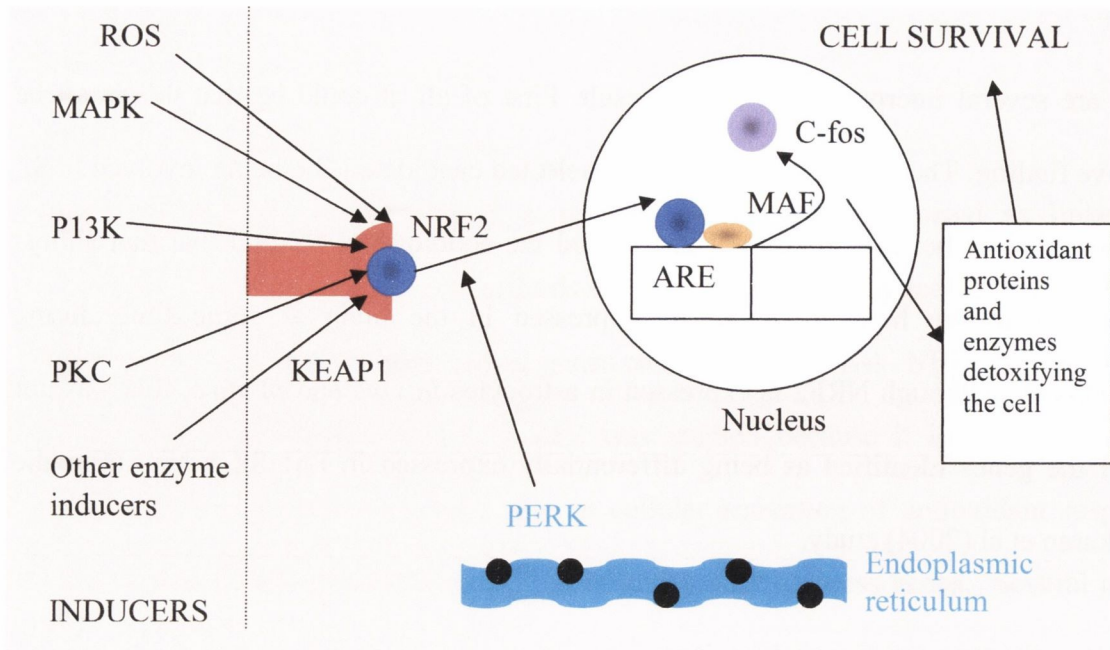


Figure 6.8: This schematic represents external cellular factors (Reactive oxidative species, ROS) and other proteins (MAPK, P13K and PKC) involved upstream in the dissociation pathway of NRF2 from KEAP1. Once free, NRF2 freely travels into the nucleus with the aid of PERK from the Endoplasmic Reticulum. NRF2 then binds with the ARE and a small Maf with the simultaneous release of C-fos. The NRF2/ARE complex then initiates transcription of genes involved in oxidative stress response. This schematic highlights the many possibilities that could be involved in the resultant altered expression of anti-oxidative proteins in post mortem SZ brains. Diagram adapted from (Itoh and Yamamoto 2004, Kwak et al 2004 and Nioi and Hayes 2004).

There is also the issue of how much variation was captured. Although Tagger identified 8 tag SNPs, the approach was to investigate common variation, defined as (MAF \geq 0.1). Therefore 6 SNPs were included in the analysis. Furthermore, two additional SNPs failed the commercial assay design and my attempts to genotype these markers using an alternative assay design (Taqman assay, Applied Biosystems). Within the parameters described for the Tagger software ($r^2 = 0.8$, LOD = 3), this study captured information on 7 of 16 SNPs. (See Table 6.2). Of these 7 SNPs, one was located in the 5' region (NRF2-18) with the remainder located in intron 1. It is possible that through the loss of information on 9 SNPs, this may explain a false negative. Although my choice of markers was based on common SNPs from public databases, not all common variation was tested. However, despite increasingly economic and higher throughput platforms for genotyping re-sequencing of potential functional candidate genes (such as NRF2) is not currently economically viable.

Taken together, this study fails to support the hypothesis that NRF2 is involved in SZ susceptibility. However, it is likely that studies at greater marker density and larger sample sizes will be required to definitively exclude smaller genetic effects at this locus.

The second hypothesis of this study was to see if the Phase I data of HapMap (which was current at the time of this study) was at a useful level to allow selection of tSNPs for association studies in the Irish population. The Phase I data aimed to have 1 SNP (with $MAF > 0.05$) per 5 Kb across the genome. This was successfully carried out (International HapMap Consortium, 2005) and the Consortium has evaluated its performance. They found that LD structure was high, with $> 75\%$ of SNPs being highly correlated ($r^2 \geq 0.8$) with one or more SNPs. Based on simulations using the ENCODE regions, Phase I currently covers 74% of all common variation ($r^2 \geq 0.8$). The Consortium's paper shows that to capture all common SNPs ($r^2 \geq 0.8$) in Phase I (CEPH) requires 290,000 tSNP. To capture 94% of all common variation in the Phase I (CEPH) sample with an $r^2 \geq 0.8$ requires 250,000 tSNPs. Therefore Phase I data already provides excellent power in the CEPH sample. HapMap is now in Phase II which aims to increase the density by genotyping an additional 4.6 million SNPs. Modelling Phase II data on ENCODE regions, it is predicted that with an $r^2 \geq 0.8$, 94% of all common variation will be captured. This will increase power and efficiency further.

One of the main uses of HapMap is to select tSNPs for association studies. The aim is to maximise efficiency of genotyping while minimising loss of information by exploiting redundancy amongst SNPs. The SNP density of Phase I data was an average of 1 SNP per 5 Kb and simulation tests using ENCODE data suggest that all SNPs genotyped in Phase I will be highly correlation ($r^2 \geq 0.8$) with at least one other SNP. Therefore, ignoring issues of marker density, Phase I has achieved the goal of reducing genotyping and maximising

efficiency, providing a provisional genome-wide association study marker set. However, as this data is only available in the CEPH sample, it does not tell us whether HapMap will be useful for inferring tSNPs in the Irish population. Therefore I had to compare LD structure and tSNP selection between the CEPH and Irish populations. Although there are several measures in use for LD analysis, I used the two most common methods, D' and r^2 . Both methods are based on the basic pairwise-disequilibrium coefficient (Lewontin's D) but measure different things and hold very different properties. I will describe my findings on D' first and then my findings on r^2 .

D' is calculated by dividing D by its maximum possible value, given the allele frequencies at the two loci. Therefore if two SNPs have not been separated by recombination, recurrent mutation or gene conversion during the sample's history this leads to a $D'=1$ (complete LD). Any values of $D' < 1$ indicates that the ancestral LD has been disrupted. However D' values < 1 have no clear interpretation.

It is known that D' values can be inflated in small sample sizes. The CEPH sample only had 60 unrelated individuals whereas the Irish sample has 92 unrelated individuals. Therefore our sample is $\sim 50\%$ bigger. On inspection of the Haploview D' output there seemed to be higher LD in the CEPH sample compared to the Irish sample. If the sample sizes had been the same this could have suggested that the Irish sample has had more recombination events than the CEPH sample since the ancestral mutations occurred. However in reality it is more probable that the apparent higher LD in the CEPH sample is due to inflation of D' values from a smaller sample size. In addition, it is known that estimates of D' are strongly inflated for SNPs with rare alleles. So, high D' values can be obtained even when markers are in fact in linkage equilibrium.

As per Figure 6.3 and 6.4, there were some differences in D' values between the samples, with D' values being lower in the Irish sample. The allele frequencies between the CEPH and Irish samples are highly correlated (Figure 6.2). However, in order to explain the low correlation of D' values between the CEPH and Irish samples, I analysed the D' raw data for a SNP with a high MAF in the Irish sample (NRF2-5 (MAF = 0.45)) and five SNPs with a low MAF in the Irish sample (NRF2-4, NRF2-6, NRF2-8, NRF2-16 and NRF2-17 (MAF = 0.08)). Essentially, I noted two things. When the same SNPs were compared in the CEPH sample the D' values were higher and less spread than in the Irish sample: In addition, both samples with SNPs having a $MAF \leq 0.1$ had higher D' values than SNPs with $MAF > 0.1$. This suggests that low MAF SNPs in both samples may have inflated D' values, leading to false calculations and the subsequent lack of D' correlation between both samples. It is also possible that there are true structural differences between both populations.

The measure r^2 has recently emerged as the measure of choice for quantifying and comparing LD in the context of mapping. It is the correlation of alleles at the two sites, and is calculated by dividing D^2 by the product of the four allele frequencies at the two loci. $r^2 = 1$ only if, there is no recombination between the two markers and have the same allele frequency. So although $r^2 = 1$ between markers 1 and 2, it may be < 1 between marker 1 and marker 3 based purely on allele frequency and is therefore not necessarily indicative of heavy ancestral recombination. One benefit of r^2 is that it shows much less inflation in small samples than does D' and is helpful in power calculations as the sample size increase in order to detect an effect is $1/r^2$ because it encompasses the effect of both D and all marker/disease allele frequencies.

It is well known that pairwise LD shows variability within populations. However, as suggested by Evans and Cardon (2005) it may be that the differences may be due to the measures themselves and not completely due to different ancestries. The dependence of D' and r^2 on allele frequencies may not be the best way to compare LD patterns between populations as different populations often differ in allele frequencies.

Evans and Cardon (2005) also suggest that individual measures of pairwise D' are likely to be of limited use in assessing LD structure between populations. From the lack of D' correlation between the CEPH and Irish data presented, this appears to hold true. My study also agrees with the statement by Evans and Cardon (2005) that pairwise r^2 values between two populations correlate much better and should be the choice method for comparing LD between populations.

To date, there have been several studies comparing the CEPH HapMap data with their own population (Evan and Cardon 2005, Willer et al 2006, Stankovich et al 2006). In the Evans and Cardon (2005) study they compared amongst others, D' and r^2 values between CEPH and UK samples along a 10Mb stretch on chromosome 20. There was a good linear correlation between r^2 values yet little correlation between the D' values. They suggest that the lack of D' correlation is in part, due to a 'ceiling effect', i.e. where $D' = 1$ in one population, it is relatively low in the other population, with the majority of points located in the upper right hand corner (see Figure 6.9 (a)). The Willer et al (2006) study compared CEPH HapMap samples with a Finnish sample along a 17.9Mb stretch of chromosome 14. Although they found small differences in SNP allele frequencies, they found a strong r^2 correlation between the CEPH and Finnish sample, concluding that the HapMap project will be of benefit to Finnish association studies (see Figure 6.9 (b) for Finnish vs. CEPH Haploview r^2 output). In the Stankovich et al (2006) study they compared Australian

samples of European decent with the CEPH sample along a 3.7Mb region of chromosome 6 and a 1.3Mb region of chromosome 10. Again they found that both sets of samples were highly correlated (242 tSNPs CEPH $r^2=0.97$, Tasmania $r^2=0.945$).

Therefore, although in comparison to the UK, Australian and Finnish studies, I only looked at a very small region of 34Kb spanning the length of NRF2, the r^2 and D' results obtained seem to concur with the larger studies. My data would suggest that Phase I information from HapMap would have been useful for selecting tSNPs for association studies in the Irish population. However, the density at that time would not have covered all common variation leading to possible missed associations. The increased density of markers used in phase II data of HapMap will capture most common variation ($MAF > 0.1$) in the Irish population. However, it is possible that there will also be population specific common SNPs that are not covered by HapMap with a $MAF > 0.1$. It will be important therefore, to bear this in mind as one possibility when explaining any future negative association study based on HapMap tSNPs. However, it is clear that HapMap will serve as a very useful tool in Irish population association studies.

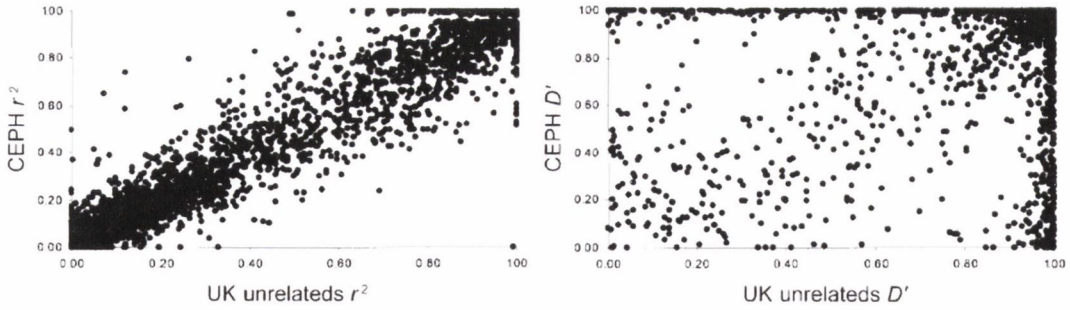


Figure 2 Comparison of r^2 (left) and D' (right) values between the CEPH and unrelated U.K. unrelated samples

(a) From Evans and Cardon 2005 AJHG

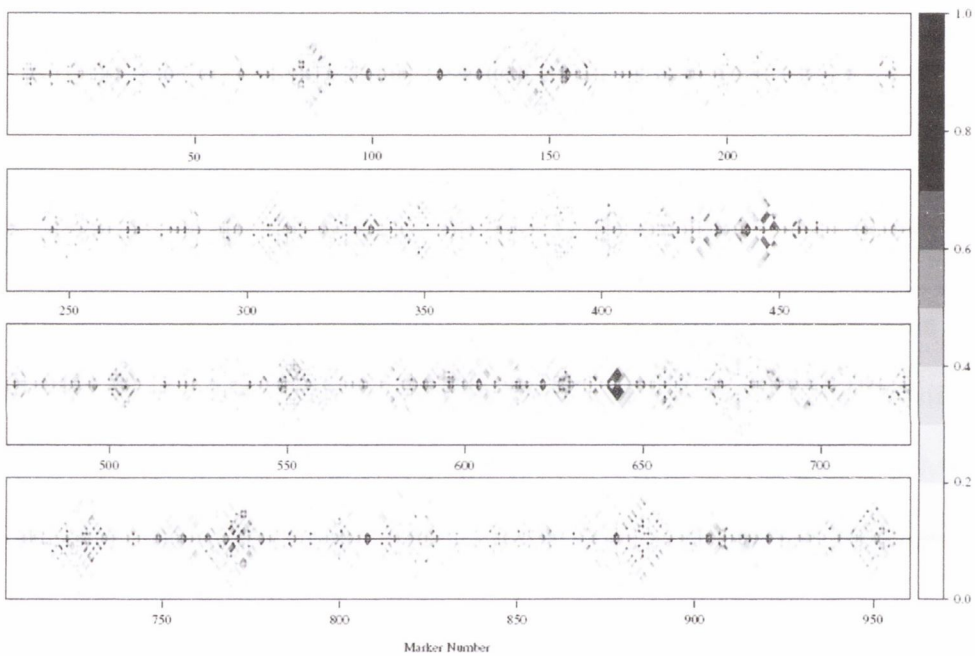


Fig. 2. Plot of estimated r^2 values for SNPs paired with each of the neighboring 20 SNPs in the Finnish (top segment of each row) and HapMap CEU (bottom segment of each row) samples. SNPs were selected based on estimated r^2 values in the HapMap CEU samples from a 17.9-Mb region on chromosome 14q.

(b) From Willer et al 2006, Genetic Epidemiology

Figure 6.9: Diagrams taken from two studies which compared LD structure in local populations with HapMap CEPH data. (a) r^2 and D' scatterplots comparing CEPH samples with UK samples. The D' has little correlations, whereas the r^2 does seem to suggest a linear correlation, similar to the Irish data (figure 6.6). (b) r^2 comparison of a 17.9 Mb region of chromosome 14 between Finnish (top segment of the row) and CEPH samples (bottom segment of the row). Again the LD measure looks very similar in both populations.

Chapter 7 - General Discussion

The overarching aim of the work undertaken over the course of this PhD was to investigate candidate genes, based on positional and functional evidence, for their putative role in schizophrenia (SZ) aetiology. All studies were carried out using an Irish SZ case-control association sample, initially consisting of 219 cases and 299 controls initially, and subsequently of 299 cases and 645 controls. The genes investigated during my studies were chosen on the basis of their relevance to two hypotheses of SZ pathogenesis; (i) the neurodevelopmental hypothesis of SZ (genes studied included the APOL family of genes, and G72 and DAAO), and (ii) the oxidative stress hypothesis of SZ (genes studied included HSPA8, GSTM3 and NRF2). Although the identification of genes involved in SZ pathogenesis has in general proved difficult, the field of SZ genetics has had some success in recent years with data emerging from several studies to suggest DTNBP1, NRG1, RGS4, DISC-1, 5HT2a and DRD3 as SZ susceptibility genes. The work described in this thesis has contributed to the literature with results from the study of the Irish SZ case-control sample indicating further support for G72 and DAAO as SZ susceptibility genes.

7.1 Review of main findings – Association Studies

7.1.1 Apolipoprotein-L 1-6

The Apolipoprotein-L (APOL) gene family were selected for investigation, as they were both positional (located on chromosome 22) and functional (Mimmack et al 2001) candidate genes for SZ with a possible role in neurodevelopment. A novel three-stage DNA pooling methodology was applied to the study of the six APOL genes. The DNA pooling

method was based on the method described by Norton et al (2003), but incorporated a novel variation to increase efficiency when correcting for pooled allele frequencies (McGhee et al 2005). The pooling experiments identified two SNPs, which met screening criteria for individual genotyping. However, after individual genotyping, neither marker met the criteria for statistically significant association with SZ ($p < 0.05$). This work did not support the involvement of the APOL gene family in SZ pathogenesis.

7.1.2 G72 and DAAO

G72 was a positional candidate gene identified by Chumakov et al (2002), and in vitro evidence indicated an interaction with DAAO. The purpose of this study was to investigate these genes for evidence of association with SZ in the Irish population. Subsequent replication studies (Williams et al 2003; Schumacher et al 2003; Addington et al 2003; Datta et al 2003; Bass et al 2003; Schumacher et al 2004; Wang et al 2004; Hall et al 2004; Korostishevsky et al 2004), identified SNPs in both genes associated with SZ susceptibility. Markers were selected across both genes based on data presented from these studies at the World Congress of Psychiatric Genetics in 2003 and in 2004. Five SNPs at G72 and four SNPS at DAAO were selected for investigation in the Irish SZ case-control sample. SNPs from both genes were found to be associated with SZ in the Irish sample (G72-M12, $p=0.005$, OR 1.34 95% CI; G72-M15, $p=0.011$, OR 1.31 95% CI; DAAO-M4, $p=0.018$, OR 1.3 95% CI; DAAO-M5, $p=0.003$, OR=1.43 95% CI). Data from this study replicates the findings of other studies (Chumakov et al 2002; Schumacher et al 2004; Wang et al 2004; Williams et al, personal communication) at these loci and provides further support for the involvement of G72 and DAAO in SZ pathogenesis.

