



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

Copyright statement

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

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

Liability statement

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

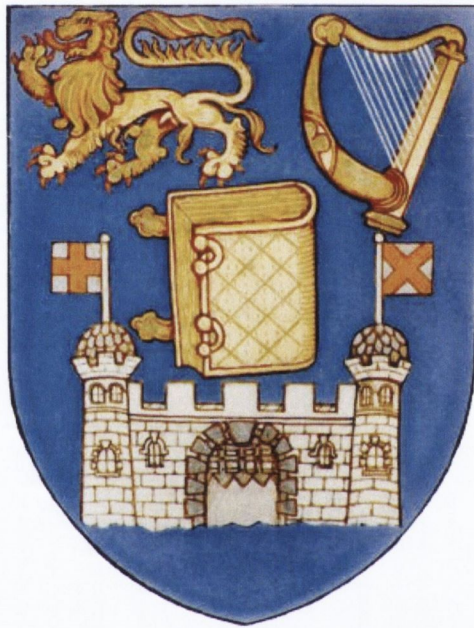
Access Agreement

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

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

**Investigation of Rare Sequence Variants in Putative
Susceptibility Genes for Schizophrenia and Autism
Spectrum Disorders**

Sarah Furlong



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

**Department of Psychiatry
University of Dublin
Trinity College – March 2013**

Declaration

University of Dublin

Trinity College – March 2013

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, except where stated otherwise.

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.

Signed: *Sarah Furlong*



Thesis 9917

Summary

Schizophrenia (SZ) and Autism Spectrum Disorders (ASD) are complex neuropsychiatric disorders that have a very large health, societal and economic burden in Ireland. They affect approximately 1% and 0.8% of the population respectively, with heritability estimated to be as high as 80%. The underlying genetic architecture and pathogenesis of each disorder remains to be elucidated. Substantial phenotypic and genetic overlap exists between SZ and ASD, indicating that they may share a common pathogenic mechanism. The aim of this study is to identify rare variants that contribute to the pathophysiology of SZ and ASD.

A systematic approach was undertaken to select candidate genes for inclusion in the study based on five separate biological categories; 1) *NRXN1*, interactors and related proteins (n = 46), 2) *DISC1* and interactors (n = 23), 3) Postsynaptic Glutamate Receptor Complexes (n = 58), 4) Neural cell Adhesion Molecules (n = 61), and 5) Functional and Positional Candidates (n = 27). Next-generation sequencing of the exonic regions of 215 putative susceptibility genes was carried out in an Irish sample of 147 cases of ASD, 273 cases of SZ and 287 controls. A multiplex target enrichment method combined DNA samples using indexes followed by enrichment of exonic regions using Agilent SureSelect system and paired-end sequencing on an Illumina Genome Analyzer II.

As the sample prep protocol in place at the beginning of this project was laborious and expensive, I successfully introduced a number of protocol modifications to improve it. For example, the single-tube processing format was converted to a 96-well format, dramatically reducing the amount of time and labour required. In addition, modifications following analysis of a pilot set of samples lead to a significant reduction in the level of duplicate sequence reads and allowed a greater number of samples to be processed, significantly improving study power and reducing project costs.

Sequencing the exons of the 215 genes in the combined dataset of SZ cases, ASD cases and controls identified 3,708 variants that passed quality control filters. As the majority of samples were part of previous GWAS, it was possible to compare GWAS SNV calls with SNVs calls from this study. The concordance

rate was greater than 99%, indicating very high specificity for variant calls in the sequence data. A large proportion of these variant were rare variants (2,861) of which 31 were novel LoF variants, which are predicted to severely disrupt the protein-coding sequence. A significant excess of SZ and ASD cases in this dataset carry these rare LoF variants compared to the controls ($p=0.02$), with the excess greater for ASD cases ($p=0.005$) than for SZ ($p=0.07$). There was an even greater excess of these singleton LoF variants in the combined case sample ($p=0.0007$) and the excess was present in both ASD ($p=0.001$) and SZ ($p=0.002$). I also observed a significant excess ($p=0.05$) of variants in the *NRXN1*-related and interacting proteins category (category 1).

A nonsense variant in *GRIN2B* in an ASD case was found to be *de novo*, adding to the growing evidence that this is an important risk gene for the disorder. Parent-of-origin analysis revealed that this mutation arose on the maternal strand. A second *de novo* mutation was found in *DISC1*. Parent-of-origin analysis indicated that this variant was on the paternal chromosome but a closer study revealed evidence of mosaicism in the father.

A total of 1,307 rare missense variants were identified in this study. I performed an assessment of annotation tools for prioritising such variants and little consensus was found to exist between them. I chose SIFT and PP2 to apply to my subsequent association analysis. Genotype based association analysis did not reveal a significant excess of these variants in cases compared to controls.

Understanding the impact of LoF variants requires further functional investigation at a cellular and molecular level. *FYN* is a receptor tyrosine kinase that regulates synaptic plasticity primarily via its interaction with NMDA receptors and has previously been associated with SZ. I carried out investigation of a nonsense LoF variant in the *FYN* gene by transfecting mouse primary cortical neurons with wildtype and mutant version of the gene. Preliminary data indicates that the LoF does not affect expression but may impact on neurite length and synaptic functioning. Altogether, these data support synapse formation and maintenance as key molecular mechanisms for SZ and ASD.

Acknowledgements

I would like to thank my supervisor Dr. Derek Morris for his continued patience, dedication and support throughout this project, and for all he has taught me. I'm truly grateful to have had him as my supervisor. I would also like to extend my gratitude to the organisers of the Molecular Medicine PhD programme for accepting me on to the programme and to the Health Research Board for funding this project.

This project would not have been possible without the contributions of Elaine and Paul, who were always willing to offer me advice. I'm particularly grateful to Elaine for the huge level of support she gave me in the early part of my project. I'd like to thank Daniela, Ines and Ciara for their time and most of all, patience in the last few months. And to all others who have kindly helped me at some point during this project including Ric, Eleisa and Carlos.

My time in the group has been made very enjoyable, particularly by all those friendly people, past and present in room 1.10! Ciara, I'd have been lost without your daily dose of ridiculousness! Graham, you have been there from the start and without you, who knows where I would have ended up! Alison, your company especially toward the end was much appreciated; it would have been a lonely task without you! And thanks to all others for proofreading, admin or simply their company during this time: Emma, Siobhan, Carol, Stefania, Elif and Geraldine.

I'd like to thank all my friends and my family; William, Bridget, Joanne, James and Eamon, and Jackie, for their kindness and for generally putting up with me, and not asking me too often, when am I going to finally leave college and get a real job. I'd especially like to thank my boyfriend Bub for encouraging me, standing by me and for being the greatest incentive I could have had to finish this thesis.

Finally, I'd like to thank my Dad, William for his support and encouragement; I dedicate this thesis to you.

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

AA	Amino Acid
AAS	Amino Acid Substitution
ADHD	Attention Deficit Hyperactivity Disorder
ADI-R	Autism Diagnostic Interview-Revised
ADOS-G	Autism Diagnostic Observation Schedule-Generic
AGP	Autism Genome Project
AMP	Ampicillin
AMPA	2-amino-3-(3-hydroxy-5-methyl-isoxazol-4-yl) propanoic acid
APOE	Apolipoprotein E
ASD	Autism Spectrum Disorders
AS-PCR	Allele-Specific Polymerase Chain Reaction
ATAV	Association Tests for Annotated Variants
BDNF	Brain-Derived Neurotrophic Factor
BioGRID	The Biological General Repository for Interaction Datasets
BLOSUM	BLOcks of Amino Acid SUBstitution Matrix
BSA	Bovine Serum Albumin
BWA	Burrows-Wheeler Aligner
CAM	Cell Adhesion Molecule
CDCV	Common Disease, Common Variant
cDNA	Complementary DNA
ChIP-Seq	Chromatin Immunoprecipitation Sequencing
Chr	Chromosome
CI	Confidence Interval
CMC	Combined Multivariate and Collapsing
CMV	Cytomegalovirus
CNV	Copy Number Variant
COSMIC	Catalogue of Somatic Mutations in Cancer
CRT	Cyclical Reversible Termination
Ct	Cycle Threshold
CVRV	Common disease, Rare Variant
DAPI	4',6-diamidino-2-phenylindole
ddNTPs	Dideoxy Nucleotide Triphosphate
DGV	Database of Genomic Variants

dH ₂ O	distilled Hydrogen Dioxide
DISC1	Disrupted-in-Schizophrenia 1
DMEM	Dulbecco's Modified Eagle Medium
DNA	Deoxyribonucleic Acid
dNTP	Deoxyribonucleotide Triphosphate
DSM-IV	Diagnostic Statistical Manual IV
DST	Dystonin
DZ	Dizygotic
EB	Elution Buffer
EEG	Electroencephalography
EPS	Extra-Pyramidal Side-effects
EtBr	Ethidium Bromide
FAM	6-carboxyfluorescein
FBS	Foetal Bovine Serum
FIS	Functional Impact Score
FYN	Tyrosine-protein kinase Fyn
GAII	Illumina Genome Analyzer II
GATK	Genome Analysis Toolkit
GFP	Green Fluorescent Protein
GO	Gene Ontology
GRANVIL	Gene or Region-based ANalysis of Variants of Intermediate and Low frequency
GRIN2B	Glutamate Receptor, Ionotropic, N-methyl D-aspartate 2B
GRIP1	Glutamate Receptor Interacting Protein 1
GWAS	Genome-Wide Association Study
H ₂ O	Hydrogen Dioxide
HMM	Hidden Markov Model
HPRD	Human Protein Reference Database
HTG	Hypertriglyceridemia
ICD-10	International Classification of Diseases-10
ID	Intellectual Disability
IDT	Integrated DNA Technologies
IGV	Integrative Genomics Viewer
INADL	InaD-like
Indel	Insertion/Deletion

IQ	Intelligence Quotient
ISC	The International Schizophrenia Consortium
ITGAV	Integrin, Alpha V
KCTD13	Potassium Channel Tetramerisation Domain Containing 13
KEGG	Kyoto Encyclopedia of Genes and Genomes
LB	Lysogeny Broth
LD	Linkage Disequilibrium
LMW	Low Molecular Weight
LOD	Log of Odds
LoF	Loss-of-Function
MACROD2	MACRO domain containing 2
MAF	Minor Allele Frequency
MAGUK	Membrane Associated Guanylate Kinase
MAP2	Microtubule-Associated Protein 2
MASS-FIS	Massessor Functional Impact Score
MBP	Myelin Basic Protein
MECP2	Methyl CpG Binding Protein 2
mGluR5	Metabotropic Glutamate Receptor 5
MHC	Major Histocompatibility Complex
MR	Mental Retardation
MRI	Magnetic Resonance Imaging
mRNA	Messenger Ribonucleic Acid
MUT	Mutant
MZ	Monozygotic
NaOAc	Sodium Acetate
NCBI	National Center for Biotechnology Information
NEB	New England Biolabs
NGS	Next Generation Sequencing
NLGN	Neurologin
NLGN3	Neurologin3
NLGN4X	Neurologin 4X
NMDA	N-methyl-D-aspartate
NMDAR	N-Methyl-D-aspartate receptor
NRXN	Neurexin
NT	Non-Transfected

OR	Odds Ratio
ORF	Open Reading Frame
PBS	Phosphate-Buffered Saline
PBST	Phosphate-Buffered Saline-Triton
PCR	Polymerase Chain Reaction
PDD	Pervasive Developmental Disorder
PDD-NOS	Pervasive Developmental Disorder – Not Otherwise Specified
PP2	Polyphen 2
PSD	Post-Synaptic Density
PSD95	Post-Synaptic Density Protein 95
PSP	Post-Synaptic Proteome
QC	Quality Control
RACE	Rapid Amplification of cDNA Ends
RB	Reaction Buffer
RNA	Ribonucleic Acid
SAP	Shrimp Alkaline Phosphatase
SCID-P	Structured Clinical Interview for DSM-IV-Patient Edition
SD	Standard Deviation
SDS	Sequence Detection System
SEMA5A	Semaphorin 5A
SIFT	Scale-Invariant Feature Transform
SKAT	Kernel Association Test
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variant
SOC	Super Optimal broth with Catabolite repression
SZ	Schizophrenia
TAE	Tris Base, Acetic Acid and EDTA
TMS	Transcranial Magnetic Stimulation
UCSC	University College, Santa Cruz
UTR	Untranslated Region
VABS	Vineland Adaptive Behaviour Scales
VEP	Variant Effect Predictor
WT	Wildtype
XTERM	XTerminator
ZNF804A	Zinc Finger Protein 804A

Table of Contents

Declarationii

Summaryiii

Acknowledgements v

List of Abbreviations..... vii

Table of Contentsxi

Tables of Figures xvii

Table of Tables.....xx

1 Introduction 1

1.1 Overview of Schizophrenia..... 1

1.2 Overview of Autism 4

1.3 Phenotypic Overlap Schizophrenia and Autism..... 9

1.4 Genetic Variation.....10

1.4.1 Single Nucleotide Variants10

1.4.2 Insertion/Deletions11

1.4.3 Structural Variants11

1.4.4 Linkage Disequilibrium12

1.5 Genetics of Common Complex Disorders12

1.6 Gene Discovery Approaches.....14

1.6.1 Linkage Analysis.....14

1.6.2 Association Analysis.....15

1.6.3 Genetic Evidence in Schizophrenia16

1.6.4 Genetic Evidence in Autism.....18

1.7 Genetic Overlap in Schizophrenia and Autism19

1.8 Next Generation Sequencing.....20

1.8.1 Targeted Resequencing20

1.8.2 Indexing21

1.9 Project Aims.....	22
2 Materials and Method.....	24
2.1 Gene Selection	24
2.2 Target Enrichment Array Design	24
2.3 Selection of Samples	27
2.3.1 Autism.....	27
2.3.2 Schizophrenia.....	27
2.3.3 Control Samples	27
2.4 Illumina Sample Preparation Optimisation	28
2.4.1 Overview	28
2.4.2 DNA Fragmentation by Sonication	30
2.4.3 Ligation Test	30
2.4.4 DNA Fragmentation using dsDNA Fragmentase.....	31
2.4.5 Purification Method Comparison	32
2.4.5.1 Qiagen Column Purification	33
2.4.5.2 Ampure XP Purification	33
2.4.5.3 Chemagic Magnetic Bead Purification.....	34
2.4.6 Indexed Adapter Annealing	34
2.5 Sample Preparation Protocol.....	35
2.5.1 Fragmentation of DNA with dsDNA Fragmentase	35
2.5.2 End Repair	36
2.5.3 Add 'A' base	36
2.5.4 Quick Ligation.....	37
2.5.5 Size Selection.....	37
2.5.6 Qiagen Gel Purification.....	38
2.5.7 Sample Preparation PCR Enrichment.....	39
2.5.8 DNA Library Quality Assessment by qPCR	39
2.5.9 SureSelect Target Enrichment.....	40

2.5.10 Target Enrichment Library Quality Assessment Using Bioanalyser	41
2.6 Next Generation Sequencing and Data Analysis	41
2.6.1 Variant Detection	41
2.6.2 Loss-of-Function Variant Validation	42
2.6.3 Determination of Parent-of-Origin of <i>De Novo</i> Variants	45
2.6.4 <i>In Silico</i> Analysis of Variants	48
2.6.4.1 <i>In Silico</i> Analysis of LOF Variants	48
2.6.4.2 <i>In Silico</i> Analysis of Missense Variants	48
2.7 <i>In vitro</i> Investigation of Loss-of-Function Variants	48
2.7.1 Transformation of Plasmids into Component <i>E.coli</i> Cells	49
2.7.1.1 Post Transformation	50
2.7.1.2 Colony Screen	50
2.7.1.3 Growth of Positive Colonies	51
2.7.1.4 Plasmid Purification	52
2.7.1.5 Check Plasmid Quality	53
2.7.1.6 Concentration of Plasmid DNA by Ethanol Precipitation	54
2.7.1.7 Sequence Plasmid Inserts	54
2.7.2 Preparation of Mouse Cell Cultures	54
2.7.3 Transfection	56
2.7.4 Immunohistochemistry	57
2.7.4.1 Primary Antibody Staining	58
2.7.4.2 Secondary Antibody Staining	58
2.7.5 Confocal Microscopy	59
2.7.6 Recipes for Broths and Media	59
2.7.6.1 Super Optimal broth with Catabolite Repression (SOC) Media	59
2.7.6.2 Lysogeny Broth (LB) Agar	59
2.7.6.3 Feeding Media	60
2.7.6.4 Ara-C Feeding Media	60

2.7.6.5 DMEM/ 10% FBS Media 60

2.7.6.6 PBS-Triton (PBS-T) 60

3 Selection of Candidate Gene and Optimisation of Multiplex Target Enrichment Method Including Application to Test Pool of Samples 61

