



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.

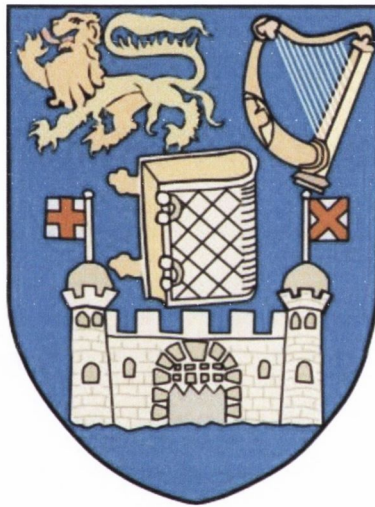
Access Agreement

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.

Characterisation of *MACROD2*, a putative autism susceptibility gene: Translating Genome-Wide Analysis to Causation

Graham Kenny



**A thesis submitted for the degree of
Doctor of Philosophy at the
University of Dublin**

**Department of Psychiatry
School of Medicine
Trinity College Dublin**

2013

Declaration

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work.

I agree to deposit this thesis in the University's open access institutional repository or allow the library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

GRAHAM KENNY

GRAHAM KENNY

Graham Kenny



Thesis 10244

Summary

Genome Wide Association studies (GWAS) have been used to identify genetic markers which have strong statistical association with a disease. Establishing a functional consequence of association is required to determine a causal link between marker and disease. In a recent GWA of autism the Autism Genome Project identified a strong association within the gene *MACROD2*. Very little is known about the function of *MACROD2*, although there are suggestions in the wider literature associating variation within *MACROD2* with neuropsychiatric disorders. This project examined the biological role of *MACROD2*, with the hypothesis that disruption of *MACROD2* will impact on the structure and/or function of the neuron.

Chapter 3 describes this effort to understand the basic biology of this gene; including investigation of the **structure** of the gene, characterise the **expression** profile, and finally to explore the role of **variation** on the expression of this gene, in the hopes that it will inform further analysis. I have documented the expression of *MACROD2* isoform 1 in the human brain, and documented the expression profile of *MACROD2* and *MACROD2-AS1* across a panel of 21 human tissues. AEI experiments recorded allele expression imbalance of *MACROD2* and an in-depth analysis of CNV at the *MACROD2* locus identified a rare exonic deletion that segregates with severity of autism diagnosis. eQTL analysis demonstrates that variation at the *MACROD2* locus may be influencing downstream gene expression.

Chapter 4 represents a functional study of a protein that has not been characterised in humans and may be a protein important to human health. In this project I have sought to answer some of the fundamental questions about the **localisation**, **interaction** and **function** of the MacroD2 protein. I have confirmed the localisation of *MACROD2* to be present in the nucleus and cytoplasm. Data presented here suggests that *MACROD2* interacts with multiple sets of proteins, with functions in cell adhesion and the cytoskeleton, chromatin modulation, DNA damage repair and transcriptional regulation. Preliminary siRNA knockdown experiments suggest that decreased *MACROD2* expression affects the expression of neuronal markers and overall cell morphology, including neurite outgrowth.

Finally in Chapter 5, the key aim was to **refine** the *MACROD2* association signal, using the AGP Stage 1 and Stage 2 data for the *MACROD2* region and **fine-mapping** of the locus by **imputation**, alongside a more detailed interrogation of genetic model and clinical subgroups to better describe the original *MACROD2* association. From this analysis it appears that the AGP Stage 1 sample is driving the association signal, although breaking down this sample into smaller subgroups based on IQ or verbal status was not found to improve the association signal. I investigated whether genes which interact with *MACROD2* show association enrichment across the GWAS sample – thereby assessing whether common biological pathways which include *MACROD2* interactors are important in ASD. This analysis identified an enrichment for *USP33*, a carboxyl-terminal hydrolase.

The work reported here represents an in depth analysis of a poorly characterised gene that has been associated with autism, and thus this report may have relevance to human health.

The author acknowledges grant support from the HRB.

Acknowledgements

I would like to take this opportunity to thank Dr. Richard Anney in his role as my supervisor for the past three years, for the countless hours of explanation and thoughtful discussion throughout this project, but most of all for his patience and sarcastic wit. I have grown considerably over the past few years, both in terms of my appreciation of science and personally in terms of my outlook on the world. For your role in that Ric, I thank you.

I want to thank everyone from Lab 1.10 for making it such a great place to work, for our work together planning experiments to our discussions about science and the often strange and frustrating world around us. I thank Dr. Emma Quinn for her help with planning my AEI experiments and Dr. Alison Merikangas for always racking my brains about statistics. Lab 1.10 wouldn't be the same without Ciara Fahey, or quoting the I.T. Crowd and Arrested Development at every opportunity. Thank you to Dr. Ines Molinos for instructing me in the dark art of immunohistochemistry and Spanish cuisine. I especially want to thank Dr. Sarah Furlong and Stefania Bellini for their friendship and for always lending an ear to my scientific or life drama.

Thanks go to the 'grown-ups' office, Dr. Elaine Kenny, Dr. Derek Morris and Dr. Paul Cormican for making time for me when I had a question that required their expertise, and to Dr. Carlos Pinto for managing our servers and complaining about Microsoft operating systems. Thank you to Dr. Daniela Tropea for your instruction with confocal imaging.

Thank you to all the new faces in the group, I hope that you will continue to make Lab 1.10 a great place to work. To Carol, Siobhan, Ronan, Jennie and Kristin, I wish you the best in your endeavours for the future. To all those who went before me, Matt, Katie, Lynne, Ziarah, Will and Elif, I thank you for reminding me that this project is but one stop on my journey. To Geraldine, thank you for making our lives so much easier. Thank you to Prof. Michael Gill, Dr. Aiden Corvin and Dr. Gary Donohoe in their roles as head of the Neuropsychiatric Genetics Group. Thank you to Dr. Ross McManus, Dr. Michael Freeley, Dr. Paula Murphy and Sheila Walshe in their roles in the PhD in Molecular Medicine.

I thank the Health Research Board for funding this work, and to all the families who donated their genetic material to us so that we may push those boundaries a little further.

This has been an invaluable experience.

Statement of Work

This work was undertaken by the author with the Neuropsychiatric Genetics research group, Department of Psychiatry and Institute of Molecular Medicine, Trinity College Dublin.

List of Abbreviations

5' RACE	5' Rapid Amplification of cDNA Ends
ADI-R	Autism Diagnostic Interview-Revised
ADOS	Autism Diagnostic Observation Schedule
AEI	Allele Expression Imbalance
AGP	Autism Genome Project
ASD	Autism Spectrum Disorder
BDNF	Brain Derived Neurotrophic Factor
cDNA	Complementary Deoxyribonucleic Acid
CEPH	Utah Residents with Northern and Western European Ancestry
CNV	Copy Number Variation
Co-IP	Co-Immunoprecipitation
DAPI	4',6-Diamidino-2-Phenylindole
dATP	Deoxy Adenosine Triphosphate
DAVID	Database for Annotation, Visualisation and Integrated Discovery
DGV	Database of Genomic Variants
DMEM	Dulbecco's Modified Eagle Medium
DNA	Deoxyribonucleic Acid
EGFP	Enhanced Green Fluorescent Protein
eQTL	Expression Quantitative Trait Locus
FBS	Fetal Bovine Serum
<i>FLRT3</i>	Fibronectin leucine rich transmembrane protein 3
GI	Gene Info identifier
GO	Gene-Ontology
GRR	Genotype Relative Risk

GRT	Gene Ratio Test
GWAS	Genome Wide Association
HGNC	Human Genome Nomenclature Committee
INRICH	INterval enRICHment analysis
LB	Lysogeny broth
LC/ESI-MS	Liquid Chromatography/Electrospray Ionization-Mass Spectrometry
LOD	Logarithm (base 10) of odds
<i>MACROD2</i>	Macro domain containing 2
<i>MACROD2-AS1</i>	<i>MACROD2</i> Anti-Sense 1
MAF	Minor Allele Frequency
mRNA	Messenger Ribonucleic Acid
ncRNA	non-coding Ribonucleic Acid
NPG	Neuropsychiatric Genetics
PBS	Phosphate Buffered Saline
PCR	Polymerase Chain Reaction
PDD	Pervasive Developmental Disorders
QC	Quality Control
RA	Retinoic Acid
RIPA	Radioimmunoprecipitation assay buffer
RNAi	Ribonucleic Acid Interference
RNAseq	Ribonucleic Acid Sequencing
SAGE	Serial Analysis of Gene Expression
siRNA	Small interfering Ribonucleic Acid
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variants
SOC	Super Optimal broth with Catabolite repression

SRT	SNP Ratio Test
TAE	Tris base, Acetic acid, EDTA
TBS-T	Tris-Buffered Saline and Tween-20
TUCC	Transmitted Untransmitted Case Control

Table of Contents

Declaration..... ii

Summary iii

Acknowledgements..... v

Statement of Work..... vi

List of Abbreviations vii

Table of Figures and Tables..... xv

Chapter 1 Introduction 1

1.1 Autism 1

1.2 Genetic Aetiology of Autism 2

1.2.1 Linkage Studies..... 4

1.2.2 Common and Rare Variant hypotheses of Autism..... 5

1.3 Investigating the Genetic Aetiology of Autism 6

1.3.1 Introduction to Genome-Wide Association Studies 6

1.3.2 Genome-Wide Association Studies in Autism..... 8

1.3.3 CNV studies of Autism..... 12

1.3.4 Sequencing studies of Autism 12

1.3.5 Epigenetic studies of Autism..... 15

1.4 Translation of Association to Function 15

1.5 MACRO-domain containing 2 (Gene: *MACROD2*, Protein: MacroD2) 16

1.5.1 *MACROD2* Anti-Sense 1 – A non-coding RNA 18

1.6 Aims of this project 18

Chapter 2 Materials and Methods 20

2.1 Molecular Biology Techniques 20

2.1.1 Quantification of Nucleic Acids and Protein 20

2.1.2	Polymerase Chain reaction	20
2.1.3	Restriction Digest of Vector Plasmid.....	22
2.1.4	Agarose Gel Extraction by Wizard SV Gel and PCR Clean-Up System (Promega).....	24
2.1.5	DNA Sequencing Protocol on ABI 3130xl Genetic Analyser.....	24
2.1.6	Ligation Reaction.....	24
2.2	Bacterial Cell Culture.....	25
2.2.1	Standard Transformation Protocol	25
2.2.2	Plasmid purification.....	26
2.3	Mammalian Cell Culture.....	27
2.3.1	Standard Cell Culture	27
2.4	Transfection of plasmid DNA into mammalian cell lines	28
2.4.1	Transfection Protocol	28
2.4.2	Luciferase Assay	28
Chapter 3	Investigation of the gene structure and expression of <i>MACROD2</i> in the human brain.....	30
3.1	Introduction.....	30
3.2	Aims of this study	34
3.2.1	Characterisation of the Structure of <i>MACROD2</i>	34
3.2.2	Characterisation of the Expression of <i>MACROD2</i>	35
3.2.3	Characterisation of the Role of Variation on the Expression of <i>MACROD2</i>	35
3.3	Materials and Methods	37
3.3.1	Characterisation of the Structure of <i>MACROD2</i>	37
3.3.2	Characterisation of the Expression of <i>MACROD2</i>	41

3.3.3	Characterisation of the Role of Variation on the Expression of <i>MACROD2</i>	43
3.3.4	Preparation of data for GWA and eQTL analysis: Quality Control of the Myers dataset.....	46
3.3.5	Preparation of data for GWA and eQTL analysis: Data pruning	48
3.4	Results	51
3.4.1	Characterisation of the Structure of <i>MACROD2</i>	51
3.4.2	Characterisation of the Expression of <i>MACROD2</i>	55
3.4.3	Characterisation of the Role of Variation on the Expression of <i>MACROD2</i>	57
3.5	Discussion.....	67
3.5.1	Characterisation of the Structure of <i>MACROD2</i>	67
3.5.2	Characterisation of the Expression of <i>MACROD2</i>	68
3.5.3	Characterisation of the Role of Variation on the Expression of <i>MACROD2</i>	70
Chapter 4 Investigation of the function of MacroD2 in the human brain.....		77
4.1	Introduction.....	77
4.1.1	Functional genomics studies in neuropsychiatric genetics.....	77
4.1.2	<i>MACROD2</i>	79
4.1.3	<i>MACROD2</i> and Autism Spectrum Disorder	82
4.1.4	Aims of this study	84
4.2	Materials and Methods	85
4.2.1	Localisation.....	85
4.2.2	Interaction	89
4.2.3	siRNA knockdown.....	95
4.3	Results	99

4.3.1	Localisation.....	99
4.3.2	Interaction	102
4.3.3	siRNA knockdown of <i>MACROD2</i> in SH-SY5Y cell lines: A preliminary study	110
4.4	Discussion.....	125
4.4.1	Localisation.....	125
4.4.2	Interaction	125
4.4.3	siRNA knockdown of <i>MACROD2</i> in SH-SY5Y cell lines: A preliminary study	128
4.4.4	Cell Model	129
4.4.5	Limitations of this work and future directions.....	130
Chapter 5 Fine-mapping of an association signal at the <i>MACROD2</i> locus		132
5.1	Introduction.....	132
5.1.1	The effect of sample size on the power of GWAS	133
5.1.2	The effect of clinical heterogeneity on the power of GWAS	134
5.1.3	The effect of SNP coverage on the power of GWAS.....	135
5.1.4	Missing Heritability.....	136
5.1.5	Pathway analysis	137
5.2	Aims of this study	138
5.3	Materials and Methods	139
5.3.1	Genotyping	139
5.3.2	Genetic Ancestry	139
5.3.3	Clinical evaluation	140
5.3.4	Imputation.....	141
5.3.5	Quality Control of GWAS	142
5.3.6	Generating the case pseudo-control dataset from trio data	143

5.3.7	Genotype imputation	144
5.3.8	Reference panels.....	145
5.3.9	Genetic Association Analyses.....	145
5.3.10	Association analysis using imputed data (dosage/threshold)	146
5.3.11	Enrichment analysis of putative protein-protein interactors identified via Co-IP experiments.....	146
5.4	Results	147
5.4.1	The effect of SNP coverage on the power of GWAS - Imputation of the <i>MACROD2</i> locus	147
5.4.2	The effect of sample size and clinical heterogeneity on power of GWAS.....	150
5.4.3	The effect of inheritance model on power of GWAS	151
5.4.4	Enrichment analysis of putative protein-protein interactors identified via Co-IP experiments.....	154
5.5	Discussion.....	156
5.5.1	The effect of SNP coverage on the power of GWAS - Imputation of the <i>MACROD2</i> locus	156
5.5.2	The effect of sample size and clinical heterogeneity on power of GWAS.....	159
5.5.3	The effect of inheritance model on power of GWAS	160
5.5.4	Enrichment analysis of putative protein-protein interactors identified via Co-IP experiments.....	161
5.5.5	Summary and future directions	161
Chapter 6	Discussion of <i>MACROD2</i> as a susceptibility gene for autism.....	163
References		172

Table of Figures and Tables

Figure 1.1 Published Genome-Wide Associations	7
Figure 1.2 Association plot for region containing rs4141463 in the AGP Stage 1 GWAS.....	9
Figure 1.3 Association plot for region containing rs4141463 in the AGP Stage 2 and combined Stage 1 + 2 GWAS.....	11
Figure 1.4 The <i>MACROD2</i> locus contains 2 major intergenic transcripts on the anti-sense strand	17
Figure 2.1 DNA ligation	23
Table 1.1 Restriction Digest - Reagents and protocol used.....	23
Table 2.2 DNA Ligation - Reference guidelines for ligation protocols used for given reagents/constructs	25
Figure 3.1 The top association from the Autism Genome Project Stage 1 GWAS study	31
Table 3.1 The <i>MACROD2</i> locus contains multiple genes related to transcriptional regulation.....	32
Table 3.2 The evolutionary conservation of <i>MACROD2</i>	33
Figure 3.2 A closer look at the <i>MACROD2</i> locus.....	34
Figure 3.3 5' RACE.....	38
Table 3.3 The oligonucleotide nested primers used in the 5' RACE experiments.....	38
Table 3.4 Sequencing reaction used with the ABI 3130XL Genetic Analyser.	40
Table 3.4 Tissue Sources for the Human Total RNA Master Panel II.....	42
Table 3.5 Minor Allele Frequency and number of representative heterozygotes for each SNP assayed for AEI.....	44
Figure 3.5 TaqMan Allelic Discrimination Reaction	44
Figure 3.6 Comprehensive QC using PLINK was performed to prepare the Myers dataset for use in subsequent analysis.....	47
Figure 3.7: A multidimensional scaling matrix plot for 193 samples derived from brain tissue in the context of the HapMap populations	48
Figure 3.8 An example of the output of RNAseq data from Ensembl.org.	52
Table 3.6 RACE experiments identify expression of <i>MACROD2</i> isoform 1.....	53
Figure 3.9 The observed sequence from the 5' RACE experiments indicates expression of isoform 1 but not isoform 2	53

Figure 3.10 (a) Mapping of the putative *MACROD2* promoter region using inserts of increasing size. (b) Differences in the activity of each putative promoter insert in a luciferase reporter assay indicate the location of regulatory elements upstream of *MACROD2*.....54

Table 3.7 Top hits from bioinformatic search of regulatory regions from the putative *MACROD2* promoter region.....55

Figure 3.11 Exon-crossing primers were used to demonstrate the expression profile of *MACROD2*.....56

Figure 3.12 Expression profile of *MACROD2* across different bioinformatic resources56

Figure 3.13 Allele Expression imbalance is observed for heterozygous coding SNPs in *MACROD2*.57

Table 3.8 Validation of missense mutation reported in *MACROD2*.58

Figure 3.14 Pattern of copy number variation in the AGP probands59

Figure 3.15: Example of a de novo CNV deletion in *MACROD2* from the AGP Collection ..60

Figure 3.16 CNV from the DGV (Control) database overlapping or downstream of exon 5 of *MACROD2*61

Table 3.9 Top unique association signals from human cortical gene expression *MACROD2* trans eQTL analysis62

Table 3.410 Analysis of genes identified from the eQTL analysis were analysed using the Gene Ratio Test.....65

Table 3.411 Analysis of genes identified from the eQTL analysis were analysed using DAVID.....66

Figure 3.17 CNV at the *MACROD2* locus from the DGV.72

Figure 4.1 Differentiated SH-SY5Y cells in culture.86

Table 4.1 Composition of the RIPA buffer.90

Table 4.2 Composition of the resolving and stacking gel for Western Blot.91

Table 4.3 HPLC gradient parameters as determined by ProteaBio for the liquid chromatography analysis.....94

Table 4.4 Peptide chains tagged with Uniprot protein accession number IDs94

Figure 4.2 Expression of neuronal markers in SH-SY5Y cell lines.....100

Figure 4.3 Localisation of *MACROD2* in SH-SY5Y neuroblastoma cell lines, visualised using confocal microscopy101

Figure 4.4 Western blot of *MACROD2* Co-IP from SH-SY5Y.....103

Figure 4.5 Silver staining of whole cell Co-IP from SH-SY5Y cell lines for *MACROD2*..... 104

Table 4.5 Top hits from the LC/ESI-MS analysis of Co-IP pulldown 1, from SH-SY5Y whole cell protein extract. 106

Table 4.6 Top hits from the LC/ESI-MS analysis of CoIP pulldown 2, from SH-SY5Y whole cell protein extract. 107

Table 4.7 Top hits from the LC/ESI-MS analysis of CoIP pulldown 2, from the cytoplasmic fraction of SH-SY5Y protein extract. 107

Table 4.8 Top hit from the LC/ESI-MS analysis of CoIP pulldown 2, from the nuclear fraction of SH-SY5Y protein extract. 108

Table 4.9 Analysis of genes identified from the LC/ESI-MS were analysed using DAVID.. 109

Figure 4.6 The fold change in gene expression for SNP1 (rs16997016). 111

Figure 4.7 The fold change in gene expression for SNP2 (rs11696319). 112

Figure 4.8 The fold change in gene expression for SNP3 (rs1238278). 113

Figure 4.9 The fold change in gene expression for SNP4 (rs1225891). 114

Figure 4.10 The fold change in gene expression for SNP5 (rs6080109). 115

Figure 4.11 Expression of *MACROD2* and neuronal markers in untreated SH-SY5Y cell lines..... 117

Figure 4.12 Expression of *MACROD2* and neuronal markers in SH-SY5Y cell lines treated with 'scrambled' siRNA (Negative control). 118

Figure 4.13 Expression of *MACROD2* and neuronal markers in SH-SY5Y cell lines treated with siRNA against *PPIB* (Positive control)..... 119

Figure 4.14 Expression of *MACROD2* and neuronal markers in SH-SY5Y cell lines treated with 5nM siRNA against *MACROD2*..... 120

Figure 4.15 Expression of *MACROD2* and neuronal markers in SH-SY5Y cell lines treated with 10nM siRNA against *MACROD2*..... 121

Figure 4.16 Expression of *MACROD2* and neuronal markers in SH-SY5Y cell lines treated with 20nM siRNA against *MACROD2*..... 122

Figure 4.17 Expression of *MACROD2* and neuronal markers in SH-SY5Y cell lines treated with 35nM siRNA against *MACROD2*..... 123

Figure 4.18 Expression of *MACROD2* and neuronal markers in SH-SY5Y cell lines treated with 50nM siRNA against *MACROD2*..... 124

Table 4.10 Protein interactors of *MACROD2* homologs identified via the STRING database. 127

Figure 5.1 Forest plot of associations reported by Curran et al (2011) for strictly defined autism cases only. 132

Figure 5.2 Power to observe association under the additive model for different sample sizes 134

Figure 5.3 Plot highlighting population ancestry. 139

Table 5.1 A breakdown of the individuals who meet criteria for analysis 141

Figure 5.4 An association plot of European strict AGP stage 1 individuals. 148

Figure 5.5 Association analysis of a 500kb region at the *MACROD2* locus. 149

Figure 5.6 An investigation of the effects of phenotype on the effect size and genome-wide significance for rs4141463 151

Figure 5.7 An investigation of the effects of inheritance model on the effect size and genome-wide significance for rs4141463..... 152

Figure 5.8 An association plot for the different inheritance models illustrates their effects on genome-wide significance for rs4141463..... 153

Figure 5.9 Association plot emphasising the D' prime linkage disequilibrium 154

Table 5.2 Interval-Gene-Target Summary. 155

Table 5.3 Interval Summary. 155

Figure 5.10 Association plot for the nominally significant hit within *MACROD2* from the PGC ASD GWAS 157

Figure 5.11 An illustration of the effects of CNV on allele intensity 158

Figure 5.12 Genotyping plot for rs4141463 finds no evidence of CNV deletions driving the signal by disruption of genotype calls..... 158

Table 6.1 A subset of proteins identified by the LC/ESI-MS analysis of the Co-IP pulldown from differentiated SH-SY5Y cell lines..... 166

Chapter 1 Introduction

1.1 Autism

Autism is a complex neurodevelopmental disorder characterized by deficits in the development and use of language, difficulty interpreting and understanding social interactions, and restricted and repetitive behaviours (Newschaffer *et al.* 2007). The wider diagnosis of autism spectrum disorder (ASD) describes a range of disorders, such as Asperger Syndrome, which share a common impairment in social interaction. ASDs can themselves be considered a part of the wider diagnostic category of Pervasive Developmental Disorders (PDDs) (Association 1994; Organization 1992), which also includes Rett Syndrome and Childhood Disintegrative Disorder. An early onset disorder, estimates of autism prevalence vary; recent estimates suggest a prevalence of 60 per 10,000 for autism spectrum disorder and PDDs (Fombonne 2005). There is a distinct sex bias in autism, with a male-to-female ratio of 4.3:1. The popular notion that the prevalence of autism is increasing has some supportive evidence, however the reason for this is typically attributed, in part, to changes in diagnostic criteria and/or increased vigilance amongst clinicians. Nonetheless, we cannot rule out the hypothesis that the prevalence of autism is indeed on the increase.

For research and clinical purposes a diagnosis of autism is typically garnered via observation and interaction with the child using the Autism Diagnostic Observation Schedule (ADOS (Lord *et al.* 2000)) or via parent interview using the Autism Diagnostic Interview-Revised (ADI-R (Lord *et al.* 1994)). There is considerable variability amongst individuals with autism, making the identification of distinct core and co-morbid deficits difficult. Consequently this impacts on the identification of distinct neurobiological phenotypes shared amongst all those with ASD. Nonetheless, it is evident that the neurodevelopment of individuals with autism is altered compared to their typically developing peers – including differences in neuronal patterning, brain growth and cortical connectivity. This is supported by the results of neuroimaging studies which have shown abnormal growth patterns in cortical white matter, frontal and temporal lobes and limbic structures (e.g. amygdala) (Pardo & Eberhart 2007). A more recent study reports physical changes in single axons below prefrontal cortical areas in autism are likely to affect synaptic function, which fits with the hypothesis that the fundamental defect in autism is at the synapse (Zikopoulos & Barbas 2010).

1.2 Genetic Aetiology of Autism

Beginning in the late 1970s with Folstein and Rutter's seminal paper 'Genetic influences and infantile autism' (Folstein & Rutter 1977) several twin studies have been performed in an attempt to estimate the concordance rate of autism. By examining the concordance rate in identical and non-identical (same-sex) pairs we can estimate how much genetic factors can be attributable to the trait under investigation. The early work of Folstein and Rutter was important as it demonstrated a clear heritability of autism and helped society to move away from the idea that autism is caused by "cold" parents who reject their children. Whilst one can appreciate that there is an environmental component to autism, these studies established the importance of genetics to autism aetiology. Folstein and Rutter performed a modestly sized twin study, and observed that 4 of their 11 pairs of identical twins (36%) were concordant for strictly defined autism, whereas 0 of their 10 pairs of fraternal twins were concordant for the same measure. Almost twenty years later, Bailey and colleagues performed a follow-up study in a larger sample, that also included the original sample from the Folstein and Rutter analysis (Bailey *et al.* 1995). Here the concordance rate in the combined sample was 60% for identical twins, and again no fraternal twins were concordant for strictly defined autism. For broader spectrum defined autism, 92% of identical twins were concordant for this definition, whereas 10% of fraternal twins met these criteria. This publication is often cited when informing the high heritability estimates for autism compared to other neuropsychiatric disorders, with a heritability estimate of up to 91-93%.

In a review of more than 30 twin studies of ASD and autistic traits (Ronald & Hoekstra 2011), Ronald and Hoekstra reported that earlier twin studies that relied upon a strictly defined autism showed a median concordance rate of 76% for identical twins, and again 0% for fraternal twins. When the estimation was expanded to include studies that used a broader definition of ASD, a median concordance rate of 88% and 31% was reported for identical twins and fraternal twins respectively. All the evidence from these studies points towards ASD having a strong genetic component. However this was not readily translated into identifiable genetic loci in linkage and association studies. If the heritability of autism was so strong, questions were raised as to why it has been so difficult to discover strong causative variants, and the idea of 'missing heritability' was discussed (Maher 2008). Some of the key discussion points surround whether the 'missing heritability' is due to an original overestimate of heritability (due to bias and confounding)

or limitations of our genetic measurements including power, association models and variants examined. Clearly all of these points have merit. Recent work by Hallmayer and colleagues (Hallmayer *et al.* 2011) utilised more stringent phenotyping via clinical diagnostic tools (such as the Autism Diagnostic Interview or ADI (Le Couteur *et al.* 1989) and the Autism Diagnostic Observation Schedule or ADOS (Lord *et al.* 1989)) in a much larger sample of 202 twin pairs from the California Twin Registry. This large sample enabled them to break down their report not only by status of autism diagnosis within identical or fraternal twins, but also of the gender of the twin pairs. For strictly defined autism in males, the concordance rate was 58% for identical twin pairs and 21% for fraternal twins; for females the concordance rate was 60% and 27% respectively. For a broader diagnosis of ASD in males, the concordance rate was 77% for identical twin pairs and 31% for fraternal twins; for females the concordance rate was 50% and 36% respectively. The authors calculated that a modest majority of the variance in liability can be attributed to the twins shared environmental factors, with a concordance of 55% for autism and 58% for ASD, with genetic heritability calculated at 37% for autism and 38% for ASD. The differences in concordance rates between identical and fraternal twins support a hypothesis that autism is not a single gene disorder. A caveat to these results is the wide confidence intervals reported; however, if these concordance rates more accurately reflect the heritability for autism, it could go some way towards explaining why the search for causative genetic variants in ASD has been so difficult. A second large study from the Twins Early Development Study (TEDS), examined ASD traits in a 12-year-old twin population (Robinson *et al.* 2011) and supports a stronger genetic component than Hallmayer's study. The study employed an extreme trait design; a group of 192 twin-pairs with at least one ASD affected twin was selected from a population of 5944 twin pairs. A strong additive genetic component was observed in males for each of the three core autism domains of social interaction at 0.72, communication impairment at 0.76 and restrictive-repetitive behaviours at 0.76. The authors also highlight that for each domain there is a considerable unique genetic component, at greater than 50%. This impact of both environmental and genetic factors would seem to foster the clinical heterogeneity observed in ASD.

Cytogenetic analyses have documented rare causative chromosomal abnormalities that confer risk of ASD-related syndromes; for example inherited

duplications in the 15q11-15q13 chromosomal region/Angelman syndrome or Fragile X syndrome each account for 1-2% of cases (Abrahams & Geschwind 2008). Although these syndromic forms provide evidence in favour of a genetic cause of ASD, the low prevalence rates of these syndromes mean that they were considered exceptional cases and therefore may not have relevance for more common cases of autism. Due to the complex nature of the autism phenotype, encompassing a wide spectrum of disorders under the label of 'ASD', one can postulate that there is likely to be a multigenic component corresponding to the various distinct phenotypes of ASD. This postulation was supported by evidence from linkage studies which suggested a multi-locus model (>15 genes) (Risch *et al.* 1999).

1.2.1 Linkage Studies

Linkage studies were perhaps the first kind of large scale study used in the search for the genetic factors contributing to autism. These utilise genetic linkage, whereby genes, and therefore any variants contained within them, that are located in close proximity to each other are more likely to be inherited or "linked" together. This non-random linkage of genes means that if a known genetic variant and a disease mutation co-localise (i.e. are linked), it may be possible to determine the approximate location of the causative mutation. The method is imperfect in that by highlighting a genic region and not specific variants within a given gene, multiple genes at a given locus can be implicated. Although there were successes with implementation of this approach in Mendelian disorders, for more complex traits such as ASD the results were not as positive. In a review of 14 genome-wide linkage studies using guidelines as outlined by Lander and Kruglyak (Lander & Kruglyak 1995), Yang and Gill (Yang & Gill 2007) report 94 loci as significant or suggestive of linkage in these genome wide analyses. However, just 8 loci meet the criteria for significant linkage (i.e. LOD $\geq$ 3.3): 1p21.1 (Risch *et al.* 1999), 2q31.1 (Anon 2001a), 3q25–27 (Auranen *et al.* 2002), 5p15.3 (Liu *et al.* 2001), 7q22.1 (International Molecular Genetic Study of Autism Consortium 2001), 7q32.2 (IMGSAC 1998), 8q24.13 (Liu *et al.* 2001) and 13q21.32 (Barrett *et al.* 1999; Anon 2001b). Of these loci, those that had the most replications from independent reports were located in regions such as 7q, 16p, 1q, 3q and 2q.

The chromosomal region 7q, which had the most independent reports from replication studies according to Yang and Gill, has been subject to much investigation with

the aim to discover candidate genes for autism. Fine mapping identified several putative susceptibility genes or loci, such as *CNTNAP2* (Arking *et al.* 2008), *EN2* (Within an intronic loci) (Gharani *et al.* 2004), *MET* (Within the promoter region) (Campbell *et al.* 2006) and *RELN* (Skaar *et al.* 2005). Whilst linkage studies are a reliable tool in identifying variants in rare mendelian disorders, they are limited under models whereby many genes are proposed. Under these models association studies are a more attractive tool due to their use of more readily available singleton cohorts and their ability to discover common variation of low penetrance. Autism, like many other common disorders with a genetic component, was thought to fit the common disease-common variant model.

1.2.2 Common and Rare Variant hypotheses of Autism

The 'common disease-common variant' hypothesis posits that variants confer a susceptibility to complex polygenic diseases via an additive model, such that multiple common variants which alone confer some small risk towards development of a disorder, when inherited together, perhaps interacting with each other and with environmental factors, result in a particular phenotype. This model is an attractive fit for autism when one considers the fact that there is robust evidence to support a familiarity of ASD. For example, first degree relatives of ASD probands have a higher prevalence of lesser forms of behavioural or cognitive features (termed the 'broader phenotype' (Pickles *et al.* 2000)) associated with autism compared to the general population (Losh *et al.* 2009). Moreover, in a recent study as part of the BabySibs cohort, the recurrence rate of ASD in the baby siblings of individuals with ASD is ~19%, compared to the population risk of <1% (Ozonoff *et al.* 2011). However, as with many complex genetic disorders, there is much discussion over whether common or rare variants are the major contributors to the genetic component of autism. This coincides with recent studies that suggest that there is increased evidence to demonstrate a role for rare variants in autism (Levy *et al.* 2011; Sanders *et al.* 2011).

The 'rare variant hypothesis' posits that a significant proportion of the inherited susceptibility to common human disease is due to the effect of low frequency, highly penetrant, dominant variants that confer a moderate to large increase in relative risk. *De novo* and inherited recurrent copy number variants (CNVs) and single nucleotide variants (SNVs) have emerged as contributors to ASDs; the increased resolution of array based approaches and major advances in DNA sequencing technologies has lead to increased

discovery of these rare variants. Interestingly, a number of studies have highlighted rare mutations in biologically plausible candidate genes that function at the synapse, including *SHANK3* (Durand *et al.* 2007), *NRXN1* (Szatmari *et al.* 2007), *SHANK2*, *SYNGAP1* (Pinto *et al.* 2010), *NLGN3* and *NLGN4* (Jamain *et al.* 2003).

These as well as many other CNVs and SNVs were estimated to account for up to 10% of cases of idiopathic autism (Geschwind 2011). However, with the considerable wealth of data from large scale CNV and next-generation sequencing efforts, this number is rapidly increasing. Indeed CNV, SNV and syndromal presentations of ASD implicates a growing number of risk loci in the hundreds (Betancur 2011; State 2010; Abrahams & Geschwind 2008). Thus investigation of the genetic aetiology of autism has proceeded via GWAS, CNV studies and whole and exome sequencing studies.

1.3 Investigating the Genetic Aetiology of Autism

1.3.1 Introduction to Genome-Wide Association Studies

Genome-wide association studies examine whether there is a statistical association of a genotype with a trait or disease. Advances in both our knowledge of genetic variation coupled with improved genotyping technology in the early part of the century facilitated a move from single variant and candidate gene investigation to large scale genome-wide association studies (GWAS) – where many thousands of markers were investigated in a single experiment. Since the first GWAS revealing a genome-wide significant association was published in the journal *Science* in 2005 (Klein *et al.* 2005) there has been a steady increase in the identification of genome-wide significant associations for a variety of traits. The NIH has monitored the success of GWAS over the years as part of its GWAS Catalogue (Hindorff LA, MacArthur J (European Bioinformatics Institute), Morales J (European Bioinformatics Institute), Junkins HA, Hall PN, Klemm AK, and Manolio TA. A Catalog of Published Genome-Wide Association Studies. Available at: www.genome.gov/gwastudies). In their latest summary, updating findings to the second quarter of 2012, the resource had identified that 1350 associated loci had been reported from GWAS to show association at less than $p = 10^{-5}$. An ideogram of these findings across 18 trait categories is shown in Figure 1.1.

The GWAS approach has some key advantages over linkage studies, most importantly the increased power to identify effect in populations (Risch & Merikangas 1996). Their recent success in identifying association has encouraged the genetics

community to embrace this method to identify risk loci. GWAS apply a ‘data driven’ approach, so called because of the fact that these studies are free from assumptions made about the underlying biology of a trait, such as variant, gene or location of putative risk. Thus GWAS can transcend barriers created by our incomplete understanding of the biological process underpinning the disease. There exists an idea amongst some in the field that GWAS are ‘hypothesis-free’; however, the term is a misnomer, a misunderstanding of ‘explicit’ and ‘implicit’ hypotheses. For example, with a microarray expression analysis between cases and controls, the implicit hypothesis is that expression changes in sets of genes (or modules of protein etc) will correlate with differences between the diseased and normal state. With GWAS (or a sequencing study), again we are comparing cases and controls resulting in the same implicit hypothesis.

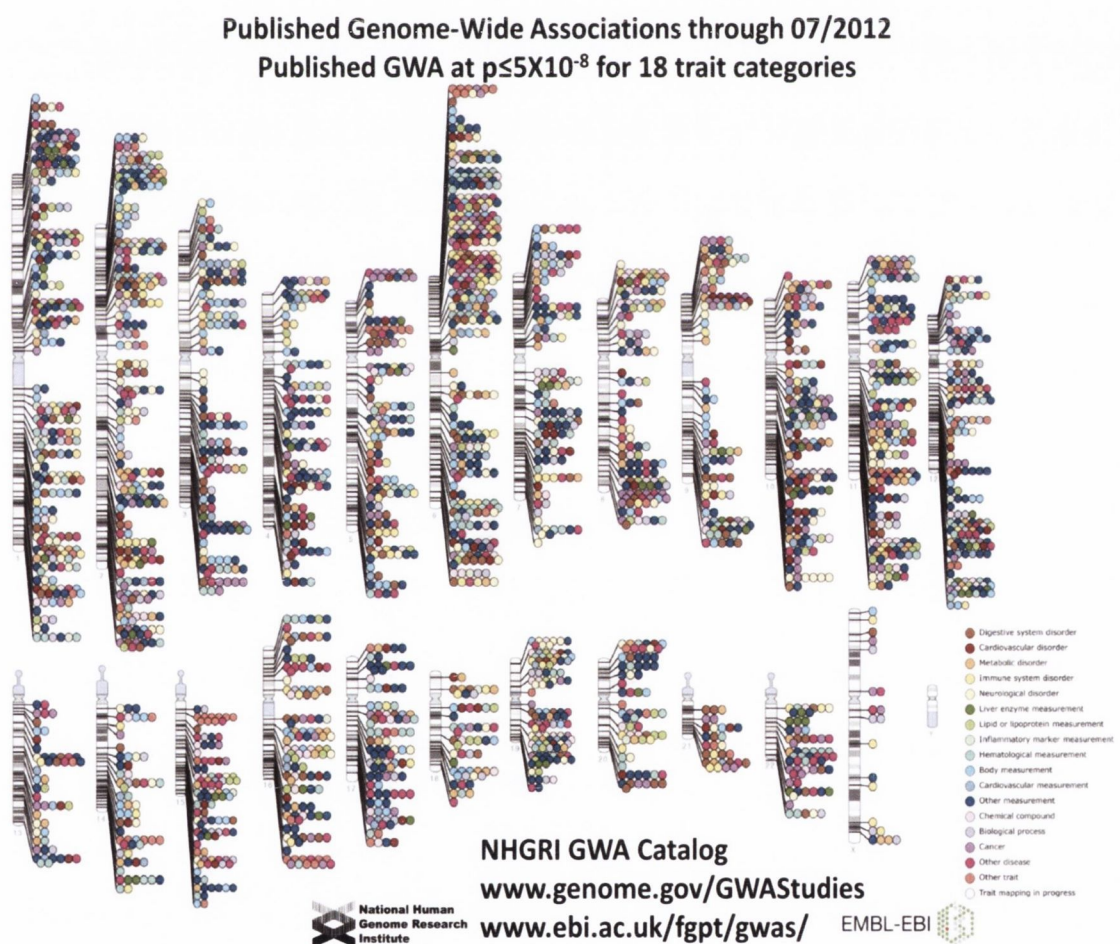


Figure 1.1 Published Genome-Wide Associations, illustrating the wide spread of genome wide significant loci across the human genome for a wide array of disorders. Credit: Darryl Leja and Teri Manolio, NHGRI; Tony Burdett, Dani Welter, and Helen Parkinson, EBI (<http://www.genome.gov/GWASStudies/>)

The study of polygenic disorders, that is, disorders that are hypothesised to involve more than one gene in their causation, has seen many different approaches

developed to test for a genetic association. At first these approaches were restricted to our knowledge of genetic variation, the genotyping technology of the day and the modest sample sizes of the study collections. This reduced the statistical power to observe anything but strong effects and combined with a limited coverage of the variation in the genome led to many true associations going unobserved. As noted above, improved SNP discovery through efforts such as the HapMap Project (www.hapmap.org) and advances in genotyping technology has significantly improved our ability to tag common variation in the genome.

1.3.2 Genome-Wide Association Studies in Autism

In 2010 Anney *et al* (Anney, Klei, Pinto, Regan, Conroy, Tiago, et al. 2010) in a publication for the Autism Genome Project (AGP) performed a family based GWAS study in a sample of 1,385 autism probands from 1,369 families. The most significant GWA hit was observed in a subset of 718 individuals of European ancestry with a diagnosis of autism on both the Autism Diagnostic Interview (ADI; (Lord et al. 1994) and Autism Diagnostic Observation Schedule (ADOS; (Lord et al. 2000)). This association was with the SNP rs4141463, located at 20p12.1 within an intronic region of the *MACROD2* (MACRO domain containing 2) gene (OR = 0.56 (95%CI 0.47 – 0.67; $p = 5 \times 10^{-8}$) (see Figure 1.2 below). This association within *MACROD2* forms the basis for this project.

A follow up study was performed by the AGP, including an additional 1301 families (660 additional probands with a Strict Diagnosis of Autism and of European Ancestry). Despite having greater than 99% power to observe a nominal association of $p < 0.05$ in the Strict European subset and given the parameters observed in Stage 1 (i.e. MAF=0.42, OR=0.65 (Risch Merikangas, 1996)) the authors did not show association at the rs4141463 for in this new collection (OR=0.91 (95%CI 0.78-1.06) $p = 0.206$) (Anney et al. 2012). The observed Stage 2 association was in the same direction, as the Stage 1 data leading to the combined analyses retained a strong association signal ($p = 1.19 \times 10^{-6}$; OR=0.77 (95%CI 0.69-0.86)) (See Figure 1.3). The strongest association for the AGP stage 2 analyses was observed within the *CNTNAP2* gene (rs1718101; $p = 7.8 \times 10^{-9}$; OR=2.13), a gene previously implicated in ASD via linkage and GWAS analyses (Arking *et al.* 2008; Bakkaloglu *et al.* 2008). Importantly, the AGP study did not replicate any of the previous GWA from other association studies in autism such as those reported by:

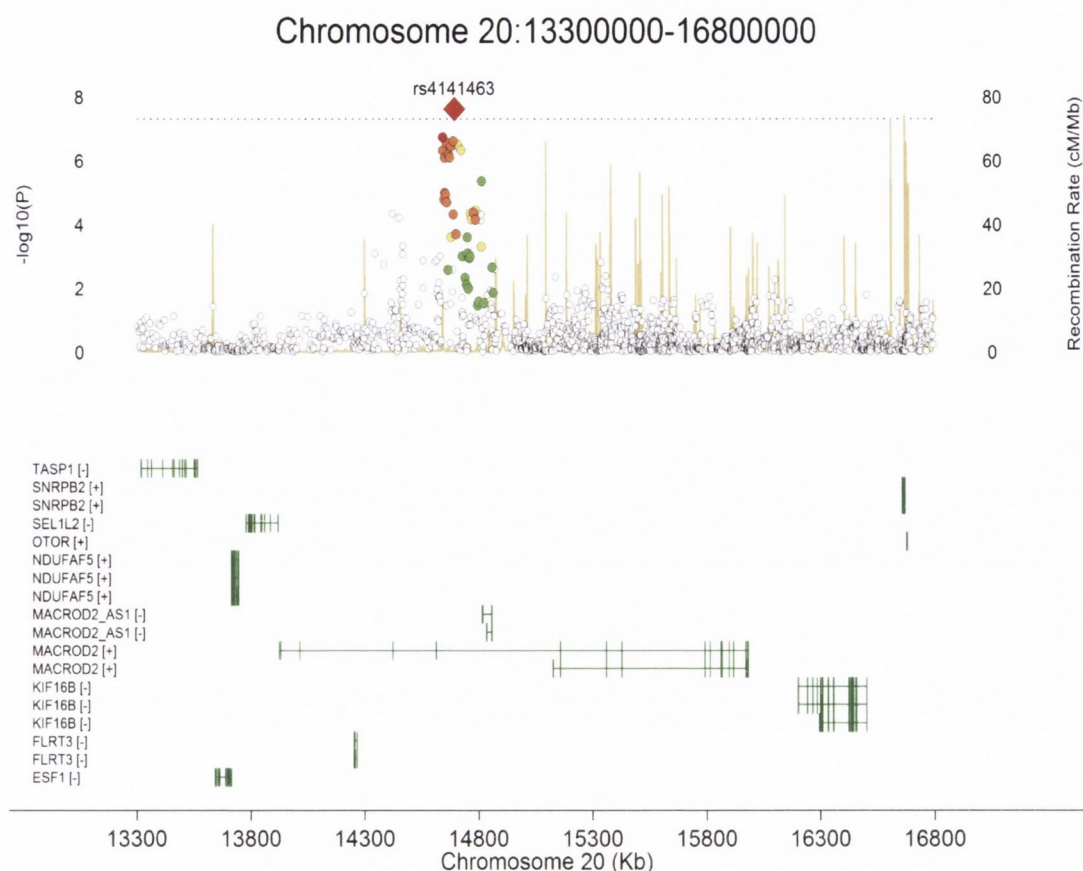
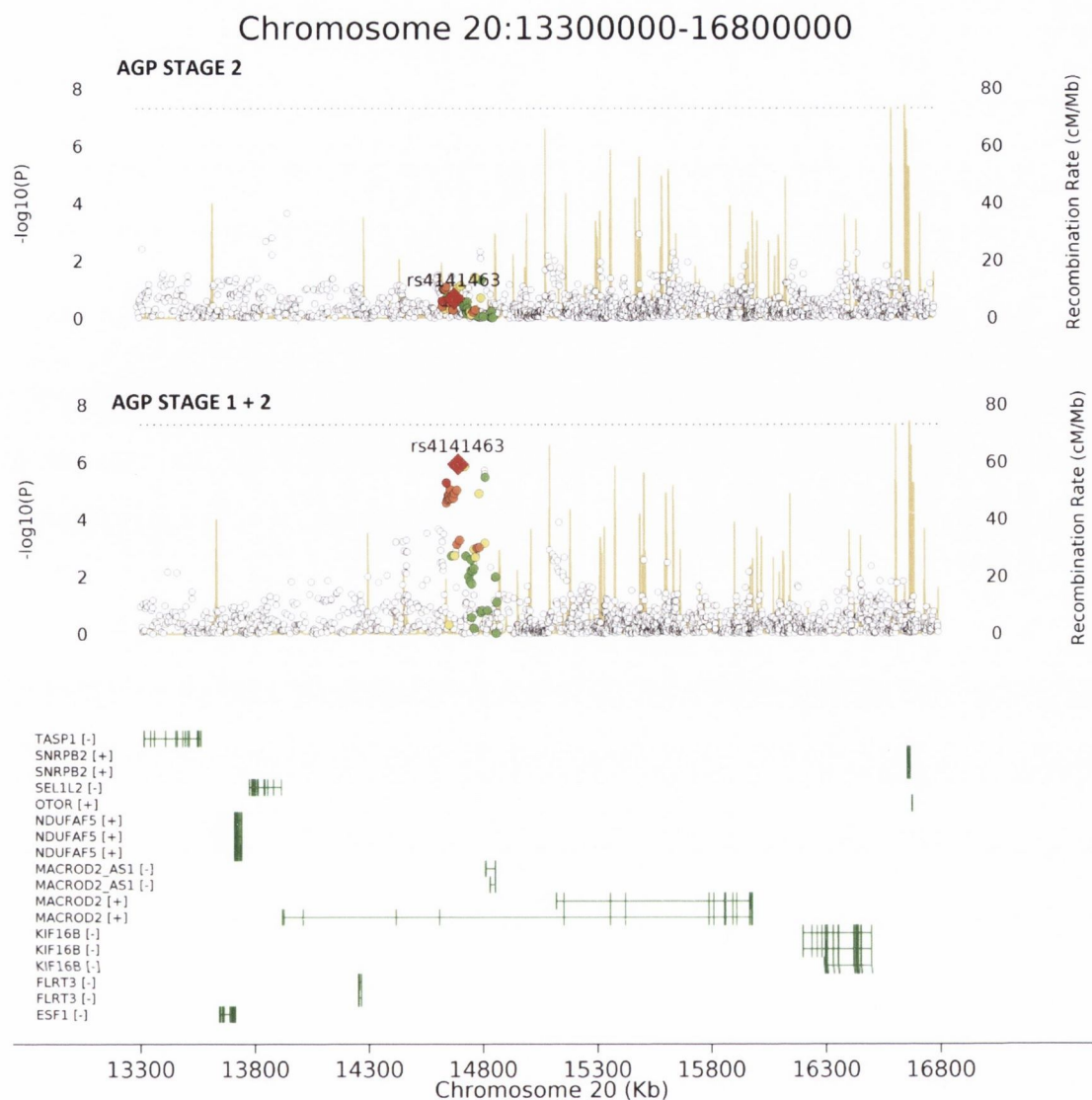


Figure 1.2 Association plot for region containing rs4141463 in the AGP Stage 1 GWAS (adapted from Anney et al., 2010).

- 1) Wang *et al* (Wang *et al.* 2009) reported a GW significant hit on chromosome 5p14.1 (rs4307059; $p=3.4 \times 10^{-8}$) in addition to suggestive association signals on 13p33.3 (*MYO16* locus), 14q21.1 (*FBXO33* and *LRFN5* locus) and Xp22.32 (*PRKXI* and *NLGN4X* locus) with an additional analysis of the Y chromosome returning an association at the *USP9Y* locus.
- 2) Ma *et al* (Ma *et al.* 2009) did not find any GW significant associations in their study of 1537 individuals, although they did find nominally significant associations residing in 5p14.1 in a validation analysis.
- 3) Weiss and colleagues (Weiss *et al.* 2009) did not fare much better, a genome-wide linkage analysis of their discovery set of 1553 affected individuals did not identify any GW significant associations, although supplementation of the discovery set with 90 probands improved the signal for their top hits. 45 SNPs from the top associated loci were genotyped across 1755 trios. The only evidence for replication came from a marker residing on 5p15, in close proximity to *TAS2R1* and

approximately 80kb from *SEMA5A* (Semaphorin-5A, a protein thought to be involved in axonal guidance during neural development due to the activity of homologous proteins in mice. (Adams *et al.* 1996)). Subsequent analysis demonstrated reduced expression of *SEMA5A* in brains from autistic patients compared to controls

Taken in the context of the failure of these studies to replicate their GW significant hits in subsequent analyses, perhaps it is not surprising that the AGP study was also unable to achieve this. However, there can be multiple reasons why a finding may fail to replicate. Most simply, the finding may be a chance occurrence, a false positive. There are multiple measures one can take to limit the chance of type I error (false positive), however these measures may then increase the chance of type II error (false negative). Typical measures taken to limit these errors involve increasing the required level of evidence and stringent quality control metrics such as limiting allele frequency, missingness and deviation from Hardy Weinberg equilibrium and adjusting for population structure. Quality control of data is necessary; however depending on the stringency of the applied thresholds one can dampen a true signal by over correction of the data. Due to the impact of applying quality control to GWAS data it is important that any replication dataset undergoes the same quality control measures. Efforts to refine genome wide associations are discussed in greater detail in Chapter 5.



Effect Size for rs4141463 across AGP study and ancestral and phenotype subsets

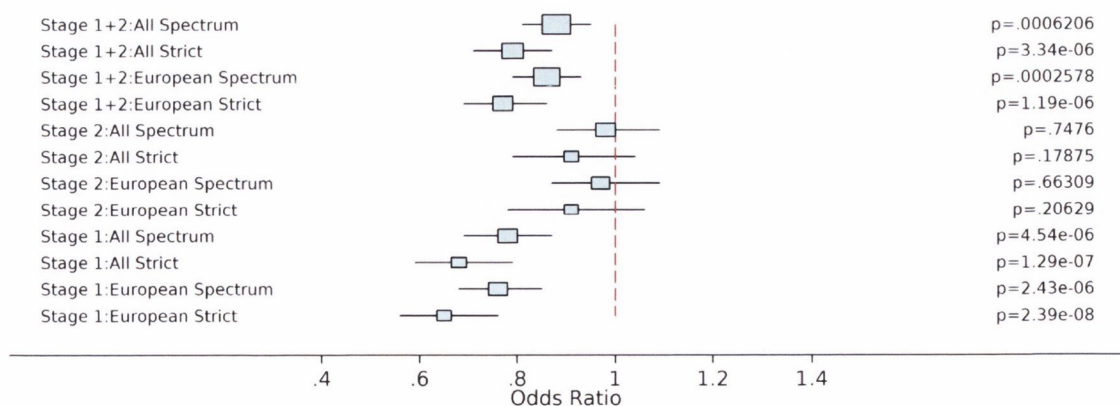


Figure 1.3 Association plot for region containing rs4141463 in the AGP Stage 2 and combined Stage 1 + 2 GWAS (adapted from Anney et al., 2012). Additional panel highlights a forest plot of association signals across stages, phenotypes and ancestries for rs4141463 in the AGP sample.

1.3.3 CNV studies of Autism

Copy Number Variation (CNV) is another way in which the human genome varies and the term CNV is used to describe structural variation, such as duplications and deletions, which affect the number of functional copies of a gene. It was anticipated that CNV studies in neuropsychiatric genetics and the identified structural variants would go some way towards explaining some of the missing heritability, considering the fact that CNVs have been reported to make a considerable contribution to the genetic causes of disease susceptibility (Estivill & Armengol 2007). CNV studies, particularly *de novo* CNVs (Sebat *et al.* 2007), have implicated several new susceptibility loci and lent additional support to previously associated loci. Glessner and colleagues identified associations with CNVs at *NRXN1* and *CNTN4*, but also identified novel loci such as *NLGN1*, *ASTN2* and identified over-representation of genes involved in ubiquitin pathways and cell adhesion (Glessner *et al.* 2009). Cell adhesion, GTPase and cytoskeletal proteins were highlighted by the AGP in their enrichment analysis of global rare CNV in autism (Pinto *et al.* 2010). The AGP grouped their findings by function, and reported any enrichment for particular biological processes. Grouping the genes together in this way can perhaps be more informative than reporting genes in isolation. More recent studies have taken this approach to reporting the results from their CNV studies, and reporting enrichment for particular gene-ontology (GO) terms. Gai and colleagues reported enrichment for biological processes in support of those reported by the AGP, but also found enrichment for genes related to post-translational protein modification, neurotransmitter transport and cell-cell signalling (Gai *et al.* 2012). However, there are design (e.g. sensitivity and resolution) and analytic issues with CNV studies (Merikangas *et al.* 2009). The pleiotropic nature of CNVs and their downstream effects requires in-depth characterisation of CNVs in order to understand the functional impact of these structural variants (O'Donovan *et al.* 2008). Annotation to establish whether a CNV not only impacts a locus but whether it disrupts an exon or is limited to intronic/intergenic material can offer important information as to whether the variant is anticipated to influence the gene product. I have conducted my own in depth analysis of CNV at the *MACROD2* locus, reported in Chapter 3.

1.3.4 Sequencing studies of Autism

Early sequencing studies focused on the interrogation of candidate genes and exons. These approaches were laborious relying upon the manual sequencing of

individually amplified PCR fragments. Indeed a whole PhD thesis could be based upon the interrogation of a single gene or a few kilobases of DNA in a small cohort of individuals. As with microarrays for GWAS, there have been considerable advances in the area of DNA sequencing.

Until recently, the expense and time to perform large scale sequencing studies of many genes was prohibitive, and thus initial sample sizes were relatively small. The idea of interrogating the entire coding region of the genome or even the entire genome in a large cohort of individuals was fanciful. Up until about the year 2007, sequencing costs were following a pattern not unlike that of Moore's law, which describes the phenomenon whereby computational power doubles every two years – in the case of sequencing we were observing a reduction in costs per megabase by a half every two years. However, since 2008 the cost of sequencing has been falling drastically as technology improves. This has given feasibility to large scale studies that sequence bundles of candidate genes or whole exomes.

The first autism exome study was carried out using a sample of 20 trios of probands with sporadic ASD (O'Roak *et al.* 2011). In this study the authors hypothesised that the study of individuals with no prior familial history of ASD would reveal more *de novo* mutations of large effect. This modestly sized study observed *de novo* mutations in 4 of the 20 trios (who also met a strict criteria for the diagnosis of autism), in the genes *FOXP1*, *GRIN2B*, *SCN1A* and *LAMC3* and one inherited potential risk variant in *CNTNAP2*. The reasoning for focusing on *de novo* events is that inherited mutations are hypothesised to effect reproductive fitness and consequently are selected against; we assume that sporadic mutations have a greater functional impact because they are not under this same level of selection and thus deserve our focus. This study by O'Roak and colleagues demonstrates that by focusing on *de novo* variants we are more likely to enrich our findings for variants that have a functional impact.

De novo mutations were again the focus of four subsequent exome sequencing studies published simultaneously in Nature (Sanders *et al.* 2012; Neale *et al.* 2012; O'Roak, Vives, Fu, *et al.* 2012) and Neuron (Iossifov *et al.* 2012)) earlier last year.

Sanders and colleagues (Sanders *et al.* 2012) utilised a sample of 238 families from the Simons Simplex Collection. Based upon estimations taking into account the *de novo* mutation rate, the number of genes under investigation and the size of the sample, it was

calculated that three independent mutations in a given gene would be considered a highly unlikely observation and two independent mutations being suggestive. In the initial analysis no genes were found to carry three independent *de novo* mutations, although two independent nonsense *de novo* mutations were observed for *SCN2A*.

O’Roak and colleagues (O’Roak *et al.* 2012) sequenced the exomes of 189 trios and 50 unaffected siblings and identified *de novo* mutations that impacted genes within the β -catenin and chromatin remodelling protein network, and also identified recurrent disruptive mutations in *NTNG1* and *CHD8*. They also carried out mutation screening of 6 candidate genes across 1,703 individuals with ASD and identified additional *de novo* loss of function variants in *GRIN2B*, *LAMC3* and *SCN1A*.

Neale and colleagues (Neale *et al.* 2012) performed exome sequencing of 175 ASD trios, and in combination with the data from the other two publications reported that *SCN2A*, *KATNAL2* and *CHD8* are strong candidates as autism susceptibility genes due to the observed rate of *de novo* nonsense, splice and frameshift mutations within these genes.

Iossifov and colleagues (Iossifov *et al.* 2012) performed exome sequencing of 343 families, with one affected child and at least one unaffected sibling and identified *de novo* indels (insertion deletions) and point substitutions. The burden of *de novo* missense mutations in probands versus unaffected siblings was not significantly different, but gene disrupting mutations were twice as frequent in probands. The authors report that many of the disrupted genes were associated with the fragile X protein, FMRP.

Taking these recent reports together, it appears that *de novo* and rare inherited mutations contribute together to the overall genetic component of ASD in any one affected individual, so there can be no doubt that autism is a polygenic disorder. However the functional impact of many of these genes is poorly characterised, although as we have seen efforts have been made to group associated genes together by GO-term to look for enrichments of particular biological processes. Despite this difficulty in identifying functional variants, there are several loci that have been identified via linkage, GWAS, CNV and exome studies that have withstood replication studies. Chapter 3 describes the results from the characterisation of the CNV pattern and the exome sequencing of *MACROD2*.

1.3.5 Epigenetic studies of Autism

The complex nature of the autism phenotype and the difficulty in elucidating the genetic architecture of ASD has led to an increasing consideration for the role of epigenetic modifications and autism. Epigenetic modifications can be at the level of DNA, such as methylation of cytosine nucleotides, which can inhibit the binding of transcription factors to that region of DNA, or at the protein level, via acetylation or ubiquitination of histone proteins which can lead to changes in chromatin structure. For example, the previous section discussed the identification of *CHD8* as a candidate gene for autism. Chromodomain/Helicase/DNA-binding protein 8 (*CHD8*) is a member of the CHD family of proteins, that have been linked to roles in transcriptional regulation, development and differentiation by remodelling of chromatin structure (Marfella & Imbalzano 2007). This remodelling of chromatin via interaction with other chromatin 'readers' and 'writers' can lead to downstream changes in gene expression. The *Rattus norvegicus* ortholog of CHD8 has been demonstrated to bind beta-catenin and negatively regulate the Wnt signalling pathway (Sakamoto et al. 2000), implicating it in processes vital to early vertebrate development and morphogenesis that have already been associated with the development of ASD. *CHD8* has been reported to be highly expressed during early embryogenesis, with *Chd8*^{-/-} mice dying during this stage of development. A recent review has highlighted the association of *CHD8* with autism from exome sequencing studies, and highlights the need for biochemical characterisation of *CHD8*-containing complexes as associated proteins may emerge as contributors to autism (Ronan et al. 2013). Study of differentially expressed genes using stem cells from autism cases versus controls found upregulation of CHD8 in cases, and highlights the utility and feasibility of using stem cells in the study of epigenetics and gene expression in neuropsychiatric disorders such as autism (Griesi-Oliveira et al. 2013).