7.1.3 HSPA8 and GSTM3

HSPA8 and GSTM3 were selected for investigation as they were both positional candidates (located at chromosomes 11q23.3-q25 and 1p13.3 respectively) and functional candidates (involved in mitochondrial function and oxidative stress response) for SZ. In addition, they were also found to be significantly down-regulated in SZ post mortem brains compared to controls (Prabakaran et al 2004). Neither of these genes had previously been studied as putative SZ genes. Mutation detection was performed across these genes using resequencing (HSPA8) and DHPLC (GSTM3) to identify common genetic variants. SNPs detected across these genes were genotyped in the case-control sample. No markers reached statistical significance criteria for association with SZ.

7.1.4 NRF2

The basic leucine transcription factor, NRF2, is a transcription factor for a number of genes, including GSTM3. NRF2 was chosen to investigate the hypothesis that genetic variation at NRF2 contributed to the reduced expression of GSTM3 mRNA found in post mortem SZ brains (Prabakaran et al 2004). In undertaking the study of NRF2 as a putative SZ susceptibility gene, the usefulness of linkage disequilibrium data produced by the International HapMap project in the context of designing genetic association studies in the Irish SZ case-control sample was also investigated. One of the main goals of HapMap is to identify genome-wide LD structure such that information on correlations between SNPs can be used to increase study efficiency and cost.

At the time this study was conducted, Phase I HapMap data from the CEPH Caucasian sample indicated that there was strong LD across the NRF2 gene. Eleven SNPs used in

HapMap across the NRF2 gene were selected for analysis. An additional five markers located across NRF2 were selected from dbSNP that had not been genotyped in HapMap. All of these SNPs were genotyped in a sample of 92 controls (termed the Irish reference panel). This allowed (a) comparison of the LD structure between the HapMap CEPH sample and the Irish reference panel for those SNPs genotyped in both samples, (b) analysis to determine if the tagging SNPs chosen on the basis of the HapMap data would have captured the additional SNPs at NRF2 chosen from dbSNP, and (c) a selection of tagging SNPs across NRF2 in the Irish reference panel for a comprehensive association study of NRF2 and SZ.

Comparison of LD data from HapMap Phase I and the Irish reference panel indicated that there is a high correlation in terms of r^2 values between the two samples. This suggests that the HapMap CEPH sample is a suitable reference panel to choose tagging SNPs for association studies in an Irish sample. However, it is important to note that the Phase I data lacked sufficient density to capture additional genetic variation at NRF2 when the extra SNPs from dbSNP were introduced into the analysis. This is likely to be less of a problem with the new HapMap Phase II data, which has a significantly higher SNP density. The SNPs chosen for full case-control analysis did not show evidence of association between NRF2 and SZ.

7.2 Strengths of the studies

The main strength of these investigations is that this is a relatively large collection of carefully phenotyped cases and controls drawn from the Irish population. This population may have some advantages in terms of genetic homogeneity when compared with samples drawn from other populations. One caveat of association studies is the possibility of false

positives due to population stratification. As highlighted in (Marchini et al 2004) even the smallest mix of two separate populations can dramatically affect allele frequencies. If ancestry specific allele frequencies are different this can confound association findings. Hoggart et al (2003) looked at extreme population stratification using African American, African Caribbean and Hispanic American samples. Over one third of markers were associated because of different ancestry specific allele frequencies. However, once they corrected for this, only a few true markers remained associated.

Another potential solution suggested by Freedman et al (2004) is that careful geographical and ancestral matching of case and controls may eliminate stratification issues. A population study by Hill et al (2000) suggested that Ireland was more genetically homogenous than the UK or USA due to its geographical location and limited immigration at least up until recently. Therefore a homogenous population such as Ireland may reduce the chance of false positives, have fewer susceptibility loci and potentially higher effect size, than more genetically heterogeneous populations (Rosenberg et al 2003).

Another strength was the strategy used to select genes for investigation. Understanding of the aetiology of schizophrenia is limited with no definitive molecular or biochemical pathways implicated. To increase the prior probability candidate genes were selected on the basis of positional (from linkage and cytogenetic studies) and functional (from expression studies) evidence in addition to compatibility with current hypotheses of aetiology. The regulation of the APOL family of genes and the oxidative stress genes (HSPA8 and GSTM3) were significantly altered in gene expression studies. In addition, they are all located in loci previously indicated from linkage studies. However, many of the genes in this study turned out to be negative. There are several reasons for this discussed in the next section (Section 7.3 Limitations of the study).

A novel addition for correcting pooled DNA allele frequencies was developed during the course of this work. The method by Norton et al (2002) suggested genotyping several samples to identify heterozygotes for each SNP. This is both time-consuming and costly. The novel method of simulating a range of heterozygote values to identify potential markers first and then confirming with true a heterozygote helps to reduce time and labour costs. For example, in the APOL study (Chapter 3) 51 SNPs were genotyped. If a panel of 15 individuals was used to identify heterozygotes, this would require 715 genotypes (15 samples x 51 SNPs). Using the simulation method, only two SNPs required correction with true heterozygotes (15 samples x 2 SNPs = 30 genotypes). This highlights the cost benefit of this novel addition to DNA pooling methodology.

The work carried out on the NRF2 gene to compare the applicability of HapMap information in future Irish case-control studies has proved useful. The finding of high r^2 correlations between the CEPH and Irish population is encouraging and has informed the use of tSNPs from HapMap data in further association studies.

Results described in this thesis support the involvement of G72 and DAAO in SZ pathogenesis. This is encouraging as association has been found in samples from other populations with these genes. However, caution is required in interpreting the results as we found that associated markers sometimes vary between population (see section 4.3) suggesting (1) they are all in high LD with an unknown functional mutation; (2) possible genetic heterogeneity; (3) or some findings may be false positives. Further functional work is warranted to assess the impact of various alleles at associated SNPs (as discussed in section 7.4).

7.3 Limitations of the Studies

A potential problem with the investigations presented is limitation in statistical power, despite the large samples. The sample sizes used for the APOL 4-6 study were 219 cases and 231 controls. Subsequent studies for G72, DAAO, HSPA8, GSTM3 and NRF2 used increased sample sizes of 299 cases and 645 controls. Table 7.1 gives an example of power. .

Table 7.1 Power Calculations for the two sample sizes used in these studies

Odds ratio	Cases = 219, Controls = 231	Cases = 299, Controls = 645
2	> 80%	> 98%
1.5	41%	67%
1.3	20%	33%

When these studies were planned the effect sizes of putative susceptibility genes (or loci) were recognised as being modest, with likely OR of <2. These previous studies informed the power calculations for our study samples. With the confident identification of risk variants for schizophrenia, for example, at NRG1 and dysbindin, estimated ORs of 1.3-1.4 have been reported. Furthermore, initial studies tend to overestimate effect sizes- an effect termed the winners curse- and subsequent replication studies report lower effect sizes (OR 1.1-1.3). As more studies with large sample sizes are conducted, it becomes increasingly likely that individual schizophrenia risk alleles or haplotypes will have OR=1.5. Despite our samples being some of the larger samples available at the time, they may have lacked statistical power to detect these very small effects.

A further potential limitation of the study was that no methods for genomic control were employed. Although the Irish population is regarded as being relatively homogenous (see section 7.2), an active debate continues on the effects of population stratification in association studies. A recent study by Helgason (2005) on an Icelandic sample (regarded as even more homogenous than Ireland) showed potential relevant effects of population stratification. They found regional subdivisions in genetic variation. Therefore it is feasible for one to assume that as the Irish population is older, larger and perhaps harbours more genetic diversity, and hence potential allelic and indeed locus heterogeneity than Iceland. Therefore consideration should be given to the effect of population stratification in future Irish association studies. In addition, as shown by Marchini et al (2004) differences in disease prevalence between subpopulations can also have an effect on population stratification. However, this is unlikely to be an issue here as epidemiological research in SZ suggests that disease prevalence is relatively globally uniform.

The HapMap project was in its infancy at the start of these investigations. By the completion of the Phase I data, the data described in this thesis and by others (e.g. Evans and Cardon 2005) showed that HapMap was useful for identifying linkage disequilibrium but lacked sufficient density to be comprehensive. Now, at the end of this studentship, the situation is considerably improved with the density of markers reaching 1 SNP per 1 Kb in most areas of the genome. This will make future association studies on candidate genes easier to perform and reduce the amount of genotyping and re-sequencing required. For example, when all the common variants are contained in the HapMap database, re-sequencing will not be required. If however, you were looking for rare variants in your population then it would still be useful. This is because HapMap is based on the common disease common variant model and does not take into account SNPs with minor allele frequency less than 5%.

There is also potential bias in sample ascertainment. The SZ samples were mainly ascertained from outpatient clinics at three Dublin hospitals and one regional hospital (Monaghan). However, a large proportion of the patients were ascertained from hospital wards. They may have been on the wards because they respond less well to treatment, or because they represent a particular subtype of schizophrenia characterised by severity of illness. The lack of availability of a true epidemiological sample may be problematic for complex genetic disorders depending on the likely effects of ascertainment bias, known and unknown, on clinical or genetic heterogeneity. Ascertainment in this study, as in most others in the field may over-sample from chronic treatment resistant patients (in hospital or attending Clozapine clinics) and from insightful, compliant individuals who attend outpatient clinics.

Phenotyping is also an issue. Although the DSM-IV and ICD-10 have face validity they do not have demonstrated biological validity. Therefore the sample collection probably contains sub-phenotypes of SZ thereby reducing power to detect associations. In light of this, the study of endophenotypes is emerging as a focus of interest in psychiatric genetics research. It must be pointed out that schizophrenia as a phenotype has been intensively investigated in genetic epidemiological research and has been confirmed as highly heritable, the heritability of endophenotypes is less well supported.

A further limitation in association studies is genotyping error. Case-control studies have no inbuilt means of detecting genotyping error as is the case with family studies. One suggested method is to genotype a subset of samples twice to estimate the error probability (Rice and Holmans 2003) and comparing the percentage of discrepancies. This method is sometimes adopted as a measure of accuracy. However, as highlighted by McDonald et al

(2005) the undetected presence of rare tertiary SNPs (SNPs with three potential alleles) may over-estimate genotyping accuracy. In addition, if discrepancies are identified there is no means of determining which is the correct classification (Kang et al 2004). Towards the end of these investigations, an opportunity to genotype a subset of the full sample arose (Chapter 6) using the suggested method by Rose and Holmans (2003). This showed no discrepancy between 92 genotypes using two different genotyping methods. It is intended to continue genotyping the Irish reference panel (n=92) using one method and genotyping the full case-control sample using another method to highlight any errors. In addition, the group now use the CEPH sample from Coriell (which have been validated and genotyped in the HapMap sample) so that we can use them to assess genotyping accuracy on different platforms.

This thesis involved the investigation of susceptibility genes and did not extend to the performance of functional studies. It is clearly important to design functional assays to further elucidate the role of identified susceptibility genes in disease aetiology. For example, one current method involves Yeast two-hybrid experiments to identify protein interactions, as has been described previously (Chumakov et al 2004). Methods such as Haplochip (Knight et al 2005) can also be used to characterise the functional significance of genetic variation. This method examines DNA transcribing into RNA in living cells using RNA polymerase II (Pol II) loading. It allows DNA variation to be examined for potential effects on gene expression levels and considers haplotypes across genes in an *in vivo* model. Animal gene-knockout studies (e.g. O'Tuathaigh et al 2006) can examine the phenotypic outcome based on genotypic manipulation. By assessing differences between knockout animals with control animals, the role of putative susceptibility genes can be studied. Furthermore, gene expression studies such as those carried out by Sabine Bahn's team in Cambridge (Mimmack et al 2002, Prabakaran et al 2004) can highlight over or

under expression of proteins giving functional evidence for putative susceptibility genes. Newer work involving MRI and fMRI to examine the effects of variation on both structure and function in specific regions of the brain will also assist in determining gene function.

There is increasing evidence that some of the genes identified in SZ are also significantly associated in Bipolar Affective Disorder. Indeed when specific subtypes of each disorder are pooled together, association increases (Williams et al, 2006). Therefore in the studies presented it is possible that the sample phenotype was either too broad or too narrow, depending on the underlying genetic architecture and this had a ‘dilution effect’ on any signal. As in the discussion about endophenotypes, future studies should incorporate clinical sub-phenotypes to aid identification of susceptibility genes. However, determining those phenotypes at present is problematic. It is also important to note that with each additional test performed on a sample without a priori hypothesis the chance of type II error occurring is increased.

7.4 Future Directions

To identify genes involved in the pathogenesis of SZ, sample size and study power issues need to be addressed. It can be concluded by the work presented in this thesis that the current sample sizes are underpowered to find $OR < 1.5$. Funding has been secured by Prof. Gill, Dr. Corvin and Dr. Morris to ascertain a further 1000 Irish cases and 3000 controls (RPGI). This sample will be available to all researchers and the returned data and analyses will be a beneficial to the psychiatric genetics field. The combined sample of this study (GASP and RPGI) is ~ 1400 cases and ~ 3000 controls. Power calculations suggest that the combined sample will have 95% power to detect an allele with $MAF = 0.25$ and $OR 1.3$ ($MAF = 0.1$ and $OR 1.3$).

An area that saw rapid change during the course of these investigations is that of LD structure and analysis. Initially, reports suggested that the genome could be divided into blocks, separated by areas of recombination (hotspots) (Gabriel et al 2002). Barrett et al (2005) developed a program called Haploview that generates quality statistics, LD information, haplotype blocks, population haplotype frequencies and single marker association statistics. It accepts data in the form of pedigree, unphased diplotypes in the standard linkage format or as a data dump direct from HapMap. LD blocks are determined by using one of three methods: (1) 95% confidence bounds on D' are generated and each comparison is called "strong LD", "inconclusive" or "strong recombination". A block is created if 95% of informative (i.e. non-inconclusive) comparisons are "strong LD"; (2) The population frequencies of the 4 possible two-marker haplotypes are computed for each marker pair. If all 4 are observed with at least frequency 0.01, a recombination is deemed to have taken place. Blocks are formed by consecutive markers where only 3 gametes are observed; (3) the third is internally developed and searches for a "spine" of strong LD running from one marker to another. This means that the first and last markers in a block are in strong LD with all intermediate markers but that the intermediate markers are not necessarily in LD with each other.

When HapMap was in its infancy the density of markers was poor. This led to all three methods producing very different LD blocks that made comparing populations problematic. An increase in the density of markers available, technological progress in genotyping and evolving statistical methods have improved understanding of LD relationships in the genome. The most recent change to LD analysis and tSNP selection was the inclusion of Tagger (de Bakker 2005) into Haploview. Tagger selects tSNPs in two different ways. (1) Greedy pairwise tagging (alleles are captured by single-marker tests at

the prescribed r^2 ; (2) prioritizing tags by the number of alleles that they can capture at a set r^2 ; it should also be noted that the maximum allowed distance between a tSNP and allele it proxies is 200 kb. It has become increasingly apparent that due to low effect size of individual risk alleles, power is a greater issue in association studies than previously thought. Block methods genotype multiple markers identifying haplotypes. However, generating haplotypes increases the degrees of freedom in statistical analysis and may decrease overall statistical power. By choosing tSNPs that act as proxies for other SNPs based on r^2 values, it may be possible to increase efficiency whilst retaining power. In addition, Tagger allows you to retrospectively see how much information previously chosen SNPs captured. A further benefit is that the calculations are based on r^2 values which have now become the measure of choice for comparing LD structure between populations. LD and HapMap comparisons with the Irish sample concur well with the current consensus. Scatterplot results showed that r^2 values between the CEPH sample and the Irish sample were highly correlated ($r^2 = 0.9369$). In contrast D' values showed much less correlation between both populations ($r^2 = 0.192$) making LD inference from HapMap difficult. This is probably due to the fact that both LD measures contain very different properties (as presented in section 1.6) and that D' is inflated when using low sample sizes.

Currently, the HapMap project has Phase II data on > 3.9 million SNPs genotyped in 30 CEPH trios. In future studies, I intend to use this information to establish local LD structure for each candidate gene and then select tagging SNPs. I will then test for single and haplotypic associations at each gene. In addition, I plan to develop skills in testing for epistatic interaction between putatively associated genes as genes are likely to interact in contributing to susceptibility (Cordell 2002, Moore 2004).

Moving on from candidate gene studies to genome-wide studies is the next logical step. Advances in genotyping technology and bioinformatics have made genome wide association studies a practical reality. Genome wide association studies combine the power and comprehensiveness of a genome-wide linkage study with the fine resolution of an association study highlighting regions of putative susceptibility. One obvious difficulty with these studies is that they will generate very large amounts of data. Computer systems, databases and statistical methods, capable of handling the vast amount of data generated in whole genome association studies need to be in place.

Already, case-control and family based methods have been developed for this type of study. As the number of markers being tested is vastly increased, methods to correct for multiple testing are being developed (Lin 2005). New methods have already been developed for statistical corrections at candidate gene level (Becker and Knapp 2004; Seaman and Muller-Myhosk 2005, Neale and Sham 2004). Collaborative efforts and multi-centre whole genome studies will be common place in the future.

During my Post-Doctoral career I intend to carry out additional candidate gene studies focusing on genes putatively involved in three areas: (1) Glutamate Neurotransmission (2) Oligodendrocyte function / myelination (3) Oxidative Stress. Selection of candidate genes will be informed by multiple sources of bioinformation. Priority will be given to SNPs located in functional or regulatory regions, regions conserved between mammalian species (<http://ecrbrowser.dcode.org>). Previous studies by our group (Morris et al 2006) have shown that re-sequencing of genes is redundant as databases such as dbSNP are adequate for SNP discovery. SNPs from dbSNP will be selected if they are heterozygous in European populations. If this information is not available and the SNP is in a regulatory region, an assay will be designed to determine heterozygosity in a DNA pool of 15 SZ

cases. 15 cases gives 95% power to detect alleles in the screened population with a MAF = 0.1. Genotyping will then be carried out using the Irish Reference Panel to determine LD and subsequent tSNPs will be genotyped in the full GASP sample.

The studies contained within this thesis have made a small but worthy contribution to psychiatric research. A positive replication in an independent sample has strengthened the original finding that G72/DAAO is involved in SZ pathogenesis. One paper has been published that increases the efficiency of DNA pooling experiments. The candidate genes investigated had been selected based on not only information from genetic studies but other biological disciplines, helping to make research more translational. I have shown through the comparison of HapMap data with the Irish population that it is a valuable tool, applicable to all complex disorders, in making association studies more efficient. I have enjoyed my time in Dublin and have gained valuable irreplaceable experience through the mentorship of Prof. Gill, Dr. Corvin and Dr. Morris. I hope that during my post-doctoral career I can now travel to other centres of excellence and develop my skills further to assist in solving the puzzle of psychiatric disorders.