3.1 Aims 61

3.2 Gene Selection 62

3.2.1 Neurexin (NRXN1, Interactors and Related Proteins) 62

3.2.2 Disrupted in Schizophrenia 1 (DISC1) and Interactors 64

3.2.3 Post-Synaptic Glutamate Receptor Complexes..... 64

3.2.4 Neural Cell Adhesion Molecules 65

3.2.5 Functional and Positional Candidates..... 65

3.3 Target Enrichment Array Design 66

3.4 Sample Preparation Optimisation 66

3.4.1 Ligation Test 66

3.4.2 Fragmentase vs. Sonication 68

3.4.3 Purification Method Optimisation 71

3.5 Analysis of Pilot Samples 73

3.6 Conclusion 79

4 Investigation of Identified Sequence Variants including Validation, *In Silico* analysis and Association Analysis 81

4.1 Aim..... 81

4.2 Full Dataset Analysis 81

4.3 Loss-of-Function Variants..... 84

4.3.1 LoF variant Validation 86

4.3.2 Determination of Parent-of-Origin of *De Novo* Variants 93

4.3.2.1 *GRIN2B* *De Novo* LoF Variant Parent-of-Origin 94

4.3.2.2 *DISC1* *De Novo* LoF Variant Parent-of-Origin 96

4.3.3 *In silico* Analysis of LoF variants..... 99

4.4 LoF variant Association Analysis..... 101

	4.5 Functional Annotation of Missense Variants	104
	4.5.1 Annovar	106
	4.5.2 Mutation Assessor (Massessor)	107
	4.5.3 MU2A	107
	4.5.4 PolyPhen-2	108
	4.5.5 Seattle Seq	109
	4.5.6 SIFT	110
	4.5.7 Use of Annotation Tools	110
	4.5.8 Investigation of Annotation Scores using SPSS	114
	4.5.8.1 Correlation	116
	4.6 Missense Variant Association Analysis	121
	4.7 Conclusion	122
5	Functional Investigation of a Loss-of-Function Variant in <i>FYN</i>	124
	5.1 Aim	124
	5.2 LoF Variant Selection	124
	5.2.1 Determination of Homology Between Human LoF Containing Transcripts with Mouse Transcripts	125
	5.3 Plasmid Design	128
	5.4 Determination of the Correct Plasmid Insert	129
	5.5 Impact on <i>FYN</i> Expression and Dendrite Length	129
	5.6 Impact on Synapse Function	138
	5.7 Conclusion	143
6	Discussion	146
	6.1 Review of the Field at the Beginning of the Project	146
	6.2 Review of this Study	148
	6.2.1 Challenges for Genes Selection	148
	6.2.2 Protocol Modifications	148
	6.2.3 Rare LoF Variant Analysis	149
	6.2.4 Analysis of Individual LoF Variants	151

6.2.5 Importance of *De Novo* Variants..... 153

6.2.6 Functional Genomics..... 156

6.2.7 Alternative Investigation of LoF Variants 159

6.2.8 Missense Variant Analysis..... 160

6.2.9 Summary of Findings..... 163

6.3 Where is the Field of Schizophrenia and Autism Genetics at in 2012?
..... 163

References 166

Appendix..... 187

Tables of Figures

Figure 1.1: Classification of Pervasive Developmental Disorder according to DSM-IV. 5

Figure 1.2: Feasibility of identifying genetic variants by risk allele frequency.14

Figure 1.3: Risk of developing SZ amongst relatives of SZ patients18

Figure 2.1 Agilent's SureSelect target enrichment system25

Figure 2.2: Expansion of small (<120bp) target exons for target enrichment.26

Figure 2.3: Illumina DNA sample sequencing preparation29

Figure 3.1: Agarose gel of dsDNA fragmentase test.68

Figure 3.2: A) Example of pile-up of sequence reads on target exons following alignment and B) Magnified image representing detection of a SNV at a target exon.74

Figure 3.3: Distribution of sequence reads amongst the 20 indexed control samples in pilot study.75

Figure 3.4: Flowchart of sequence information in two test sets of 20 indexed control samples based on the Pre-modifications and Post-modifications sample prep protocol.76

Figure 3.5: Percentage target coverage per sample in pool of 20 indexed controls in pilot study.....77

Figure 3.6: Distribution of sequence reads amongst the 20 indexed control samples post protocol modifications.78

Figure 3.7: Percentage target coverage per sample in pool of 20 indexed controls after protocol amendment.....79

Figure 4.1: Box plots of the sequence coverage for each of the three phenotype groups in the study.82

Figure 4.2: Flowchart of breakdown of variants in full dataset.....83

Figure 4.3: Illustration of the affect of LoF variants85

Figure 4.4: Electropherogram image of a nonsense mutation in *FYN*.88

Figure 4.5: Electropherogram image of a *de novo* nonsense mutation in *GRIN2B*.89

Figure 4.6: Electropherogram image of a *de novo* 1bp deletion in *DISC1*.90

Figure 4.7: Electropherogram image of a 1bp insertion in *ITGAV*.91

Figure 4.8: Electropherogram image of a donor splice site mutation in *GRIP1*.91

Figure 4.9: Electropherogram image of a stop loss mutation in *MBP*. 92

Figure 4.10: Determining parent-of-origin of *de novo* variants by allele-specific PCR..... 94

Figure 4.11: Electropherogram images of AS-PCR of *GRIN2B* *de novo* carrier and parents at the informative SNV Site. 96

Figure 4.12: Electropherogram images of AS-PCR of *DISC1* *de novo* carrier and parents 98

Figure 4.13: Location of the nonsense LoF variant in *DLG5* visualised in UCSC genome browser. 100

Figure 4.14: Plots of the distribution of missense variants across annotation scores..... 116

Figure 4.15: Correlation plots in annotation tools sets with R values ≥ 5 120

Figure 5.1: Expression clone with CMV promoter 128

Figure 5.2: Schematic of the plasmids used in transfection 129

Figure 5.3: Confocal Image of mouse cortical cells stained with A. DAPI, B. GFP C. *FYN* and D. *MAP2* and E. an overlap of all four stains. 131

Figure 5.4: Expression in *FYN* in transfected and non transfected neurons following attempted transfection with *GFP*, *FYN* WT and *FYN* MUT plasmids. 134

Figure 5.5: *FYN* Expression in neurons transfected with A. *GFP* B. *FYN* WT and C. *FYN* MUT 135

Figure 5.6: Dendrite length in transfected and non transfected neurons following attempted transfection with *GFP*, *FYN* WT and *FYN* MUT plasmids. 136

Figure 5.7: Dendrite Length in neurons transfected with A. *GFP* B. *FYN* WT and C. *FYN* MUT 137

Figure 5.8: Confocal image of mouse prefrontal cortex cells stained with A. DAPI, B. *GFP* C. Synapsin and D. an overlap of all four stains. 139

Figure 5.9: Synapsin intensity with more stringent (*GFP* intensity in neuron $\geq 3X$ background) transfection threshold. 140

Figure 5.10: Synapsin intensity with less stringent (*GFP* intensity in neuron $\geq 2X$ background) transfection threshold. 141

Figure 5.11: Localisation of synapsin measured by determining intensity of synapse signal in dendrites with more stringent (*GFP* intensity in neuron $\geq 3X$ background) transfection threshold..... 142

Figure 5.12: Localisation of synapsin measured by determining intensity of synapse signal in dendrites with less stringent (*GFP* intensity in neuron $\geq 2X$ background) transfection threshold..... 142

Figure 5.13: Synapsin localisation in neurons transfected with A. *FYN* WT and B. *FYN* MUT.....143

Figure 6.1: Diagram of *INADL* protein with functional domains and sites of variants identified in this study152

Figure 6.2: Diagram of *GRIP1* protein with functional domains and sites of variants.....153

Figure 6.3: Diagram of *GRIN2B* protein with functional domains and sites of variants.....154

Figure 6.4: Diagram of *FYN* protein with functional domains and sites of variants identified in this study157

Table of Tables

Table 1.1: Criteria for diagnosis of schizophrenia according to the Diagnostics Statistics Manual Classification IV (DSM-IV)3

Table 1.2: Criteria for diagnosis of schizophrenia according to the International Classification of Disease 10 (ICD-10)4

Table 1.3: Criteria for the diagnosis of autism spectrum disorders according to the Diagnostics Statistics Manual IV (DSM-IV).....7

Table 1.4: Criteria for the diagnosis of autism spectrum disorders according to the International Classification of Disease 10 (ICD-10)8

Table 2.1: Protein interaction databases used in gene selection 24

Table 2.2: Quick ligase reaction protocol 30

Table 2.3: T4 DNA ligase reaction components 31

Table 2.4: Test of dsDNA fragmentase incubation period 32

Table 2.5: Protocol for annealing of indexed adapters..... 35

Table: 2.6: Protocol for fragmentation of DNA with dsDNA fragmentase..... 36

Table 2.7: Protocol for end repair 36

Table 2.8: Protocol for A' base addition 37

Table 2.9: Protocol for quick ligation of adapters 37

Table 2.10: Sample preparation PCR..... 39

Table 2.11: Sample preparation PCR programme..... 39

Table 2.12: DNA library assessment by qPCR 40

Table 2.13: qPCR protocol to check sample quality..... 40

Table 2.14: PCR of loss-of-function variants for validation..... 42

Table 2.15: PCR product purification using ExoSAP-IT 43

Table 2.16: BigDye® Terminator v1.1 sequencing preparation 43

Table 2.17: BigDye® Terminator v1.1 sequencing reaction 44

Table 2.18: SAM X-Term purification of Big Dye sequencing reaction product ..
..... 44

Table 2.19: Allele-specific primers 47

Table 2.20: PCR colony screen 51

Table 3.1: Overnight ligation test 67

Table 3.2: Two hour ligation test 67

Table 3.3: Fragmentase vs. sonication sample preparation..... 70

Table 3.4: Purification method comparison..... 72

Table 3.5: Sample prep purification method comparison 72

Table 4.1: Breakdown of all LoF variants.....87

Table 4.2: Genotypes of *GRIN2B* SNVs used in AS-PCR.....95

Table 4.3: Alleles found at rs7312845 by AS-PCR from *GRIN2B de novo* SNV...
.....95

Table 4.4: Genotypes of *DISC1* SNVs used in AS-PCR.97

Table 4.5: Alleles found at *DISC1 de novo* SNV site by AS-PCR from rs821616
.....98

Table 4.6: Summary of LoF variants.....101

Table 4.7: Carrier-based analysis of LoF variants in the combined case sample
versus controls.....102

Table 4.8: Carrier-based analysis of LoF carriers in SZ cases versus controls
.....102

Table 4.9: Carrier-based analysis of LoF carriers in ASD cases versus controls
.....102

Table 4.10: Carrier-based analysis of singleton LoF variants in the combined
case sample versus controls.....103

Table 4.11: Carrier-based analysis of LoF singleton variants in SZ cases versus
controls.....103

Table 4.12: Carrier-based analysis of LoF singleton variants in ASD cases
versus controls.....103

Table 4.13: Carrier-based analysis of LoF variants in the individual gene
categories in the combined case sample versus controls.104

Table 4.14: Tools used for annotation of SNVs.....105

Table 4.15: Massessor Functional Impact Score categories107

Table 4.16: PolyPhen-2 probability score categories109

Table 4.17: Grantham score categories.....109

Table 4.18: SIFT score categories110

Table 4.19: Annotation tool scores and information used to generate each .111

Table 4.20: Breakdown SNVs into functional categories and the rate of missing
data in individual categories for each annotation score.....113

Table 4.21: Correlation between pairs of annotation scores calculated using
Spearman's rho in SPSS. Values are R.....118

Table 4.22: Heat map shading scheme for table 4.21118

Table 4.23: Genotype based association analysis of all rare variants using chi-
squared test122

Table 5.1: Information on LoF variant transcripts shortlisted for potential functional investigation 127

1 Introduction

Schizophrenia (SZ) and Autism Spectrum Disorders (ASD) are complex neuropsychiatric disorders that have a very large health, societal and economic burden in Ireland. The pathogenesis of these disorders is not well understood and as such current treatments are not sufficient to ease the suffering of a lot of patients and families. Recent genetic data suggests some shared aetiology between the disorders related to neurodevelopment and the function of the synapse, the connection point between nerve cells. Further new genetic discoveries can help elucidate the pathobiology of these disorders and could improve treatment strategies and diagnostic tools.

1.1 Overview of Schizophrenia

Schizophrenia is a severe and debilitating disorder and symptoms can be dividing into three broad categories; positive symptoms (hallucinations, delusions), negative symptoms (apathy, social withdrawal, disorganised speech) and cognitive deficits (attention, working memory and problem solving) (Friedrichs and Stoll 2006). Diagnosis is based on Diagnostic and Statistics Manual of Mental Disorder edition 4 (DSM-IV) or International Classification of Disease (ICD-10), as outlined in Table 1.1 and Table 1.2 (American Psychiatric Association 1994). Disease onset is typically between the ages of 16 and 30 and is usually earlier in males (Picchioni and Murray 2007). The average life expectancy of affected individuals is reduced by 20-25 years due to lifestyle choices and poor health awareness associated with illness, comorbid medical disorders and an increased rate of suicide (Nechiporuk, Fernandez, and Vasioukhin 2007). Across Europe less than 20% of people with SZ are in employment (Nechiporuk, Fernandez, and Vasioukhin 2007). The worldwide prevalence of the disorder is estimated at approximately 1% (Nechiporuk, Fernandez, and Vasioukhin 2007) and it is 1.4 times more prevalent in males than females (Abrahams and Geschwind 2008).

Biochemical theories that may explain the biology of SZ include the dopamine and glutamate hypotheses. The dopamine hypothesis attributes the positive symptoms of the disorder to a hyperactive dopaminergic system in the brain (Scarr and Dean 2009). Support for this is based on the anti-psychotic effects

of drugs that block dopamine receptors and the psychosis-inducing effect of dopamine-releasing drugs such as amphetamines. The glutamate hypothesis on the other hand implicates a hypoactive glutamate signalling via NMDA receptors (Bailey et al. 1995). This hypothesis is supported by studies that show reduced levels of glutamate receptors in the brains of SZ patients post mortem (Bourgeron 2009). In addition, glutamate antagonists such as phencyclidine and ketamine can produce some of the positive symptoms of SZ as well as attention and memory deficits (Woo, Kim, and Viscidi 2008).

It is a widely held belief that SZ is a neurodevelopmental disorder and early brain insults either pre or perinatally may disturb brain development and lead to abnormalities in the mature brain that cause the disorder at a later age. In support of a neurodevelopmental hypothesis, SZ patients have a higher incidence of dermatoglyphic and minor physical anomalies (Fatjo-Vilas et al. 2008). Environmental risk factors for SZ include prenatal events such as infections, cytokine exposure, maternal malnutrition whilst pregnant during famine and obstetric events (Insel 2010; van Os, Kenis, and Rutten 2010). It is possible that such events affect early neural development but the impact on brain function does not manifest itself as SZ until early adulthood. Other risk factors include living in urban areas and socio-economic factors (Pedersen and Mortensen 2001). In terms of pathology, magnetic resonance imaging (MRI) of SZ patients reveals that there is an enlargement of the lateral ventricles and decrease in total brain and hippocampus volume (Steen et al. 2006). Although the disorder does not develop until early adulthood, there may be early signs such as delayed developmental milestones (Sorensen et al. 2010).

There is no cure for SZ but symptoms can be treated with medication and behavioural and cognitive therapy. The primary treatment is anti-psychotic drugs of which there are two types; typical and atypical. Typical psychotics were the first generation of anti-psychotics to be developed and only act on the positive symptoms and examples include chlorpromazine and haloperidol. These drugs block receptors in the dopamine neurotransmitter pathway and are associated with many adverse side-effects, particularly extra-pyramidal side effects (EPS) such as rigidity or tremors. Atypical anti-psychotics such as olanzapine and clozapine were developed later and also act on neurotransmitter systems but are less likely to cause motor control problems.

The long-term use of any anti-psychotic drug is associated with lower mortality compared to no use of antipsychotic drugs (Nechiporuk, Fernandez, and Vasioukhin 2007).

Table 1.1: Criteria for diagnosis of schizophrenia according to the Diagnostics Statistics Manual Classification IV (DSM-IV)

A. Characteristic symptoms: Two (or more) of the following, each present for a significant portion of time during a 1-month period (or less if successfully treated):
a) delusions
b) hallucinations
c) disorganized speech (e.g. frequent derailment or incoherence)
d) grossly disorganized or catatonic behaviour
e) negative symptoms, i.e. affective flattening, alogia, or avolition
Note: Only one Criterion A symptom is required if delusions are bizarre or hallucinations consist of a voice keeping up a running commentary on the person's behaviour or thoughts, or two or more voices conversing with each other.
B. Social/occupational dysfunction: For a significant portion of the time since the onset of the disturbance, one or more major areas of functioning such as work, interpersonal relations, or self-care are markedly below the level achieved prior to the onset (or when the onset is in childhood or adolescence, failure to achieve expected level of interpersonal, academic or occupational achievement).
C. Duration: Continuous signs of the disturbance persist for at least 6 months. This 6-month period must include at least 1 month of symptoms (or less if successfully treated) that meet Criterion A (i.e. active-phase symptoms) and may include periods of prodromal or residual symptoms. During these prodromal or residual periods, the signs of the disturbance may be manifested by only negative symptoms or two or more symptoms listed in Criterion A present in an attenuated form (e.g. odd beliefs, unusual perceptual experiences).
D. Schizoaffective and Mood Disorder exclusion: Schizoaffective Disorder and Mood Disorder With Psychotic Features have been ruled out because either (1) no Major Depressive, Manic or Mixed Episodes have occurred concurrently with the activephase symptoms; or (2) if mood episodes have occurred during active-phase symptoms, their total duration has been brief relative to the duration of the active and residual periods.
E. Substance/general medical condition exclusion: The disturbance is not due to the direct physiological effects of a substance (e.g., a drug of abuse, a medication) or a general medical condition.
F. Relationship to a Pervasive Developmental Disorder: If there is a history of Autistic Disorder or another Pervasive Developmental Disorder, the additional diagnosis of Schizophrenia is made only if prominent delusions or hallucinations are also present for at least a month (or less if successfully treated).