1.4 Translation of Association to Function

As discussed previously, GWAS have been used to identify genetic markers which have strong statistical association with a disease. However, statistical likelihood does not always directly translate to markers that have significant biological importance. Therefore it is imperative to establish a functional consequence of any reported associations and to determine a causal link between marker and disease. Similarly, CNV discovery studies have been used to identify potentially significant rare variants in diseases and disorders that are believed to have a multi-gene risk and require functional follow-up. As I have

discussed, there is much debate in the field regarding whether rare or common variants play the more significant role in such diseases; functional validation of identified variants will aid in settling this debate. In a recent GWA of autism the Autism Genome Project identified a strong association within the gene *MACROD2* (Anney, Klei, Pinto, Regan, Conroy, Tiago, et al. 2010), and CNVs in this gene have also been associated with autism (Pinto *et al.* 2010). At the outset of this report, very little was known about the function of *MACROD2*, although there were suggestions in the wider literature associating variation within *MACROD2* with neuropsychiatric disorders (Xu *et al.* 2009) and neurological phenotypes (Baranzini *et al.* 2009).

1.5 MACRO-domain containing 2 (Gene: *MACROD2*, Protein: MacroD2)

MACROD2 is one of the largest genes in the human genome, at over 2 Mb, although the mRNA length is only 4863 bp. Located on chromosome 20 (20p12.1), the full transcript of the gene includes 17 exons. Two isoforms of the gene have been documented. To date there are 2 reported nested transcripts within *MACROD2* (Figure 1.1), *MACROD2-AS1* (*MACROD2* Anti-Sense 1) a non-coding RNA encoded on the reverse strand within intron 5, and *FLRT3* (Fibronectin leucine rich transmembrane protein 3), a nested gene within intron 3 of *MACROD2*, related to cell adhesion and/or receptor signalling. A number of neighbouring genes also surround this locus, including: *SEL1L2* (Sel-1 suppressor of lin-12-like 2), *KIF16B* (Kinesin family member 16), *NDUFAF5* (NADH dehydrogenase (ubiquinone) complex I, assembly factor 5) and multiple pseudogenes. An in depth investigation of the structure and variation at the *MACROD2* locus is described in Chapter 3.

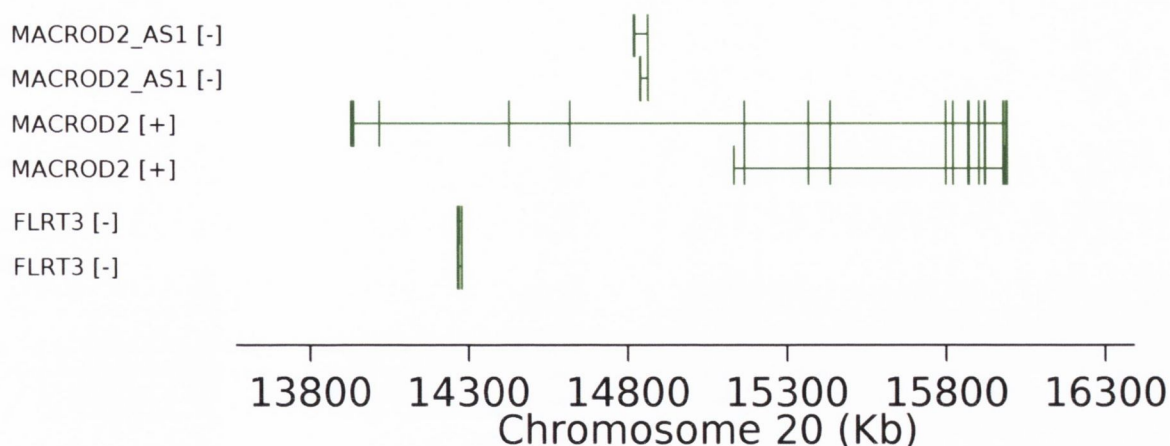


Figure 1.4 The *MACROD2* locus contains 2 major intergenic transcripts on the anti-sense strand (Indicated by [-]), *FLRT3* and *MACROD2-AS1*. It is unknown if these transcripts have any interaction with *MACROD2*.

The role of *MACROD2* in humans at the outset of this report was largely unknown although structural variations in this gene and specifically the region harbouring the AGP ASD association signal have been implicated in schizophrenia (Xu, Woodroffe *et al.* 2009; Kirov, Pocklington *et al.* 2012) and epilepsy (Lesca, Rudolf *et al.* 2012). However, this same region has been identified as a hotspot for deletions in the genome (Bradley, Raelson *et al.* 2010). SNPs within *MACROD2* (under the positional name *C20orf133*) were associated with risk of MRI defined brain infarct (St  phanie Debette *et al.* 2010) and reduced parenchymal volume (Baranzini *et al.* 2009). This association with brain volume has been replicated in other populations (Kohannim *et al.* 2012). Variation at *MACROD2* has also been associated with severe defects in brain patterning (Bendavid *et al.* 2009; Mercier *et al.* 2011) and epileptic encephalopathies (Lesca *et al.* 2012). Despite all these compelling observations, until recently there remained very little information regarding the function of *MACROD2* in humans or in model organisms

The MACROD-like proteins are highly conserved across evolutionary time, which may indicate an essential role. The MACRO-domain is an ADP-ribose binding module (Karras *et al.* 2005). Consequently, as a MACRO-domain containing protein, *MACROD2* has been implicated in the ADP-ribosylation of proteins, an important post-translational modification that occurs in a variety of biological processes such as DNA repair, transcription, chromatin biology and long-term memory formation (Cohen-Armon *et al.* 2004). As we have discussed above, the results of linkage, GWAS, CNV and sequencing studies have implicated post-translational regulation and chromatin biology in autism

aetiology, but *MACROD2* has not been shown to interact with cytoskeletal and cell adhesion molecules that have been associated with autism from the multiple studies discussed above.

1.5.1 *MACROD2* Anti-Sense 1 – A non-coding RNA

The association signal observed at rs4141463, albeit tagged to the *MACROD2* gene, resides in an intronic region near an intragenic non-protein-coding-RNA, *NCRNA00186* (*MACROD2-AS1*). Two *MACROD2-AS1* transcripts have been reported of 673bp and 1230bp in length, located on the reverse strand between exon 5 and 6 of *MACROD2*. Non-coding RNAs (ncRNA) are biologically functional RNA molecules that are not translated into a protein product. ncRNAs typically target transcriptional activators or repressors, and thus exert an upward or downward regulation effect on a transcript near to their locus. Anti-sense RNAs also typically interact with neighbouring genes (Katayama *et al.* 2005), resulting in transcriptional or post-transcriptional effects and have been linked to brain development and plasticity (Qureshi and Mehler 2012). However, unlike *MSNP1AS* described for the 5p14.1 association observed by Wang and colleagues (K. Wang *et al.* 2009) a function for this non-coding RNA has not yet been reported.

1.6 Aims of this project

As we have discussed, *MACROD2* has been associated with autism via the AGP GWAS study and CNV studies of neuropsychiatric disorders, but attempts at replication of this original association has been poor. Variation at this locus has been associated with brain volume, brain patterning and epilepsy, but an in depth study of this locus has not been performed. A function for the *MACROD2* protein as an AADPR deacetylase and mono-ADP-ribosylhydrolase implicates it in a diverse array of biological processes, but the interaction of this enzyme with other proteins has not been demonstrated, including how this protein could have an impact upon neuronal biology or biological processes such as cell adhesion, cell signalling or chromatin modulation, that have been implicated in the aetiology of autism by other studies.

The work reported here represents an in depth analysis of a poorly characterised gene that has been associated with autism, and thus may have relevance to human health, across three chapters.

Chapter 3 describes this effort to understand the basic biology of this gene; including investigation of the **structure** of the gene, characterise the **expression** profile,

and finally to explore the role of **variation** on the expression of this gene, in the hopes that it will inform further analysis.

Chapter 4 represents a functional study of a protein that has not been characterised in humans and may be a protein important to human health. In this project I have sought to answer some of the fundamental questions about the **localisation**, **interaction** and **function** of the *MACROD2* protein.

Finally in Chapter 5, the key aim is to **refine** the *MACROD2* association signal, using the AGP Stage 1 and Stage 2 data for the *MACROD2* region and **fine-mapping** of the locus by **imputation**, alongside a more detailed interrogation of genetic model and clinical subgroups to better describe the original *MACROD2* association. Moreover, I will investigate whether genes which interact with *MACROD2* show association **enrichment** across the GWAS sample – thereby assessing whether common biological pathways which include *MACROD2* interactors are important in ASD.

Chapter 2 Materials and Methods

This chapter describes the standard materials and methods that were utilised across the different aspects of the project. Methods specific to chapters are described in their respective chapter.

2.1 Molecular Biology Techniques

2.1.1 Quantification of Nucleic Acids and Protein

Quantification of nucleic acids and protein was performed using either a NanoDrop 2000 (Thermo Scientific) apparatus, or a Qubit Fluorometer (Invitrogen). Before any quantification was performed, the samples were thawed completely at room temperature and gently mixed on a desktop vortex.

The NanoDrop 2000 is a micro-volume spectrophotometer that can read the concentration of nucleic acids and proteins in a very small sample volume i.e. 1-2 μ l. Thus 2 μ l of each sample to be quantified was loaded in triplicate onto the micro-volume pedestal of the machine. The machine reads the absorption of eight samples at a time at an absorbance of 260 nm (for DNA and RNA) and 280 nm (for proteins), which gives a measurement for the concentration of nucleic acids in the sample and also the purity of the samples by calculating the 260/280 ratio. The average of the three readings was then taken as the best estimate for the concentration of the sample.

The Qubit 2.0 fluorometer uses fluorescent dyes that selectively bind to nucleic acids and proteins in quantification. The Qubit fluorometer was used with proprietary 'Quant-iT' reagent and 'Quant-iT' buffer at a ratio of 1:199 to make up 'Quant-iT' working solution. Between 1 and 10 μ l of sample was added to this working solution to a total volume of 200 μ l per sample, the samples were vortexed and incubated at room temperature for at least 2 minutes (and at least 15 minutes for proteins) before reading on the fluorometer.

2.1.2 Polymerase Chain reaction

The polymerase chain reaction (PCR) is an enzymatically catalysed reaction whereby short sequences of DNA, typically less than 1000 base pairs long, are amplified to many tens of millions of copies from small amounts of starting material. PCR requires DNA nucleotides free in solution and short DNA sequences to prime the reaction, simply called primers, which are complimentary in sequence to the target region of DNA. The

primers are the target site for the second and most important requirement for the PCR, *Taq* polymerase. *Taq* polymerase is a thermostable DNA polymerase originally isolated from the thermophilic bacterium *Thermus aquaticus* and thus proteins isolated from this organism are stable at temperatures that would typically denature proteins from non-thermophilic organisms. This simple fact is exploited during PCR, whereby the reaction volume is first heated to ~95°C to denature the DNA and then cooled to ~60°C, at which point the primers anneal and then finally heated again to ~72°C to enable the *Taq* polymerase to extend and copy, using the DNA nucleotides free in solution, the DNA fragment. This heating and cooling is repeated in cycles multiple times, a 'chain reaction', leading to amplification of the starting material. PCR is the staple technique of any molecular biologist and its principles are used in techniques like RACE and DNA sequencing, described elsewhere.

The standard PCR reaction for this project used 'Hot Star *Taq* Polymerase' (Qiagen) and the reaction buffer supplied by the manufacturer. All primers were purchased from IDT Technologies and diluted to a working stock concentration of 20mM in double distilled water.

2.1.2.1 Primer Design

DNA sequence for a target region was acquired from UCSC genome browser (<http://genome.ucsc.edu/cgi-bin/hgGateway>). The reference sequence returned by UCSC was then used as input into the Primer3 primer design tool (<http://frodo.wi.mit.edu/>), which returns oligonucleotides from the input sequence under the parameters, such as T_m (primer melting temperature) or GC nucleotide content, that have been specified by the user. The ideal T_m is typically between 52 and 58 °C with a maximum of 65 °C. GC content indicates the number of guanine and thymine nucleotides expressed as a percentage of the total number of bases in the primer, with the optimum range being 40-60%. Forward and reverse primers are then chosen in pairs that most closely meet the specified PCR conditions.

2.1.2.2 PCR optimisation

Whilst PCR is a very robust and simple technique, some DNA sequences are more difficult to amplify than others. This can be due to the use of primers that were designed outside of the ideal GC nucleotide content and T_m (Melting temperature) parameters, length of the DNA fragment to be amplified or otherwise problematic DNA sequences

(e.g. sequences including repeats). Modifications to the PCR protocol can ameliorate the reaction and provide a better PCR product. Throughout this project any problematic PCR was first repeated at alternate annealing temperatures, varying around the predicted T_m of the primer in question, and then the PCR was performed using a series of 12 proprietary optimisation buffers (Simply labelled 'A-L', Cambridge Bioscience). For example, a primer with a predicted T_m of 60°C would be tested at 56, 58, 60, 62 and 64°C across the full panel of 12 optimisation buffers. The combination that gave the best band was then recorded and used for any following reactions using that primer.

Whilst the composition of the optimisation buffers is not reported by the manufacturer, they contain various concentrations of magnesium chloride, formamide, glycerol or other stabilising additives such as bovine serum albumin, which are the traditionally utilised additives used to modify problematic PCR reactions.

2.1.3 Restriction Digest of Vector Plasmid

Restriction digestion is a technique used to prepare DNA strands for ligation and utilises enzymes which recognise and 'cut' specific DNA motifs called 'restriction sites'. Each restriction enzyme recognises a specific restriction site, which are usually between four and eight base pairs long. These sites are also palindromic, which in the context of DNA means that the motif reads the same on the forward as it does on the reverse strand. When the restriction enzymes cleave the DNA, the cleavage site can be either 'blunt' or 'sticky'. A 'sticky end' has a short nucleotide overhang; ligating DNA strands with complimentary 'sticky ends' that bind readily is easier to achieve than 'blunt ends', which are straight cut. Ligating strands with 'sticky ends' can also lend more specificity to a ligation reaction as it is more likely that a complementary strand will be ligated at the target site (Figure 2.1). By using two different restriction enzymes, with different restriction sites, at either end of an insert, and the same two restriction enzymes on the vector into which the insert is to be ligated – the ligation reaction can be tailored towards specifically ligating the two strands together. Thus the restriction enzymes NCO1 and SAC1 were used to cut the PCR product insert from the T-Easy Vector Plasmid (Promega) and to cut the pGL3 Luciferase Reporter Vector (Promega) to prepare it for ligation. NCO1 and SAC1 (New England Biolabs, NEB) were chosen for the restriction digest reactions as sites for these enzymes were shared between both plasmids within their respective multiple cloning regions. The restriction digest reaction was performed under the

conditions described in Figure 2.1. Following digestion each reaction was gel separated on a TAE buffer/1% agarose gel.

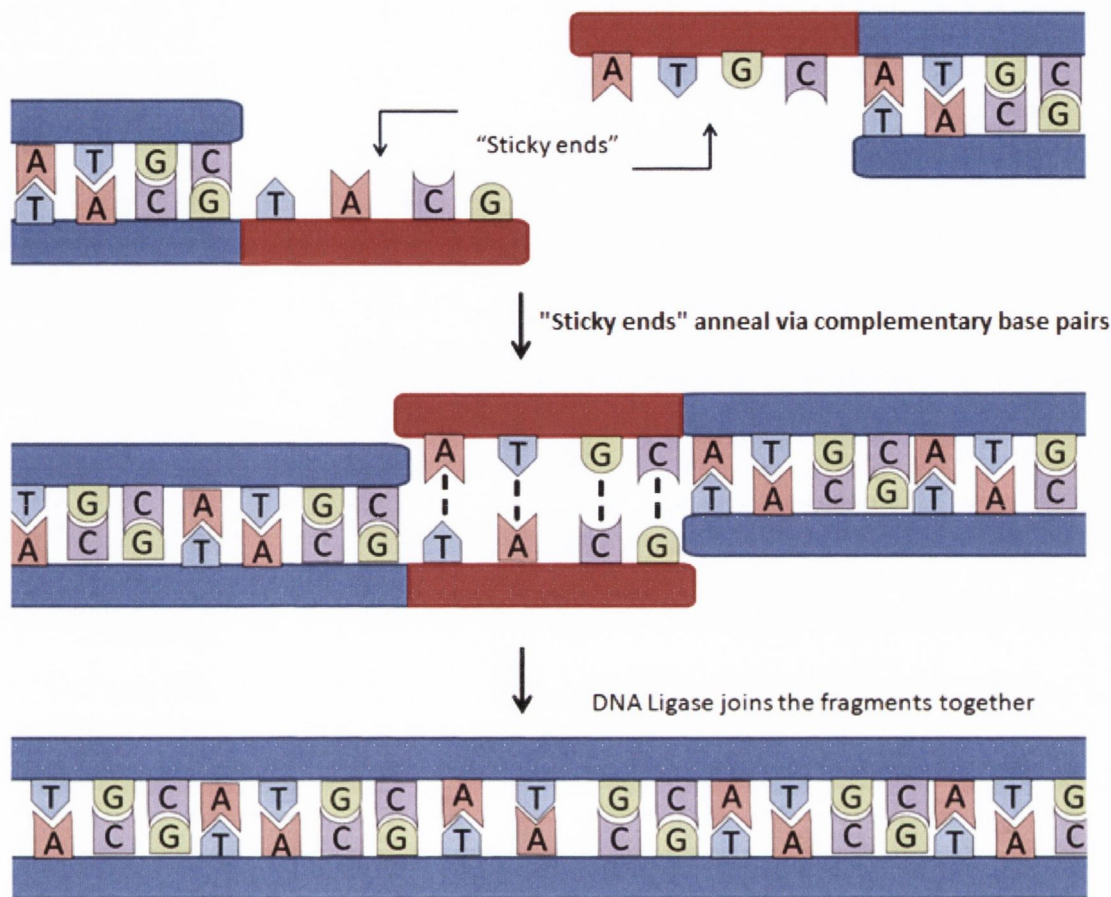


Figure 2.1 DNA ligation. DNA is cut using restriction enzymes, which creates nucleotide overhangs known as 'sticky ends'. DNA ligation occurs via annealing of these complementary 'sticky ends'.

Table 1.1 Restriction Digest - Reagents and protocol used. The restriction digest reaction was optimised using 2X digestion buffer and the addition of bovine serum albumin.

Restriction Digest Reaction	Reagents (µl)		Program
	Digestion Buffer 1 (NEB)	2	
	Restriction Enzymes NCO1 and SAC1	0.5 + 0.5	
	Bovine Serum Albumin	0.5	
	Vector Plasmid	10	
	H ₂ O	1.5	
	Total	15	
			37°C for 4 hours

2.1.4 Agarose Gel Extraction by Wizard SV Gel and PCR Clean-Up System (Promega)

Following electrophoresis and identification of the appropriate restriction fragment, the band was excised from the gel and placed in a 1.5 ml Eppendorf tube. 10 µl Membrane Binding Solution per 10mg of gel slice was added to the tube and incubated at 50-60°C until the gel slice had dissolved. The dissolved gel mixture was transferred to a Minicolumn assembly and incubated at room temperature for 1 minute. The column was then centrifuged at 16,000 x g for 1 minute, and the flow-through discarded. 700 µl Membrane Wash Solution was added to the column and centrifuged at 16,000 x g for 5 minutes. This wash step was repeated with 500 µl Membrane Wash Solution and centrifuged at 16,000 x g for 5 minutes. The Collection Tube was emptied and the column was centrifuged for 1 minute with the column lid open to allow evaporation of any residual ethanol. The Minicolumn was transferred to a clean 1.5 ml Eppendorf, 50 µl of Nulcease Free Water was added to the Minicolumn and then centrifuged at 16,000 rpm for 1 minute. The column was discarded and the eluate, containing purified PCR product, was stored at 4°C or -20°C.

2.1.5 DNA Sequencing Protocol on ABI 3130xl Genetic Analyser

Dye-terminator sequencing reactions were performed using BigDye Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems) using a PCR program specific to the primer set being used. Sequencing reaction clean-up was performed using the BigDye XTerminator purification kit (Applied Biosystems) whereby 45 µl SAM™ solution and 10 µl BigDye XTerminator solution was added to each sample and placed on a shaker for 30 minutes. The samples were then centrifuged at 1000 x g for 2 minutes. Sequence analysis was performed on the ABI 3130xl Genetic Analyser (Applied Biosystems) – a machine which utilises capillary electrophoresis from 96 well plates to sequence DNA up to ~700 base pairs accurately. Sequencing for any insert was conducted in duplicate – i.e. for the forward and reverse strands. Aligning the forward and reverse sequence (using the freely available program BioEdit) against each other increases confidence in the sequence reads.

2.1.6 Ligation Reaction

Ligation is an enzymatically catalysed reaction linking two DNA strands together by forming two covalent phosphodiester bonds between the 3' hydroxyl end of one strand of DNA and the 5' phosphate end of another strand – for example the linking of purified PCR product into a vector plasmid that has been 'cut' by restriction digest.

Ligation of purified PCR product and purified ‘cut’ plasmid was calculated using the formula below, whereby a molar ratio of insert and vector is calculated for addition to the ligation reaction:

$$Insert\ mass\ =\ 3\ x\ \frac{Insert\ Length}{Vector\ Length}\ x\ Vector\ Mass$$

This formula takes into account the mass (i.e. the length) of the insert and the vector, using the required vector mass and a predetermined ratio to calculate the mass of the vector to be used. The T-Easy Vector (Promega) was optimised to a ratio of 3:1 insert:vector and the ligation reaction was performed under the conditions described in Table 2.2.

Table 2.2 DNA Ligation - Reference guidelines for ligation protocols used for given reagents/constructs. Ligation of insert into the pGL3 vector was optimised at a ratio of 10:1 insert:vector.

	Reagents	T-Easy / PGL3	Incubation Program
Ligation Reaction	T-Easy Vector	1 / 2 µl	16°C for 4 hours
	T4 Ligase	0.5 / 1 µl	
	PCR product	From formula 3:1 / 10:1	
	T4 DNA Ligase buffer (10X)	1 / 2 µl	
	H ₂ O	Make up to 10 / 20 µl	

2.2 Bacterial Cell Culture

Competent *E. coli* (HB101, JM109) were purchased from Promega Ltd., UK. Cultures were selected on LB-agar plates supplemented with 100 ng antibiotic per litre of media (Carbenicillin solution, Bioline) and selected colonies were grown in SOC Media at 37°C with shaking. Lysogeny broth (LB) Agar and Super Optimal broth with Catabolite repression (SOC) Media were prepared according to the protocols described in Sambrook *et al.* (Sambrook, Fritsch & Maniatis, 2001).

2.2.1 Standard Transformation Protocol

Bacterial transformation is a technique used to introduce foreign DNA (typically a bacterial plasmid) into bacterial cells. Heat shock is the most common method used to

achieve this, whereby competent cells, i.e. cells that are naturally able to take up DNA from their environment, are heated briefly and then cooled rapidly, disrupting the cell membrane and encouraging uptake of DNA. As such, all materials were kept on ice before use. JM109 competent cells (Promega) were used in transformation protocols for all experiments in this report.

1. The cells were removed from -70 °C and placed on ice for 5 minutes to thaw.
2. Once thawed, 50 µl of cells was used per transformation reaction.
3. 10-20ng of plasmid DNA was added per 50 µl of competent cells and immediately incubated on ice for ten minutes.
4. The cells were heat shocked for 45-50 seconds in a water bath at exactly 42 °C.
5. The samples were then immediately placed on ice for 2 minutes.
6. 900 µl of warm (37 °C) SOC media was added to each tube and incubated at 37 °C for 60 minutes.
7. Following incubation, the samples were centrifuged (13,000rpm, 2 minutes) and the supernatant was discarded by pouring.
8. The pellet was resuspended in the remaining supernatant (~50 µl) and inoculated onto agar plates, spread evenly across the plate using a plate spreader (Sarstedt).

2.2.2 Plasmid purification

Plasmid purification was performed using the PureYield Plasmid Midiprep system (Promega). Before plasmid preparation, a single clone colony was picked from a freshly streaked LB agar plate and inoculated into 5ml of selective SOC medium (i.e. with Carbenicillin), and incubated on an orbital shaker for approximately 8 h at 37°C with vigorous shaking (275 rpm).

1. This starting culture was diluted 1/500 to 1/1000 into selective SOC medium and grown at 37°C with vigorous shaking (275 rpm) for 12-16 h.
2. The bacterial cells were harvested by centrifugation at 5000 x g for 10 minutes at room temperature, ensuring to remove all traces of supernatant.
3. The bacterial pellet was resuspended in 12 ml of cell resuspension buffer, mixed by vortexing or pipetting and incubated at room temperature for 3 minutes.
4. 12 ml of cell lysis buffer was then added, mixed by gentle inversion of the sealed tube, and incubated at room temperature for 3 min.
5. 12 ml of chilled neutralisation buffer was added and immediately mixed by gentle inversion, ensuring complete precipitation of cellular debris.
6. The lysate was centrifuged at 7000 x g for 30 minutes at room temperature, following which the cell lysate was clear of debris.

7. The lysate was filtered through the PureYield Clearing Column system (Promega) and eluted using a vacuum manifold.
8. The eluate, containing purified plasmid, was stored at 4°C or -20°C.

2.3 Mammalian Cell Culture

2.3.1 Standard Cell Culture

Culturing cells denotes the method by which cells from a specific tissue of a human or model organism are grown *in vitro* for use in experiments. Originally the method used cells isolated from cancerous tumours, as these cells were naturally immortalised (i.e. cells that have mutations that effect a cells normal senescence pathways, and thus proliferate indefinitely), but this meant dependency on biopsy from tissue. Methods to generate immortalised cell lines using a virus (e.g. the Epstein Barr virus) to mutate cells means that theoretically any cell line can be immortalised. Cell lines in culture are useful as a tool in that they enable experiments in a model environment. They enable the investigation of the molecular and cell biology particular to a given cell line, and reproducibility of results due to the identical nature of these populations of cells. However, due to the fact that the cell lines are mutated, this may alter the biology of the cell and is a caveat to any analysis performed that utilises them.

This project used the SH-SY5Y cell line in all cell culture led experiments, a human derived neuroblastoma cell line which originated as a subclone of the SK-N-SH cell line. This cell line was established from a female patient and thus the cell line is genetically female. The SH-SY5Y cells used throughout this project were purchased from the European Collection of Cell Cultures (ECACC).

1. Cells were plated in Dulbecco's Modified Eagle Medium (DMEM, 2.5 mM Glutamine, HyClone Labaoratories), 15% Foetal Bovine Serum (FBS) at 37°C, 5% CO₂ in 25 cm² or 225 cm² vented tissue culture flasks from Sarstedt.
2. Cells were harveseted and split by trypsinisation using trypsin (10X, Sigma) diluted to 1X in PBS (Phosphate Buffered Saline, Sigma).
3. The cells were centrifuged at 1200 rpm for 2 minutes and the supernatant discarded.
4. The cells were then resuspended in 10ml of fresh media and then 0.5 ml of this cell suspension was plated into a 25 cm² flask containing 9.5 ml of fresh media (or 4.5 ml into 85.5 ml of fresh media).

2.4 Transfection of plasmid DNA into mammalian cell lines

2.4.1 Transfection Protocol

Transfection is a method whereby nucleic acids, such as a plasmid, are introduced into a eukaryotic cell. There are several methods to achieve this but the liposomal method is described here. The recombinant and control pGL3 vectors and the internal control vector Renilla were prepared via bacterial transformation, described in Section 2.2 above. Renilla was used as an internal control to measure the success of the transfection and to normalise against the experimental vector for gene reporter assays, described in Chapter 3. First the DNA was diluted in 50 μ l serum free RPMI media (Sigma) and then mixed with 3 μ l Lipofectamine (Invitrogen). At this stage the lipofectamine forms liposomes that contain the DNA – the liposomes will then fuse with any cell membranes into which they come in contact and release the DNA into the cell. 24 well cell plates were used to grow the cells, and were first washed with 400 μ l PBS per well. 400 μ l DMEM (10% FBS, 2.5 mM Glutamine) was added to each well, followed by 100 μ l DNA-Lipofectamine mix. The cells are then incubated at 37°C for 48 hours before subsequent experiments were carried out.

2.4.2 Luciferase Assay

The luciferase gene can be used as a reporter gene in promoter assays, used to assess the effect of a putative or known promoter element upon gene expression. Here, the activity of the promoter effects the expression of the luciferase protein, which in turn fluoresces in the presence of luminol substrate – thus the more the promoter enhances expression of luciferase, the brighter the reaction volume will glow, which can be read using a fluorometer. The renilla vector, described above, was used as an internal control, against which the fluorescence was normalised – thereby corrected for variances in cell density and transfection efficiency between wells/reactions. The Dual Luciferase Reporter Assay System (Promega) was used here to analyse the effect of the *MACROD2* putative promoter region upon expression of the luciferase reporter gene in the SH-SH5Y cell line. The cells are first washed in 400 μ l PBS, ensuring all residual media was removed. 100 μ l of lysis buffer was added to each well and incubated for 20 minutes at room temperature (~21°C) and then stored at -20°C overnight to improve cell lysis. The luciferase assay was then performed using a plate reading spectrophotometer (Viktor). A fresh opaque 96-well plate was used with the spectrophotometer. 100 μ l luciferase assay reagent was added to

each required well and then mixed with 20 μ l of lysed cells per well. The luciferase luminescence was then measured on the plate reader. 100 μ l 'Stop-n-Glo' solution was added to each well and mixed by pipetting. The renilla luminescence was then measured on the plate reader. These two values – the luciferase luminescence and the renilla luminescence – are used in calculations to determine the relative luminescence of the cells and therefore the relative expression of the luciferase gene, which was determined by the activity of the putative promotor element.

Chapter 3 Investigation of the gene structure and expression of *MACROD2* in the human brain

3.1 Introduction

The strongest association from the Stage 1 Autism Genome Project GWAS (Anney, Klei, Pinto, Regan, Conroy, Magalhaes, et al. 2010) was observed at a locus at chromosome 20p12.1. The association was located in the intron of an uncharacterised transcript *c20orf133*. *C20orf133* was later termed *MACROD2* (MACRO domain containing 2). The strongest associated marker (rs4141463) is located between exon 5 and 6 of *MACROD2*. The strongest association does not necessarily equate to causation, with the possibility of linkage disequilibrium influencing this signal such that a “functional” variation is located at an adjacent site. At the time of discovery the only transcripts discovered in the region immediately surrounding the associated locus were *MACROD2* and *FLRT3* (Fibronectin leucine rich transmembrane protein 3). A non-coding RNA was subsequently identified, which has now been named as *MACROD2* Anti-Sense 1, or *MACROD2-AS1* for short. It is unknown if this anti-sense RNA plays any regulatory role in the translation of the *MACROD2* mRNA, although it should be noted that antisense RNAs typically interact with neighbouring genes (Katayama et al. 2005). A number of neighbouring genes also surround this locus, including: *SEL1L2* (Sel-1 suppressor of lin-12-like 2), *KIF16B* (Kinesin family member 16), *NDUFAF5* (NADH dehydrogenase (ubiquinone) complex I, assembly factor 5) and multiple pseudogenes (see Figure 3.1).

A summary of the genes at this locus are given in Table 3.1. These genes include many whose biological function is in the regulation of transcription. Whether this represents synteny at this locus is unknown. However, it is plausible that any causative variation is likely to influence one or more of these gene. Of these genes, the most biologically plausible in terms of the association were determined to be *FLRT3*, *KIF16B*, *MACROD2* and *NDUFAF5*.

FLRT3 had previously been identified as a transmembrane protein, with a role in cell adhesion and receptor signalling. In the context of autism however, it is the observation that this gene is important in the promotion and regulation of neurite outgrowth that highlights it as an interesting candidate (Tsuji et al. 2004; Robinson et al. 2004). *KIF16B* was reported with a function in intracellular transport of endosomes, with

implications for cell signalling (Hoepfner *et al.* 2005), and a member of the same protein superfamily has been reported to have a role in axonal transport of synaptic vesicles (Okada *et al.* 1995).

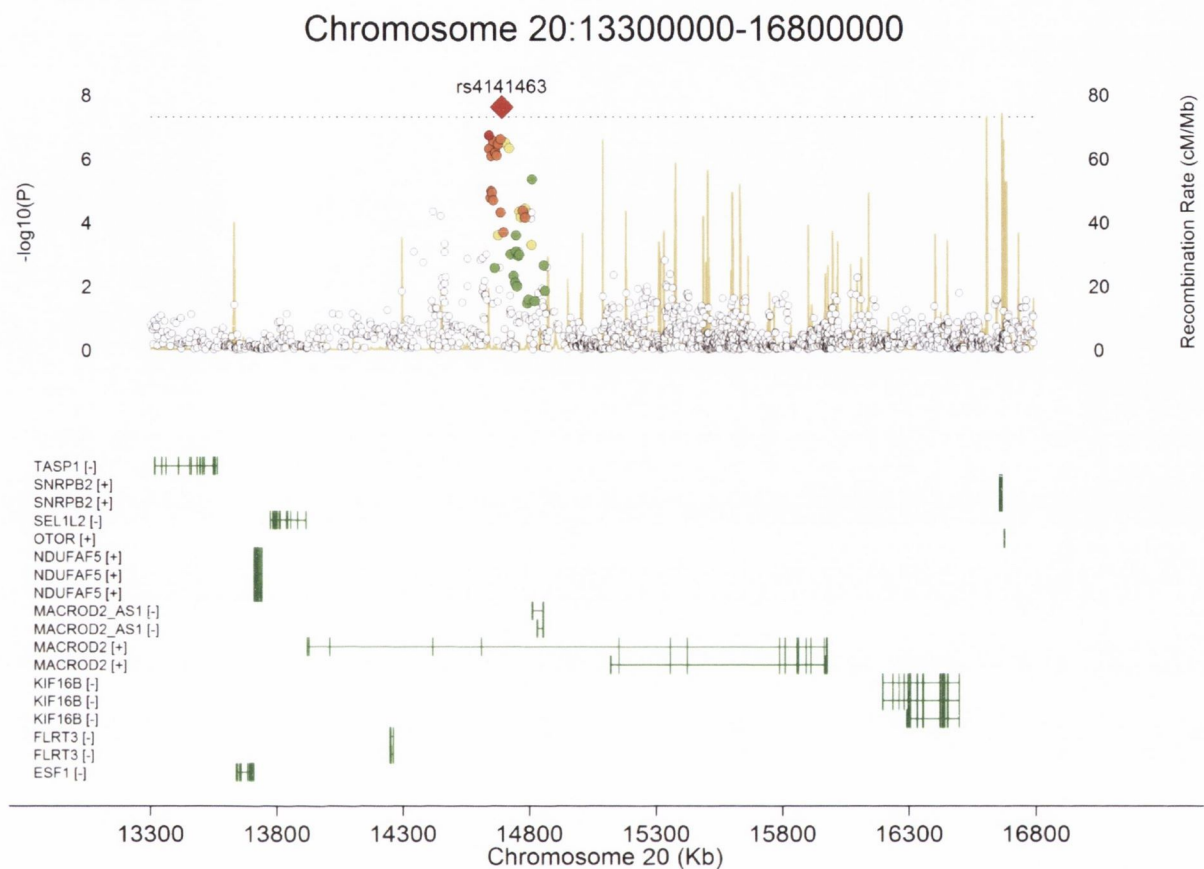


Figure 3.1 The top association from the Autism Genome Project Stage 1 GWAS study (adapted from Anney *et al.*, 2010). Each circle on these plots indicates an individual marker SNP. On the left y-axis of the association plot is the $-\log_{10}(P)$ scale of genome-wide significance. The red diamond is the most significant marker, which was rs4141463 for the European strict analysis. On the right y-axis is the recombination rate, indicated by tan lines on the association plot. Associated markers in linkage disequilibrium with the most significant marker are highlighted with colour, in order of significance from orange to yellow. A plot of the genes at this locus and the genomic location are mapped below the association plot. Forward and reverse strands are indicated by [+] and [-] respectively.

Table 3.1 The *MACROD2* locus contains multiple genes related to transcriptional regulation. A literature search to investigate the function of the genes at the *MACROD2* locus finds multiple genes with roles in transcriptional regulation and development. At the outset of this project, no functional role for *MACROD2* had been documented, however this table has been updated to represent recent developments.

Gene name	Full name	Protein Function
<i>TASP1</i>	Taspase (Threonine aspartase 1)	Hox gene expression (Hsieh <i>et al.</i> 2003)
<i>SNRPB2</i>	U2 small nuclear ribonucleoprotein	Pre mRNA splicing (Hong <i>et al.</i> 1997)
<i>SEL1L2</i>	Sel-1 suppressor of lin-12-like 2	Unknown
<i>OTOR</i>	Otoraplin	Cartilage development and hearing (Robertson <i>et al.</i> 2000)
<i>NDUFAF5</i>	NADH dehydrogenase (ubiquinone) complex I, assembly factor 5	Mitochondrial complex I (Sugiana <i>et al.</i> 2008)
<i>MACROD2</i>	Macro domain containing 2	O-acetyl-ADP-ribose deacetylase (Chen <i>et al.</i> 2011)
<i>KIF16B</i>	Kinesin family member 16	Intracellular transport of endosomes (Hoepfner <i>et al.</i> 2005)
<i>FLRT3</i>	Leucine-rich repeat transmembrane protein	Cell adhesion, excitatory synapse development (O’Sullivan <i>et al.</i> 2012)
<i>ESF1</i>	Nucleolar pre-rRNA processing protein, homolog	Transcriptional regulation (Oda <i>et al.</i> 2004)

More recently *KIF16B* has been highlighted in a GWAS of intelligence (Loo *et al.* 2012) which could indicate a function in a disorder with reduced intellectual functioning such as autism. *NDUFAF5*, under the positional name *C20orf7*, has been demonstrated to be a crucial component of mitochondrial complex assembly, with mutation at this locus causing lethal neonatal mitochondrial disease (Sugiana *et al.* 2008). Defective mitochondrial complex I assembly due to mutation within *NDUFAF5* has also been associated with the rare neurodegenerative disorder Leigh syndrome (Gerards *et al.* 2010).

However, the primary focus of this thesis has been the study of the gene *c20orf133*, latterly termed *MACROD2*. The gene is highly conserved across different model organisms as seen in Table 3.2. This may indicate an essential role but to date no knock-out of this gene has been performed to investigate this possibility. The function of this gene was largely unknown, although there were some hints from the literature to associate it with neurological and neuropsychiatric conditions (Maas *et al.* 2007; Bendavid *et al.* 2009; Baranzini *et al.* 2009) including associations of recurrent CNV with schizophrenia (Xu *et al.* 2009). In this chapter I focused on approaches to extend additional knowledge on the structure and function of the gene previously known as *c20orf133*.

Table 3.2 The evolutionary conservation of MACROD2: MACROD2 is a highly conserved protein, particularly among mammals, which could indicate an essential role. This data was retrieved from:
http://www.ncbi.nlm.nih.gov/homologene?cmd=Retrieve&db=homologene&dopt=Alignments&cores&list_uids=85987

Species	Gene Symbol	Protein Identity (%)	DNA Identity (%)
<i>H. Sapiens</i> (Humans)	MACRO D2	N/A	N/A
vs. <i>P. troglodytes</i> (Chimpanzee)	MACRO D2	98.1	99.1
vs. <i>M. mulatta</i> (Macaque)	MACRO D2	96.9	97
vs. <i>C. lupus</i> (Dog)	MACRO D2	84	87.9
vs. <i>M. musculus</i> (Mouse)	Macro d2	81.9	86.3
vs. <i>R. norvegicus</i> (Rat)	Macro d2	80.6	84.9
vs. <i>G. gallus</i> (Chicken)	LOC421259	40.1	60.3
vs. <i>D. rerio</i> (Zebrafish)	macro d2	52.3	57.6
vs. <i>C. elegans</i> (Nematode worm)	B0035.3	46.2	52.3
vs. <i>A. thaliana</i> (Thale cress)	AT2G40600	50.3	55.9
vs. <i>O. sativa</i> (Rice)	Os03g0336500	45.7	51.6

3.2 Aims of this study

The work reported here represents an analysis of the gene structure and expression of a poorly characterised gene that has been associated with autism, and thus may have relevance to human health. This chapter describes an effort to understand the basic biology of this gene, in the hopes that it will inform further analysis.

The three core aims of this study were to confirm and expand the knowledge of the **structure** of the gene, characterise the **expression** profile, and finally to explore the role of **variation** on the expression of this gene.

3.2.1 Characterisation of the Structure of *MACROD2*

According to the NCBI RefGene database the *MACROD2* gene locus is one of the largest in the human genome, at 2,057,828 base pairs in length. The gene is encoded by 17 exons, encompassing an mRNA of only 4863 base pairs; only 0.23% of the *MACROD2* locus actually encodes the *MACROD2* mRNA. Two antisense transcripts have been identified within the *MACROD2* locus: *FLRT3* and *MACROD2-AS1* (See Figure 3.2).

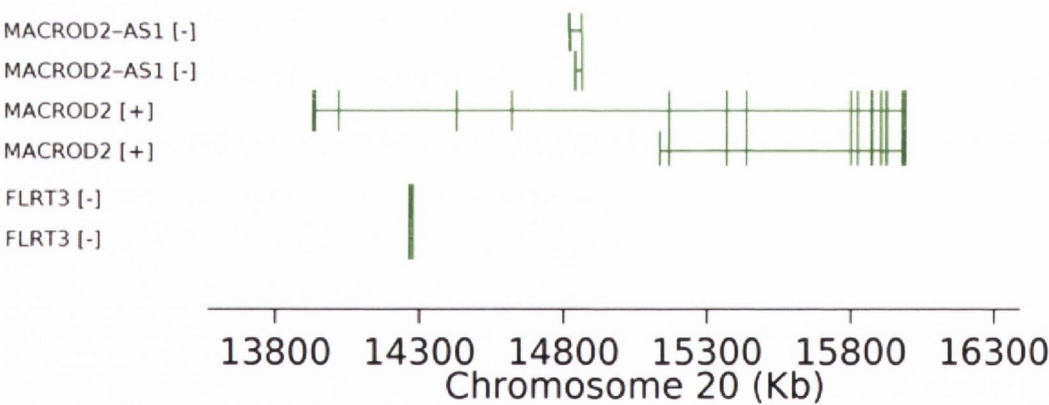


Figure 3.2 A closer look at the *MACROD2* locus, illustrating the vast length of the gene and the number of exons and intergenic transcripts. Forward and reverse strands are indicated by [+] and [-] respectively.

Historically, the 5' structure of any gene, specifically the untranslated regions can often be misrepresented in databases. The observation of the second isoform of the *MACROD2* gene led to the hypotheses that this represents an abridged cDNA as opposed to a different isoform. To test this and also whether this is a true isoform amongst other isoforms I sequenced the 5' region of the gene's cDNA using the 5'Rapid Amplification of cDNA Ends approach in human brain cDNA.

Moreover, to better understand the regulation of the gene, I also performed promoter mapping of the 5' flanking region of the cDNA ends to confirm promoter activity and identify putative loci which strongly influence gene expression.

3.2.2 Characterisation of the Expression of *MACROD2*

The expression profile of *MACROD2* is of interest because depending upon where the protein is expressed it can give some hints as to the role of the protein in overall biology. For a protein that confers putative susceptibility to autism or other neuropsychiatric disorders, we would hypothesise that neuronal and brain expression is necessary. The expression profile of *MACROD2* has been characterised via in situ hybridisation in mice which illustrated that murine *MACROD2* is fairly ubiquitously expressed, although it is highly expressed in the developing brain (Maas *et al.* 2007). I will experimentally validate presence of *MACROD* transcripts across a range of human cDNA panels.

3.2.3 Characterisation of the Role of Variation on the Expression of *MACROD2*

One of the hypotheses regarding association of non-coding variants with a disorder is that these variants may impact the expression of the gene. I will also endeavour to examine the role of expression on the *MACROD2* transcript.

I assessed the influence of allele on allele-specific expression using an Allele Expression Imbalance (AEI) approach. AEI experiments enable us to look for preferential expression of one allele over another - for example, a variant in a promoter region may increase or decrease expression of a gene by affecting transcription factor binding at that site.

Variation in the coding sequence itself can have drastic effects upon the expression of a gene, for example, introduction of a stop codon prematurely or frame-shift mutations can completely knock out the function of a gene. I will examine the range of coding variants in the *MACROD2* gene by sequencing of individuals with ASD as part of our local targeted resequencing project. I will also examine the CNV pattern at *MACROD2*, from data available to me from the AGP, but also via literature searches and use of the DGV (Database of Genomic Variants).

Finally, I will explore the influence of variation on *MACROD2* expression in brain tissue and the influence of variation at the *MACROD2* locus on the expression of other genes. Using public domain genome-wide SNP data I will perform association analysis on expression data from Myers *et al.* (2007). Myers and colleagues generated expression and genotype data from the human cortex from control brains. Genotyping was performed using the Affymetrix 500K array, whilst expression levels were assayed using the Illumina HumanRefSeq-8 expression bead chip. I will use this data to look for *cis* or *trans* effects of variation at *MACROD2*. A *cis* eQTL is a locus that effects expression of a gene at that locus, whereas a *trans* eQTL is a locus that effects expression of a gene anywhere else in the genome.

Expression: To examine the expression profile of transcripts at the *MACROD2* locus, with a focus on *MACROD2* and the brain.

Structure: To identify what isoforms of *MACROD2* exist and whether there is a brain specific isoform of *MACROD2*, and also to characterise the regulatory elements of *MACROD2*.

Variation: To examine what kinds of variation exists at the *MACROD2* locus, with a focus on whether *de novo* and rare mutation at this locus exists in autism samples. I also seek to examine whether common variation at *MACROD2* influences *cis* or *trans* gene expression.

3.3 Materials and Methods

3.3.1 Characterisation of the Structure of *MACROD2*

3.3.1.1 5' Rapid Amplification of cDNA Ends

I sought to identify what isoform of *MACROD2* is expressed in the human brain. This was accomplished using 5' RACE (Rapid Amplification of cDNA Ends), an advanced PCR based technique that uses nested primers specific to cDNA to amplify large fragments of the cDNA from a given tissue in the 5' direction. RACE has previously been used to identify the gene structure at the 5' and 3' end of genes (e.g. R. J. Anney *et al.* 2002). Briefly, first strand cDNA synthesis is generated using a primer (SP1) specific to the transcript of interest. The resultant cDNA is dATP tailed and a corresponding poly-T primer is used with a second specific primer (SP2) which is nested 5' to SP1 to amplify the transcript. The resultant amplicons can then be further enriched using an additional specific nested primer, SP3, and an anchor primer. The primer sequences are provided for reference in Table 3.3. A diagram of the 5'RACE protocol is shown in Figure 3.3.

For *MACROD2*, 5'RACE was used to characterise the transcripts expressed in brain sourced RNA. I designed the SP1 primer to target the transcript 3' of the exons where the known isoform 1 and isoform 2 deviate to determine which isoform is expressed in each tissue. The resulting fragments from the RACE were cloned into a T-vector (pGEM-T-easy, Promega) and their DNA sequence determined and aligned to the human genome sequence template to infer gene structure. The resulting cDNA fragments were sequenced and compared against the reference sequence of *MACROD2*. cDNA derived from human total RNA was used in these experiments across whole brain, foetal brain, cerebellum and testis.

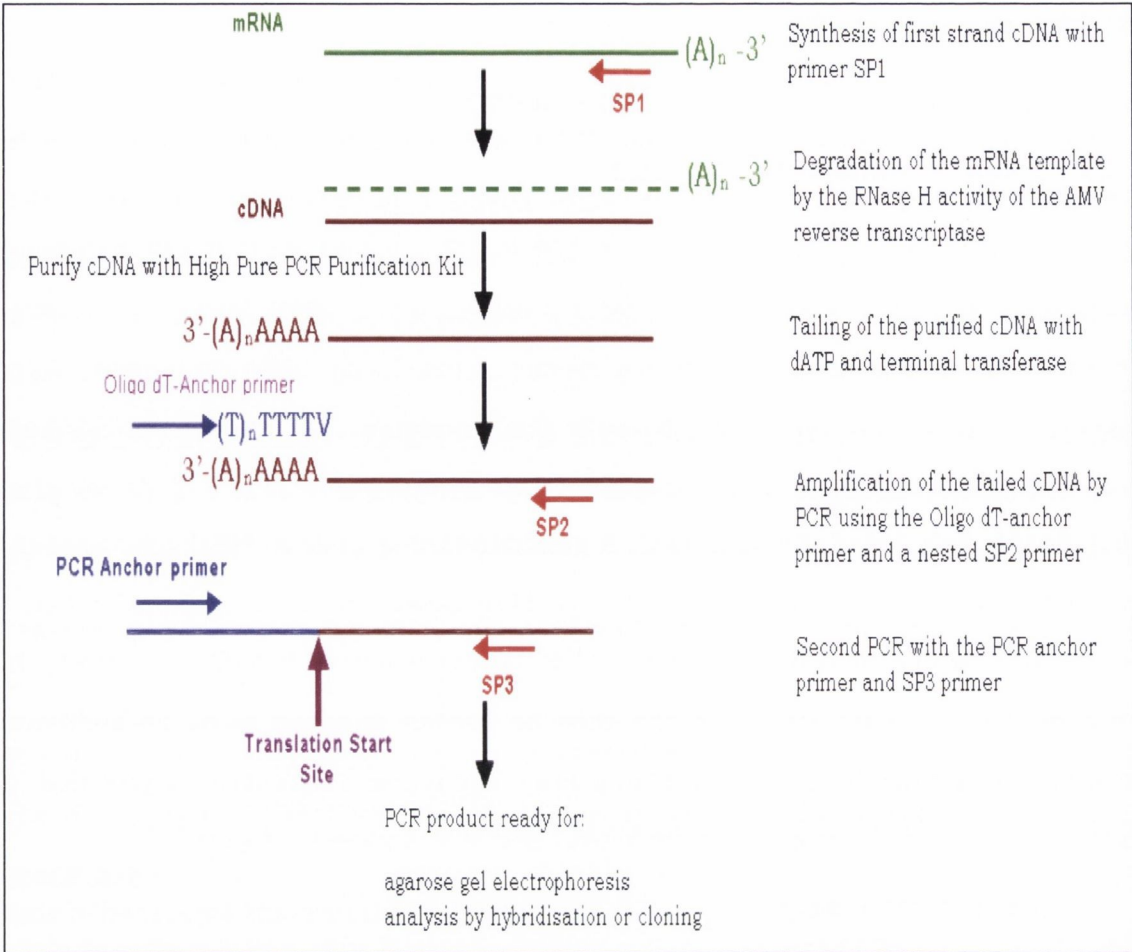


Figure 3.3 5' RACE is a specialised PCR-based technique that involves the use of three nested primers (SP1, SP2 and SP3) to amplify the full length of a cDNA in the 5' direction. The length of the synthesised cDNA is thus limited only by the effectivity of the polymerase used and not by the use of a reverse primer. However, this lack of a reverse primer means that downstream analysis of the 5' RACE product is dependent on ligation into a plasmid vector for sequencing.

Table 3.3 A table of the oligonucleotide nested primers used in the 5' RACE experiments. Three nested primers are used as part of the standard 5' RACE method, as illustrated in Figure 3.3.

Primer	Primer sequence
MACROD2 Exon 9 SP1	CGGAGAAAACTCATTCATTTTC
MACROD2 Exon 9 SP2	TTTTTGTAGATTTTGAAGTCAACTTC
MACROD2 Exon 9 SP3	CACAGAAAATGATCCGATCCA

3.3.1.2 Promoter Mapping

I sought to characterise the regulatory elements of *MACROD2*, by isolating fragments of varying size from the putative regulatory region of *MACROD2*, which was determined as the 1000 base pairs immediately prior to the start codon of exon 1. These fragments were cloned into a vector plasmid containing a luciferase reporter gene and transfected into a neuronal cell line. The expression of the reporter gene is measured and normalised against an internal control. This is repeated across multiple experiments and the data is analysed and reports the relative expression effect of each putative regulatory region DNA fragment.

3.3.1.2.1 Isolation of the putative regulatory region of *MACROD2* and Cloning into the vector plasmid

Primer sets were designed using Primer3 (<http://frodo.wi.mit.edu/primer3/>) to amplify five fragments of increasing size from the putative regulatory region of *MACROD2*. This resulted in 5 PCR products of increasing base pair length: MD1: 408bp, MD2: 505bp, MD3: 620bp, MD4: 740bp and MD5: 890bp. These fragments were visualised on a 1.5% agarose and TAE buffer gel, extracted and purified using the Wizard SV Gel and PCR Clean-Up System (Promega), described in Chapter 2. The resulting fragments underwent a Poly-A tailing reaction and once more were visualised on a 1.5% TAE buffer gel, with each band being excised and purified using the Wizard SV Gel and PCR Clean-Up System (Promega). Each fragment was then ligated into T-Easy Vector Plasmid (Promega) and then transformed into competent *E. coli* cells (JM109, Promega). The transformed cells were inoculated onto SOC agar plates and cultured for 16-18 hours at 37°C. 24 colonies from each plate were then chosen and inoculated into 100 µL SOC media in a 96-well plate and cultured for 14-16 hours at 37°C. A colony screen PCR was performed using a forward primer specific to the T-Easy plasmid (pUC/M13 forward primer) and a reverse primer specific to the *MACROD2* inserts. The resulting PCR products were visualized on a 1.5% TAE agarose gel. Where a band was visible, this indicated a colony that had the *MACROD2* fragment containing plasmid. Three to four of such colonies were selected and inoculated into 10mL SOC media and incubated for 16-18 hours at 37°C, with shaking, at 150 rpm. The cultures were then harvested for vector plasmid containing *MACROD2* insert using the MIDI Prep system (Promega) and

quantified on a NanoDrop 2000 (Thermo Scientific). All purified plasmids concentrations were adjusted to 100 ng/ μ L.

3.3.1.2.2 Sequence confirmation of insert and re-insertion in reporter plasmid

Following cloning of PCR products into a vector plasmid, it is necessary to confirm the presence of the insert in the plasmid and that the insert is in the right direction, i.e. that the fragments were inserted in the 5' -> 3' direction. This was achieved here using DNA sequencing. Each vector plus *MACROD2* insert was prepared using the sequencing reaction described in Table 3.4 and resulting products were purified using the DyeEx 2.0 spin columns system (Qiagen) and sequenced on an Applied Biosystems 3130XL Genetic Analyser. The results were analysed using the BioEdit Sequence Alignment Tool to confirm presence and orientation of the insert. Restriction digest of the plasmids containing the *MACROD2* inserts was then performed and the products were visualised on an agarose gel. The digest resulted in two bands: a large band representing the vector plasmid and a smaller band representing the insert. This smaller band was excised and purified using the Wizard SV Gel and PCR Clean-Up System (Promega) and then ligated into the pGL3 reporter plasmid and cloned into *E. coli* competent cells. The sequencing protocol was then repeated as above for the pGL3 plasmid containing the *MACROD2* inserts to confirm the presence and orientation of the insert.

This two-step plasmid insertion protocol was used due to the poor performance of ligating directly into the pGL3 vector, with the T-Easy vector plasmid enabling insertion of blunt end DNA fragments.

Table 3.4 Sequencing reaction used with the ABI 3130XL Genetic Analyser. The sequencing reaction was performed as advised by the manufacturers guidelines, optimised with the use of between 200 and 400 ng plasmid DNA in total.

Reagents (μ L)		Program	
BigDye Reaction Mix	4	95°C for 30 seconds 95°C for 10 seconds 55°C for 10 seconds 60°C for 3 minutes 10°C hold } Cycle 24 times	
10X BigDye Terminator v3.1	1		
Sequencing Buffer			
Forward and Reverse Primers	0.5 + 0.5		
Vector Plasmid (50-100ng/ μ L)	4		
Total	10		

3.3.1.2.3 Transfection and Characterisation of the regulatory elements of *MACROD2* using reporter gene assays

Recombinant and control pGL3 vectors and the internal control vector Renilla were prepared via bacterial transformation, described in Chapter 2. The Renilla vector was used as the internal control against which the luciferase is normalised. SH-SY5Y cell lines were chosen as the experimental cell model, and for each transfection experiment a 24-well plate was prepared. The transfection protocol proceeded as described in Chapter 2.

3.3.1.2.4 Normalisation and statistical analyses

The luciferase and renilla luminescence values are used to determine the relative luminescence of the cells and therefore the relative expression of the luciferase gene, the expression of which is determined by the activity of the putative promotor element inserted into the plasmid. As such the relative expression is calculated by normalising the luminescence of the pGL3 luciferase against the luminescence of the renilla luciferase internal control, which gives the normalised relative activity. The mean relative activity for each *MACROD2* insert is calculated against that of the control pGL3 vectors.

3.3.1.2.5 *In silico* analysis of the putative regulatory region of *MACROD2*

MatInspector (<http://www.genomatix.de>) was used to examine the putative regulatory region of *MACROD2* to look for core transcription factor and other regulatory domains. The ~900 base pairs that were cloned into the PGL3 vector was used as the input for the MatInspector analysis, which used the NCBI build 37 of the human genome for reference. Similarity scores are based upon matches to the core and the full promoter sequence.

3.3.2 Characterisation of the Expression of *MACROD2*

Due to the lack of experimentally validated information surrounding the expression profile of *MACROD2*, I sought to characterise this by using exon-crossing primers and cDNA from human tissues.

3.3.2.1 Human Total RNA Master Panel

Human RNA was sourced from a Human Total RNA Master Panel II (Clontech). This panel contains RNA from 20 different human tissues. Each sample is a pool of tissue from several individuals of varying gender and age, as detailed in Table 3.4.

Table 3.4 Tissue Sources for the Human Total RNA Master Panel II (Clontech), according to the supporting documentation that came with the panel. The majority of samples are pooled from multiple individuals across a wide age range.

Tissue	Source
Adrenal Gland	Pooled from 62 male/female Caucasians, ages 15-61
Colon w/ Mucosa	Pooled from a 23 year old female Caucasian
Brain, cerebellum	Pooled from 10 male/female Caucasians, ages 22-68
Brain (whole)	18 year old male Caucasian
Foetal brain	Pooled from 59 spontaneously aborted male/female Caucasian foetus es, ages 20-33 weeks
Foetal liver	Pooled from 63 spontaneously aborted male/female Caucasian foetuses, ages 22-40 weeks
Heart	Pooled from 3 male Caucasians, ages 30-39
Kidney	40 year old female Caucasian
Liver	51 year old male Caucasian
Lung	Pooled from 3 male/female Caucasians, ages 32-61
Placenta	Pooled from 15 Caucasians, ages 19-33
Prostate	Pooled from 12 Caucasians, ages 20-58
Salivary gland	Pooled from 24 male/female Caucasians, ages 16-60
Skeletal Muscle	Pooled from 2 male/female Caucasians, ages 43 and 46
Spleen	Pooled from 12 male/female Caucasians, ages 18-54
Testis	Pooled from 39 Caucasians, ages 14-64
Thymus	Pooled from 2 male Caucasians, ages 18 and 57
Thyroid gland	Pooled from 64 male/female Caucasians, ages 15-61
Trachea	Pooled from 22 male/female Caucasians, ages 18-54
Uterus	Pooled from 8 Caucasians, ages 23-63

3.3.2.2 RNA isolation from cell culture

To complement the master RNA panel, RNA was isolated from lymphoblasts and SH-SY5Y cells in culture using the RNeasy Plus Micro with QIA Shredder kit (Qiagen) following the manufacturer’s guidelines.

3.3.2.3 Expression profiling of the *MACROD2* message across human tissues

Exon-crossing primers were designed using Primer3 (<http://frodo.wi.mit.edu/primer3/>) to demonstrate the expression profile of *MACROD2* and *MACROD2-AS1* across a panel of cDNAs derived from 21 human tissues (Clontech, Palo Alto) using PCR, with genomic DNA as an internal negative control. Where a band corresponding to a specific tissue is visible on the agarose gel following PCR, it suggests that the gene the primers target is expressed in those tissues. However, the primers are designed in such a way that they will not return a discrete band with genomic DNA, as the fragment to be amplified, when including introns, would be too big for successful amplification by PCR.