References

2003. The International HapMap Project. *Nature*. 426, 789-796.

2005. The International HapMap Consortium. *Nature*. 437(7063):1299-320

Abecasis,G.R., Noguchi,E., Heinzmann,A., Traherne,J.A., Bhattacharyya,S., Leaves,N.I., Anderson,G.G., Zhang,Y., Lench,N.J., Carey,A., Cardon,L.R., Moffatt,M.F., and Cookson,W.O., 2001. Extent and distribution of linkage disequilibrium in three genomic regions. *Am J Hum Genet*. 68, 191-197.

Abecasis,G.R., Burt,R.A., Hall,D., Bochum,S., Doheny,K.F., Lundy,S.L., Torrington,M., Roos,J.L., Gogos,J.A., and Karayiorgou,M., 2004. Genomewide scan in families with schizophrenia from the founder population of Afrikaners reveals evidence for linkage and uniparental disomy on chromosome 1. *Am J Hum Genet*. 74, 403-417.

Addington,A.M., Gornick,M., Sporn,A.L., Gogtay,N., Greenstein,D., Lenane,M., Gochman,P., Baker,N., Balkissoon,R., Vakkalanka,R.K., Weinberger,D.R., Straub,R.E., and Rapoport,J.L., 2004. Polymorphisms in the 13q33.2 gene G72/G30 are associated with childhood-onset schizophrenia and psychosis not otherwise specified. *Biol Psychiatry*. 55, 976-980.

Andolfatto,P. and Nordborg,M., 1998. The effect of gene conversion on intralocus associations. *Genetics*. 148, 1397-1399.

Andreasson,S., Allebeck,P., Engstrom,A., and Rydberg,U., 1987. Cannabis and schizophrenia. A longitudinal study of Swedish conscripts. *Lancet*. 2, 1483-1486.

Ardlie,K.G., Kruglyak,L., and Seielstad,M., 2002. Patterns of linkage disequilibrium in the human genome. *Nat Rev Genet*. 3, 299-309.

- Arinami,T., Ohtsuki,T., Takase,K., Shimizu,H., Yoshikawa,T., Horigome,H., Nakayama,J., and Toru,M., 2001. Screening for 22q11 deletions in a schizophrenia population. *Schizophr Res.* 52, 167-170.
- Badner,J.A. and Gershon,E.S., 2002. Meta-analysis of whole-genome linkage scans of bipolar disorder and schizophrenia. *Mol Psychiatry.* 7, 405-411.
- Baron,M., 2001. Genetics of schizophrenia and the new millennium: progress and pitfalls. *Am J Hum Genet.* 68, 299-312.
- Barrett,J.C., Fry,B., Maller,J., and Daly,M.J., 2005. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics.* 21, 263-265.
- Bassett,A.S., Chow,E.W., and Weksberg,R., 2000. Chromosomal abnormalities and schizophrenia. *Am J Med Genet.* 97, 45-51.
- Bassett,A.S., Chow,E.W., AbdelMalik,P., Gheorghiu,M., Husted,J., and Weksberg,R., 2003. The schizophrenia phenotype in 22q11 deletion syndrome. *Am J Psychiatry.* 160, 1580-1586.
- Becker T, Knapp M. *Am J Hum Genet.* 2004 Oct;75(4):561-70. Epub 2004 Jul 30. A powerful strategy to account for multiple testing in the context of haplotype analysis.
- Blackwood,D.H., Fordyce,A., Walker,M.T., St Clair,D.M., Porteous,D.J., and Muir,W.J., 2001. Schizophrenia and affective disorders-- cosegregation with a translocation at chromosome 1q42 that directly disrupts brain-expressed genes: clinical and P300 findings in a family. *Am J Hum Genet.* 69, 428-433.

- Blouin, J.L., Dombroski, B.A., Nath, S.K., Lasseter, V.K., Wolyniec, P.S., Nestadt, G., Thornquist, M., Ullrich, G., McGrath, J., Kasch, L., Lamacz, M., Thomas, M.G., Gehrig, C., Radhakrishna, U., Snyder, S.E., Balk, K.G., Neufeld, K., Swartz, K.L., DeMarchi, N., Papadimitriou, G.N., Dikeos, D.G., Stefanis, C.N., Chakravarti, A., Childs, B., Housman, D.E., Kazazian, H.H., Antonarakis, S., and Pulver, A.E., 1998. Schizophrenia susceptibility loci on chromosomes 13q32 and 8p21. *Nat Genet.* 20, 70-73.
- Boehnke, M., 2000. A look at linkage disequilibrium. *Nat Genet.* 25, 246-247.
- Boska P. 2004. Animal models of obstetric complications in relation to schizophrenia. *Brain Res Brain Res Rev.* Apr;45(1):1-17
- Bray, N.J. and Owen, M.J., 2001. Searching for schizophrenia genes. *Trends Mol Med.* 7, 169-174.
- Brzustowicz, L.M., Honer, W.G., Chow, E.W., Little, D., Hogan, J., Hodgkinson, K., and Bassett, A.S., 1999. Linkage of familial schizophrenia to chromosome 13q32. *Am J Hum Genet.* 65, 1096-1103.
- 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, 657-666.
- Camp, N.J., Neuhausen, S.L., Tiobech, J., Polloi, A., Coon, H., and Myles-Worsley, M., 2001. Genomewide multipoint linkage analysis of seven extended Palauan pedigrees with schizophrenia, by a Markov-chain Monte Carlo method. *Am J Hum Genet.* 69, 1278-1289.
- Cannon, M., Jones, P.B., and Murray, R.M., 2002. Obstetric complications and schizophrenia: historical and meta-analytic review. *Am J Psychiatry.* 159, 1080-1092.

- Cannon,M. and Clarke,M.C., 2005. Risk for schizophrenia--broadening the concepts, pushing back the boundaries. *Schizophr Res.* 79, 5-13.
- Cannon,T.D., Rosso,I.M., Hollister,J.M., Bearden,C.E., Sanchez,L.E., and Hadley,T., 2000. A prospective cohort study of genetic and perinatal influences in the etiology of schizophrenia. *Schizophr Bull.* 26, 351-366.
- Cannon,T.D., Bearden,C.E., Hollister,J.M., Rosso,I.M., Sanchez,L.E., and Hadley,T., 2000. Childhood cognitive functioning in schizophrenia patients and their unaffected siblings: a prospective cohort study. *Schizophr Bull.* 26, 379-393.
- Cannon,T.D., Huttunen,M.O., Lonnqvist,J., Tuulio-Henriksson,A., Pirkola,T., Glahn,D., Finkelstein,J., Hietanen,M., Kaprio,J., and Koskenvuo,M., 2000. The inheritance of neuropsychological dysfunction in twins discordant for schizophrenia. *Am J Hum Genet.* 67, 369-382.
- Cannon,T.D., van Erp,T.G., Rosso,I.M., Huttunen,M., Lonnqvist,J., Pirkola,T., Salonen,O., Valanne,L., Poutanen,V.P., and Standertskjold-Nordenstam,C.G., 2002. Fetal hypoxia and structural brain abnormalities in schizophrenic patients, their siblings, and controls. *Arch Gen Psychiatry.* 59, 35-41.
- Cardno,A.G. and Gottesman,I.I., 2000. Twin studies of schizophrenia: from bow-and-arrow concordances to star wars Mx and functional genomics. *Am J Med Genet.* 97, 12-17.
- Cardno,A.G., Holmans,P.A., Rees,M.I., Jones,L.A., McCarthy,G.M., Hamshere,M.L., Williams,N.M., Norton,N., Williams,H.J., Fenton,I., Murphy,K.C., Sanders,R.D., Gray,M.Y., O'Donovan,M.C., McGuffin,P., and Owen,M.J., 2001. A genomewide linkage study of age at onset in schizophrenia. *Am J Med Genet.* 105, 439-445.
- Cardno,A.G., Rijdsdijk,F.V., Sham,P.C., Murray,R.M., and McGuffin,P., 2002. A twin study of genetic relationships between psychotic symptoms. *Am J Psychiatry.* 159, 539-545.

Cardno, A., McGuffin, P. 2002 Quantitative Genetics. Chpt 2. 31-54, Psychiatric Genetics and Genomics, Eds. McGuffin, P., Owen, M.J., Gottesman, I.I.

Carlsson, A. 1988. The current status of the dopamine hypothesis of schizophrenia. *Neuropsych.* 1, 179-86

Chakraborty, R. and Weiss, K.M., 1988. Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. *Proc Natl Acad Sci U S A.* 85, 9119-9123.

Chanas, S.A., Jiang, Q., McMahon, M., McWalter, G.K., McLellan, L.I., Elcombe, C.R., Henderson, C.J., Wolf, C.R., Moffat, G.J., Itoh, K., Yamamoto, M., and Hayes, J.D., 2002. Loss of the Nrf2 transcription factor causes a marked reduction in constitutive and inducible expression of the glutathione S-transferase *Gsta1*, *Gsta2*, *Gstm1*, *Gstm2*, *Gstm3* and *Gstm4* genes in the livers of male and female mice. *Biochem J.* 365, 405-416.

Chen, Y.S., Akula, N., Detera-Wadleigh, S.D., Schulze, T.G., Thomas, J., Potash, J.B., DePaulo, J.R., McInnis, M.G., Cox, N.J., and McMahon, F.J., 2004. Findings in an independent sample support an association between bipolar affective disorder and the *G72/G30* locus on chromosome 13q33. *Mol Psychiatry.* 9, 87-92.

Chumakov, I., Blumenfeld, M., Guerassimenko, O., Cavarec, L., Palicio, M., Abderrahim, H., Bougueleret, L., Barry, C., Tanaka, H., La Rosa, P., Puech, A., Tahri, N., Cohen-Akenine, A., Delabrosse, S., Lissarrague, S., Picard, F.P., Maurice, K., Essioux, L., Millasseau, P., Grel, P., Debailleul, V., Simon, A.M., Caterina, D., Dufaure, I., Malekzadeh, K., Belova, M., Luan, J.J., Bouillot, M., Sambucy, J.L., Primas, G., Saumier, M., Boubkiri, N., Martin-Saumier, S., Nasroune, M., Peixoto, H., Delaye, A., Pinchot, V., Bastucci, M., Guillou, S., Chevillon, M., Sainz-Fuertes, R., Meguenni, S., Aurich-Costa, J., Cherif, D., Gimalac, A., Van Duijn, C., Gauvreau, D., Ouellette, G., Fortier, I., Raelson, J., Sherbatich, T., Riazanskaia, N., Rogaev, E., Raeymaekers, P., Aerssens, J., Konings, F., Luyten, W., Macciardi, F., Sham, P.C., Straub, R.E.,

- Weinberger,D.R., Cohen,N., and Cohen,D., 2002. Genetic and physiological data implicating the new human gene G72 and the gene for D-amino acid oxidase in schizophrenia. *Proc Natl Acad Sci U S A.* 99, 13675-13680.
- Cordell, H.J. 2002. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum Mol Gen.* 11, 20, 2463-2468
- Corvin,A. and Gill,M., 2003. Psychiatric genetics in the post-genome age. *Br J Psychiatry.* 182, 95-96.
- 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, 208-213.
- Craddock,N., O'Donovan,M.C., and Owen,M.J., 2005. The genetics of schizophrenia and bipolar disorder: dissecting psychosis. *J Med Genet.* 42, 193-204.
- Craddock,N., O'Donovan,M.C., and Owen,M.J., 2005. The genetics of schizophrenia and bipolar disorder: dissecting psychosis. *J Med Genet.* 42, 193-204.
- Craddock,N., O'Donovan,M.C., and Owen,M.J., 2006. Genes for schizophrenia and bipolar disorder? Implications for psychiatric nosology. *Schizophr Bull.* 32, 9-16.
- Crow,T., 2003. Genes for schizophrenia. *Lancet.* 361, 1829-1830.
- 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, 229-232.
- Datta SR, McQuillin A, Rizig MA, Thirumalai S, Pimm J, Moorey H, Quedsted D, Kalsi G, Bass n, Lawrence J, Choudury K, Puri V, Curtis D, Grling HMD. 2003. P8.18 Further

genetic analysis of the *pcm1* gene association with schizophrenia on chromosome 8p21 and tests of the G72, dysbindin, RGS4, calcineurin, COMT, frizzled 3, *mrds1*, *akt1* and *capon* associations. *Amer Journ. Med Gen part B.* 130B, 86.

David, A.S., Malmberg, A., Brandt, L., Allebeck, P., and Lewis, G., 1997. IQ and risk for schizophrenia: a population-based cohort study. *Psychol Med.* 27, 1311-1323.

Dawson, E., Abecasis, G.R., Bumpstead, S., Chen, Y., Hunt, S., Beare, D.M., Pabial, J., Dibling, T., Tinsley, E., Kirby, S., Carter, D., Papaspyridonos, M., Livingstone, S., Ganske, R., Lohmussaar, E., Zernant, J., Tonisson, N., Remm, M., Magi, R., Puurand, T., Vilo, J., Kurg, A., Rice, K., Deloukas, P., Mott, R., Metspalu, A., Bentley, D.R., Cardon, L.R., and Dunham, I., 2002. A first-generation linkage disequilibrium map of human chromosome 22. *Nature.* 418, 544-548.

de Bakker PI, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, Altshuler D. Efficiency and power in genetic association studies. *Nat Genet.* 2005 Nov;37(11):1217-23. Epub 2005 Oct 23.

Dean, B., Sundram, S., Bradbury, R., Scarr, E., and Copolov, D., 2001. Studies on [³H]CP-55940 binding in the human central nervous system: regional specific changes in density of cannabinoid-1 receptors associated with schizophrenia and cannabis use. *Neuroscience.* 103, 9-15.

DeLisi, L.E., 1997. Is schizophrenia a lifetime disorder of brain plasticity, growth and ageing? *Schizophr Res.* 23, 119-129.

DeLisi, L.E., Shaw, S.H., Crow, T.J., Shields, G., Smith, A.B., Larach, V.W., Wellman, N., Loftus, J., Nanthakumar, B., Razi, K., Stewart, J., Comazzi, M., Vita, A., Heffner, T., and Sherrington, R., 2002. A genome-wide scan for linkage to chromosomal regions in 382 sibling pairs with schizophrenia or schizoaffective disorder. *Am J Psychiatry.* 159, 803-812.

- Detera-Wadleigh, S.D., McMahon, F.J. 2006. G72/G30 in schizophrenia and bipolar disorder: review and meta-analysis. *Biol Psychiatry*. Jul 15;60(2):106-14
- Devlin, B. and Roeder, K., 1999. Genomic control for association studies. *Biometrics*. 55, 997-1004.
- DSM-III. 1980. Diagnostic and Statistical Manual of mental disorders 3rd Ed. American Psychiatric Association
- DSM-IV. 2001. Diagnostic and Statistical Manual of mental disorders 4th Ed. American Psychiatric Association
- DSM-IV-TR. 2004. Diagnostic and Statistical Manual of mental disorders 4th Ed. American Psychiatric Association
- Edwards, J.H. 1965. The meaning of associations between blood groups and disease. *Annals of Human Genetics*. 29, 77-83
- Ekelund, J., Lichtermann, D., Hovatta, I., Ellonen, P., Suvisaari, J., Terwilliger, J.D., Juvonen, H., Varilo, T., Arajärvi, R., Kokko-Sahin, M.L., Lonnqvist, J., and Peltonen, L., 2000. Genome-wide scan for schizophrenia in the Finnish population: evidence for a locus on chromosome 7q22. *Hum Mol Genet*. 9, 1049-1057.
- Ekelund, J., Hovatta, I., Parker, A., Paunio, T., Varilo, T., Martin, R., Suhonen, J., Ellonen, P., Chan, G., Sinsheimer, J.S., Sobel, E., Juvonen, H., Arajärvi, R., Partonen, T., Suvisaari, J., Lonnqvist, J., Meyer, J., and Peltonen, L., 2001. Chromosome 1 loci in Finnish schizophrenia families. *Hum Mol Genet*. 10, 1611-1617.
- Dawson E, Abecasis GR, Bumpstead S, Chen Y, Hunt S, Beare DM, Pabial J, Dibling T, Tinsley E, Kirby S, Carter D, Papaspyridonos M, Livingstone S, Ganskek R, Lohmussaar E, Zernant J, Tonisson N, Remmq M, Maägi R, Puurand T, Vilo J, Kurg A, Rice K, Deloukas P,

Mott R, Metspalu A, Bentley DR, Cardon LR, Dunham I. 2002. A first-generation linkage disequilibrium map of human chromosome 22. *Nature*. 418, 544-548

Erlenmeyer-Kimling,L., Rock,D., Roberts,S.A., Janal,M., Kestenbaum,C., Cornblatt,B., Adamo,U.H., and Gottesman,I.I., 2000. Attention, memory, and motor skills as childhood predictors of schizophrenia-related psychoses: the New York High-Risk Project. *Am J Psychiatry*. 157, 1416-1422.

Evans,D.M. and Cardon,L.R., 2005. A comparison of linkage disequilibrium patterns and estimated population recombination rates across multiple populations. *Am J Hum Genet*. 76, 681-687.

Falconer,D.S., 1967. The inheritance of liability to diseases with variable age of onset, with particular reference to diabetes mellitus. *Ann Hum Genet*. 31, 1-20.

Falkai,P., Schneider-Axmann,T., Honer,W.G., Vogele,K., Schonell,H., Pfeiffer,U., Scherk,H., Block,W., Traber,F., Schild,H.H., Maier,W., and Tepest,R., 2003. Influence of genetic loading, obstetric complications and premorbid adjustment on brain morphology in schizophrenia: a MRI study. *Eur Arch Psychiatry Clin Neurosci*. 253, 92-99.

Faraone,S.V., Skol,A.D., Tsuang,D.W., Bingham,S., Young,K.A., Prabhudesai,S., Haverstock,S.L., Mena,F., Menon,A.S., Bisset,D., Pepple,J., Sautter,F., Baldwin,C., Weiss,D., Collins,J., Keith,T., Boehnke,M., Tsuang,M.T., and Schellenberg,G.D., 2002. Linkage of chromosome 13q32 to schizophrenia in a large veterans affairs cooperative study sample. *Am J Med Genet*. 114, 598-604.