Table 1.2: Criteria for diagnosis of schizophrenia according to the International Classification of Disease 10 (ICD-10)	
Diagnosis requires the presence of one symptom from section 1 or two symptoms from section 2, for a period of at least one month	
1.	a) Thought echo, insertion, withdrawal or broadcasting b) Delusions of control, influence or passivity c) Hallucinatory voices d) Persistent delusions of other kinds which are culturally inappropriate or implausible.
2.	a) Persistent hallucinations, accompanied by delusions without clear affective content, persistent over-valued ideas b) Breaks of interpolations in the train of thought resulting in incoherence or irrelevant speech or neologisms c) Catatonic behaviour such as excitement, posturing, negativism, mutism and stupor d) Negative symptoms such as marked apathy, paucity of speech and blunting of emotional responses e) Change in the overall quality of aspects of personal behaviour manifest as loss of interest, aimlessness, idleness, self-absorbed attitude and social withdrawal

1.2 Overview of Autism

There is a wide range symptoms associated with ASD that encompass cognitive, emotional and behavioural abnormalities. The key features are impairments in social interaction and communication, understanding and production of verbal and non-verbal IQ, and restricted interests and repetitive patterns of behaviour. Diagnosis is based on DSM-IV or ICD-10, as outlined in Table 1.3 and Table 1.4. The presentation of the disorder is highly variable and ranges from mild to severe with much variation in the cognitive abilities; some cases have above average abilities whilst others have severe intellectual disability. It is lifetime disorder with symptoms generally presenting before the age of three and is usually diagnosed when language delays become apparent. Onset is gradual but approximately 30% have developmental regression (Miles 2011).

According to the DSM-IV and ICD 10, autism spectrum disorders (ASDs) are part of a larger classification known as Pervasive Developmental Disorders (PDDs) which includes Rett’s and childhood disintegrative disorder. ASDs encompass autistic disorder, Asperger’s syndrome, pervasive developmental delay not otherwise specified (PDD-NOS) (Figure 1.1). Autism is the prototypic

form of this group of conditions and is considered as a strict diagnosis, whilst the larger group of conditions are considered as a broad diagnosis. A new version of the DSM (DSM-V) is due to be released in May 2013 and will reportedly include one diagnostic category; ASD which will include autism, Asperger’s syndrome and PDD-NOS. Rett’s syndrome will be listed as a separate disorder (Stankovic, Lakic, and Ilic 2012).

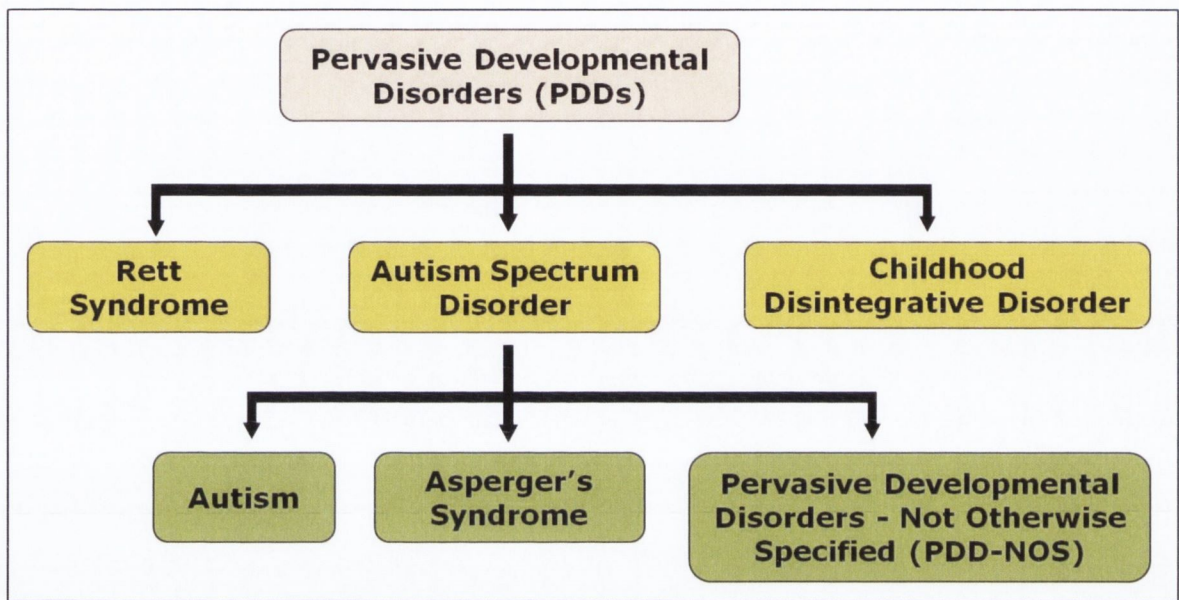


Figure 1.1: Classification of Pervasive Developmental Disorder according to DSM-IV.

ASDs are part of a larger classification known as Pervasive Developmental Disorders (PDDs), which includes Rett’s syndrome and Childhood Disintegrative Disorder. ASDs include autism, Asperger’s syndrome and Pervasive Developmental Disorders Not Otherwise Specified (PDD-NOS).

The prevalence of strict autism is approximately 20 cases per 10,000 (Fombonne 2009), whilst for broad diagnosis autism, it is estimated to be 60-100 per 10,000 (Fernell and Gillberg 2010; Baron-Cohen et al. 2009). There is a 4.2:1 male to female ratio with this increasing to 11:1 in Asperger’s syndrome (Gillberg et al. 2006). Autistic features are present in many other genetic disorders including Fragile X syndrome and Tuberous Sclerosis (Miles 2011). Approximately 25% of children with ASD develop seizures and many are affected by sleep disturbances and gastrointestinal problems (Baron-Cohen et al. 2009). Over 70% of individuals with ASD have intellectual disability (ID) (Tuchman and Rapin 2002). A common feature of ASD is brain overgrowth which may not limited to specific brain regions. In particular, brain circuitry related to social cognition language is likely to be involved

(Geschwind 2009). ASD is a neurodevelopmental disorder and imaging studies support a dysconnectivity in the brains of ASD individuals (Schmitt et al. 2011).

Treatment of ASD generally includes a combination of behavioural and other therapies, education and medication. Medications typically target the cognitive and psychotic elements of the disorder and include antidepressants (selective serotonin reuptake inhibitors (SSRIs) and tricyclics), anti-psychotics and anti-anxiety drugs (Myers, Johnson, and American Academy of Pediatrics Council on Children With Disabilities 2007). Atypical antipsychotics are often used to treat ASD patients with serious behavioural disturbances.

The underlying cause of ASD is unknown but is believed to be as a result of genetic and environmental factors. The genetics of ASD will be discussed in detail below (section 1.6.4). Environmental risk factors with the greatest supporting evidence are exposure to the antiepileptic drug valproate, maternal infection during gestation and maternal stress (Bromley et al. 2008; Rapoport, Giedd, and Gogtay 2012). It may be that such exposures cause deficiencies in fetal neurodevelopment.

Table 1.3: Criteria for the diagnosis of autism spectrum disorders according to the Diagnostics Statistics Manual IV (DSM-IV)

A. A total of six (or more) items from (1), (2), and (3), with at least two from (1), and one each from (2) and (3):
1. Qualitative impairment in social interaction, as manifested by at least two of the following: Marked impairment in the use of multiple nonverbal behaviours such as eye-to-eye gaze, facial expression, body postures, and gestures to regulate social interaction Failure to develop peer relationships appropriate to developmental level A lack of spontaneous seeking to share enjoyment, interests, or achievements with other people (e.g. by a lack of showing, bringing, or pointing out objects of interest) Lack of social or emotional reciprocity
2. Qualitative impairments in communication as manifested by at least one of the following: Delay in, or total lack of, the development of spoken language (not accompanied by an attempt to compensate through alternative modes of communication such as gesture or mime) In individuals with adequate speech, marked impairment in the ability to initiate or sustain a conversation with others Stereotyped and repetitive use of language or idiosyncratic language Lack of varied, spontaneous make-believe play or social imitative play appropriate to developmental level
3. Restricted repetitive and stereotyped patterns of behaviour, interests, and activities, as manifested by at least one of the following: Encompassing preoccupation with one or more stereotyped and restricted patterns of interest that is abnormal either in intensity or focus Apparently inflexible adherence to specific, nonfunctional routines or rituals Stereotyped and repetitive motor mannerisms (e.g. hand or finger flapping or twisting, or complex whole body movements) Persistent preoccupation with parts of objects
B. Delays or abnormal functioning in at least one of the following areas, with onset prior to age 3 years: (1) social interaction, (2) language as used in social communication or (3) symbolic or imaginative play.
C. The disturbance is not better accounted for by Rett's Disorder or Childhood Disintegrative Disorder.

Table 1.4: Criteria for the diagnosis of autism spectrum disorders according to the International Classification of Disease 10 (ICD-10)
A. Abnormal or impaired development is evident before the age of 3 years in at least one of the following areas:
1. Receptive or expressive language as used in social communication; 2. The development of selective social attachments or of reciprocal social interaction; 3. Functional or symbolic play.
B. A total of at least six symptoms from (1), (2) and (3) must be present, with at least two from (1) and at least one from each of (2) and (3)
1. Qualitative impairment in social interaction are manifest in at least two of the following areas: a. Failure adequately to use eye-to-eye gaze, facial expression, body postures and gestures to regulate social interaction; b. Failure to develop (in a manner appropriate to mental age, and despite ample opportunities) peer relationships that involve a mutual sharing of interests, activities and emotions; c. Lack of socio-emotional reciprocity as shown by an impaired or deviant response to other people's emotions; or lack of modulation of behaviour according to social context; or a weak integration of social, emotional, and communicative behaviours; d. Lack of spontaneous seeking to share enjoyment, interests, or achievements with other people (e.g. a lack of showing, bringing, or pointing out to other people objects of interest to the individual).
2. Qualitative abnormalities in communication as manifest in at least one of the following areas: a. Delay in or total lack of, development of spoken language that is not accompanied by an attempt to compensate through the use of gestures or mime as an alternative mode of communication (often preceded by a lack of communicative babbling); b. Relative failure to initiate or sustain conversational interchange (at whatever level of language skill is present), in which there is reciprocal responsiveness to the communications of the other person; c. Stereotyped and repetitive use of language or idiosyncratic use of words or phrases; d. Lack of varied spontaneous make-believe play or (when young) social imitative play.
3. Restricted, repetitive, and stereotyped patterns of behaviour, interests, and activities are manifested in at least one of the following: a. An encompassing preoccupation with one or more stereotyped and restricted patterns of interest that are abnormal in content or focus; or one or more interests that are abnormal in their intensity and circumscribed nature though not in their content or focus; b. Apparently compulsive adherence to specific, non-functional routines or rituals; c. Stereotyped and repetitive motor mannerisms that involve either hand or finger flapping or twisting or complex whole body movements; d. Preoccupations with part-objects of non-functional elements of play materials (such as their odour, the feel of their surface, or the noise or vibration they generate).

1.3 Phenotypic Overlap Schizophrenia and Autism

SZ and ASD are highly heterogeneous and the line of demarcation is particularly unclear between the broad phenotypes of these disorders. Interestingly, up until the 1970's ASD was often diagnosed as childhood schizophrenia. One of the major achievements of the DSM-III published in the 1980's was to separate ASD and SZ into different diagnostic categories (American Psychiatric Association 1980).

There are many features of SZ that are common to ASD e.g. unusual pre-occupations, odd-speech, lack of close relationships and social anxiety. The main distinguishing feature between the two disorders is paranoia. Cognitive deficits are a common element of both disorders. Theory of mind can be described as an inability to understand the perspective of others and is a feature of SZ and ASD (King and Lord 2011). The severity of these deficits is related to presence and severity of certain symptoms of SZ including delusion. A component of theory of mind in ASD is difficulty with reciprocal communication.

Mirror neurons are a system of neurons in the prefrontal cortex that fire both when an action is performed and upon observation of an equivalent action performed by another individual. Brain imaging studies of this system compare the activation during both scenarios and many have reported reduced activation of the system in ASD children compared to typically developing children in response to observation of actions performed by others (Rizzolatti and Fabbri-Destro 2010). Imaging studies of SZ patients have also observed an abnormal pattern of activation of this system (Enticott et al. 2008). However, it may be that these abnormalities are due to an overall connectivity deficit in the brains of individuals with ASD and SZ. A wealth of neuro-imaging (MRI and transcranial magnetic stimulation (TMS)), electrophysiology (electroencephalography; EEG) and molecular biology methods, support neural dysconnectivity in both ASD and SZ (King and Lord 2011).

1.4 Genetic Variation

The human genome is composed of approximately 3 billion nucleotides and is highly variable among individuals. DNA sequence variants range in size from small single base pair changes to large variations in copy number of sections of the genome. Variants can be neutral, which is most often the case, they can be beneficial, which contributes to evolution or they can be harmful and contribute to disease. The impact of a variant depends not only on the type of variant it is but where it resides relative to functional elements of the genome. I will detail the most well understood human genetic variation in terms of disease genetics.

1.4.1 Single Nucleotide Variants

Single Nucleotide Variants (SNVs) are sequence variation resulting from a nucleotide substitution that occurs at a single base. SNVs are considered to be common if they occur at a frequency of $>1\%$ in the general population and are deemed single nucleotide polymorphism (SNPs). The International HapMap Consortium is an organisation which, through genotyping of SNPs, has had the goal of cataloguing common variation throughout the genome and providing this information as a public resource in order to accelerate population genetics and medical genetics research (www.hapmap.org/). The 1000 Genomes (1KG) project is an international collaboration that aims to extend resources such as HapMap, by sequencing in order to develop a resource of almost all variants in the human genome (<http://www.genome.gov/>). A pilot study set out to achieve this by whole genome sequencing 179 unrelated individuals across four diverse populations. In doing so, this project has identified approximately 15 million SNPs and estimates that this accounts for most of the common SNPs in the human genome (Genomes Project Consortium 2010). There is a far greater number of low-frequency (MAF 1-5%) and rare variants (MAF $<1\%$) but it is not possible to systematically identify them all at this time. However, the 1KG project has estimated that there are approximately 3 million SNV sites per individual genome.

SNVs can occur in coding regions of genes in which case they can be synonymous or nonsynonymous. Synonymous or silent SNVs do not change the amino acid sequence of the resulting protein. Conversely, non-

synonymous mutations do change the protein sequence and are composed of missense and nonsense mutations. Missense mutations result in a single amino acid change, whereas nonsense mutations result in a change from an amino acid codon to a stop codon, thus truncating the polypeptide sequence. Non-synonymous SNVs can have an obvious impact on protein function and expression. SNVs can reside in untranslated regions (UTRs) or non-coding regions where they can have an effect on genes regulation, e.g. by affecting the binding of transcription factors. Another type of SNVs are those that occur at the highly conserved splice donor and acceptor sites at either end of intron-exon boundaries. Splice acceptor sites are located at the 2bp of intronic sequence directly upstream of the 5' end of exons and usually consist of an AG sequence. Splice donor sites are directly downstream of the 3' end of exons and usually include a GT sequence. Recognition of these sites is required by splicing machinery for post-transcriptional regulation. Variants at these sites can result in exon skipping, retained introns or protein truncation.

1.4.2 Insertion/Deletions

Small insertions or deletions (indels) cause a substantial amount of human genetic variation. Indels are very difficult to call with great accuracy. The best estimate currently can be garnered from the 1KG project which has identified an average of 350,000 indels per human genome (Genomes Project Consortium 2010). As with SNVs, they can occur in non-coding regions where they may have a regulatory effect or they can occur in coding regions where their effect can be more easily interpreted. In coding regions, unless the length of an indel is a multiple of 3 nucleotides, it produces a frameshift mutation, which typically results in the introduction of a premature stop codon and thus protein truncation. Indels that are multiples of 3 nucleotides result in the loss or gain of amino acids but the subsequent amino acid sequence remains intact. These mutations can also occur at donor or acceptor splice sites, thereby altering gene splicing.

1.4.3 Structural Variants

Structural variants are deletions, insertions, inversions or duplications of segments of DNA of greater than 1kb in size within the genome. They are

collectively termed copy number variants (CNVs). The size, category and location of these variants determine their functional impact. CNVs can occur in non-coding regions where they may impact regulatory regions. Where CNVs impact genes, they can affect a single exon, multiple exons, the entire gene or often large numbers of genes located adjacent to each other on a chromosome. Approximately 12-15% of the human genome varies in copy number and as such CNVs contribute greatly to genetic variation (Sonnenberg and Liem 2007; Leung et al. 2001).