3.3.3 Characterisation of the Role of Variation on the Expression of *MACROD2*

3.3.3.1 Allelic Expression Imbalance

Allelic expression imbalance is a real time PCR based method that calculates any difference in the relative expression of two alleles. Briefly, the dosage of alleles is calculated from cDNA of a heterozygous individual. This is normalised to the individual's genomic DNA (i.e an expected 50:50 dose). Where the SNP exerts a functional effect, you will observe a deviation from the 50:50 ratio in the cDNA. In extreme cases such as complete imprinting at the locus, we would observe mono-allelic expression in the cDNA but bi-allelic data from the genomic DNA.

In the experiment I used cell-lines derived from the CEPH HapMap Collection. The limiting factor for choosing a SNP is the allele frequency and consequently the number of heterozygous individuals to examine. SNPs within the *MACROD2* coding sequence with at least 5 heterozygotes represented in the HapMap sample pool were examined for AEI. All HapMap lymphoblasts had previously been sourced from the Coriell institute by the NPG group, cultured in-house and RNA was isolated using the procedures described above.

To select for suitable SNPs, first all coding SNPs within *MACROD2* were selected using dbSNP, which returned 27 SNPs, the majority of which were located within the 5' and 3' UTR of the gene. 11 SNPs met the allele frequency cut-off of at least five heterozygotes represented in the available HapMap sample pool, which were then selected based upon linkage disequilibrium using Haploview to return the most informative SNPs from the region. In total, 5 linkage independent SNPs met the criteria and corresponding genotyping assays were available from Applied Biosystems (Table 3.5). The assays were first tested using genomic DNA to confirm their efficiency. All available heterozygotes for each SNP from the pool of HapMap CEPH individuals were analysed in triplicate (for both cDNA and genomic DNA for the individual) on an ABI 7900HT Fast Real-Time PCR system (Applied Biosystems) using the reagent mix described in Figure 3.5. Each assay contained probes labelled with 'FAM' and 'VIC' dyes. Due to poor performance of these assays during optimisation the protocol was adjusted such that reagents were used at 2X final concentration, which was found to improve results.

Table 3.5 Minor Allele Frequency and number of representative heterozygotes for each SNP assayed for AEI. Due to overall poor SNP documentation within *MACROD2*, very few SNPs with enough representative heterozygotes were available from within the HapMap CEU sample pool.

Marker	Location in MACROD2	Minor Allele Frequency	HapMap CEU Heterozygotes
rs1225891	3' UTR	0.486	33
rs1238278	3' UTR	0.292	19
rs6080109	3' UTR	0.087	9
rs11696319	3' UTR	0.018	11
rs16997016	3' UTR	0.037	5

Taqman Allelic Discrimination	Reagents (µl)		Program	
	Universal Master Mix	2.5	50°C for 2 minutes	} Cycle 40 times
	40X Taqman Assay Mix	1	95°C for 10 minutes	
	DNA	2	95°C for 15 seconds	
	H ₂ O	1.875	60°C for 1 minute	
	Total	7.375		

Figure 3.5 TaqMan Allelic Discrimination Reaction. The reaction was performed as per the manufacturers guidelines, however the protocol was adjusted such that reagents were used at 2X final concentration, which was found to improve results.

Following RT-PCR, the SDS software package (Applied Biosystems) was used to perform an analysis of the expression from each allele, which is reported as threshold amplification cycle data, or ‘Ct’. It is the ‘Δ-Ct’ that is used to calculate the relative allele frequency (rF), i.e. the Δ-Ct is the difference in the threshold cycle at which each allele is recorded. The equation used to calculate rF (adapted from Sham *et al.* 2002) is as follows:

(1) $\Delta Ct' = (Ct \text{ Allele 1 (cDNA)} - Ct \text{ Allele 2 (cDNA)}) - (\Delta Ct \text{ gDNA})$
(2) $rF = (1 / (2^{\Delta Ct'} + 1))$

The average Ct and standard deviation for each allele was calculated from the Ct values from cDNA for each individual, which was then normalised against the Ct values reported for each allele from genomic DNA. Essentially, any difference in the ΔCt for the genomic DNA is subtracted from the difference for the cDNA, which has been demonstrated previously to correct for any differences in the amplification of an allele arising from the assay itself (Hoogendoorn *et al.* 2000). rF values range from 0 to 1, with rF = 0 representing total expression of Allele 1, rF = 1 representing total expression of Allele 2 and rF = 0.5 representing equal expression of both alleles. For this analysis, allelic expression imbalance was denoted as any value for rF over 0.4 or below 0.6, i.e. a 20% difference in expression levels.

3.3.3.2 Validation of Missense variants from the Exome sequencing study

The exome sequencing study performed by the NPG group identified 3 mis-sense variants within *MACROD2* that were brought forward for validation. Primers were designed to amplify short (~300bp) fragments surrounding the identified mutation and first optimised using control DNA using PCR optimisation buffers (CamBio). Once the right PCR conditions were found, PCR was performed using DNA from individuals in which the mutations had been identified via exome sequencing, and also using DNA from the parents of each proband. Raw PCR product was sent to Source Bioscience (Dublin), who purified and sequenced the products. Data was returned as a sequencing chromatogram, which was analysed using the freely available BioEdit software (<http://www.mbio.ncsu.edu/bioedit/bioedit.html>).

3.3.3.3 Characterisation of *MACROD2* CNV pattern

The *MACROD2* locus is a known hotspot for CNV but I wished to characterise this variation in greater detail. CNV data from the AGP sample was examined for CNVs that crossed exons or were *de novo* CNVs, and cross-referenced with phenotype to see if these CNVs segregated with severity of the disorder. The DGV (<http://projects.tcag.ca/variation/>) was then used to identify the CNV pattern at this locus in the wider population.

3.3.4 Preparation of data for GWA and eQTL analysis: Quality Control of the Myers dataset

Before eQTL analysis the data was run through a quality control (QC) process to maximise reliability of the analysis. The Myers dataset in .ped and .map format was downloaded from the Myers lab data website at: <http://labs.med.miami.edu/myers/LFuN/data%20ng.html> and updated to the human genome version 135, build 37 as reference. The Myers dataset includes 193 individuals sourced from the National Institute of Aging Alzheimer Center and the Miami Brain Bank. The individuals were self-defined as ethnically of European descent with no clinical history of stroke, cerebro-vascular disease, Lewy bodies or co-morbidity with any other known neurological disease. Ethnic outliers and samples that were possibly related were also excluded. The sample was genotyped on an Affymetrix GeneChip Human Mapping 500K Array. Before quality control (QC) of the data using the whole genome analysis toolset PLINK (S. Purcell et al. 2007), there were 499,571 markers and 193 individuals (107 males, 86 females). Only markers with a minor allele frequency of >0.05 were considered, due to the modest sample size of 193 individuals. Following this QC pipeline (described in Figure 3.6), the cleaned Myers dataset now included 387,163 markers and 191 individuals (106 males, 85 females).

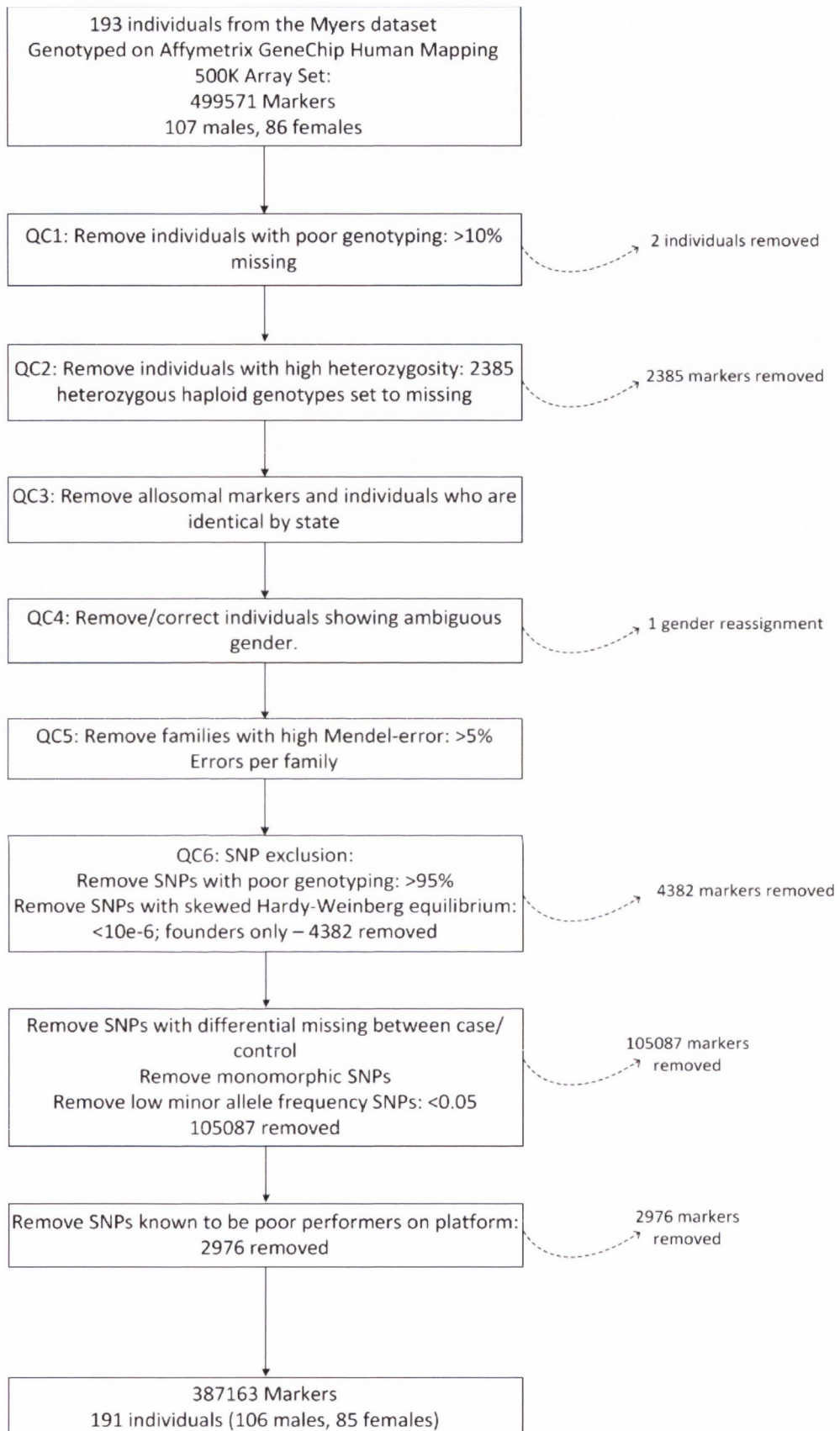


Figure 3.6 Comprehensive QC using PLINK was performed to prepare the Myers dataset for use in subsequent analysis. The QC checks are performed to remove problematic samples or SNPs that may create 'noise' in the analysis. In total, 1 individual with ambiguous gender was reassigned; 2 individuals and 112408 SNPs were removed.

3.3.5 Preparation of data for GWA and eQTL analysis: Data pruning

To identify regions of the genome that may be influencing or are being influenced by *MACROD2*, a genome-wide association analysis of normalised gene expression was performed to identify expression quantitative trait loci (eQTL). An eQTL is a genomic loci that regulates expression levels of mRNAs or proteins. Those eQTLs that are mapped in close proximity to the transcript-of-origin are referred to as *cis*-eQTLs, whereas those eQTLs which map far from their transcript-of-origin are referred to as *trans*-eQTLs. The data source was whole-genome genotyping (Affymetrix GeneChip Human Mapping 500K Array Set) and expression analysis (Illumina HumanRefseq-8 Expression BeadChip) on a series of 193 neuropathologically normal human brain samples available at <http://labs.med.miami.edu/myers/LFuN/data%20ng.html>, as described in Section 3.3.4 above and reported by Myers and colleagues (Myers *et al.* 2007).

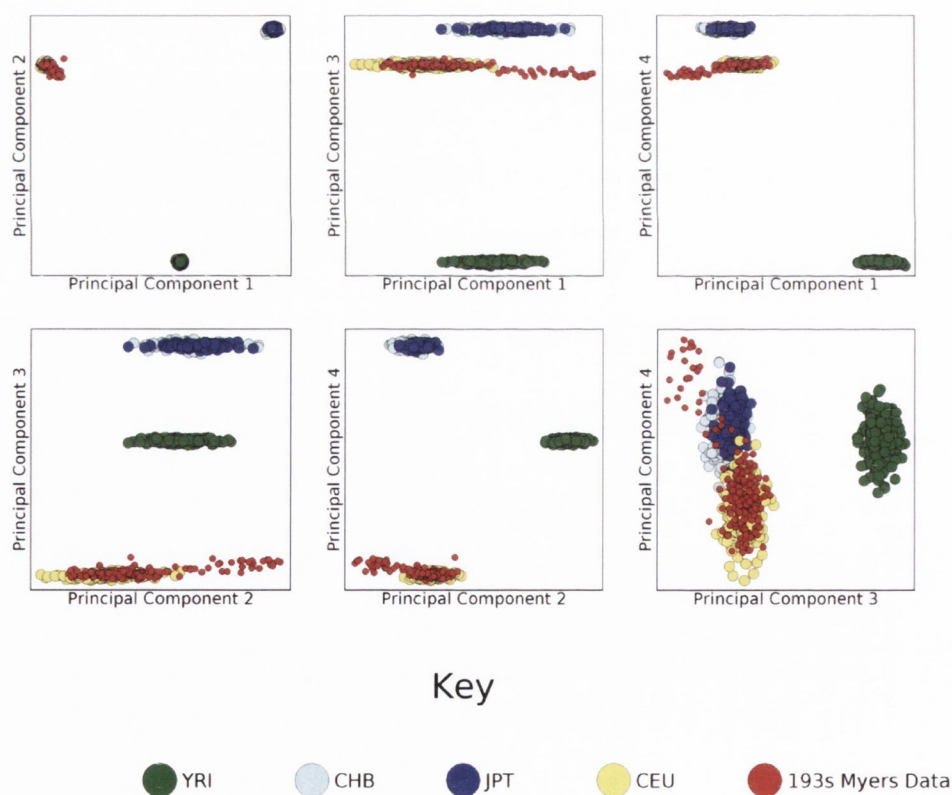


Figure 3.7: A multidimensional scaling matrix plot for 193 samples derived from brain tissue in the context of the HapMap populations. YRI: Yoruba in Ibadan, Nigeria; CHB: Han Chinese in Beijing, China; JPT: Japanese in Tokyo, Japan; CEU: Utah residents with Northern and Western European Ancestry. The 193s Myers data (in red) overlaps with European individuals from the CEU sample (in yellow), demonstrating this sample is indeed one sample of European descent.

To reduce bias due to population structure, covariates were generated based on population distance (Figure 3.7). Briefly, the dataset was LD pruned to generate a more LD independent dataset; LD pruning was performed using PLINK applying a 200 SNP 5-sliding window and an R^2 of 0.25. A derived dataset of 78519 autosomal markers was used to generate population-based covariates using PLINK.

Prior to analysis I confirmed that phenotypes were transformed to be normally distributed. All expression probes were mapped to gene names using Human Genome Nomenclature Committee (HGNC) standards, and thus any probe that failed to map to a gene name was removed. Association analyses, assuming an additive model, was performed for approximately 14,000 expression probes using linear regression analysis adjusting for 4-dimensions of population ancestry and sex. Data reduction clumping was performed within experiments (for a single probe); i.e. association findings were combined where a second marker is also associated but non-independent of the first signal.

3.3.5.1 Characterisation of *MACROD2*-related trans-eQTL in human control brains

In order to identify trans-eQTL related to *MACROD2*, GWA data was parsed to identify transcripts which are associated with the associated *MACROD2* region on chromosome 20. All data manipulation was performed using PLINK. The phenotype files were prepared using Stata 10 to parse out expression data for each expression probe by probe ID, which was also $\log_{10}$ transformed and all missing variables were re-annotated as ‘-9’ to enable PLINK to identify them as such. The covariate data was downloaded from the Myers web page: <http://labs.med.miami.edu/myers/LFuN/data%20ng.html>. This file contained the co-variables of gender, age at death, post-mortem interval, the transcript detection rate, hybridisation date, sample source and brain region (frontal, parietal and temporal); and was converted to PLINK format using Stata. The association analysis was then performed using the 193s Myers dataset, that had previously been imputed against 1000 genomes data as reference (as described in Section 5.4.7 in Chapter 5). The association analysis was run in triplicate using the --pheno routine in PLINK, using the imputed data for the imputation based on: 1) The most likely call, 2) A defined probability threshold of 0.6 and 3) Using the .dosage files from PLINK for the imputation. This resulted in a file reporting SNPs across the defined length of *MACROD2* in association with probes, indicated by GI (Gene Info identifier) accession numbers and reported a p-

value and standard error for each association. The threshold for imputation was set at any $p < 10^{-5}$, which was determined to be an appropriate threshold for interpretation due to the genome-wide scale of this analysis. Phenotypic quality control limited analyses to those probes expressed in >80% of individuals. Out of a starting dataset of 19,919 expression probes, 8349 passed QC.

3.3.5.2 Gene Enrichment Analysis of *MACROD2*-related trans-eQTL in human control brains using the Gene Ratio Test and DAVID

Using self-authored scripts I examined whether there is an enrichment of associated genes in populated GO-terms. This Gene Ratio Test (GRT) proceeded as follows:

1. Identified the total number of genes in the dataset that map to GO-terms
2. Identified the genes that are associated
3. Calculated the number of hits per term of the associated genes
4. Simulated gene-lists from the list of the total number of genes that are equal to the associated genes
5. Calculated the number of hits per term for each simulated list (Repeat for 10,000 permutations)
6. Calculated the simulated p-value equal to hits in the simulated set that are $\geq$ the observed
7. Additional limits were placed on the analyses:
 - (a) Minimum (& maximum) number of genes per term (Default: 20)
 - (b) Minimum number of association hits per term (Default: 3)

A secondary enrichment analysis was performed using DAVID (Database for Annotation, Visualisation and Integrated Discovery) v6.7, an online gene ontology database that enables functional annotation of genes, and grouping of these genes together by gene ontology term (Huang *et al.* 2009). Gene-ontology enrichment analysis can discover whether a list of identified genes is enriched for a particular function. The list of genes from the eQTL association analysis was input using the web form at <http://david.abcc.ncifcrf.gov/tools.jsp>. This returned a grouping of genes by functional annotation, and demonstrated any enrichment for particular biological processes.

3.4 Results

3.4.1 Characterisation of the Structure of *MACROD2*

The *MACROD2* locus on NCBI describes two main isoforms of *MACROD2*; *MACROD2* isoform 1 encodes a 17 exon transcript, which encompasses the whole coding region; *MACROD2* isoform 2 encodes a shorter 9 exon transcript that lacks exon 1 to 8 (<http://www.ncbi.nlm.nih.gov/gene/140733>). The tissue expression profile of the two isoforms is unknown and specifically whether there is differential or isoform specific expression of these transcripts within the human brain. The macro domain of the protein, which confers the gene its name and is the putative important functional domain of the protein, is encoded by exons 4 through to 6. The macro domain is an ADP-ribose binding module, and the ADP-ribosylation of proteins has been identified as an important post-translational modification with a role in multiple biological processes, including DNA repair, transcription, chromatin biology and long-term memory formation (Karras *et al.* 2005). The lack of this domain in the product of isoform 2 leads to the question as to whether the MACRO domain is the main functional site of the protein, and if so, what is the function of the *MACROD2* isoform 2.

3.4.1.1 Online Investigation of *MACROD2* Structure

RNAseq data enables us to examine the structure of a gene based on the sequence reads across the mRNA of the gene. I used Illumina's Human BodyMap 2.0 project online application (Accessible as additional data on the gene viewer at <http://www.ensembl.org>) to examine the RNAseq reads across *MACROD2*. Figure 3.8 is an example output of RNAseq data, across Exon 9 of *MACROD2*, which finds reads in support of isoform 1 being expressed in human brain but not isoform 2.

Chromosome 20: 15,843,267-15,843,594

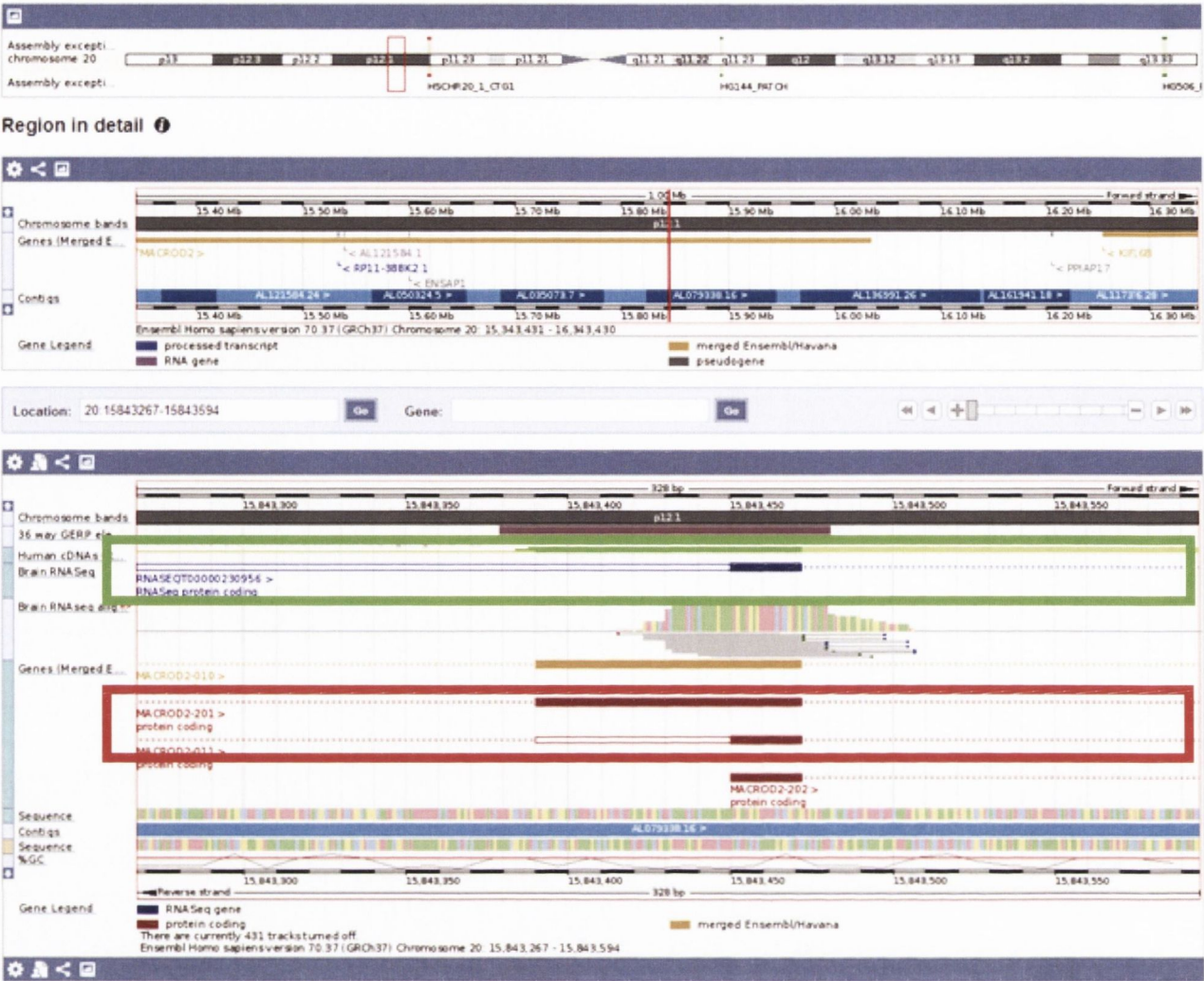


Figure 3.8 An example of the output of RNAseq data from Ensembl.org. This is a ‘zoomed in’ view of exon 9 of *MACROD2*, demonstrating RNAseq reads (highlighted in green) in support of isoform 1 (highlighted in red).

3.4.1.2 5’ Rapid Amplification of cDNA Ends

Two isoforms of *MACROD2* are reported in online databases, and I wished to confirm the existence and number of these isoforms, which may inform further study (such as whether isoform expression is tissue dependent). RefSeq (<http://www.ncbi.nlm.nih.gov/refseq/>) recorded two validated isoforms for *MACROD2*: *MACROD2* isoform 1 encodes the longer full 17 exon transcript, whereas isoform 2 is shorter than isoform 1, encoding a shorter transcript that lacks exon 1 to 8, creating a shorter N-terminus (<http://www.ncbi.nlm.nih.gov/gene/140733>). 5’ RACE experiments were therefore designed starting at exon 9, which would demonstrate which of the isoforms were expressed in the tissues assayed, human adult and foetal whole brain,

cerebellum and testis. 5' RACE using cDNA from brain tissues and testis created fragments of different lengths that were then assayed using DNA sequencing.

Following DNA sequencing of each fragment, the presence or absence of each exon of *MACROD2* within that fragment was recorded, using the reference sequence of *MACROD2* for comparison. As the primers for the 5' RACE were designed within exon 9, if any fragment was observed in exon 8 and backwards towards exon 1, it would indicate expression of the longer isoform 1. Table 3.6 and Figure 3.9 record the results of sequencing from 5' RACE cDNA fragments from different human tissues: whole brain of both adult and foetal origin, cerebellum and testis. Due to the presence of exons 2 to 8 in these fragments, the isoform expressed in human brain is likely the longer isoform 1, a transcript which contains these exons.

Table 3.6 RACE experiments identify expression of *MACROD2* isoform 1 but not isoform 2.

cDNA	Brain, whole (adult)	Brain, (foetal)	Cerebellum	Testis
<i>MACROD2</i> – Isoform1	Yes	Yes	Yes	Yes
<i>MACROD2</i> – Isoform 2	No	No	No	No

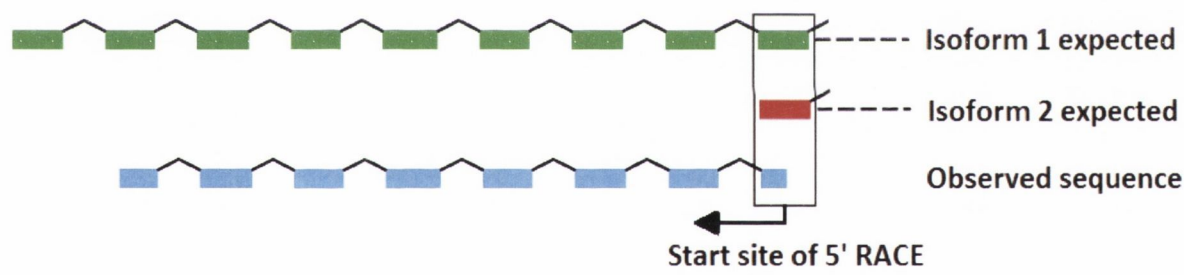


Figure 3.9 The observed sequence from the 5' RACE experiments indicates expression of isoform 1 but not isoform 2. Primers were designed at exon 9 of *MACROD2*, as this was the site where the isoforms were reported to deviate in the RefSeq database.

3.4.1.3 Promoter Mapping

In an effort to study the structure of the regulatory elements of *MACROD2*, primers were designed to amplify five fragments of steadily increasing size from the putative *MACROD2* promoter region directly 5' of the transcription start site sequence. PCR was used to amplify these fragments, which were visualised on an agarose gel, excised and purified. The fragments were then cloned into the pGL3 basic plasmid (Promega) to examine if they had an effect upon expression of the luciferase reporter gene, which is encoded on this plamid, when transfected into the SH-SY5Y cell line. This approach examines any promoter or enhancer/repressor effects and also gives an

indication of the location of such elements, and thus aids in building a picture of the gene structure in the region examined.

Differential expression of luciferase was observed across the five putative regulatory regions. From this data it would appear that a promoter or enhancer element of *MACROD2* lies ~400-500bp upstream from the transcription start site, within fragment 'MD2'. Further out from the transcription start site, at ~600-900bp, there appears to be a repressor element that is affecting the expression of luciferase, within fragments 'MD3', 'MD4' and 'MD5'.

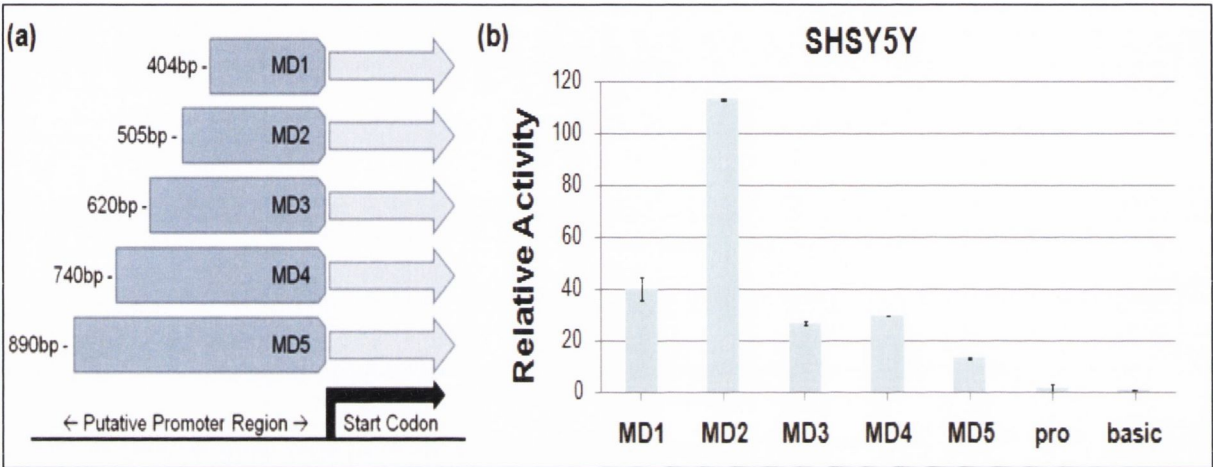


Figure 3.10 (a) Mapping of the putative *MACROD2* promoter region using inserts of increasing size (bp). **(b)** Differences in the activity of each putative promoter insert in a luciferase reporter assay indicate the location of regulatory elements upstream of *MACROD2*.

To assess the presence of putative transcription factor binding domains within the region specific to MD2 but not MD1 (promoter or enhancer elements) and MD3, MD4 and MD5 (repressor elements) was performed using Genomatix MatInspector online search tool. The results of this output are described in Table 3.7.

Table 3.7 Top hits from bioinformatic search of regulatory regions from the putative *MACROD2* promoter region: A search using MatInspector returned a list of transcription factor binding sites. Presented here are the sites that had a 100% sequence match for known transcription factor binding sites. The base pair positions are from the start codon of *MACROD2*. For the sequence panel, the core promoter sequence is highlighted in red.

Binding site family	From	To	Anchor	Strand	Sequence
SWI/SNF related nucleophosphoproteins with a RING finger DNA binding motif	180	190	185	(+)	a tg tACTT gtc
Pleomorphic adenoma gene	267	289	278	(-)	aa GGGGg caaaa tagcaa tgcta
SWI/SNF related nucleophosphoproteins with a RING finger DNA binding motif	420	430	425	(-)	atgt ACTT att
NKX homeodomain factors	567	585	576	(-)	ga cctc AAGT gatccaaa
PAX homeodomain bindingsites	828	842	835	(+)	a caaaa ATT Acctta
Pleomorphic adenoma gene	904	926	915	(-)	aa GGGGg ga ggccccaaaaagaa
Pleomorphic adenoma gene	332	354	343	(-)	ga GGGGg cggagcca ggga gcga
Pleomorphic adenoma gene	353	375	364	(-)	ga GGGGg cggagcca ggga gcga
Pleomorphic adenoma gene	311	333	322	(-)	ga GGGGg ctggagctgtctcga
GATA binding factors	64	76	70	(+)	a tca GATT a cctg
RNA polymerase II transcription factor II B	298	304	301	(+)	ccg CGCC

3.4.2 Characterisation of the Expression of *MACROD2*

The expression profile of *MACROD2* in humans is available to view on the online databases (Illumina BodyMap RNAseq data, ensemble.org and Serial Analysis of Gene Expression, <http://cgap.nci.nih.gov/SAGE/>) where it also displays a near ubiquitous expression pattern, including expression in the brain. However, each database shows a somewhat different expression profile and these profiles have not been experimentally validated, thus there is no confirmed expression profile of *MACROD2* when comparing these databases.

3.4.2.1 Expression profiling of *MACROD2*

Expression profiling of *MACROD2* reveals that the gene is almost ubiquitously expressed across human tissues, including brain, kidney, placenta, skeletal muscle, testis and thyroid gland. I did not observe expression in heart, adult liver, lung or thymus (Figure 3.11). Importantly, expression is observed in adult and foetal brain as well as in the SH-SY5Y neuronal cell-line. Genomic DNA, as a negative control for the cDNA primers was derived from lymphoblastoid cells. The expression profile of *MACROD2-AS1* was also investigated using this method, and was detected to be expressed in the brain and testis only.

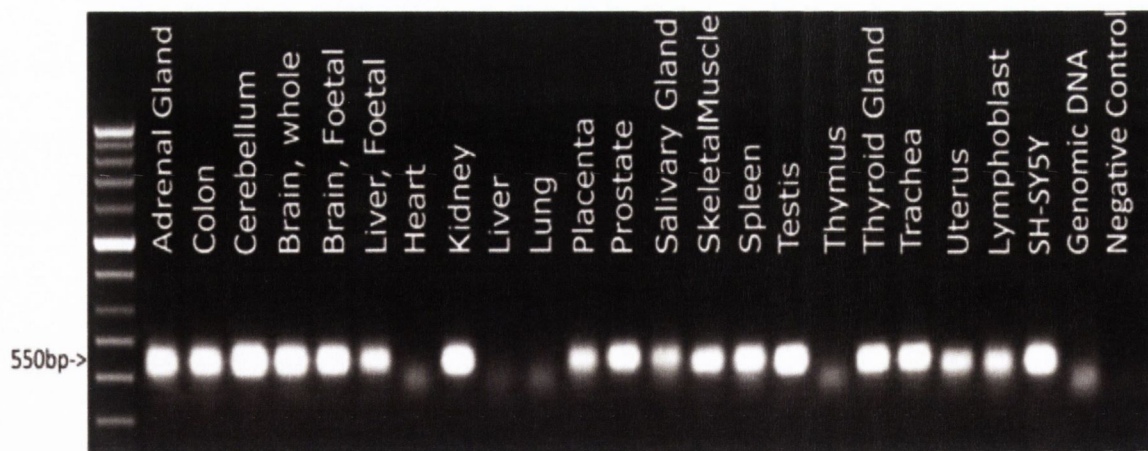


Figure 3.11 Exon-crossing primers were used to demonstrate the expression profile of *MACROD2* using RNA derived from 21 human tissues. The expected amplicon size was ~550 base pairs. Wherever a band is visible, this suggests that *MACROD2* is expressed in that tissue.

Database resources have recently been developed to describe tissue specific expression. RNAseq and SAGE databases document expression of *MACROD2* (Figure 3.12). These databases corroborate the results of the expression profiling of *MACROD2*, with expression of *MACROD2* recorded in the brain.

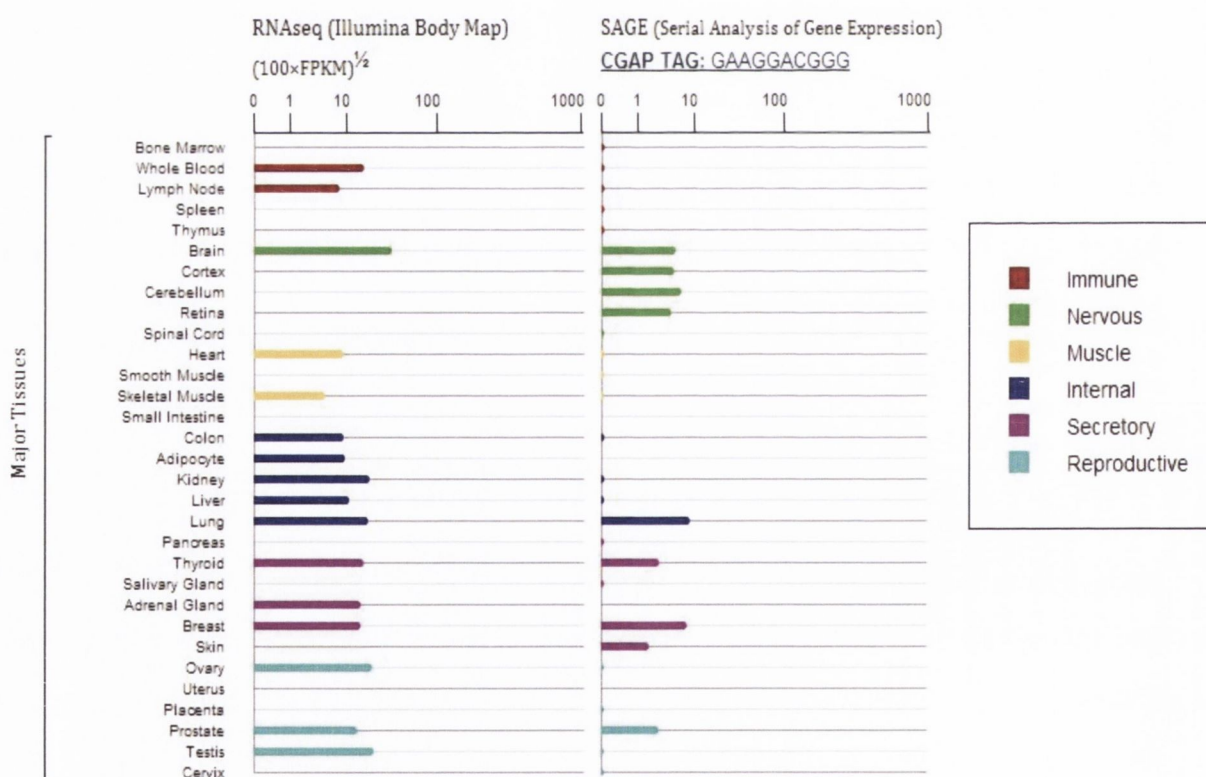


Figure 3.12 Expression profile of *MACROD2* across different bioinformatic resources. Recent RNAseq data has become available that lends support to the results of the expression profile experiments. The Illumina Body Map project and the Serial Analysis of Gene Expression databases document expression of *MACROD2* across the brain and nervous system.

3.4.3 Characterisation of the Role of Variation on the Expression of *MACROD2*

3.4.3.1 Allelic Expression Imbalance

An allele with balanced expression should have a relative expression value of approximately 0.5, indicating half of the expression is coming from that allele. A value greater or less than 0.5 indicates a difference in expression between the alleles. Given the variability and sensitivity of the genotyping assays it is accepted that AEI is observed if the relative expression skews to a level below 0.4 or greater than 0.6.

Figure 3.13 illustrates the allelic expression imbalance across the five SNPs assayed within *MACROD2*. Two markers showed strong evidence for AEI: For the assay using the marker SNP rs1225891, 19 out of a total 33 heterozygous individuals showed a positive result for AEI; for the assay using the marker SNP rs1238278, 17 of a total 19 heterozygous individuals were positive for AEI. For the remaining three marker SNPs assayed, although the sample size was small, there is suggestive evidence of AEI, with ~50% of the heterozygous individuals for each of these three SNPs positive for AEI.

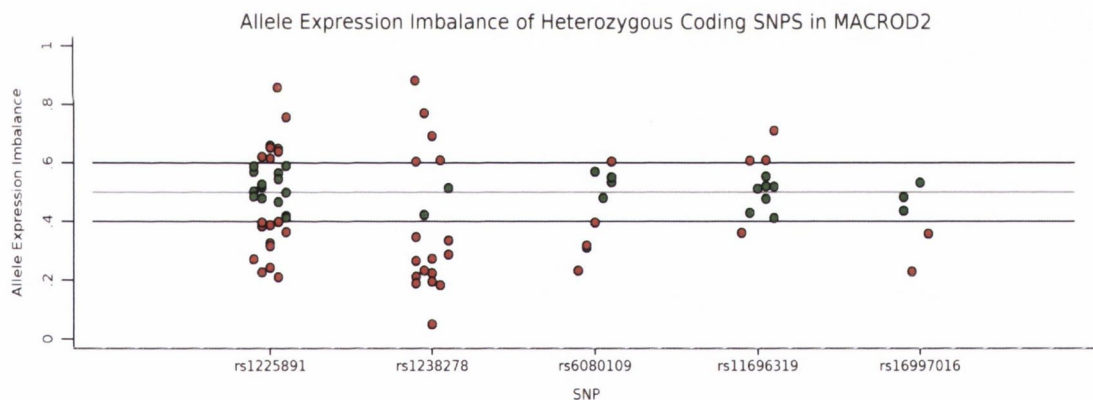


Figure 3.13 Allele Expression imbalance is observed for heterozygous coding SNPs in *MACROD2*. An allele with balanced expression should have a relative expression value of approximately 0.5, indicating half of the expression is coming from that allele. A value greater or less than 0.5 indicates a difference in expression between the alleles. Given the variability and sensitivity of the genotyping assays it is accepted that AEI is observed if the relative expression skews to a level below 0.4 or greater than 0.6. Thus the evidence for AEI at the *MACROD2* locus is strongest from the marker SNPs rs1225891 and rs1238278, which could indicate imprinting or RNA induced silencing of this gene resulting in this imbalance in expression.

3.4.3.2 Validation of Missense variants from the Exome sequencing study

The NPG group conducted an exome sequencing study of 147 autism and 273 schizophrenia cases across 215 putative susceptibility genes, of which *MACROD2* was included based on the data from the AGP and this work (Kenny, E. *et al* 2013, under

review). Several missense variants within *MACROD2* were recorded and were highlighted for this work for validation (see Table 3.8).

Primers were designed around each reported SNP and PCR was performed from the DNA of each proband, together with that of the parents, which would enable investigation of whether the SNPs were *de novo*. PCR products were sent to Source Bioscience for sequencing and the resulting DNA sequence was analysed using the BioEdit software. Each sequence read was compared against the reference sequence retrieved from UCSC genome browser. In total there were three missense SNPs checked for validation. From the sequencing results, all three SNPs were true SNPs, with no reported *de novo* SNPs.

Table 3.8 Validation of missense mutation reported in *MACROD2*. Missense mutations that had been identified by next generation sequencing of autism probands were validated by capillary sequencing. The probands were from a trio sample, which enabled sequencing of these variants in the parents to check if they were *de novo*. However all three missense variants were inherited from a parent.

Identifier	Position	Ref	Alt	Observed	Validated	Present in Father	Present in Mother
AS097C1	15967383	G	T	G/T	Yes	Yes	No
AS113C1	15967719	T	C	T/C	Yes	No	Yes
AS030C1	15967724	C	G	C/G	Yes	Yes	No

3.4.3.3 Characterisation of *MACROD2* CNV pattern

Investigation of the CNV pattern at the *MACROD2* locus was enabled by data coming from the AGP and DGV. The AGP dataset also includes detailed records of multiple autism related phenotypic scores for metrics such as IQ and age at first phrase. The AGP dataset is also of trio design, which enables comparison of CNV patterns between probands and parents to look for *de novo* deletions or duplications. Figure 3.14 illustrates the CNV pattern at the *MACROD2* locus for individuals from the AGP sample. The majority of CNVs in this sample are located in the intron between exon 5 and 6. However, there are two *de novo* CNVs present that cross exon 5. When the three individuals who carry these CNVs were cross referenced for severity of autism phenotype, they were found to have a more severe autism diagnosis, indicated by delayed or absent speech and low IQ. Figure 3.15 is an example of one of these *de novo* CNVs, indicating that neither parent has a deletion at this locus.

Chromosome 20:14300000-15300000

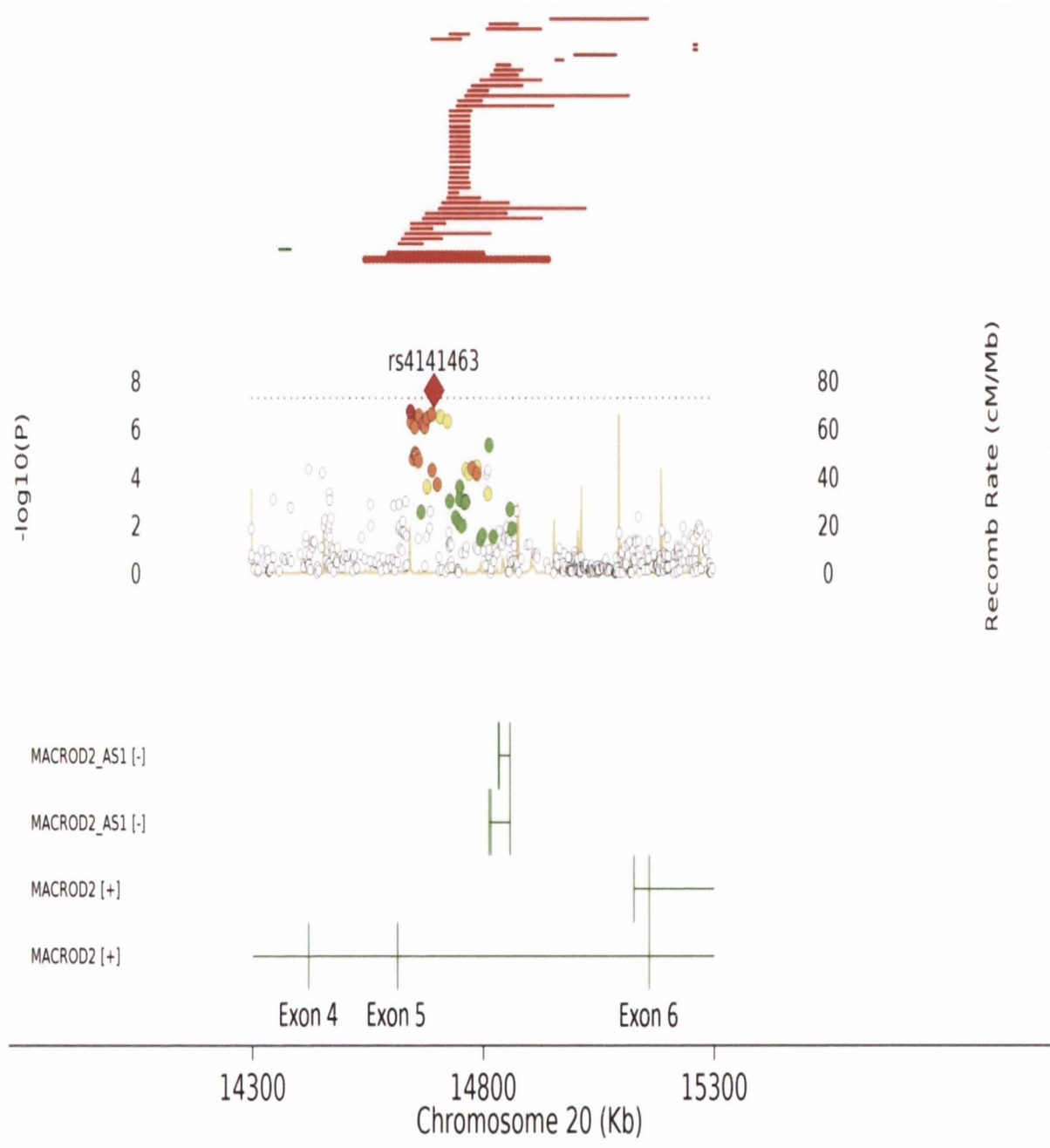


Figure 3.14 Pattern of copy number variation in the AGP probands. Top panel: CNV deletions from the AGP sample are indicated in red, with the thicker red lines denoting *de novo* deletions; Middle panel: The association plot of the top hit, rs4141463 from the AGP European strict sample; Bottom panel: A map of the MACROD2 gene, indicating the location of exon 5. The large CNV burden at this locus is clear, however the *de novo* deletions hit exon 5 while the other CNV deletions do not. Note that the most significant association from the AGP GWAS is adjacent to exon 5.

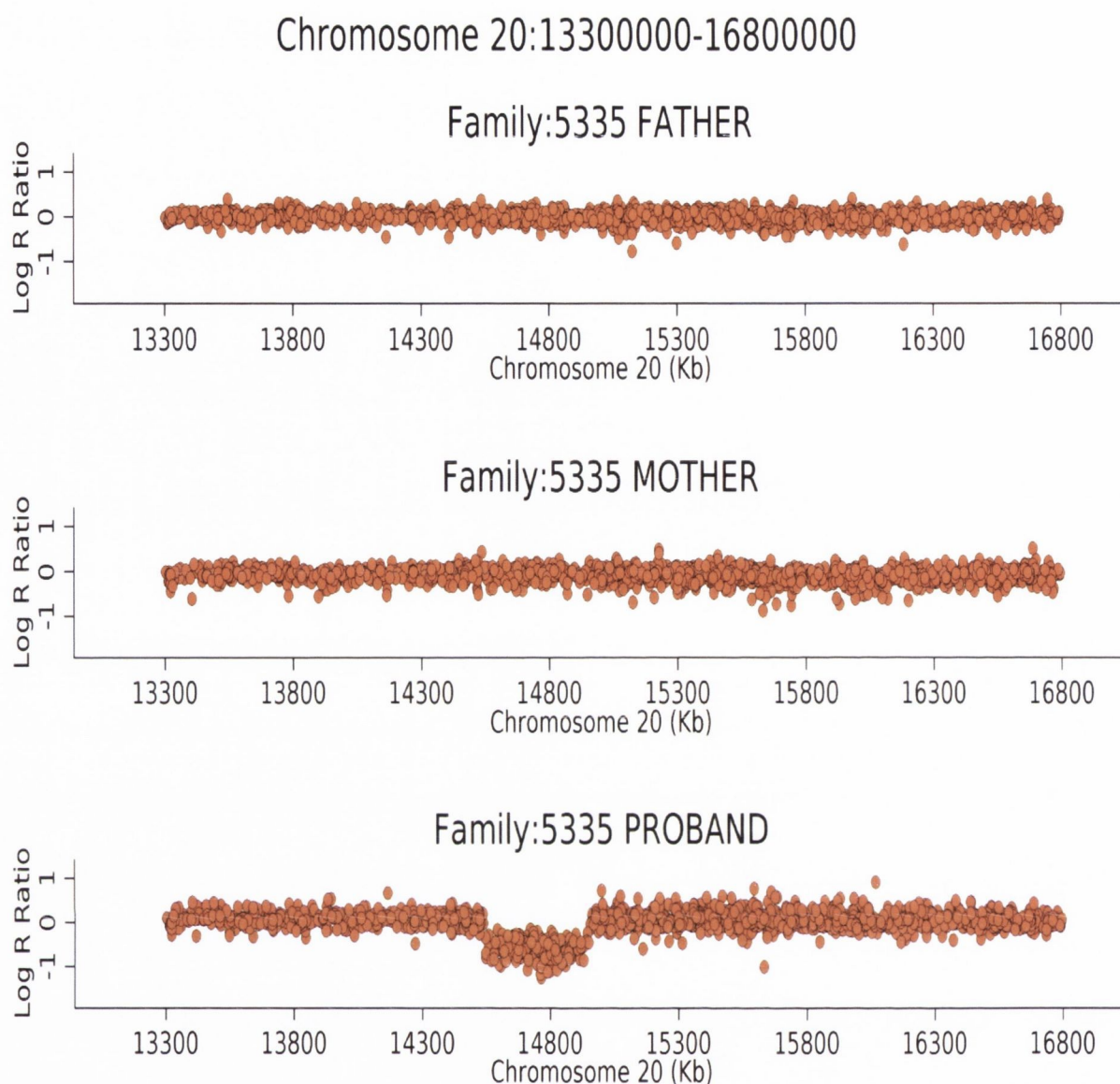


Figure 3.15: Example of a de novo CNV deletion in *MACROD2* from the AGP Collection. This is a genotype intensity plot, with each point on the plot indicating a different marker SNP. Any Log R Ratio value at 0 implies a normal copy number, whereas a value at 1 implies a duplication and a value at -1 implies a deletion. Here a deletion that crosses exon 5 in the proband is clearly visible that has not been inherited from the parents.

Turning again to Figure 3.14 and the recurrent CNVs located in the intron between exon 5 and 6 of *MACROD2*, I sought to investigate if this large burden of CNVs in this intronic space is present in other datasets, and thus I used the DGV to investigate CNV burden from control data, illustrated in Figure 3.16. The data from the DGV mirrors that from the AGP, with the vast majority of CNVs at this locus being intronic deletions, with a small number of CNVs that delete exonic material (exon 5).

Chromosome 20:14600000-14700000

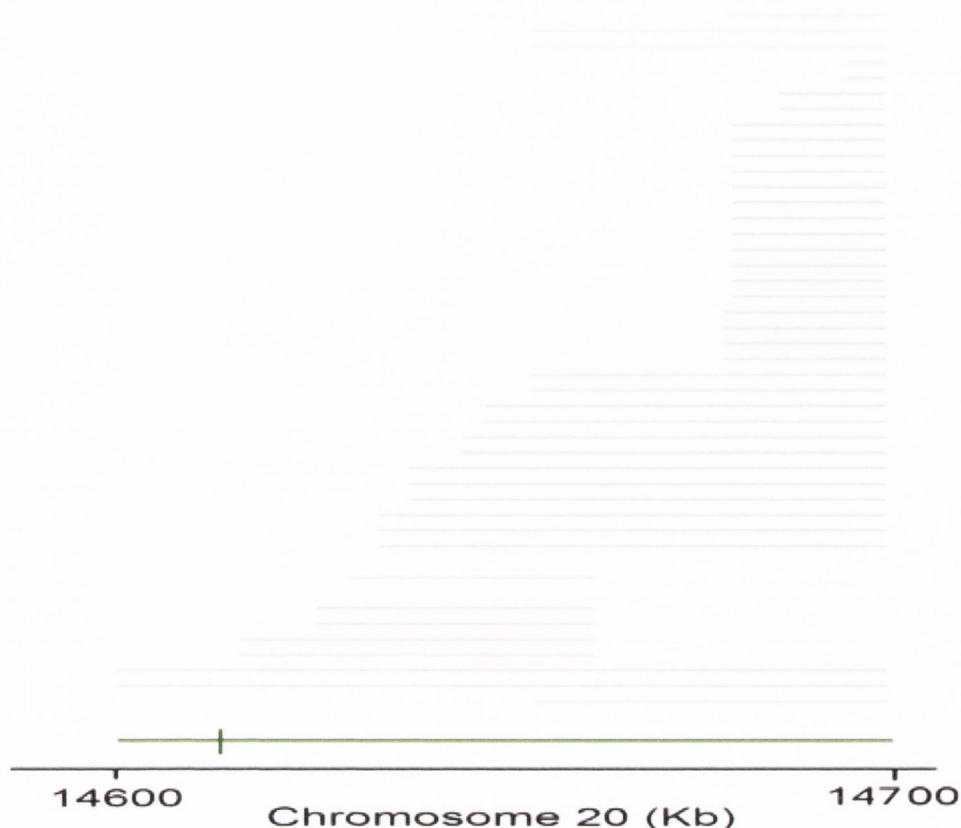


Figure 3.16 CNV from the DGV (Control) database overlapping or downstream of exon 5 of *MACROD2*. The DGV data displays the same pattern as the CNVs documented by the AGP, with exonic deletions at exon 5 being rare, whereas other intronic deletions at this locus are common. To date CNVs at this locus have been regarded as benign due to the high number of intronic deletions. Mapping of these deletions to elucidate possible functional impact is required, as mapping of the *MACROD2* locus demonstrates.

3.4.3.4 Characterisation of *MACROD2*-related trans-eQTL in human control brains

Association analyses of the human cortical gene expression dataset did not identify genome-wide significant association signals in this region. 54 out of 8349 probes revealed at least one SNP associated at $p < 10^{-5}$, which were then mapped to gene IDs (See Table 3.9).