Fekete A, Treszl A, Tóth-Heyn P, Vannay A, Tordai A, Tulassay T, Vásárhelyi B. Association between heat shock protein 72 gene polymorphism and acute renal failure in premature neonates. *Pediatr Res*. 2003 Oct;54(4):452-5.

- Freedman,M.L., Reich,D., Penney,K.L., McDonald,G.J., Mignault,A.A., Patterson,N., Gabriel,S.B., Topol,E.J., Smoller,J.W., Pato,C.N., Pato,M.T., Petryshen,T.L., Kolonel,L.N., Lander,E.S., Sklar,P., Henderson,B., Hirschhorn,J.N., and Altshuler,D., 2004. Assessing the impact of population stratification on genetic association studies. *Nat Genet.* 36, 388-393.
- Gabriel,S.B., Schaffner,S.F., Nguyen,H., Moore,J.M., Roy,J., Blumenstiel,B., Higgins,J., DeFelice,M., Lochner,A., Faggart,M., Liu-Cordero,S.N., Rotimi,C., Adeyemo,A., Cooper,R., Ward,R., Lander,E.S., Daly,M.J., and Altshuler,D., 2002. The structure of haplotype blocks in the human genome. *Science.* 296, 2225-2229.
- Gill,M., Vallada,H., Collier,D., Sham,P., Holmans,P., Murray,R., McGuffin,P., Nanko,S., Owen,M., Antonarakis,S., Housman,D., Kazazian,H., Nestadt,G., Pulver,A.E., Straub,R.E., MacLean,C.J., Walsh,D., Kendler,K.S., Delisi,L., Polymeropoulos,M., Coon,H., Byerley,W., Lofthouse,R., Gershon,E., Read,C.M., and ., 1996. A combined analysis of D22S278 marker alleles in affected sib-pairs: support for a susceptibility locus for schizophrenia at chromosome 22q12. Schizophrenia Collaborative Linkage Group (Chromosome 22). *Am J Med Genet.* 67, 40-45.
- Gillman and Myatt 1998 (Epicalc) <http://www.brixtonhealth.com/epicalc.html> (2006)
- Gottesman,I.I. and Shields,J., 1967. A polygenic theory of schizophrenia. *Proc Natl Acad Sci U S A.* 58, 199-205.
- Gottesman, I.I., Shields, J., 1972. *Schizophrenia and genetics: a twin study vantage point.* New York: Academic Press
- Gottesman ,I I. 1991. *Schizophrenia genesis: The origins of madness.* Freeman. New York
- Gottesman,I.I. and Bertelsen,A., 1989. Confirming unexpressed genotypes for schizophrenia. Risks in the offspring of Fischer's Danish identical and fraternal discordant twins. *Arch Gen Psychiatry.* 46, 867-872.

- Grima G, benz B, Parpura V, Cuenod M, Do KQ. 2003. Dopamine-induced oxidative stress in neurons with glutathione deficit: implication for schizophrenia. *Schizophr Res*. Aug 1;62(3):213-24
- Gurling,H.M., Kalsi,G., Brynjolfson,J., Sigmundsson,T., Sherrington,R., Mankoo,B.S., Read,T., Murphy,P., Blaveri,E., McQuillin,A., Petursson,H., and Curtis,D., 2001. Genomewide genetic linkage analysis confirms the presence of susceptibility loci for schizophrenia, on chromosomes 1q32.2, 5q33.2, and 8p21-22 and provides support for linkage to schizophrenia, on chromosomes 11q23.3-24 and 20q12.1-11.23. *Am J Hum Genet*. 68, 661-673.
- Hakak,Y., Walker,J.R., Li,C., Wong,W.H., Davis,K.L., Buxbaum,J.D., Haroutunian,V., and Fienberg,A.A., 2001. Genome-wide expression analysis reveals dysregulation of myelination-related genes in chronic schizophrenia. *Proc Natl Acad Sci U S A*. 98, 4746-4751.
- Halliwell B.1992. Reactive oxidative species and the central nervous system. *J Neurochem*. Nov;59(5):1609-23
- Harrison,P.J., 1999. Neurochemical alterations in schizophrenia affecting the putative receptor targets of atypical antipsychotics. Focus on dopamine (D1, D3, D4) and 5-HT2a receptors. *Br J Psychiatry Suppl*. 12-22.
- Harrison,P.J. and Weinberger,D.R., 2005. Schizophrenia genes, gene expression, and neuropathology: on the matter of their convergence. *Mol Psychiatry*. 10, 40-68.
- Harrison,P.J. and Weinberger,D.R., 2005. Schizophrenia genes, gene expression, and neuropathology: on the matter of their convergence. *Mol Psychiatry*. 10, 40-68.

Hashimoto E, Ozawa H, Saito T, Gsell W, Takahata N, Riederer P, Frölich L. Impairment of G(salpa) function in human brain cortex of Alzheimer's disease: comparison with normal aging. *J Neural Transm.* 2004 Mar;111(3):311-22.

Hattori,E., Liu,C., Badner,J.A., Bonner,T.I., Christian,S.L., Maheshwari,M., Detera-Wadleigh,S.D., Gibbs,R.A., and Gershon,E.S., 2003. Polymorphisms at the G72/G30 gene locus, on 13q33, are associated with bipolar disorder in two independent pedigree series. *Am J Hum Genet.* 72, 1131-1140.

Helgason A, Yngvadóttir B, Hrafnkelsson B, Gulcher J, Stefánsson K. An Icelandic example of the impact of population structure on association studies. *Nat Genet.* 2005 Jan;37(1):90-5.

Heston,L.L., 1966. Psychiatric disorders in foster home reared children of schizophrenic mothers. *Br J Psychiatry.* 112, 819-825.

Hill,E.W., Jobling,M.A., and Bradley,D.G., 2000. Y-chromosome variation and Irish origins. *Nature.* 404, 351-352.

Hoggart , C.J., Parra EJ, Shriver MD, Bonilla C, Kittles RA, Clayton DG, McKeigue PM. 2003. Control of confounding of genetic association in stratified populations. *Am.J.Hu.Genet.* 72, 1492-1504.

Hoogendoorn B, Norton N, Kirov G, Williams N, Hamshere ML, Spurlock G, Austin J, Stephens MK, Buckland PR, Owen MJ, O'Donovan MC. Cheap, accurate and rapid allele frequency estimation of single nucleotide polymorphisms by primer extension and DHPLC in DNA pools. *Hum Genet.* 2000 Nov;107(5):488-93.

- Hovatta,I., Varilo,T., Suvisaari,J., Terwilliger,J.D., Ollikainen,V., Arajärvi,R., Juvonen,H., Kokko-Sahin,M.L., Vaisanen,L., Mannila,H., Lonnqvist,J., and Peltonen,L., 1999. A genomewide screen for schizophrenia genes in an isolated Finnish subpopulation, suggesting multiple susceptibility loci. *Am J Hum Genet.* 65, 1114-1124.
- Hultman,C.M., Sparen,P., Takei,N., Murray,R.M., and Cnattingius,S., 1999. Prenatal and perinatal risk factors for schizophrenia, affective psychosis, and reactive psychosis of early onset: case-control study. *BMJ.* 318, 421-426.
- Hutchinson,G., Takei,N., Fahy,T.A., Bhugra,D., Gilvarry,C., Moran,P., Mallett,R., Sham,P., Leff,J., and Murray,R.M., 1996. Morbid risk of schizophrenia in first-degree relatives of white and African-Caribbean patients with psychosis. *Br J Psychiatry.* 169, 776-780.
- ICD-10. 1990. International Classification of Diseases. World Health Organisation.
- Ingraham,L.J. and Kety,S.S., 2000. Adoption studies of schizophrenia. *Am J Med Genet.* 97, 18-22.
- Itoh,K., Chiba,T., Takahashi,S., Ishii,T., Igarashi,K., Katoh,Y., Oyake,T., Hayashi,N., Satoh,K., Hatayama,I., Yamamoto,M., and Nabeshima,Y., 1997. An Nrf2/small Maf heterodimer mediates the induction of phase II detoxifying enzyme genes through antioxidant response elements. *Biochem Biophys Res Commun.* 236, 313-322.
- Jeffreys, A.J., May, C.A., 2004. Intense and highly localised gene conversion activity in human meiotic crossover hot spots. *Nat. Genet.* 36, 151-156.
- 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, 233-237.

- Jones AC, Austin J, Hansen N, Hoogendoorn B, Oefner PJ, Cheadle JP, O'Donovan MC. Optimal temperature selection for mutation detection by denaturing HPLC and comparison to single-stranded conformation polymorphism and heteroduplex analysis. *Clin Chem*. 1999 Aug;45(8 Pt 1):1133-40. Jeffreys,A.J. and May,C.A., 2004. Intense and highly localized gene conversion activity in human meiotic crossover hot spots. *Nat Genet*. 36, 151-156.
- Jones,P., Rodgers,B., Murray,R., and Marmot,M., 1994. Child development risk factors for adult schizophrenia in the British 1946 birth cohort. *Lancet*. 344, 1398-1402.
- Jorde,L.B., 2000. Linkage disequilibrium and the search for complex disease genes. *Genome Res*. 10, 1435-1444.
- Jurewicz,I., Owen,R.J., O'Donovan,M.C., and Owen,M.J., 2001. Searching for susceptibility genes in schizophrenia. *Eur Neuropsychopharmacol*. 11, 395-398.
- Kang,M.I., Kobayashi,A., Wakabayashi,N., Kim,S.G., and Yamamoto,M., 2004. Scaffolding of Keap1 to the actin cytoskeleton controls the function of Nrf2 as key regulator of cytoprotective phase 2 genes. *Proc Natl Acad Sci U S A*. 101, 2046-2051.
- Karayiorgou,M., Morris,M.A., Morrow,B., Shprintzen,R.J., Goldberg,R., Borrow,J., Gos,A., Nestadt,G., Wolynec,P.S., Lasseter,V.K., and ., 1995. Schizophrenia susceptibility associated with interstitial deletions of chromosome 22q11. *Proc Natl Acad Sci U S A*. 92, 7612-7616.
- Kendler, K, S. 2000. Schizophrenia genetics, in B.J. Sadock and V.A. Sadock (eds) Kaplan and Sadock's comprehensive textbook of psychiatry, Vol 1, Philadelphia: Lippincott, Williams and Wilkins

- Kendler,K.S., Karkowski,L.M., and Walsh,D., 1998. The structure of psychosis: latent class analysis of probands from the Roscommon Family Study. *Arch Gen Psychiatry*. 55, 492-499.
- Kendler,K.S., MacLean,C.J., Ma,Y., O'Neill,F.A., Walsh,D., and Straub,R.E., 1999. Marker-to-marker linkage disequilibrium on chromosomes 5q, 6p, and 8p in Irish high-density schizophrenia pedigrees. *Am J Med Genet*. 88, 29-33.
- Kety,S.S., Wender,P.H., Jacobsen,B., Ingraham,L.J., Jansson,L., Faber,B., and Kinney,D.K., 1994. Mental illness in the biological and adoptive relatives of schizophrenic adoptees. Replication of the Copenhagen Study in the rest of Denmark. *Arch Gen Psychiatry*. 51, 442-455.
- Knight JC, Keating BJ, Rockett KA, Kwiatkowski DP. In vivo characterization of regulatory polymorphisms by allele-specific quantification of RNA polymerase loading. *Nat Genet*. 2003 Apr;33(4):469-75.
- Korostishevsky,M., Kaganovich,M., Cholostoy,A., Ashkenazi,M., Ratner,Y., Dahary,D., Bernstein,J., Bening-Abu-Shach,U., Ben Asher,E., Lancet,D., Ritsner,M., and Navon,R., 2004. Is the G72/G30 locus associated with schizophrenia? single nucleotide polymorphisms, haplotypes, and gene expression analysis. *Biol Psychiatry*. 56, 169-176.
- Kringlen,E. and Cramer,G., 1989. Offspring of monozygotic twins discordant for schizophrenia. *Arch Gen Psychiatry*. 46, 873-877.
- Kruglyak,L., 1999. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat Genet*. 22, 139-144.
- Kwak,M.K., Itoh,K., Yamamoto,M., Sutter,T.R., and Kensler,T.W., 2001. Role of transcription factor Nrf2 in the induction of hepatic phase 2 and antioxidative enzymes

in vivo by the cancer chemoprotective agent, 3H-1, 2-dimethiole-3-thione. *Mol Med.* 7, 135-145.

Kwak,M.K., Wakabayashi,N., Greenlaw,J.L., Yamamoto,M., and Kensler,T.W., 2003. Antioxidants enhance mammalian proteasome expression through the Keap1-Nrf2 signaling pathway. *Mol Cell Biol.* 23, 8786-8794.

Lander,E. and Kruglyak,L., 1995. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat Genet.* 11, 241-247.

Lasseter,V.K., Pulver,A.E., Wolyniec,P.S., Nestadt,G., Meyers,D., Karayiorgou,M., Housman,D., Antonarakis,S., Kazazian,H., Kasch,L., and ., 1995. Follow-up report of potential linkage for schizophrenia on chromosome 22q: Part 3. *Am J Med Genet.* 60, 172-173.

Lee,J.M., Shih,A.Y., Murphy,T.H., and Johnson,J.A., 2003. NF-E2-related factor-2 mediates neuroprotection against mitochondrial complex I inhibitors and increased concentrations of intracellular calcium in primary cortical neurons. *J Biol Chem.* 278, 37948-37956.

Lee,J.M., Anderson,P.C., Padgitt,J.K., Hanson,J.M., Waters,C.M., and Johnson,J.A., 2003. Nrf2, not the estrogen receptor, mediates catechol estrogen-induced activation of the antioxidant responsive element. *Biochim Biophys Acta.* 1629, 92-101.

Lee,J.M., Calkins,M.J., Chan,K., Kan,Y.W., and Johnson,J.A., 2003. Identification of the NF-E2-related factor-2-dependent genes conferring protection against oxidative stress in primary cortical astrocytes using oligonucleotide microarray analysis. *J Biol Chem.* 278, 12029-12038.

Leung,L., Kwong,M., Hou,S., Lee,C., and Chan,J.Y., 2003. Deficiency of the Nrf1 and Nrf2 transcription factors results in early embryonic lethality and severe oxidative stress. *J Biol Chem.* 278, 48021-48029.

Lewis,C.M., Levinson,D.F., Wise,L.H., DeLisi,L.E., Straub,R.E., Hovatta,I., Williams,N.M., Schwab,S.G., Pulver,A.E., Faraone,S.V., Brzustowicz,L.M., Kaufmann,C.A., Garver,D.L., Gurling,H.M., Lindholm,E., Coon,H., Moises,H.W., Byerley,W., Shaw,S.H., Mesen,A., Sherrington,R., O'Neill,F.A., Walsh,D., Kendler,K.S., Ekelund,J., Paunio,T., Lonnqvist,J., Peltonen,L., O'Donovan,M.C., Owen,M.J., Wildenauer,D.B., Maier,W., Nestadt,G., Blouin,J.L., Antonarakis,S.E., Mowry,B.J., Silverman,J.M., Crowe,R.R., Cloninger,C.R., Tsuang,M.T., Malaspina,D., Harkavy-Friedman,J.M., Svrakic,D.M., Bassett,A.S., Holcomb,J., Kalsi,G., McQuillin,A., Brynjolfson,J., Sigmundsson,T., Petursson,H., Jazin,E., Zoega,T., and Helgason,T., 2003. Genome scan meta-analysis of schizophrenia and bipolar disorder, part II: Schizophrenia. *Am J Hum Genet.* 73, 34-48.

Lewontin, R.C. 164. The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics.* 49, 49-67

Li,T., Ball,D., Zhao,J., Murray,R.M., Liu,X., Sham,P.C., and Collier,D.A., 2000. Family-based linkage disequilibrium mapping using SNP marker haplotypes: application to a potential locus for schizophrenia at chromosome 22q11. *Mol Psychiatry.* 5, 77-84.

Lieberan, J.A.. 1999. Is schizophrenia a neurodegenerative disorder? A clinical and neurobiological perspective. *Biol. Psychiatry.* 46, 729-739

Lin,D.Y., 2005. An efficient Monte Carlo approach to assessing statistical significance in genomic studies. *Bioinformatics.* 21, 781-787.

Lin,M.W., Curtis,D., Williams,N., Arranz,M., Nanko,S., Collier,D., McGuffin,P., Murray,R., Owen,M., Gill,M., and ., 1995. Suggestive evidence for linkage of schizophrenia to markers on chromosome 13q14.1-q32. *Psychiatr Genet.* 5, 117-126.

- Lin, M.W., Sham, P., Hwu, H.G., Collier, D., Murray, R., and Powell, J.F., 1997. Suggestive evidence for linkage of schizophrenia to markers on chromosome 13 in Caucasian but not Oriental populations. *Hum Genet.* 99, 417-420.
- Liu, X., He, G., Wang, X., Chen, Q., Qian, X., Lin, W., Li, D., Gu, N., Feng, G., and He, L., 2004. Association of DAAO with schizophrenia in the Chinese population. *Neurosci Lett.* 369, 228-233.
- 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, 177-182.
- Macario, A.J., Conway D.E., Macario, E. 2005. Sick chaperones, cellular stress and disease. *N Engl J Med* Oct 6;353(14):1489-501
- Magin, T.M. 2003. A keeper and a striker maintain epidermal homeostasis. *Nat Genet.* Nov;35(3)202-4
- Mahadik, S.P., Evans, D., Lal, H. 2001. Oxidative stress and role of antioxidant and omega-3 essential fatty acid supplementation in schizophrenia. *prog. Neuro-Psychopharmacol. & Biol. Psychiat.* 25, 463-493
- Mahadik, S.P. and Mukherjee, S., 1996. Cultured skin fibroblasts as a cell model for investigating schizophrenia. *J Psychiatr Res.* 30, 421-439.
- Maeir, W. Schwab, S. 1998. Molecular genetics of schizophrenia. *Cur. Opin. Psych.* 11(1), 19-25
- Malmberg, A., Lewis, G., David, A., and Allebeck, P., 1998. Premorbid adjustment and personality in people with schizophrenia. *Br J Psychiatry.* 172, 308-313.