1.4.4 Linkage Disequilibrium

Alleles at different loci in the genome are not always independent of each other because they often display non-random assortment, i.e. an allele at one locus can be found to co-occur with another allele at another locus more often than expected by chance. This is termed linkage disequilibrium (LD) and groups of alleles in LD with each other are termed haplotypes. A number of processes contribute to the presence of LD and influence the level of LD across the genome, including rate of mutation, recombination and genetic drift (Ardlie, Kruglyak et al. 2002). By genotyping millions of SNPs in 100's of individuals from different reference panels, HapMap has developed LD maps across the genome in different populations. This data is publicly available at www.hapmap.org.

1.5 Genetics of Common Complex Disorders

Rare diseases that are caused by a single gene such as cystic fibrosis or Huntington's disease are known as monogenic Mendelian diseases (Riordan et al. 1989; Gusella et al. 1983). Although such diseases may be complicated by factors such as incomplete penetrance, they generally display a clear mode of Mendelian inheritance. Common complex disorders such as cardiovascular disease, diabetes or neuropsychiatric disorders are caused by a combination of genetic and environmental risk factors. It is more difficult to map genes for these disorders because of factors such as genetic heterogeneity (multiple risk genes and potentially multiple risk variants at each gene), incomplete penetrance (individuals carry a risk variant but do not develop disease) and phenocopies (individuals develop illness due to a non-genetic cause).

There are two opposing hypotheses for the inheritance of complex disorders: the “common disease, common variant” hypothesis (CDCV) and “common disease, rare variant” hypothesis (CDRV). The CDCV model posits that disease is caused by the presence of many common variants (minor allele frequency (MAF) $>5\%$), each with a small effect whereas the CDRV model posits that disease is caused by few rare variants (MAF $<5\%$) of large effect. Figure 1.2 illustrates the feasibility of identifying genetic variants by risk allele frequency and strength of genetic effect (odds ratio (OR)). Rare variants with very large effects are easily detectable by linkage studies and are responsible for Mendelian disorders. There are however, very few examples of common variants of large effect. An example exception is APOE (Apolipoprotein E) and Alzheimer’s Disease where the e4 allele has a frequency of $\sim 17\%$ in Caucasian populations and e4/e4 genotype has an associated OR of 14.9 for disease (Farrer et al. 1997). To date, the main focus of research into common disorders has been on common variants using genome-wide association studies (GWAS) as discussed below in section 1.6.2. Rare variants are not well captured by GWAS (Manolio et al. 2009) and I will describe below how rare variants are now being studied using next-generation sequencing. It is likely that the genetic architecture of illness could be different for different complex disorders, and within any given disorder there is likely to be a mixture of common and rare risk variants.

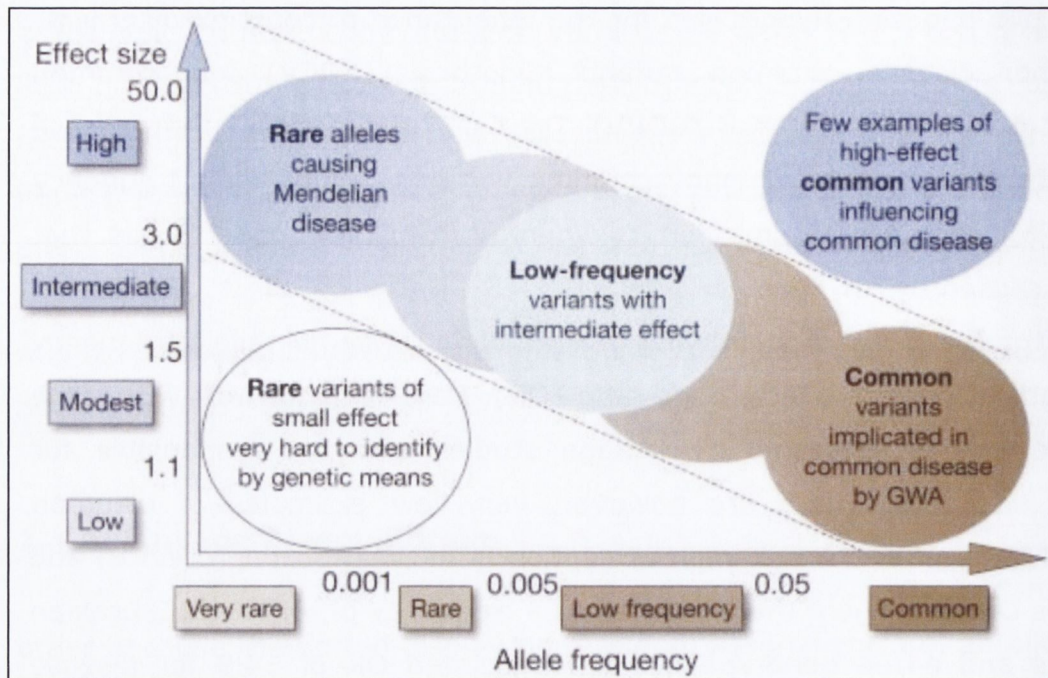


Figure 1.2: Feasibility of identifying genetic variants by risk allele frequency

(Manolio et al. 2009)

1.6 Gene Discovery Approaches

1.6.1 Linkage Analysis

Linkage studies are a hypothesis free approach for identifying chromosomal regions containing causative mutation for disease. These studies determine if there is a co-segregation of genetic markers such as SNPs or microsatellite markers (repeating sequences of 2-6bps) with disease. These markers are genotyped in a very large multi-generational family or a number of smaller families and a test statistic called a LOD score (logarithm (base 10) of odds) is calculated to determine the level of segregation of different markers with disease in these families. Studies can either be performed as genome-wide screens with a dense set of markers or in particular chromosomal regions. Once linkage regions of interest are identified, fine-mapping is required to identify causal variants. A major challenge for such studies is that large families are required and these can be difficult to identify. It is more applicable to monogenic disorders in which rare highly penetrant variants are involved and has been hugely successful for disorders such as cystic fibrosis and Huntington's disease (Riordan et al. 1989; Gusella et al. 1983).

1.6.2 Association Analysis

Association studies are an alternative approach to linkage studies for gene mapping. Instead of using large family samples they generally use population-based case (affected) and control (unaffected) samples, although family-based association methods are sometimes used, but mostly for childhood disorders. For any genetic variant, its frequency is calculated in affected individuals and compared to the variant frequency in the unaffected individuals. A significant difference in frequency can identify a variant as being associated with disease. Association studies were originally undertaken at candidate regions that were highlighted by linkage studies or at genes selected for their biological function. Such studies were largely unsuccessful because either the linkage results were incorrect, or the biological candidate genes were incorrect, or even if a true susceptibility region was being studied, the small samples sizes in use did not have enough power to detect variants of low effect.

With technological advances it became possible to examine SNPs across the entire genome using genome-wide association studies (GWAS). Up to millions of SNPs can be captured on a single genotyping array. This dense coverage captures a very high proportion of common human genetic variation (International HapMap Consortium et al. 2007). Furthermore, algorithms were developed that could use a dense reference dataset in conjunction with the genotyped samples to impute calls for variants that were not directly typed. This approach took advantage of the LD patterns in the reference panel to predict the genotypes in the test samples. Imputation was widely used in the analysis of GWAS data to boost study power and associations and to facilitate the combination of data across different studies in meta-analyses (Marchini and Howie 2010).

GWAS is a hypothesis free approach and thus is not affected by lack of understanding of disease pathogenesis. The causal SNP is rarely directly genotyped, it is more often the case that the SNPs are in LD with the true causal variant. In this case, the typed SNPs in LD with the causal SNP will each only show moderate effects since they are not in perfect LD with the causal variant. Very large sample numbers are required to detect associations that have low effect sizes and can survive for correction for multiple testing. Independent replication is required to confirm any findings. GWAS have

yielded quite dramatic results with many risk loci identified for hundreds of traits and diseases (<http://www.genome.gov/gwastudies/>).

A second type of genetic variant that could be assayed by SNP genotyping arrays used in GWAS are CNVs. If a number of SNPs in close proximity on a chromosome have significantly increased or decreased probe signal intensity, a duplication or deletion event can be detected. Specialized arrays for CNV detection have also been developed using a method known as array comparative genomic hybridization (aCGH). This method is also based on increased or decreased intensity signals following the hybridization of DNA to a microarray. A change in copy number can be detected at a specific locus when the test DNA sample is compared to a reference panel mix of DNA samples (Pinkel et al. 1998). Population and family-based association analysis of CNVs have delivered important new discoveries in neuropsychiatric disorders (see section 1.6.3 and 1.6.4 below).

1.6.3 Genetic Evidence in Schizophrenia

SZ has a strong genetic component and this is demonstrated in families where the co-occurrence of the disorder is correlated to the degree of genetic relatedness between family members (Figure 1.3). Twin studies compare the rates of a disorder in monozygotic (MZ) twins (that share 100% of their genes) to dizygotic (DZ) twins (that share ~50% of their genes), which both have a shared environment. The concordance rates in twin studies can be used to determine the heritability of a disorder. The heritability of SZ has been estimated at approximately 80% (Cardno and Gottesman 2000). This indicates that ~80% of the variance in the phenotype can be attributed to genetic factors.

Pre-GWAS, the strongest evidence for a risk gene from linkage analysis in SZ was the identification of segregation of major psychiatric illness with a chromosome 1 to chromosome 11 translocation in a large Scottish pedigree (St Clair et al. 1990). This mutation impacted a gene on chromosome 1 that was then called *disrupted in schizophrenia 1* (*DISC1*). *DISC1* has since become the most widely-studied candidate gene for SZ (Myers et al. 2011). Numerous other linkage findings have been reported for SZ but few have been

consistently replicated. Similarly, candidate gene association studies delivered few consistent findings for SZ before GWAS.

As I began my PhD studies, the first GWAS for SZ were being reported. Genes containing SNPs that reached the genome-wide significance threshold of $p \leq 5 \times 10^{-8}$ included *ZNF804A* (O'Donovan et al. 2008), *TCF4* and *NRG1* (Stefansson et al. 2009) and the Major Histocompatibility Complex (MHC) region on chromosome 6 (Stefansson et al. 2009; Shi et al. 2011; International Schizophrenia Consortium et al. 2009)

At the same time GWAS was producing very interesting results for CNVs in SZ. Large but rare chromosomal deletions (100Kb – 3Mb) were found to occur more frequently in SZ cases compared to normal healthy controls (International Schizophrenia Consortium 2008). Most of these deletions are spread across the genome and did not implicate specific genes. However, some SZ deletions did cluster at certain regions, implicating specific genes in the illness. Specific loci known as deletion regions that were consistently identified include regions at chromosomes 1q21.1, 15q11.2, 15q13.3 and 22q11.2 (Stefansson et al. 2008; International Schizophrenia Consortium 2008). All regions were new except 22q11, which had previously been identified as a rare risk locus for SZ (Bassett et al. 2008).

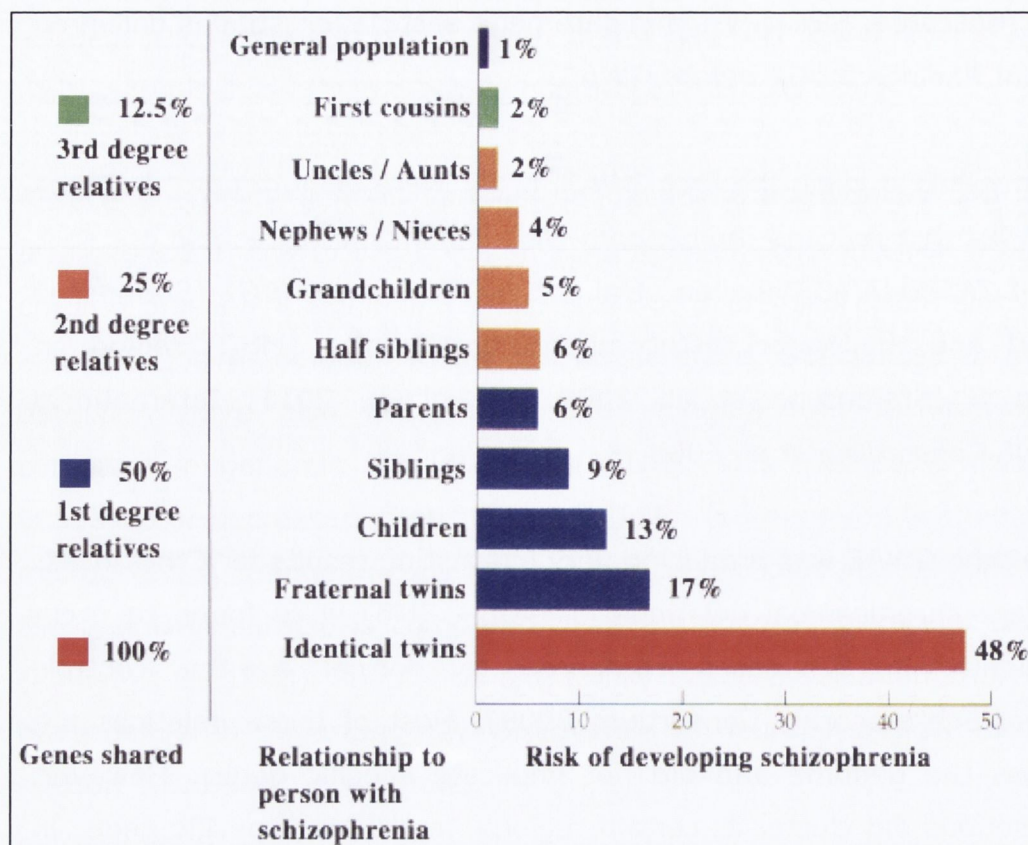


Figure 1.3: Risk of developing SZ amongst relatives of SZ patients (Gottesman and Erlenmeyer-Kimling 2001)

1.6.4 Genetic Evidence in Autism

There is a very strong genetic component to ASD with heritable estimates as high as 92% (Bailey et al. 1995). However, a recent publication argues a far lower heritability estimate of 40% (Hallmayer et al. 2011). A wealth of linkage studies have been performed in ASD and at least nine regions have been identified in at least two independent studies (Freitag et al. 2010). Some of the best supported regions include 7q, 10p and 17p (Trikalinos et al. 2006; Autism Genome Project et al. 2007).

Many of these regions were followed up by fine mapping and candidate association studies. Such studies have yielded inconsistent findings for ASD due to a variety of reasons as discussed previously in section 1.6.4. To date, three large independent GWAS studies have been performed in ASD. Two reported significant association at two different loci: 5p14.1 and 5p15.2 (Weiss et al. 2009; Wang, Zhang, et al. 2009). The 5p14.1 region harbours two cadherin genes (*CHD9* and *CHD10*). The 5p15.2 region encodes two genes;

taste receptor 2 (*TAS2R1*) and semaphorin 5A (*SEMA5A*). The third study by the Autism Genome Project (AGP) consortium identified an association at 20p12.1 which contains the gene MACRO domain containing 2 (*MACROD2*) (Anney et al. 2010).

There is an increased rate of CNVs in individuals with autism (Sebat et al. 2007). A number of recurrent CNVs have been identified including regions on chromosomes 1q21.1 (Pinto et al. 2010), 15q11-13 (Glessner et al. 2009) and 16p11.2 (Weiss et al. 2008). Glessner et al. identified an increased rate of recurrent CNVs affecting genes involved in neuronal cell adhesion and upon replication they further implicated neuronal cell adhesion and ubiquitin degradation (Glessner et al. 2009).

1.7 Genetic Overlap in Schizophrenia and Autism

There is an increased rate of co-occurrence of SZ and ASD in families (Devlin and Scherer 2012). Many studies have identified rare risk variants that are found in both disorders (Hamada and Yagi 2001; Carroll and Owen 2009). In particular, many CNVs have been identified that are common to both disorders e.g. 1q21.1 (International Schizophrenia Consortium 2008; Mefford et al. 2008), 15q11.2 (Hogart et al. 2010; Stefansson et al. 2008), 15q13.3 (International Schizophrenia Consortium 2008; Miller et al. 2009), 16p11.2 (McCarthy et al. 2009; Weiss et al. 2008) and 17q12 (Kirov et al. 2009). It is important to note however that there is substantial heterogeneity at these sites in terms of type (deletion or duplication), penetrance and size of the risk CNVs. These sites often contain many genes; therefore it is difficult to identify the specific risk gene(s). One example where just a single gene is implicated in both disorders is Neurexin 1 (*NRXN1*). Multiple studies have reported that rare deletions at *NRXN1* are risk factors for SZ and ASD (Autism Genome Project et al. 2007; Kirov et al. 2008; Kim et al. 2008; Marshall et al. 2008; Vrijenhoek et al. 2008; Walsh et al. 2008; Glessner et al. 2009; Rujescu et al. 2009).