Table 3.9 Top unique association signals from human cortical gene expression *MACROD2* trans eQTL analysis

GENEID	GENE	GENE_NAME_HGNC	GI_ACCESSION	SNP	A1	FRQ	INFO	BETA	SE	P
5834	PYGB	phosphorylase, glycogen; brain	GI_34577123	rs6135221	C	0.06	0.96	-0.14	0.025	8.05E-08
7917	BAG6	BCL2-associated athanogene 6	GI_18375629	rs6135221	C	0.06	0.96	-0.142	0.025	8.06E-08
25930	PTPN23	protein tyrosine phosphatase, non-receptor type 23	GI_24308072	rs6135221	C	0.06	0.96	-0.102	0.019	2.01E-07
10399	GNB2L1	guanine nucleotide binding protein (G protein), beta polypeptide 2-like 1	GI_24475893	rs6135221	C	0.06	0.96	-0.188	0.035	2.77E-07
9150	CTDP1	CTD (carboxy-terminal domain, RNA polymerase II, polypeptide A) phosphatase, subunit 1	GI_16554588	rs10485773	G	0.11	0.57	-0.116	0.022	3.70E-07
26168	SEN3	SUMO1/sentrin/SMT3 specific peptidase 3	GI_21361498	rs6135221	C	0.06	0.96	-0.104	0.02	4.71E-07
6749	SSRP1	structure specific recognition protein 1	GI_28416943	rs6135221	C	0.06	0.96	-0.108	0.021	5.05E-07
10594	PRPF8	PRP8 pre-mRNA processing factor 8 homolog (S. cerevisiae)	GI_17999536	rs6135221	C	0.06	0.96	-0.097	0.019	6.69E-07
51650	MRPS33	mitochondrial ribosomal protein S33	GI_16950596	rs1324394	C	0.16	0.98	0.0731	0.014	6.98E-07
23352	UBR4	ubiquitin protein ligase E3 component n-recogin 4	GI_24416001	rs6135221	C	0.06	0.96	-0.136	0.026	7.47E-07
5034	P4HB	prolyl 4-hydroxylase, beta polypeptide	GI_20070124	rs6135221	C	0.06	0.96	-0.149	0.029	1.05E-06
7415	VCP	valosin containing protein	GI_7669552	rs6135221	C	0.06	0.96	-0.144	0.029	1.17E-06
26528	DAZAP1	DAZ associated protein 1	GI_25470885	rs10485773	G	0.11	0.57	-0.126	0.025	1.25E-06
6829	SUPT5H	suppressor of Ty 5 homolog (S. cerevisiae)	GI_20149523	rs6135221	C	0.06	0.96	-0.138	0.028	1.42E-06
8666	EIF3G	eukaryotic translation initiation factor 3, subunit G	GI_4503516	rs6135221	C	0.06	0.96	-0.125	0.025	1.47E-06
7386	UQCRCF1	ubiquinol-cytochrome c reductase, Rieske iron-sulfur polypeptide 1	GI_5174742	rs16995184	A	0.16	0.94	0.0934	0.019	1.64E-06
83590	TMUB1	transmembrane and ubiquitin-like domain containing 1	GI_31543199	rs6135221	C	0.06	0.96	-0.127	0.026	1.84E-06
994	CDC25B	cell division cycle 25 homolog B (S. pombe)	GI_11641410	rs10485773	G	0.11	0.57	-0.137	0.028	1.87E-06
10844	TUBGCP2	tubulin, gamma complex associated protein 2	GI_5729839	rs6135221	C	0.06	0.96	-0.135	0.027	2.05E-06
114049	WBSCR22	Williams Beuren syndrome chromosome region 22	GI_23199994	rs6135221	C	0.06	0.96	-0.097	0.02	2.19E-06
23193	GANAB	glucosidase, alpha; neutral AB	GI_38202256	rs6135221	C	0.06	0.96	-0.094	0.019	2.41E-06
10781	ZNF266	zinc finger protein 266	GI_37622342	rs6042970	G	0.19	0.94	0.0718	0.015	2.44E-06
81605	URM1	ubiquitin related modifier 1	GI_13569869	rs6135221	C	0.06	0.96	-0.121	0.025	2.44E-06
2879	GPX4	glutathione peroxidase 4	GI_4504106	rs6135221	C	0.06	0.96	-0.155	0.032	2.48E-06
4832	NME3	NME/NM23 nucleoside diphosphate kinase 3	GI_37693992	rs10485773	G	0.11	0.57	-0.131	0.027	2.70E-06

56270	WDR45L	WDR45-like	GI_19923554	rs1040740	C	0.93	0.87	0.1213	0.025	2.78E-06
8818	DPM2	dolichyl-phosphate mannosyltransferase polypeptide 2, regulatory subunit	GI_24497593	rs6135221	C	0.06	0.96	-0.126	0.026	3.13E-06
10938	EHD1	EH-domain containing 1	GI_30240931	rs10485773	G	0.11	0.57	-0.131	0.027	3.14E-06
1716	DGUOK	deoxyguanosine kinase	GI_18426962	rs4814320	G	0.14	0.94	0.0524	0.011	3.26E-06
9016	SLC25A14	solute carrier family 25 (mitochondrial carrier, brain), member 14	GI_13259542	rs1040740	C	0.93	0.87	-0.095	0.02	3.54E-06
51150	SDF4	stromal cell derived factor 4	GI_18699731	rs2423873	G	0.22	0.70	-0.084	0.018	3.54E-06
375260	WASH2P	WAS protein family homolog 2 pseudogene	GI_39573729	rs6135221	C	0.06	0.96	-0.151	0.032	3.91E-06
51167	CYB5R4	cytochrome b5 reductase 4	GI_21314659	rs4814320	G	0.15	0.92	-0.11	0.023	3.98E-06
84256	FLYWCH1	FLYWCH-type zinc finger 1	GI_14150048	rs6135221	C	0.06	0.96	-0.125	0.026	4.07E-06
4176	MCM7	minichromosome maintenance complex component 7	GI_33469967	rs10485773	G	0.11	0.57	-0.197	0.042	4.68E-06
152926	PPM1K	protein phosphatase, Mg2+/Mn2+ dependent, 1K	GI_31542532	rs6074801	A	0.73	0.96	0.1227	0.026	4.87E-06
83862	TMEM120A	transmembrane protein 120A	GI_13994299	rs6135221	C	0.06	0.96	-0.136	0.029	5.09E-06
26502	NARF	nuclear prelamin A recognition factor	GI_14165459	rs6135221	C	0.06	0.96	-0.111	0.024	5.41E-06
348793	WDR53	WD repeat domain 53	GI_32699063	rs2423804	A	0.45	0.97	-0.075	0.016	5.87E-06
84152	PPP1R1B	protein phosphatase 1, regulatory (inhibitor) subunit 1B	GI_31415879	rs6135221	C	0.06	0.96	-0.166	0.036	6.01E-06
6124	RPL4	ribosomal protein L4	GI_16579884	rs6135221	C	0.06	0.96	-0.107	0.023	6.38E-06
140735	DYNLL2	dynein, light chain, LC8-type 2	GI_18087854	rs16995184	A	0.15	0.89	0.1002	0.021	7.22E-06
10572	SIVA1	SIVA1, apoptosis-inducing factor	GI_11277469	rs6135221	C	0.06	0.96	-0.11	0.024	7.22E-06
10726	NUDC	nuclear distribution C homolog (A. nidulans)	GI_31543300	rs6135221	C	0.06	0.96	-0.127	0.028	7.61E-06
57827	C6orf47	chromosome 6 open reading frame 47	GI_10863984	rs10485773	G	0.11	0.57	-0.103	0.022	7.99E-06
9524	TECR	trans-2,3-enoyl-CoA reductase	GI_4759061	rs6135221	C	0.06	0.96	-0.134	0.029	8.13E-06
7171	TPM4	tropomyosin 4	GI_4507650	rs1040740	C	0.93	0.87	0.1683	0.037	8.13E-06
22909	FAN1	FANCD2/FANCI-associated nuclease 1	GI_7662449	rs6135298	A	0.18	0.99	0.0416	0.009	8.26E-06
7376	NR1H2	nuclear receptor subfamily 1, group H, member 2	GI_11321629	rs1040740	C	0.93	0.87	0.1168	0.025	8.60E-06
896	CCND3	cyclin D3	GI_16950657	rs6135221	C	0.06	0.96	-0.15	0.033	8.72E-06
11253	MAN1B1	mannosidase, alpha, class 1B, member 1	GI_6005807	rs6135221	C	0.06	0.96	-0.121	0.026	8.91E-06
5257	PHKB	phosphorylase kinase, beta	GI_4505782	rs1324394	C	0.16	0.98	-0.09	0.02	9.60E-06
9479	MAPK8IP1	mitogen-activated protein kinase 8 interacting protein 1	GI_20986517	rs6135221	C	0.06	0.96	-0.142	0.031	9.86E-06

3.4.3.5 Gene Enrichment Analysis of *MACROD2*-related trans-eQTL in human control brains

Gene Enrichment Analyses of the associated markers using the Gene Ratio Test did not identify GO-terms that show significant enrichment after correcting for multiple comparisons i.e. Bonferroni correction determined a significant threshold of 3.7×10^{-5} (Table 3.10). The secondary enrichment analysis using DAVID used 51 of the 54 most significant genes from the eQTL analysis for use in the gene ontology enrichment analysis. DAVID reports a p-value for the significance of the enrichment. A list of gene ontology terms with an uncorrected p-value <0.05 is reported in Table 3.11. The most significant gene ontology terms are related to the biological processes of acetylation, post-translational and other such modification of proteins and cytoplasmic and nuclear localisation terms. Acetylation is the only GO term that remains significant following Bonferroni correction, at $p = 0.0015$.

Table 3.410 Analysis of genes identified from the eQTL analysis were analysed using the Gene Ratio Test. The most significant Gene Ontology terms are reported here, although none reached the significance threshold of 3.7×10^{-5} as determined by Bonferroni correction. Count refers to the number of genes from the input that are associated with a given GO term. Bonferroni refers to the corrected p-value based on the number of multiple comparisons.

Gene Ontology Term	Simulated P
negative regulation of cellular protein metabolic process	0.0015
vesicle-mediated transport	0.0019
negative regulation of protein metabolic process	0.0024
establishment of localization in cell	0.0026
protein binding	0.0033
phosphoric ester hydrolase activity	0.0054
cytoplasmic membrane-bounded vesicle	0.0067
membrane-bounded vesicle	0.0069
multicellular organismal reproductive process	0.0096
cytoplasmic vesicle	0.0111
identical protein binding	0.0111
vesicle	0.0118
establishment of localization	0.0122
phosphorus metabolic process	0.0126
phosphate-containing compound metabolic process	0.0126
reproductive process	0.0189
cell projection	0.0191
regulation of cellular protein metabolic process	0.0204
intracellular membrane-bounded organelle	0.0244
membrane-bounded organelle	0.0245
transport	0.0274
transferase activity, transferring phosphorus-containing groups	0.029
neuron projection	0.0313
cellular component assembly at cellular level	0.0321
regulation of protein metabolic process	0.0324
hydrolase activity	0.033
apoptotic process	0.0354
interspecies interaction between organisms	0.0378
programmed cell death	0.0386
intracellular organelle	0.0392
phosphotransferase activity, alcohol group as acceptor	0.0397
receptor binding	0.0399
organelle	0.0405
intracellular transport	0.0457
kinase activity	0.0463

Table 3.411 Analysis of genes identified from the eQTL analysis were analysed using DAVID. The most significant Gene Ontology terms are reported here. Count refers to the number of genes from the input that are associated with a given GO term. Bonferroni refers to the corrected p-value based on the number of multiple comparisons. A single Gene Ontology term remains significant after Bonferroni correction - Acetylation. This result is in support of the identification of *MACROD2* as an O-Acetyl-ADP-Ribose deacetylase (as discussed in Chapter 1) and could indicate co-expression of *MACROD2* with other acetylases.

Gene Ontology Term	Count	P-Value	Bonferroni
acetylation	20	1.13E-05	0.001573324
GO:0070013~intracellular organelle lumen	15	8.59E-04	0.103341155
GO:0043233~organelle lumen	15	0.00108	0.12820987
GO:0031974~membrane-enclosed lumen	15	0.001313	0.153636745
GO:0005829~cytosol	12	0.002469	0.269443218
protein phosphatase	4	0.00443	0.460505896
phosphoprotein	29	0.004507	0.466259875
GO:0006470~protein amino acid dephosphorylation	4	0.006558	0.963458214
cytoplasm	17	0.006683	0.606261742
GO:0016311~dephosphorylation	4	0.009797	0.992931048
hsa00510:N-Glycan biosynthesis	3	0.011172	0.301982954
GO:0006091~generation of precursor metabolites and energy	5	0.01223	0.997949347
mutagenesis site	12	0.012892	0.943170612
GO:0004721~phosphoprotein phosphatase activity	4	0.013178	0.893750856
GO:0031981~nuclear lumen	11	0.014244	0.838297693
nucleus	19	0.015105	0.879438963
GO:0005654~nucleoplasm	8	0.020815	0.930843806
endoplasmic reticulum	6	0.034377	0.992268973
compositionally biased region:Asp/Glu-rich (acidic)	3	0.03457	0.999579915
domain:CS	2	0.035312	0.999645563
GO:0016791~phosphatase activity	4	0.038379	0.998658279
IPR007052:CS domain	2	0.045138	0.997750151
IPR017447:CS	2	0.045138	0.997750151
glycogen metabolism	2	0.054599	0.999592013

3.5 Discussion

3.5.1 Characterisation of the Structure of *MACROD2*

The observation from the literature that there were two reported isoforms of *MACROD2*, together with the observation from my own experiments that *MACROD2* was expressed across multiple human tissues, led me to investigate whether there was a brain specific isoform of *MACROD2*. I also wanted to investigate the putative regulatory region of *MACROD2*, as identification of regulatory regions here would inform further experimentation, for example, the impact of genetic variation across regulatory elements upon gene expression.

Using publicly available RNA-seq data from Illumina's Human BodyMap 2.0 project, it appears that isoform 1 is the dominant transcript across all tissues assayed, with sequence reads present across the full 17 exon length of *MACROD2*.

Data from my direct interrogation of the cDNA species using 5'RACE also suggests that expression of *MACROD2* in the human brain is from the longer full length transcript, isoform 1, and that there is no difference in isoform expression between whole brain, foetal brain, cerebellum and testis. Expression in the non-brain cDNA would also indicate that expression of isoform 1 is not brain specific. It would be of interest to expand these analyses across the available library of 21 human tissues to reveal the full extent of different splice patterns across tissues. However, time limitations and a focus to determine the profile in brain precluded these analyses.

5' RACE, whilst a simple technique in principle, was dependent here upon the success of the ligation of the RACE product into the T-Easy plasmid, which proved inconsistent in terms of transformation success. This experiment was also limited by sequencing quality. As capillary sequencing reads are limited to ~500-600 base pairs of good quality sequence, it was not possible to read the full length of *MACROD2* from exon 9, where the RACE reaction began, out to the full length of the transcript. However, some reads were very clear, with sequence reads from exon 9 back to exon 2. Future experiments could compensate for this by designing primers closer to exon 1 in an effort to sequence the full length transcript.

3.5.1.1 Promoter Mapping

Promoter mapping is a useful tool in the search for regulatory elements surrounding a candidate gene, and can highlight interesting loci for further analysis. Promoter mapping does not only search for promoter elements, but in fact can also reveal enhancer and repressor elements. From the results reported here, it appears that a promoter and possibly an enhancer element is present ~500bp from the start codon of *MACROD2*, whereas relative expression is seen to fall greatly as the inserts increase in size above 500bp. This large drop in relative expression could be indicative of a repressor element at this locus. An unusual observation about these results is the fact that relative expression from the *MACROD2* inserts is many times greater than the relative expression from the control pGL3 'basic' vector. The data presented here was calculated using twenty-seven individual replicates for each experimental and control plasmid, and thus I am confident in the accuracy of the measurements of relative expression. However, as this experiment was conducted using cell lines that have been immortalised, these results carry a caveat that these cells may not correctly model the expression of genes *in vivo* due to the genetic changes that may have been introduced into these cells.

3.5.2 Characterisation of the Expression of *MACROD2*

Due to the lack of information about the expression of genes at the *MACROD2* locus it was necessary to perform expression profiling as a preliminary investigation to determine if these genes were indeed expressed within the human brain. My focus on the brain was due to the neuropsychiatric focus of this project, I determined that it would be more informative if I could source RNA from specific brain regions, e.g. amygdala, hippocampus, septum, mammillary bodies and cerebellum. However, it proved difficult to source this material commercially and thus the best available material that was also cost effective was the Human Total RNA Master Panel II (Clontech). This panel was chosen as although it did not have RNA sourced from specific brain regions bar cerebellum, it did have RNA sourced from foetal brain which would give an indication as to whether the gene(s) under investigation were expressed in the developing as well as the adult brain.

3.5.2.1 Expression profiling of *MACROD2*

The expression profile of *MACROD2* was found to be near ubiquitous across the tissues assayed, and demonstrated that *MACROD2* is indeed expressed in both the adult

and developing human brain. I also demonstrated that *MACROD2* was expressed in the SH-SY5Y cell line, which at the time of experiment I was investigating whether that would be a suitable cell line for use in further investigation of *MACROD2*. Bioinformatic searches for *MACROD2* at the time of this experiment did not reveal a full expression pattern for *MACROD2*, and there were no experimentally validated results. More recently however, there exist full expression profiles for *MACROD2* from RNAseq and SAGE analysis, as described in Figure 3.12. The expression profile differs across these two databases and across my own results; this could be due to the observation (from the online database of Protein Abundance Across Organisms: <http://pax-db.org>) that *MACROD2* is not a highly expressed protein in most tissues. Interesting to note is that *MACROD2-AS1* expression was only observed in the brain and testis, an observation that is mirrored in RNAseq and SAGE data. Further study could discover whether *MACROD2-AS1* has a functional role in the brain.

It would have been advantageous for this analysis to source RNA from specific brain regions, as very detailed expression profiles for *MACROD2* in the murine brain exist (Maas *et al.* 2007) and it would have been interesting to demonstrate if the expression profile of *MACROD2* in the human brain mirrors that of the mouse.

As autism is thought of as a neurodevelopmental disorder, it would also have been interesting to investigate the expression of *MACROD2* across early development; however I could not source RNA from human foetal brains across different stages of development. However the online transcriptome database, The Developing Human Brain (<http://www.developinghumanbrain.org/home.html>) presents data to suggest that *MACROD2* expression levels are greatest during foetal development, in particular in the caudate nucleus and nucleus accumbens. So while the tissue localisation of the protein message has been difficult to detect within the scope of this project, as the group does not have direct access to RNA from specific brain regions, there is some evidence to suggest that *MACROD2* is highly expressed in specific regions of the human brain during development.

3.5.3 Characterisation of the Role of Variation on the Expression of *MACROD2*

3.5.3.1 Allelic Expression Imbalance

A key component to understanding how common variants contribute to disease pathophysiology is by understanding how these variants affect normal gene function. Thus AEI experiments were performed to investigate whether there were *cis*-acting variants affecting expression of *MACROD2*. AEI experiments were carried out for 5 independent SNPs within the 3' UTR of *MACROD2*. Unfortunately due to the low minor allele frequency of SNPs at the *MACROD2* locus, it meant that whilst there were documented exonic, missense and intronic variants recorded in dbSNP, these SNPs were not suitable for AEI in the HapMap CEU sample pool available. Of the 5 SNPs assayed, all 5 showed evidence of AEI, with rs1225891 and rs1238278 showing the most convincing evidence of AEI due to the increased sample size compared to the three other SNPs. When annotation of the variance at this region improves hopefully the variant that is driving this difference in allele expression can be discovered.

3.5.3.2 Validation of missense variants from the Exome sequencing study

I did not observe any de novo mutations in the *MACROD2* gene. Also of note is that *MACROD2* has not been highlighted in exome sequencing studies of autism cases (O'Roak, Vives, Fu, et al. 2012; Sanders et al. 2012; Neale et al. 2012; Iossifov et al. 2012), but as seems to be the case to date with autism, the sample sizes remain too small to discover associations with enough confidence.

3.5.3.3 Characterisation of *MACROD2* CNV pattern

The *MACROD2* locus has previously been identified as a hotspot for large (~0.5 Mb) CNV deletions (Bradley *et al.* 2010) but I wished to characterise this variation in greater detail. Deletions, and not duplication, were of more interest for this analysis due to the observation that deletions are more likely to have a functional impact (Pinto *et al.* 2010). Bradley and colleagues noted that these deletions occur at this locus up to 100 times more frequently than in the majority of the human genome, and thus this locus was named as a deletion hotspot. The authors note several interesting findings. The first being that of the 13 genomic regions identified as regions with a high frequency of rare deletions in 4 population samples, nine of these regions contain only one gene, of which 7 (including *MACROD2*) rank within the 35 longest genes in the human genome. The

authors note that the length of the deletions they record roughly correspond with the length of chromatin loops, which they posit may indicate a mechanism for this high frequency of deletions at these sites, and that they typically reside within very long introns. However, the authors also note that there is no abundance of sequence motifs or functional elements known to be involved in chromosomal instability, or even proximity to fragile sites. The *MACROD2* locus is of particular interest as it was shown to host the most independent deletions across the 4 study groups examined (ADHD, Schizophrenia, Longevity and Endometriosis), and also host to the most *de novo* deletions, which seems to suggest fragility or some underlying mechanism causing such deletions at this locus. It is possible that large deletions at this locus are simply tolerated with no effect to overall organism fitness and those deletions that affect exons are rarer. The large number of intronic versus exonic deletions at this locus observed in the DGV again lends credence to this idea of the *MACROD2* locus being a hotspot for deletions. It is also possible that deletion on one allele may expose a variant on the other allele.

Deletions at the *MACROD2* locus have also been identified in tumour-derived cell lines and primary tumours, which supports this idea of this site as a deletion hotspot further still (Davison *et al.* 2005). If the *MACROD2* locus is indeed a deletion ‘hotspot’, what does this mean for the associations made between CNVs at this locus and disease? Bradley uses the example of *NRXN1* and autism to explain that we need more sensitivity in our analysis of CNVs. While deletions in *NRXN1* were associated with ASD (Kim *et al.* 2008) a larger family based study that reported deletions at *NRXN1* found that these deletions did not segregate with autism (Weiss *et al.* 2008). However, Bradley and colleagues explain that a close assessment of exon dosage is required before one can rule out an association at this locus, due to the observation from their own data that CNVs that affected exons within *NRXN1* were only found within the schizophrenia and ADHD samples, and not the endometriosis and longevity samples. Thus not all CNVs are created equal, we should examine whether CNVs hit intronic or exonic sequence when making decisions about the possible functional impact of CNVs.

It can be argued that this same increase in sensitivity is required when we look at CNVs at *MACROD2*, due in part to the observation of this locus as a deletion hotspot. Figure 3.14 illustrates the distribution of CNVs at the *MACROD2* locus from the AGP probands, which clearly shows an abundance of CNVs between exon 5 and 6. However, of

the de novo CNVs at the locus, these CNVs cross exon 5, an important exon in terms of the *MACROD2* protein function as it encodes the central part of the macro domain (The function of *MACROD2* is discussed in greater detail in Chapter 4). Examination of the DGV control data also shows most of the CNVs in the region overlap with intron 5 (between exon 5 and 6, see Figure 3.17 below) whereas only a small number overlap exon 5.

Chromosome 20:14500000-15000000

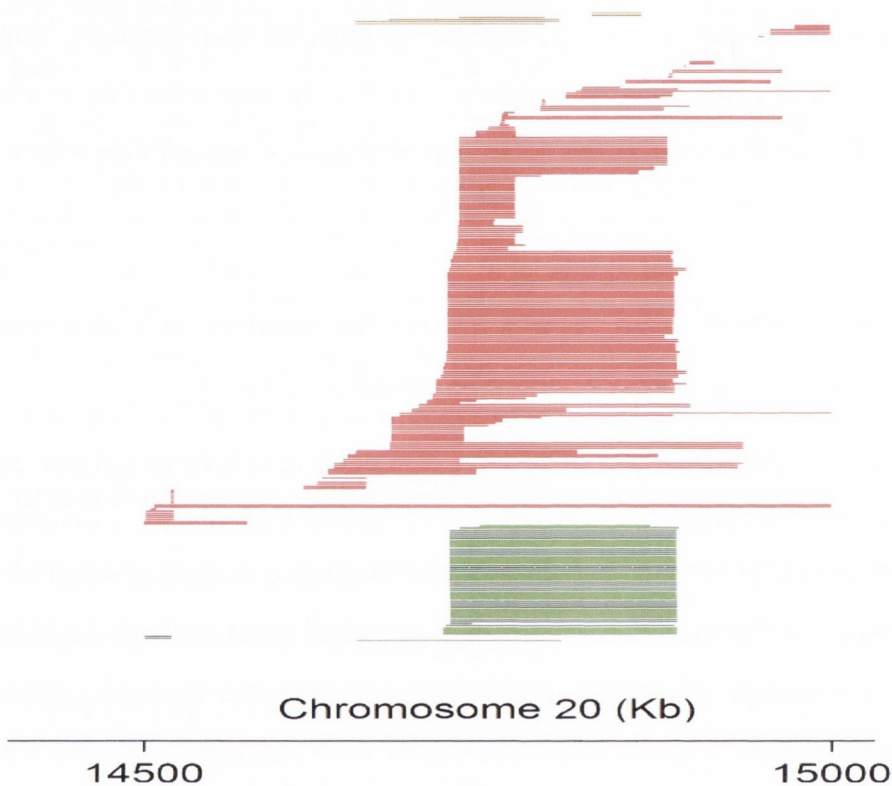


Figure 3.17 CNV at the *MACROD2* locus from the DGV. Note that the current resolution of image equates that each band is ~5 CNVs. Visualizing the CNVs at this locus in this way illustrates the idea of this locus being a deletion hotspot.

When the phenotypic severity of AGP individuals with de novo deletions crossing exon 5 was examined it was found that these individuals had a more severe autism diagnosis with high ADOS severity score. There is evidence from other publications that CNVs that affect the *MACROD2* coding sequence have a role in neurological conditions. An Array-CGH analysis of 111 patients with holoprosencephaly (HPE) identified CNVs within *MACROD2* that cross exons, in particular exon 5 (Bendavid *et al.* 2009). A larger analysis identified similar deletions of exons at *MACROD2*, against a backdrop of mutations in other HPE associated genes (Mercier *et al.* 2011). A *de novo* deletion of exon 5 of *MACROD2* has also been associated with the neurodevelopmental disorder Kabuki syndrome (Maas *et al.* 2007). Recent reports have demonstrated that exon 5 encodes

core conserved amino acids within the macro domain, which was demonstrated to be the active site of the protein (Jankevicius *et al.* 2013; Rosenthal *et al.* 2013).

MACROD2 does harbour many recurrent CNVs, however, the supposition that incidence of CNVs per se at *MACROD2* are not clinically relevant is too broad a view. A closer look at the breakdown of CNVs at this locus reveals that whilst it is a common site of deletion in the genome, it appears that deletions that affect exons, and thus are more likely to have a functional impact, are rare. The observation of segregation of *de novo* CNVs at exon 5 of *MACROD2* with severity of autism diagnosis, whilst anecdotal due to the sample size, does merit further investigation. It should also be noted that some of the CNVs at the *MACROD2* locus from the AGP sample, including some *de novo* CNVs, also cross *MACROD2-AS1*. If *MACROD2-AS1* is truly a regulatory element for *MACROD2*, it is unclear if CNVs that hit this anti-sense RNA have an impact upon expression of *MACROD2*. In fact there has been little investigation into the effect of variation upon expression of *MACROD2* overall.

3.5.3.4 Characterisation of *MACROD2*-related trans-eQTL in human control brains

I attempted to identify cortically expressed genes using the publically available Myers brain expression dataset (Myers *et al.* 2007) which was imputed against 1000 genomes data as reference. This was performed in an attempt to highlight new plausible candidate genes whose expression is associated with the *MACROD2* locus. The associated region of *MACROD2* contains an intra-genic anti-sense RNA species, *MACROD2-AS1*, which I had previously detected to be brain expressed via expression profiling experiments. Thus I sought to investigate this region in greater detail using brain expression data as anti-sense RNA are known to regulate cis and trans genes (Brantl 2002). Considering what I had hypothesised at the outset of this study regarding a role for *MACROD2* in histone biology it was possible a protein with such a role may have many downstream effects. I had hypothesised that such an analysis could provide additional information to implicate *MACROD2* in ASD.

66 genes showed strong association at this locus, and thus were brought forward for investigation using Gene Ontology Enrichment analyses. A caveat to the interpretation of these results is that during QC, any probes that failed to map to known genes were excluded. It is possible that by removing these probes the analysis was excluding

informative loci. Overall only half of the probes passed QC, which means the study was furthermore underpowered. As more expression data from brain derived tissues becomes available, an analysis such as the one described here will hopefully increase in statistical power and thus give greater confidence in any variants associated with *cis* and *trans* effects on expression.

3.5.3.5 Gene Enrichment Analysis of *MACROD2*-related trans-eQTL in human control brain

Exploration of association enrichment of these genes within biologically related gene-sets (gene-ontology terms) by the GRT did not identify statistically robust processes. Although unlikely to be a major regulator of those gene-sets tested, the limited sample size precludes conclusive evidence to exclude *MACROD2-AS1* from further investigation on possible functional underpinning of the ASD association. In addition to this, the number of individuals from the Myers dataset with recorded *MACROD2* expression was low, reducing the power to observe an effect even further. Input of the same gene list into the DAVID functional annotation tool however, did yield significant enrichments for biological processes, with one process, acetylation, maintaining significance even after correction for multiple comparisons (Bonferroni correction). The difference between the two enrichment analyses is likely due to the differences in how the enrichment score is calculated. The GRT uses multiple permutations (10,000 in this case) to calculate a simulated p-value based on the likelihood of observing a particular gene set. DAVID uses a modified Fisher's exact test (EASE score: Expression Analysis Systemic Explorer (Hosack *et al.* 2003)) to calculate whether lists of genes are more enriched for particular GO terms than by random chance. Enrichment for the acetylation biological process is biologically plausible in terms of the known enzymatic function of *MACROD2* as an AADPR deacetylase, but how changes in *MACROD2* expression could impact other genes from the same biological process is unknown and warrants further investigation. It is possible that the shared sequence similarity between different acetylase/deacetylases (Rosenthal *et al.* 2013; Jankevicius *et al.* 2013) means that an anti-sense RNA (e.g. *MACROD2-AS1*) that targets the expression of one such enzyme can also affect the expression of other homologous genes.

3.5.3.6 Summary and Future Directions

The work presented here has begun to build up a picture of *MACROD2* at the gene level, and also represents some of the only experimentally validated information for this gene at present. I have performed RNA expression profiling of *MACROD2* across 21 human tissues and discovered that the *MACROD2* gene is expressed in the vast majority of the tissues analysed, including human adult and foetal brain. As such this result is not as informative as if the gene expression was only detected in neuronal tissue, but it does indicate near ubiquitous expression. Taken into the context of the evolutionary conservation of this gene, this could indicate an essential role for this protein. However the observation that *MACROD2-AS1* is primarily brain expressed highlights this intronic loci for further study using brain expression data, although (publically available) sample sizes are limited at present.

I have performed gene reporter assays, which have demonstrated two things. Firstly that the *MACROD2* promoter has activity in a neuronal cell line, in concordance with the result that *MACROD2* is expressed in human brain. Secondly due to the use of fragments of the putative promoter region, the location of *MACROD2* regulatory elements in the putative promoter have been identified. The MatInspector analysis described in Table 3.7 highlights core promoter regions that may be the active regulatory domains.

This project also involved the use of *in silico* based techniques, such as GWAS and eQTL analysis, which the NPG and AGP use extensively in carrying out their research. In this way one can appreciate the two tiered approach to much of the analysis described in this report: to use *in silico* based methods to give added support, expand upon or lend credence to *in vitro* results.

For example, as more and more next-generation sequencing data becomes available this data can be used to support the results from the RACE experiments and build up a more complete picture of the gene structure, on a scale that is perhaps unrivalled. The eQTL analysis, although limited by the size of the Myers dataset (at 191 individuals following QC), demonstrates that such an analysis can highlight particular genes and biological processes that are affected by expression at *cis* or *trans* loci.

The observation of missense variants in autism cases from the Irish exome sequencing study and *de novo* CNVs across exon 5 of *MACROD2* in a small number of individuals from the AGP sample that co-segregate with severity of the disorder are interesting observations, but the sample size is very limited. *MACROD2* continues to appear as one of the top findings in large meta-analyses of autism (Psychiatric Genomics Consortium – ASD Working Group, personal communication) and so still remains an interesting susceptibility gene. This is discussed in greater detail in Chapter 5. It is imperative that biological relevance of associated variation is demonstrated to ascertain causal links. It is for this reason that study of the associated variant in greater detail is required, and is discussed at length in Chapter 4. This study is built upon this premise to translate the statistical evidence from GWAS to biological role, beginning with investigation of the *MACROD2* locus and the gene structure, expression and variation, which has informed the work described in Chapter 4 and 5. This comprehensive study of *MACROD2* has shed some light on a poorly characterised gene, but it is clear that much further study is required to understand how variance at this locus may contribute to neuropsychiatric disorders and human health in general.

Chapter 4 Investigation of the function of MacroD2 in the human brain

4.1 Introduction

At the time of the initial observation of the genome-wide significant association of SNPs at the *MACROD2* (Macro domain containing 2) locus there was very little known about the basic biology of *MACROD2*, let alone any evidence from the literature to link it to playing a role in the development of autism. This chapter describes an effort to improve our understanding of the function of *MACROD2* in humans, with a focus on the human brain. Whilst the other chapters that form a part of this work focus on basic biology and generating more support for the association of *MACROD2* with autism, this chapter is perhaps more informative when it comes to a role for *MACROD2* with neurodevelopmental disorders, via focusing on the localisation and molecular function of the protein. The dearth of functional information on *MACROD2*, let alone experimentally validated information, highlights the need to provide supporting evidence for the function of the protein and any possible role in neuropsychiatric disorder.

In order to gain further knowledge about the aetiology of disorder from association studies it is important to understand the function of proteins under these association signals. In the case of *MACROD2*, where the literature is lacking in information on gene and protein function, I have instigated investigations to gather information to help inform these early association findings. Once we have this information we can make further inference regarding these associations. The ultimate aim of research into the genetic aetiology of ASD is to inform the diagnosis, treatment and prognosis in autism, but this is a piecemeal process. At present we are in the interim period between basic science - the discovery of causative variants and discovery of the molecular pathways leading to disease - and clinical treatments that can be developed using what we have learned about the molecular mechanisms of the disorder. In the same way that identifying the molecular mechanism behind the syndromic causes of autism is leading to the development of new treatments, functional genomics studies in autism are revealing new clues as to how we might treat non-syndromic forms.

4.1.1 Functional genomics studies in neuropsychiatric genetics

Functional genomics is the study of the basic biology and function of a gene and its protein product. A greater understanding of the function will aid our understanding of

how the protein may impact on disease. There are a number of examples in psychiatry where the identification of the genetic aetiology of a disorder has informed diagnosis and treatment. For example, Fragile X Syndrome, the most common single gene defect in autism, accounting for up to 5% of cases (Budimirovic & Kaufmann 2011) was found to be most commonly associated with the expansion of a trinucleotide repeat in the 5' UTR of the *FMR1* gene, resulting in a failure to express FMRP (Fragile X mental retardation protein). Study of FMRP revealed that it is needed for the transport of mRNA from the nucleus to the dendrites of neurons, which leads to impairment in synapse formation and ultimately in impaired neural plasticity (Bassell & Warren 2008). Whilst this understanding has not yet led to better treatment, it has highlighted possible therapeutic approaches (Santoro *et al.* 2012).

There are also more promising hopes for treatment of a syndromic form of developmental delay with autism-like features, Rett syndrome. Rett syndrome has some symptoms that are very similar to autism that has been demonstrated to be caused by mutations in the gene *MECP2* (Methyl CpG binding protein 2) (Amir *et al.* 1999). *MECP2* binds to methylated CpG sites, which triggers silencing of that region of DNA via condensation of chromatin and interaction with histone deacetylases or by blocking transcription factors. *MECP2* has been implicated in the transcriptional regulation of imprinted genes, such as *UBE3A*, a gene that has been linked to two disorders with autism-like features: Angelman syndrome and Prader-Willi syndrome (LaSalle 2007). Thus mutations in the *MECP2* gene that affects the silencing activity of the protein mean that genes that are normally regulated by *MECP2* binding are no longer regulated. Further functional studies in a mouse model of Rett syndrome revealed that *MECP2* has its main effect in the noradrenergic system in the locus ceruleus of the brain, which affects cognition among other effects (Taneja *et al.* 2009). Mouse models for Rett syndrome have also demonstrated that treatment with an active peptide fragment of IGF-1 can ameliorate the symptoms of Rett syndrome, including increasing brain weight (Tropea *et al.* 2009).

These two examples demonstrate that understanding the underlying mechanism behind autism or autism-like phenotypes is leading to new treatments. It appears that transport to the synapse and disruption of the regulation of gene expression may be important mechanisms in autism (Bourgeron 2009). These studies also highlight the

utility of a ‘top-down’ approach, whereby you start with the disorder and work backwards to understanding the pathological mechanisms. However the ‘bottom-up’ approach to understanding neuronal function has identified some interesting candidate genes, such as the SHANK family of proteins, via investigation of important neuronal protein complexes (e.g. PSD-95) and neuronal receptors (e.g. NMDA receptor) (Naisbitt *et al.* 1999). *SHANK3* was associated with 22q13 deletion syndrome (Wilson 2003) and further study highlighted variation at *SHANK3* as the cause of the major neurological ASD-like features associated with 22q13 deletion syndrome (Bonaglia *et al.* 2011). Disruption of *SHANK3* in mutant mice mirrored these ASD-like features, as these mice displayed deficits affecting social interaction, social communication and overall synaptic function (Bozdagi *et al.* 2010). *SHANK3* is also a target of FMRP (Darnell *et al.* 2011), which not only supports a role for the disruption of glutamatergic signalling in ASD, but also demonstrates how functional studies of individual proteins and their effects slowly build up a picture of the molecular networks and mechanisms behind a disorder.

4.1.2 MACROD2

At the outset of this project the role of *MACROD2* in human biology was unknown and therefore its clinical relevance was debateable. The observation that CNVs at the *MACROD2* locus were common in the normal population has been used to infer that variation at this locus was likely benign (Reddy *et al.* 2013). However, in the interim period between the start and end of this project, and including the work described herein, a better picture of the role of *MACROD2* in human cell biology and its potential role in neurological disorders is coming to the fore.

As described in Chapter 3, a growing body of evidence suggests that variation at the *MACROD2* locus is associated with neuronal disease and disorder via association studies and in particular with CNVs in schizophrenia. However, this information only gives us hints as to what processes the protein may function or influence. It does tell us that we should look at brains and neurons to explore the protein function, and a growing number of studies of brain phenotypes, such as brain volume, and disease phenotypes of the brain, such as encephalopathies, have found associations with *MACROD2*. In an Array-CGH analysis of 111 patients with holoprosencephaly (HPE), a congenital disorder of the brain in which the forebrain fails to develop into two hemispheres, Bendavid and colleagues identified CNVs within *MACROD2* at a higher rate than in the general

population (Bendavid *et al.* 2009). A more recent study of HPE in “about 1500 patients” identified CNVs in *MACROD2* in seven cases - again a higher rate than in the general population (Mercier *et al.* 2011). It is not clear if this larger data set contains data from the smaller Bendavid paper, but there does appear to be a higher rate of CNV in *MACROD2* in HPE patients from this data. These patients also had mutations in the genes *SHH*, *ZIC2*, *SIX3* and *TGIF*. It is unknown if *MACROD2* interacts with these proteins in any way.

There is some evidence from the literature to link encephalopathies with autism. A study of CNVs in 61 patients with CWSSS and LKS (Continuous spike and Waves during Slow wave Sleep Syndrome and Landau-Kleffner Syndrome, two forms of epileptic encephalopathies) discovered CNVs in multiple genes associated with ASD, speech or language impairment - including 2 CNVs within *MACROD2* (Lesca *et al.* 2012). Epilepsy is the medical condition most highly associated with ASD, affecting up to a third of patients (Muhle *et al.* 2004), although the risk of comorbidity between ASD and epilepsy varies with the patients age, cognitive level or type of language disorder (Tuchman & Rapin 2002). However, to date an investigation of *MACROD2* in a sample of patients with ASD and epilepsy has not been carried out. Whilst the clinical relevance of variation at *MACROD2* has been disputed due to the presence of common CNVs (as discussed in Chapter 3), neurologists have highlighted associations with this protein and brain volume. SNPs within *MACROD2* (under the positional name *C20orf133*) were associated with risk of MRI defined brain infarct (S Debette *et al.* 2010) and reduced parenchymal volume (Baranzini *et al.* 2009).

The strongest evidence for the role of *MACROD2* in brain structure and function comes from an association study of MRI-defined temporal lobe volume across 729 elderly North American subjects from the Alzheimer’s Disease Neuroimaging Initiative (ADNI). Associations were calculated using LASSO (Least Absolute Shrinkage and Selection Operator) regression, whereby the combined effect of many correlated variables is assessed; essentially SNPs within a gene are grouped together and assessed as one. *MACROD2* was the most significant association out of a total of 22 GW significant associations, which included other ASD candidate genes, such as *GRIN2B* and *NRXN3* (Kohannim *et al.* 2012). The authors were able to replicate this association with

MACROD2 in a sample of 564 healthy young adult twins and siblings from Brisbane, although the effect on temporal lobe volume was not as extreme in this sample.

Despite this recent strong evidence suggesting a role for *MACROD2* in brain structure, no evidence was found from a recent high-profile investigation of the transcriptomic differences between autism cases and controls. In a study of 19 autism cases and 17 controls from the Autism Tissue Project, consistent differences in gene expression between autistic and normal brains was reported using gene co-expression network analysis (Voineagu *et al.* 2011). The authors did not report any finding for *MACROD2*, however, the sample size is small and the authors only used tissue from regions they believed to be the most significant in autism: superior temporal gyrus (the primary auditory cortex), prefrontal cortex (executive function) and cerebellar vermis (posture and locomotion). *MACROD2* has been reported as primarily expressed, in the developing foetal and adult brain, across the cortex as a whole, but more specifically at the subventricular zone of striatum (emotional and motivational aspects of behaviour) and inferior colliculus of the tectum (auditory processing and relaying information to the superior temporal gyrus) (Lesch *et al.* 2012). Therefore the lack of a finding for *MACROD2* could be due to the bias of the study towards specific brain regions.

In fact looking at the regions of expression in murine brain of the mouse homolog for *MACROD2* reveals the expression pattern in the mouse brain:

“At E12.5, significant expression levels were seen in the neural tube and in the ganglia of the peripheral nervous system (the dorsal root and cranial ganglia)” ... “In the cranial region, increased levels of C20orf133 expression were seen in the epithelial and mesenchymal components of the tooth-bud condensations, the epithelium of the primitive nasal cavity, in cells lining the vestibulocochlear and cochlear ducts, and the cranial ganglia.” ... “In addition, the brain, and in particular the ventricular zone, which is also seen in the neural tube, became positive for C20orf133 mRNA at this developmental stage. The expression in the brain was also clear at E18.” ... “Just before birth, at E18.5, C20orf133 continued to be expressed in the brain, with relatively high levels of expression in discrete regions: the subventricular zone of striatum and olfactory lobe, the cortical plate, cerebellar primordium and the inferior colliculus of the tectum.” ... “Strong expression was seen in stage E18.5 and adult brain. Low, but marked expression

levels were seen in the E11.5 embryo, kidneys of embryonic stage E18.5 and adult mice, and E18.5 eyes and bladder. Expression seen in the heart at stage E18.5 was 45 times lower than in the brain at stage E18.5.” (Maas *et al.* 2007)

From this data by Maas and colleagues of in situ hybridisation and RT-PCR analysis it appears that *MACROD2* is highly expressed in the developing brain of the mouse.

Due to the variability amongst individuals with autism, it is difficult to describe distinct core and co-morbid deficits - and consequently a single neurobiological phenotype shared amongst all those with ASD. Nonetheless, the neurodevelopment of individuals with autism is altered compared to their typically developing peers – including differences in neuronal patterning, brain growth and cortical connectivity. This is supported by the results of neuroimaging studies which have shown abnormal growth patterns in cortical white matter, frontal and temporal lobes and limbic structures (e.g. amygdala) (Pardo & Eberhart 2007). Moreover, physical changes in single axons below prefrontal cortical areas in autism are likely to affect synaptic function, which fits with the hypothesis that the fundamental defect in autism is at the synapse (Zikopoulos & Barbas 2010). Brain structure and volume appears to be an important phenotype in autism, with some authors going so far as to suggest they can use MRI to assist in the diagnosis of autism (Ecker *et al.* 2010).

4.1.3 *MACROD2* and Autism Spectrum Disorder

The role of *MACROD2* in influencing brain development and structure remains unclear. Until recently there was very little information regarding the function of *MACROD2* in humans and model organisms; however, MACROD-like proteins are highly conserved across evolutionary time, which may indicate an essential role. The MACRO-domain is an ADP-ribose binding module (Karras *et al.* 2005). Consequently, as a MACRO-domain containing protein, *MACROD2* has been implicated in the ADP-ribosylation of proteins, an important post-translational modification that occurs in a variety of biological processes such as DNA repair, transcription, chromatin biology and long-term memory formation (Cohen-Armon *et al.* 2004). A role for *MACROD2* in DNA repair has some experimental evidence; the *MACROD2* macro domain recruits to sites of PARP1 activity *in vivo* after a DNA damage event (Timinszky *et al.* 2009; Supplemental). There is some evidence to suggest that *MACROD2* directly interacts with TRF2 (Telomere Repeat binding Factor 2), together with other proteins involved with DNA repair and chromatin

remodelling (Giannone *et al.* 2010). Chen and colleagues demonstrated that several human MACRO-domain containing proteins (including human *MACROD2* and its homolog MacroD1) demonstrate O-Acetyl-ADP-ribose (AADPR) deacetylase activity when cloned into yeasts (Chen *et al.* 2011) and more recent studies have identified *MACROD2* as a mono-ADP-ribosylhydrolase (Rosenthal *et al.* 2013; Jankevicius *et al.* 2013), implicating it further in post-translational protein modification that regulates diverse biological processes such as cell surface interactions, trafficking of endosomal vesicles and cell signalling, among many other processes (Berger *et al.* 2004; Corda & Di Girolamo 2003; Bürkle 2005). The cellular localisation of recombinant EGFP-tagged *MACROD2* was also described as being both cytoplasmic and nuclear in HeLa cell lines (Neuvonen & Ahola 2009), which could indicate a diverse function or transcriptional role. AADPR, a product of NAD-dependent histone deacetylation, has a potential role in heterochromatin formation, histone modification and sirtuin biology (Liou *et al.* 2005; Hoff & Wolberger 2005), which supports this idea of transcriptional activity.

Disruption of this deacetylase activity has not yet been conclusively linked to disease; however, considering the connection that has been made between protein deacetylation and energy metabolism (Imai & Guarente 2010), it is possible that an energy hungry organ such as the brain is vulnerable to disruptions in these enzymes. The supposition that *MACROD2* could be involved in neuropsychiatric disorders such as ASD is supported by GWA studies linking it to disorders of the central nervous system, but not necessarily supported by the study associating it with DNA repair mechanisms and histone modification. If *MACROD2* interacts with histone proteins and thus affects transcriptional regulation, it could have large downstream effects, in much the same way as the interaction of *MeCP2* with histone deacetylases has been documented to affect the expression of other ASD-related genes, as discussed in Section 4.1.1 above.

However any effect of *MACROD2* upon gene expression has not been assessed. Following the GWAS findings for *MACROD2*, one hypothesis is that *MACROD2* has a function in the survivability of neurons and/or the connections between them. However, the hypothesis need not necessarily be neuron specific - there are examples of mutations presenting in intellectually disabled patients that do not affect neuron specific processes (L. Abbasi Moheb *et al.* 2011), but affect (downstream) general cellular processes. One can

argue that the brain is a very vulnerable organ which may explain why such mutations affect the brain more severely, impacting intellect and behaviour.

4.1.4 Aims of this study

This work represents a functional study of a protein that has not been characterised in humans and may be a protein important to human health. In this project I have sought to answer some of the fundamental questions about the localisation, interaction and function of the *MACROD2* protein.

The aims of this study were:

Localisation: To establish the localisation of *MACROD2* within neuronal cells. I hypothesise that *MACROD2* is located in the nucleus and the cytoplasm of human neurons due to the results from the literature (Neuvonen & Ahola 2009).

Interaction: To identify what interactions exist between the *MACROD2* protein to better characterise the protein function and identify putative biological networks that the protein may work within. There is some evidence to suggest *MACROD2* interacts with histone proteins and sirtuins (Chen et al. 2011; Timinszky et al. 2009). However, I also hypothesise that *MACROD2* is linked, either by protein-protein interaction or by sharing a molecular process, to other putative autism susceptibility genes.

siRNA Knockdown: To examine the gross impact of *MACROD2* dose in human neurons using siRNA knockdown methodology. I hypothesise that *MACROD2* will have a function important to the survivability of the neuron due to its association with multiple neurological and neuropsychiatric disorders and the observation that it is highly expressed in the developing brain of the mouse.

4.2 Materials and Methods

4.2.1 Localisation

Due to the lack of information surrounding the localisation of endogenous *MACROD2*, I sought to experimentally validate the localisation of the protein in neuronal cells. This required generation of a neuronal cell model and immunohistochemistry to target and visualise the *MACROD2* protein.

4.2.1.1 Generating a Neuronal Cell Model

In order to examine the localisation of *MACROD2* in the human neuron, it was necessary to create a cell model that was a suitable model for neurons *in vivo*. An appropriate cell model gives us confidence that any observations that are made using the cell model would also be made in human tissue *in vivo*. A neuronal cell model was required as I wanted to investigate the localisation *MACROD2* in the human neuron.

Therefore I created a differentiated SH-SY5Y neuroblastoma cell line model in-house for use in subsequent experiments (Figure 4.1). This protocol was adapted from Encinas (Encinas *et al.* 2000). Neuronal cells require a substrate upon which to grow. For this reason cells were grown on poly-L-lysine or collagen type IV coated cover slips/flasks or in pre-treated 'Cell+' flasks (Sarstedt). I determined that using pre-treated flasks or poly-L-lysine coated cover slips was the most cost effective method for cell growth. Treating the cells with 10 μ M Retinoic Acid (RA) for a period of 5 days, and then following washing, incubating the cells in 50ng/mL BDNF (Brain Derived Neurotrophic Factor) containing media (without FBS or RA) for a period of 3-10 days results in a more 'neuron-like' phenotype in these cells. The protocol was performed as follows:

1. Media was emptied from the confluent flask and the cells were washed with 5 mL warm (37°C) PBS, rocking gently to wash the entire surface of the flask.
2. 1 mL 10x Trypsin (Sigma) was diluted to 2x using 4 mL PBS.
3. PBS was removed from the cells and 5ml trypsin 2X was added to each flask. The flasks were incubated at 37°C in an incubator until the cells had detached (Approximately 5 minutes incubation).
4. 5 mL warm DMEM was added to each flask to halt the trypsin. The cell suspension was then removed from each flask to a fresh 15 mL tube and centrifuged at 1200rpm for 2 minutes.
5. The cells were re-suspended in 10 mL DMEM (provided with 2.5 mM L-glutamine and supplemented with 15% foetal bovine serum) and counted on a haemocytometer. Approximately 10^6 cells were diluted 1:20 (e.g. 0.5 mL cell suspension in 9.5 mL media) and plated in coated culture flasks. Cells were seeded at 1×10^4 for 24-well plates.

6. Retinoic Acid ($10\text{ }\mu\text{M}$ in DMEM with 15% FBS) was added the day after plating and incubated for 5 days for a 175cm^2 flask, 24-48 hours for a 24-well plate.
7. Following the incubation in the presence of RA, the cells were washed 3 times with DMEM (serum free) and incubated with 50ng/ml BDNF in DMEM (without serum) for 5 days in a 175cm^2 flask, or 24-96 hours for a 24-well plate.

When diluting retinoic acid and BDNF, each supplement was first dissolved in a small volume of DMEM media and then filtered through a $0.2\text{ }\mu\text{m}$ filter using a syringe into the final volume of media. This step was included to ensure sterility of the cultures. Cells were not cultured beyond passage 30 or if an overt phenotypic change was observed, such as a change in cell morphology.

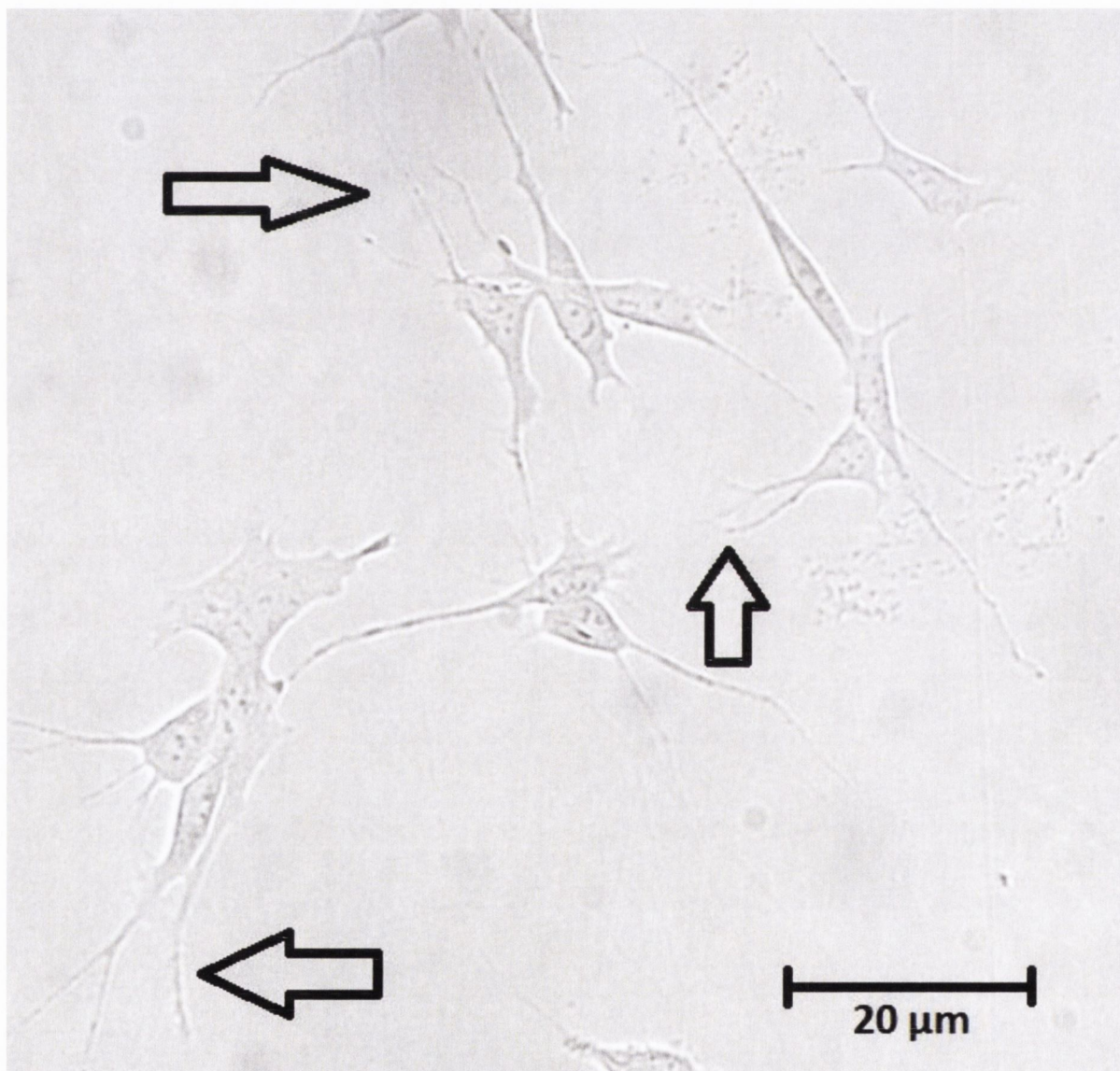


Figure 4.1 Differentiated SH-SY5Y cells in culture. SH-SY5Y neuroblastoma cells are differentiated to a more neuron-like phenotype (defined as cells that are phase bright with neurite outgrowth) and are then used in subsequent experiments. Neurite outgrowth is highlighted with black arrows.

4.2.1.2 Immunohistochemistry for visualisation of neuronal markers (PSD95, Synapsin and MAP2) by Confocal Microscopy in the SH-SY5Y cell model

In order to verify that the cell model was an appropriate model for neuronal function, immunohistochemistry was used to confirm the expression of several markers of neuronal function in differentiated but otherwise untreated SH-SY5Y cell lines. It was also required to verify that the differentiated cells were a homogenous population of neuronal cells. I used the neuronal markers PSD95, Synapsin and MAP2 to establish this; PSD95 and Synapsin are used to visualise the synapse, whereas MAP2 is used to visualise the cell body and dendrites. The mounting medium contains DAPI (4',6-Diamidino-2-Phenylindole) which stains the nuclei, enabling easy identification of individual cells. SH-SY5Y neuroblastoma cells were plated at a density of 5×10^4 per 24-well cell culture plate. The cells were differentiated following the protocol described in section 4.2.1 above. However, for this method the cells were grown on cover slips coated with Poly-L-Lysine. Cultures with the best 'neuron like' morphology, determined by light microscopy, were chosen for immunostaining as follows:

1. The cells were fixed in 4% PFA in PBS for 20 minutes.
2. The cultures were washed 3 times in PBS and placed on a shaker for 5 minutes each time at 250 rpm, room temperature.
3. Non-specific binding was blocked by adding 10% Goat serum in 0.1% PBST, 400 μ l/well. Samples were incubated on a shaker for 1 hour, 250 rpm at room temperature.
4. The primary antibody mix was prepared in 1% Goat serum in 0.1% PBST (PBS + Tween-20), 350 μ l/well and incubated overnight at 4°C at the following ratios:
 - 1:400 rabbit anti-mouse* PSD95 (Cell Signalling)
 - 1:1000 guinea pig anti-mouse* Synapsin (SYSY)
 - 1:200 mouse anti-mouse* MAP2 (Millipore)*: These antibodies demonstrate reactivity with human peptides and thus were suitable for use with the human SH-SY5Y cell lines.
5. The samples were washed 3 times with PBS, 5 minutes each time, 250rpm at room temperature.
6. Secondary antibody in 0.05% PBST was added, 350 μ l/well. The plates were covered with aluminium foil due to the light sensitivity of the secondary antibodies and incubated for 2 hours:
 - 1:500 goat anti-rabbit DL488
 - 1:500 goat anti-guinea pig Cy3
 - 1:500 goat anti-mouse DL649
7. The samples were washed 3 times in PBS, 5 minutes each time, 250rpm, at room temperature while covered with aluminium foil.
8. Each coverslip was then mounted onto a slide using Vectashield Hard set with DAPI and stored at 4°C.

4.2.1.3 Immunohistochemistry for visualisation of *MACROD2* by Confocal Microscopy in the siRNA treated SH-SY5Y cell model

The immunohistochemistry protocol was adapted for use with the *MACROD2* antibody. Therefore for this protocol, anti-PSD95 was not used as only three fluorescent dyes may be used at a time for the best confocal microscopy results; of the two synaptic markers, Synapsin was found to visualise better than PSD-95 in the differentiated SH-SY5Y cell lines. SH-SY5Y neuroblastoma cells were plated at a density of 5×10^5 in a 24-well cell culture plate and transfected with siRNA according to the protocol described in Section 4.2.3.2 below. After 72-96 hours, cells were treated for immunostaining as follows:

1. The cells were fixed in PFA.
2. The cultures were washed 3 times in PBS and placed on a shaker for 5 minutes each time at 250rpm, room temperature.
3. Non-specific binding was blocked by adding 10% Donkey serum in 0.1% PBST, 400 μ l/well. Samples were incubated on a shaker for 1 hour, 250 rpm at room temperature.
4. The primary antibody mix was prepared in 10% Donkey serum in 0.1% PBST, 350 μ l/well and incubated overnight at room temperature at the following ratios:
 - 1:500 goat anti-human *MACROD2* (Santa Cruz)
 - 1:1000 guinea pig anti-mouse* Synapsin (SYSY)
 - 1:200 mouse anti-mouse* MAP2 (Millipore)*: These antibodies demonstrate reactivity with human peptides and thus were suitable for use with the human SH-SY5Y cell lines.
5. The samples were washed 3 times with PBS, 5 minutes each time, 250rpm at room temperature.
6. Secondary antibody in 0.05% PBST was added, 350 μ l/well. The plates were covered with aluminium foil due to the light sensitivity of the secondary antibodies and incubated for 2 hours:
 - 1:500 goat anti-guinea pig (Cy3)
 - 1:500 goat anti-mouse (DL649)
 - 1:500 goat anti-goat (DL488)
7. The samples were washed 3 times in PBS, 5 minutes each time, 250rpm, at room temperature while covered with aluminium foil.
8. Each coverslip was then mounted onto a slide using Vectashield Hard set with DAPI and stored at 4°C.

4.2.1.4 Confocal Microscopy

Confocal microscopy was performed under the supervision of Dr. Ines Molinos, using a ZEISS LSM 510 confocal microscope coupled to a digital camera. Image analysis was performed with Zeiss LSM 510 operating software and the freely available ImageJ imaging software (<http://rsbweb.nih.gov/ij/>). Cells were analysed for expression of the two neuronal markers, Synapsin and MAP-2 (PSD-95 was assessed only during validation

of the neuronal cell model). For the siRNA knockdown experiments, the gross expression and localisation of *MACROD2* was assessed across different concentrations of siRNA.

4.2.2 Interaction

To elucidate the role of *MACROD2* in overall cell biology, and to investigate if the hypothesis that *MACROD2* interacts with histone or sirtuin proteins in the nucleus, I used co-immunoprecipitation and analysis of the subsequent eluent to demonstrate what, if any, protein partners *MACROD2* had in a differentiated neuronal cell line.

4.2.2.1 Protein Extraction from Cell Culture by Detergent Lysis

Protein was extracted from SH-SY5Y cells in culture for use in subsequent experiments. RIPA (Radioimmunoprecipitation assay buffer, see Table 4.1) was found to be the best performing buffer for extraction of *MACROD2* protein from SH-SY5Y cells in culture. Due to the use of protease inhibitor tablets that are recommended to be prepared freshly before use, the buffer was always prepared immediately prior to protein extraction and performed as follows, with the cell culture plates maintained on ice for the duration of the protocol:

Preparation of whole cell extract:

1. The culture medium was discarded and the cells were washed twice with cold phosphate buffered saline (PBS, Sigma), and all residue discarded.
2. RIPA buffer was added to each flask (4ml per 175cm² flask) and the flasks were then incubated for 10 minutes on ice.
3. The flasks were then scraped using a chilled cell scraper (Sarstedt).
4. The lysate was collected into 1,5ml eppendorf tubes and sonicated (using a Diagenode Bioruptor sonication bath) 3 times for 15 seconds each time.
5. The tubes were centrifuged at 13,000 rpm for 20 minutes at 4°C.
6. The supernatant, containing the proteins, was collected into a fresh eppendorf tube and stored at -80°C (or used immediately in co-immunoprecipitation).

Preparation of cytoplasmic and nuclear extract

1. The culture medium was discarded and the cells were washed twice with cold phosphate buffered saline (PBS, Sigma), and all residue discarded.
2. RIPA buffer was added to each flask (4ml per 175cm² flask) and the flasks were then incubated for 30 minutes on ice.
3. The flasks were then scraped using a chilled cell scraper (Sarstedt).
4. The lysate was collected into 1,5ml eppendorf tubes, homogenised and then centrifuged at 1000 rpm for 10 minutes at 4°C.
5. The supernatant was retained separately, this is the cytoplasmic fraction.
6. The pellet was resuspended in 500 µL RIPA buffer and homogenised and sonicated for 2 minutes (10 seconds on, 10 seconds off). This is the nuclear fraction.

7. The fractions were collected into fresh eppendorf tubes and stored at -80°C (or used immediately in co-immunoprecipitation).

Table 4.1 Composition of the RIPA buffer. This buffer is designed to preserve proteins during lysis - EDTA, PMSF and the protease inhibitor tablets are added to deactivate proteolytic enzymes that would otherwise degrade the proteins under study. Sodium orthovanadate is added to deactivate phosphatases, and thus preserve the phosphorylation of proteins of interest.