- Marchini,J., Cardon,L.R., Phillips,M.S., and Donnelly,P., 2004. The effects of human population structure on large genetic association studies. *Nat Genet.* 36, 512-517.
- Marchini,J., Cardon,L.R., Phillips,M.S., and Donnelly,P., 2004. The effects of human population structure on large genetic association studies. *Nat Genet.* 36, 512-517.
- Martin,R.H., 1999. Sperm chromosome analysis in a man heterozygous for a paracentric inversion of chromosome 14 (q24.1q32.1). *Am J Hum Genet.* 64, 1480-1484.
- McGhee,K.A., Morris,D.W., Schwaiger,S., Nangle,J.M., Donohoe,G., Clarke,S., Meagher,D., Quinn,J., Scully,P., Waddington,J.L., Gill,M., and Corvin,A., 2005. Investigation of the apolipoprotein-L (APOL) gene family and schizophrenia using a novel DNA pooling strategy for public database SNPs. *Schizophr Res.* 76, 231-238.
- McGue,M. and Gottesman,I.I., 1989. A single dominant gene still cannot account for the transmission of schizophrenia. *Arch Gen Psychiatry.* 46, 478-480.
- McGuffin,P., Owen,M.J., and Farmer,A.E., 1995. Genetic basis of schizophrenia. *Lancet.* 346, 678-682.
- Mednick,S.A., Mura,E., Schulsinger,F., and Mednick,B., 1971. Perinatal conditions and infant development in children with schizophrenic parents. *Soc Biol.* 18, S103-S113.
- Middleton FA, Mirnics K, Pierri JN, Lewis DA, Levitt P. Gene expression profiling reveals alterations of specific metabolic pathways in schizophrenia. *J Neurosci.* 2002 Apr 1;22(7):2718-29.
- 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, 1415-1423.

- Milner, C.M., Campbell, R.D. 1990. Structure and expression of the three MHC-linked HSP70 genes 1. *Immunogenetics*. 3232(4):242-51
- Mimmack, M.L., Ryan, M., Baba, H., Navarro-Ruiz, J., Iritani, S., Faull, R.L., McKenna, P.J., Jones, P.B., Arai, H., Starkey, M., Emson, P.C., and Bahn, S., 2002. Gene expression analysis in schizophrenia: reproducible up-regulation of several members of the apolipoprotein L family located in a high-susceptibility locus for schizophrenia on chromosome 22. *Proc Natl Acad Sci U S A*. 99, 4680-4685.
- Miyamoto, S., LaMantia, A.S., Duncan, G.E., Sullivan, P., Gilmore, J.H., and Lieberman, J.A., 2003. Recent advances in the neurobiology of schizophrenia. *Mol Interv*. 3, 27-39.
- Miyoshi, K., Honda, A., Baba, K., Taniguchi, M., Oono, K., Fujita, T., Kuroda, S., Katayama, T., and Tohyama, M., 2003. Disrupted-In-Schizophrenia 1, a candidate gene for schizophrenia, participates in neurite outgrowth. *Mol Psychiatry*. 8, 685-694.
- Monajemi, H., Fontijn, R.D., Pannekoek, H., and Horrevoets, A.J., 2002. The apolipoprotein L gene cluster has emerged recently in evolution and is expressed in human vascular tissue. *Genomics*. 79, 539-546.
- Moore J.H. 2004. Computational analysis of gene-gene interactions using multifactor dimensionality reduction. *Expert Rev Mol Diagn*. Nov, 496):795-803.
- Morris DW, Murphy K, Kenny N, Purcell SM, McGhee KA, Schwaiger S, Nangle JM, Donohoe G, Clarke S, Scully P, Quinn J, Meagher D, Baldwin P, Crumlish N, O'Callaghan E, Waddington JL, Gill M, Corvin AP. Dysbindin (DTNBP1) and the biogenesis of lysosome-related organelles complex 1 (BLOC-1): main and epistatic gene effects are potential contributors to schizophrenia susceptibility. *Epub* 2007 Jul 9.
- Mortensen, P.B., Pedersen, C.B., Westergaard, T., Wohlfahrt, J., Ewald, H., Mors, O., Andersen, P.K., and Melbye, M., 1999. Effects of family history and place and season of birth on the risk of schizophrenia. *N Engl J Med*. 340, 603-608.

- Morton, N.E., 1955. Sequential tests for the detection of linkage. *Am J Hum Genet.* 7, 277-318.
- Mothet JP, Parent AT, Wolosker H, Brady RO Jr, Linden DJ, Ferris CD, Rogawski MA, Snyder SH. D-serine is an endogenous ligand for the glycine site of the N-methyl-D-aspartate receptor. *Proc Natl Acad Sci U S A.* 2000 Apr 25;97(9):4926-31.
- Mueser, K.T. and McGurk, S.R., 2004. Schizophrenia. *Lancet.* 363, 2063-2072.
- Mulle JG, Chowdari KV, Nimgaonkar V, Chakravarti A. No evidence for association to the G72/G30 locus in an independent sample of schizophrenia families. *Mol Psychiatry.* 2005 May;10(5):431-3.
- Murphy, K.C., Jones, L.A., and Owen, M.J., 1999. High rates of schizophrenia in adults with velo-cardio-facial syndrome. *Arch Gen Psychiatry.* 56, 940-945.
- Murphy, K.C., 2002. Schizophrenia and velo-cardio-facial syndrome. *Lancet.* 359, 426-430.
- Neale, B.M. and Sham, P.C., 2004. The future of association studies: gene-based analysis and replication. *Am J Hum Genet.* 75, 353-362.
- Norton N, Williams NM, Williams HJ, Spurlock G, Kirov G, Morris DW, Hoogendoorn B, Owen MJ, O'Donovan MC. Universal, robust, highly quantitative SNP allele frequency measurement in DNA pools. *Hum Genet.* 2002 May;110(5):471-8. Epub 2002 Mar 23.
- Nyholt, D.R. 2004. A simple correction for multiple testing for single nucleotide polymorphisms in linkage disequilibrium with each other. *Am.J.Hum.Genet.* 74, 765-769
- O'Donovan MC, Oefner PJ, Roberts SC, Austin J, Hoogendoorn B, Guy C, Speight G,

- Upadhyaya M, Sommer SS, McGuffin P. Blind analysis of denaturing high-performance liquid chromatography as a tool for mutation detection. *Genomics*. 1998 Aug 15;52(1):44-9.
- O'Tuathaigh,C.M., O'Sullivan,G.J., Kinsella,A., Harvey,R.P., Tighe,O., Croke,D.T., and Waddington,J.L., 2006. Sexually dimorphic changes in the exploratory and habituation profiles of heterozygous neuregulin-1 knockout mice. *Neuroreport*. 17, 79-83.
- Owen, M.J., O'Donovan, M.C., Gottesman, I.I. 2002. Schizophrenia. Chapter 10. 247-266. *Psychiatric Genetics and Genomics*. Oxford
- Owen,M.J., 2000. Molecular genetic studies of schizophrenia. *Brain Res Brain Res Rev*. 31, 179-186.
- Owen,M.J., Williams,N.M., and O'Donovan,M.C., 2004. The molecular genetics of schizophrenia: new findings promise new insights. *Mol Psychiatry*. 9, 14-27.
- Ozeki,Y., Tomoda,T., Kleiderlein,J., Kamiya,A., Bord,L., Fujii,K., Okawa,M., Yamada,N., Hatten,M.E., Snyder,S.H., Ross,C.A., and Sawa,A., 2003. Disrupted-in-Schizophrenia-1 (DISC-1): mutant truncation prevents binding to NudeE-like (NUDEL) and inhibits neurite outgrowth. *Proc Natl Acad Sci U S A*. 100, 289-294.
- Page,N.M., Butlin,D.J., Lomthaisong,K., and Lowry,P.J., 2001. The human apolipoprotein L gene cluster: identification, classification, and sites of distribution. *Genomics*. 74, 71-78.
- Pastinen T, Partanen J, Syvänen AC. Multiplex, fluorescent, solid-phase minisequencing for efficient screening of DNA sequence variation. *Clin Chem*. 1996 Sep;42(9):1391-7.

- Perlman,W.R., Weickert,C.S., Akil,M., and Kleinman,J.E., 2004. Postmortem investigations of the pathophysiology of schizophrenia: the role of susceptibility genes. *J Psychiatry Neurosci.* 29, 287-293.
- Perlson, G.D., Marsh, L. 1999. Structural Brain imaging in schizophrenia: A selective review. *Biol Psychiatry.* 46, 627-649
- Prabakaran,S., Swatton,J.E., Ryan,M.M., Huffaker,S.J., Huang,J.T., Griffin,J.L., Wayland,M., Freeman,T., Dudbridge,F., Lilley,K.S., Karp,N.A., Hester,S., Tkachev,D., Mimmack,M.L., Yolken,R.H., Webster,M.J., Torrey,E.F., and Bahn,S., 2004. Mitochondrial dysfunction in schizophrenia: evidence for compromised brain metabolism and oxidative stress. *Mol Psychiatry.* 9, 684-97, 643.
- Pritchard,J.K. and Rosenberg,N.A., 1999. Use of unlinked genetic markers to detect population stratification in association studies. *Am J Hum Genet.* 65, 220-228.
- Pritchard,J.K. and Przeworski,M., 2001. Linkage disequilibrium in humans: models and data. *Am J Hum Genet.* 69, 1-14.
- Pulver,A.E., Nestadt,G., Goldberg,R., Shprintzen,R.J., Lamacz,M., Wolyniec,P.S., Morrow,B., Karayiorgou,M., Antonarakis,S.E., Housman,D., and ., 1994. Psychotic illness in patients diagnosed with velo-cardio-facial syndrome and their relatives. *J Nerv Ment Dis.* 182, 476-478.
- Rapoport JL, Giedd J, Kumra S, Jacobsen L, Smith A, Lee P, Nelson J, Hamburger, S. Childhood-onset schizophrenia. Progressive ventricular change during adolescence. *Vol. 54 No. 10, October 1997*
- 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, 199-204.

- Reich,D.E. and Goldstein,D.B., 2001. Detecting association in a case-control study while correcting for population stratification. *Genet Epidemiol.* 20, 4-16.
- Rice K.M., Holmans, P. 2003. Allowing for genotyping error in analysis of unmatched case-control studies. *Ann hum genet.* Mar;67(Pt 2):165-74
- Risch,N., 1990. Linkage strategies for genetically complex traits. I. Multilocus models. *Am J Hum Genet.* 46, 222-228.
- Risch,N., 1990. Genetic linkage and complex diseases, with special reference to psychiatric disorders. *Genet Epidemiol.* 7, 3-16.
- Rosenberg NA, Li LM, Ward R, Pritchard JK(2003) Informativeness of genetic markers for inference of ancestry. *American Journal of Human Genetics* 73: 1402-1422.
- Rosenthal,D., Wender,P.H., Kety,S.S., Welner,J., and Schulsinger,F., 1971. The adopted-away offspring of schizophrenics. *Am J Psychiatry.* 128, 307-311.
- Rosso,I.M., Cannon,T.D., Huttunen,T., Huttunen,M.O., Lonnqvist,J., and Gasperoni,T.L., 2000. Obstetric risk factors for early-onset schizophrenia in a Finnish birth cohort. *Am J Psychiatry.* 157, 801-807.
- Sambrook, J., Fritsch, E.F., Maniatis, T. 1982. *Molecular Cloning, A Laboratory Manual* (Cold Spring Harbor)
- Sargent CA, Dunham I, Trowsdale J, Campbell RD. Human major histocompatibility complex contains genes for the major heat shock protein HSP70. *Proc Natl Acad Sci U S A.* 1989 Mar;86(6):1968-72.
- Schulze,K., McDonald,C., Frangou,S., Sham,P., Grech,A., Touloupoulou,T., Walshe,M., Sharma,T., Sigmundsson,T., Taylor,M., and Murray,R.M., 2003. Hippocampal volume

in familial and nonfamilial schizophrenic probands and their unaffected relatives. *Biol Psychiatry*. 53, 562-570.

Schumacher,J., Jamra,R.A., Freudenberg,J., Becker,T., Ohlraun,S., Otte,A.C., Tullius,M., Kovalenko,S., Bogaert,A.V., Maier,W., Rietschel,M., Propping,P., Nothen,M.M., and Cichon,S., 2004. Examination of G72 and D-amino-acid oxidase as genetic risk factors for schizophrenia and bipolar affective disorder. *Mol Psychiatry*. 9, 203-207.

Schumacher,J., Jamra,R.A., Freudenberg,J., Becker,T., Ohlraun,S., Otte,A.C., Tullius,M., Kovalenko,S., Bogaert,A.V., Maier,W., Rietschel,M., Propping,P., Nothen,M.M., and Cichon,S., 2004. Examination of G72 and D-amino-acid oxidase as genetic risk factors for schizophrenia and bipolar affective disorder. *Mol Psychiatry*. 9, 203-207.

Schwab,S.G., Knapp,M., Mondabon,S., Hallmayer,J., Borrmann-Hassenbach,M., Albus,M., Lerer,B., Rietschel,M., Trixler,M., Maier,W., and Wildenauer,D.B., 2003. Support for association of schizophrenia with genetic variation in the 6p22.3 gene, dysbindin, in sib-pair families with linkage and in an additional sample of triad families. *Am J Hum Genet*. 72, 185-190.

Seaman SR, Müller-Myhsok B. Rapid simulation of P values for product methods and multiple-testing adjustment in association studies. *Am J Hum Genet*. 2005 Mar;76(3):399-408. Epub 2005 Jan 11.

Sham, P., McGuffin, P. Linkage and association. Chapter 3, 55-76. *Psychiatric Genetics and Genomics*. Oxford.

Shaw,S.H., Kelly,M., Smith,A.B., Shields,G., Hopkins,P.J., Loftus,J., Laval,S.H., Vita,A., De Hert,M., Cardon,L.R., Crow,T.J., Sherrington,R., and DeLisi,L.E., 1998. A genome-wide search for schizophrenia susceptibility genes. *Am J Med Genet*. 81, 364-376.

- Shaw,S.H., Kelly,M., Smith,A.B., Shields,G., Hopkins,P.J., Loftus,J., Laval,S.H., Vita,A., De Hert,M., Cardon,L.R., Crow,T.J., Sherrington,R., and DeLisi,L.E., 1998. A genome-wide search for schizophrenia susceptibility genes. *Am J Med Genet.* 81, 364-376.
- Shih,A.Y., Imbeault,S., Barakauskas,V., Erb,H., Jiang,L., Li,P., and Murphy,T.H., 2005. Induction of the Nrf2-driven antioxidant response confers neuroprotection during mitochondrial stress in vivo. *J Biol Chem.* 280, 22925-22936.
- Sklar,P., 2002. Linkage analysis in psychiatric disorders: the emerging picture. *Annu Rev Genomics Hum Genet.* 3, 371-413.
- Stankovich,J., Cox,C.J., Tan,R.B., Montgomery,D.S., Huxtable,S.J., Rubio,J.P., Ehm,M.G., Johnson,L., Butzkueven,H., Kilpatrick,T.J., Speed,T.P., Roses,A.D., Bahlo,M., and Foote,S.J., 2006. On the utility of data from the International HapMap Project for Australian association studies. *Hum Genet.* 119, 220-222.
- Stefan, M., Travis, M. and Murray, R.M. (Eds.) (2002) *An atlas of schizophrenia.* London
- 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, 877-892.
- 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, 83-87.

- Stephens, J.C., Briscoe, D., and O'Brien, S.J., 1994. Mapping by admixture linkage disequilibrium in human populations: limits and guidelines. *Am J Hum Genet.* 55, 809-824.
- Stephens, J.C., Schneider, J.A., Tanguay, D.A., Choi, J., Acharya, T., Stanley, S.E., Jiang, R., Messer, C.J., Chew, A., Han, J.H., Duan, J., Carr, J.L., Lee, M.S., Koshy, B., Kumar, A.M., Zhang, G., Newell, W.R., Windemuth, A., Xu, C., Kalbfleisch, T.S., Shaner, S.L., Arnold, K., Schulz, V., Drysdale, C.M., Nandabalan, K., Judson, R.S., Ruano, G., and Vovis, G.F., 2001. Haplotype variation and linkage disequilibrium in 313 human genes. *Science.* 293, 489-493.
- Straub, R.E., Jiang, Y., MacLean, C.J., Ma, Y., Webb, B.T., Myakishev, M.V., Harris-Kerr, C., Wormley, B., Sadek, H., Kadambi, B., Cesare, A.J., Gibberman, A., Wang, X., O'Neill, F.A., Walsh, D., and Kendler, K.S., 2002. Genetic variation in the 6p22.3 gene DTNBP1, the human ortholog of the mouse dysbindin gene, is associated with schizophrenia. *Am J Hum Genet.* 71, 337-348.
- Sutcliffe, J.G. and Thomas, E.A., 2002. The neurobiology of apolipoproteins in psychiatric disorders. *Mol Neurobiol.* 26, 369-388.
- Syvanen AC, 2001. Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nat Rev Genet.* Dec 29(12):930-42
- Takahashi, S., Cui, Y.H., Kojima, T., Han, Y.H., Zhou, R.L., Kamioka, M., Yu, S.Y., Matsuura, M., Matsushima, E., Wilcox, M., Arinami, T., Shen, Y.C., Faraone, S.V., and Tsuang, M.T., 2003. Family-based association study of markers on chromosome 22 in schizophrenia using African-American, European-American, and Chinese families. *Am J Med Genet B Neuropsychiatr Genet.* 120, 11-17.
- Terwilliger, J.D., Zollner, S., Laan, M., and Paabo, S., 1998. Mapping genes through the use of linkage disequilibrium generated by genetic drift: 'drift mapping' in small populations with no demographic expansion. *Hum Hered.* 48, 138-154.