1.8 Next Generation Sequencing

In recent years there has been a massive shift away from Sanger DNA sequencing, a method that was extremely valuable since the 1970's, to new ultra-high throughput DNA sequencing using next generation sequencing (NGS). The main advantages of NGS are the ability to produce many accurate sequence reads in a short space of time at a relatively cheap cost. These new technologies consist of a number of common elements; template preparation, sequencing and imaging, and sequence alignment and assembly (Metzker 2010). Many companies have released their own version of NGS platforms, with unique combinations of the above steps and this determines the type of data produced and how some of these technologies are more suited to different applications. Some of the companies that have released NGS platforms include Illumina/Solexa, Roche/454, Helico Biosciences and Life Technologies/Solid (Metzker 2010). Most methods involve fragmentation of DNA to smaller sizes from which either single fragment templates or mate-pair templates are created. These templates are generally bound to or immobilized to a solid surface and are spatially separated allowing up to millions of sequencing reactions to take place in parallel. There are four different methods that are used for both single templates and clonally amplified templates; cyclical reversible termination (CRT), ligation, pyrosequencing and real time sequencing. CRT involves the incorporation of fluorescent reversible nucleotide terminator complement to the template base, followed by fluorescent imaging and cleavage of the terminator, in a cyclical manner. The Illumina Genome Analyzer platform uses a four colour CRT method whereby the four nucleotide groups with separate labelled dyes each, are incorporated and imaged simultaneously (Metzker 2010).

1.8.1 Targeted Resequencing

The relevance of NGS to genetic studies of SZ and ASD is that it permits the study of rare variants to extend beyond large CNVs to smaller CNVs, indels and single base substitutions that may be important contributors to risk of illness. This will allow a better understanding of the genetic variation underlying these complex phenotypes. While NGS makes it possible to sequence the entire human genome in a quick and cheap manner, performing this analysis on a large number of individuals is still prohibitively expensive.

Therefore, it is practical to target regions of interest within the genome for sequencing. Much effort has gone into developing methods of target enrichment (also known as genomic partitioning). Genomic regions are selectively targeted and enriched before sequencing. Sequences of interest can be exons or full genes or the whole exome (Mamanova et al. 2010). Exome sequencing involves target enrichment of the exons of most genes in the genome. It has become popular as the cost of sequencing has dropped but remains a costly approach. It has been very successful in the study of rare single gene disorders e.g. identification of *MYH3* as the causative gene for Freeman-Sheldon syndrome (Ng et al. 2009).

Resequencing target regions is much more cost and time efficient than whole genome sequencing, and also makes it possible to sequence many more individuals. For example, sequencing 1Mb of targeted DNA (e.g. 500 genes) in 3,000 people or 30MB (whole exome sequencing) in 100 people is likely be far more informative for mapping a gene for a genetic disorder compared to sequencing 3,000Mb (a genome) in 1 individual.

A number of methods have been developed in recent years for target enrichment. Some performance metrics that determine the efficiency of these methods include; capture specificity, uniformity, sensitivity, allelic bias, input requirement, multiplexity, scalability and perhaps most importantly cost. Methods that have high specificity and uniformity will require less sequencing to provide adequate coverage for end data analysis (Mamanova et al. 2010). There are a number of strategies for target enrichment including; multiplex PCR, target circularisation-based methods and hybridisation-based methods (both on-array or in-solution methods). As I started my project, my group were testing the SureSelect in-solution based hybridisation method by Agilent. This was the method of target enrichment recommended by Illumina and the NGS platform operated by my group was an Illumina Genome Analyzer II.

1.8.2 Indexing

Target enrichment is a very effective way of reducing sequencing costs despite increasing the amount of time and labour required for sample preparation. Despite this overall increase in efficiency, sequencing a large number of

samples remains an expensive task. Thus, a significant amount of effort has been devoted to modifying and improving sequencing approaches to allow multiple samples to be analyzed at the same time. This is the process of multiplexing samples. Specifically, methods have been developed to index or barcode samples allowing enrichment and/or sequencing of samples in parallel (Craig et al. 2008). Craig et al. developed a framework for multiplexing samples on the Illumina Genome Analyzer by ligating short identifying oligonucleotides to fragments prior to sequencing. The six-base indexes have built-in redundancy for error correction so at least one sequencing error could be tolerated before an index could mutate into a different valid index.

This method has been further developed by Kenny et al. in my research group to index samples prior to SureSelect target enrichment at high efficiency (Kenny et al. 2011). A unique 6bp index is incorporated into the universal adapter sequences. Sequencing libraries with multiple DNA samples can be indexed, combined in equal amounts, enriched using SureSelect target enrichment and sequenced in a single sequencing run. Samples can easily be identified in downstream data analysis using the unique indexes. Kenny et al. have demonstrated the ability of this approach to detect SNPs accurately without a compromise on uniformity of coverage. Indexing significantly reduces enrichment and sequencing costs in studies containing a large number of samples.

1.9 Project Aims

The underlying pathogenesis of SZ and ASD are poorly understood and there is a great need for diagnostic tools and new therapeutic interventions. Both have a very strong genetic component but the genetic architecture of them remains to be elucidated. The search for common variants contributing to these disorders has had some success for SZ but has not been very fruitful for ASD. Spurred on by the unexpected success with CNV studies in SZ and ASD, the field is expanding to study other rare variants that may contribute to disease. This has become possible with the advent of next generation sequencing and target enrichment technologies.

There is a strong phenotypic and genetic overlap between these disorders indicating shared pathogenesis. In particular, a number of risk variants related to synaptic function and glutamate signalling have been identified. As discussed above, there are many CNVs identified in both disorders, for example, CNVs affecting *NRXN1*, the neuronal cell adhesion molecule that plays a major role in synaptic function. Additional evidence that abnormal synapse formation and maintenance is part of the pathogenesis of both SZ and ASD comes from other CNV studies in SZ (Walsh et al. 2008) and ASD (Pinto et al. 2010), transcriptomic analysis of the brain in SZ (Mudge et al. 2008). Where SZ and ASD have been combined for CNV (Guilmatre et al. 2009), or sequencing analysis (Myers et al. 2011), data supports shared biological pathways for the disorders in synaptogenesis and glutamate neurotransmission.

The aim of this study is to identify rare variants that contribute to pathophysiology of ASD and SZ in genes involved in synaptic functioning and in glutamate signalling. The exons of a total of 215 candidate susceptibility genes related to these functions will be sequenced in 150 cases of ASD, 300 cases of SZ, and 300 controls. The resultant data on rare genetic variants will be used to test for association between these genes and the disorders.

Chapter 2 will detail the material and methods employed throughout this study. Chapter 3 describes the selection of candidate genes for sequencing, the design of the target enrichment array, the optimisation of the multiplex target enrichment protocol and application of it to a test pool of samples. Chapter 4 explores the findings from sequencing of the full dataset of SZ, ASD and control samples. Particular emphasis is placed on rare variants, which undergo annotation, *in silico* analysis and association analysis. Chapter 5 outlines the functional investigation in primary neuronal cultures from mice of a loss-of-function variant identified in an interesting candidate gene. And finally, chapter 6 provides a discussion of the work, future directions and my conclusions.

2 Materials and Method

2.1 Gene Selection

The aim of the study is to sequence the exons of approximately 200 genes in approximately 150 cases of ASD, 300 cases of SZ and 300 controls in an attempt to identify rare mutations that contribute to the pathogenesis of one or both disorders. Specifically, genes with a role in synaptic function and glutamate signalling will be investigated. Gene selection was based on 5 categories; 1) *NRXN1* and interactors, 2) *DISC1* and interactors, 3) Neural cell adhesion molecules, 4) Post-synaptic Glutamate Receptor Complexes (*NMDA*, *mGluR5* and *AMPA*) and 5) Functional and Positional Candidates. The primary approaches employed to select genes were; an extensive literature review, use of protein interaction databases (Table 2.1) and KEGG pathways (<http://www.genome.jp/kegg/pathway.html>). A comprehensive description of this process is located in section 3.2. The total number of genes included in the study was 215, which can be found in Tables A-E in the appendix, along with their genomic location and reasons for selection. Some genes are members of multiple categories but are listed based on method of first selection.

Table 2.1: Protein interaction databases used in gene selection	
Interaction Database	Website address
BioGrid	http://thebiogrid.org/
HPRD	http://www.hprd.org/
IntAct	http://www.ebi.ac.uk/intact/
String	string-db.org/

2.2 Target Enrichment Array Design

In order to maximize the number of genes included in the study, exonic regions only are sequenced. This is achieved using the technique of target enrichment as described in section 1.8.1. In this study, genomic regions of interest are selectively targeted and enriched for prior to sequencing using a method of target enrichment developed by Agilent Technologies called SureSelect (Gnirke et al. 2009). This is a solution-based hybridisation method as outlined in Figure 2.1.

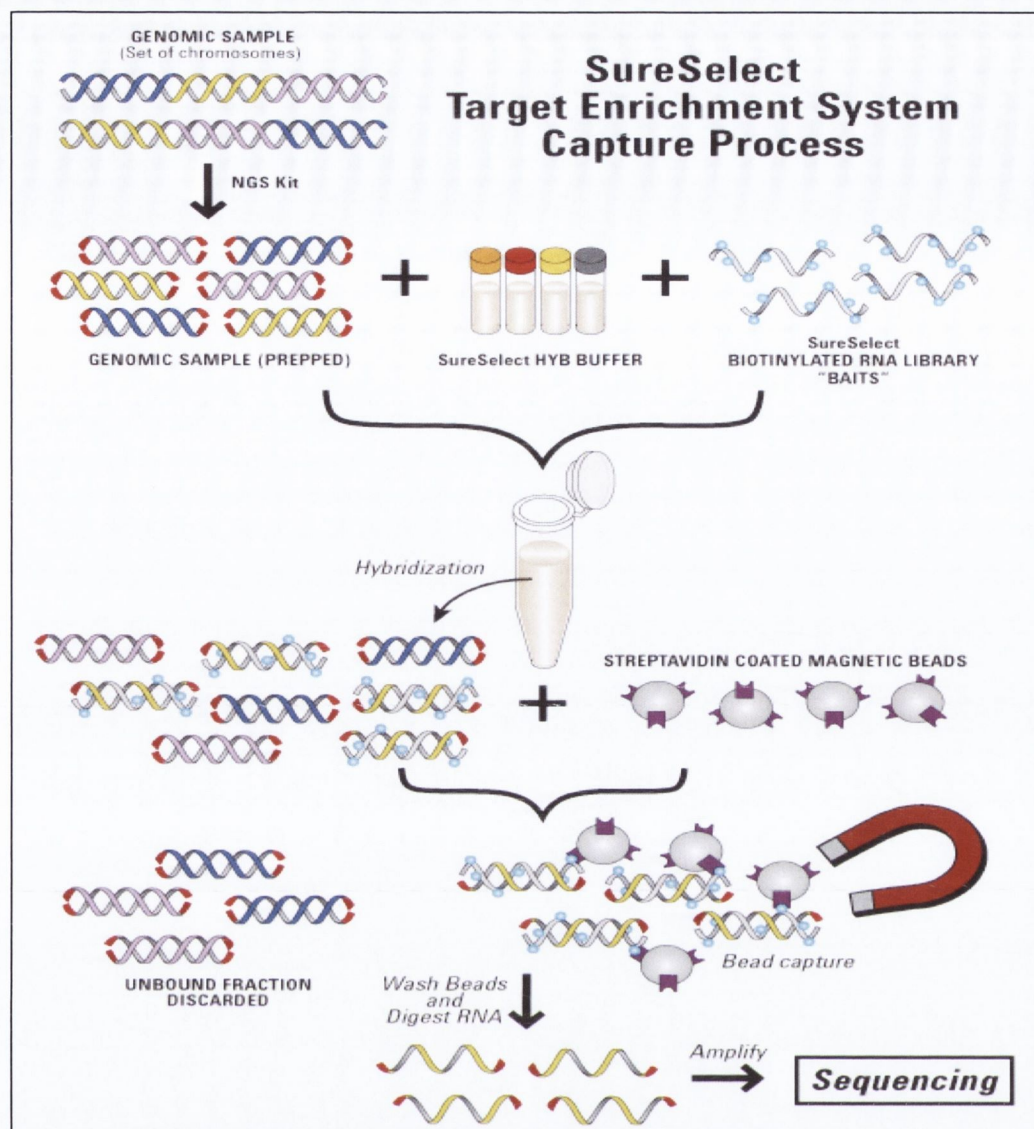


Figure 2.1 Agilent's SureSelect target enrichment system

A DNA library is fragmented and universal adapters ligated to fragment ends. The library hybridises to biotinylated RNA probes upon mixing in solution. Following this, streptavidin beads are added to the solution, which bind to the biotinylated probes allowing targets to be pulled down through the use of magnetics. Captured targets are then amplified by PCR using primers complementary to the universal adapters before sequencing (earray.chem.agilent.com).

I used Agilent's Technologies online design programme eArray (<https://earray.chem.agilent.com/>) to design cRNA baits. To begin with, I uploaded the finalised list of 215 genes to UCSC genome browser (genome.ucsc.edu/) to determine hg18 gene coordinates, which I sent via my web browser to the data manipulation package Galaxy (<http://main.g2.bx.psu.edu>). Exon coordinates were extracted in Galaxy and the generated bed file uploaded to eArray. I set the parameters to default and submitted the design. Following processing, result files are generated

containing all necessary information about the baits that will be generated from this design, i.e. number, genomic coordinates etc.

Upon inspection of exons that are the same size or smaller than baits (i.e. $\leq 120\text{bp}$) it was noted that they were only covered by a single bait (Figure 2.2A). This could possibly lead to underrepresentation of these small exons if the single target bait did not function well. Therefore, I artificially inflated any target exons that were $\leq 120\text{bp}$ to 121bp by extension of the start and stop coordinates on either side of the target sequence. I redesigned the array with these extended exons and uploaded the new extended exon coordinates and baits to UCSC and set them as a custom tracks. The small exons were now covered by at least 3 baits (Figure 2.2B).

Using default settings, baits could be generated for 3,672 of 3,709 target exons (99%). The total quantity of target sequence was 1,064,238bp. The bait library consisted of 14,091 baits, which accounts for 24.4% of the total array occupancy of 57,750 baits, so the library was replicated four times to make use of 97.6% of the available space on the array.

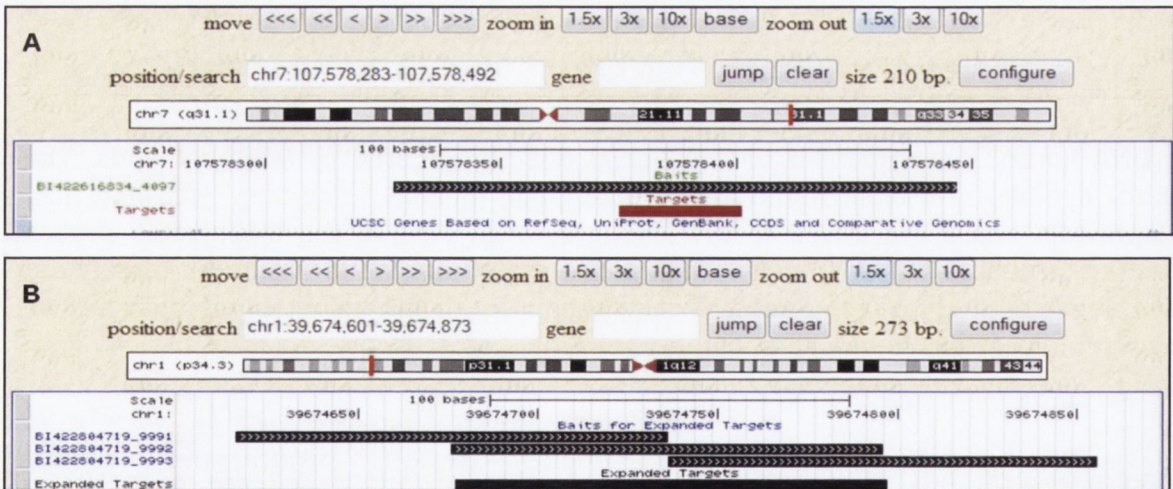


Figure 2.2: Expansion of small (<120bp) target exons for target enrichment.

A) UCSC image of bait coverage on a small target exon (<120bp). B) UCSC image of bait coverage on an artificially expanded target exon.

2.3 Selection of Samples

2.3.1 Autism

ASD case samples (n=168) were recruited by Prof. Louise Gallagher and colleagues in the Neuropsychiatric Genetics Research Group at TCD through schools, parent support groups and clinician referral with local ethics approval. Autism diagnoses were confirmed using Autism Diagnostic Interview – Revised (ADI-R) (Lord, Rutter, and Le Couteur 1994) and the Autism Diagnostic Observation Schedule - Generic (ADOS-G) (Lord et al. 1989). Individuals with known medical causes of autism (for example, tuberous sclerosis, extreme prematurity and congenital rubella) were excluded. Subjects met the ADI-R criteria for autism, and the ADOS-G criteria for autism or ASD. All individuals had an IQ >35 or a mental age (as defined by the Vineland Adaptive Behaviour Scale >18 months (Sparrow, Cicchetti, and Balla 2005). DNA was extracted from blood or buccal swabs. The final sample used for analysis was 83.1% male.