Reagents	Final Concentration
Tris-HCl pH 7.5	50mM
NaCl	150mM
EDTA	10mM
Triton X-100	1%
SDS	0.1%
PMSF	1mM
Sodium Orthovanadate	1mM
Protease Inhibitor Tablets	1 per 10mL
ddH ₂ O	-

4.2.2.2 Antibody Coupling Protocol

The co-immunoprecipitation method used in this study required the creation of antibody coated magnetic beads, which are used to bind to the target protein and are then ‘pulled down’ out of solution using a magnet. Any experiment that uses antibodies is going to experience a certain level of ‘background noise’, i.e. non-specific binding of proteins to the antibody. The antibody coupling protocol is designed to limit the amount of non-specific binding and thus improve the quality and specificity of the Co-IP. Magnetic beads suitable for use in covalent antibody binding (Dynabeads M-270 Epoxy, Invitrogen) were mixed with *MACROD2* antibody (Santa Cruz, sc-87990, goat polyclonal mapping near the C-terminus of *MACROD2* of human origin) and washed several times using a magnet rack (Invitrogen) and buffers C1 and C2 from the antibody coupling kit (Invitrogen). The beads were incubated with antibody and buffer C2 for 16-24 hours at 37°C overnight, with rotation. The beads were then used in co-immunoprecipitation experiments or stored at 4°C. Antibody coupling was optimised at 7 µg antibody per mg of beads.

4.2.2.3 Western Blot

Western blot was used to confirm the activity of the *MACROD2* antibody (Santa Cruz) and to confirm the presence of *MACROD2* in the co-immunoprecipitate. All western blots were prepared in the same manner, and visualised using Fusion FX (Vilber Lourmat) image acquisition system.

Table 4.2 The resolving and stacking gels were created as per Table 4.2. Samples were loaded in a 1:1 ratio of sample to buffer (Buffer was 6.25 µl Sample buffer, 2.5 µl Reducing Agent and 3.75 µl H₂O), with at least 50ng of protein extract loaded per lane of each gel for comparison. As the aim of these experiments was to confirm the activity of the antibody and the presence of the *MACROD2* protein in sufficient quantities from SH-SY5Y cell extract co-immunoprecipitate, only a minimum/maximum total protein content was set for these samples, as the total protein content was low within the samples originating from Co-IP (e.g. 0.2 mg/mL for Co-IP samples compared to 20mg/mL for protein extract). Precision Plus Protein Standards (Bio-Rad) were loaded to every gel. The transfer was performed on a dry transfer apparatus at 160V for 60 minutes. Following transfer the membrane was washed in TBS-T (3 times for 5 minutes each time) and then stained using red ponceau. The experimental procedure only continued at this point if protein was stained and visible in the positive control lanes, otherwise the membrane was discarded. The membrane was washed again in TBS-T and then blocking with non-fat milk (from powder, 5% in 0.1mM TBS-T) was performed for 60 minutes, with shaking. Primary *MACROD2* antibody (Santa Cruz, sc-87990) was diluted 1:2000 in milk and incubated at 4°C overnight. The membrane was then washed and incubated with secondary Anti-hRP antibody (Santa Cruz) diluted to 1:2000 in milk, washed a final time in TBS-T and then imaged using a Fusion FX (Vilber Lourmat) image acquisition system.

Table 4.2 Composition of the resolving and stacking gel for Western Blot.

Reagents	Resolving Gel	Stacking Gel
H ₂ O	4.0 µl	2.7 µl
30% Acrylamide Mix	3.3 µl	0.67 µl
Tris (pH 8.8)	2.5 µl of 1.5mM	0.5 µl of 0.5mM
10% SDS	0.1 µl	0.04 µl
10% Ammonium Persulfate	0.05 µl	0.04 µl
TEMED	0.005 µl	0.004 µl
Total	~10 ml	~10 ml

4.2.2.4 Co-Immunoprecipitation Protocol and Mass Spectrometry analysis

Co-immunoprecipitation is a powerful technique used to examine interactions between a target protein and any protein partners. The goal of Co-IP is isolation of intact protein complexes and binding partners from protein extract, for example from cell lines or tissue biopsy, using an antibody specific to a protein of interest. The antibody binds to the known protein of interest, which is bound to known or unknown members of protein complexes, and is extracted from the protein extract. This can result in the identification

of protein-protein interactions via 'pull-down' of intact protein complexes and subsequent analysis of the isolate.

Co-immunoprecipitation was performed using the Dynabeads Co-Immunoprecipitation Kit (Invitrogen) following the default manufacturers protocol and RIPA buffer. 1.5mg of antibody coupled Dynabeads were used per 0.05-1.5g of cell sample. The cell lysate was incubated for one hour with antibody coated beads, instead of the recommended 30 minutes according to the manufacturer's protocol. The manufacturers state that this kit is able to pick up transient or labile complexes, thus at the final step, the eluent was immediately stored at -80°C to ensure maximal preservation of any protein complexes.

Four samples in total were sent for analysis: Two samples representative of the whole cell protein extract, one sample from a cytoplasmic fraction of whole cell protein extract and one sample from a nuclear fraction of whole cell protein extract. This separation of samples enabled analysis of *MACROD2* across the whole cell, and then to look for localisation effects of protein-protein interactions across the nucleus and cytoplasmic fractions. The samples were then sent to ProteaBio for analysis, who consulted with me about how to proceed with the analysis. I decided upon grouping of bands from defined molecular weight ranges to be analysed together in an attempt to harvest the maximum amount of material.

4.2.2.5 Preparation of proteins for Mass Spectrometry analysis

Following validation of the Co-IP samples they were sent to Protea Biosciences for analysis. This required three steps: Separation via 1D-SDS PAGE, gel digestion with trypsin and protein ID by LC/ESI-MS. The following sections (From 4.2.2.6 to 4.2.2.9) describe work carried out by Protea Biosciences:

4.2.2.6 Gel Loading

The Co-IP sample was separated using 1D SDS-PAGE analysis. 15µL of each sample was mixed with 15µL Laemmli buffer +BME (1:1 dilution) and incubated at 90 °C for 5 minutes. Samples (30 µL), BSA (10 µL), and molecular weight markers (10 µL) were loaded onto a BioRAD Mini-Protean TGX Gel (10%) and the gel was placed in a gel tank filled with an adequate volume of 1X TGS running buffer. The gel was run at 200V for 33 minutes. After electrophoresis, gel was removed from the cassette and rinsed by shaking in water for 5 minutes. After decanting off the water, the gel was fixed and silver stained.

4.2.2.7 In-Gel Digestion

Gel bands were cut into 1mm² pieces and placed into labeled microcentrifuge tubes. Bands of the same molecular weight were pooled together. 250 µL of 150mM ammonium bicarbonate (AmBic) was added to the samples and incubated at room temperature (RT) for no less than 5 minutes. The AmBic was removed. 250 µL of acetonitrile was added to each sample and incubated at room temperature for no less than 5 minutes. The acetonitrile was removed and samples dried on a lyophiliser.

250 µL of 10 mM dithiothreitol (DTT) in 50mM AmBic was added to each sample and incubated at 56°C for 45-60 minutes. The samples were allowed to cool to RT, centrifuged at 4000 RPM for 30 seconds at RT, and the DTT was removed. 250 µL of 55mM iodoacetamide (IA) in 50mM AmBic was added, and incubated at room temperature in the dark for 45 to 60 minutes. IA solution was removed. Added 250 µL of acetonitrile to dehydrate gel pieces and incubated for no less than 5 minutes at room temperature. Acetonitrile was removed and the samples were dried using a lyophiliser. Added 250 µL of 12.5ng/µL trypsin in 50mM AmBic to gel pieces and incubated on ice for at least 45 minutes. The trypsin was removed and 250 µL of 50mM AmBic was added to the samples. The samples were incubated overnight at 37°C.

The samples were removed from 37°C incubation and centrifuged at room temperature for 30 seconds at 4000RPM. The supernatants were transferred into pre-labeled clean microcentrifuge vials. The gel pieces were dehydrated in 250 µL of acetonitrile for 20 minutes and the supernatants were pooled into the corresponding vials. The samples were frozen at -80°C for approximately 30 minutes and placed on the lyophilizer to dry. Each of the samples was reconstituted in 50 µL of 0.1M acetic acid and placed in a freezer at -80°C for approximately 30 minutes. Samples were dried on a lyophilizer. The samples were then reconstituted in 40 µL of acetonitrile/water/formic acid (5%/95%/0.1%).

4.2.2.8 LC/ESI-MS Analysis

This work was outsourced to ProteaBio due their experience with LC/ESI-MS analysis. Therefore all instrument parameters were decided upon by their technicians and are reported here for reference:

4.2.2.8.1 Liquid chromatography parameters

The LC instrument used to perform the analysis was a Shimadzu LC-20AD HPLC Instrument running control software Analyst 1.5. The LC column was a Kinetex 100 x 2.1 mm C18 (100 Å, 1.7µm) column by Phenomenex (Torrance, CA), held at 40°C. The LC Separation was a 120 minute gradient, detailed in Table 4.3.

Table 4.3 HPLC gradient parameters as determined by ProteaBio for the liquid chromatography analysis.

Time (Minute)	Percentage Mobile Phase A (0.1% Formic Acid in Water)	Percentage Mobile Phase B (0.1% Formic Acid in ACN)	Flow Rate (µl/Minute)
0.01	98	2	200
3.00	98	2	200
90.0	60	40	200
100	10	90	200
114	10	90	200
115	98	2	200
120	98	2	200

4.2.2.8.2 ESI mass spectrometer parameters

The ESI-MS instrument used to perform the analysis was a QTrap5500 (AB Sciex Toronto, Canada) running data acquisition software Analyst 1.5. The ionization method was electrospray ionization in positive ion mode with a spray voltage at 5.0kV and a mass range of 100-1000 Da.

4.2.2.8.3 Database Search Parameters

The LC/ESI-MS analysis identified chains of peptides, which were then tagged with reference IDs following the parameters outlined in Table 4.4.

Table 4.4 Peptide chains were tagged with Uniprot protein accession number IDs using the parameters outlined below.

Program for MS/MS data processing	ABI Protein ProteinPilot software 4.0.
Search Engine	Paragon.
Sample Type	Identification.
Digestion Enzymes	Trypsin.
Special Factors	Gel-based ID.
Species	<i>Homo sapiens</i> .
I.D Focus	None.
Databases	uniprot-human.

4.2.2.9 Annotation of proteins from Mass Spectrometry analysis

Following completion of the Mass Spectrometry analysis, Protea Biosciences made all of the resulting data available as data summaries for each excised band or sample. I

culled this list for only those proteins that had at least 1 representative peptide fragment at 95% confidence being brought forward for further analysis. This short-list was annotated with updated Protein names, UniProt protein IDs and Gene Ontology terms culled from online databases.

4.2.2.10 Gene Ontology enrichment analysis using DAVID

DAVID (Database for Annotation, Visualisation and Integrated Discovery) v6.7 is an online gene ontology database that enables functional annotation of genes, and grouping of these genes together by gene ontology term (Huang *et al.* 2009). Gene-ontology enrichment analysis can discover whether lists of identified genes are enriched for a particular function. The list of genes from the LC/ESI-MS analysis of Co-IP pulldown was input using the web form at <http://david.abcc.ncifcrf.gov/tools.jsp>. This returned a grouping of genes by functional annotation, and demonstrated any enrichment for particular biological processes.

4.2.3 siRNA knockdown

To investigate the impact of a defect in *MACROD2* expression in human neurons, I sought to perform some preliminary experiments to knockdown the expression of *MACROD2* in the differentiated SH-SY5Y cell model using siRNA. The working hypothesis for these experiments was that if normal *MACROD2* expression is important to the structure or function of the neuron, that knockdown of this gene will have a noticeable impact upon overall cell morphology and/or expression of neuronal markers. Unfortunately due to time constraints, these experiments were limited in scale and replication. In an effort to enable comparison of results between RT-PCR and immunohistochemistry experiments, cells in culture were plated at the same time from the same seeding flask, using the same siRNA preparations. For the RT-PCR analysis, GAPDH was used as an internal control in an effort to control for differences arising from the assay itself. For the immunohistochemistry analysis, the same microscopy settings were used across the analyses to enable comparison between images. However, the experiments did differ regarding the incubation time following transfection: ~48 hours for mRNA detection and ~96 hours for protein detection.

4.2.3.1 Design of siRNA and resuspension protocol

The ON-TARGETplus SMARTpool siRNA system (Dharmacon/Thermo Scientific) was chosen for use in this study, due to its reported high silencing potency and low off-target effects. This siRNA contains 4 different siRNA constructs for the target gene. siRNA for *MACROD2* targeting the 3' UTR of the gene were readily available from this manufacturer. siRNA requires careful resuspension, as follows:

1. The tubes containing siRNA were centrifuged to ensure that the siRNA pellet was collected at the bottom of the tube.
2. 200 μ L of 5X siRNA buffer was diluted to 1X in 800 μ L RNase free water.
3. A working stock solution of 5 μ M siRNA was prepared for each siRNA in 1X siRNA buffer, e.g. 5nmol siRNA + 1000 μ L siRNA buffer = 5 μ M stock solution.
4. The solution was mixed by pipetting the solution up and down 3-5 times, avoiding the introduction of bubbles.
5. The solution was then placed on an orbital shaker for 30 minutes at room temperature.
6. The siRNA solution was centrifuged to ensure that the solution was collected at the bottom of the tube.
7. The concentration of the siRNA was verified at 260nm.
8. The siRNA was aliquoted into small volumes and stored at -20°C to -80°C. Freeze-thaw was limited to 5 times.

4.2.3.2 Transfection of siRNA into neuronal cell lines

The use of siRNA is a sensitive technology and thus this required optimisation of the transfection protocol. According to the manufacturer, 'DharmaFECT 1' was the transfection reagent most suitable for use with SH-SY5Y cell lines, and thus this was chosen for use in this project. Each experiment included the following samples in triplicate:

1. Untreated cells.
2. Positive control siRNA, targeting Cyclophilin B (*PPIB*).
3. Negative control siRNA, non-targeting 'scrambled' siRNA.
4. *MACROD2* test siRNA.

All steps of the protocol were performed in a laminar flow cell culture hood (Clean Air EN12469) using sterile techniques. Cell density was optimised at 5×10^5 cells in each well of a 24-well cell culture plate. This was 10-fold more concentrated than cells were typically plated at due to loss of cells from the transfection process, which is cytotoxic. 500 μ L of cells were plated in antibiotic free complete medium containing 10 μ M retinoic

acid into each well of a 24-well plate and incubated at 37°C with 5% CO₂ overnight. The cells were then transfected as follows:

1. siRNA was diluted to the required concentration in serum free media, to a total volume of 175 µL.
2. 1.75 µL DharmaFECT 1 transfection reagent was diluted in 173.25 µL serum free media.
3. The contents of each tube were gently mixed by pipetting, and incubated for 5 minutes at room temperature.
4. The siRNA was added to the transfection reagent, mixed gently by pipetting and incubated for 20 minutes at room temperature.
5. 1400 µL antibiotic free complete medium containing BDNF was added to the siRNA/transfection reagent mix, for a final volume of 1750 µL.
6. Culture medium was removed from the wells of the 24-well plate and washed gently with warm (37°C) PBS. 500 µL of transfection medium containing siRNA was added to each well.
7. The cells were incubated at 37°C in 5 % CO₂ for 24-48 hours (for mRNA analysis by qPCR) or 48-96 hours (for analysis of the effect of knockdown at the protein level on cell morphology by immunohistochemistry).

Transfection was optimised by testing the concentration of the *MACROD2* siRNA at 5nM, 10nM, 20nM, 35nM and 50nM. The 'scrambled' negative control and *PPIB* positive control were transfected at a concentration of 25nM, according to the manufacturer's guidelines. For best results, samples that showed >80 % viability were chosen for further analysis.

4.2.3.3 Relative Quantification of Gene Expression using Real-Time Quantitative PCR

In order to confirm the mRNA knockdown by the *MACROD2* siRNA construct, Taqman assays targeting the 3' UTR of *MACROD2* (As described in Chapter 3, *MACROD2* SNP1 - SNP5) and Exon 5 of GAPDH were used to assess the relative expression of *MACROD2* when normalised against the expression of GAPDH. Normalisation also required comparison against untreated cells. Analysis was performed using the SDS 2.4 and RQ Manager 1.2.1 software (Applied Biosystems) and Excel 2007 (Microsoft).

4.2.3.4 RNA isolation from Cell Culture and Reverse Transcription Reactions

The highly sensitive 'Single Cell-to-CT' RNA extraction kit (Ambion) was used to extract RNA from siRNA treated SH-SY5Y cells in culture. The protocol varied somewhat from the manufacturer's recommended protocol:

1. Individual coverslips were scraped using cell scrapers (Sarstedt) and the resulting cell suspension was collected into 1.5 mL eppendorf tubes.
2. The cells were then pelleted at 800 rpm for 5 minutes at 4°C.

3. 1 μ L of the resulting cell pellet was added to 10 μ L of Single Cell Lysis/DNase I solution and incubated for 5 minutes at room temperature.
4. 1 μ L of stop solution was added to each lysis reaction and incubated for 2 minutes at room temperature. 4.5 μ L of RT Mix was added to each reaction volume and incubated for:
 - 10 minutes at 25 °C
 - 60 minutes at 42 °C
 - 5 minutes at 85 °C
5. 11 μ L of PreAmp Mix/pooled TaqMan assays was added to each reaction volume and incubated for:
 - 10 minutes at 95 °CFollowed by 14 cycles of:
 - 15 seconds at 95 °C
 - 4 minutes at 60 °C
6. Each reaction volume was quantified at 260nm using a NanoDropper.
7. Each reaction volume was diluted 1:20 with 1X TE Buffer, pH 8.0 and used to perform real-time PCR.

4.3 Results

4.3.1 Localisation

As discussed in section 4.1, according to the wider scientific literature there has been no experiment to examine the localisation of endogenous *MACROD2* expression, and no results to document the localisation of *MACROD2* in human neurons. Thus I sought to demonstrate the localisation of *MACROD2* in a human neuronal cell model.

4.3.1.1 Differentiated SH-SY5Y cell lines express PSD-95, Synapsin and MAP2

In order to examine the effect of *MACROD2* depletion upon the cell and upon neurite outgrowth, it was necessary to first create a cell model within which experiments to test the effect of such variation can be conducted. The SH-SY5Y neuroblastoma cells were differentiated using retinoic acid and BDNF. The retinoic acid treatment causes the cells to withdraw from the cell cycle and display neurite outgrowth, whereas the BDNF treatment makes the cells trophic dependent, and thus more similar to neurons *in vivo*.

I confirmed the differentiation of SH-SY5Y cells was complete experimentally via examination of the expression of the neuronal markers PSD-95, MAP2 and Synapsin using immunohistochemistry. Visualisation of the cells was performed by confocal microscopy and confirmed that the SH-SY5Y cell lines had differentiated. All of PSD-95, MAP-2 and Synapsin were shown to be expressed (see Figure 4.2).

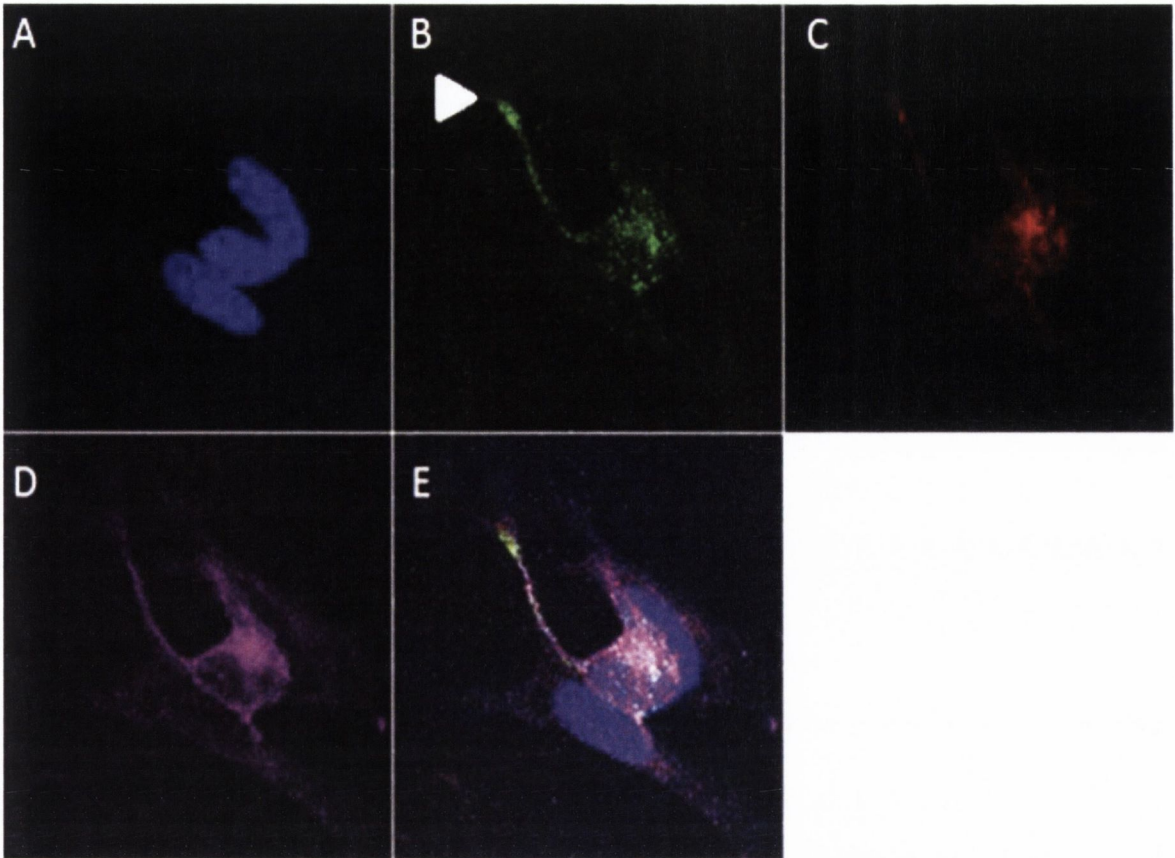


Figure 4.2 Expression of neuronal markers in SH-SY5Y cell lines. Panel A: DAPI staining of the nuclei, three cells are visible; Panel B: Staining of PSD-95, the bright point at the end of the neurite is determined to be a neuronal puncta by elevated local expression of PSD-95; Panel C: Synapsin staining; Panel D: MAP2 staining; Panel E: Overlaid staining of Panels A-D.

4.3.1.2 Identification of the localisation of *MACROD2* using Immunohistochemistry

Following development of the cell model and confirmation that the cell model was suitable for use in further experiments due to the expression of neuronal markers in the SH-SY5Y cell line, it was then necessary to confirm that *MACROD2* could be visualised in these cells. Prior to this analysis, it was reported that EGFP-tagged recombinant *MACROD2* (here named as *MDO2*) was expressed in the cytoplasm and nucleus (Neuvonen & Ahola 2009). From these results (Illustrated in Figure 4.3) it would appear that endogenous *MACROD2* is expressed across the entire cell (cytoplasm, neurites and nucleus), with a higher rate of expression in the cytoplasm compared to the nucleus.

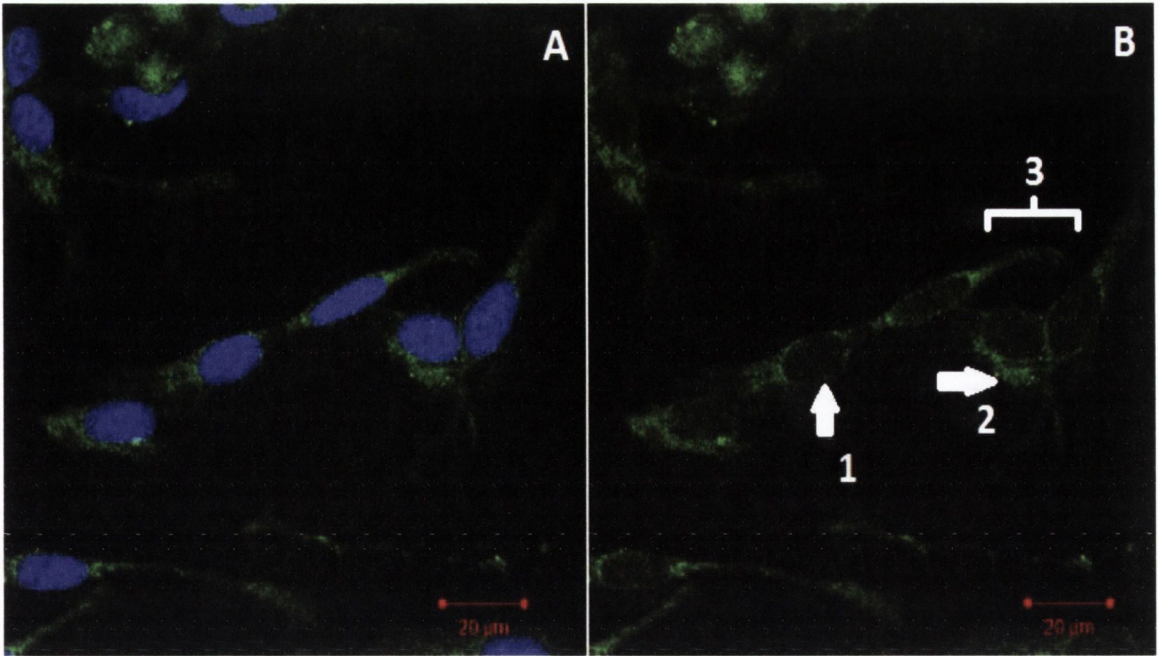


Figure 4.3 Localisation of *MACROD2* in SH-SY5Y neuroblastoma cell lines, visualised using confocal microscopy. Panel A: DAPI staining of the nuclei in blue, with *MACROD2* expression overlaid in green. Panel B: *MACROD2* expression is detected in the nucleus (1), cytoplasm (2) and neurites (3).

4.3.2 Interaction

4.3.2.1 Isolation of *MACROD2* interacting proteins from SH-SY5Y cell lines using Co-Immunoprecipitation

Co-immunoprecipitation was used to extract *MACROD2* and any binding partners from SH-SY5Y cell lysate. To verify the specificity of each Co-IP experiment I performed a western blot using the *MACROD2* antibody. Figure 4.4 illustrates the Ponceau staining of SH-SY5Y protein extracts on a western blot membrane (Panel A) and the corresponding binding of the *MACROD2* antibody (Panel B). For the Ponceau stain, lanes 1, 5 and 7 represent protein extract from the whole cell, cytoplasmic fraction and nuclear fraction respectively. Lanes 2-4, 6 and 8 corresponding to the Co-IP from each protein extract do not show measurable Ponceau staining. This was anticipated and reflects low levels of protein from Co-IP pulldown.

The Western Blot with the *MACROD2* antibody highlights protein across all lanes; with enrichment for proteins at the 50 kDa range (highlighted on Figure 4.4 by a black border), the molecular weight of *MACROD2*. The large bands visible at ~75 and ~25 kDa for the Co-IP samples, may represent enrichment of protein pulldown for proteins at this molecular weight range or poor specificity of the antibody-coated magnetic beads which affected the protein pulldown. Note that these large bands are not visible in the lanes representing protein extract. This western blot confirms the presence of proteins at a molecular weight range that corresponds with that of *MACROD2* and thus these samples were brought forward for LC/ESI-MS analysis.

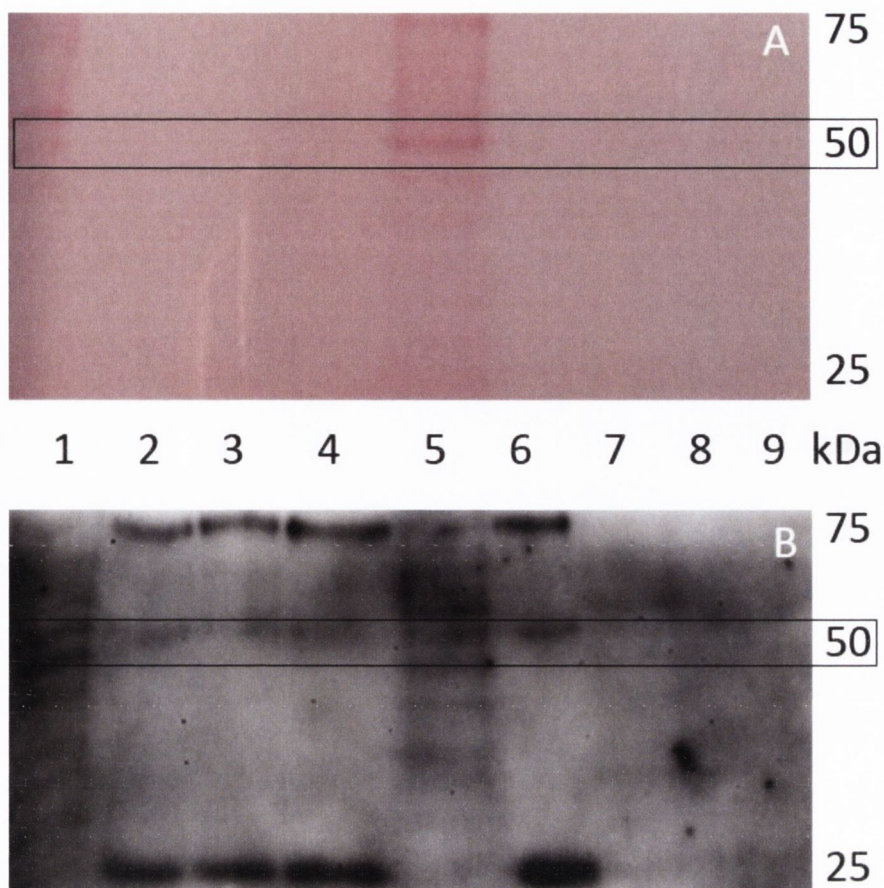


Figure 4.4 Western blot of *MACROD2* Co-IP from SH-SY5Y. Ponceau Stain (Panel A), *MACROD2* Antibody Staining (Panel B). Lanes 1 SH-SY5Y whole cell protein extract; Lanes 2-4 SH-SY5Y whole cell Co-IP (for 1 hour, 2 hours and 3 hours time-points); lane 5 SH-SY5Y cytoplasmic protein extract; Lane 6 SH-SY5Y cytoplasmic Co-IP; Lane 7 SH-SY5Y nuclear protein extract; Lane 8 Nuclear Co-IP; Lane 9 No protein negative control. *MACROD2* is a 50 kDa protein; the 50kDa molecular weight range is highlighted with a black border. Lanes 1 and 5 are clearly positive for protein content from the Ponceau stain; Lanes 2-4 and 6-8 appear to be negative for protein content. However, looking at the Western Blot image for Lanes 2-4 and 6-8, the Co-IP has pulled down protein at 50 kDa, with additional strong bands at 75 and 25 kDa. Additional bands are expected for Co-IP samples. Note that in lanes 1 and 5 there is smearing of protein bands; this could represent poor specificity of the *MACROD2* antibody or fragmentation of protein from the cell lysis and protein extraction protocol.

4.3.2.2 Identification of interacting proteins from LC/ESI-MS analysis of Co-IP pull-down

The Co-IP pulldown was visualised using SDS-PAGE and silver staining by Protea Bio, and an image of the stain was sent to me to make decisions about how to proceed (Figure 4.5). As discussed in Section 4.2.2.4, I decided upon grouping of bands from defined molecular weight ranges (Labelled as A to F in Figure 4.5) to be analysed together in an attempt to harvest the maximum amount of material. The set of bands labelled as 'D', which corresponds with the molecular weight range of *MACROD2*, contained the most protein material based on the Silver stain.

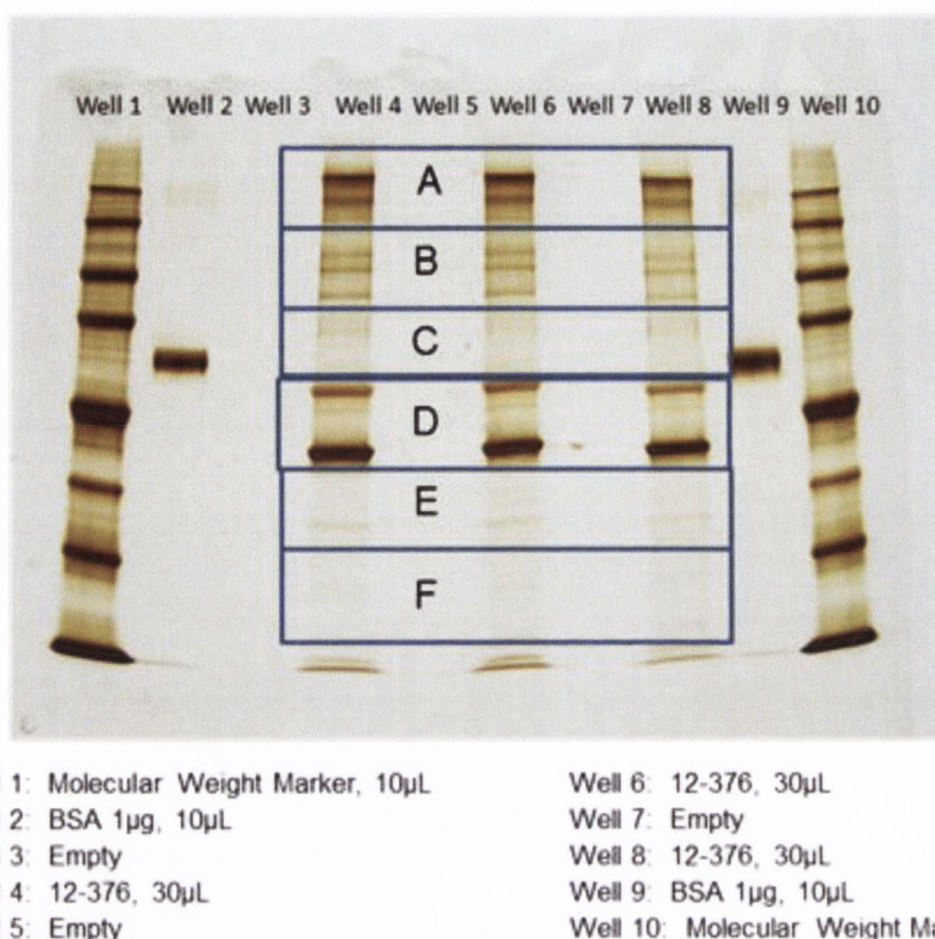


Figure 4.5 Silver staining of whole cell Co-IP from SH-SY5Y cell lines for *MACROD2*. The most protein is present in section D, as indicated by the strength of the stain. Bands of different molecular weight are expected to be detected for a Co-IP pulldown, representing different proteins that are bound to the target protein, *MACROD2*. 12-376 is the identifier code used by Protea Bio for the SH-SY5Y Co-IP sample.

LC/ESI-MS analysis was performed for several *MACROD2* Co-IP samples in an effort to identify protein-protein interactions between *MACROD2* and any partner proteins. Each Co-IP sample was trypsinised to break down the protein(s) in the sample into simpler peptide fragments. It is these fragments that were identified via the LC/ESI-MS analysis, which were then annotated with protein IDs based on similarity to known protein chains from the UniProt database of human proteins. The list of proteins was large, with peptides recognised at varying levels of confidence. Therefore I decided upon a significance threshold of at least one representative peptide, identified at 95% confidence by ProteaBio, for each protein partner. Four samples were analysed in total, across two separate Co-IP experiments:

Co-IP pulldown 1:

- One sample representative of the whole cell (~1.2mg/mL). This was the primary sample for analysis.

Co-IP pulldown 2:

- One sample representative of the whole cell (~0.1mg/mL).
- One sample representative of the cytoplasm (~0.15mg/mL).
- One sample representative of the nucleus (~0.2mg/mL).

These samples for the second Co-IP pulldown were secondary samples to support the primary analysis. The results for the LC/ESI-MS analysis for each Co-IP sample are presented below. Due to shared peptide chains among proteins from the same protein family, the LC/ESI-MS analysis often returned multiple 'hits' for a single observed peptide. Thus the results reported below are an abbreviated list of the identified proteins, whereby one representative protein from each protein family is reported. Due to the lower protein content of the samples from Co-IP pulldown 2, there were fewer peptides identified from LC/ESI-MS for these samples.

4.3.2.2.1 Results of Co-IP pulldown 1:

The LC/ESI-MS analysis of Co-IP pulldown 1, which was derived from an SH-SY5Y whole cell protein extract identified 168 peptides that tagged to genes with over 95% confidence. When duplicates on this list were removed, the list contained 83 genes (Table 4.5 overleaf). The genes identified from this analysis are dominated by proteins that interact with actin, enzymes with hydrolase or phosphorylase activity and signal transduction proteins. Subsets of these genes have functions in transcriptional regulation/RNA binding and chromatin formation.

Table 4.5 Top hits from the LC/ESI-MS analysis of Co-IP pulldown 1, from SH-SY5Y whole cell protein extract.

Gene name	Full name	Protein Function
<i>ACTB</i>	Beta actin	Cell structure and motility
<i>ACTG1</i>	Actin, gamma 1	Cell motility
<i>ACTN1</i>	Alpha actinin	Actin binding
<i>CAPZA1, CAPZA2, CAPZB</i>	F-actin capping protein	Actin filament regulation
<i>CORO1C</i>	Coronin, actin binding protein, 1C	Signal transduction, gene regulation.
<i>CTNNA1</i>	Catenin (cadherin-associated protein), alpha 1	Interacts with cadherins.
<i>CTSA</i>	Cathepsin A	Carboxypeptidase
<i>DBN1</i>	Drebrin	Actin-binding, neuronal growth (Shirao 1995)
<i>FLNA, FLNB, FLNC, FAM101B</i>	Filamin	Actin binding and regulation
<i>GLB1</i>	Galactosidase, beta 1	Beta galactosidase
<i>GNA12, GNA13, GNAI1, GNAI2, GNAI3, GNAL, GNAO1, GNAT1, GNAT2, GNAT3, GNAS</i>	Guanine nucleotide-binding proteins	Signal transduction
<i>HIST1H1A, HIST1H1B, HIST1H1C, HIST1H1D, HIST1H1E, HIST1H1T</i>	Histone H1.1	Chromatin compaction
<i>HNRNPC, LOC440563, LOC649330</i>	Heterogenous nuclear ribonuclearproteins	RNA binding, pre-mRNA processing (König <i>et al.</i> 2010)
<i>HSPA8</i>	Heat shock 70 kDa protein 8	ATPase, vesicle disassembly, protein folding.
<i>KRT1, KRT10, KRT13, KRT14, KRT15, KRT16, KRT2, KRT5, KRT6A, KRT6B, KRT6C, KRT75, KRT79, KRT9</i>	Keratin	Cytoskeleton
<i>LGALS8</i>	Galectin-8	Galactosidase binding, cell adhesion, apoptosis
<i>LIMA1</i>	LIM domain and actin-binding protein	Inhibition of actin filament depolymerisation
<i>MYH10, MYH9, MYL6, MYL6B, MYO1B</i>	Myosin	Actin binding, actin-based motility, ATP hydrolysis
<i>NEFL</i>	Neurofilament light polypeptide	Interacts with MAP2 (Frappier <i>et al.</i> 1991)
<i>NES</i>	Nestin	Radial growth of axons, co-assembles with vimentin
<i>PLEC</i>	Plectin	Interaction between intermediate filaments and microtubules, interacts with spectrin and vimentin
<i>PPP1CA, PPP1CB, PPP1CC, PPP1R9B</i>	Protein phosphatase 1/Spinophilin/Neurabin-2	Protein phosphorylation, regulation of formation and function of dendritic spines (Feng <i>et al.</i> 2000), suppressor of learning and memory (Genoux <i>et al.</i> 2002)
<i>SPTAN1, SPTBN1, SPTBN2</i>	Spectrin alpha chain	Actin crosslinking, epileptic encephalopathy (Mastrangelo & Leuzzi 2012)
<i>SVIL</i>	Supervillin	Actin crosslinking
<i>SYNPO2</i>	Synaptopodin-2	Regulation of actin bundling and dendritic spine formation (Asanuma <i>et al.</i> 2005)
<i>TJP1</i>	Tight junction protein 1	Signal transduction
<i>TMOD1, TPM1, TPM3, TPM4</i>	Tropomodulin/Tropomyosin	Regulation of actin filament length

VIM	Vimentin	Intermediate filament, astrocyte development
XRCC6	Ku70	Non-homologous end joining of DNA repair, V(D)J recombination, telomere length maintenance and subtelomeric gene silencing
ZKSCAN5	Zinc finger protein with KRAB and SCAN domains 5	Transcriptional regulation

4.3.2.2.2 Results of Co-IP Pulldown 2 (Whole Cell):

Table 4.6 Top hits from the LC/ESI-MS analysis of CoIP pulldown 2, from SH-SY5Y whole cell protein extract.

Gene name	Full name	Protein Function
CKAP5	Cytoskeleton associated protein 5	Microtubule plus end dynamics (Barr & Gergely 2008)
USP33	Ubiquitin specific peptidase 33	Carboxyl-terminal hydrolase
ZNF101	Zinc finger protein 101	Transcriptional regulation (Bellefroid <i>et al.</i> 1993)

LC/ESI-MS analysis of Co-IP pulldown 2, derived from an SH-SY5Y whole cell protein extract identified 3 peptides that tagged to genes with over 95% confidence. The genes identified from this analysis have functions in cytoskeletal dynamics, hydrolase activity and transcriptional regulation.

4.3.2.2.3 Results of Co-IP Pulldown 2 (Cytoplasmic fraction):

Table 4.7 Top hits from the LC/ESI-MS analysis of CoIP pulldown 2, from the cytoplasmic fraction of SH-SY5Y protein extract.

Gene name	Full name	Protein Function
AKAP2	A-kinase anchor protein 2	Protein kinase compartmentalisation (Alto <i>et al.</i> 2003)
DNAH8	Dynein, axonemal, heavy chain 8	Cytoskeletal dynamics, ATPase.

LC/ESI-MS analysis of Co-IP pulldown 2, derived from the cytoplasmic fraction of SH-SY5Y protein extract identified 2 peptides that tagged to genes with over 95% confidence. The genes identified from this analysis have functions in compartmentalisation of cAMP protein kinase and ATPase dependent cytoskeletal dynamics.

4.3.2.2.4 Results of Co-IP Pulldown 2 (Nuclear Fraction):

Table 4.8 Top hit from the LC/ESI-MS analysis of CoIP pulldown 2, from the nuclear fraction of SH-SY5Y protein extract.

Gene name	Full name	Protein Function
HIST1H1A, HIST1H1B, HIST1H1C, HIST1H1D, HIST1H1E, HIST1H1T	Histone H1.1	Chromatin compaction

The LC/ESI-MS analysis of Co-IP pulldown 2, which was derived from the nuclear fraction of SH-SY5Y protein extract, identified 1 peptide that mapped to histone H1.1, a protein crucial to chromatin formation.

4.3.2.3 Results of Gene Ontology Enrichment Analysis using DAVID:

Of the 93 genes identified from the LC/ESI-MS analysis, DAVID used 78 for gene ontology enrichment analysis. DAVID reports a p-value for the significance of the enrichment. A list of gene ontology terms with a p-value <0.05 is reported in Table 4.9. The most significant gene ontology terms are all related to either the cytoskeleton and actin binding, or GTPase or GTP binding activity.

Table 4.9 Analysis of genes identified from the LC/ESI-MS were analysed using DAVID. The most significant Gene Ontology terms are reported here. Count refers to the number of genes from the input that are associated with a given GO term.

Term	Count	P Value
GO:0005200~structural constituent of cytoskeleton	17	2.27E-21
GO:0003779~actin binding	24	4.90E-19
GO:0008092~cytoskeletal protein binding	27	3.74E-18
GO:0005198~structural molecule activity	25	1.05E-13
GO:0051015~actin filament binding	7	6.57E-07
GO:0003924~GTPase activity	10	4.13E-06
GO:0005525~GTP binding	11	6.56E-05
GO:0032561~guanyl ribonucleotide binding	11	8.18E-05
GO:0019001~guanyl nucleotide binding	11	8.18E-05
GO:0032555~purine ribonudeotide binding	23	7.38E-04
GO:0032553~ribonucleotide binding	23	7.38E-04
GO:0030898~actin-dependent ATPase activity	3	9.50E-04
GO:0017076~purine nudeotide binding	23	0.001332
GO:0003774~motor activity	6	0.001533
GO:0019901~protein kinase binding	6	0.001785
GO:0008307~structural constituent of musde	4	0.00195
GO:0019900~kinase binding	6	0.00418
GO:0000166~nucleotide binding	24	0.004501
GO:0043531~ADP binding	3	0.005561
GO:0005516~calmodulin binding	5	0.009548
GO:0031852~mu-type opioid receptor binding	2	0.011827
GO:0016887~ATPase activity	7	0.014233
GO:0031821~metabotropic serotonin receptor binding	2	0.017689
GO:0031628~opioid receptor binding	2	0.017689
GO:0032403~protein complex binding	5	0.029022
GO:0004722~protein serine/threonine phosphatase activity	3	0.029161
GO:0019899~enzyme binding	8	0.035167
GO:0017166~vinculin binding	2	0.040794
GO:0050780~dopamine receptor binding	2	0.052144

4.3.3 siRNA knockdown of *MACROD2* in SH-SY5Y cell lines: A preliminary study

This preliminary study was designed to investigate the impact of knockdown of *MACROD2* using siRNA upon overall cell morphology in the differentiated SH-SY5Y neuronal cell model. The success of the knockdown was assessed at the mRNA level, using real time PCR, and at the protein level, by visualising *MACROD2* expression using immunohistochemistry.

4.3.3.1 Results of siRNA knockdown of *MACROD2* - confirmation of mRNA knockdown by RT-PCR:

To assess the knockdown of the *MACROD2* mRNA in SH-SY5Y cells, the expression of *MACROD2* was measured using real time PCR and five previously designed Taqman assays (SNP1 - SNP5, as described in Chapter 3). The fold change in gene expression is the difference in gene expression across different concentrations of *MACROD2* siRNA (5nM, 10nM, 20nM, 35nM and 50nM) and control siRNAs (Negative control non-targeting 'scrambled' siRNA and positive control siRNA against *PP1B*) compared to untreated cells, normalised against an internal control (GAPDH expression). Figure 4.6 to 4.10 record the fold change in expression across the five Taqman assays used. However, each Taqman assay was used with the same RNA starting material and thus do not represent independent experiments. Please note the different scale across each figure.

The siRNA knockdown of *MACROD2* at the mRNA level appears to have been successful, although affectivity of the siRNA does not appear to correlate with increased concentration. From this preliminary set of experiments it would appear that *MACROD2* can be knocked down at the mRNA level using this siRNA treatment, with a concentration of 5nM being the most effective.

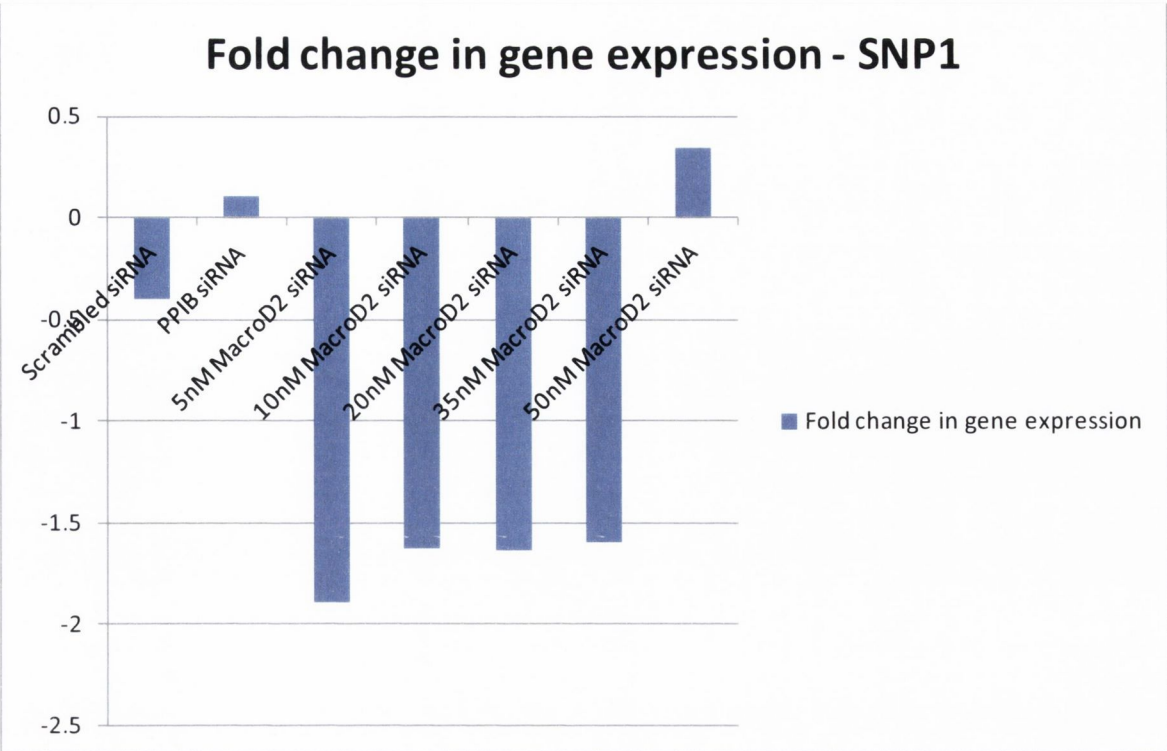


Figure 4.6 The fold change in gene expression for SNP1 (rs16997016). For this series of figures (4.6 to 4.10), all values have been normalised against endogenous *Gapdh* expression, and the non-targeting 'scrambled' siRNA can be seen as a measure of the effect the transfection process has upon the expression of *MACROD2*. Here the control siRNAs, 'scrambled' and *PPIB* demonstrate a small fold change in gene expression. However, the majority of the *MACROD2* siRNA display over 1.5 fold decrease in gene expression. However, the 50nM concentration of *MACROD2* siRNA in this case appears to have induced an increase in gene expression compared to untreated cells.

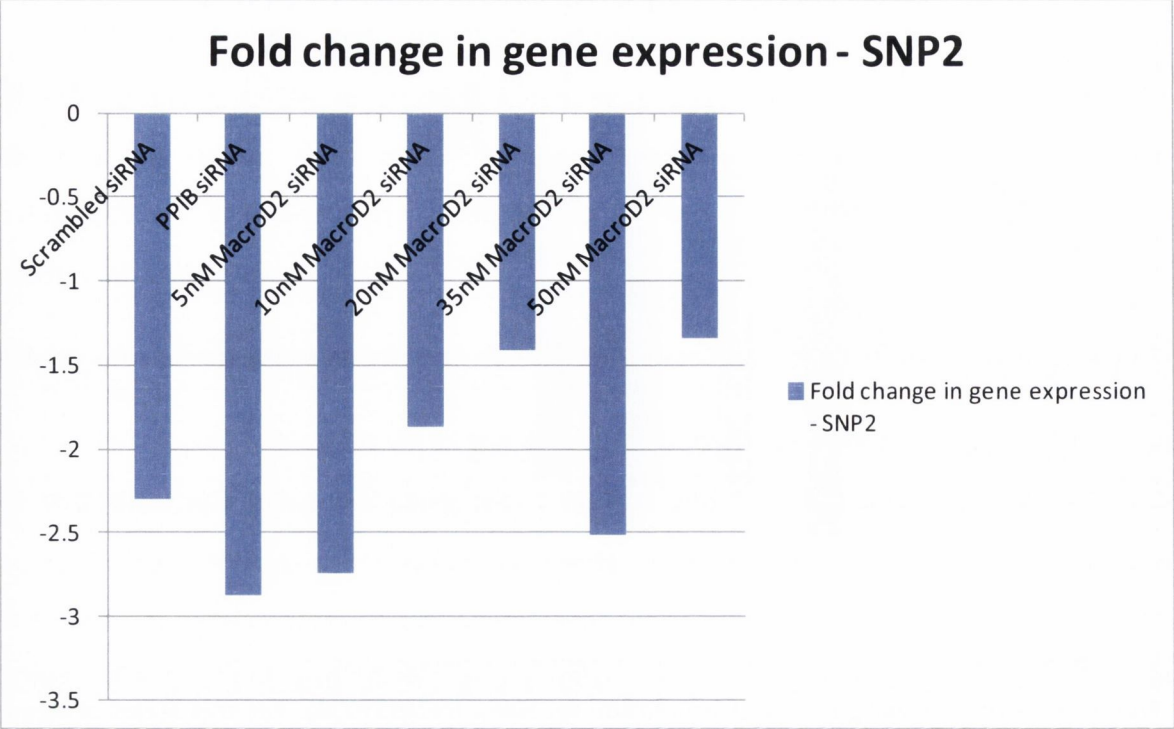


Figure 4.7 The fold change in gene expression for SNP2 (rs11696319). For this experiment, the control siRNAs and all 5 concentrations of *MACROD2* have resulted in reduced expression of *MACROD2*, thus regardless of treatment, the transfection process has affected gene expression.

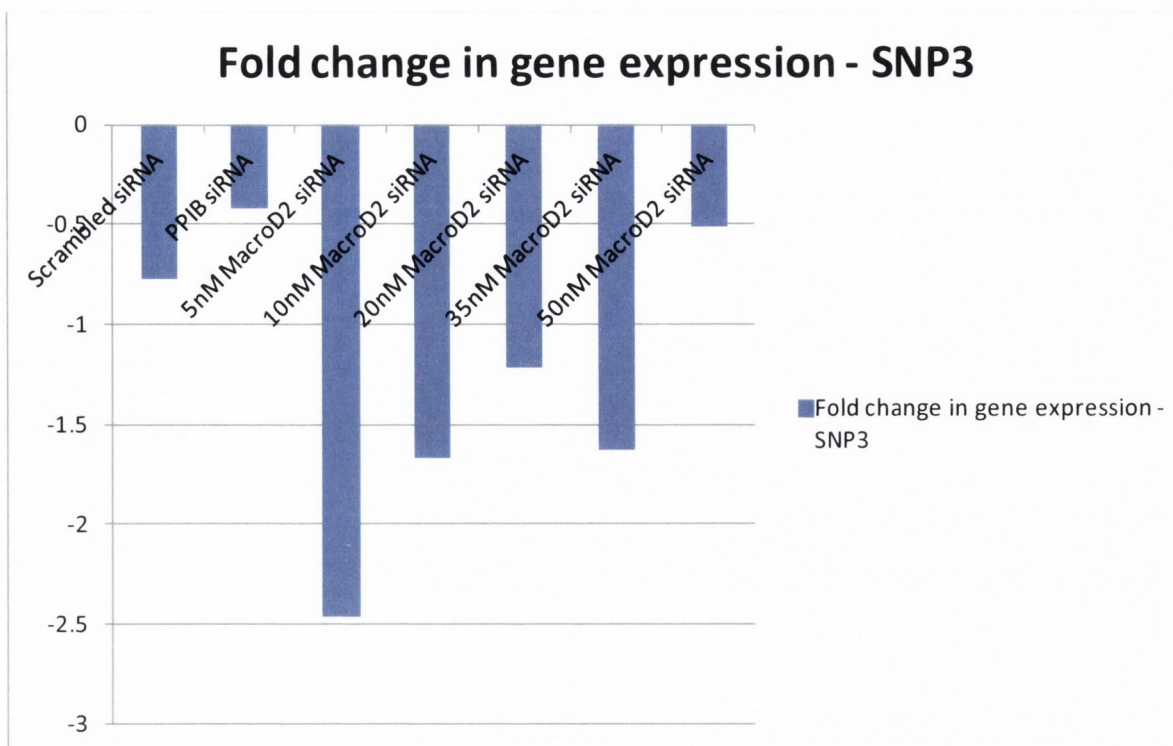


Figure 4.8 The fold change in gene expression for SNP3 (rs1238278). The control siRNAs, 'scrambled' and *PPIB* demonstrate a small fold decrease in gene expression. The *MACROD2* siRNA have caused a decrease in gene expression, with the 5nM concentration being the most effective.

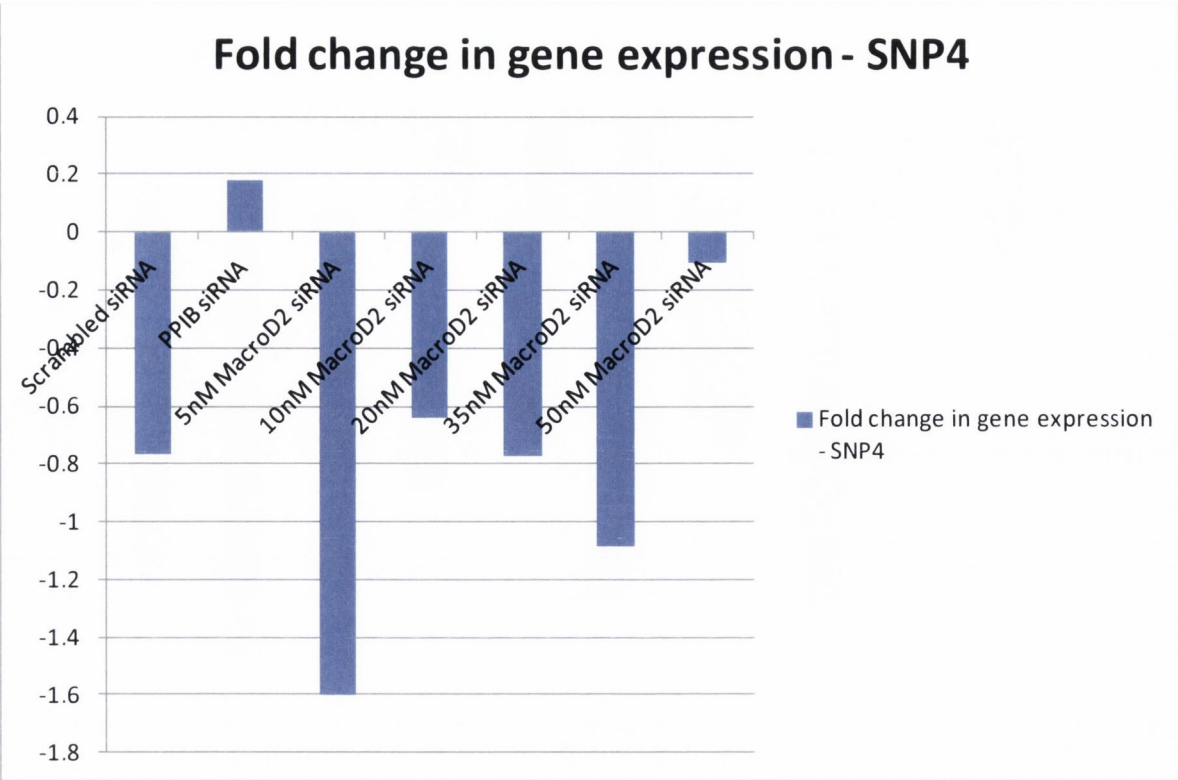


Figure 4.9 The fold change in gene expression for SNP4 (rs1225891). The control siRNAs, 'scrambled' and *PPIB* demonstrate changes in gene expression of an increase and decrease respectively. The siRNAs against *MACROD2* have all resulted in a decrease in gene expression. The 5nM concentration of *MACROD2* siRNA is the most effective siRNA treatment.



Figure 4.10 The fold change in gene expression for SNP5 (rs6080109). Here the control siRNAs, ‘scrambled’ and *PPIB* have caused a decrease in gene expression. The siRNA knockdown of *MACROD2* has been successful, with up to a 2.5 fold decrease in gene expression. The 5nM concentration was the most effective treatment for *MACROD2* knockdown.

4.3.3.2 Results of siRNA knockdown of *MACROD2* - confirmation of protein knockdown by immunohistochemistry and effects of knockdown on neuronal marker expression:

To assess the knockdown of the *MACROD2* protein in SH-SY5Y cells, siRNA treated SH-SY5Y cells in culture were visualised using immunohistochemistry. The effect of *MACROD2* knockdown on neuronal marker expression and cell morphology was also assessed. The *MACROD2* siRNA was tested at five concentrations: 5nM, 10nM, 20nM, 35nM and 50nM. The results of these experiments are illustrated in Figure 4.11 to 4.18. For analysis, all treated cells are compared against untreated cells, and are assessed based on overall cell morphology and intensity of the protein stains. For the 'scrambled' siRNA and positive control siRNA against *PPIB* treated cells, the staining of *MACROD2*, MAP2 and Synapsin is similar in intensity to the untreated cells, indicating that expression of these proteins has not been affected. The cell morphology is normal when compared to the untreated cells and features such as neurite outgrowth are clearly visible.