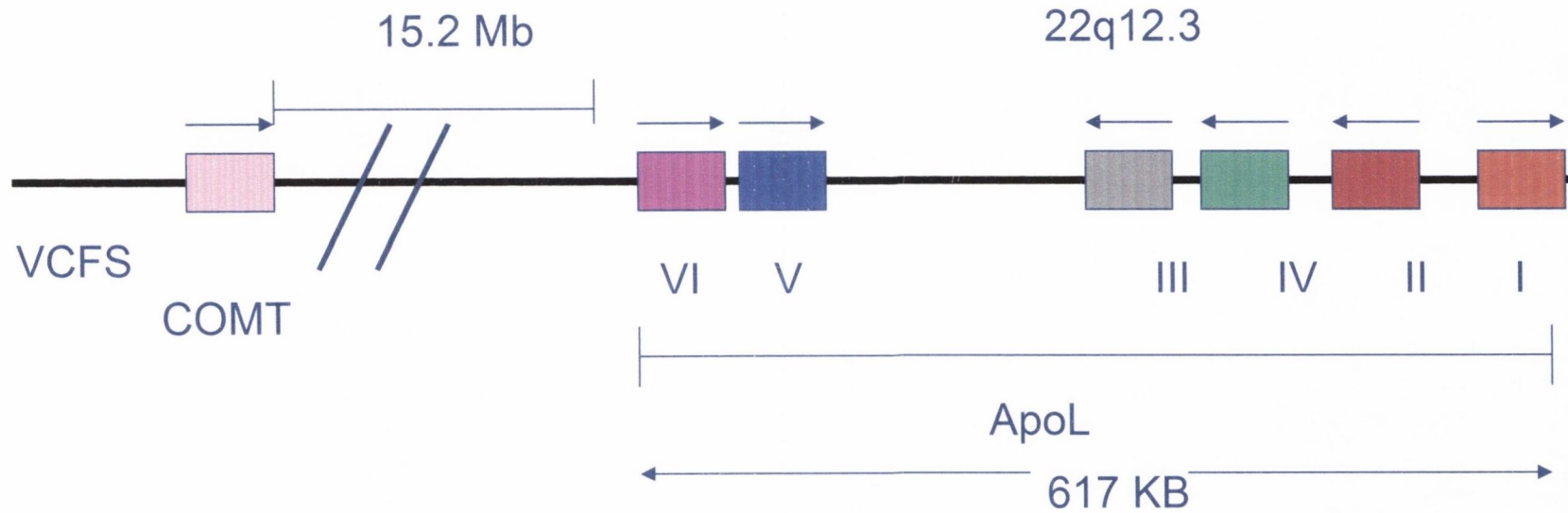
- Tienari,P., Wynne,L.C., Moring,J., Laksy,K., Nieminen,P., Sorri,A., Lahti,I., Wahlberg,K.E., Naarala,M., Kurki-Suonio,K., Saarento,O., Koistinen,P., Tarvainen,T., Hakko,H., and Miettunen,J., 2000. Finnish adoptive family study: sample selection and adoptee DSM-III-R diagnoses. *Acta Psychiatr Scand.* 101, 433-443.
- 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. *Lancet.* 362, 798-805.
- Tsuang,M.T., Taylor,L., and Faraone,S.V., 2004. An overview of the genetics of psychotic mood disorders. *J Psychiatr Res.* 38, 3-15.
- Underhill PA, Jin L, Lin AA, Mehdi SQ, Jenkins T, Vollrath D, Davis RW, Cavalli-Sforza LL, Oefner PJ. Detection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography. *Genome Res.* 1997 Oct;7(10):996-1005.
- Wang,X., He,G., Gu,N., Yang,J., Tang,J., Chen,Q., Liu,X., Shen,Y., Qian,X., Lin,W., Duan,Y., Feng,G., and He,L., 2004. Association of G72/G30 with schizophrenia in the Chinese population. *Biochem Biophys Res Commun.* 319, 1281-1286.
- Weinberger,D.R., 1995. From neuropathology to neurodevelopment. *Lancet.* 346, 552-557.
- Weiss,K.M. and Clark,A.G., 2002. Linkage disequilibrium and the mapping of complex human traits. *Trends Genet.* 18, 19-24.
- Welch P.J., Gambetti, P. 1998. Chaperoning brain diseases. *Nature.* 392, 23-24
- Wender,P.H., Rosenthal,D., Kety,S.S., Schulsinger,F., and Welner,J., 1974. Crossfostering. A research strategy for clarifying the role of genetic and experiential factors in the etiology of schizophrenia. *Arch Gen Psychiatry.* 30, 121-128.

- Wender,P.H., Rosenthal,D., Kety,S.S., Schulsinger,F., and Welner,J., 1974. Crossfostering. A research strategy for clarifying the role of genetic and experiential factors in the etiology of schizophrenia. *Arch Gen Psychiatry.* 30, 121-128.
- Wijsman,E.M., Rosenthal,E.A., Hall,D., Blundell,M.L., Sobin,C., Heath,S.C., Williams,R., Brownstein,M.J., Gogos,J.A., and Karayiorgou,M., 2003. Genome-wide scan in a large complex pedigree with predominantly male schizophrenics from the island of Kosrae: evidence for linkage to chromosome 2q. *Mol Psychiatry.* 8, 695-705, 643.
- Willer,C.J., Scott,L.J., Bonnycastle,L.L., Jackson,A.U., Chines,P., Pruim,R., Bark,C.W., Tsai,Y.Y., Pugh,E.W., Doheny,K.F., Kinnunen,L., Mohlke,K.L., Valle,T.T., Bergman,R.N., Tuomilehto,J., Collins,F.S., and Boehnke,M., 2006. Tag SNP selection for Finnish individuals based on the CEPH Utah HapMap database. *Genet Epidemiol.* 30, 180-190.
- Williams,N.M., Cardno, A.G., Murphy, K.C., Jones, L.A., Asherson, P., McGuffin, P., Pwen, M.J. 1997. Association between schizophrenia and a microsatellite polymorphism at the dopamine D5 receptor gene. *Psychiatr Genet,* 7(2):83-5
- Williams, N.M., Norton,N., Williams,H., Ekholm,B., Hamshere,M.L., Lindblom,Y., Chowdari,K.V., Cardno,A.G., Zammit,S., Jones,L.A., Murphy,K.C., Sanders,R.D., McCarthy,G., Gray,M.Y., Jones,G., Holmans,P., Nimgaonkar,V., Adolfson,R., Osby,U., Terenius,L., Sedvall,G., O'Donovan,M.C., and Owen,M.J., 2003. A systematic genomewide linkage study in 353 sib pairs with schizophrenia. *Am J Hum Genet.* 73, 1355-1367.
- Williams,N.M., Preece,A., Morris,D.W., Spurlock,G., Bray,N.J., Stephens,M., Norton,N., Williams,H., Clement,M., Dwyer,S., Curran,C., Wilkinson,J., Moskvina,V., Waddington,J.L., Gill,M., Corvin,A.P., Zammit,S., Kirov,G., Owen,M.J., and O'Donovan,M.C., 2004. Identification in 2 independent samples of a novel schizophrenia risk haplotype of the dystrobrevin binding protein gene (DTNBP1). *Arch Gen Psychiatry.* 61, 336-344.

- Williams,N.M., Preece,A., Spurlock,G., Norton,N., Williams,H.J., Zammit,S., O'Donovan,M.C., and Owen,M.J., 2003. Support for genetic variation in neuregulin 1 and susceptibility to schizophrenia. *Mol Psychiatry*. 8, 485-487.
- Wise,L.H., Lanchbury,J.S., and Lewis,C.M., 1999. Meta-analysis of genome searches. *Ann Hum Genet*. 63 (Pt 3), 263-272.
- Wright IC, Rabe-Hesketh S, Woodruff PWR, David AS, Murray RM, Bullmore ET. Meta-analysis of regional brain volumes in schizophrenia. *Am J Psychiatry* 2000; 157: 16–25.
- Wyatt,R.J., Henter,I., Leary,M.C., and Taylor,E., 1995. An economic evaluation of schizophrenia--1991. *Soc Psychiatry Psychiatr Epidemiol*. 30, 196-205.
- Yamada, K. 2004 *Am J Med Genet B Neuropsychiatr Genet*. 2004 Sep 15;130B(1):1-179
- Zhao JH, Lissarrague S, Essioux L, Sham PC. GENECOUNTING: haplotype analysis with missing genotypes. *Bioinformatics*. 2002 Dec;18(12):1694-5.
- Zhao,J.H. and Sham,P.C., 2002. Faster haplotype frequency estimation using unrelated subjects. *Hum Hered*. 53, 36-41.
- Zou,F., Li,C., Duan,S., Zheng,Y., Gu,N., Feng,G., Xing,Y., Shi,J., and He,L., 2005. A family-based study of the association between the G72/G30 genes and schizophrenia in the Chinese population. *Schizophr Res*. 73, 257-261.

Appendix

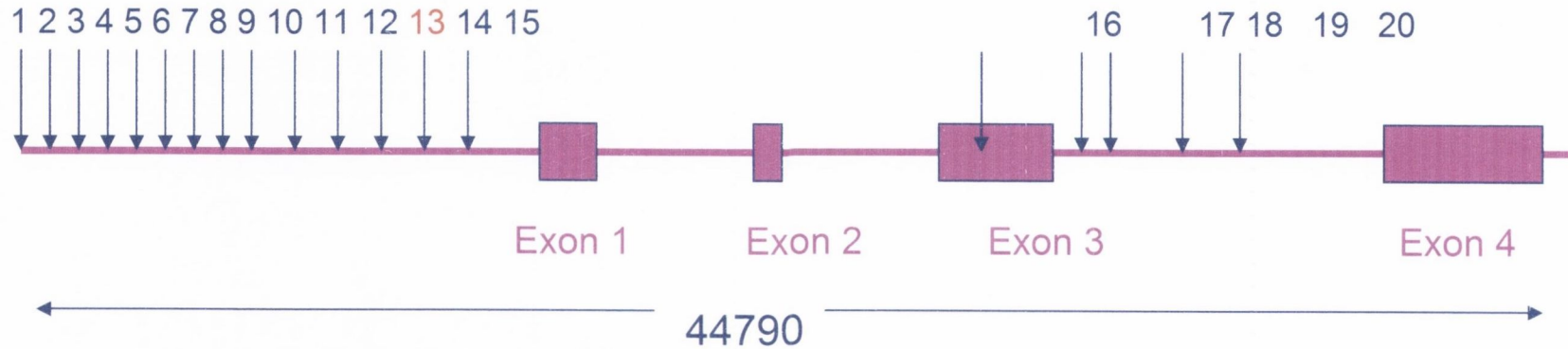
Apolipoprotein L Family I - VI



Adapted from Mimmack *et al* 2002

Figure A.1: Schematic of the location of the APOL genes in relation to the VCFS region.

Apolipoprotein VI



Apolipoprotein V

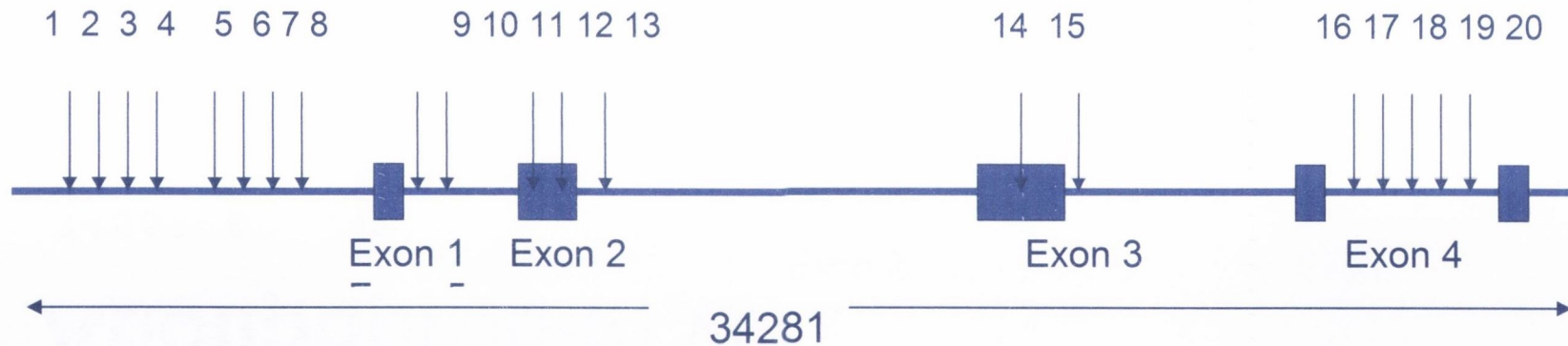
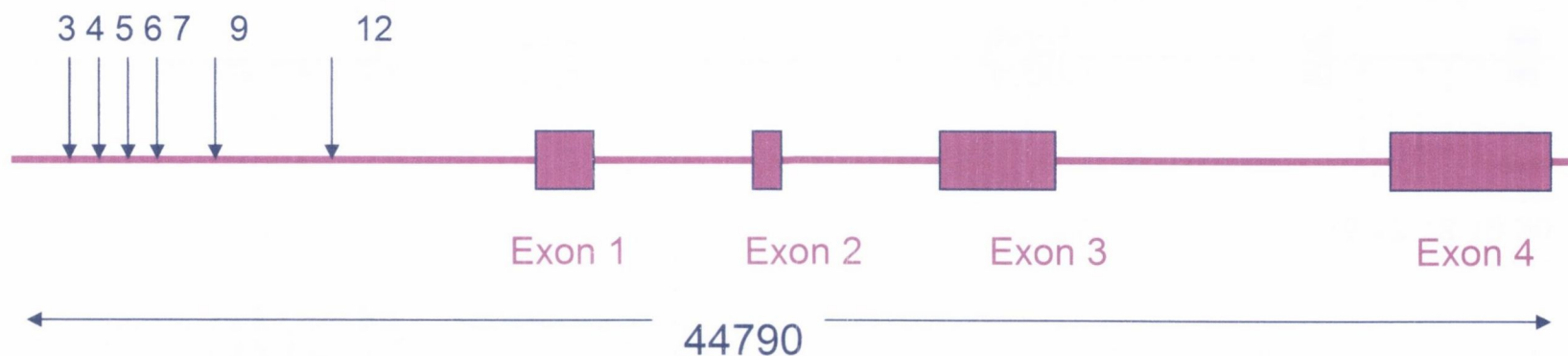


Figure A.2: Schematic of the SNPs identified from dbSNP spanning APOL6 and APOL5. All these SNPs were selected for

Apolipoprotein VI

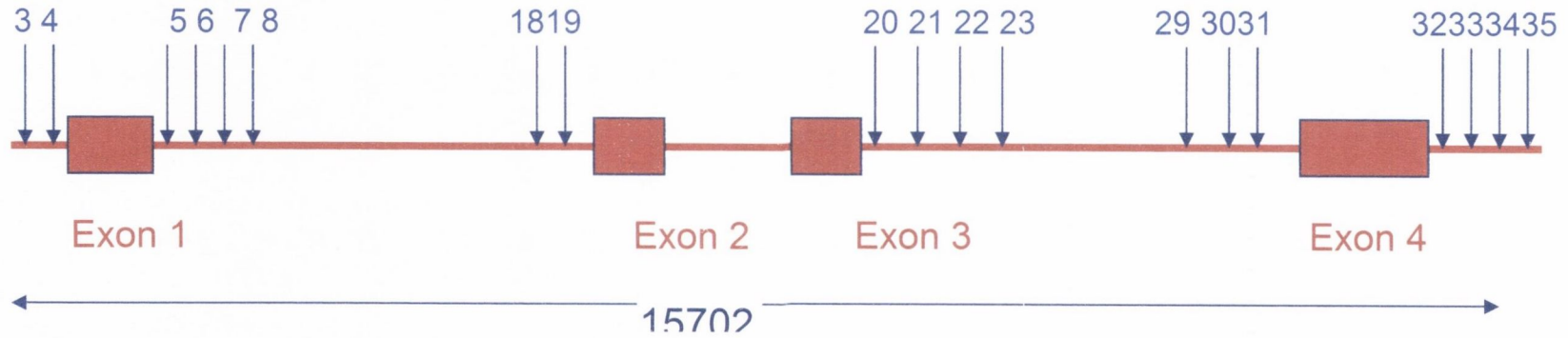


Apolipoprotein V



Figure A.3: Schematic of the SNPs that were heterozygous in the Irish Sample. Association analysis was carried out on all of these SNPs using DNA pooling techniques.

Apolipoprotein I



Apolipoprotein III

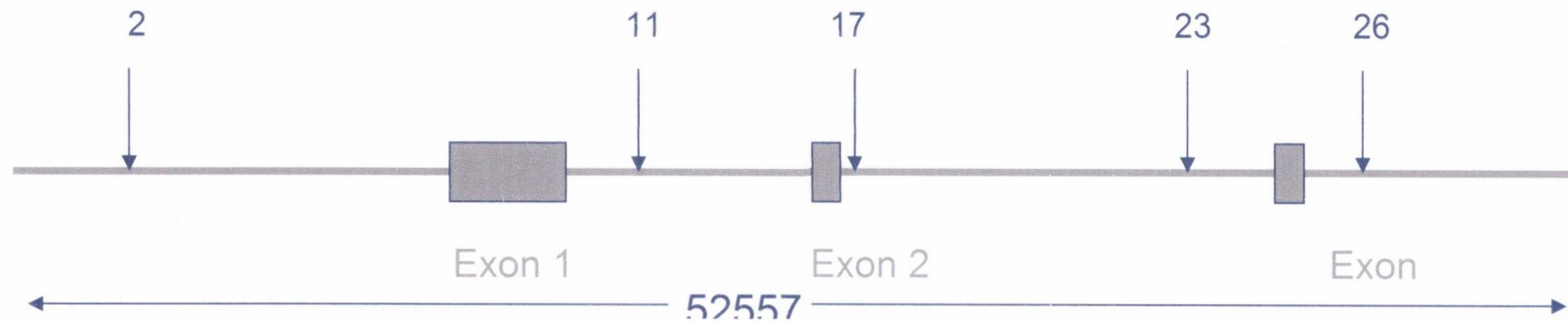
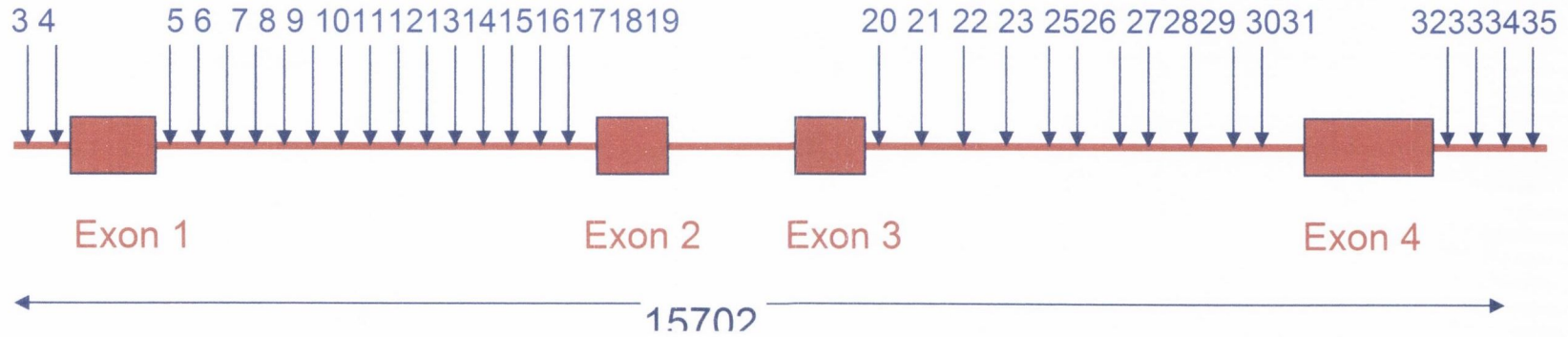


Figure A.5: Schematic of the SNPs selected for initial pooling studies to identify heterozygosity in the Irish sample for