2.3.2 Schizophrenia

SZ case samples (n=306) were recruited by Prof. Aiden Corvin and colleagues at TCD through community mental services and inpatient units in the Republic of Ireland with local ethics approval. All participants were interviewed using a structured clinical interview (Structured Clinical Interview for DSM-IV (SCID-P). Diagnosis of a major psychotic disorder was made by the consensus lifetime best estimate method using DSM-IV criteria with all available information (interview, family or staff report and chart review). All cases were over 18 years of age, of Irish origin (being of Irish parents and having all four grandparents born in Ireland or the United Kingdom) and had been screened to exclude substance-induced psychotic disorder or psychosis due to a general medical condition. DNA for all samples was extracted from blood. The final sample used for analysis is 65.2% male and had a mean age at collection of 47.1 years (standard deviation (SD) = 19.4).

2.3.3 Control Samples

Control samples (n=304) were ascertained with informed consent from the Trinity Biobank and represent blood donors from the Irish Blood Transfusion

Service recruited in the Republic of Ireland. They met the same ethnicity criteria as cases but were not specifically screened for psychiatric illness. Individuals taking regular prescribed medication are excluded from blood donation in Ireland and donors are not financially remunerated. As the lifetime prevalence of schizophrenia or autism is relatively low (<1%), and there no obvious reason for individuals with either disorder to be over represented in the controls, the fraction of putative case individuals in the control collection is expected to be small. DNA for all samples was extracted from blood. Control samples were matched to case samples for gender. The final sample used for analysis was 65.9% male and had a mean age at collection of 34.0 years (SD = 12.6).

2.4 Illumina Sample Preparation Optimisation

2.4.1 Overview

DNA samples must be prepared for sequencing on the Illumina platform. Prior to this project the standard Illumina sample preparation method was employed as per the paired-end sequencing sample preparation guide (September 2009) (Figure 2.3) with two exceptions as outlined below. The first exception is that input genomic DNA (1µg) is fragmented by sonication as per section 2.4.2, instead of with the Covaris system recommend by Illumina. The fragments are blunt ended and phosphorylated, and a single 'A' nucleotide is added to the 3' ends of the fragments. Indexed-adapters that have a single-base 'T' overhang are ligated to the A' overhang. Adapter ligated fragments are size-selected by agarose gel electrophoresis. Size-selected DNA is PCR amplified to enrich for adapter-ligated fragments. The second exception is that there is no second size selection step. DNA samples are purified in between each step by Qiagen column purification (section 2.4.5.1).

This was a labour intensive task when large samples are required so it was necessary to make this method more amenable to large sample numbers. I will describe optimisation of the protocol in the upcoming sections.

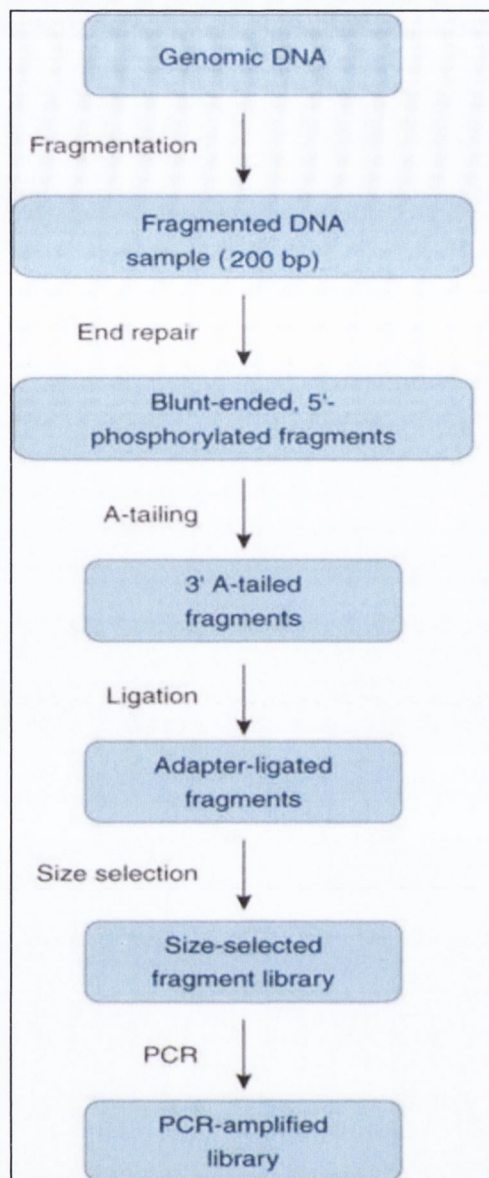


Figure 2.3: Illumina DNA sample sequencing preparation

DNA is fragmented to approximately 200bp; ends are repaired, an A' overhang is added to the blunt ends, adapters with indexes incorporated are ligated, unligated adaptors are removed, size selection is achieved by gel excision and PCR enrichment of targets is carried out using primers specific to the adaptors. The sample is then ready for target enrichment followed by sequencing. DNA samples are purified after each step of the protocol.

There are three particular aspects of the protocol that I targeted for amendments. Firstly, I attempted to alter the ligation of adapters to a cheaper alternative (section 2.4.3). Secondly, I attempted to change the fragmentation method to a more scalable option (section 2.4.4). And finally, I investigated alternative purification methods to column purification that would allow for samples to be processed in 96-well format (section 2.4.5).

2.4.2 DNA Fragmentation by Sonication

DNA samples were sheared by subjecting them to brief periods of sonication in a bioruptor (UCD-200 Diagenode). For each sample 1µg in 150µl of DNA was added to a centrifuge tube and sonicated, alternating between 30 seconds at a high level and 30 seconds off, for 30 mins. Ice was added to the bioruptor initially and every ten minutes thereafter, to maintain a cool temperature.

2.4.3 Ligation Test

A very costly element of the sample preparation protocol is ligation of the indexed adapters to DNA fragments. In the existing protocol, adapter ligation is achieved using quick ligase (New England Biolabs (NEB); 2000U/µl) as per Table 2.2. Each reaction contains a total of 5000 units of DNA ligase. A more cost efficient enzyme, T4 DNA ligase (NEB; 400U/µl) was tested as a potential alternative. This enzyme was tested in two different ways. The standard sample prep protocol was applied to three identical DNA samples. One sample underwent ligation by the standard method (Table 2.3). The other two samples underwent modified ligation reactions. One was set up with 25 units of T4 DNA ligase in the ligation reaction and an incubation of 20°C for 2 hrs followed by 16°C for 16 hours as recommended in a study by Craig et al. (Table 2.3) (Craig et al. 2008). The other was set up with 1000 units of T4 DNA ligase in the reaction and incubated for 2hrs at 20°C as recommended in the NEB supplied protocol for the enzyme (Table 2.3). Each sample was quantified once the full preparation protocol was complete. The PCR selectively amplifies adapter ligated fragments, thus the quantity of DNA post-PCR indicates the efficiency of adapter ligation.

Table 2.2: Quick ligase reaction protocol	
Reaction components	
DNA sample	10µl
DNA ligase 2X buffer	25µl
Adapter oligo Mix	2µl
Quick DNA ligase (1000U/µl)	5µl (5000U)
dH ² O	8µl
Total	50µl
Incubate for 15 mins at 20°C	

Table 2.3: T4 DNA ligase reaction components		
Reaction	1	2
DNA sample	10µl	10µl
DNA ligase buffer 2X	25µl	25µl
Adapter oligo Mix	2µl (for 1µg DNA)	2µl (for 1µg DNA)
T4 DNA ligase (400U/µl)	-	2.5µl (1000U)
T4 DNA ligase (400U/µl) 1:100 dilution (4U/µl)	6.25µl (25U)	-
dH²O	6.75µl	10.5µl
Total	50µl	50µl
Incubation	20°C for 2 hrs followed by 16°C for 16 hours*	2hr at 20°C**
*as per Craig et al. protocol		
**as per recommended protocol for T4 DNA ligase		

2.4.4 DNA Fragmentation using dsDNA Fragmentase

An enzymatic method of DNA fragmentation; NEBNext dsDNAFragmentase (New England Biolab) was tested and compared to the existing method of sonication. NEBNext dsDNA Fragmentase generates dsDNA breaks in a time-dependent manner. It contains two enzymes, one randomly generates nicks on dsDNA and the other recognizes the nicked site and cuts the opposite DNA strand across from the nick, producing dsDNA breaks. The resulting DNA fragments contain short overhangs, 5'-phosphates, and 3'-hydroxyl groups. Genomic DNA was quantified using a Qubit fluorometer and 1µg DNA added to each reaction. The following steps are undertaken to set up the dsDNA fragmentase reaction:

1. Four reactions are set up as per Table 2.4.
2. Add 5µl 0.5M EDTA (Ethylenediaminetetraacetic acid; Applied Biosystems) to each sample.
3. Take 10µl of each DNA sample and add 2µl 6X loading dye to each sample.
4. Prepare 50ml of a 2% agarose gel (Sigma Aldrich) with distilled water and TAE.
5. Add 3.5µl ethidium bromide (EtBr) after the TAE-agarose has cooled.
6. Add 1.5µl low molecular weight ladder (LMW) and the DNA samples to separate wells of the gel.
7. Run for 40 mins at 100 Volts.

Table 2.4: Test of dsDNA fragmentase incubation period				
Reaction components				
Reaction	1	2	3	4
DNA	15.8µl	15.8µl	15.8µl	15.8µl
Fragmentase Reaction Buffer	2µl	2µl	2µl	2µl
Fragmentase	-	2µl	2µl	2µl
BSA	0.2µl	0.2µl	0.2µl	0.2µl
dH ₂ O	2µl	-	-	-
Total	20µl	20µl	20µl	20µl
Incubation	37°C 20mins	0mins	37°C 20mins	37°C 30mins

To compare the efficiency of fragmenting samples by fragmentase to sonication, the full sample preparation protocol was applied to four samples derived from the same genomic DNA sample as per Illumina’s protocol: two samples underwent fragmentation by sonication and two with fragmentase. This was carried out according to the standard Illumina protocol with sonication performed as above (section 2.4.2) and purification by column purification (section 2.4.5.1). Samples were quantified after each step to compare the DNA yield of the two methods.

2.4.5 Purification Method Comparison

A major component of sample preparation is sample purification which is achieved using Qiagen column purification in the existing protocol. Two magnetic bead-based purification methods were tested as potentially cheaper and more time efficient methods of purification; Chemagen Chemagic PCR Pure magnetic beads and Agencourt Ampure XP beads. These purification methods can be carried out in 96-well format. The magnetic beads in binding buffer solution bind to DNA in the wells and by placing the 96-well plate into a magnetic plate, a series of washes allows removal of all but the magnetic beads bound to DNA. Upon removing the plate from the magnet and adding elution buffer (EB, Qiagen) the DNA separates from the beads and by returning the plate to the magnet the DNA can be removed leaving behind the beads. To compare the efficiency of the three purification kits, the full sample preparation protocol was applied to three samples derived from the same genomic DNA sample, each being purified with a separate kit as per manufacturer’s recommended protocols (section 2.4.5.1 - 2.4.5.3).

2.4.5.1 Qiagen Column Purification

The following steps describe purification using Qiagen's spin column purification kits:

1. Add DNA volume X 3 of buffer QG to each sample.
2. Add to spin column and centrifuge for 1 min at 14,000g.
3. Discard the flow-through and place the column back in the same collection tube.
4. Add 750µl of Buffer PE (ensure ethanol has been added) to the column, leave to stand for 1-2 mins and centrifuge for 1 min at 14,000g.
5. Discard the flow-through and centrifuge for an additional 1 min at 14,000g.
6. Place the column in a new labelled centrifuge tube.
7. Let air dry $\leq$ 5mins with the lid open.
8. Add 30µl EB and leave to stand for 2 mins.
9. Centrifuge for 1 min at 14,000g.
10. Remove samples to a 96 well plate.

2.4.5.2 Ampure XP Purification

The following steps describe purification using Ampure XP magnetic beads:

1. Shake the Agencourt AMPure XP bottle to resuspend any magnetic particles that may have settled.
2. Add sample volume X 1.8 of Agencourt AMPure XP to each reaction and vortex gently for 30 secs.
3. Incubate at room temperature for 5 mins.
4. Place the reaction plate onto an Agencourt SPRI Super Magnetic Plate for 2 mins to separate beads from solution.
5. Aspirate the supernatant from the reaction plate and discard.
6. Dispense 200µl of fresh 70% ethanol into each well and incubate at room temperature for at least 30 secs.
7. Aspirate out the ethanol and discard.
8. Allow the ethanol to evaporate by placing the plate on a 37°C heat block for $\leq$ 5 mins. Be careful not to allow beads to overdry (bead ring appears cracked).
9. Repeat for a total of two washes.

10. Add required volume of elution buffer and vortex gently for 30 secs.
11. Incubate at room temperature for 2 mins.
12. Spin plate using a plate centrifuge up to 2000g.
13. Place the reaction plate onto an Agencourt SPRI Super Magnetic Plate for 1 min to separate beads from solution.
14. Transfer purified product to a new plate.

2.4.5.3 Chemagic Magnetic Bead Purification

The following steps describe purification using Chemagen's Chemagic magnetic beads:

1. Add 6µl Magnetic Beads and 79µl Binding Buffer 1 per 25µl DNA sample volume in a 96 well plate and mix by pipetting 4-5 times.
2. Incubate at room temperature for 5 mins.
3. Place the plate onto a magnet for 30 secs.
4. Aspirate the supernatant from the reaction plate and discard.
5. Remove the plate from the magnet and dispense 300µl of Wash Buffer 2 to each sample and resuspend the sample in the wash buffer by pipette mixing.
6. Place the plate onto the magnet for 30 secs and discard the supernatant.
7. Remove the plate from the magnet and repeat steps 5-6 twice using Wash Buffer 3.
8. After removing all traces of Wash Buffer 3, leave the plate on the magnet for ten mins with the lids open.
9. Add suitable volume of Elution Buffer 4 to the tube and resuspend by pipetting.
10. Incubate for 5 mins.
11. Place the reaction plate onto a Magnet Plate for 1 min to separate beads from solution.
12. Transfer purified product to a new plate.

2.4.6 Indexed Adapter Annealing

Indexed adapters were custom made by Dr. Elaine Kenny and ordered from Integrated DNA Technologies (IDT; <http://eu.idtdna.com/site>). They contained the original Illumina adapter sequence along with a 6bp index with built-in

redundancy for error correction (as described in section 1.8.2). A total of 42 6bp indexed adapters were ordered from IDT and were provided as single stranded oligos which require annealing. Index sequences can be found in Table F in the Appendix. For each adapter set, 10µl of each 100µM adapter were combined in one tube and put in a thermocycler (Peltier PTC-225 Tetrad, MJ Research) with conditions as per Table 2.5. After annealing, 80µl of 1x ligase buffer (40µl H₂O + 40µl 2x ligase buffer) are added to attain a 10µM adapter solution.

Table 2.5: Protocol for annealing of indexed adapters		
Step	Temperature	Time
1	95°C	2 mins
2	94°C	1 min
	Go to 2. and repeat -1°C/cycle for 70 cycles	
3	4°C	Hold

2.5 Sample Preparation Protocol

The following protocol describes preparation of 24 DNA samples in parallel. DNA sample stocks are quantified using the Qubit flourometer (Invitrogen). For each set of samples 15.8µl of 1µg genomic DNA is placed into assigned wells of a 96-well plate ready for the fragmentation reaction. These are stored at -20°C. The proceeding protocol describes reaction components for one reaction and for a master mix with enough reagents for 27 reactions to ensure adequate volumes for 24 samples.

2.5.1 Fragmentation of DNA with dsDNA Fragmentase

The following steps describe the set-up of the DNA fragmentation with dsDNA fragmentase.

1. Prepare a master mix as per table 2.6.
2. Add 2.2µl master mix to each well individually and incubate on ice for 5 mins.
3. Add 2µl dsDNA fragmentase to each well and vortex for 10 secs.
4. Incubate at 37°C for 20 mins.
5. Add 5µl of 0.5M EDTA to each reaction to stop activity of the enzyme.
6. Purify samples using Ampure XP and elute in 50µl EB.

- Run 3µl of each sample on a 2% agarose gel for 40 mins at 90 Volts to ensure adequate fragmentation.

Table: 2.6: Protocol for fragmentation of DNA with dsDNA fragmentase		
Reaction components		
	X1	Master Mix (X27)
Fragmentase RB	2µl	54µl
BSA	0.2µl	5.4µl
Total	2.2µL	59.4µl

2.5.2 End Repair

The end repair reaction converts the overhangs resulting from fragmentation into blunt ends, using T4 DNA polymerase and DNA polymerase I Klenow fragment. The 3' to 5' exonuclease activity of these enzymes removes 3' overhangs and the polymerase activity fills in the 5' overhangs. The following steps describe the set-up of the end repair reaction.

- Prepare a master mix as per Table 2.7.
- Add 53µl master mix to each well containing 47µl DNA.
- Incubate at 20°C for 30 mins on a thermocycler.
- Purify samples using Ampure XP and elute in 42µl EB.