For the 5nM *MACROD2* siRNA treated cells, the staining of *MACROD2*, MAP2 and Synapsin is reduced in intensity compared to the untreated cells, indicating that expression of these proteins has been reduced. The cell morphology is abnormal when compared to the untreated cells, with no apparent neurite outgrowth. This is in contrast to cells treated with 10nM *MACROD2* siRNA. The staining of *MACROD2*, MAP2 and Synapsin is similar in intensity to the untreated cells, indicating that expression of these proteins has not been affected. The cell morphology is normal when compared to the untreated cells and features such as neurite outgrowth are clearly visible. The cells that have been treated with *MACROD2* siRNA at concentrations of 20nM, 35nM and 50nM all display a similar staining pattern and cell morphology. The intensity of the stain for all three proteins is reduced, indicating reduced expression of these proteins due to the siRNA treatment. The cell morphology is also abnormal when compared to the untreated cells, with reduced cytoplasmic volume, although neurite outgrowth has been maintained. From these results it would appear that the 5nM *MACROD2* siRNA treatment has been the most effective at knockdown of *MACROD2*. Compared to all the other treatments, the 5nM *MACROD2* siRNA treatment has had the greatest effect upon cell morphology and expression of the neuronal markers MAP2 and Synapsin.

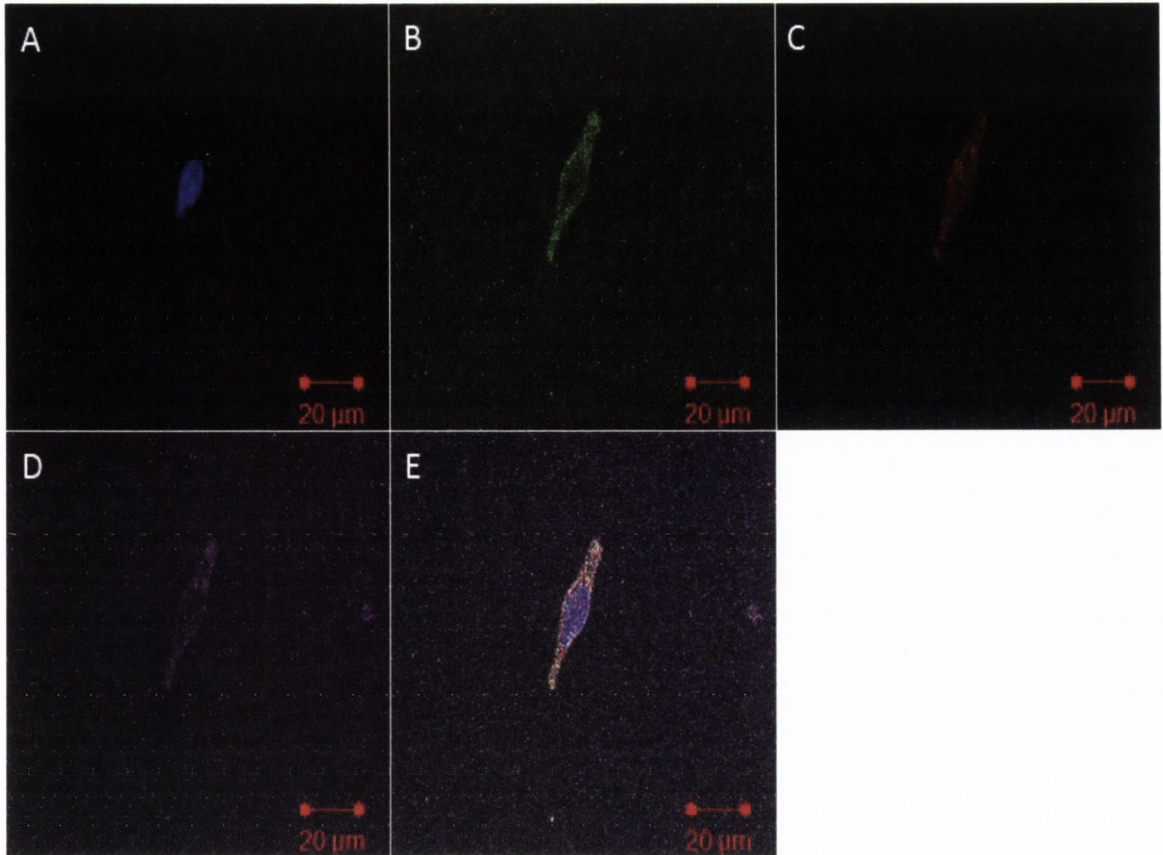


Figure 4.11 Expression of *MACROD2* and neuronal markers in untreated SH-SY5Y cell lines. Panel A: DAPI staining of the nucleus in blue; Panel B: *MACROD2* is stained in green; Panel C: Synapsin expression is stained in red; Panel D: MAP2 expression is stained in purple, Panel E: Overlaid staining of Panels A-D. Expression of all three proteins is clearly visible, indicating healthy neuron-like cells.

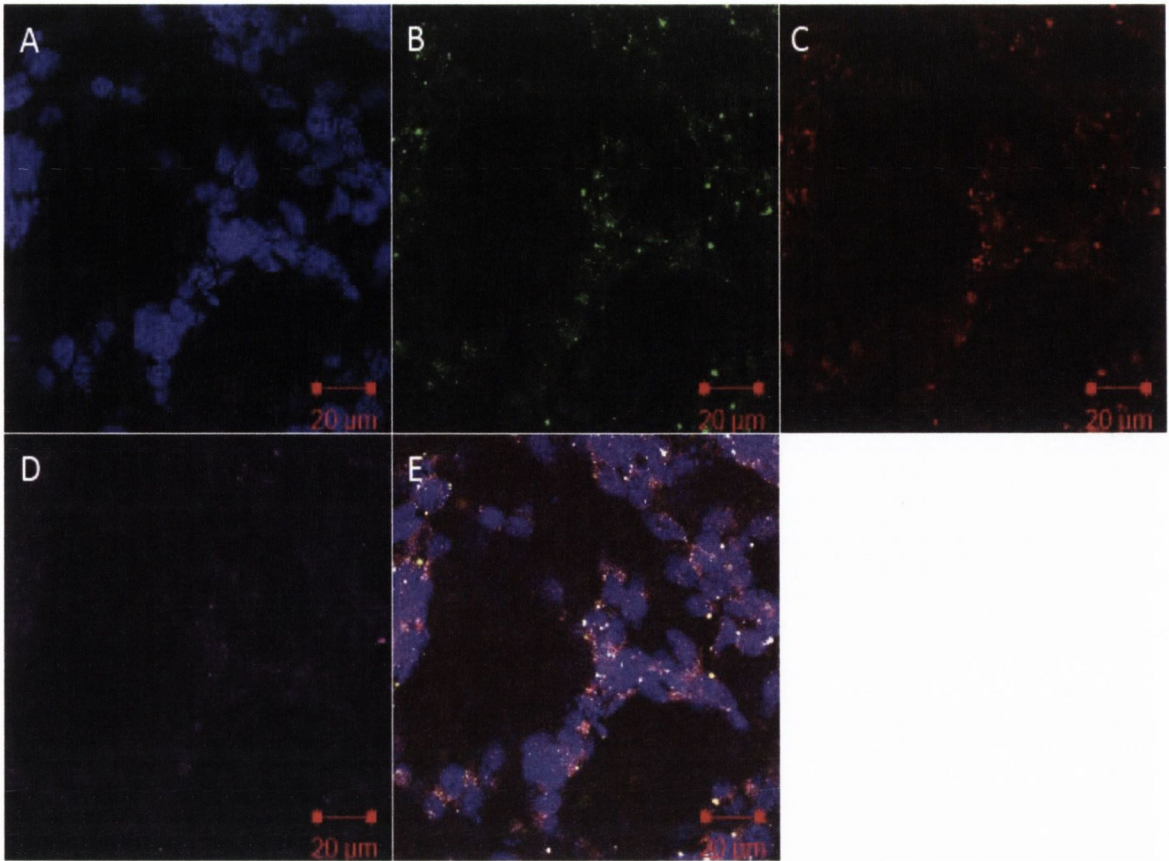


Figure 4.12 Expression of *MACROD2* and neuronal markers in SH-SY5Y cell lines treated with 'scrambled' siRNA (Negative control). Panel A: DAPI staining of the nucleus in blue; Panel B: *MACROD2* is stained in green; Panel C: Synapsin expression is stained in red; Panel D: MAP2 expression is stained in purple, Panel E: Overlaid staining of Panels A-D. Expression of all three proteins is clearly visible and appears unaffected by the siRNA transfection.

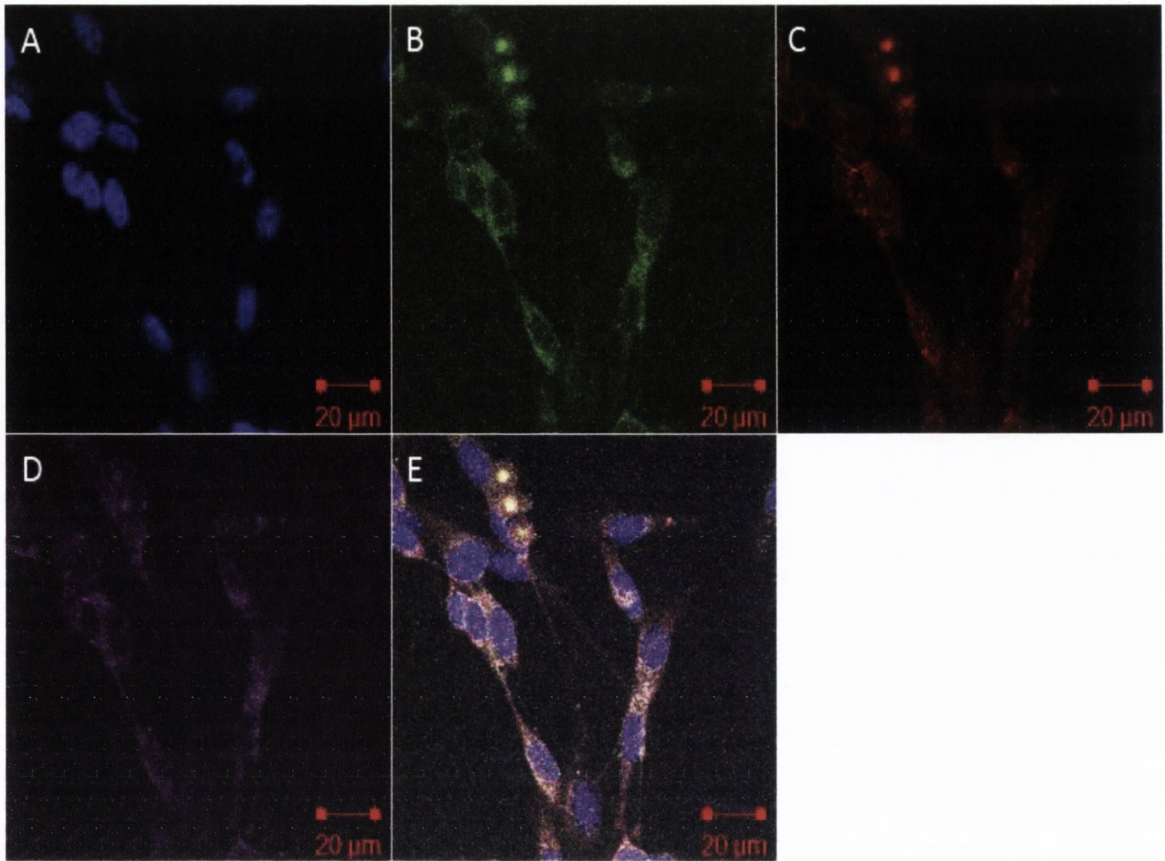


Figure 4.13 Expression of *MACROD2* and neuronal markers in SH-SY5Y cell lines treated with siRNA against *PPIB* (Positive control). Panel A: DAPI staining of the nucleus in blue; Panel B: *MACROD2* is stained in green; Panel C: Synapsin expression is stained in red; Panel D: MAP2 expression is stained in purple, Panel E: Overlaid staining of Panels A-D. Expression of all three proteins is clearly visible and appears unaffected by the siRNA transfection.

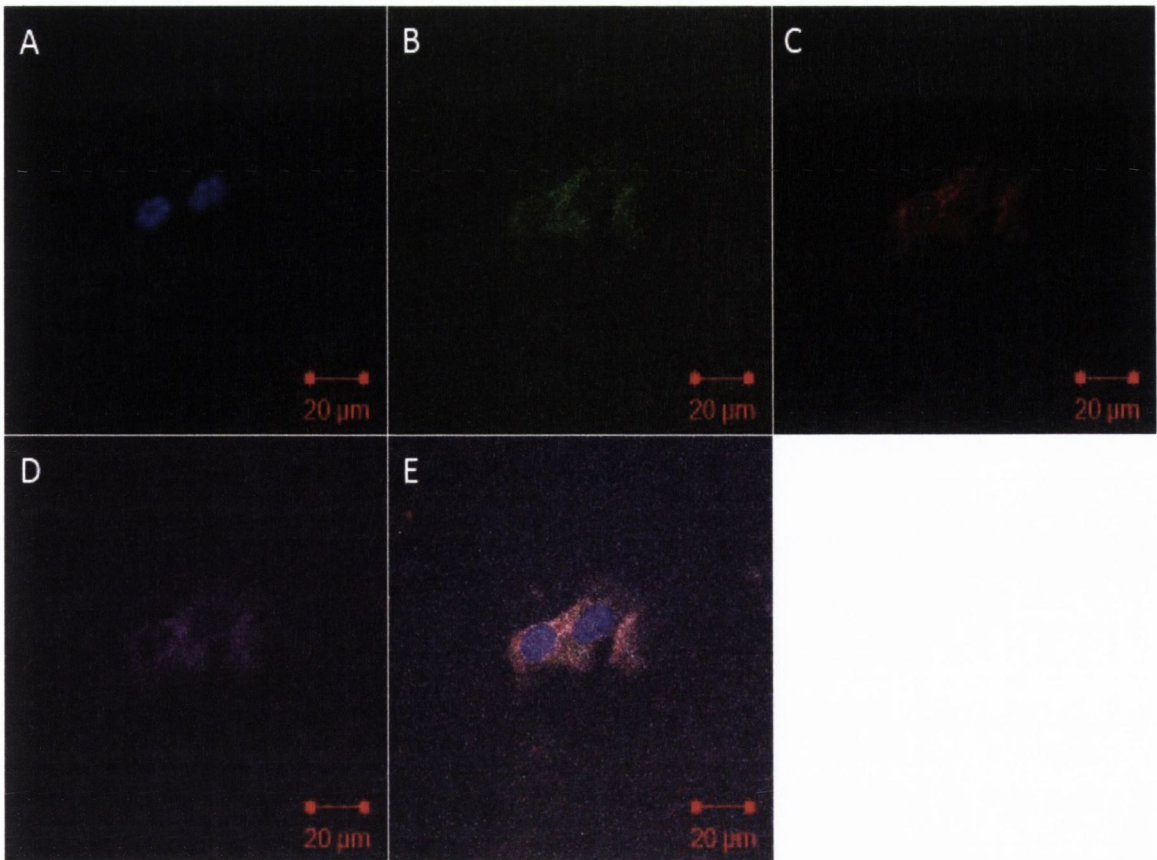


Figure 4.14 Expression of *MACROD2* and neuronal markers in SH-SY5Y cell lines treated with 5nM siRNA against *MACROD2*. Panel A: DAPI staining of the nucleus in blue; Panel B: *MACROD2* is stained in green; Panel C: Synapsin expression is stained in red; Panel D: MAP2 expression is stained in purple, Panel E: Overlaid staining of Panels A-D. The cell morphology appears affected compared to the untreated cells, with reduced neurite outgrowth. The intensity of the stain, an indicator of protein expression, is not as strong compared to the untreated cells.

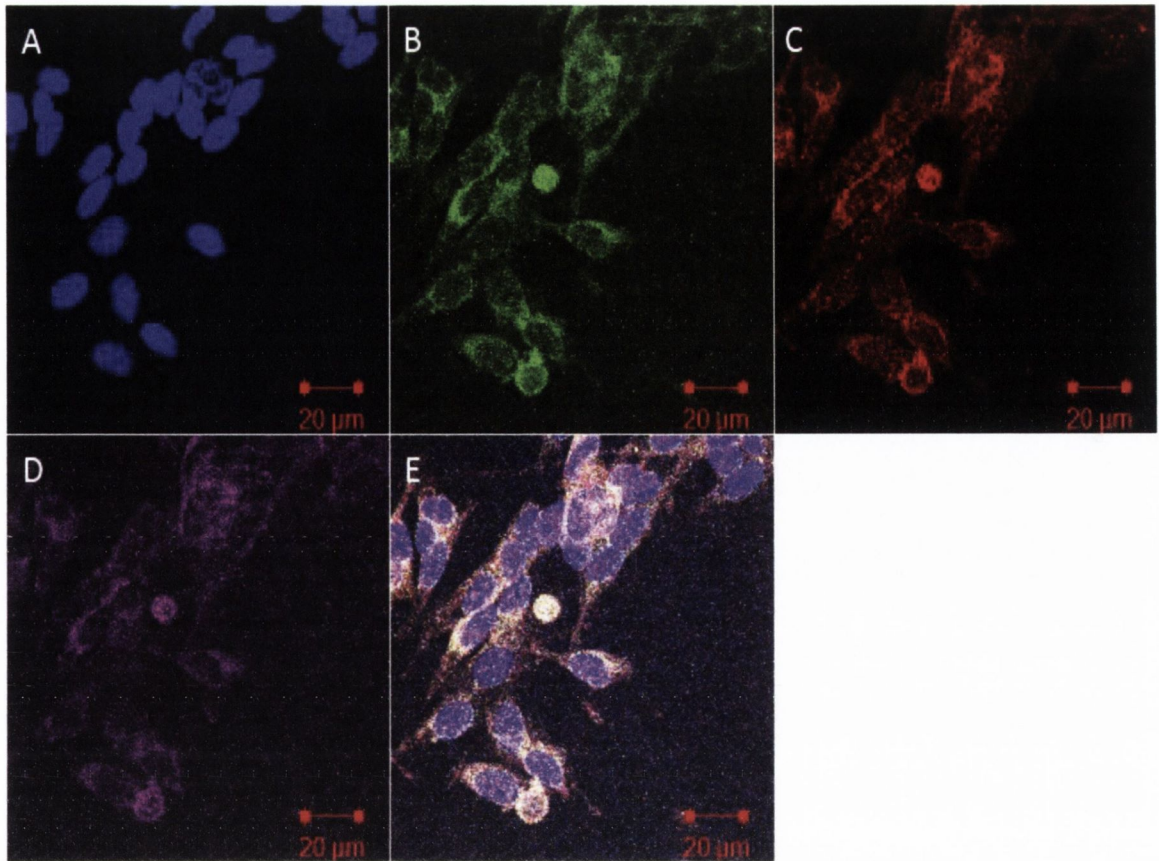


Figure 4.15 Expression of *MACROD2* and neuronal markers in SH-SY5Y cell lines treated with 10nM siRNA against *MACROD2*. Panel A: DAPI staining of the nucleus in blue; Panel B: *MACROD2* is stained in green; Panel C: Synapsin expression is stained in red; Panel D: MAP2 expression is stained in purple, Panel E: Overlaid staining of Panels A-D. Expression of all three proteins is clearly visible, and appears unaffected by *MACROD2* siRNA transfection. Cell morphology appears normal when compared to untreated cells.

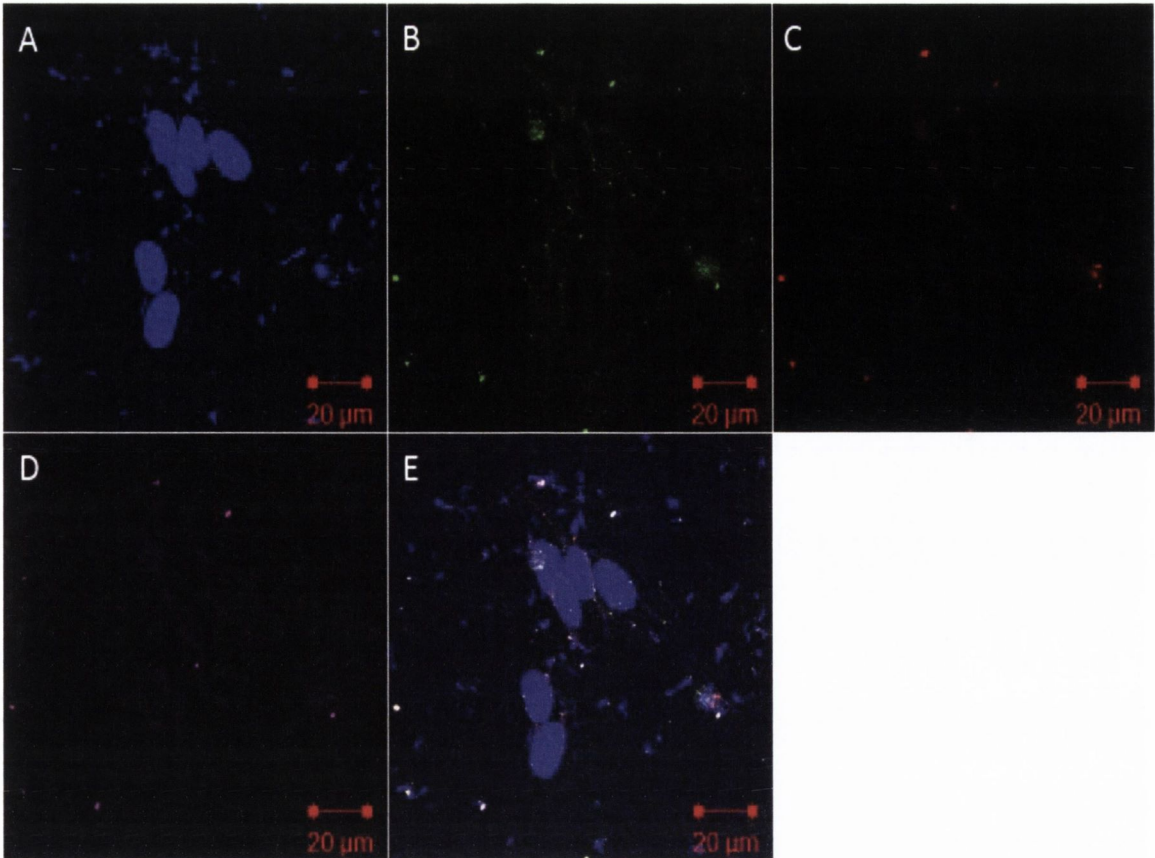


Figure 4.16 Expression of *MACROD2* and neuronal markers in SH-SY5Y cell lines treated with 20nM siRNA against *MACROD2*. Panel A: DAPI staining of the nucleus in blue; Panel B: *MACROD2* is stained in green; Panel C: Synapsin expression is stained in red; Panel D: MAP2 expression is stained in purple, Panel E: Overlaid staining of Panels A-D. Expression of all three proteins is visible, however the cell morphology appears affected compared to the untreated cells, with reduced cytoplasmic volume. The intensity of the stain for all three proteins is not as strong compared to the untreated cells, indicating expression of these proteins is reduced due to the *MACROD2* siRNA treatment.

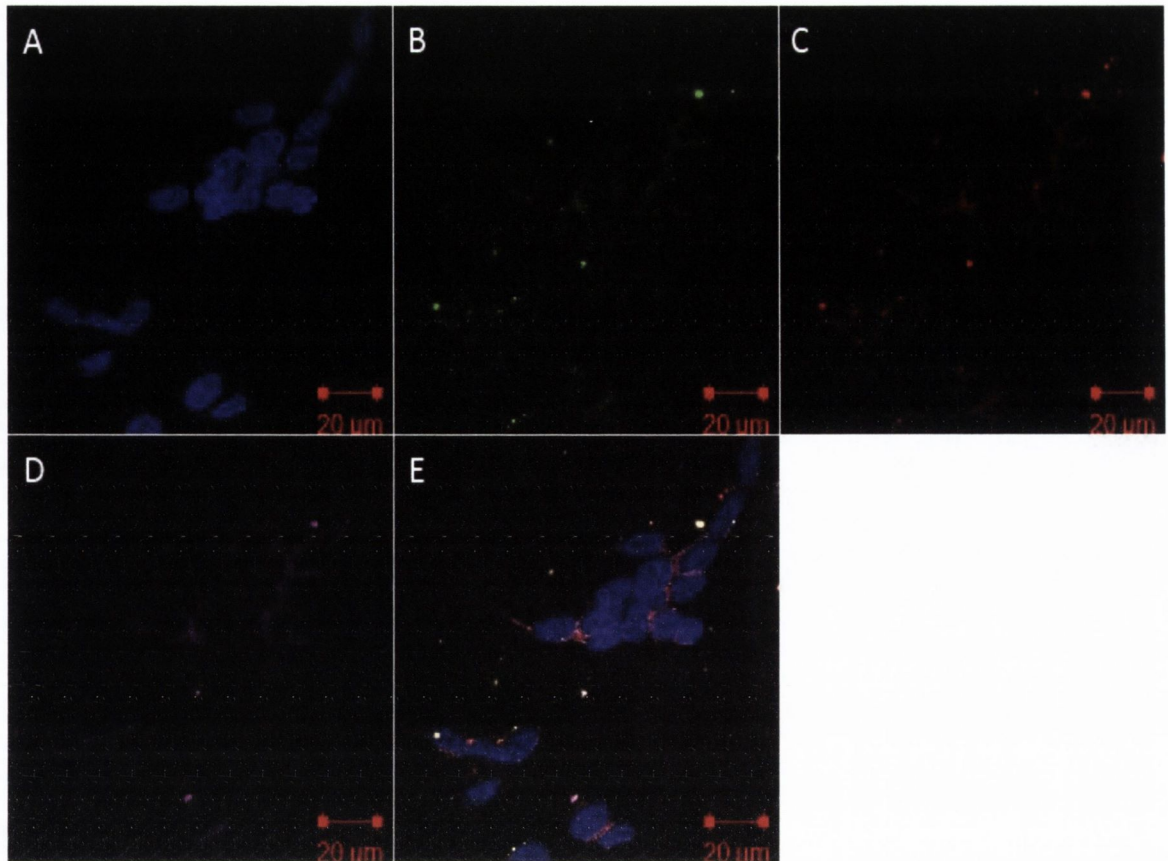


Figure 4.17 Expression of *MACROD2* and neuronal markers in SH-SY5Y cell lines treated with 35nM siRNA against *MACROD2*. Panel A: DAPI staining of the nucleus in blue; Panel B: *MACROD2* is stained in green; Panel C: Synapsin expression is stained in red; Panel D: MAP2 expression is stained in purple, Panel E: Overlaid staining of Panels A-D. Expression of all three proteins is visible, however the cell morphology appears affected compared to the untreated cells, with reduced cytoplasmic volume. The intensity of the stain for all three proteins is not as strong compared to the untreated cells, indicating expression of these proteins is reduced due to the *MACROD2* siRNA treatment.

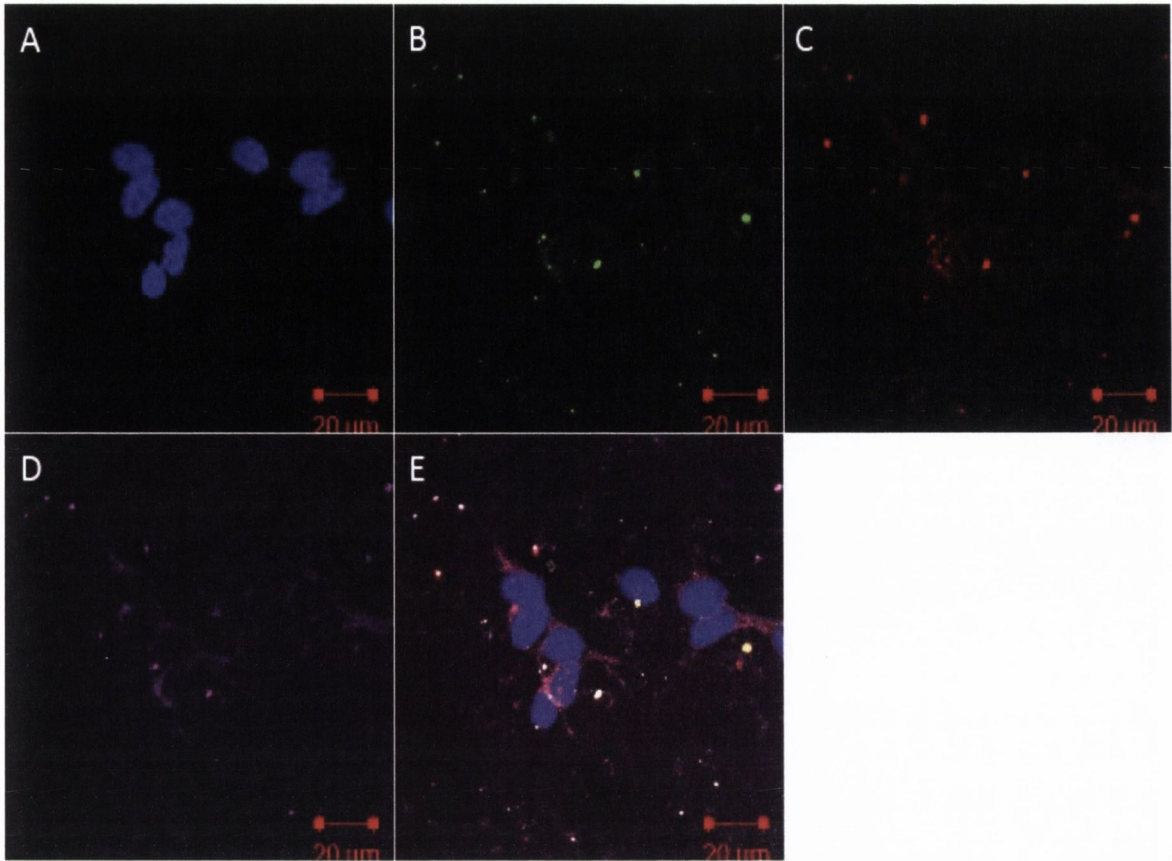


Figure 4.18 Expression of *MACROD2* and neuronal markers in SH-SY5Y cell lines treated with 50nM siRNA against *MACROD2*. Panel A: DAPI staining of the nucleus in blue; Panel B: *MACROD2* is stained in green; Panel C: Synapsin expression is stained in red; Panel D: MAP2 expression is stained in purple, Panel E: Overlaid staining of Panels A-D. Expression of all three proteins is visible, however the cell morphology appears affected compared to the untreated cells, with reduced cytoplasmic volume. The intensity of the stain for all three proteins is not as strong compared to the untreated cells, indicating expression of these proteins is reduced due to the *MACROD2* siRNA treatment.

4.4 Discussion

4.4.1 Localisation

From the data presented here that documents the localisation and expression of endogenous *MACROD2* in a neuronal cell model, I find an expression profile in agreement of that reported for EGFP-tagged *MACROD2* in HeLa cell lines (Neuvonen & Ahola 2009). These results suggest that endogenous *MACROD2* is expressed in both the cytoplasm and the nucleus, with an apparent higher rate of expression in the cytosol. This result represents the first time, to the best of my knowledge, that the localisation of endogenous *MACROD2* in human cells has been identified by immunohistochemistry. This result confirms the hypothesis that endogenous *MACROD2* is localised in the nucleus and cytoplasm. This hypothesis was based upon previous reports, which identified the activity of *MACROD2* in the nucleus following DNA damage (Timinszky *et al.* 2009) and have demonstrated that *MACROD2* is an O-Acetyl-ADP-Ribose Deacetylase and a potential regulator of cellular OAADPR produced by NAD⁺-dependent deacetylation, implicating *MACROD2* and other macro domain containing proteins in sirtuin activity (Chen *et al.* 2011). The cellular localisation of sirtuin proteins is diverse, located within the nucleus, nucleous, cytoplasm and mitochondria (North & Verdin 2004).

4.4.2 Interaction

The failure of the protein analysis to identify *MACROD2* among the proteins identified from the Co-IP pulldown creates an obstacle for interpretation, as one would expect *MACROD2* to be contained within the list of identified genes, particularly due to the observation of bands at 50 kDa, the molecular weight of *MACROD2*, across the silver stain and western blot. Looking at Figure 4.5, the set of bands labelled as 'D', which corresponds with the molecular weight range of *MACROD2*, contained the most protein material based on the Silver stain. This result was as expected as the protein pulldown should have been enriched for *MACROD2*. Whilst it remains possible that *MACROD2* contained within the analysed Co-IP eluate but at a level too low to be detected by LC/ESI-MS analysis, it is a weakness of these results that *MACROD2* was not identified across any of the Co-IP pulldown analyses. Whilst I hypothesised that some of the bands visible on the western blot/silver stain were non-specific, this assertion could not be made without further investigation and re-optimisation of the Co-IP protocol.

However, the results of the Co-IP analysis do fit in with the original hypothesis, that *MACROD2* is associated with histone proteins or sirtuins. Histone proteins were identified across both Co-IP pulldown analyses, and many sirtuins interact with cytoplasmic cytoskeletal proteins (North & Verdin 2004), particularly *SIRT7* (Sirtuin 7). Microtubule assembly is an essential step for neurogenesis and have been demonstrated to regulate brain size (Bond & Woods 2006), which is interesting considering the association of *MACROD2* with brain volume (Kohannim *et al.* 2012). Of particular interest from the LC/ESI-MS analysis is Spectrin alpha chain, which has been associated with early onset epileptic encephalopathies (Mastrangelo & Leuzzi 2012), which as has been discussed in the introduction of this chapter, is a disorder that has also been associated with CNVs in *MACROD2* (Bendavid *et al.* 2009). The identification of Ku70 as an interactor of *MACROD2* is of interest considering its role in non-homologous DNA end repair considering the reports of recruitment of *MACROD2* to sites of DNA damage (Timinszky *et al.* 2009). The observation from the DAVID analysis of enrichment for dopamine receptor binding is also interesting in the context of neuropsychiatric disorders.

In an attempt to investigate putative binding partners across homologs of *MACROD2* in model organisms, I used the online database of known and predicted protein-protein interactions 'STRING' (string-db.org/) to examine documented protein interaction networks across the different model organisms that have entries for *MACROD2* and its homologs. I focused solely on protein interactors with reported co-occurrence or co-expression as there were no protein interactors reported with experimental evidence for interaction, and excluded entries based upon text-mining alone. There are no experimentally validated PPI networks documented for human or murine *MACROD2* at the time of the search (March 2013), however, there are PPI network listings for *C. elegans*, *F. catus* and *A. Carolinensis* documented in Table 4.10. Interesting to note is that Synaptopodin 2 was also identified by the LC/ESI-MS analysis of the sample 'Co-IP pulldown 1, from SH-SY5Y whole cell protein extract', in addition to other histone proteins of which H2AFY is a member. GDAP1 is associated with the same disorder as NEFL (Also identified from 'Co-IP pulldown 1, from SH-SY5Y whole cell protein extract'), Charcot-Marie-Tooth disease, a disorder of the peripheral nervous system. This database search would appear to provide some additional support for the proteins identified by the LC/ESI-MS analysis.

Whilst there are some caveats to interpretation of the LC/ESI-MS analysis, it does appear to give some clues as to how *MACROD2* could have a role in neuropsychiatric conditions and lends credence to the hypothesis that *MACROD2* is associated with histone proteins directly.

Table 4.10 Protein interactors of *MACROD2* homologs identified via the STRING database. Although protein interactors of *MACROD2* have not been documented in humans in online databases, such information is available for model organisms which can supplement the analyses reported herein.

Organism	Homologous gene	Putative protein interactor	Function of protein interactor
<i>C. elegans</i>	B0035.3	Synaptopodin 2	Synaptopodin is an actin-associated protein that may play a role in actin-based cell shape and motility (Mundel et al. 1997).
		Hus1 DNA damage checkpoint protein	DNA damage-induced cell cycle arrest and apoptosis, telomere maintenance, and hence, genome stability and development.
<i>F. catus</i>	<i>MACROD2</i>	GDAP2 ganglioside induced differentiation associated protein 2	No functional annotation exists, although GDAP1 is involved in the signal transduction pathway during neuronal development
		H2AFY H2A histone family, member Y	Chromatin modulation
<i>A. carolinensis</i>	<i>MACROD2</i>	H2AFY H2A histone family, member Y	Chromatin modulation

4.4.3 siRNA knockdown of *MACROD2* in SH-SY5Y cell lines: A preliminary study

Due to the nature of these experiments as a preliminary study, one must hesitate in the interpretation of the results until there are more experimental replicates. However, from the modest set of results reported here, it seems that it is possible to knockdown *MACROD2* using siRNA in the cell model, and that knockdown of *MACROD2* has a noticeable phenotypic impact, in terms of the expression of neuronal markers but also upon overall cell morphology. A caveat to these results is that siRNA treatment is known to be cytotoxic, so the results for concentrations at 20nM, 35nM and 50nM could be put under question. It is possible that the phenotype of these cells is purely due to stress caused by the transfection of siRNA. The 5nM *MACROD2* siRNA appears to have been the most successful treatment however, with reduced expression of *MACROD2* and abnormal cell morphology, including loss of neurite outgrowth. The 5nM treatment is a concentration 5 times lower than that of the positive and negative control siRNAs, which displayed normal cell morphology and neuronal marker expression. This lends credence to the fact that the knockdown of *MACROD2*, and not the siRNA transfection itself, resulted in reduced expression of the neuronal markers MAP2 and Synapsin, and the apparent drastically affected cell morphology.

This work requires expansion in terms of experimental replicates to be able to determine the impact of *MACROD2* knockdown on neurons and neuronal markers, including direct measurement of the protein expression levels and the cells themselves, but these results are promising and indicate that further study of the impact of *MACROD2* knockdown on neuronal biology is warranted.

A search for model organisms with knockouts of the homolog for *MACROD2* (i.e. *MACROD2* in *M. Musculus* and B0035.3 in *C. Elegans*) was performed across different online databases:

- *M. Musculus*: The Jackson Laboratory: jaxmice.jax.org/strain/018570.html, Mouse Genome Informatics: informatics.jax.org/searches/allele_report.cgi?_Marker_key=56879, International Gene Trap Consortium: genetrap.org/cgi-bin/annotation.py?mgi=MGI:1920149 and International Knockout Mouse Consortium: knockoutmouse.org/search_results?criteria=MACROD2

- *C. elegans*: The *C. elegans* Gene Knockout Consortium:
celeganskoconsortium.omrf.org/genedetails.aspx?GeneID=B0035.3).

Although these searches indicated that knockouts for the *MACROD2* homolog are available, any associated phenotypes have not been assessed, including documentation of the viability and fertility of homozygotes for the knockout in mice. It appears these knockouts have been generated as part of recent unpublished work (See <http://www.informatics.jax.org/reference/J:188991> for reference), and thus it is possible in time the mouse knockout at The Jackson Laboratory will be assessed for impact on overall phenotype. Results from the literature for *C. elegans* did not fare much better, as although the *MACROD2* homolog B0035.3 was knocked out using RNAi, no associated phenotype was reported for this homologous transcript (Kamath *et al.* 2003, Supplementary Table 3), although genes immediately downstream (B0035.5 and B0035.7) were associated with a non-viable phenotype. The functional class of B0035.3 was indicated as 'Chromatin' (Kamath *et al.* 2003, Supplementary Table 2). B0035.3 has not been highlighted as an essential gene via study of *C. elegans* development (Green *et al.* 2011; Zhang & Lin 2009). It is clear from this database search that *MACROD2* and its homologs remain poorly characterised, supporting the idea that further characterisation of this gene is still warranted.

4.4.4 Cell Model

I have created a neuronal model via use of the SH-SY5Y cell line which were differentiated to a neuron-like cell type with neurite outgrowth using BDNF and retinoic acid. As my hypothesis was that disruption of *MACROD2* in the cell will impact on the structure and/or function of the neuron, it was imperative that I had a means to test this *in vitro*. A limitation of confocal analysis is that you are restricted to visualisation of only a small number of proteins. Therefore we restricted our analyses to Synapsin, PSD-95 and MAP2. The use of Synapsin as a neuronal marker enables visualisation of pre-synaptic and post-synaptic neurons and is a more general marker for neuronal function. As we have discussed previously, glutamatergic transmission is hypothesised to be very important in autism etiology. PSD-95 is a marker for excitatory synapses, including glutamatergic transmission, thus having a cell model that expresses this marker in a study of a gene implicated with autism is most useful as it enables visualisation of these synapses. Finally,

MAP2 is a marker for the dendrites, enabling visualisation and thus quantification of the dendrites. These three diverse neuronal markers were chosen to give the best overall assessment of the function of the neuron, as it was unknown what cellular processes *MACROD2* would interact with or influence. Following assessment of these three markers in the cell model, Synapsin and MAP2 were brought forward as neuronal markers for the siRNA knockdown experiments.

My preliminary analysis indicated that the cell model is a good model for examining synaptic genes because these markers were shown to be affected by the siRNA knockdown of *MACROD2*, but not the siRNA knockdown of the positive control siRNA against *PPIB*. The cell model is most effective when the cells are less densely populated, which can be difficult to control for, and the experiments are vastly dependent upon the effectiveness of the antibodies used, and thus thorough optimisation is required before analysis can begin.

4.4.5 Limitations of this work and future directions

In this study I have begun to build up a picture of *MACROD2* at the protein level. I have performed immunostaining of *MACROD2* in the SH-SY5Y cell model and discovered that the *MACROD2* protein is present in both the nucleus and cytoplasm, a result in support of previous results from yeasts (Chen et al. 2011) and recombinant *MACROD2* in HeLa cells (Neuvonen & Ahola 2009). However this is the first time the localisation of endogenous *MACROD2* has been demonstrated. The hypothesis that *MACROD2* interacts with histone proteins appears to have been confirmed by the Co-IP and LC/ESI-MS analysis, although these results do come with caveat that *MACROD2* was not itself identified in the analysis. The association with Ku70, Spectrin alpha chain and microtubule assembly may begin to explain the associations of *MACROD2* with DNA repair, epileptic encephalopathies and brain volume.

This study represents the first in depth study of *MACROD2* at the protein level in a neuronal cell model. I have confirmed the localisation of *MACROD2* in neuronal cells for the first time, and also examined any protein-protein interactors of this poorly characterised gene, which appears to give context to some of the associations that have been made between *MACROD2* and human disorders. Whilst this study was constrained due to the limited time of a PhD project, it represents a strong stepping stone upon which

to move forward towards understanding the role of *MACROD2* in human biology overall. Functional studies of associated genes from GWAS are important, as these associations mean little without putting them into context of wider human biology. It can even enable us to ascertain causal links. It is for this reason that a study of the associated variant in greater detail, such as the study reported here, is required. This study was built upon this premise to translate the statistical evidence from GWAS to biological role, and has given some clues as to the role of this poorly characterised gene. It is my hope that with further study we will come to understand how variance at this locus may contribute to neuropsychiatric disorders and human health in general.

Chapter 5 Fine-mapping of an association signal at the *MACROD2* locus

5.1 Introduction

Independent attempts at replication of the association signal within *MACROD2* have been mixed. Curran *et al* (Curran *et al.* 2011) presented a case-control replication study of rs4141463 in an ASD sample of 1,170 cases from across 5 European sites and 35,307 controls. They found some support for the association in their German (225 Strict diagnosis vs 1145 controls, (OR=0.807 (95%CI 0.65–0.98) $p = 0.039$), but not in their UK, Dutch, Italian or Icelandic samples or for a combined meta-analysis (For summary of autism-only analyses see Figure 5.1). The authors note that they do not have the power from their sample size to replicate the original AGP finding at a genome-wide significance level. However, based on a population disease prevalence of approximately 0.2%, the observed disease allele frequency of 0.42 and an effect size of OR=0.65, one can calculate the UK (93%), German (94%), Dutch (80%), and Icelandic (>90%) studies all have a greater than 80% power to observe a nominal association at this marker. Only the Italian sample has limited power (<40%) in this endeavour.

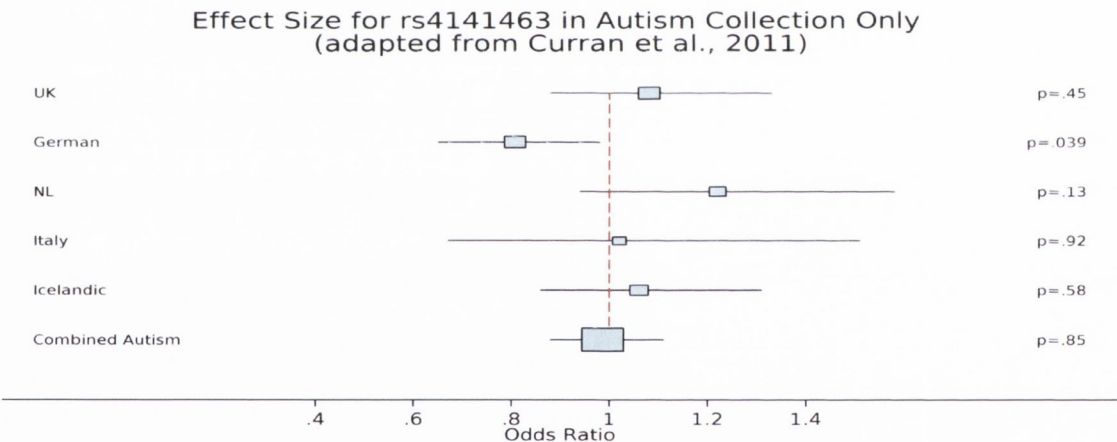


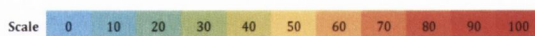
Figure 5.1 Forest plot of associations reported by Curran et al (2011) for strictly defined autism cases only.

While the results of GWA studies of autism are not as striking in terms of reported significance thresholds as has been observed in the studies of other multigenic diseases, such as type 2 diabetes and coronary artery disease (W.T.C.C. Consortium 2007), it is possible that this is due to the heterogeneity within autism samples. Individuals with autism are typically grouped together by phenotype, but could in fact be members of separate phenotypes with different severity and slightly different presentation. This

supposition is supported by the study by Anney *et al* (Anney, Klei, Pinto, Regan, Conroy, Tiago, et al. 2010) where the autism sample was split into 'spectrum' (which includes less severe cases of autism with those individuals with a more severe presentation of symptoms) and 'strict' (which only includes those individuals who score highly on the ADOS and ADI-R, and therefore have a more severe presentation of symptoms); separating the cohort in this fashion yielded a stronger association between the variant identified in *MACROD2* and the 'strict' dataset. In fact there are several ways in which GWA studies of autism can be improved, such as by increasing the sample size, limiting clinical heterogeneity and by improving SNP coverage.

5.1.1 The effect of sample size on the power of GWAS

The power of a genetic study to observe an effect is governed by a number of factors, the population prevalence of the disorder, the frequency of the disease allele, the linkage disequilibrium between the test and disease allele, the effect size and the most tractable and amendable variable being the study sample size (Risch Merikangas, 1996). Excluding studies prone to the misclassification of case or control-status, it is generally true that increasing the size of a GWAS study will in turn improve the power to observe an effect. Polygenic disorders were initially predicted to be driven by variants of modest effect $GRR > 2$ (Genotype Relative Risk). However, over time, and with experience, it has become clear that the data would suggest that risk alleles are likely to have much lower effect sizes $GRR < 1.5$. For variants which have a lower effect size ($GRR \sim 1.1 - 1.2$), the sample size required to achieve even modest power to observe an association requires sample collections in the thousands to the tens of thousands of cases and controls. To illustrate this point, a set of heat maps (created using CATS: www.sph.umich.edu/csg/abecasis/CaTS/) showing the power to observe association under varying conditions is shown in Figure 5.2. Of note even in studies of 1000 cases and controls, there is very little power to observe an effect at $p < 10^{-5}$ of $GRR = 1.25$.



Power based on: Population Prevalence = 1%, $\alpha = 0.00001$

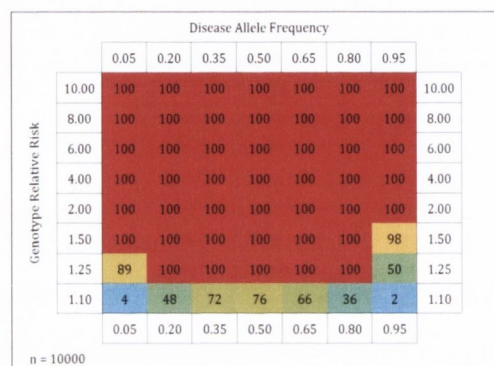
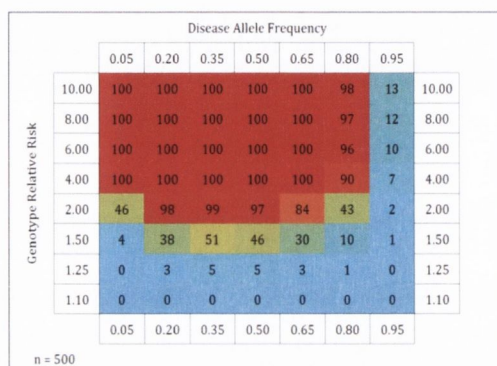
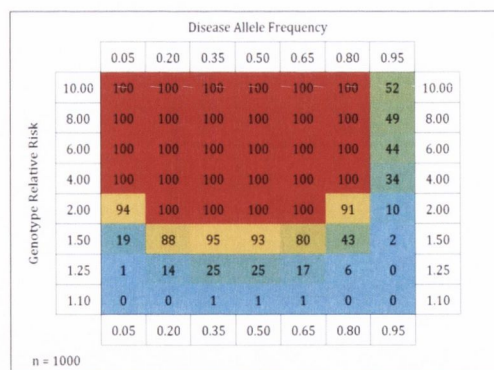
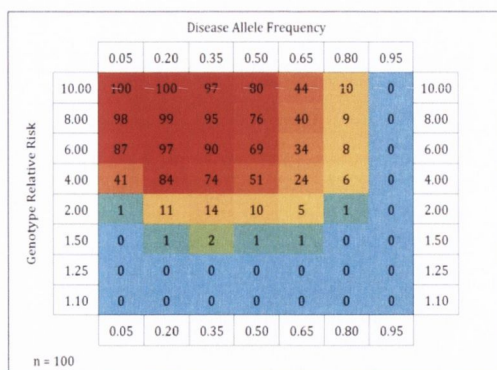


Figure 5.2 Power to observe association under the additive model for different sample sizes (n=100, 500, 1000 and 10000). Calculated for varying genotype relative risk and disease allele frequency using CATS (www.sph.umich.edu/csg/abecasis/cats/) based on population prevalence of 1% and $\alpha = 0.00001$.

5.1.2 The effect of clinical heterogeneity on the power of GWAS

Although the most amenable method for increasing the power of a study is to increase the sample size – this is not necessarily a panacea. One of the most challenging aspects of research into neuropsychiatric illness is clinical heterogeneity. Patients with autism may present with a wide array of symptoms and behaviours – from mild to severe intellectual disability, verbal and non-verbal status, and low, moderate and high IQs (Baird & August 1985). Szatmari hypothesised that clinical heterogeneity in autism is mirrored with genetic heterogeneity (Szatmari 1999). If this is the case heterogeneity in the sample will reduce effect sizes and limit the power of the study.

Whilst sample size is an obvious concern, it is possible that case heterogeneity is one of the reasons why discovering significant associations in autism has been so difficult to date. Studying individuals based not only spectrum and strict diagnoses of autism, but also phenotypes that may be more informative, such as cognitive ability or verbal status may reduce the genetic heterogeneity in the study. The additional partitioning of the case

sample will inevitably reduce the sample size and consequently may reduce the power of the study to observe an effect (ϵ). However, It is hypothesised that reduced clinical heterogeneity will increase the underlying ϵ within the case set such that $\epsilon_{\text{homogeneous}}$ is much greater $\epsilon_{\text{heterogeneous}}$.

5.1.3 The effect of SNP coverage on the power of GWAS

The ability of the researcher to interrogate more of the variation in the genome was improved by the development of microarrays, yet it is often the case that despite increased coverage, much of the genetic variation is measured via linkage disequilibrium. The power of a study to observe an association and an unobserved causative marker is reduced if the linkage disequilibrium between the observed and unobserved causative marker is low. Imputation is a method whereby the unobserved genotypes are estimated from the observed data at many markers and a reference panel of known haplotypes. By imputing genotypes in this way, you enrich the data by reducing missingness at observed markers and unobserved markers. By directly testing unobserved markers you significantly increase the statistical power of GWAS and improve the fine mapping of the causal variant (Marchini & Howie 2010). Imputation can also be deployed to compensate for problems related to coverage of different SNP arrays and can make it easier to combine genetic information from different arrays when imputed against the same reference panel. In this way imputation allows mega- or meta-analyses of data generated from different genotyping platforms. First described by Servin and Stephens, they developed a Bayesian regression method to estimate unmeasured genotypes using statistical modelling of patterns of LD from a reference dataset (such as HapMap SNP data), and made this publically available in the software package Bim-Bam (Servin & Stephens 2007). In the short time since then, multiple programs implementing different statistical tweaks have been developed to aid in such analyses, including IMPUTE (Marchini *et al.* 2007), MaCH (Li *et al.* 2010), BEAGLE (Browning & Browning 2009) and SHAPEIT (Delaneau *et al.* 2013). Imputation software typically reports a Bayes factor (indicating the strength of the evidence for an association), and not a p -value, in calculations to assess genotype-phenotype association, but each software can differ in how the LD pattern is calculated to infer haplotype. Typically HapMap haplotypes were used as a reference panel, however Huang and colleagues have demonstrated that the

1000 genomes dataset is the new benchmark for imputation, demonstrating that this large reference panel can improve and add additional support to a GW significant finding (Huang *et al.* 2012).

5.1.4 Missing Heritability

As we have discussed in Chapter 1, it is possible that the heritability estimates for ASD have been overstated. However, it is clear that there is a case of ‘missing heritability’ when it comes to autism susceptibility. ‘Missing heritability’ describes the phenomenon whereby the genetic variants that have been discovered by GWAS are of modest risk and explain only a small proportion of the heritability. Visscher, Hill and Wray (Visscher *et al.* 2008) offer a formal definition of heritability as “...a ratio of variances, specifically as the proportion of total variance in a population for a particular measurement, taken at a particular time or age, that is attributable to variation in additive genetic or total genetic values — termed the narrow-sense heritability (or just heritability, h^2) and the broad-sense heritability (H^2).” If we think of a phenotype as being the sum of the genetic and environmental factors, then:

$$\text{Phenotype } (P) = \text{Genotype } (G) + \text{Environment } (E)$$

The broad-sense heritability, H^2 , which encompasses all the genetic contributions to variance in the given phenotype, such as additive, dominant and epistatic genetic effects, is then given by:

$$H^2 = \text{Var } (G) / \text{Var } (P)$$

The narrow-sense heritability, h^2 , describes the variance in a given phenotype due to additive genetic factors, given by:

$$h^2 = \text{Var } (A) / \text{Var } (P)$$

Where A is the additive genetic effects. Narrow-sense heritability is more pertinent for GWAS and other family-based studies because phenotypic similarity between relatives is mostly driven by additive genetic variance (Wray & Visscher 2008). The value for heritability will always lie between 0 and 1. ‘Missing heritability’ therefore is apparent when all the contributing factors and effects do not add up to 1.

Manolio and colleagues suggest multiple explanations for how this ‘missing heritability’ is made manifest and paint a picture of genetic variation that is more

complex than previously anticipated. They suggest that a greater number of variants that confer small effects may yet remain to be found, and that rarer variants, which may or may not be of larger effect, are invisible to genotyping arrays that have focused on typing variants present in 5% or more of the population (Manolio *et al.* 2009). Structural variants (such as CNVs) may also be captured poorly by existing arrays. This is complicated further by the fact that the study design itself may be lacking, with low power to detect gene-gene interactions and environmental effects. These points all ring true for the study of autism and other neurodevelopmental disorders, as many of the detected variants thus far have a modest calculated effect size, e.g. OR <1.5.

However, despite the additive model capturing, to some extent, aspects of the recessive and dominant models it is likely that direct investigation of these models alongside models examining gender may better explain the association findings in ASD in a similar way that clinical homogeneity is also anticipated to fine-tune the association findings.

5.1.5 Pathway analysis

Chapter 3 has discussed the utility of grouping associated genes by gene ontology (GO) term by function, in the hopes of discovering enrichment for particular biological processes or pathways. This approach has been taken before by the AGP in their analysis of the AGP dataset using a modified version of the SNP Ratio Test (SRT) (O'Dushlaine *et al.* 2009) called 'pedigree SRT' (pedSRT) (Anney *et al.* 2011). The SRT examines the ratio of the amount of associated SNPs from a given GWAS dataset to the total number of SNPs in a pre-defined set of genes. Mapping the genes that are enriched for associated SNPs to GO terms can thus potentially reveal links between the associated genes in terms of biological process or pathway. However, this analysis is dependent on the GO term information being complete and correct for each gene under analysis. Anney and colleagues were unable to include *MACROD2* in their report due to the lack of GO term information for the gene at the time of analysis.

Chapter 4 describes the efforts, using co-immunoprecipitation analysis, to discover protein-protein interactions with *MACROD2*. This interaction data could be used to inform future analysis using the recently developed interval-based enrichment analysis tool for GWAS, INRICH or Interval enRICHment analysis (Lee *et al.* 2012). INRICH works by taking a set of independent, nominally associated genomic intervals and then looks for

enrichment of pre-defined gene sets, such as those identified by protein-protein interaction analysis. A genomic interval in this context is a genomic region of SNP association defined by linkage disequilibrium from a genome-wide scan. INRICH has been used previously to examine data from a GWAS of bipolar disorder, and discovered enrichment for calcium ion channel genes, which is a biologically plausible gene set (Sklar et al. 2011). Thus INRICH could be a useful tool for pathway analysis.

5.2 Aims of this study

In this chapter the key aim is to refine the *MACROD2* association signal. I will use the AGP genotype data from Stage 1 and Stage 2 for the *MACROD2* region and enhance the genotype coverage by applying a fine-mapping (imputation) approach alongside a more detailed interrogation of the genetic model and clinical subgroups to better describe the *MACROD2* association reported in Anney et al., (2010). Imputation in addition to achieving greater marker density will also reduce genotype missingness for typed markers in the region. Moreover, I will investigate whether genes which interact with *MACROD2* show association enrichment across the GWAS sample – thereby assessing whether common biological pathways which include *MACROD2* interactors are important in ASD.

5.3 Materials and Methods

5.3.1 Genotyping

The Autism Genome Project data is described in detail elsewhere (Anney, Klei, Pinto, Regan, Conroy, Tiago, et al. 2010; Anney et al. 2012). Briefly, the Stage 1 sample was genotyped using the Illumina Infinium 1M-single SNP microarray; the Stage 2 sample was genotyped on either the Illumina Infinium 1M-single or the Illumina 1M-duo microarray. The use of two microarray platforms, although from the same company were noted to have some between array discrepancies that may introduce errors. To assess potential biases between arrays, markers from subjects inferred to be of European ancestry showing $F_{st} > 0.02$ across recruitment sites were excluded from analyses. This assessment was performed by Lambertus Klei at the University of Pittsburgh and a list of exclusion lists were provided to AGP analysts. A total of 947,233 SNPs passed QC for the combined data sets and were used for downstream analyses.

5.3.2 Genetic Ancestry

Individuals were also described by their genetic similarity to reference ancestry populations. Specifically, individuals were classified as European if they showed similarity to the CEU HapMap reference population. Ancestry was estimated for both Stage 1 and Stage 2 samples by using 5,176 SNPs, which had a genotype completion rate of $\geq 99.9\%$, a minor allele frequency ≥ 0.05 and were separated by at least 500 kb using SpectralGem. Ancestry information was estimated by Lambertus Klei and Bernie Devlin within the core AGP analysis team. The classification of individuals by ancestry over the first three principle components from SpectralGem are shown in Figure 5.3.

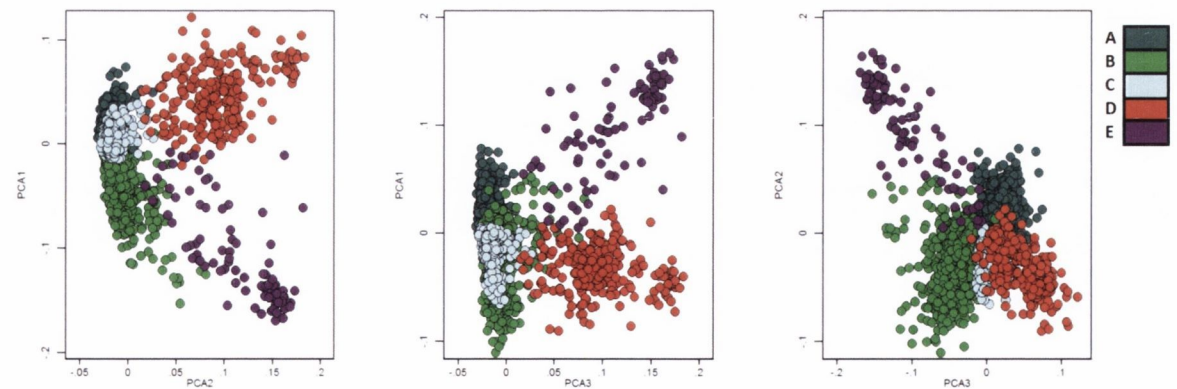


Figure 5.3 Plot highlighting population ancestry. Groups A through C delineate 1285, 771, and 486 families of European ancestry as determined by representation of European recruitment sites. All other families were assigned to be non-European for purposes of association analysis (from Anney et al., 2012).