Apolipoprotein I



Apolipoprotein III

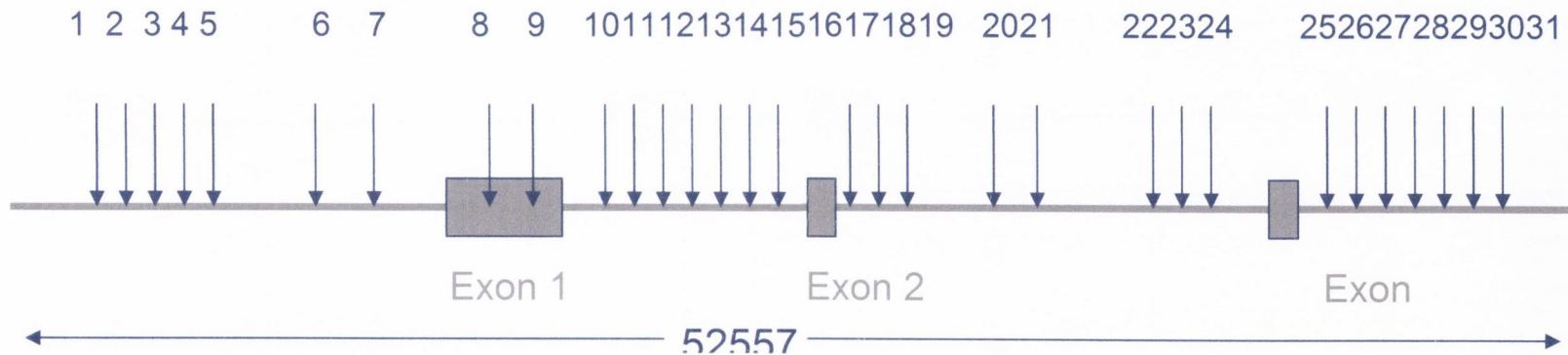
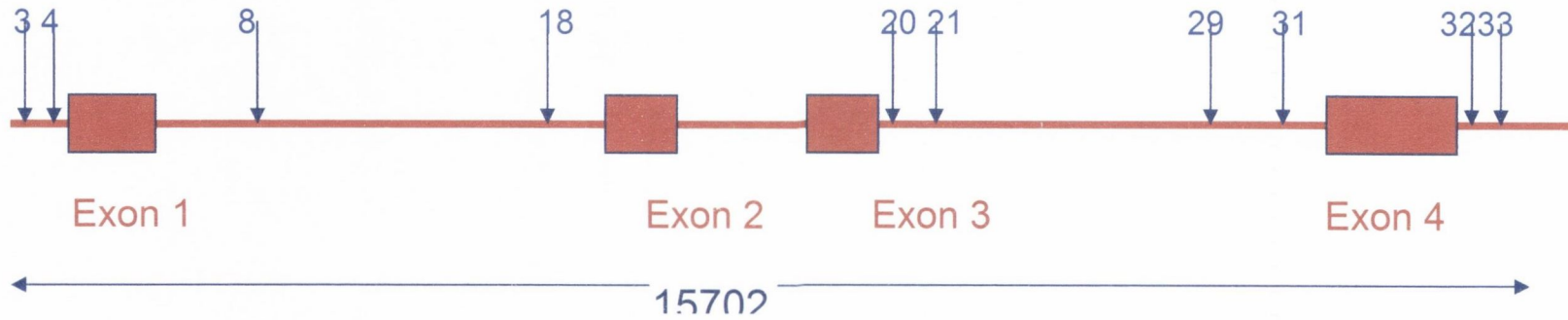
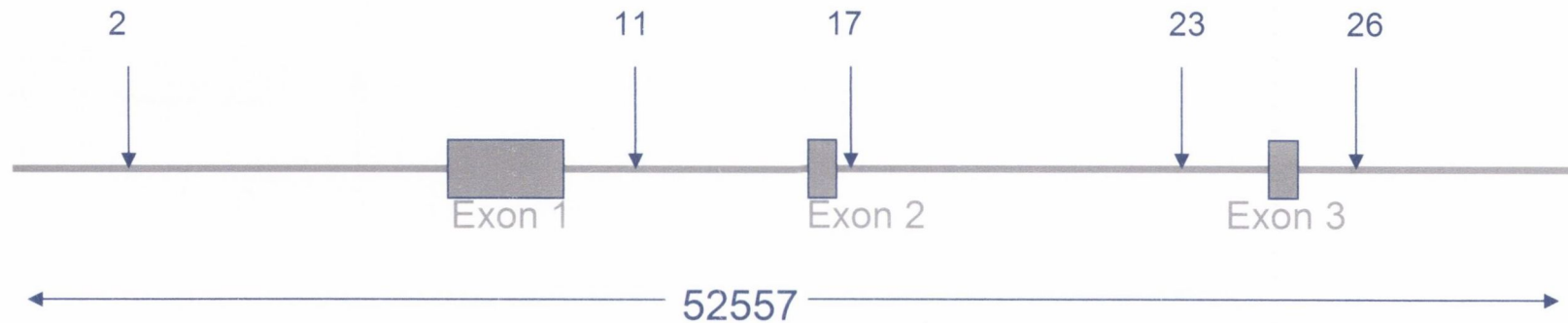


Figure A.4: Schematic of the SNPs identified from dbSNP spanning APOL1 and APOL3.

Apolipoprotein I



Apolipoprotein III



Apolipoprotein IV

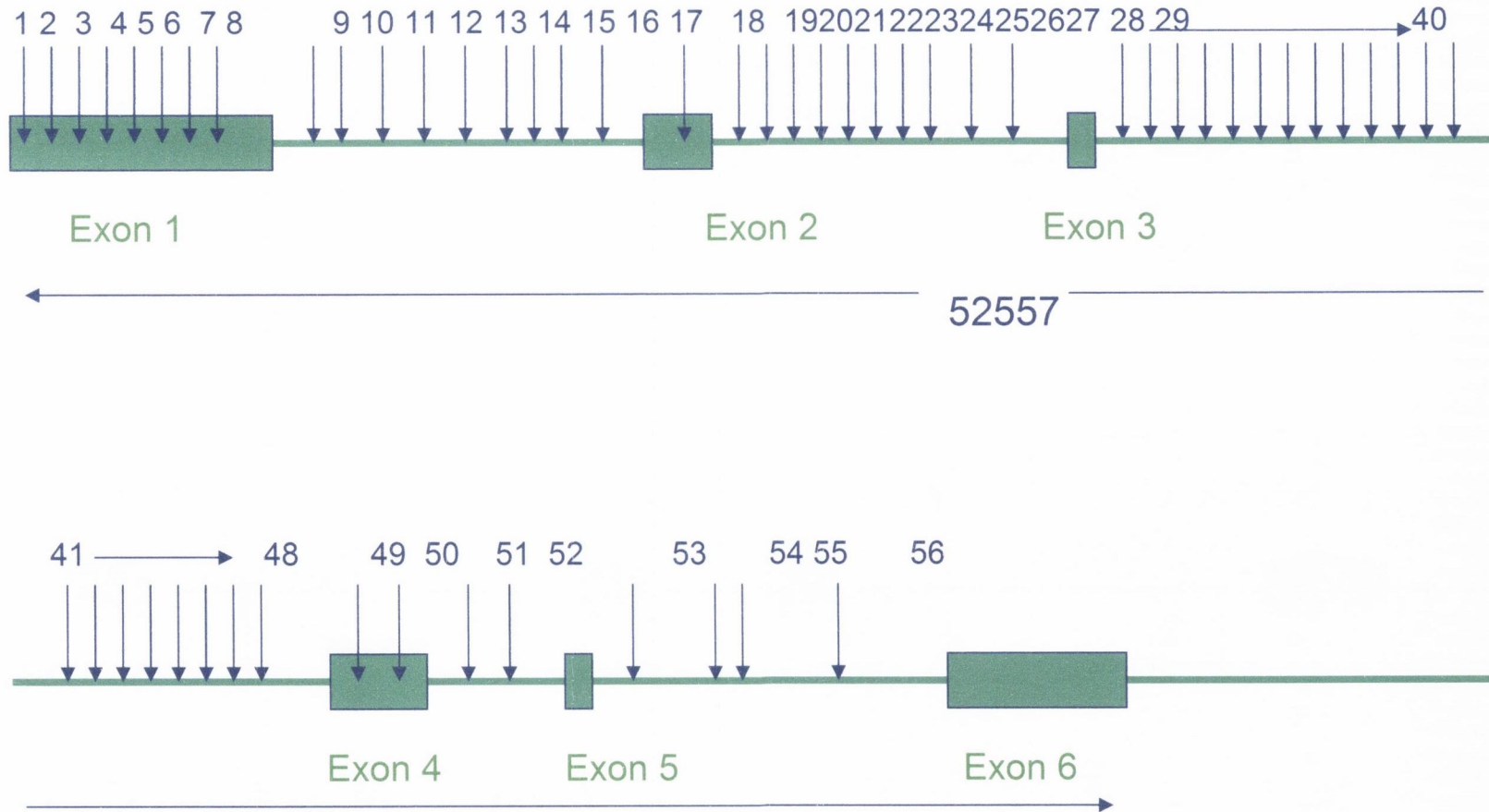
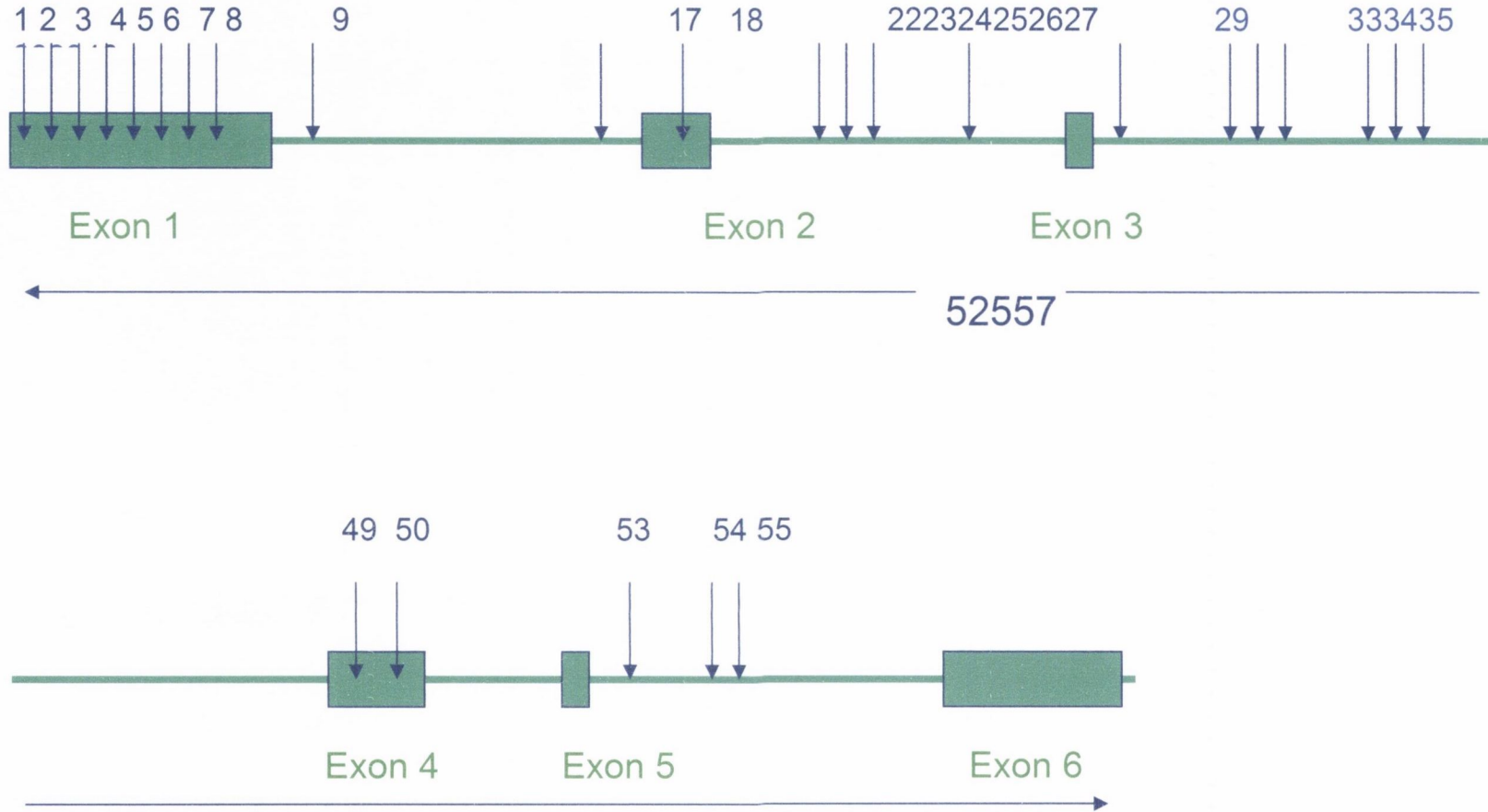


Figure A.7: Schematic of the SNPs identified from dbSNP spanning APOL4.

Apolipoprotein IV



Apolipoprotein IV

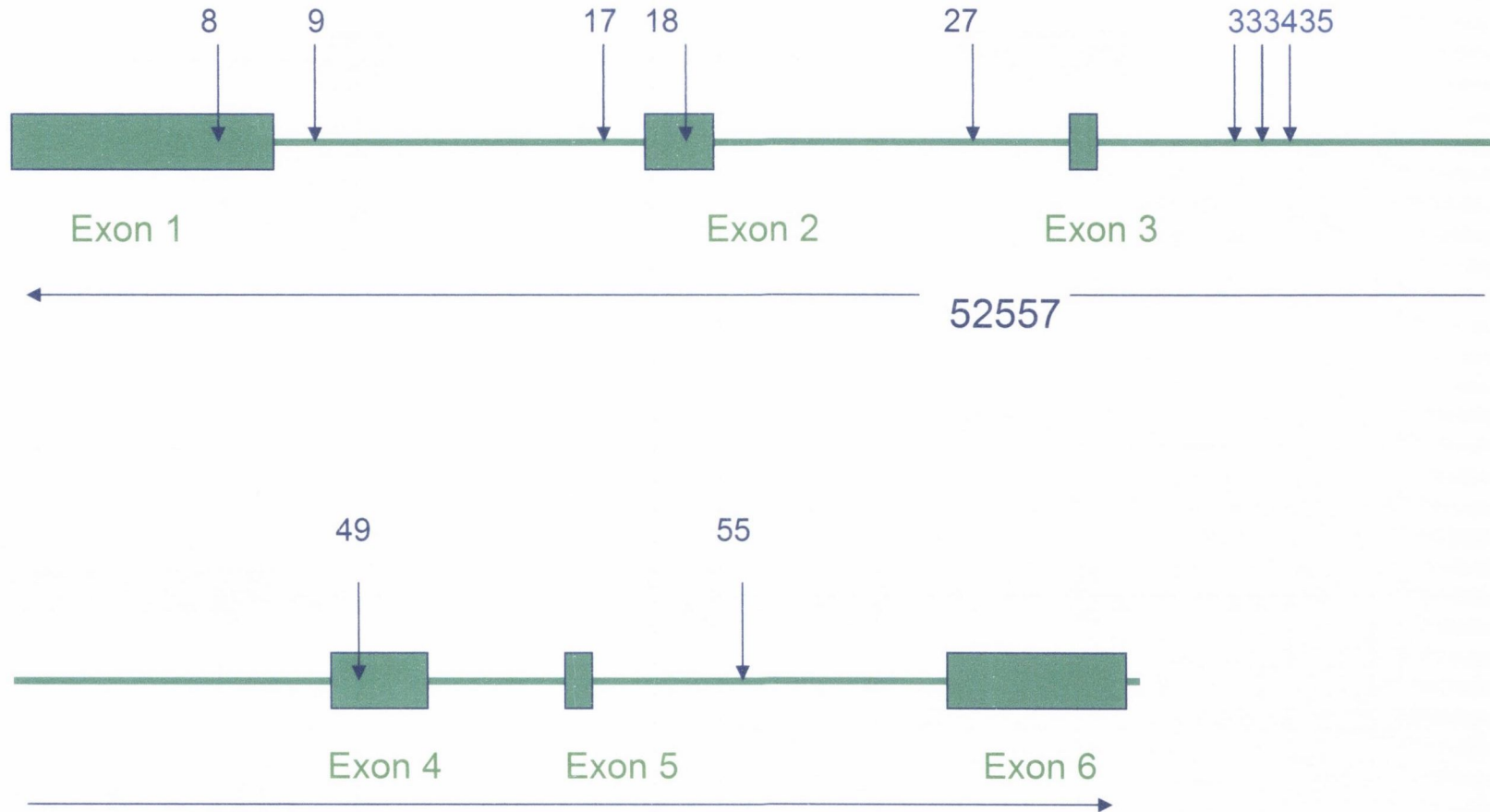


Figure A.9: Schematic of the SNPs that were heterozygous in the Irish Sample. Association analysis was carried out on all of these SNPs using DNA pooling techniques.