Table 2.7: Protocol for end repair		
Reaction components		
	X1	Master Mix (X27)
DNA	47µl	-
NEBNext End Repair Reaction Buffer (10X)	10µl	270µl
NEBNext End Repair Enzyme Mix	5µl	135µl
Sterile H₂O	37.5µl	1012.5µl
E.coli DNA ligase for fragmentase	0.5µl	13.5µl
Total	100µl	1458µl

2.5.3 Add 'A' base

An 'A' base that allows ligation to the 'T' overhang of the adapters is added to the 3' end of the blunt phosphorylated DNA fragments, using the polymerase activity of Klenow fragment (3' to 5' exo minus). The following steps are undertaken:

1. Prepare a master mix as per Table 2.8.
2. Add 8µl to each well containing 33µl DNA.
3. Incubate at 37°C for 30 mins.
4. Samples are purified using Ampure XP and eluted in 42µl EB.

Table 2.8: Protocol for A' base addition		
Reaction components		
	X1	Master Mix (X27)
DNA	42µl	-
NEBNext dA-Tailing Reaction Buffer (10X)	5µl	135µl
Klenow Fragment (3'→5' exo)	3µl	81µl
Total	50µl	216µl

2.5.4 Quick Ligation

Adapters are ligated to the 'A' base overhang of fragments via a 'T' base overhang at their 3' end. The following steps are undertaken:

1. Prepare a master mix as per Table 2.9.
2. Add 15µl to each well containing 42µl DNA.
3. Add 2µl of separate indexed adapters to each sample.
4. Incubate at 20°C for 30mins.
5. Purify samples using Ampure XP and elute in 25µl EB.

Table 2.9: Protocol for quick ligation of adapters		
Reaction components		
	X1	Master Mix (X27)
DNA	33µl	-
Quick Ligase Reaction Buffer (5X)	10µl	270
Quick T4 DNA Ligase (2000U/µl)	5µl	135
Adapter	2µl	
Total	50µl	405

2.5.5 Size Selection

Samples undergo gel excision to select a size-range of templates for sequencing on the Illumina platform. The following steps describe gel excision with Qiagen's Qiaquick gel extraction kit:

1. Prepare 300ml of 2% Certified Low Range Ultra Agarose (Alpha Technologies) gel with distilled water and TAE.
2. Add 10µl ethidium bromide (EtBr) after the TAE-agarose has cooled.
3. Add 9µl 6X loading dye to each sample.
4. Load 20µl LMW (Low Molecular Weight) marker to the first and last well of each row.
5. Load 17µl sample in two adjacent lanes leaving a two lane gap between samples
6. Run @ 250 volts for 1 hr.
7. Place gel on blue light box, using a p1000 pipette and gel excision tips excise bands of 300-350bp and at 400bp for each sample and place in labelled centrifuge tubes.
8. Store the 400bp slice, un-purified at -20°C

2.5.6 Qiagen Gel Purification

The following steps describe purification of the gel excised slices using Qiaquick gel extraction kit (Qiagen):

1. Add 450µl Buffer QG to each sample and vortex until the agarose is fully dissolved.
2. Add 150µl isopropanol to sample and mix by inverting several times.
3. Add to spin columns (~650µl) and centrifuge for 1 min at 14,000g.
4. Discard the flow-through and place the column back in the same collection tube.
5. Add 500µl of Buffer QG to the spin column and centrifuge for 1 min at 14,000g.
6. Add 750µl of Buffer PE (ensure ethanol has been added) to the column, leave to stand for 1-2mins and centrifuge for 1 min at 14,000g.
7. Discard the flow-through and centrifuge for an additional 1 min at 14,000g.
8. Place the column in a new labelled centrifuge tube.
9. Let air dry ≤ 5 mins with the lid open.
10. Add 30µl EB and leave to stand for 2 mins.
11. Centrifuge for 1 min at 14,000g.
12. Remove samples to a 96 well plate.

2.5.7 Sample Preparation PCR Enrichment

PCR is performed to selectively enrich for DNA fragments that have adapter molecules on both ends, and to amplify the amount of DNA in the library. The PCR is performed with two primers that anneal to the ends of the adapters. The following steps describe the set-up of the PCR reaction.

- 1. Prepare a master mix as per Table 2.10.
- 2. Add 23µl of the purified gel excised samples to a 96-well plate.
- 3. Add 27µl of the master mix reagent in Table 2.10
- 4. Run PCR programme as per Table 2.11.
- 5. Purify samples using Ampure XP and elute in 25µl EB.

Table 2.10: Sample preparation PCR		
Reaction components		
	X1	Master Mix (X27)
DNA	23µl	-
Phusion DNA polymerase	25µl	675µl
Primer 1.0 (25µM)	1µl	27µl
Primer 2.0 (25µM)	1µl	27µl
Total	27µl	729ul

Table 2.11: Sample preparation PCR programme		
Step	Temperature	Time
1	98°C	30 secs
2	98°C	10 secs
3	65°C	30 secs
4	72°C	30 secs
Repeat step 2 through 4 for a total of 7 times.		
5	72°C	5 mins
6	4°C	Hold

2.5.8 DNA Library Quality Assessment by qPCR

Following sequencing prep individual DNA samples were assessed by real time quantitative PCR (qPCR) on an ABI Prism 7900HT Sequence Detection System (Applied Biosystems) using a FAM detector and analyzed with SDS2.2.2 Absolute Quantification (standard curve) software. Each DNA sample was diluted 1 in 500 using distilled H₂O (dH₂O) and a master mix prepared as per Table 2.12 in triplicate. Standards at 1 pmol (picomole), 10 pmol and 100 pmol were also prepared in triplicate along with a negative control. To

separate wells, 2.5µL diluted DNA, standard or dH₂O was added. The same primers used in the sample prep protocol are used. The probe was previously designed and supplied by Applied Biosystems with a FAM detector.

Table 2.12: DNA library assessment by qPCR	
Reaction components	
	X1
Diluted DNA library/ H ₂ O/ Standard	2.5µl
2X Taqman Universal Master Mix (Applied Biosystems)	12.5µl
Probe	2.5µl
Primer 1.0 (25µM)	2.5µl
Primer 2.0 (25µM)	2.5µl
Distilled H ₂ O	2.5µl
Total	25µl

Table 2.13: qPCR protocol to check sample quality		
Step	Temperature	Time
1	98°C	10 mins
2	95°C	10 secs
3	95°C	15 secs
4	72°C	32 secs
Repeat step 2 through 4 for a total of 40 times.		
5	72°C	5 mins
6	4°C	Hold

2.5.9 SureSelect Target Enrichment

The total DNA library input requirement for target enrichment is 500ng equivalent to 3.4µl of 147ng/µl library. DNA samples were combined in pools in equal quantities and calculations were based on a 550ng requirement to account for potential DNA loss whilst pipetting. No two samples with the same index were added to the same pool and cases and controls were pooled separately where possible. Following pooling, samples were evaporated either at room temperature or in a speed vac vacuum concentrator (DNA 1100 Savant) at a maximum temperature of 42°C. Pools were reconstituted in 3.4µl of EB.

Individuals sample pools were enriched using one custom SureSelect enrichment reaction which was performed as per the manufacturer’s

instructions (v1.2 April 2009) with the exception of the PCR which was performed with 10µl DNA input and 5-10 cycles.

2.5.10 Target Enrichment Library Quality Assessment Using Bioanalyser

The quality, size and concentration of the target enriched libraries were assessed on Agilent's bioanalyser using the highly sensitive DNA assay kit as per manufacturer's protocol (Edition 05/2009). Target enriched libraries were stored at a stock concentration of 2-10nM ready for sequencing.

2.6 Next Generation Sequencing and Data Analysis

2.6.1 Variant Detection

A total of 18pmol (picomole) of target enriched libraries were sequenced on the Illumina Genome Analyzer II using 80bp paired end reads following the manufacturers protocol by Ciara Fahey. Once sequence data was generated, samples were separated according to index and aligned to a human reference genome using BWA (Burrows-Wheeler Aligner; (<http://bio-bwa.sourceforge.net/>)). Read calibration, realignment and SNV calling was carried out using GATK (Genome Analysis Toolkit; v1.0.5506) (DePristo et al. 2011). Functional annotation of variants was carried out using VEP (Variant Effect Predictor; http://www.ensembl.org/Homo_sapiens/UserData/UploadVariations). As these samples were previously part of various GWAS studies, it was possible to compare the GWAS SNV data with SNVs identified from the sequence data using GATK and determine the level of concordance. Various quality control measures were put in place as outlined in section 4.2. These tasks were performed by Dr. Paul Cormican and Dr. Elaine Kenny. I was provided with Plink format PED and MAP files for all single nucleotide variants. Plink v1.07 was used to determine summary data statistics and for data management throughout the project (<http://pngu.mgh.harvard.edu/~purcell/plink/>). The web-based data manipulation package Galaxy was also used for data management (<http://main.g2.bx.psu.edu/>). I was also provided with annotation of data for each SNV and indel along with a list of Loss-of-Function Variants (LoFs) deemed likely to affect protein function. A detailed description of such variants can be found in section 4.3.

2.6.2 Loss-of-Function Variant Validation

A total of 33 LoF variants were detected. As next generation sequencing can generate false positives, it is necessary to confirm identified variants that are likely to be followed up using an alternative method, in this case capillary sequencing. I custom designed primers to capture these variants using primer 3 (<http://frodo.wi.mit.edu/>). The PCR programme and conditions were optimised for each primer set, which were set up as per Table 2.14. Initially 3 programmes were tested with annealing temperatures of 56°C, 58°C and 60°C, using conditions detailed in Table 2.11. A 5µl aliquot of each PCR product was separated on a 2% agarose gel to ensure a single tight band of the correct size was produced. If the PCR was unsuccessful, 2X Failsafe optimisation buffers A-L (CAMBIO) were used instead of the Qiagen 10X buffer and tested with the programme that gave the closest result to one distinct band on a gel that was the correct size. Once appropriate conditions were determined, I performed PCR on DNA from the individual carrying the mutation as per Tables G and H in the appendix. Again, a 5µl aliquot of each PCR product was separated on a 2% agarose gel before proceeding to the next step to ensure that the PCR was successful and gave a single band of the correct size. The task of optimising PCR conditions and PCR of sample DNA was aided by Graham Kenny.

Table 2.14: PCR of loss-of-function variants for validation		
Reaction components		
	X1 with Qiagen Buffer	X1 with Failsafe buffers
DNA (10ng/µl)	3µl	2.9µl
10X buffer (containing 15mM MgCl ₂ :Qiagen)	1µl	-
Failsafe Buffer (Cambio;FS99060)	-	5µl
Forward Primer (25µM)	1µl	1µl
Reverse Primer (25µM)	1µl	1µl
dNTPs (10mM each)	1µl	-
Taq (5U/µl; Qiagen)	0.1µl	0.1µl
Total	10µl	10µl

After PCR of the region containing the variant, Exo1 (New England Biolabs) and Shrimp alkaline phosphatase (SAP, USB Corp) are used to remove excess dNTPs and primers respectively. This was performed using the reaction set-up detailed in Table 2.15.

Table 2.15: PCR product purification using ExoSAP-IT	
Reaction components	
PCR product	3µl
SAP	1µl
Exo 1	0.2µl
Dilution Buffer	1µl
H ₂ O	2µl
37°C for 1 hour followed by 80°C for 20 mins	

Purified PCR products were quantified by Qubit and two sequencing reactions were prepared, separately using forward and reverse primers, as per Table 2.16 using the ABI BigDye Terminator v3.1 Cycle Sequencing kit (Applied Biosystems). These reactions were performed on a thermocycler as detailed in Table 2.17. In the DNA sequencing reaction, fluorescently labelled dideoxynTPs (ddNTPs) that prevent addition of any more dNTPs are used in addition to standard dNTPs used in PCR. This causes the extension of the primer to terminate randomly as a ddNTP is incorporated into the sequence, creating a different length fragment for each base in the sequence.

Table 2.16: BigDye® Terminator v1.1 sequencing preparation	
Reaction components	
Purified DNA	10ng
BigDye® Terminator v13.1 Sequencing Buffer (5X)	2µl
Big Dye	3µl
5µM primer	1µl
H ₂ O	Up to 10µl

Table 2.17: BigDye® Terminator v1.1 sequencing reaction		
Step	Temperature	Time
1	96°C	1 min
2	96°C	10 secs
3	50°C	10 secs
4	60°C	4 mins
Repeat step 2 through 4 for a total of 25 times.		
5	72°C	5 mins
6	4°C	Hold

This sequence product is purified using a BigDye XTerminator Purification Kit (Applied Biosystems) which contains an XTerminator Solution that scavenges unincorporated dye terminators and frees salts from the post-sequencing reaction and SAM solution which enhances the performance of the XTerminator Solution and stabilizes the post-purification reactions. This is set up as per Table 2.18. DNA samples were then sequenced on an ABI 3130XL Genetic Analyzer (Applied Biosystems). Resulting electropherograms were visualised using BioEdit v7.1.3 (<http://www.mbio.ncsu.edu/bioedit/bioedit.html>).

Table 2.18: SAM X-Term purification of Big Dye sequencing reaction product	
Reaction components	
Sequencing Reaction Product	10µl
X-Term	10µl
SAM	45µl
Vortex gently for 30 mins	

I began sequencing samples as described on a local 3130xl as described above but to scale up speed and reduce project costs, I outsourced the remainder of the sequencing to Source Bioscience (<http://www.sourcebioscience.com/>). For each sample 1ng/µl of PCR product per 100bp fragment was sent to Source Bioscience for PCR purification and sequencing. Parent DNA was available for the Autism samples so these samples were also sequenced for the particular variant carried by the proband in order to determine parent-of-origin or if the variants were *de novo*.

2.6.3 Determination of Parent-of-Origin of *De Novo* Variants

Two *de novo* variants were detected in ASD cases, one in sample AS084C1 at chr1:232144803 (*DISC1*) and the other in sample AS155C1 at chr12:13724778 (*GRIN2B*). The approach undertaken to determine parental origin of the *de novo* variants was allele-specific PCR. A heterozygous SNV must be identified in the proband that is near enough to the *de novo* SNV to allow both variants to be amplified in a single PCR product. This neighbouring SNV must also have different genotypes in the parental samples. When designing the PCR, one of the variants is positioned at the 3' end of the forward primer and two versions of the primer are designed, each containing a different allele of the variant. The basis for this method is that a mismatch at the 3' end of a primer leads to an insufficient elongation during PCR by Taq polymerase. A primer that matches a particular variant allele at its 3' end, and mismatches to the other allele, results in amplification of only the strand containing the matching allele. In this way, it is possible to discriminate between parental haplotypes in the carrier and to determine which haplotype contains the *de novo* variant.

This first step was to identify a nearby heterozygous SNV in the case for which the parents have different genotypes. Using the data generated from the sequencing study I searched for SNVs within 15kb of the two *de novo* variants in the carriers; AS084C1 (*DISC1* chr1:232144803) and AS155C1 (*GRIN2B* chr12:13724778). This was achieved using Plink commands e.g.:

```
"plink --file AUTSZ215 --keep AS084C1.txt --from 232129803 --to 232159803  
--recode --out AS084C1_DISC1."
```

The `--file` command opens the PED and MAP files that contain all SNVs identified in the study that have passed QC filters. The `--keep` command excludes all samples except the *de novo* carrier identified in the file AS084C1.txt. The `--from` and `--to` commands select only SNVs within 15kb of the variant of interest based on the positions I selected. The `--recode` command generates new PED and MAP files that include SNVs in this 30kb window in only the carrier sample. The `--out` command names the PED and MAP files.

In the *DISC1* *de novo* mutation carrier, just two SNV were identified; (i) an A/T SNV at genomic position chr1: 232144598 with SNV ID rs821616, located 205bp from the variant, and (ii) a homozygous G SNV at chr1: 232144739 with SNV ID rs115112816, located 64bp from the variant. In *GRIN2B* a total of 15 SNVs were identified by this method, of which none were heterozygous in the case sample.

When previously validating the *de novo* variant, 260bp surrounding the mutation were sequenced in the *DISC1* proband and 301bp were sequenced in the *GRIN2B* proband. These regions are not limited to exonic sequence, as was the case for the sequencing study data, so could inform on SNVs that the sequencing study did not detect. However, upon inspecting the resulting electropherograms no SNVs were discovered.

The *GRIN2B* *de novo* case and parents were included in a previous GWAS study, which again are not limited to exonic sequences so could inform on SNVs that this study did not capture. I attained information on any SNVs +/- 20kb of the variant from Dr. Richard Anney in my research group. A total of 36 SNVs were genotyped within this region and 9 of these were heterozygous in the carrier with different genotypes in the parents. The closest variant to the *de novo* variant was at position chr1: 13723127, located 1651bp away. This distance is too far for a single PCR fragment that can be sequenced because the size limit with Sanger sequencing is approximately 1100bp.

No informative SNV had been identified for *GRIN2B* and for *DISC1* a heterozygous SNV within close proximity of the *DISC1* *de novo* variant was identified but sequencing was required to determine the genotypes of the parents at this site. Thus, I proceeded to sequence upstream and downstream of the *de novo* variants as far as possible in the carrier and parents in order to identify potentially informative SNVs. This would inform of any SNVs within this distance of the variants, not just limited to exons as in our study or SNVs on GWAS arrays. I designed primers that would amplify an 1100bp region upstream and downstream of the *de novo* variant with 60bp overlapping with the fragment previously sequenced during validation. Primers were designed and PCR performed as previously described and primer sequences and PCR programmes are located in Table I in the Appendix.