5.3.3 Clinical evaluation

Participants in this AGP study were recruited at centres in North America and Europe. Subjects with known karyotypic abnormalities and fragile X mutations were typically excluded. Enrolled individuals were assessed using the Autism Diagnostic Interview-Revised (Lord et al. 1994) and Autism Diagnostic Observation Schedule (Lord et al. 1989). Cognitive functioning was established for the majority of subjects using a range of cognitive measures from which a categorical classification of intellectual capacity was derived for the analyses, as described in Anney *et al.* (Anney, Klei, Pinto, Regan, Conroy, Tiago, et al. 2010).

I followed the primary diagnostic classifications used in the primary AGP manuscripts, namely individuals were described within two non-independent diagnostic classes. Where more than one ADI or ADOS was reported for an individual, the most recent measure was taken:

Strict: Defined as individuals who met ADI-R and ADOS criteria for autism

Spectrum: Defined as individuals who met strict criteria plus individuals with an ASD diagnosis as determined by responses to the ADI-R and ADOS. Individuals were excluded if on either the ADI-R or ADOS there was evidence contrary to an autism diagnosis.

Cognitive ability was assessed using a range of measures. Integration of multiple measures of Verbal, Performance and Full-Scale IQ was performed by the AGP Phenotype Analysis Committee. Using broad categorical bands, individuals were described as having **higher IQ** if they reported a categorical best-estimate IQ > 80. Individuals were described as having **lower IQ** if their best-estimate IQ was shown to be between 25 and 70. IQ categorization was determined using available measures of verbal, non-verbal (performance) and full-scale IQ assessments. A score of greater than 80 for any of these three measures resulted in classification in the higher IQ group; otherwise, individuals classified into the lower IQ group had IQ estimates in the range 25 – 70.

Verbal status was described using the language level question on the ADI and ADOS. Specifically individuals were defined as being **verbal** if they have functional phrases. **Non-verbal** individuals were defined as having no functional use of 3 word phrases and less than 5 words total. Individuals were classed as verbal if they met verbal

criteria on either measure. Individuals were classed as non-verbal if they met non-verbal status on both measures or met non-verbal status on one measure and were missing on the second measure. A summary of the number of individuals who meet criteria for analysis are given in Table 5.1.

Table 5.1 A breakdown of the individuals who meet criteria for analysis, including the number of individuals from each grouping across AGP stage 1, stage 2 and stage 1 and 2 combined. %Probands describes the percentage of probands contained within that group compared to the overall proband sample.

Group	Stage 1	Stage 2	Stage 1 + 2	%Probands
Male	1140	1158	2298	86%
Female	223	138	361	14%
Verbal	1187	1178	2365	89%
Non-Verbal	175	117	292	11%
Higher IQ	656	610	1266	48%
Lower IQ	444	356	800	30%
Spectrum	1363	1296	2659	100%
Strict	829	623	1452	55%
European	1230	1087	2317	87%
European Spectrum	1230	1087	2317	87%
European Strict	746	521	1267	48%
European Strict (Verbal)	669	497	1166	44%
European Strict (Non-verbal)	77	24	101	4%
European Strict (Higher IQ)	361	289	650	24%
European Strict (Lower IQ)	246	145	391	15%
European Strict (Male)	626	464	1090	41%
European Strict (Female)	120	57	177	7%

5.3.4 Imputation

Prior to imputation data was prepared through an in-house quality control pipeline. All pipelines were written in Stata 10 and use data extracted from the NCBI ftp-site (ftp://ftp.ncbi.nih.gov/snp/organisms/human_9606/database/organism_data/).

Genotype nomenclature was updated to the latest build of dbSNP using **RSMergeArch**. SNPs that were merged or dropped from dbSNP were excluded from further analyses. Chromosome and position on reference genome were assigned using the **SNPChrPosOnRef** file. Genotype strand was matched to the phased reference haplotypes used for imputation. Briefly, ambiguous genotypes (A/T or G/C) were dropped as were genotypes that did not conform to expected complement (e.g. drop if A/G in the reference does not equal A/G or C/T). Where genotypes were complementary the strand

was flipped in the test data to conform to the reference data (e.g. C/A converted to G/T). Strand flip was performed using PLINK (Purcell et al. 2007).

5.3.5 Quality Control of GWAS

Genotype quality control was performed using an in-house STATA pipeline which uses PLINK and Stata 10. A standard quality control pipeline was adhered to when preparing a GWAS dataset for analysis. This pipeline was designed to remove any problematic markers or individuals that could otherwise impact in further analysis i.e. correct for genotyping error, technical replicates or markers with missing information, for example. For this project, the QC protocol was performed, in two stages: The first checks for problematic individuals in the sample, the second looks for problematic markers in the sample.

- 1) Remove individuals with poor genotyping data [Threshold <90%]. Some individuals or samples may have just performed poorly on the array and should be removed.
- 2) Remove individuals showing high heterozygosity [Threshold >3SD of the mean]. High heterozygosity can be indicative of samples that were mixed or contaminated.
- 3) Remove individuals who are identical by state i.e. individuals who are related or who may have been processed twice [Threshold 0.9 = Identical; 0.125 = Cousin; 0.08 = "Unrelated"]. Calculated using an autosomes file only and the --genome routine in PLINK.
- 4) Remove or correct individuals showing ambiguous gender i.e. individuals who have either been miscoded for gender or are highlighted for ambiguous gender by the --check-sex routine in PLINK [Threshold <0.2 = Female >0.8 = Male].
- 5) Remove families with high Mendel-error – with trio data the Mendel-error can be calculated i.e. alleles in an individual that could not have come from either parent, and is defined using the --mendel routine in PLINK [Threshold >5% errors per family].
- 6) Remove SNPs with poor genotyping i.e. with greater than 5% missing, calculated in PLINK using the --geno command [Threshold <95%].

- 7) Remove SNPs with skewed Hardy-Weinberg values i.e. SNPs that are not in Hardy Weinberg equilibrium. PLINK calculates this using only information from the founders, thus any offspring are ignored [Threshold $< 10 \times 10^{-6}$]. It is performed using the `--hwe` command.
- 8) Remove SNPs with differential missing between cases and controls and performed using the `--test-missing` command in PLINK. [Threshold $p < 10 \times 10^{-6}$].
- 9) Remove monomorphic SNPs: Those SNPs for which only a single form or allele can be identified in the sample population.
- 10) Remove SNPs with very low minor allele frequency using the command `--maf` [Threshold: Defined at association level according to power, but for this study was typically $>5\%$].

5.3.5.1 Summary of QC pipeline scripts

PLINK --bfile input --mind 0.1 --make-bed (1)

PLINK --bfile input --het --make-bed (2)

PLINK --bfile input --genome (3)

PLINK --bfile input --check-sex (4)

PLINK --bfile input --mendel (5)

PLINK --bfile input --out input --geno 0.95 --make-bed (6)

PLINK --bfile input --hwe 0.00001 --make-bed (7)

PLINK --bfile input --test-missing (8)

PLINK --bfile input --out input --exclude PLINK.nof --make-bed (9)

PLINK --bfile input --maf 0.05 --make-bed (10)

5.3.6 Generating the case pseudo-control dataset from trio data

It is sometimes necessary to examine case individuals from a family-based association study as part of a case-control design. For imputation routines as standard practice we impute within a case-control framework. In the absence of a matched control dataset we can generate a matched control from the untransmitted alleles from the parents. The untransmitted alleles are termed the pseudo-control and act as a matched

control for the purpose of the association. Pseudo-controls were generated using PLINK using the following three stage routine.

```
PLINK --bfile input --me 1 1 --set-me-missing --recode (1)
```

```
PLINK --bfile input --tucc --me 1 1 --set-me-missing --make-bed (2)
```

```
PLINK --ped PLINK.tucc.ped --map PLINK.map --make-bed (3)
```

5.3.7 Genotype imputation

Imputation was performed using BEAGLE (Browning & Browning 2009) and PLINK. The latest version of BEAGLE (3.3.2) was downloaded from the University of Washington website: <http://faculty.washington.edu/browning/BEAGLE/BEAGLE.html>. BEAGLE is a software package designed for analysis of the large scale datasets, in terms of markers and samples, generated for GWAS. BEAGLE was chosen for this analysis because it was compatible with the format of the genotype data in use and was also compatible with PLINK.

The data formats that each program require are somewhat different, and thus there were some additional steps and adjustments needed to run the protocol: BEAGLE does not accept the individual IDs ('iid') and family IDs ('fid') of the .gprobs and .dosage file format used by PLINK and so the 'iid' and 'fid' was combined into a new variable 'cid' prior to running the analysis. Due to the computational load of running the imputation, it was required to run the analysis for only a 500 kb region surrounding rs4141463 within the *MACROD2* locus. This boundary was held consistent across all analyses. Three separate output files were created from the imputation that was performed:

- 1) Imputation based on the most likely call.
- 2) Imputation based on a defined threshold [Probability > 0.6, Missing = 0].
- 3) Imputation using the .dosage files from PLINK.

Prior to imputation the sample is pre-phased and imputed to remove missing data at typed markers. The pre-phasing stage uses the input data as a reference and therefore makes the imputation susceptible to batch effects. Therefore to limit batch effects all samples were imputed simultaneously. Following the pre-phasing routine, a second imputation round was used alongside a reference dataset to impute un-typed markers. All

imputation routines were computationally intensive and performed using the Trinity Centre for High Performance Computing cluster.

5.3.8 Reference panels

Imputation was performed against the 1000 genomes reference panel phase 1 version 3, available as phased genotype data in BEAGLE format. This data was acquired from:

http://bochet.gcc.biostat.washington.edu/BEAGLE/1000_Genomes.phase1_release_v3/.

5.3.9 Genetic Association Analyses

In the original AGP manuscript association analyses were performed using the family based association tests called the Transmission Disequilibrium Test (TDT). TDT was first described by Spielman, McGinnis and Ewens, in their study of the genetics of diabetes mellitus. TDT is a family based association test designed to compensate for the effect of population structure on association (Spielman *et al.* 1993). TDT gained favour as it reports genetic linkage only in the presence of genetic association and can cope with variance in population structure. TDT can only be performed using family based data, such as trio data, as is the case with this study. TDT can only detect linkage between the locus of the associated marker and the disease locus if the association is due to linkage disequilibrium.

However, in the context of exploring the imputed data we converted the family-based design to a case - pseudo-control analyses. All association analyses were performed in PLINK using either the default --assoc function, which corresponds to a logistic regression, or using the --linear function, which calculates a linear regression using covariates specified by the user. Regression analysis describes the statistical method used to model the relationship between a dependent variable and an explanatory variable. Linear regression uses a linear model, calculated from the explanatory variables (from the covariate file in this case) and is used to explain variance in the dependent variable. It is used to quantify the strength of the relationship between these sets of variables and is used in the context of genetic studies to model the effect of genetic variance upon quantitative traits, such as brain volume or enzymatic activity. Logistic regression uses a logarithmic linear model, calculated from predictor variables used to explain the (continuous) variance within a categorical independent variable. It is used to quantify the strength of the relationship between a categorical dependent variable (i.e. disease status)

and a continuous independent variable (e.g. genotype, gender) and typically reports an odds ratio, i.e. the contribution of the independent variable, say a particular mutation, towards a categorical variable, such as disease status.

5.3.10 Association analysis using imputed data (dosage/threshold)

BEAGLE enables generation of “dosage” SNP datasets, which contain information about the expected allele counts for each imputed allele. BEAGLE also enables specification of the imputation threshold; here I used a threshold of 0.6 as a supplementary analysis, which meant that BEAGLE would impute an untyped allele if it had >60% confidence in the likelihood of that allele being the true allele. As previously stated, BEAGLE and PLINK were used in this analysis due to their interoperability, and therefore the ‘--dosage’ routine was used in PLINK to perform association analysis that also uses the dosage information to inform the analysis.

5.3.11 Enrichment analysis of putative protein-protein interactors identified via Co-IP experiments

Chapter 4 includes descriptions of a set of co-immunoprecipitation experiments designed to isolate proteins that interact with *MACROD2*. This resulted in a list of genes that had been identified from LC/ESI-MS analysis of the Co-IP eluent. The sets of genes from the different experiments were combined with all peptides with at least an 80% match between peptide chains and known proteins from the UniProt database was retained for further analyses. INRICH (INterval enRICHment Test, P. H. Lee *et al.* 2012), is a PLINK format compatible program that examines pre-defined sets of genetic variants (i.e. the list of genes from the Co-IP pulldown) for enrichment of association signals (i.e. associated variants from the AGP GWAS). INRICH groups together GWAS associations according to a user defined threshold, which was set at $P < 10^{-2}$, $P < 10^{-3}$ and $P < 10^{-4}$, using autosomes only, and defines regions based on a linkage disequilibrium clumping algorithm for these associations. INRICH then assesses whether these defined regions overlap with the pre-defined list of genes more often than would be expected by chance, and reports significance as a p-value.

5.4 Results

5.4.1 The effect of SNP coverage on the power of GWAS - Imputation of the *MACROD2* locus

In an effort to improve upon and add supporting evidence for association at the *MACROD2* locus, I performed imputation of a 500kb window surrounding the original association signal at rs4141463. Using the 1000 genomes dataset as reference increased the marker density (i.e. SNP coverage) at this locus greatly, as illustrated in Figure 5.4. Imputation at this locus improves the original AGP signal slightly, meeting a significance of 2.12×10^{-9} , but did not identify any new associated SNPs with significance above rs4141463. This association study replicates the association identified in the 2010 AGP study, and imputation has also identified additional supporting markers. However, there was no unique association signal observed above the previously described association at rs4141463.

Imputing using BEAGLE also enabled generation of dosage files and specification of a confidence threshold in allele calls. The results of these different analyses for the AGP stage 1 strict data are illustrated in Figure 5.5. Imputation has vastly increased the marker density across the imputed window, and in tandem with this there has also been an increase of marker SNPs in support of the association at rs4141463, although no additional markers exceed the genome-wide significance level or that of at the rs4141463 marker. The genome-wide significance of rs4141463 does not differ significantly across the different association analyses tested.

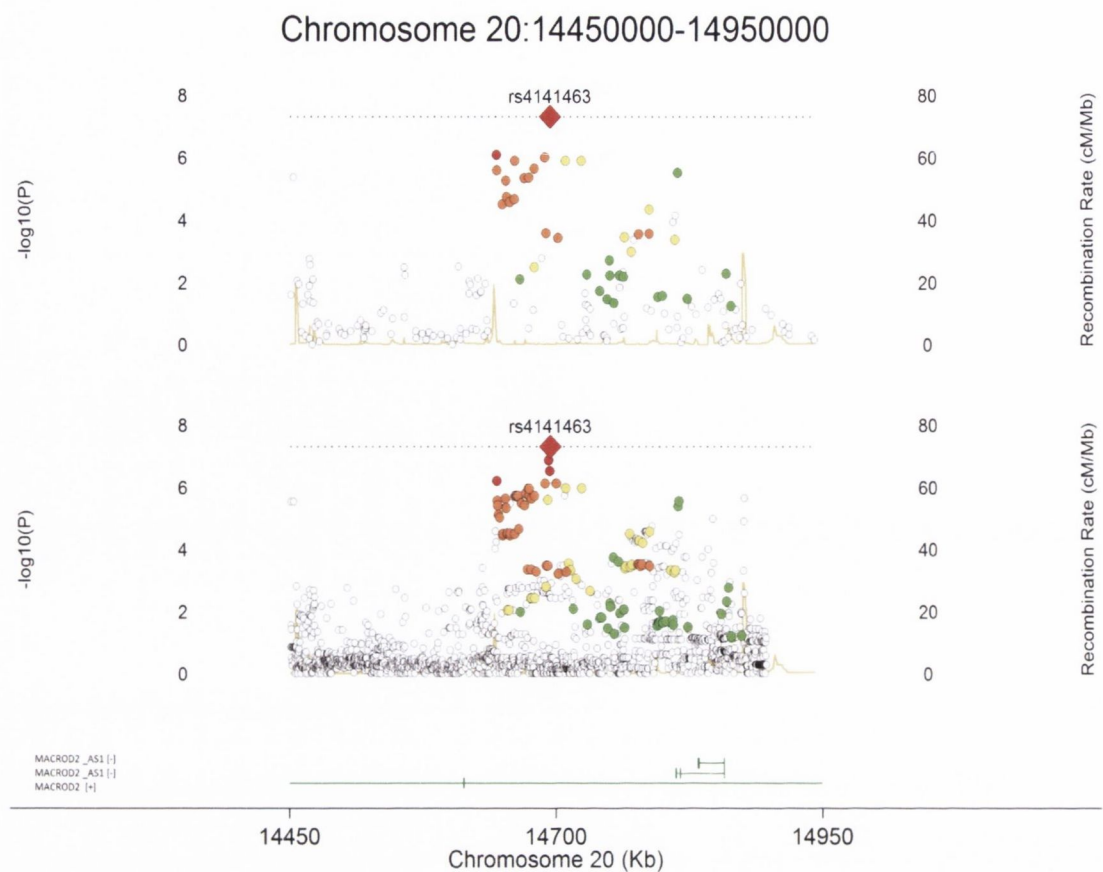


Figure 5.4 An association plot of European strict AGP stage 1 individuals. The top panel is the raw association, the bottom panel is the imputed association. Each circle on these plots indicates an individual marker SNP. On the left of each association plot is the $-\log_{10}(P)$ scale of genome-wide significance, indicated by the height and colour of each marker SNP. The red diamond is the most significant marker, which was rs4141463 across all analyses, imputed and non-imputed. On the right is the recombination rate, indicated by the tan line on the association plot, and demonstrates that the associated markers appear to be in linkage disequilibrium.

Chromosome 20:14400000-14900000

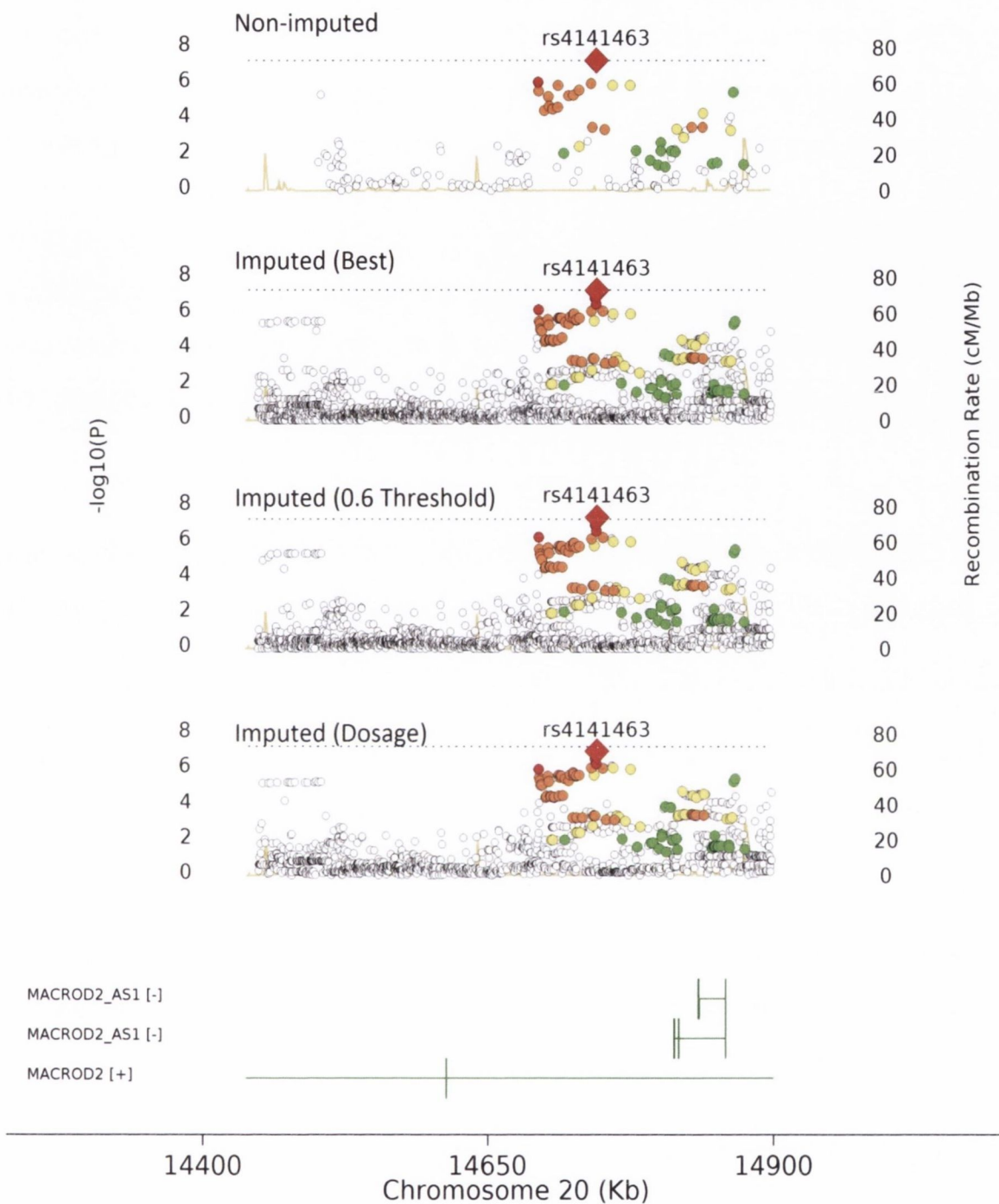


Figure 5.5 Association analysis of a 500kb region at the *MACROD2* locus. ‘Non-imputed’ refers to an association analysis using the original stage 1 and stage 2 AGP dataset. ‘Imputed (Best)’ refers to an association analysis using imputed stage 1 and stage 2 AGP data that was imputed against 1000 genomes as reference, using a confidence threshold of 0.9 for allele calls. ‘Imputed (0.6 Threshold)’ was a similar analysis to ‘Imputed (Best)’ but the confidence threshold for allele calls was set at 0.6. Finally, ‘Imputed (Dosage)’ used the dosage files generated by BEAGLE, which contain information about the expected allele counts which is used to inform the association analysis.

5.4.2 The effect of sample size and clinical heterogeneity on power of GWAS

In an effort to increase the sample size I used the combined AGP stage 1 and stage 2 dataset for imputed association analysis, which could then be compared against the AGP stage 1 and stage 2 data individually. As discussed in the introduction of this chapter, refining the clinical heterogeneity was of interest to me. Although parsing the data into smaller groups matched by phenotype affects the sample size and thus the statistical power to observe an effect, if one of the phenotypic subgroups is driving the signal than this will become apparent when dividing the sample up in this way. This was compensated for somewhat by the use of a pseudo case-control method, as described in Section 5.4.6, which creates matched controls from trio data using the untransmitted allele information from the parent data.

5.4.2.1 Comparison of association in AGP Stage 1

The AGP stage 1 dataset is the discovery dataset reported by the AGP for the original association at *MACROD2* (Anney, Klei, Pinto, Regan, Conroy, Tiago, et al. 2010). Figure 5.6 illustrates the phenotypic breakdown of the AGP ‘strict’ samples, which are described in greater detail in Section 5.4.3. Note that the effect size for the strict groups, indicating a more severe autism diagnosis, was always greater than that of the corresponding spectrum group, which includes the individuals from the strict diagnosis and also any individuals who had a less severe diagnosis of autism. I sought to assess whether refinement of the association signal within the stage 1 subset would significantly reduce (or increase) the observed effect. Plotting the data in this way, it appears that male probands from the strict sample are driving the association signal, with female and non-verbal phenotypic subgroups appearing to be poor predictors for association. However, note that parsing the data into these subgroups has created groupings of limited sample size.

From Figure 5.6 I demonstrate that the strict European individuals from the Stage 2 data show poor association at this locus. Moreover, division by cognitive ability does little to alter the effect size, although verbal status has a greater association than that of non-verbal status. The association signal in males is again more significant than for females. Although the difference in effect size is not large, I see some evidence for increased effect for the male and verbal subdivisions of the stage 1 data; however this is small and not dramatically different from the undivided strict European groupings.

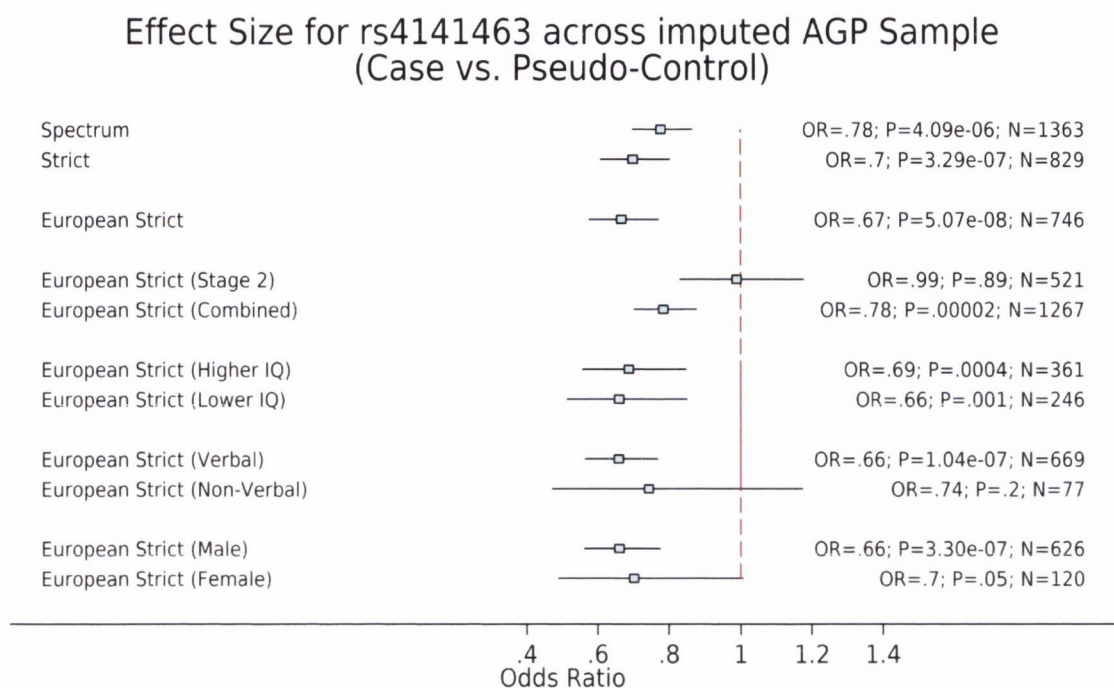


Figure 5.6 An investigation of the effects of phenotype on the effect size and genome-wide significance for rs4141463, the most significant association at the *MACROD2* locus, across the AGP stage 1 and stage 2 data. Parsing the data into sub-groups has no significant effect on enriching the genome-wide significance.

5.4.3 The effect of inheritance model on power of GWAS

PLINK enables investigation of the effect of different inheritance models on association, including dominant, recessive and additive models for association. Figure 5.7 illustrates the effect of these different inheritance models on association at rs4141463. I sought to seek support of the idea that the additive model (and thus the narrow-sense heritability) as being more pertinent for GWAS, as discussed in Section 5.2.4. For this analysis the effect size for the dominant analysis is weaker compared to the additive analysis, although the genome-wide significance is only modestly impacted. The recessive model is also slightly stronger performing than the dominant model, however the differences are nominal. Figure 5.8 illustrates this effect on genome-wide significance in greater detail, and indicates a difference in the markers supporting a dominant and recessive model in comparison to the additive model. Note that recessive markers (Figure 5.8, bottom panel) are present but not strongly linked to the index SNP rs4141463, although there is stronger support for recessive alleles in the region. This difference for

the recessive model may reflect an allele frequency difference due to using R^2 over D' , and therefore I examined the association plot using D' linkage disequilibrium (Figure 5.9). When D' is high, it indicates alleles are good surrogates for each other, but does not necessarily reflect that the markers are highly correlated for one another. So although rs4141463 does not appear as significant using the recessive model, Figure 5.9 demonstrates that there are many more significant markers in LD with rs4141463 from the recessive model. Therefore from this data we can surmise that the recessive model may give additional associated loci at this locus.

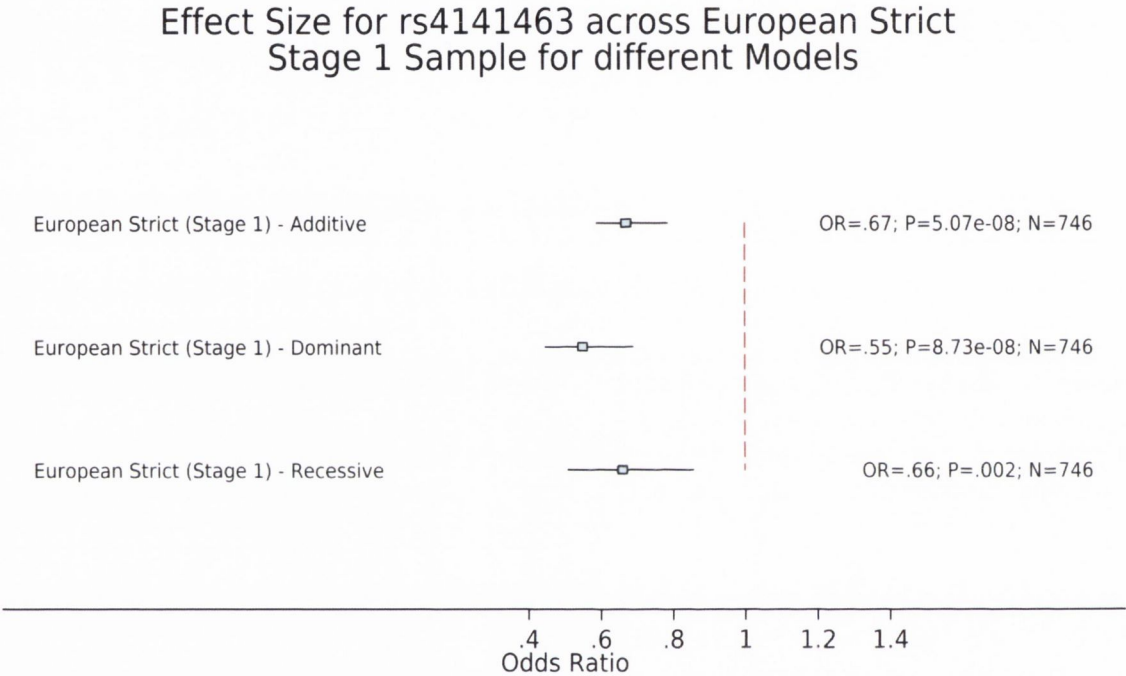


Figure 5.7 An investigation of the effects of inheritance model on the effect size and genome-wide significance for rs4141463, the most significant association at the *MACROD2* locus, across the AGP stage 1 data. The additive model appears to be the best model of inheritance to be used with this data.

Chromosome 20:14400000-14900000

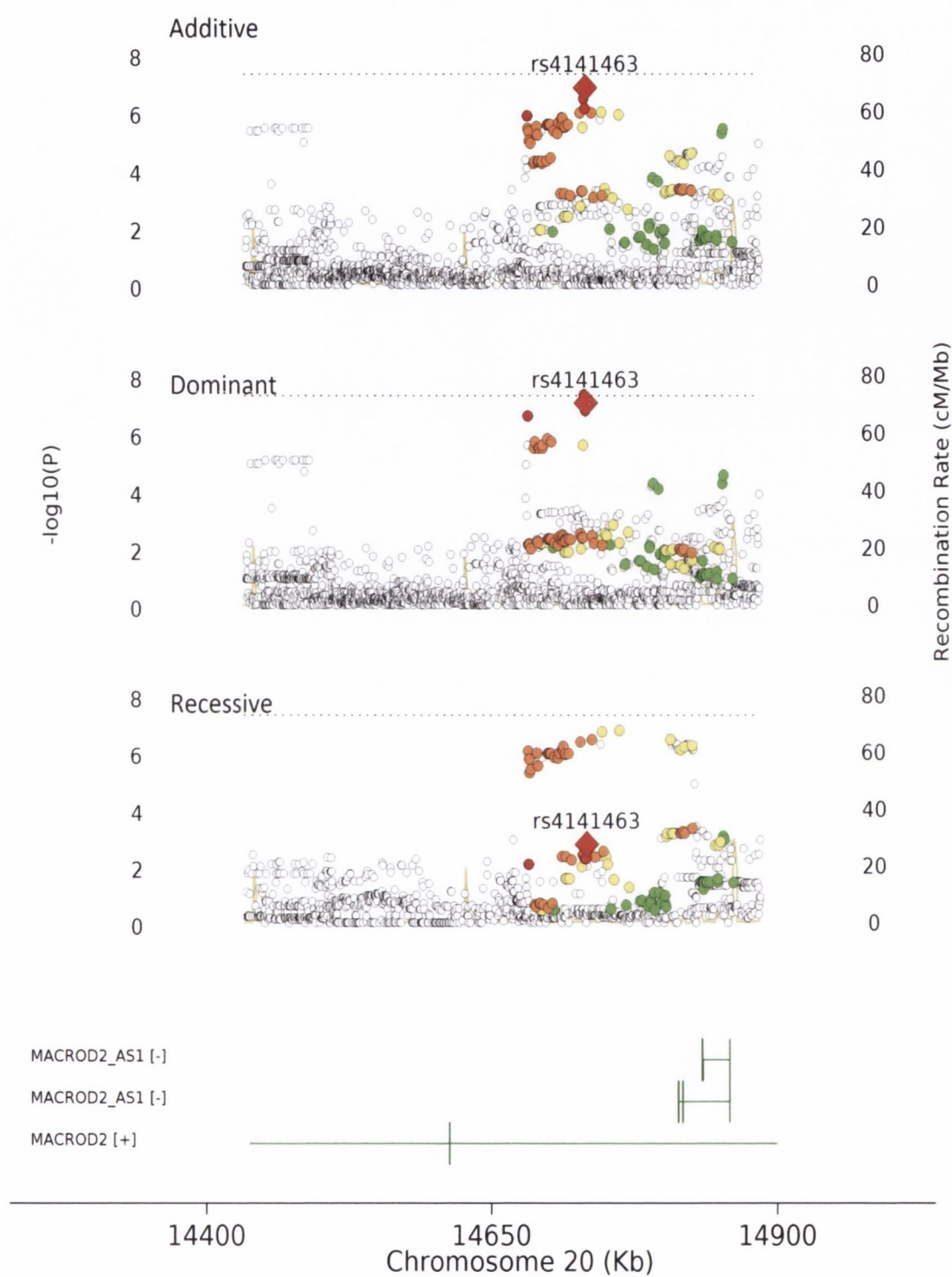


Figure 5.8 An association plot for the different inheritance models illustrates their effects on genome-wide significance for rs4141463, the most significant association at the *MACROD2* locus, across the AGP stage 1 data.

Chromosome 20:14400000-14900000

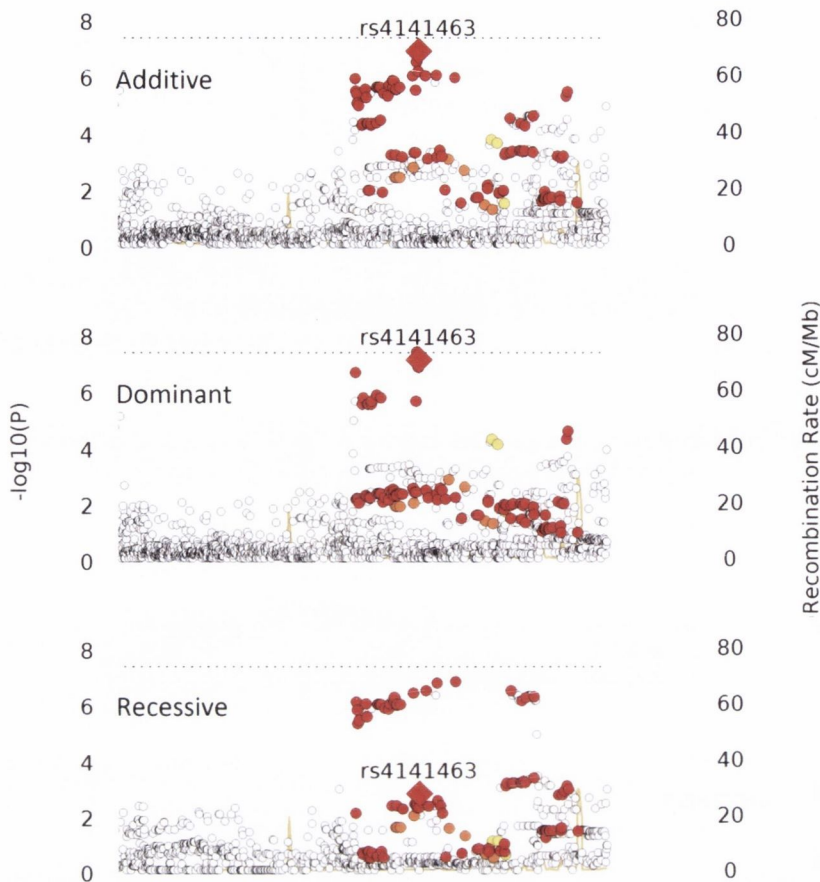


Figure 5.9 Association plot emphasising the D' prime linkage disequilibrium. Markers in high-LD (i.e. high D') are indicated in red. The additive model is the best performing inheritance model overall, however there are more markers approaching genome-wide significance in LD with rs4141463 for the recessive model, indicating that this model may be useful for capturing additional associated markers.

5.4.4 Enrichment analysis of putative protein-protein interactors identified via Co-IP experiments

Enrichment analysis failed to identify any significant enrichment for genes from the LC/ESI-MS analysis of the Co-IP eluent and associated variants from the AGP sample. 84 genes were analysed in total, across 3 different p-value thresholds for the genomic intervals: $P < 10^{-2}$, $P < 10^{-3}$ and $P < 10^{-4}$. INRICH analyses were used to determine whether this group of genes show enrichment for association. Analyses were run with and without *MACROD2* in the model to determine the role of the putative interacting proteins alone. The results of this alternate analysis are outlined in Table 5.2 and 5.3. Here we find an enrichment from the 'Co-IP pulldown from whole cell protein extract' (as described in Chapter 4) for *MACROD2* and *USP33* for the $P < 10^{-3}$ interval. *USP33* (Ubiquitin specific peptidase 33) is a carboxyl-terminal hydrolase.

Table 5.2 Interval-Gene-Target Summary.

GENE_LOC	GENE_ID	GENE_DESC	TARGET	P
chr1:77934264..77998152	USP33	USP33	ALL2 ALL2	0.0333933
chr20:13924146..15981841	MACROD2	MACROD2	ALL2 ALL2	0.0333933

Table 5.3 Interval Summary.

INTERVAL	N_SNP	N_ALTLOC	N_GENE	GENE_ID	GENE_DESC
int46:chr1:77830973..78055571	67	1072	1	USP33	USP33
int595,int596,int597,int598,int599,int600,int601,int602:chr20:14347525..15138270	338	5104	1	MACROD2	MACROD2

5.5 Discussion

5.5.1 The effect of SNP coverage on the power of GWAS - Imputation of the *MACROD2* locus

In an attempt to improve and refine the original signal from the AGP GWAS study, I sought to implement imputation of the AGP stage 1 and stage 2 datasets using BEAGLE, using the publically available 1000 genomes data as reference. While imputation at the *MACROD2* locus did not drastically improve the association signal at *MACROD2*, the number of supporting markers at this locus was greatly increased. This in itself is a positive result, as for an associated marker, one has more confidence in a significant loci if the surrounding markers, particularly those that are in the same LD-block as the associated marker, are supportive of the association. It was also possible that imputing the missingness away from the top signal would not have improved the significance of the top hit at rs4141463. In order to run imputation, all analyses were run under a case-pseudo-control association model as opposed to the TDT model. This approach did not significantly alter the TDT-based findings reported in the original manuscript. The results presented here are in support of the utility of imputation using 1000 genomes data to provide additional support for significant loci and demonstrate the impact of increasing SNP coverage.

Improving the SNP coverage increases the likelihood of identifying association of ungenotyped variation. Array designs utilise linkage disequilibrium so that in theory an array can tag potential causative variants – the global coverage of the array refers to the number of common SNPs (i.e. those SNPs with a minor allele frequency ≥ 0.05) across the genome that are tagged on that array or that are in linkage disequilibrium (e.g. with an r^2 threshold of >0.8) with the SNPs on the array. Whilst the issue of coverage is not as much of a problem in present times as it was in the past, due to the increased size of the arrays and the improvements in imputation, one must keep these caveats in mind when comparing present studies to studies performed on older arrays. Ideally replication studies should be performed using the same method (i.e. genotyping array), but this is not always the case. The results presented here demonstrate that imputation can be used to replicate an association signal.

The PGC Autism GWAS, which combines array data from 10 studies using imputation across a combined study of 6,495 cases and 6,495 controls retain a nominally

significant hit for *MACROD2* - the 11th top hit from that analysis (PGC ASD Steering Committee, personal communication, see Figure 5.10). However this combined sample does include the AGP stage 1 which could be driving the signal, as seen with the results reported in Section 5.5.2 for the AGP stage 1 and 2 combined analyses. Thus the jury is still out on the significance of the association within *MACROD2*.

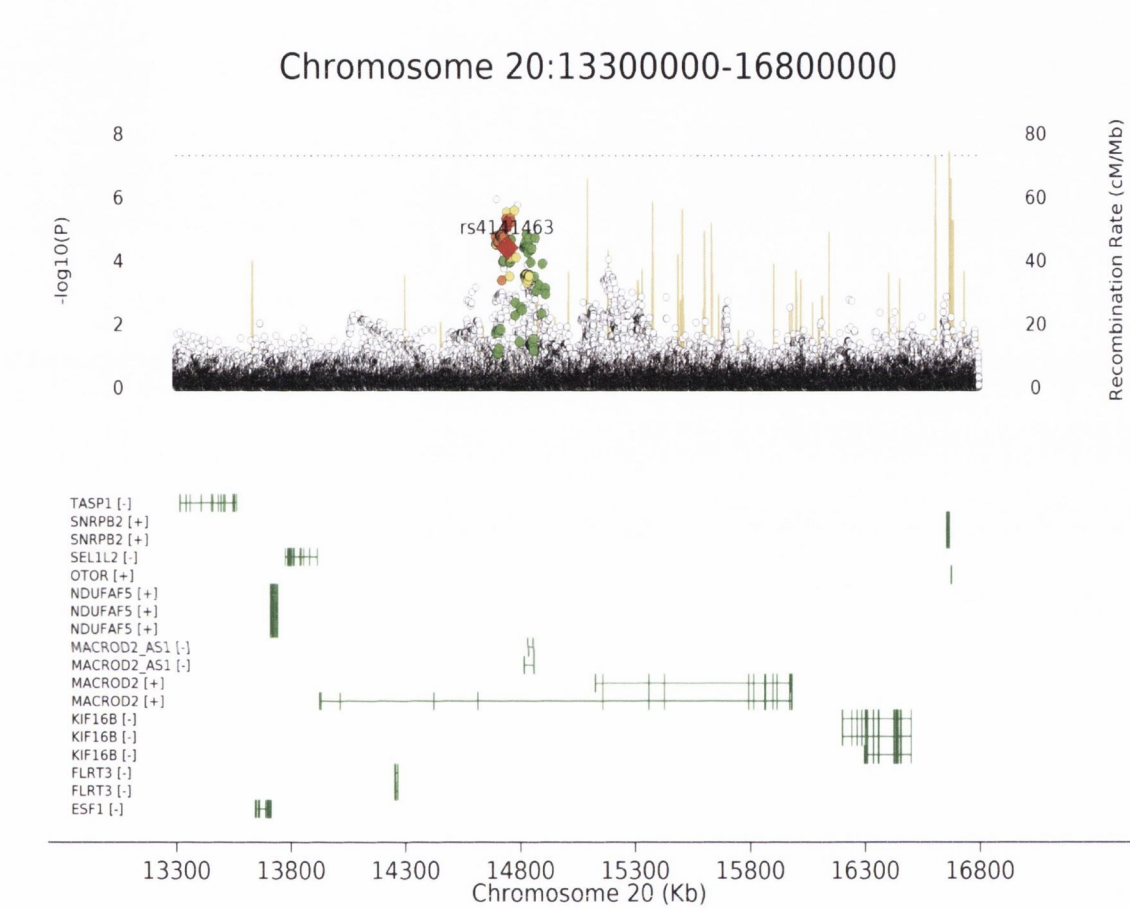


Figure 5.10 Association plot for the nominally significant hit within *MACROD2* from the PGC ASD GWAS demonstrating that the association at *MACROD2* remains nominally significant. However the PGC dataset contains the AGP Stage 1 strict sample (from which the original association with *MACROD2* was made) which could be driving this signal.

It is possible, due to the observation of multiple CNVs at the associated *MACROD2* locus (See Chapter 3), that these large deletions could be affecting genotyping at this locus. In the presence of CNVs, the clustering of the SNP genotyping instead of forming 3 distinct genotype groups tends to split these groups into smaller clusters resulting in a distinct pattern (see Figure 5.11). However, by plotting the raw allele intensities I found no evidence supporting an overlapping CNV at this allele therefore I am confident that

the association is not due to a CNV driving the signal directly or as a result of reduction of the genotyping quality (See Figure 5.12).

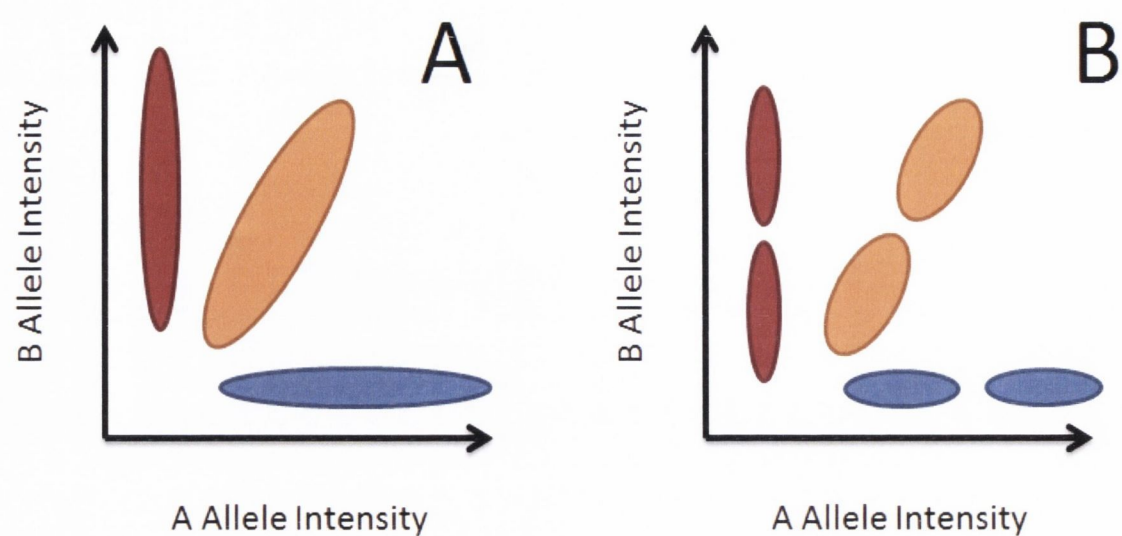


Figure 5.11 An illustration of the effects of CNV on allele intensity. A) An allele intensity plot for an allele with no impact upon genotype calls. B) An allele intensity plot that indicates presence of a CNV due to the distinct pattern of allele calls.

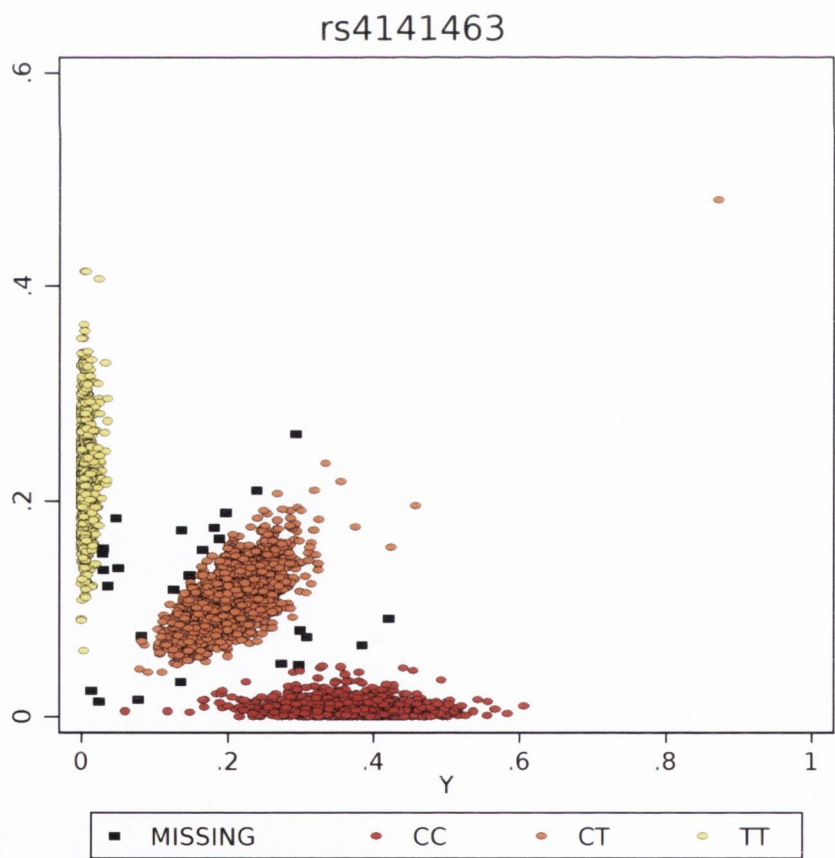


Figure 5.12 Genotyping plot for rs4141463 finds no evidence of CNV deletions driving the signal by disruption of genotype calls.

5.5.2 The effect of sample size and clinical heterogeneity on power of GWAS

One can appreciate that limitations of sample size are clearly not an isolated problem in the study of ASD. The common assertion that the larger the sample size the better is perhaps especially difficult to realize when it comes to association studies of complex/polygenic traits, due to the sheer volume of samples required, and the expense, both monetarily and temporally, that it takes to process such volumes of data. The more recent results from the PGC Schizophrenia GWAS (PGC Schizophrenia Steering Committee, Presented at Hamburg World Congress of Psychiatric Genetics (Ripke, Daly, O'Donovan et al., 2012)) of over 52,000 individuals (25,785 cases, 28,441 controls) have demonstrated, not unsurprisingly, that once a sample size of the right size (as calculated by power calculations) is utilised one is better able to identify risk loci. Mark Daly goes on to estimate using inference from the results of GWAS in other common complex disorders, that autism GWAS of corresponding power to the PGC SZ GWAS will require up to 60,000 individuals (https://pgc.unc.edu/worldwide/presentations/Daly_Jan_11_2013.pptx). One could suggest that the AGP GWAS, and any previous GWAS and replication study was just too small to be able to identify robust association findings. At present, autism samples for GWAS or exome sequencing are limited to under 7000 individuals, and there is a considerable need to improve these if we wish to accurately and robustly identify common risk of the effect sizes of $OR=1.1-1.3$. However one would hope that when sample sizes reach critical mass that multiple genome wide significant loci (or in the case of exome studies, multiple significant loss of function variants) will be identified. It is also likely that the GW significant loci and even the nominally significant loci identified to date will strengthen and become more robust in much larger analyses.

My own attempts reported here to increase the sample size by combination of the AGP stage 1 and stage 2 data and generation of pseudo case-controls demonstrate that a signal from one group or sub-phenotype can drive the signal even when combined into a larger sample. The association from the PGC autism GWAS for *MACROD2* as I have suggested, could be driven by the AGP stage 1 data, although the current PGC autism GWAS has not been performed without the AGP stage 1 data. It would be important to test the association completely independent of the AGP Stage 1 data.

As such it is possible that the original association from the AGP stage 1 data for rs4141463 is a case of the 'winners curse', the phenomenon of regression towards the mean that typically affects samples of relatively small size, discussed at length by others (Iles 2008; Zhong & Prentice 2010). For a study involving many tests on a subset of the full population, the thresholds for significance make it likely that the first significant association (the 'winner') will thus report an effect size much larger than is likely to be seen in subsequent replication studies (Ioannidis 2008).

Although the sample sizes in use were quite small for my investigation of the effects of clinical heterogeneity on significance, there are some interesting observations. It makes intuitive sense that the individuals with the more severe diagnosis for autism (e.g. the 'strict' subset) will form a more homogeneous group and thus may enrich for GW significant loci, which became my working hypothesis. I appear to have observed data in support of this hypothesis with the comparison of the AGP stage 1 'strict' data to the other AGP data. I had also hypothesised that markers of a more severe autism presentation, such as non-verbal status and low-IQ would represent a more homogenous sample and thus enrich for GW significant loci, but this was not the case for rs4141463. From the results presented here it would seem that male probands are more informative when looking for association, which could indicate sex-linked effects. Autism is more prevalent in males however which meant that the sample of female probands was much smaller than the males, which could be influencing the difference in GW significance for rs4141463 across male and female probands.

5.5.3 The effect of inheritance model on power of GWAS

The results presented here do not find a striking difference in power between the additive, dominant and recessive inheritance models, although the results favour either the additive or recessive models. The results presented here suggest that the dominant model has poor power to detect significant loci, at least for the AGP dataset and rs4141463. Of note is the fact that the recessive model highlights many more significantly associated markers compared to the other two models, although no individual marker, including rs4141463 meets genome-wide significance. Therefore it would appear that there is utility in using the recessive model in addition to the additive model to identify significant markers at this locus.

5.5.4 Enrichment analysis of putative protein-protein interactors identified via Co-IP experiments

The results of the enrichment analysis show that *MACROD2* and its interacting proteins reveal enrichment in association at the genes *MACROD2* and *USP33* but not for the complete gene list. *USP33* is known to have a role in cell migration and has been shown to interact with *ROBO1* (Yuasa-Kawada *et al.* 2009). *ROBO1* axon guidance receptors have been associated with autism (Anitha *et al.* 2008; Casey *et al.* 2012) and thus this association is biologically plausible. However, as has been discussed in Chapter 4, a repeat of the Co-IP experiments to reduce potential non-specific binding and enrichment for *MACROD2* should be performed. It is important to note that in these data *MACROD2* was not observed and as such was manually added to the enrichment list. As the INRICH program uses association data to create the genomic interval files, the power of the INRICH analysis will increase as the power of the input GWAS improves.

5.5.5 Summary and future directions

This report has provided some additional support for the association at rs4141463 within the *MACROD2* locus and autism, and demonstrated how further investigation of the different phenotypes within the sample examined can reveal subgroups driving the association. I have used imputation to examine whether the addition of increased coverage would identify additional refinement of the ASD signal. The imputation of the *MACROD2* locus in this study demonstrates that imputation using 1000 genomes data can enrich the original association signal and provide additional supporting associated markers. I have also performed an enrichment analysis using the AGP association data and my own data from Co-IP protein pulldown experiments to look for enrichment for particular sets of genes. Whilst I did not identify large numbers of related genes, I have highlighted *USP33* as a biologically plausible candidate.

The association at *MACROD2* does have support from the first AGP study, but as we have discussed, this could be a case of the ‘winner’s curse’ and furthermore the association within the latest PGC ASD GWAS for *MACROD2* could be driven by the AGP stage 1 ‘strict’ sample. However, as we have discussed throughout this report, *MACROD2* is an interesting candidate gene in terms of the variation observed at this locus, which includes large scale common CNVs and smaller scale rare deletions that cross exons. The observation from this report that the signal from the AGP stage 1 analysis is driven by a smaller subset of the data could indicate that association at *MACROD2* is not indicative of

a common variant of small effect size, but could indeed be a rarer variant of small to modest effect. It has been suggested that common variants of large effect do not exist for ASD (Devlin & Scherer 2012) which the results presented here would support, and thus we are brought back to the need for increased sample sizes in autism to be able to answer these burning questions about the importance of common and rare variance.

Over the course of this project the focus of autism genetic studies have moved from studies of common variance via GWAS studies to studies of rare variance via exome and whole genome sequencing and back again. At the moment it would appear that these approaches complement each other, and it is my hope that with increasing sample size, array size and annotation of genes via gene-ontology terms these studies will begin to reveal more candidate genes and biological processes found to be disrupted in ASD probands.

However what we do know from our studies thus far is that the effect sizes of the associated variants are modest and thus we can determine that gene-gene and gene-environment interactions are likely to be important to the genetic aetiology of autism. Unfortunately we are not at a level where we can analyse such large scale interactions in humans, although there are signs from model organisms that such large scale understanding is possible (Bloom *et al.* 2013).

With each passing year we are moving closer towards understanding the genetic contribution to ASD, and this study represents a small contribution to that forward momentum. An in depth analysis of associated variants is always warranted; as any good statistician knows it is only by parsing your data in multiple ways that you can extract the most information from it, and thus this report was such an attempt at extracting additional information from the AGP stage 1 and stage 2 data. This comprehensive study of the *MACROD2* association has provided some supportive evidence of association, but it is clear that much further study and increasing sample sizes are required to understand how variance at this locus may contribute to neuropsychiatric disorders such as ASD and human health in general.

Chapter 6 Discussion of *MACROD2* as a susceptibility gene for autism

The overarching hypothesis of this project was highlighting *MACROD2* as a candidate locus for autism susceptibility and that variation within this gene could be revealed to have an impact upon the neuron and the synapse. Moreover, links between *MACROD2* and processes previously highlighted to be implicated in autism progression would be observed. However, due to the poor characterisation of this gene, much of the basic biology was undocumented, and thus much of this project represents documenting the basic biology of this putative susceptibility gene. Thus this project aimed to investigate the structure, location and function of *MACROD2*, with the aim that this would paint a better picture as to how this gene could have an impact in autism.

Chapter 4 describes these efforts to demonstrate the localisation of *MACROD2* in the neuronal cell model, and I have confirmed that endogenous *MACROD2* is expressed in the cytoplasm and the nucleus, which is in support of the findings from the Co-IP analysis that identified proteins across cytoplasmic and nuclear fractions. The need for confirmation of the localisation of *MACROD2* is illustrated by recent publications that have no confirmation of this fact (Posavec *et al.* 2013). The preliminary study of the impact of *MACROD2* knockdown using siRNA suggests that in cells where the knockdown was the most successful, the associated phenotype is one of reduced expression of neuronal markers and a cell morphology that would indicate an impact on neurite outgrowth and synapse formation, thus providing evidence of the original hypothesis, albeit in a limited analysis. Future study should attempt to replicate these results and document any changes in greater detail, e.g. direct measurement of expression level and neurite length.

Follow up association studies of the *MACROD2* locus and index SNP have revealed at best modest replication. This failure to replicate association is often laid at the feet of sample size, model assessed, study heterogeneity and that of linkage disequilibrium between the genotyped and causative risk locus clouding the true association. I have tried to address each of these commonly embraced apologies for non-replication by exploring association in larger samples, under varying models of association and clinical subgroupings and by using imputation to better characterise this finding. From the data described in Chapter 5 it would appear that the *MACROD2* signal from the AGP stage 1

sample is from a subset of individuals with a more severe autism presentation, within the 'strict' group.

I have also noted, albeit in a small sample from the AGP data, that CNV at the *MACROD2* locus can be broken down into common intronic deletions and rare exonic deletions that segregate with a more severe phenotype. Expanding out my search across the wider literature reveals that in other disorders where CNVs at *MACROD2* have been observed, those CNVs that delete exonic material particularly that of exon 5, also co-segregate with severe phenotype. In fact I have been able to supplement a lot of my 'wet lab' work with supporting information from the wealth of online experimental genomics data deposited in databases for *in silico* analyses. For example, investigation of homologous genes for *MACROD2* across different model organisms suggests that due to the evolutionary conservation of this gene, *MACROD2* may have an essential role. The reasoning for this is that variance in evolutionary conserved regions is often of detriment to the evolutionary fitness of an organism and thus we conclude that there are selective pressures on coding sequence evolution (Eyre-Walker & Keightley 2007; Drummond & Wilke 2008). This led to my investigation of whether 'knock-out' model organisms for *MACROD2* were available. Although knockouts for the *MACROD2* homolog are available, any associated phenotypes have not been assessed, including documentation of the viability and fertility of homozygotes for the knockout in mice.