Apolipoprotein II

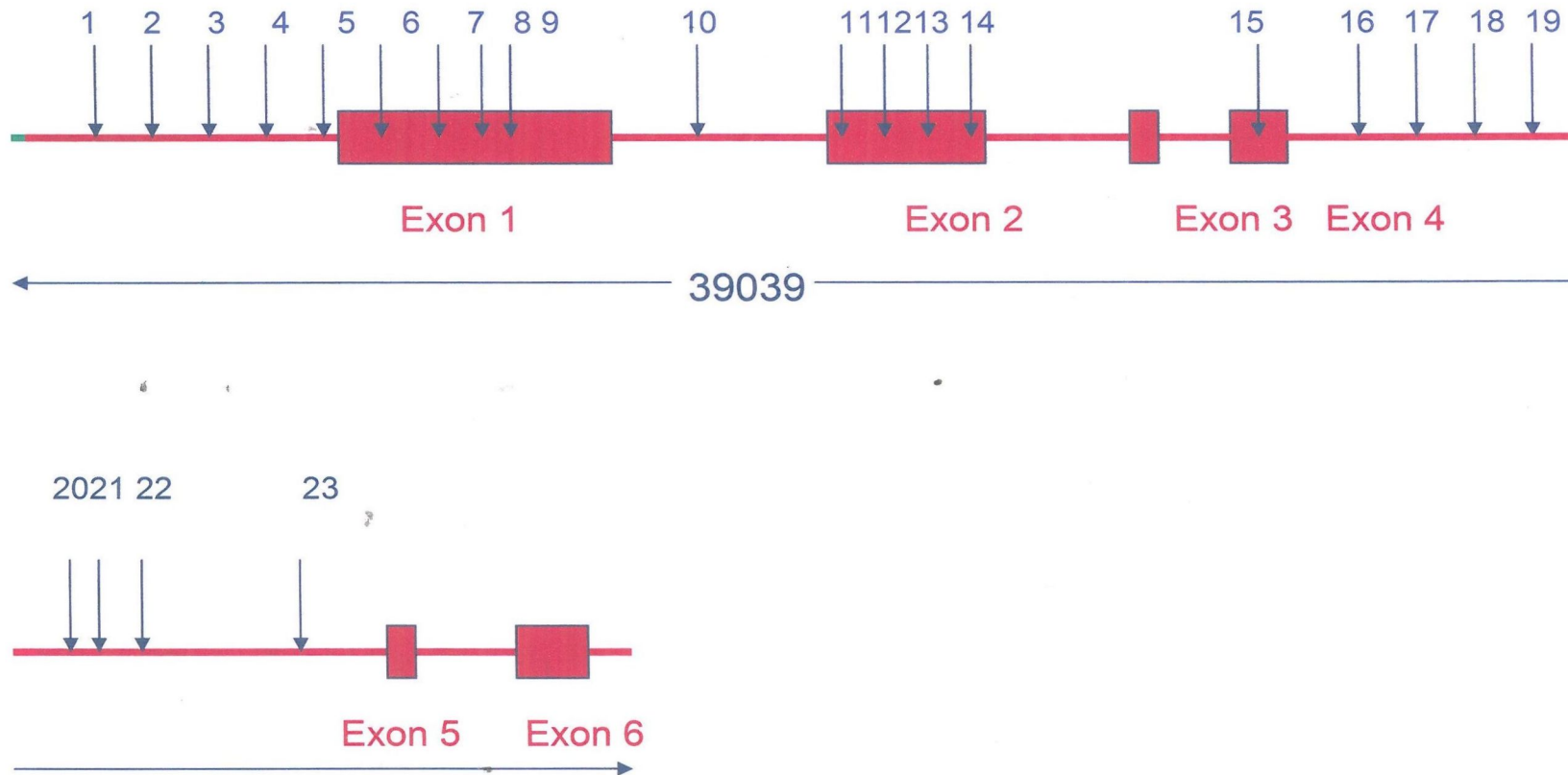


Figure A.10: Schematic of the SNPs identified from dbSNP spanning APOL2. All these SNPs were selected for initial pooling studies to identify heterozygosity in the Irish sample for APOL2.

Apolipoprotein II

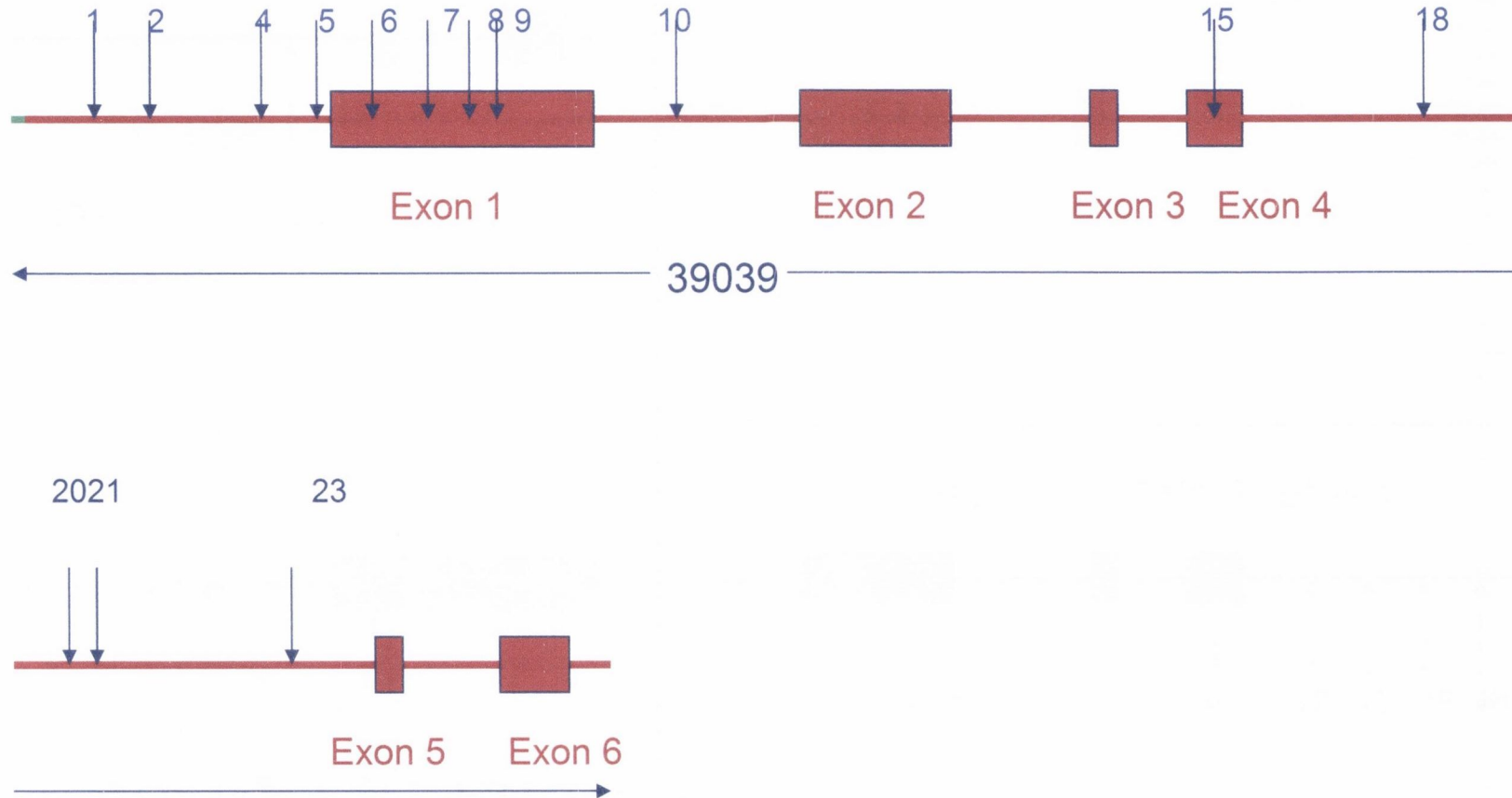


Figure A.11: Schematic of the SNPs that were heterozygous in the Irish Sample. Association analysis was carried out on all of these SNPs using DNA pooling techniques.

Table A.1 - APOL 6 table PCR fragment information

Fragment Name	PCR Primer Sequence	PCR Fragment Size (b.p.)	PCR Protocol
Apol6-1+2+3	F: AGTGGCCAGCACATAGTAGG R: ATACATGGACACACGCACAC	356	57Q + 3.0 MgCl ₂ /0.3U Taq
Apol6-4+5	F: CCCTGGCCACATTTACTATC R: AGTCTGCAAATGTGGGTTTT	255	57Q
Apol6-6	F: AGGCAAGGTTTGCTATGTGC R: GGCTTCAGGGATTCAGCTC	214	TQ60
Apol6-7+8+9	F: TGCAAAGTGACCAAGAAAAA R: CTCAGCCTTCAGGAATGAGT	352	57Q
Apol6-10+11	F: CCTTACACAGGCATGATGG R: TTAGCCACACCCGAGTTTAT	204	57R
Apol6-12+14	F: TTTACTGCTCTCATCGCTCA R: CATTTTCACCAAACAAAGCA	267	57R + 2.5 MgCl ₂
Apol6-15	F: NGGAAGCTGCGTAACAAAAC AGGATCCTTGTAACCCAAGAC R:	294	57Q
Apol6-16	F: AACCCACAAGAAATTCACCA TTGGGCTTCTTTATTTTGG R:	211	57R
Apol6-17+18+19	F: GGAAGGGGGTCTTATTCCTA R: TGAGTTCTGAGATCCTCTCCA	352	57Q
Apol6-20	F: TTCTTGTTTCTATGTGCTGGAA R: AACAGGGCTTGTCTCTCTGA	214	57Q

Table A.2 - APOL 5 table PCR information

Fragment Name	PCR Primer Sequence	Fragment Size (b.p.)	PCR Protocol
Apol5-1+2	F: TTTCATCCACTTTTCTATTGG R: TAACCGAACTCCCAAATGAC	291	T65
Apol5-4+5+6	F: TTTCATGAACCTCAGATCCAG R: CTGGGCCAAAAACAAATCTA	482	57Q
Apol5-7+8	F: CATCCTTCCAAAAGCCATAG R: GCTATATCTCCACCCAGCAA	616	57Q
Apol5-9+10	F: TGAGTCTTGCCTTGAGATCC R: AGCATTGCAACCCACTTTAG	185	57Q
Apol5-11	F: AGCTATCCAGGGCATCAAG GGAAGCTGCTCATGTCTTTC R:	227	62Q
Apol5-12+13	F: AAAGACATGAGCAGCTTCCT R: AAGGGGCAGCTGTAGTTATG	329	57Q
Apol5-14	F: CAGGGACTCATGTTCCACA R: TTCCTTGCGTCCCACTTA	221	57Q
Apol5-15	F: CTCTGTA CTTCCCTGCCAGATG R: CAGGGTGACCAACTAGGGTA	273	57Q
Apol5-16+17+18	F: GGCCTGTGCAACCTAAAATA R: AGCACTGAACCTACCTCAA	284	57Q
Apol5-19+20			

Table A.3 - APOL 4 table PCR information

Fragment Name	PCR Primer Sequence	PCR Fragment Size (b.p.)	PCR Protocol
Apol4-1+2+3	F: GAGCTGAAAATGGAGAAAAGC	494	Dropped
	R: ACTTACACACAGGGCACTCA		
Apol4-4+5+6	F: ATTGAAAAGGTCCACAGAGG	380	57Q
	R: CGTCCAACAGTGGCTTTAG		
Apol4-7+8+9	F: GACACGAAGCCTTTCTATGG	645	57Q
	R: TGTTGTACAAAACATATTTAACCTTG		
Apol4-17+18	F: TATTCTTCCTGCACTGCTGA	236	59Q
	R: CAGAAGAAAGTTAGCCCAGTG		
Apol4-22+23+24	F: GACCACTCCCAGCAG	922	Dropped
	R: GACCACTCCCAGCAG		
Apol4-27	F: TTGCTAGGTGTCAGGGTAGG	328	59Q
	R: TCTTCCTCAAGGGTGCAG		
Apol4-29	F: CTCAGGCTGTCTTGAActCC	830	Dropped
	R: GAGATGGAGTCTCGCTCTGT		
Apol4-33+34+35	F: TGAACAAAGGAGAAGGAGGA	380	57Q
	R: GGGCCTGAAATGGTATTGT		
Apol4-38+39+40	F: TCTCAGGGTTCAGACACACA	321	57Q
	R: TCCCAGTCTACTGGCTCAT		
Apol4-49	F: ATCCTCCTGGTCATTGTTG	296	57Q
	R: AGCTGCTTTGCTGAAAATCT		
Apol4-50	F: ATCCTCCTGGTCATTGTTG	401	57Q
	R: AGCTGCTTTGCTGAAAATCT		
Apol4-53+54+55	F: TAAACCGTTTTGCTGTGTCA	1580	T60 + 1.5 MgCl ₂
	R: ATAGAGCTGCCAAAAGTCCA		

Table A.4 - APOL6 extension primers

SNP	Compliment	Size (b.p.)	Extension Primers
1 = dbSNP932468 A/G	F	23	5' AACAAACACTGACTGAATTTTCAG 3'
2 = dbSNP932469 G/T	F	20	5' AGAGCGCATGTGTGTGTGTG 3'
3 = dbSNP926754 C/G	F	17	5' TGTGTAAGTGTGTGTGC 3'
4 = dbSNP2103768 G/C	F	20	5' ATTTCAAATCATTGTATT 3'
5= dbSNP762910 C/T	F	17	5' TTGCATCCCAACTCTGC 3'
6= dbSNP1883986 C/T	F	17	5' AGTCCTGCAAAGCACTT 3'
7 = dbSNP1894469 A/G	F	17	5' ACAAAAAATATACATTG 3'
8 = dbSNP1894468 C/G	F	20	5' CATCTCTGTCGCTTTGTATC 3'
9 = dbSNP2413359 A/G (C/T)	R	17	5' AAAACAAGGGGATCTCC 3'
10 = dbSNP1540294 A/C	F	17	5' TGGTAACTCAATCTCC 3'
11 = dbSNP1540295 C/T	F	20	5' TTGTTTCCCCAGACAAGCAG 3'
12 = dbSNP2010168 A/C	F	20	5' CTCTCTCTCTCTCTCTCT 3'
14 = dbSNP1001325 C/T	F	17	5' ACACATATATATACACA 3'
15= dbSNP1540296 C/T (A/G)	R	23	5' AATAATTCTCAAAGGAGTGGCTT 3'
16= dbSNP2413360 C/T (A/G)	R	23	5' TTGACCAGCGGTGGAGAGCAGCA 3'
17 = dbSNP2413364 C/T	F	15	5' TAGTCCCTACTGCTG 3'
18 = dbSNP2213390 G/C	F	20	5' GGAAGTTAACTTTAAAATA 3'
19 = dbSNP2213391 A/C	F	17	5' AAGAACAGGGGAGCTGA 3'
20 = dbSNP728816 G/T	F	17	5' CTCCTGCAGGGAGGGG 3'

Table A.5 - APOL5 extension primers

SNP	Compliment	Size (b.p.)	Extension Primers
1= dbSNP1997883 A/G	F	17	5' TTATAGGAATTCTTCAC 3'
2= dbSNP1997884 A/G	F	20	5' GAGTAAAGGAGGAGAGTGGA 3'
4= dbSNP1970778 C/T(G/A)	R	23	5' TGGATCACTTGAGGTCAGGAGTT 3'
5= dbSNP1970777 C/T	F	17	5' AGAATCTTGCTCTGTCA 3'
6= dbSNP1540297 A/G	F	20	5' TAGGCACTTCTGGAATTCTC 3'
7= dbSNP2413369 A/G	F	20	5' ATCGCAAAAAAAAAAAAAAAAAA 3'
8= dbSNP2899256 C/T	F	17	5' AGAAAAAATACGTTACA 3'
9 = dbSNP2009168 A/G	F	20	5' TCCCACAGAGCTCAGCTGCA 3'
10= dbSNP2009169 C/T	F	17	5' CTGCTGCCTAACAGAGG 3'
11= dbSNP2076671 T/C	F	17	5' ACACATCCCTTTCTGGA 3'
12= dbSNP2076671 C/T	F	17	5' ACACATCCCTTTCTGGA 3'
13= dbSNP2073198 A/G	F	20	5' CCATGAGGGTGGGGGGCGAC 3'
14= dbSNP2076673 G/C	F	20	5' GACACCAAAGAGGACAGTCT 3'

15= dbSNP2016586 G/T	F	17	5' TGAATAAGACAGAAAAG 3'
16= dbSNP879680 C/T	F	20	5' TTTTGGAACTGTGCGAGAC 3'
17= dbSNP2024641 A/G	F	23	5' TGTGAGCCGGGAAACCAATTTCC 3'
18= dbSNP2024642 C/T	F	17	5' GGAAACCAATTTCCGAA 3'

Table A.6 - APOL4 extension primers

SNP	Compliment	Size (b.p.)	Extension Primers
1 = dbSNP 12781 G/A (C/T)	R	20	GTTTAGACATTGGGGGAGGC
2 = dbSNP 1053983 A/G	F	20	CCACTCTCCC TTGTCCTCCC
3 = dbSNP 1053982 A/G	F	23	TCCTTCCCTGGTGATGGTCTCTC
4 = dbSNP 2227167 A/G	F	23	GCTTCGTCAAAATCAAGTGCAAA
5 = dbSNP 2227168 C/T	F	17	CATTGGGTGTGATGTCA
6 = dbSNP 132700 T/C	F	20	TGCTGTAAATGGTGCCAACA
7 = dbSNP 3075364 C/T(G/A)	R	23	CAGTTTAGGGAGTGGTTTTTGAA
8 = dbSNP 2227169 C/T	F	20	GCTTCATAGAGAGCATCTGC

9 = dbSNP 2097466 C/T(G/A)	R	17	CAAGCATGAGCCACCGC
17 = dbSNP 132704 T/C	F	17	GGGCTGCCTGGAGGAGG
18 = dbSNP 2007468 A/G	F	20	GGCAATTCAGCCACACGCAC
22 = dbSNP 132707 T/C	F	17	CGTTAGTAAACCACAGA
23 = dbSNP 132708 G/C	F	23	GAGGGAGAGCTTCTTCCTTGCC
24 = dbSNP 132709 A/T	F	20	GGATTCTGGGATTGGCTGAG
27 = dbSNP 132712 A/G	F	17	CTTGGGGCAGCCTCATC
29 = dbSNP 132714 T/C	F	20	GTCGCCACATTCGGCTAATT
33 = dbSNP 132719 T/G	F	17	TTGTCATTTTTTAAAGA
34 = dbSNP 132720 A/G	F	20	AGCTACAAAGTTGCTAATGG
35 = dbSNP 132721 G/A	F	23	GAATGTTAGGAAAAGGTGGAGAG
38 = dbSNP 132724 -/G	F	17	CCCAGTCCTGGGCAGCA
39 = dbSNP 132725 G/A	F	23	ACCAGGGGAGTATGCAGAGGGGC
40 = dbSNP 132726 G/C	F	20	CAGGAGAGGGGCATCCCTTC
49 = dbSNP 132734 G/A	F	20	TTAGCCTCAACTAGGACACA
50 = dbSNP 132735 G/T	R	17	GACTCGCCTAGAGGGGA
53 = dbSNP 132737 G/C	F	20	TTTTAATAAACCGTTTTGCT

54 = dbSNP 132738 T/C	F	17	AGCATCTCCTGTGTTTC
55 = dbSNP 1807673 A/G	F	23	CCTGGGCAACAGAGCGAGACTCC

Table A.7: PCR program cycles for T;Q; and R programs

<i>TX program</i>		
°C	Time	
95	3 min	
95	15 sec	
X	15 sec	
72	30 sec	
95	15 sec	
X-1	15 sec	
72	30 sec	
95	15 sec	
X-2	15 sec	
72	30 sec	
95	15 sec	
X-3	15 sec	
72	30 sec	
95	15 sec	29 cycles
X-4	15 sec	
72	30 sec	
72	10 min	

Table A.8: PCR program cycles for T;Q; and R program cont/d.

<i>XQ program</i>		
95	3 min	
95	20 sec	41 cycles
X	20 sec	
72	30 sec	
72	10 min	

<i>XR program</i>		
°C	Time	
95	3 min	
95	20 sec	35 cycles
X	20 sec	
72	30 sec	
72	10 min	