Whilst it is disappointing that there has been no phenotypic assessment of *MACROD2* knockouts across model organisms, I was able to turn to the large scale public domain genomics projects such as the DGV, UCSC genome browser and Ensembl.org. This meant that I could analyse the CNV burden across multiple studies at a glance. The value of these bioinformatic databases cannot be overstated, particularly in the case of a poorly characterised gene such as *MACROD2*. I have used many other bioinformatic tools extensively in this project, such as the transcription factor search similarity utility 'MatInspector' to supplement the promoter mapping experiments and RNAseq from Ensembl.org data to lend support to the 5' RACE experiments which looked at the gene structure of *MACROD2*. My own experiments and the RNAseq data highlight that *MACROD2* isoform 2 lacks the macro domain. The lack of this domain in the protein product of isoform-2 leads to the question as to whether the MACRO domain is the main

functional site of the protein, and if so, what the function of the *MACROD2* isoform 2 could be. Further study is required to answer this question.

Bioinformatic resources were particularly useful when it came to investigation of the expression of *MACROD2* in the developing and adult human brain. I had attempted to source RNA from specific brain regions, e.g. amygdala, hippocampus, septum, mammillary bodies and cerebellum with an aim to perform a detailed expression profile of this gene across these different regions of the brain. Although it proved difficult to source this material commercially I was able to supplement my experiments by using the online transcriptome database, The Developing Human Brain (<http://www.developinghumanbrain.org/home.html>) which presents data to suggest that *MACROD2* expression levels are greatest during foetal development, in particular in the caudate nucleus and nucleus accumbens. So while the NPG does not have direct access to RNA from specific brain regions, I was still able to investigate the expression of *MACROD2* in specific regions of the human brain during development thanks to these resources.

As we have discussed in Chapter 1, the results of linkage, GWAS, CNV and sequencing studies have implicated cadherins, catenins, cell adhesion and other cytoskeletal molecules, hydrolase enzymes (GTPases), post-translational regulation and chromatin biology in autism aetiology. The eQTL, Co-IP pulldown and LC/ESI-MS analysis, siRNA knockdown, localisation experiments and enrichment analyses were all performed in an effort to elucidate the function or functions of the *MACROD2* protein in the context of neurons and neuronal function through my use of the Myers cortical expression dataset and the creation of the SH-SY5Y differentiated cell model. This enabled cross referencing of highlighted genes and proteins from the LC/ESI-MS and eQTL analyses against lists of previously identified autism susceptibility genes and biological processes thought to be important for the development of autism, and the preliminary knockdown work implicates deficiency of *MACROD2* in the decreased expression of neuronal markers and gross changes in neuronal morphology.

It is unfortunate that doubt is cast over the Co-IP experiments due to the failure to identify *MACROD2* among the list of identified peptide chains. However, the results of this analysis fit extraordinarily well with the previously outlined functions of the macro domain, and highlight a list of genes from biological processes already highlighted in autism. From the Co-IP analysis, *MACROD2* has been shown to interact with cytoskeletal

and cell adhesion molecules that have been associated with autism from the multiple studies discussed in Chapter 1. Karras and colleagues (Karras *et al.* 2005) described the function of the macro domain as an ADP-ribose binding module, an important post-translational modification of proteins that occurs in a variety of biological processes such as DNA repair, transcription, chromatin biology and long-term memory formation (Cohen-Armon *et al.* 2004). If we take each of these functions listed by Cohen-Armon, the LC/ESI-MS analysis has identified protein interactors that meet each of these functions (See Table 6.1). I report an interaction with Protein Phosphatase 1 (also named Spinophilin/Neurabin-2) highlighted as a suppressor of learning and memory. Histone proteins appeared across separate Co-IP pulldowns, in agreement with the highlight of chromatin biology by Cohen-Armon. I report interaction with multiple genes identified as transcriptional and translational regulators, *CORO1C*, *HNRNPC*, *ZKSCAN5* and *ZNF101*. The identification of Ku70 from the LC/ESI-MS is of particular interest, not only because of its role in DNA repair, but because recruitment of *MACROD2* to regions of DNA damage has been previously described (Timinszky *et al.* 2009). hnRNP C Has been found to promote APP (Amyloid precursor protein) translation by competing with FMRP, implicating hnRNP C in neuronal biology (Lee *et al.* 2010).

Table 6.1 A subset of proteins identified by the LC/ESI-MS analysis of the Co-IP pulldown from differentiated SH-SY5Y cell lines match functions of Poly-ADP ribose modifying proteins, highlighted by Cohen-Armon and colleagues (Cohen-Armon *et al.* 2004).

Gene name	Full name	Protein Function
CORO1C	Coronin, actin binding protein, 1C	Signal transduction, gene regulation.
HIST1H1A, HIST1H1B, HIST1H1C, HIST1H1D, HIST1H1E, HIST1H1T	Histone H1.1	Chromatin compaction
hnRNP C, LOC440563, LOC649330	Heterogenous nuclear ribonuclearproteins	RNA binding, pre-mRNA processing (König <i>et al.</i> 2010)
PPP1CA, PPP1CB, PPP1CC, PPP1R9B	Protein phosphatase 1/Spinophilin/Neurabin-2	Protein phosphorylation, regulation of formation and function of dendritic spines (Feng <i>et al.</i> 2000), suppressor of learning and memory (Genoux <i>et al.</i> 2002)
XRCC6	Ku70	Non-homologous end joining of DNA repair, V(D)J recombination, telomere length maintenance and subtelomeric gene silencing
ZKSCAN5	Zinc finger protein with KRAB and SCAN domains 5	Transcriptional regulation
ZNF101	Zinc finger protein 101	Transcriptional regulation (Bellefroid <i>et al.</i> 1993)

Thus from the results presented here not only have I supported the working hypothesis that *MACROD2* would bind to histone proteins, I have also demonstrated support for the idea that macro-domain containing proteins interact with proteins that have roles in the processes highlighted by Cohen-Armon and colleagues. I had also hypothesised that *MACROD2* would interact with sirtuin proteins, due to the observation by others that *MACROD2* has a function as an AADPR-deacetylase (Chen *et al.* 2011). Although I did not observe any sirtuin proteins in the LC/ESI-MS analysis, many sirtuins interact with cytoplasmic and cytoskeletal proteins (North & Verdin 2004), which I have recorded an abundance for in my analyses. Thus I cannot rule out the hypothesis that *MACROD2* interacts with sirtuin proteins. However, the association of *MACROD2* with cadherins, catenins and other cytoskeletal and cell adhesion proteins is interesting considering that these proteins have been implicated in the aetiology of autism by other studies.

If we look at the list of identified proteins and cross reference that with the publications that link *MACROD2* to phenotypes outside of autism, I found proteins that could explain an association with *MACROD2*. The best example of this is the finding of Spectrin alpha chain, previously associated with early onset epileptic encephalopathies (Mastrangelo & Leuzzi 2012), a disorder associated with CNVs in *MACROD2* (Bendavid *et al.* 2009). Therefore I defend the Co-IP and LC/ESI-MS results in that the proteins identified are in support of many of the previous findings for *MACROD2* and also may begin to explain what proteins *MACROD2* interacts with to affect these phenotypes or biological processes. It is my hope that future study will further elucidate these different putative interactions.

As I have discussed previously, a function for the *MACROD2* protein as an AADPR deacetylase has been documented. The latest research describes another role for *MACROD2* as a mono-ADP-ribosyl-hydrolase, and highlights *MACROD2* as a regulator of GSK3 (Rosenthal *et al.* 2013). There is much evidence to implicate GSK3 in autism susceptibility. Glycogen synthase kinase-3 (GSK3) is hyperactive in the brains of *Fmr1* knockout mice, a mouse model for autism, and this impaired inhibitory regulation of GSK3 was reported to influence social preference and anxiety-related behaviours in these mice (Mines *et al.* 2010). Furthermore, drugs that act on serotonin neurotransmission have been found to regulate GSK3 in mouse models, with GSK3 inhibition rescuing behavioural

abnormalities in serotonin deficient mice (Latapy *et al.* 2012). If we cross reference the findings of the eQTL analysis against these latest publications, glycogen phosphorylase, brain (*PYGB*) was the most significant eQTL identified, which could suggest that *MACROD2* (or indeed *MACROD2-AS1*) expression can influence glycogen metabolism in the brain. I had hypothesised that disruption of the deacetylase activity of *MACROD2* could impact an energy hungry organ such as the brain, due to the connection that has been made between protein deacetylation and energy metabolism (Imai and Guarente 2010). It would appear that recent publications lend credence to this hypothesis, which had little grounding previously.

Cross over of genes from the eQTL analysis and LC/ESI-MS analysis identifies several genes, including TPM4 (tropomyosin 4) and multiple protein phosphatases from the Protein phosphatase 1 family. Taking this recent work and my own work presented here it would suggest that a much clearer role for *MACROD2* in cell biology and disease susceptibility is being realised. Further investigation of the role of *MACROD2* in cytoskeletal and cell adhesion biochemistry, potentially via NAD⁺ metabolism is warranted, and a role for *MACROD2* in phosphorylation of proteins, an important post translational modification, and of glycogen metabolism, a seemingly important process in terms of behaviour, is needed in future studies. This is not the first time that glycobiology has been highlighted in autism (van der Zwaag *et al.* 2009).

Since the beginning of the study of polygenic disorders, that is, disorders that are hypothesised to involve more than one gene in their causation, many different approaches have been developed to test for an association between variation in a set of genes and a disorder. Early studies were of modest sample size and discovery was restricted to highly penetrant variants. However, for variants which have a lower effect size the sample size required to have sufficient power to discover and confirm these variants is in the thousands to the tens of thousands of individuals. The collection of samples on this scale is an expensive endeavour and has led to numerous collaborative collections where the early modest studies have been combined and added to.

The work presented in this study came about as a result of the GWAS by the Autism Genome Project, an international consortium of over 50 centres, who combined data in a study of over 1500 families with a child with autism (Anney *et al.*, 2010). This GWAS identified *MACROD2* as a susceptibility locus for ASD. Replication of this results

from GWAS has been poor, most likely is that the sample sizes for discovery and validation are still too limited to detect the variants of small effect that are hypothesised to be behind the genetic aetiology of autism. Consequently, the common disease-common variant hypothesis tested for by GWAS has been questioned by many to have been disproved. The lack of common alleles of modest to strong effect alongside recent discoveries from rare variant studies has fuelled debate about the success of GWAS. For a time the 'death' of GWAS was heralded as the focus was turned to CNV studies, and more recently the promise of exome and whole genome sequencing studies (Cirulli & Goldstein 2010). However, writing these obituaries is premature. Recent results from the other complex traits including samples of over 45,000 individuals from the PGC schizophrenia GWAS have demonstrated, not unsurprisingly, that highly powered studies will yield strong and robust association findings. Disappointingly the early promise of identifying rare variants of large effect has not been realised. Research into CNV and SNV variation suggests more modest effect sizes for these very rare variants and as such these studies are equally burdened by a need for greater sample sizes (Nature Editorial, 2013).

Therefore one of the critical limits to identifying risk variants in the study of autism is sample size. The PGC GWAS studies have been a great boon to the use of GWAS in the field of neuropsychiatric genetics, as we are now seeing sample sizes large enough to detect the hypothesised common variants of modest to weak effect. Recent results from the PGC autism GWAS (European), which is the largest scale GWA in autism performed to date, across 5305 cases and 5303 controls, finds association at the *MACROD2* locus, where *MACROD2* is reported as the 11th most significant hit. The 9th most significant hit is PARP 11 (poly-ADP-ribose polymerase family, member 11), and so it seems ADP-ribose metabolism is being highlighted in autism not just from the associations with *MACROD2*.

Recently, the ENCODE project revealed that a much larger amount of the human genome (80.4%) displays some degree of functionality, which has implications for the study of neuropsychiatric genetics (Kavanagh *et al.* 2013). This is of interest to me in light of the fact that the association at *MACROD2* is at an intronic site, and the eQTL analysis was an attempt at investigating how variation at this locus, which contains the non-coding RNA *MACROD2-AS1*, may be influencing gene expression. The ENCODE studies highlight the utility of eQTL analyses in investigating the regulatory effects of intronic loci in this way. Supplement to this, in Chapter 5 I discussed the phenomenon of 'missing

heritability'. When we also consider that we have yet to explore gene-gene and gene-environment effects it remains clear that there are many layers to the genetic contribution to disease susceptibility that we have yet to investigate in detail.

A challenge for genomic studies is our ability to accurately annotate and predict the impact of variation. This problem is particularly pertinent to exome and genome sequencing studies. For genome studies, researchers are generally comfortable annotating the exome, the function of variation in which is best described if they give rise to loss of function variants i.e. those which introduce a stop codon or truncating mutation. This means that variants that fall outside of exomes are poorly investigated. Advances in ENCODE-like studies are improving understanding of regulatory variation; however the differential function of alleles will require experiments such as AEI and eQTL analyses as utilised in this study.

As I have discussed, there is much debate in the field regarding whether rare or common variants play the more significant role in such diseases; functional validation of identified variants will aid in settling this debate. This study is a functional characterisation of one gene, and as I have documented throughout this report, the annotation and supporting information for this gene remains poor. This project has revealed to me the limits of our understanding of many genes, and makes it clear that larger scale functional genomics is still in its infancy. Perhaps the future may yield technologies for large scale proteomics and functional annotation of genes, but I would suggest that detailed databases and sharing of information from functional characterisation of all transcribed regions, not just of genes, will yield the results that we need to map complex disease.

At the outset of this study, with little evidence bar the findings from association and CNV studies, one hypothesis was that *MACROD2* has a function in the survivability of neurons and/or the connections between them. Variation at this locus had been associated with brain volume, brain patterning and epilepsy, but an in depth study of this locus had not been performed. The preliminary siRNA knockdown study, albeit limited in scale, finds evidence in support of this hypothesis. In fact the results across all three experimental chapters find evidence for a role for *MACROD2* in overall brain biology and in neuropsychiatric conditions such as autism. However this increase in the basic knowledge of this protein only leads to more questions which should drive further study. The past three and a half years of this project have seen great advances in the field of

psychiatric genetics and I anticipate the advancement of the field over the coming years with excitement at what may be uncovered. The brain is a fascinating organ and the genetic impact on behaviour is complex and poorly understood at present. In time, it is my hope that as we build up our knowledge of the brain and behaviour, that we will discover new and better treatments for neuropsychiatric conditions such as autism spectrum disorder.

References

- Abrahams, B.S. & Geschwind, D.H., 2008. Advances in autism genetics: on the threshold of a new neurobiology. *Nat Rev Genet*, 9(5), pp.341–355. Available at: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=18414403.
- Adams, R.H., Betz, H. & Püschel, A.W., 1996. A novel class of murine semaphorins with homology to thrombospondin is differentially expressed during early embryogenesis. *Mechanisms of development*, 57(1), pp.33–45. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/8817451> [Accessed January 18, 2013].
- Alto, N.M. et al., 2003. Bioinformatic design of A-kinase anchoring protein-in silico: a potent and selective peptide antagonist of type II protein kinase A anchoring. *Proceedings of the National Academy of Sciences of the United States of America*, 100(8), pp.4445–50. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=153575&tool=pmcentrez&rendertype=abstract> [Accessed March 20, 2013].
- Amir, R.E. et al., 1999. Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. *Nature genetics*, 23(2), pp.185–8. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/10508514> [Accessed February 17, 2013].
- Anitha, A. et al., 2008. Genetic analyses of roundabout (ROBO) axon guidance receptors in autism. *American journal of medical genetics. Part B, Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics*, 147B(7), pp.1019–27. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/18270976> [Accessed March 6, 2013].
- Anney, R., Klei, L., Pinto, D., Regan, R., Conroy, J., Magalhaes, T.R., et al., 2010. A genome-wide scan for common alleles affecting risk for autism. *Human Molecular Genetics*, 19(20), pp.4072–4082. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2947401&tool=pmcentrez&rendertype=abstract>.
- Anney, R., Klei, L., Pinto, D., Regan, R., Conroy, J., Tiago, R., et al., 2010. A genome-wide scan for common alleles affecting risk for autism. *Human molecular genetics*, 19(20), pp.4072–82. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/20663923> [Accessed November 16, 2012].
- Anney, R. et al., 2012. Individual common variants exert weak effects on the risk for autism spectrum disorderspi. *Human molecular genetics*, 21(21), pp.4781–92. Available at: <http://hmg.oxfordjournals.org/content/21/21/4781.short> [Accessed March 1, 2013].
- Anney, R.J. et al., 2002. Characterisation, mutation detection, and association analysis of alternative promoters and 5' UTRs of the human dopamine D3 receptor gene in schizophrenia. *Mol Psychiatry*, 7(5), pp.493–502. Available at:

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=12082567 [Accessed January 16, 2013].

Anney, R.J.L. et al., 2011. Gene-ontology enrichment analysis in two independent family-based samples highlights biologically plausible processes for autism spectrum disorders. *Eur J Hum Genet*, 19(10), pp.1082–9. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3190264&tool=pmcentrez&rendertype=abstract> [Accessed November 16, 2012].

Anon, 1998. A full genome screen for autism with evidence for linkage to a region on chromosome 7q. International Molecular Genetic Study of Autism Consortium. *Human molecular genetics*, 7(3), pp.571–8. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/9546821> [Accessed January 17, 2013].

Anon, 2001a. A genomewide screen for autism: strong evidence for linkage to chromosomes 2q, 7q, and 16p. *American journal of human genetics*, 69(3), pp.570–81. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1235486&tool=pmcentrez&rendertype=abstract> [Accessed January 17, 2013].

Anon, 2001b. An autosomal genomic screen for autism. *American journal of medical genetics*, 105(8), pp.609–15. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/11811142> [Accessed January 17, 2013].

Arking, D.E. et al., 2008. A common genetic variant in the neurexin superfamily member CNTNAP2 increases familial risk of autism. *American journal of human genetics*, 82(1), pp.160–4. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2253968&tool=pmcentrez&rendertype=abstract> [Accessed November 2, 2012].

Asanuma, K. et al., 2005. Synaptopodin regulates the actin-bundling activity of alpha-actinin in an isoform-specific manner. *The Journal of clinical investigation*, 115(5), pp.1188–98. Available at: <http://www.jci.org/articles/view/23371> [Accessed March 20, 2013].

Association, A.P., 1994. *Diagnostic and statistical manual of mental disorders: 4th edition (DSM-IV)*, Washington, DC: American Psychiatric Press, Inc.

Auranen, M. et al., 2002. A genomewide screen for autism-spectrum disorders: evidence for a major susceptibility locus on chromosome 3q25-27. *American journal of human genetics*, 71(4), pp.777–90. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=378535&tool=pmcentrez&rendertype=abstract> [Accessed January 17, 2013].

Bailey, A. et al., 1995. Autism as a Strongly Genetic Disorder - Evidence from a British Twin Study. *Psychological Medicine*, 25(1), pp.63–77. Available at: <Go to ISI>://A1995QL43300007.

- Baird, T.D. & August, G.J., 1985. Familial heterogeneity in infantile autism. *J Autism Dev Disord*, 15(3), pp.315–321. Available at: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=4030664.
- Bakkaloglu, B. et al., 2008. Molecular cytogenetic analysis and resequencing of contactin associated protein-like 2 in autism spectrum disorders. *American journal of human genetics*, 82(1), pp.165–73. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2253974&tool=pmcentrez&rendertype=abstract> [Accessed January 18, 2013].
- Baranzini, S.E. et al., 2009. Genome-wide association analysis of susceptibility and clinical phenotype in multiple sclerosis. *Hum Mol Genet*, 18(4), pp.767–778. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/19010793> [Accessed November 2, 2012].
- Barr, A.R. & Gergely, F., 2008. MCAK-independent functions of ch-Tog/XMAP215 in microtubule plus-end dynamics. *Molecular and cellular biology*, 28(23), pp.7199–211. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2593372&tool=pmcentrez&rendertype=abstract> [Accessed March 10, 2013].
- Barrett, S. et al., 1999. An autosomal genomic screen for autism. Collaborative linkage study of autism. *American journal of medical genetics*, 88(6), pp.609–15. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/10581478> [Accessed January 17, 2013].
- Bassell, G.J. & Warren, S.T., 2008. Fragile X syndrome: loss of local mRNA regulation alters synaptic development and function. *Neuron*, 60(2), pp.201–14. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/18957214> [Accessed February 11, 2013].
- Bellefroid, E.J. et al., 1993. Clustered organization of homologous KRAB zinc-finger genes with enhanced expression in human T lymphoid cells. *The EMBO journal*, 12(4), pp.1363–74. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=413348&tool=pmcentrez&rendertype=abstract> [Accessed March 20, 2013].
- Bendavid, C. et al., 2009. Array-CGH analysis indicates a high prevalence of genomic rearrangements in holoprosencephaly: an updated map of candidate loci. *Hum Mutat*, 30(8), pp.1175–1182. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/19431187> [Accessed January 16, 2013].
- Berger, F., Ramírez-Hernández, M.H. & Ziegler, M., 2004. The new life of a centenarian: signalling functions of NAD(P). *Trends in biochemical sciences*, 29(3), pp.111–8. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/15003268> [Accessed March 24, 2013].
- Betancur, C., 2011. Etiological heterogeneity in autism spectrum disorders: more than 100 genetic and genomic disorders and still counting. *Brain research*, 1380, pp.42–77. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/21129364> [Accessed March 4, 2013].

- Bloom, J.S. et al., 2013. Finding the sources of missing heritability in a yeast cross. *Nature*, 494(7436), pp.234–7. Available at: <http://www.nature.com/nature/journal/v494/n7436/full/nature11867.html#> [Accessed March 1, 2013].
- Bonaglia, M.C. et al., 2011. Molecular mechanisms generating and stabilizing terminal 22q13 deletions in 44 subjects with Phelan/McDermid syndrome. N. B. Spinner, ed. *PLoS genetics*, 7(7), p.e1002173. Available at: <http://dx.plos.org/10.1371/journal.pgen.1002173> [Accessed February 24, 2013].
- Bond, J. & Woods, C.G., 2006. Cytoskeletal genes regulating brain size. *Current opinion in cell biology*, 18(1), pp.95–101. Available at: <http://dx.doi.org/10.1016/j.jceb.2005.11.004> [Accessed March 8, 2013].
- Bourgeron, T., 2009. A synaptic trek to autism. *Current opinion in neurobiology*, 19(2), pp.231–4. Available at: <http://dx.doi.org/10.1016/j.conb.2009.06.003> [Accessed February 21, 2013].
- Bozdagi, O. et al., 2010. Haploinsufficiency of the autism-associated Shank3 gene leads to deficits in synaptic function, social interaction, and social communication. *Molecular autism*, 1(1), p.15. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3019144&tool=pmcentrez&rendertype=abstract> [Accessed February 19, 2013].
- Bradley, W.E. et al., 2010. Hotspots of large rare deletions in the human genome. *PLoS One*, 5(2), p.e9401. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/20195527>.
- Brantl, S., 2002. Antisense-RNA regulation and RNA interference. *Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression*, 1575(1-3), pp.15–25. Available at: [http://dx.doi.org/10.1016/S0167-4781\(02\)00280-4](http://dx.doi.org/10.1016/S0167-4781(02)00280-4) [Accessed March 26, 2013].
- Browning, B.L. & Browning, S.R., 2009. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *American journal of human genetics*, 84(2), pp.210–23. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2668004&tool=pmcentrez&rendertype=abstract> [Accessed November 5, 2012].
- Budimirovic, D.B. & Kaufmann, W.E., 2011. What can we learn about autism from studying fragile X syndrome? *Developmental neuroscience*, 33(5), pp.379–94. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3254037&tool=pmcentrez&rendertype=abstract> [Accessed February 20, 2013].
- Bürkle, A., 2005. Poly(ADP-ribose). The most elaborate metabolite of NAD⁺. *The FEBS journal*, 272(18), pp.4576–89. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/16156780> [Accessed March 24, 2013].
- Campbell, D.B. et al., 2006. A genetic variant that disrupts MET transcription is associated with autism. *Proceedings of the National Academy of Sciences of the United States of*

America, 103(45), pp.16834–9. Available at:
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1838551&tool=pmcentrez&rendertype=abstract> [Accessed December 3, 2012].

Casey, J.P. et al., 2012. A novel approach of homozygous haplotype sharing identifies candidate genes in autism spectrum disorder. *Human genetics*, 131(4), pp.565–79. Available at:
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3303079&tool=pmcentrez&rendertype=abstract> [Accessed October 1, 2013].

Chen, D. et al., 2011. Identification of macrodomain proteins as novel O-acetyl-ADP-ribose deacetylases. *Journal of Biological Chemistry*, 286(15), pp.13261–13271. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/21257746> [Accessed December 21, 2012].

Cohen-Armon, M. et al., 2004. Long-term memory requires polyADP-ribosylation. *Science*, 304(5678), pp.1820–1822. Available at:
<http://www.ncbi.nlm.nih.gov/pubmed/15205535> [Accessed January 16, 2013].

Consortium, W.T.C.C., 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145), pp.661–678. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/17554300>.

Corda, D. & Di Girolamo, M., 2003. Functional aspects of protein mono-ADP-ribosylation. *The EMBO journal*, 22(9), pp.1953–8. Available at:
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=156081&tool=pmcentrez&rendertype=abstract> [Accessed March 24, 2013].

Le Couteur, A. et al., 1989. Autism diagnostic interview: a standardized investigator-based instrument. *J Autism Dev Disord*, 19(3), pp.363–387. Available at:
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=2793783.

Curran, S. et al., 2011. No association between a common single nucleotide polymorphism, rs4141463, in the MACROD2 gene and autism spectrum disorder. *American journal of medical genetics. Part B, Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics*, 156B(6), pp.633–9. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/21656903> [Accessed January 16, 2013].

Darnell, J.C. et al., 2011. FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell*, 146(2), pp.247–61. Available at:
<http://dx.doi.org/10.1016/j.cell.2011.06.013> [Accessed February 13, 2013].

Davison, E.J. et al., 2005. Deletion at chromosome band 20p12.1 in colorectal cancer revealed by high resolution array comparative genomic hybridization. *Genes, chromosomes & cancer*, 44(4), pp.384–91. Available at:
<http://www.ncbi.nlm.nih.gov/pubmed/16110499> [Accessed March 7, 2013].

- Debette, S. et al., 2010. Genome-wide association studies of MRI-defined brain infarcts: meta-analysis from the CHARGE Consortium. *Stroke*, 41(2), pp.210–217. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/20044523>.
- Debette, S. et al., 2010. Genome-wide association studies of MRI-defined brain infarcts: meta-analysis from the CHARGE Consortium. *Stroke; a journal of cerebral circulation*, 41(2), pp.210–7. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2923092&tool=pmcentrez&rendertype=abstract> [Accessed November 2, 2012].
- Delaneau, O., Zagury, J.-F. & Marchini, J., 2013. Improved whole-chromosome phasing for disease and population genetic studies. *Nature methods*, 10(1), pp.5–6. Available at: <http://dx.doi.org/10.1038/nmeth.2307> [Accessed February 1, 2013].
- Devlin, B. & Scherer, S.W., 2012. Genetic architecture in autism spectrum disorder. *Current opinion in genetics & development*, 22(3), pp.229–37. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/22463983> [Accessed February 27, 2013].
- Drummond, D.A. & Wilke, C.O., 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell*, 134(2), pp.341–52. Available at: <http://dx.doi.org/10.1016/j.cell.2008.05.042> [Accessed March 11, 2013].
- Durand, C.M. et al., 2007. Mutations in the gene encoding the synaptic scaffolding protein SHANK3 are associated with autism spectrum disorders. *Nat Genet*, 39(1), pp.25–27. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2082049&tool=pmcentrez&rendertype=abstract> [Accessed November 6, 2012].
- Ecker, C. et al., 2010. Describing the brain in autism in five dimensions--magnetic resonance imaging-assisted diagnosis of autism spectrum disorder using a multiparameter classification approach. *J Neurosci*, 30(32), pp.10612–10623. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/20702694> [Accessed October 26, 2012].
- Encinas, M. et al., 2000. Sequential treatment of SH-SY5Y cells with retinoic acid and brain-derived neurotrophic factor gives rise to fully differentiated, neurotrophic factor-dependent, human neuron-like cells. *J Neurochem*, 75(3), pp.991–1003. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/10936180>.
- Estivill, X. & Armengol, L., 2007. Copy number variants and common disorders: filling the gaps and exploring complexity in genome-wide association studies. E. M. C. Fisher, ed. *PLoS genetics*, 3(10), pp.1787–99. Available at: <http://dx.plos.org/10.1371/journal.pgen.0030190> [Accessed March 13, 2013].
- Eyre-Walker, A. & Keightley, P.D., 2007. The distribution of fitness effects of new mutations. *Nature reviews. Genetics*, 8(8), pp.610–8. Available at: <http://dx.doi.org/10.1038/nrg2146> [Accessed March 1, 2013].

- Feng, J. et al., 2000. Spinophilin regulates the formation and function of dendritic spines. *Proceedings of the National Academy of Sciences*, 97(16), pp.9287–9292. Available at: <http://www.pnas.org/content/97/16/9287.full> [Accessed March 20, 2013].
- Folstein, S. & Rutter, M., 1977. Infantile autism: a genetic study of 21 twin pairs. *J Child Psychol Psychiatry*, 18(4), pp.297–321. Available at: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=562353.
- Fombonne, E., 2005. Epidemiology of autistic disorder and other pervasive developmental disorders. *J Clin Psychiatry*, 66 Suppl 1(suppl 10), pp.3–8. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/16401144>.
- Frapppier, T., Stetzkowski-Marden, F. & Pradel, L.A., 1991. Interaction domains of neurofilament light chain and brain spectrin. *The Biochemical journal*, 275 (Pt 2, pp.521–7. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1150082&tool=pmcentrez&rendertype=abstract> [Accessed March 20, 2013].
- Gai, X. et al., 2012. Rare structural variation of synapse and neurotransmission genes in autism. *Molecular psychiatry*, 17(4), pp.402–11. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3314176&tool=pmcentrez&rendertype=abstract> [Accessed March 11, 2013].
- Genoux, D. et al., 2002. Protein phosphatase 1 is a molecular constraint on learning and memory. *Nature*, 418(6901), pp.970–5. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/12198546> [Accessed March 1, 2013].
- Gerards, M. et al., 2010. Defective complex I assembly due to C20orf7 mutations as a new cause of Leigh syndrome. *Journal of medical genetics*, 47(8), pp.507–12. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2921275&tool=pmcentrez&rendertype=abstract> [Accessed March 6, 2013].
- Geschwind, D.H., 2011. Genetics of autism spectrum disorders. *Trends Cogn Sci*, 15(9), pp.409–416. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/21855394> [Accessed October 29, 2012].
- Gharani, N. et al., 2004. Association of the homeobox transcription factor, ENGRAILED 2, 3, with autism spectrum disorder. *Molecular psychiatry*, 9(5), pp.474–84. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/15024396> [Accessed December 3, 2012].
- Giannone, R.J. et al., 2010. The protein network surrounding the human telomere repeat binding factors TRF1, TRF2, and POT1. *PloS one*, 5(8), p.e12407. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2928292&tool=pmcentrez&rendertype=abstract> [Accessed February 14, 2013].
- Glessner, J.T. et al., 2009. Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. *Nature*, 459(7246), pp.569–73. Available at:

<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2925224&tool=pmcentrez&rendertype=abstract> [Accessed March 15, 2013].

Green, R.A. et al., 2011. A high-resolution *C. elegans* essential gene network based on phenotypic profiling of a complex tissue. *Cell*, 145(3), pp.470–82. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3086541&tool=pmcentrez&rendertype=abstract> [Accessed March 13, 2013].

Griesi-Oliveira, K. et al., 2013. Stem Cells as a Good Tool to Investigate Dysregulated Biological Systems in Autism Spectrum Disorders. *Autism research : official journal of the International Society for Autism Research*, pp.1–8. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/23801657> [Accessed August 6, 2013].

Hallmayer, J. et al., 2011. Genetic Heritability and Shared Environmental Factors Among Twin Pairs With Autism. *Arch Gen Psychiatry*. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/21727249>.

Hoepfner, S. et al., 2005. Modulation of receptor recycling and degradation by the endosomal kinesin KIF16B. *Cell*, 121(3), pp.437–50. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/15882625> [Accessed November 6, 2012].

Hoff, K.G. & Wolberger, C., 2005. Getting a grip on O-acetyl-ADP-ribose. *Nat Struct Mol Biol*, 12(7), pp.560–561. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/15999106>.

Hong, W. et al., 1997. Association of U2 snRNP with the spliceosomal complex E. *Nucleic acids research*, 25(2), pp.354–61. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=146436&tool=pmcentrez&rendertype=abstract> [Accessed March 5, 2013].

Hoogendoorn, B. et al., 2000. Cheap, accurate and rapid allele frequency estimation of single nucleotide polymorphisms by primer extension and DHPLC in DNA pools. *Human genetics*, 107(5), pp.488–93. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/11140947> [Accessed March 12, 2013].

Hosack, D.A. et al., 2003. Identifying biological themes within lists of genes with EASE. *Genome biology*, 4(10), p.R70. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=328459&tool=pmcentrez&rendertype=abstract> [Accessed March 18, 2013].

Hsieh, J.J.-D., Cheng, E.H.-Y. & Korsmeyer, S.J., 2003. Taspase1: a threonine aspartase required for cleavage of MLL and proper HOX gene expression. *Cell*, 115(3), pp.293–303. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/14636557> [Accessed March 5, 2013].

Huang, D.W., Sherman, B.T. & Lempicki, R.A., 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols*, 4(1), pp.44–57. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/19131956> [Accessed February 27, 2013].

- Huang, J. et al., 2012. 1000 Genomes-based imputation identifies novel and refined associations for the Wellcome Trust Case Control Consortium phase 1 Data. *European journal of human genetics EJHG*, 20(7), pp.1–5. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/22293688>.
- Iles, M.M., 2008. What can genome-wide association studies tell us about the genetics of common disease? E. M. C. Fisher, ed. *PLoS genetics*, 4(2), p.e33. Available at: <http://dx.plos.org/10.1371/journal.pgen.0040033> [Accessed March 4, 2013].
- Imai, S. & Guarente, L., 2010. Ten years of NAD-dependent SIR2 family deacetylases: implications for metabolic diseases. *Trends in pharmacological sciences*, 31(5), pp.212–20. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/20226541> [Accessed November 28, 2012].
- International Molecular Genetic Study of Autism Consortium, 2001. Further characterization of the autism susceptibility locus AUTS1 on chromosome 7q. *Human Molecular Genetics*, 10(9), pp.973–982. Available at: <http://hmg.oxfordjournals.org/content/10/9/973.short> [Accessed January 17, 2013].
- Ioannidis, J.P.A., 2008. Why most discovered true associations are inflated. *Epidemiology (Cambridge, Mass.)*, 19(5), pp.640–8. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/18633328> [Accessed March 9, 2013].
- Iossifov, I. et al., 2012. De novo gene disruptions in children on the autistic spectrum. *Neuron*, 74(2), pp.285–99. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/22542183> [Accessed February 27, 2013].
- Jamain, S. et al., 2003. Mutations of the X-linked genes encoding neuroligins NLGN3 and NLGN4 are associated with autism. *Nat Genet*, 34(1), pp.27–29. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/12669065>.
- Jankevicius, G. et al., 2013. A family of macrodomain proteins reverses cellular mono-ADP-ribosylation. *Nature structural & molecular biology*, advance on. Available at: <http://dx.doi.org/10.1038/nsmb.2523> [Accessed March 15, 2013].
- Kamath, R.S. et al., 2003. Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature*, 421(6920), pp.231–7. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/12529635> [Accessed February 27, 2013].
- Karras, G.I. et al., 2005. The macro domain is an ADP-ribose binding module. *Embo Journal*, 24(11), pp.1911–1920. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/15902274> [Accessed December 16, 2012].
- Katayama, S. et al., 2005. Antisense transcription in the mammalian transcriptome. *Science (New York, N.Y.)*, 309(5740), pp.1564–6. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/16141073> [Accessed March 5, 2013].
- Kim, H.-G. et al., 2008. Disruption of neurexin 1 associated with autism spectrum disorder. *American journal of human genetics*, 82(1), pp.199–207. Available at:

<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2253961&tool=pmcentrez&rendertype=abstract> [Accessed March 7, 2013].

Klein, R.J. et al., 2005. Complement factor H polymorphism in age-related macular degeneration. *Science (New York, N.Y.)*, 308(5720), pp.385–9. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1512523&tool=pmcentrez&rendertype=abstract> [Accessed September 19, 2013].

Kohannim, O. et al., 2012. Discovery and Replication of Gene Influences on Brain Structure Using LASSO Regression. *Frontiers in neuroscience*, 6(August), p.115. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3412288&tool=pmcentrez&rendertype=abstract> [Accessed November 19, 2012].

König, J. et al., 2010. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nature structural & molecular biology*, 17(7), pp.909–15. Available at: <http://dx.doi.org/10.1038/nsmb.1838> [Accessed February 28, 2013].

L. Abbasi Moheb L. R. Jensen, L. Nouri Vahid, S. Ghadami, K. Wrogemann, H. Hu, R. Weissmann, M. Mohseni, K. Kahrizi, A. Tzschach, D. Wieczorek, M. Garshasbi, S. Cirak, S. Abedini, H. Najmabadi, H. Ropers, S. Sigrist, A. Kuss, S.M., 2011. Mutations in the NSUN2 and ZNF526 genes cause autosomal recessive intellectual disability in Middle Eastern populations with elevated frequency. *Annual Meeting of the European Society of Human Genetics, Amsterdam, The Netherlands*.

Lander, E. & Kruglyak, L., 1995. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nature genetics*, 11(3), pp.241–7. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/7581446> [Accessed December 6, 2012].

LaSalle, J.M., 2007. The odyssey of MeCP2 and parental imprinting. *Epigenetics : official journal of the DNA Methylation Society*, 2(1), pp.5–10. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1866173&tool=pmcentrez&rendertype=abstract> [Accessed February 20, 2013].

Latapy, C. et al., 2012. Selective deletion of forebrain glycogen synthase kinase 3 β reveals a central role in serotonin-sensitive anxiety and social behaviour. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 367(1601), pp.2460–74. Available at: <http://rstb.royalsocietypublishing.org/content/367/1601/2460.short> [Accessed February 28, 2013].

Lee, E.K. et al., 2010. hnRNP C promotes APP translation by competing with FMRP for APP mRNA recruitment to P bodies. *Nature structural & molecular biology*, 17(6), pp.732–9. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2908492&tool=pmcentrez&rendertype=abstract> [Accessed March 21, 2013].

- Lee, P.H. et al., 2012. INRICH: interval-based enrichment analysis for genome-wide association studies. *Bioinformatics (Oxford, England)*, 28(13), pp.1797–9. Available at: <http://bioinformatics.oxfordjournals.org/content/28/13/1797.abstract> [Accessed March 12, 2013].
- Lesca, G. et al., 2012. Epileptic encephalopathies of the Landau-Kleffner and continuous spike and waves during slow-wave sleep types: genomic dissection makes the link with autism. *Epilepsia*, 53(9), pp.1526–38. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/22738016> [Accessed January 16, 2013].
- Lesch, K.P. et al., 2012. Dances with black widow spiders: Dysregulation of glutamate signalling enters centre stage in ADHD. *European neuropsychopharmacology : the journal of the European College of Neuropsychopharmacology*, null(null). Available at: <http://dx.doi.org/10.1016/j.euroneuro.2012.07.013> [Accessed February 5, 2013].
- Levy, D. et al., 2011. Rare de novo and transmitted copy-number variation in autistic spectrum disorders. *Neuron*, 70(5), pp.886–897. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/21658582> [Accessed November 1, 2012].
- Li, Y. et al., 2010. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic epidemiology*, 34(8), pp.816–34. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3175618&tool=pmcentrez&rendertype=abstract> [Accessed February 4, 2013].
- Liou, G.G. et al., 2005. Assembly of the SIR complex and its regulation by O-acetyl-ADP-ribose, a product of NAD-dependent histone deacetylation. *Cell*, 121(4), pp.515–527. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/15907466>.
- Liu, J. et al., 2001. A genomewide screen for autism susceptibility loci. *American journal of human genetics*, 69(2), pp.327–40. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1235325&tool=pmcentrez&rendertype=abstract> [Accessed January 17, 2013].
- Loo, S.K. et al., 2012. Genome-wide association study of intelligence: additive effects of novel brain expressed genes. *Journal of the American Academy of Child and Adolescent Psychiatry*, 51(4), pp.432–440.e2. Available at: <http://dx.doi.org/10.1016/j.jaac.2012.01.006> [Accessed March 11, 2013].
- Lord, C. et al., 1989. Autism diagnostic observation schedule: A standardized observation of communicative and social behavior. *Journal of Autism and Developmental Disorders*, 19(2), pp.185–212. Available at: <http://www.springerlink.com/index/10.1007/BF02211841> [Accessed January 17, 2013].
- Lord, C. et al., 2000. The autism diagnostic observation schedule-generic: a standard measure of social and communication deficits associated with the spectrum of autism. *J Autism Dev Disord*, 30(3), pp.205–223. Available at: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=11055457.

- Lord, C., Rutter, M. & Le Couteur, A., 1994. Autism Diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *J Autism Dev Disord*, 24(5), pp.659–685. Available at: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=7814313.
- Losh, M. et al., 2009. Neuropsychological profile of autism and the broad autism phenotype. *Archives of general psychiatry*, 66(5), pp.518–26. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2699548&tool=pmcentrez&rendertype=abstract>.
- Ma, D. et al., 2009. A Genome-wide Association Study of Autism Reveals a Common Novel Risk Locus at 5p14. 1. *Annals of human ...*, 73(Pt 3), pp.263–273. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/19456320> [Accessed January 17, 2013].
- Maas, N.M.C. et al., 2007. The C20orf133 gene is disrupted in a patient with Kabuki syndrome. *Journal of medical genetics*, 44(9), pp.562–9. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2597955&tool=pmcentrez&rendertype=abstract> [Accessed January 16, 2013].
- Maher, B., 2008. Personal genomes: The case of the missing heritability. *Nature*, 456(7218), pp.18–21. Available at: <http://www.nature.com/news/2008/081105/full/456018a.html> [Accessed November 10, 2012].
- Manolio, T.A. et al., 2009. Finding the missing heritability of complex diseases. *Nature*, 461(7265), pp.747–53. Available at: <http://dx.doi.org/10.1038/nature08494> [Accessed January 29, 2013].
- Marchini, J. et al., 2007. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature genetics*, 39(7), pp.906–13. Available at: <http://dx.doi.org/10.1038/ng2088> [Accessed February 28, 2013].
- Marchini, J. & Howie, B., 2010. Genotype imputation for genome-wide association studies. *Nature reviews. Genetics*, 11(7), pp.499–511. Available at: <http://dx.doi.org/10.1038/nrg2796> [Accessed January 29, 2013].
- Marfella, C.G.A. & Imbalzano, A.N., 2007. The Chd family of chromatin remodelers. *Mutation research*, 618(1-2), pp.30–40. Available at: <http://dx.doi.org/10.1016/j.mrfmmm.2006.07.012> [Accessed August 6, 2013].
- Mastrangelo, M. & Leuzzi, V., 2012. Genes of early-onset epileptic encephalopathies: from genotype to phenotype. *Pediatric neurology*, 46(1), pp.24–31. Available at: <http://dx.doi.org/10.1016/j.pediatrneurol.2011.11.003> [Accessed February 27, 2013].
- Mercier, S. et al., 2011. New findings for phenotype-genotype correlations in a large European series of holoprosencephaly cases. *Journal of medical genetics*, 48(11), pp.752–60. Available at:

<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3386902&tool=pmcentrez&rendertype=abstract> [Accessed January 16, 2013].

- Merikangas, A.K., Corvin, A.P. & Gallagher, L., 2009. Copy-number variants in neurodevelopmental disorders: promises and challenges. *Trends in genetics : TIG*, 25(12), pp.536–44. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/19910074> [Accessed November 12, 2012].
- Mines, M.A. et al., 2010. GSK3 influences social preference and anxiety-related behaviors during social interaction in a mouse model of fragile X syndrome and autism. *PLoS one*, 5(3), p.e9706. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2838793&tool=pmcentrez&rendertype=abstract> [Accessed March 30, 2013].
- Muhle, R., Trentacoste, S. V. & Rapin, I., 2004. The genetics of autism. *Pediatrics*, 113(5), pp.e472–86. Available at: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=15121991 [Accessed November 5, 2012].
- Mundel, P. et al., 1997. Synaptopodin: an actin-associated protein in telencephalic dendrites and renal podocytes. *The Journal of cell biology*, 139(1), pp.193–204. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2139823&tool=pmcentrez&rendertype=abstract> [Accessed March 25, 2013].
- Myers, A.J. et al., 2007. A survey of genetic human cortical gene expression. *Nat Genet*, 39(12), pp.1494–1499. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/17982457> [Accessed November 1, 2012].
- Naisbitt, S. et al., 1999. Shank, a novel family of postsynaptic density proteins that binds to the NMDA receptor/PSD-95/GKAP complex and cortactin. *Neuron*, 23(3), pp.569–82. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/10433268> [Accessed February 24, 2013].
- Neale, B.M. et al., 2012. Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature*, 485(7397), pp.242–5. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/22495311> [Accessed February 27, 2013].
- Neuvonen, M. & Ahola, T., 2009. Differential activities of cellular and viral macro domain proteins in binding of ADP-ribose metabolites. *Journal of molecular biology*, 385(1), pp.212–25. Available at: <http://dx.doi.org/10.1016/j.jmb.2008.10.045> [Accessed February 27, 2013].
- Newschaffer, C.J. et al., 2007. The epidemiology of autism spectrum disorders. *Annu Rev Public Health*, 28, pp.235–258. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/17367287> [Accessed November 6, 2012].
- North, B.J. & Verdin, E., 2004. Sirtuins: Sir2-related NAD-dependent protein deacetylases. *Genome biology*, 5(5), p.224. Available at:

<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=416462&tool=pmcentrez&rendertype=abstract> [Accessed March 18, 2013].

- O'Donovan, M.C., Kirov, G. & Owen, M.J., 2008. Phenotypic variations on the theme of CNVs. *Nature genetics*, 40(12), pp.1392–3. Available at: <http://dx.doi.org/10.1038/ng1208-1392> [Accessed March 23, 2013].
- O'Dushlaine, C. et al., 2009. The SNP ratio test: pathway analysis of genome-wide association datasets. *Bioinformatics (Oxford, England)*, 25(20), pp.2762–3. Available at: <http://bioinformatics.oxfordjournals.org/content/25/20/2762.short> [Accessed March 21, 2013].
- O'Roak, B.J. et al., 2011. Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nature genetics*, 43(6), pp.585–9. Available at: <http://dx.doi.org/10.1038/ng.835> [Accessed February 28, 2013].
- O'Roak, B.J., Vives, L., Fu, W., et al., 2012. Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. *Science (New York, N.Y.)*, 338(6114), pp.1619–22. Available at: <http://www.sciencemag.org/content/338/6114/1619.abstract> [Accessed March 5, 2013].
- O'Roak, B.J., Vives, L., Girirajan, S., et al., 2012. Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature*, 485(7397), pp.246–50. Available at: <http://dx.doi.org/10.1038/nature10989> [Accessed February 27, 2013].
- O'Sullivan, M.L. et al., 2012. FLRT proteins are endogenous latrophilin ligands and regulate excitatory synapse development. *Neuron*, 73(5), pp.903–10. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/22405201> [Accessed October 29, 2012].
- Oda, T. et al., 2004. ABT1-associated protein (ABTAP), a novel nuclear protein conserved from yeast to mammals, represses transcriptional activation by ABT1. *Journal of cellular biochemistry*, 93(4), pp.788–806. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/15660422> [Accessed March 5, 2013].
- Okada, Y. et al., 1995. The neuron-specific kinesin superfamily protein KIF1A is a unique monomeric motor for anterograde axonal transport of synaptic vesicle precursors. *Cell*, 81(5), pp.769–80. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/7539720> [Accessed March 18, 2013].
- Organization, W.H., 1992. *International classification of diseases: 10th edition (ICD-10)*, Geneva: World Health Organization.
- Ozonoff, S. et al., 2011. Recurrence risk for autism spectrum disorders: a Baby Siblings Research Consortium study. *Pediatrics*, 128(3), pp.e488–95. Available at: <http://pediatrics.aappublications.org/content/128/3/e488.short> [Accessed March 25, 2013].

- Pardo, C.A. & Eberhart, C.G., 2007. The neurobiology of autism. *Brain Pathol*, 17(4), pp.434–447. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/17919129> [Accessed November 14, 2012].
- Pickles, a et al., 2000. Variable expression of the autism broader phenotype: findings from extended pedigrees. *J Child Psychol Psychiatry*, 41(4), pp.491–502. Available at: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=10836679.
- Pinto, D. et al., 2010. Functional impact of global rare copy number variation in autism spectrum disorders. *Nature*, 466(7304), pp.368–372. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3021798&tool=pmcentrez&rendertype=abstract> [Accessed October 26, 2012].
- Posavec, M., Timinszky, G. & Buschbeck, M., 2013. Macro domains as metabolite sensors on chromatin. *Cellular and molecular life sciences : CMLS*. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/23455074> [Accessed March 6, 2013].
- Purcell, S. et al., 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, 81(3), pp.559–575. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1950838&tool=pmcentrez&rendertype=abstract> [Accessed October 25, 2012].
- Reddy, K.S. et al., 2013. A systematic analysis of small supernumerary marker chromosomes using array CGH exposes unexpected complexity. *Genetics in medicine : official journal of the American College of Medical Genetics*, 15(1), pp.3–13. Available at: <http://dx.doi.org/10.1038/gim.2012.78> [Accessed February 20, 2013].
- Risch, N. et al., 1999. A genomic screen of autism: evidence for a multilocus etiology. *American journal of human genetics*, 65(2), pp.493–507. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1377948&tool=pmcentrez&rendertype=abstract> [Accessed January 17, 2013].
- Risch, N. & Merikangas, K., 1996. The future of genetic studies of complex human diseases. *Science (New York, N.Y.)*, 273(5281), pp.1516–7. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/8801636> [Accessed February 6, 2013].
- Robertson, N.G. et al., 2000. A novel conserved cochlear gene, OTOR: identification, expression analysis, and chromosomal mapping. *Genomics*, 66(3), pp.242–8. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/10873378> [Accessed March 5, 2013].
- Robinson, E.B. et al., 2011. A Multivariate Twin Study of Autistic Traits in 12-Year-Olds: Testing the Fractionable Autism Triad Hypothesis. *Behav Genet*, 42(2), pp.245–55. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/21927971> [Accessed November 8, 2012].

- Robinson, M. et al., 2004. FLRT3 is expressed in sensory neurons after peripheral nerve injury and regulates neurite outgrowth. *Molecular and cellular neurosciences*, 27(2), pp.202–14. Available at: <http://dx.doi.org/10.1016/j.mcn.2004.06.008> [Accessed March 5, 2013].
- Ronald, A. & Hoekstra, R.A., 2011. Autism spectrum disorders and autistic traits: a decade of new twin studies. *American journal of medical genetics. Part B, Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics*, 156B(3), pp.255–74. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/21438136> [Accessed November 8, 2012].
- Ronan, J.L., Wu, W. & Crabtree, G.R., 2013. From neural development to cognition: unexpected roles for chromatin. *Nature reviews. Genetics*, 14(5), pp.347–59. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/23568486> [Accessed August 9, 2013].
- Rosenthal, F. et al., 2013. Macrodomein-containing proteins are new mono-ADP-ribosylhydrolases. *Nature structural & molecular biology*, advance on. Available at: <http://dx.doi.org/10.1038/nsmb.2521> [Accessed March 15, 2013].
- Sakamoto, I. et al., 2000. A novel beta-catenin-binding protein inhibits beta-catenin-dependent Tcf activation and axis formation. *The Journal of biological chemistry*, 275(42), pp.32871–8. Available at: <http://www.jbc.org/content/275/42/32871.short> [Accessed August 10, 2013].
- Sambrook, J., Fritsch, E.F. & Maniatis, T., 2001. *Molecular Cloning: A Laboratory Manual* J. Argentine, ed., Cold Spring Harbor Laboratory Press. Available at: www.molecularcloning.com.
- Sanders, S.J. et al., 2012. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature*, 485(7397), pp.237–41. Available at: <http://dx.doi.org/10.1038/nature10945> [Accessed February 27, 2013].
- Sanders, S.J. et al., 2011. Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron*, 70(5), pp.863–885. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/21658581> [Accessed November 2, 2012].
- Santoro, M.R., Bray, S.M. & Warren, S.T., 2012. Molecular mechanisms of fragile X syndrome: a twenty-year perspective. *Annual review of pathology*, 7, pp.219–45. Available at: <http://www.annualreviews.org/doi/full/10.1146/annurev-pathol-011811-132457> [Accessed February 20, 2013].
- Sebat, J. et al., 2007. Strong association of de novo copy number mutations with autism. *Science (New York, N.Y.)*, 316(5823), pp.445–9. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2993504&tool=pmcentrez&rendertype=abstract> [Accessed March 1, 2013].

- Servin, B. & Stephens, M., 2007. Imputation-based analysis of association studies: candidate regions and quantitative traits. D. B. Allison, ed. *PLoS genetics*, 3(7), p.e114. Available at: <http://dx.plos.org/10.1371/journal.pgen.0030114> [Accessed November 5, 2012].
- Sham, P. et al., 2002. DNA Pooling: a tool for large-scale association studies. *Nature reviews. Genetics*, 3(11), pp.862–71. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/12415316> [Accessed March 11, 2013].
- Shirao, T., 1995. The roles of microfilament-associated proteins, drebrins, in brain morphogenesis: a review. *Journal of biochemistry*, 117(2), pp.231–6. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/7608104> [Accessed March 20, 2013].
- Skaar, D.A. et al., 2005. Analysis of the RELN gene as a genetic risk factor for autism. *Molecular psychiatry*, 10(6), pp.563–71. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/15558079> [Accessed January 17, 2013].
- Sklar, P. et al., 2011. Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nature Genetics*, 43(10), pp.977–83. Available at: <http://discovery.ucl.ac.uk/1323238/>.
- Spielman, R.S., McGinnis, R.E. & Ewens, W.J., 1993. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *American journal of human genetics*, 52(3), pp.506–16. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1682161&tool=pmcentrez&rendertype=abstract> [Accessed February 10, 2013].
- State, M.W., 2010. The genetics of child psychiatric disorders: focus on autism and Tourette syndrome. *Neuron*, 68(2), pp.254–69. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3292208&tool=pmcentrez&rendertype=abstract> [Accessed March 4, 2013].
- Sugiana, C. et al., 2008. Mutation of C20orf7 disrupts complex I assembly and causes lethal neonatal mitochondrial disease. *American journal of human genetics*, 83(4), pp.468–78. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2561934&tool=pmcentrez&rendertype=abstract> [Accessed March 5, 2013].
- Szatmari, P., 1999. Heterogeneity and the genetics of autism. *Journal of psychiatry & neuroscience : JPN*, 24(2), pp.159–65. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1188998&tool=pmcentrez&rendertype=abstract> [Accessed February 12, 2013].
- Szatmari, P. et al., 2007. Mapping autism risk loci using genetic linkage and chromosomal rearrangements. *Nat Genet*, 39(3), pp.319–328. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/17322880> [Accessed November 8, 2012].
- Taneja, P. et al., 2009. Pathophysiology of locus ceruleus neurons in a mouse model of Rett syndrome. *The Journal of neuroscience : the official journal of the Society for*

Neuroscience, 29(39), pp.12187–95. Available at:
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2846656&tool=pmcentrez&rendertype=abstract> [Accessed February 20, 2013].

Timinszky, G. et al., 2009. A macrodomain-containing histone rearranges chromatin upon sensing PARP1 activation. *Nat Struct Mol Biol*, 16(9), pp.923–929. Available at:
<http://www.ncbi.nlm.nih.gov/pubmed/19680243> [Accessed December 16, 2012].

Tropea, D. et al., 2009. Partial reversal of Rett Syndrome-like symptoms in MeCP2 mutant mice. *Proceedings of the National Academy of Sciences of the United States of America*, 106(6), pp.2029–34. Available at:
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2644158&tool=pmcentrez&rendertype=abstract> [Accessed February 20, 2013].

Tsuji, L. et al., 2004. FLRT3, a cell surface molecule containing LRR repeats and a FNIII domain, promotes neurite outgrowth. *Biochemical and Biophysical Research Communications*, 313(4), pp.1086–1091. Available at:
<http://linkinghub.elsevier.com/retrieve/pii/S0006291X03026445> [Accessed December 13, 2012].

Tuchman, R. & Rapin, I., 2002. Epilepsy in autism. *The Lancet Neurology*, 1(6), pp.352–358. Available at: [http://dx.doi.org/10.1016/S1474-4422\(02\)00160-6](http://dx.doi.org/10.1016/S1474-4422(02)00160-6) [Accessed February 23, 2013].

Visscher, P.M., Hill, W.G. & Wray, N.R., 2008. Heritability in the genomics era--concepts and misconceptions. *Nature reviews. Genetics*, 9(4), pp.255–66. Available at:
<http://www.ncbi.nlm.nih.gov/pubmed/18319743> [Accessed January 30, 2013].

Voineagu, I. et al., 2011. Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature*, 474(7351), pp.380–384. Available at:
<http://www.ncbi.nlm.nih.gov/pubmed/21614001> [Accessed October 27, 2012].

Wang, K. et al., 2009. Common genetic variants on 5p14.1 associate with autism spectrum disorders. *Nature*, 459(7246), pp.528–533. Available at:
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2943511&tool=pmcentrez&rendertype=abstract> [Accessed November 2, 2012].

Weiss, L. a et al., 2009. A genome-wide linkage and association scan reveals novel loci for autism. *Nature*, 461(7265), pp.802–808. Available at:
<http://www.ncbi.nlm.nih.gov/pubmed/19812673> [Accessed November 2, 2012].

Weiss, L.A. et al., 2008. Association between microdeletion and microduplication at 16p11.2 and autism. *The New England journal of medicine*, 358(7), pp.667–75. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/18184952> [Accessed March 7, 2013].

Wilson, H.L., 2003. Molecular characterisation of the 22q13 deletion syndrome supports the role of haploinsufficiency of SHANK3/PROSAP2 in the major neurological

symptoms. *Journal of Medical Genetics*, 40(8), pp.575–584. Available at: <http://jmg.bmj.com/content/40/8/575.full> [Accessed February 24, 2013].

Wray, N.R. & Visscher, P.M., 2008. Estimating Trait Heritability. *Nature Education*, p.1. Available at: <http://www.nature.com/scitable/topicpage/estimating-trait-heritability-46889> [Accessed March 27, 2013].

Xu, B. et al., 2009. Elucidating the genetic architecture of familial schizophrenia using rare copy number variant and linkage scans. *Proc Natl Acad Sci U S A*, 106(39), pp.16746–16751. Available at: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=19805367.

Yang, M.S. & Gill, M., 2007. A review of gene linkage, association and expression studies in autism and an assessment of convergent evidence. *International journal of developmental neuroscience : the official journal of the International Society for Developmental Neuroscience*, 25(2), pp.69–85. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/17236739> [Accessed January 17, 2013].

Yuasa-Kawada, J. et al., 2009. Deubiquitinating enzyme USP33/VDU1 is required for Slit signaling in inhibiting breast cancer cell migration. *Proceedings of the National Academy of Sciences of the United States of America*, 106(34), pp.14530–5. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2732860&tool=pmcentrez&rendertype=abstract> [Accessed March 28, 2013].

Zhang, R. & Lin, Y., 2009. DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes. *Nucleic acids research*, 37(Database issue), pp.D455–8. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2686491&tool=pmcentrez&rendertype=abstract> [Accessed March 3, 2013].

Zhong, H. & Prentice, R.L., 2010. Correcting “winner’s curse” in odds ratios from genomewide association findings for major complex human diseases. *Genetic epidemiology*, 34(1), pp.78–91. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2796696&tool=pmcentrez&rendertype=abstract> [Accessed March 28, 2013].

Zikopoulos, B. & Barbas, H., 2010. Changes in prefrontal axons may disrupt the network in autism. *J Neurosci*, 30(44), pp.14595–14609. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3073590&tool=pmcentrez&rendertype=abstract> [Accessed November 21, 2012].

Van der Zwaag, B. et al., 2009. Gene-network analysis identifies susceptibility genes related to glycobiology in autism. *PloS one*, 4(5), p.e5324. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2683930&tool=pmcentrez&rendertype=abstract> [Accessed March 6, 2013].