After detection of a heterozygous SNV in the proband with different genotypes in the parents, I designed the allele-specific PCR primers as per Table 2.19. PCR programmes are located in Table J in the Appendix. For *GRIN2B*, allele-specific PCR was carried out using the rs7312845 SNV (C/A) for which the father was heterozygous and the mother was homozygous for the C allele. Primer design was difficult due to high GC content in this region and the lack of flexibility for primer positioning. Thus, primers were designed such that the *de novo* G/A variant was at the 3' end of the forward primer and thus would be the site of allele specificity.

For *DISC1*, allele-specific PCR was carried out using rs821616 (A/T) for which the father was heterozygous and the mother was homozygous for the A allele. I designed primers as per Table 2.19 and PCR programmes are located in Table J in the Appendix. Allelic discrimination can be improved by incorporating a mismatch 1bp upstream of the allele-specific base at the 3' end of the forward primer (Kana et al. 2012). The allele-specific bases at the 3' end of the forward primers are marked in bold and the mismatches are in red. In some cases the primers were extended in order to increase or decrease the GC content.

Table 2.19: Allele-specific primers	
Primer	Sequence
<i>GRIN2B</i> FWD primer for G allele	GCATCATCTACACCCCTCT G
<i>GRIN2B</i> FWD primer for G allele including mismatch	GCATCATCTACACCCCTC AG
<i>GRIN2B</i> FWD primer for A allele	AATGCATCATCTACACCCCTCT A
<i>GRIN2B</i> FWD primer for A allele including mismatch	GCATCATCTACACCCCTC CA
<i>GRIN2B</i> REV primer	AAAGGTGACTAGCCCACTGC
<i>DISC1</i> FWD primer for A allele	GCTTGTCGATTGCTTATCCAG A
<i>DISC1</i> FWD primer for A allele including mismatch	GCTTGTCGATTGCTTATCCAT A
<i>DISC1</i> FWD primer for T allele	GCTTGTCGATTGCTTATCCAG T
<i>DISC1</i> FWD primer for T allele including mismatch	GCTTGTCGATTGCTTATCCAT TT
<i>DISC1</i> REV primer	AAATATGCCAAAAGTTGGGTTTT

2.6.4 *In Silico* Analysis of Variants

2.6.4.1 *In Silico* Analysis of LOF Variants

For each LoF variant I was provided with a list chromosomal positions and the reference and alternate variant calls. I input this information into the variant effect predictor (http://useast.ensembl.org/Homo_sapiens/UserData/UploadVariations) to determine the impact of the variant on all protein coding transcripts of the gene carrying the variant. Of the transcripts that the variant caused truncation or a frameshift, I selected the one corresponding to the largest protein as determined using Ensembl (www.ensembl.org/). I used UCSC genome browser and Ensembl to catalogue the impact of each LoF variant. The protein domain structure was determined in Pfam (pfam.sanger.ac.uk/) using the Uniprot ID (<http://www.uniprot.org/>) corresponding to the transcript selected for impact analysis (pfam.sanger.ac.uk; http://www.uniprot.org).

2.6.4.2 *In Silico* Analysis of Missense Variants

I searched for tools to functionally annotate sequence variants using Pubmed (www.ncbi.nlm.nih.gov/pubmed) and Google (www.google.com). I assessed each tool identified based on ease of use (e.g. web-based, download required, Linux or Perl required), input required (genomic coordinates or protein information), year released, number of citations, usefulness of output and underlying method employed.

2.6.4.2.1 Analysis of scores using SPSS

The SPSS statistical software (Version 19.0, SPSS Science, Chicago, IL) was used to analyze the level of correlation between scores generated from different annotation methods. More detail is provided in section 4.5.8.

2.7 *In vitro* Investigation of Loss-of-Function Variants

Many LoF variants were detected and the impact of these variants requires functional investigation. As it is not feasible to investigate all variants, just one was selected; a nonsense mutation in *FYN* at chr6: 112025283 in a SZ sample. An explanation for selection this particular variant is provided in section 5.2. The approach to investigate this variant is to design plasmids

containing green fluorescent protein (GFP)-tagged wildtype and mutant versions of the gene and transfect them into primary cortical neurons from newborn mice. Immunohistochemistry and confocal microscopy could then be employed to determine the localisation of the wildtype and mutant proteins and their affect on neurite structure and number, and on synaptic function. A detailed explanation of LoF variant selection and of plasmid design can be found in chapter 5.

2.7.1 Transformation of Plasmids into Component *E.coli* Cells

As limited plasmid stocks are provided, it is necessary to increase the amount of plasmid using a technique known as transformation, which allows the introduction of foreign DNA into bacterial cells. Plasmids were transformed into competent *E.coli* cells (JM109 Promega), using a 'heat-shock' protocol as outlined below. The plasmid is mixed with the competent cells, cooled on ice and then subjected to a 'heat shock', which disrupts the cell surface allowing entry of the plasmid. Not all bacterial cells will be successfully transformed thus a selection marker is required. In this case the plasmids contain an ampicillin gene so only successfully transformed cells will grow in media or agar containing carbenicillin (ampicillin confers resistance to carbenicillin). The transformed cells are allowed to multiply and the plasmids are then purified from the cells. This is achieved as outlined below.

1. Label and chill eppendorf tubes (1 per insert plasmid, one positive control and one negative control).
2. Warm the S.O.C. medium (Super Optimal broth with Catabolite repression Media; section 2.7.6.1) to 37°C in a water bath and set another water bath to 42°C.
3. Remove JM109 cells from the -80°C freezer and thaw on ice for 5 mins: enough for 50µl per plasmid insert transformation and 25µl for each of the positive and negative control.
4. Once thawed, pipette 50µl into the plasmid insert tubes and 25µl into the control tubes.
5. Make a 15ng/µl dilution of each plasmid with nuclease free water in new eppendorf tubes.

6. Add 1µl of the plasmid dilutions to the corresponding tubes containing 50µl JM109 cells. For the positive control add 1µl of the control plasmid provided with the JM109 cells and for the negative control add 1µl H₂O.
7. Heat shock the cells for 45 secs at 42°C in a water bath. Immediately place the tubes on ice for 2 mins.
8. Add 400µl warm (37°C) S.O.C. medium to each tube (prepared as per section 2.7.6.1).
9. Incubate for 1hr at 37°C in a water bath. Allow LB agar plates (prepared as as per section 2.7.6.2) to reach room temperature.
10. For the insert plasmids inoculate plates with 200µl, 100µl, 50µl and 25µl of the 400µl transformation media, spreading evenly across the plate with a plate spreader (this number of dilutions is performed in the initial transformation to determine the dilution that gives an adequate number of well-spaced colonies, just two dilution are used in future transformations).
11. For the positive and negative control inoculate one agar plate each with 50µl of the 400µl transformation media. Discard the remaining media.
12. Incubate the plates in a 37°C oven overnight.

2.7.1.1 Post Transformation

Presence of colonies on the agar plates indicates successful transformation of plasmids into the bacterial cells. These colonies are selected and grown up in media as below.

1. Warm S.O.C media to 37°C in a water bath.
2. Ensure the positive control plate has colony growth and the negative control does not.
3. Add 200µl warm S.O.C media to six wells per plasmid of a 24-well cell culture plate (Nunclon. Fisher Scientific TKT-190-010Y).
4. Select four transformants and add to the wells of the culture plate using an inoculation loop.
5. Incubate at 150 rpm in a shaking incubator at 37°C for 8 hours.

2.7.1.2 Colony Screen

A colony screen must be performed to ensure that the bacteria cells do contain the plasmids and that the plasmids have the correct insert. Primers were

designed using primer3. The maximum amplicon size was set to 900bp and they overlapped by a minimum of 70bp to ensure near complete coverage of the insert sequence. As the plasmid sequence was not available it was necessary to design the primers as close as possible to the start and end of the insert sequence to achieve maximum sequence coverage. PCR was performed using the programme and annealing temperatures provided in Table K in the appendix along with the primer sequences. For *FYN* wildtype, 3 primer sets were designed, one of which amplified the entire shorter *FYN* mutant insert.

A PCR master mix was set up for each separate primer set as per Table 2.20, with adequate volumes for the number of colonies being screened. A pipette tip is dipped into the broth and then into the PCR reaction. This is sufficient as too much broth will impair the PCR reaction. PCR is then performed as per Table 2.21. For the wildtype plasmid, PCR reactions are set up for all 3 primer sets. For the mutant plasmid, a PCR reaction is set up for primer set 1 that covers the plasmid and also with primer set 2, which should not amplify. This is a control measure to ensure that the plasmid contains the mutant sequence.

Table 2.20: PCR colony screen	
Reaction components	
Transformed Plasmid	<0.5µl
10X buffer(containing 15mM MgCl ₂ :Qiagen)	1µl
Forward Primer (25µM)	1µl
Reverse Primer (25µM)	1µl
dNTPs (2.5mM of each dNTP: NEB)	1µl
Taq (5U/µl; Qiagen)	0.1µl
dH ² O	5.5µl
Total	10µl

Following PCR, the products are analysed on a gel to ensure that they have a band of the correct size for all 3 primers sets for the wildtype insert and just for primer set 1 for the mutant insert.

2.7.1.3 Growth of Positive Colonies

Once it is confirmed that the colonies contain the plasmids with the correct insert, they are grown up to attain greater levels of the plasmids. This is achieved as described below:

1. Warm S.O.C media to 37°C in a water bath.
2. Add 200ml warm media to a conical flask (2 per plasmid).
3. Select two positive cultures with the cleanest and brightest bands on the gel for each plasmid.
4. Take the 200µl media from the starter culture of these colonies and add to the labelled conical flasks. To the remaining positive cultures add 500µl media (to be stored as a back-up). Incubate all at 100-150 rpm in a shaking incubator at 37°C overnight.

2.7.1.4 Plasmid Purification

Once large quantities of transformed cells are attained, it is necessary to purify the plasmids from the cells. This is achieved using the Promega midi-prep kit. This system purifies plasmid DNA using a silica membrane column and a variety of wash buffers which remove protein, RNA and endotoxin contaminant from the plasmid DNA. This is performed with the following steps:

1. Remove the conical flasks from the shaking incubator.
2. Pour 200ml of culture to four 50ml falcon tubes.
3. Pellet cells at 4,000*g* for 10 mins. Discard the supernatant.
4. Resuspend one cell pellet thoroughly in 12ml of Cell Resuspension Solution by pipetting and transfer to new 50ml tubes and so on until all four pellets have been resuspended and combined in 12ml of Cell Resuspension Solution in one 50ml tube.
5. Add 12ml of Cell Lysis Solution. Invert gently 3–5 times to mix and incubate for 3 mins at room temperature.
6. Add 12ml of Neutralisation Solution. Invert gently 10–15 times to mix. Do not overmix as this may cause the centrifuged lysate to be cloudy which can cause clogging of the columns.
7. Centrifuge the lysate at 7,000*g* for 30 mins at room temperature using a fixed angle rotor. If the lysate is not sufficiently clear add to a fresh 50ml tube and centrifuge for another 5-10 mins.
8. Assemble the blue PureYield™ Clearing Column and white PureYield™ Maxi Binding Column in a stack, with the clearing column on top. Place this column stack on the vacuum manifold.
9. Pour one half of the lysate into the blue PureYield™ Clearing Column.

10. Apply maximum vacuum until the lysate has passed through both the clearing and binding columns. Then add the remaining lysate and maintain vacuum until the liquid has cleared both columns.
11. Slowly release the vacuum. Then remove and discard the blue PureYield™ Clearing Column, leaving the PureYield™ Maxi Binding Column on the vacuum manifold.
12. Add 5ml of Endotoxin Removal Wash to the PureYield Maxi Binding Column, apply a vacuum and allow the solution to be pulled through the column.
13. Add 20ml of Column Wash (ensure the recommended volume of ethanol has been pre-added) to the binding column, and allow the vacuum to pull the solution through the column.
14. Dry the membrane by applying a vacuum for 8-10 mins.
15. Remove the PureYield™ Maxi Binding Column from the vacuum manifold, and tap the tip of the column on a paper towel.
16. Place a 1.5ml microcentrifuge tube into the base of the Eluator™ Vacuum Elution Device.
17. Assemble the Eluator™ Vacuum Elution Device, and insert the DNA binding column into the device and when secure place the assembly onto a vacuum manifold.
18. Add 500µl of Nuclease-Free Water to the DNA binding membrane in the binding column. Wait for 1 min. Apply maximum vacuum for 1 min and remove the microcentrifuge tube.

2.7.1.5 Check Plasmid Quality

The quality of DNA in the purified plasmids must be checked by Nanodrop and by agarose gel electrophoresis:

1. Quantify the plasmids using the Nanodrop and check the 260/280 quality ratios.
2. Make a 1% 60ml agarose gel with 3µl EtBr.
3. Make a 10µl ~10ng/µl dilution of each purified plasmid. Add 2µl 6X loading dye and load to gel.
4. Add Hind III marker, which yields bands of 2 - 23kb.
5. Run for 40 mins at 100Volts.

2.7.1.6 Concentration of Plasmid DNA by Ethanol Precipitation

If the concentration of the purified plasmids is lower than that required for transfection, it is necessary to precipitate the DNA. This is achieved by ethanol precipitation as below:

1. Prepare 3M Sodium acetate (NaOAc) solution.
2. Add 11 μ l NaOAc for every 100 μ l DNA.
3. Add 2 X DNA volume of ice cold ethanol.
4. Precipitate for at least 30 mins at -20°C.
5. Centrifuge for 10 mins at 14,000g.
6. Pour off the ethanol and wash the sides of the tubes with ice cold 70% ethanol.
7. Centrifuge for 10 mins at 14,000g.
8. Pour off the ethanol and allow to air dry until all the ethanol has evaporated.
9. Resuspend in an adequate volume of TE.

2.7.1.7 Sequence Plasmid Inserts

Once it is determined that the plasmid is of adequate concentration and quality (based on 260/280 value being in the range of 1.8-2.2), the insert must be sequenced to ensure the correct insert is present and no errors were introduced during transformation. PCR products from the colony screen were sent to Source Bioscience for sequencing to ensure the plasmids contain the correct sequence. The resulting sequences were aligned to reference cDNA using T-Coffee (www.tcoffee.org/).

2.7.2 Preparation of Mouse Cell Cultures

Preparation of mouse cell primary cultures from mice primary cortical neurons was performed by Ines Molinos as below. All pipettes, tubes, tips, surgical tools and Petri dishes are UV irradiated for 20 mins before use.

1. Put cold DPBS (Dulbecco's Phosphate-Buffered Saline) (Invitrogen 14190-094) in Petri dishes.
2. Cut pup head into one Petri dish. Remove skin and clear as much blood as possible.

3. Take a new Petri dish with DPBS and place the head into to using a #7 forceps. Cut the skull longitudinally and make two cuts from the base of the skull to the eyes.
4. Remove the skull and using the curved #7 forceps remove the brain.
5. Carefully, remove the meninges using the #5 forceps tips.
6. Place the whole brain in a clean Petri dish with DPBS, remove the olfactory bulb and dissect the prefrontal cortex/visual cortex with the scalpel blade. Remove the brain and mince finely the prefrontal cortex using the tips of the forceps or a blade.
7. With 200µl DPBS pick up the pieces of brain, place them in an eppendorf tube and keep on ice.
8. Add 1µl DNase I and 20µl 10% warmed Trypsin-EDTA (Invitrogen R-001-100). Leave the samples in the incubator at 37°C for 10 mins.
9. Block the reaction by adding double the volume (440µl) of DMEM (Dulbecco's Modified Eagle Medium) media + 10 % FBS (Fetal Bovine Serum; section 2.7.6.5).
10. Dissociate the tissue by pipetting up and down no more than 15 times with a P1000 pipette tip. If there are too many large pieces, let them settle down for a while and pipette up and down no more than 10 times with a P200 pipette tip. Avoid bubbles.
11. Let the samples settle down for 5-10 mins. Transfer the supernatant (~550µl) to clean eppendorf tubes.
12. Centrifuge at 800 rpm for 5 mins at 10°C.
13. Remove as much supernatant as possible ($\geq 400\mu\text{l}$) without disturbing the pellet.
14. Resuspend the pellet in the same volume of DMEM + 10% FBS equal to the volume of supernatant removed.
15. Place 10µl per sample on a haemocytometer and count the cells under the microscope (10X).
16. Dilute the samples in DMEM + 10% FBS to get 1×10^6 cells/ml.
17. Put 100µl diluted sample on each coverslip (Sparks, MIC3366). Incubate for 2 hours in a 5% CO₂ incubator at 37°C.
18. Add 400µl warmed Neurobasal media and supplements and return to the incubator overnight.
19. The following day, remove 200µl media and add 200µl of 7µM feeding media (section 2.7.6.3), to make final concentration of 3.5µM feeding

media. Continue to change the media every two days with 3.5 μ M feeding media.

2.7.3 Transfection

Transfection is a technique that allows the introduction of foreign DNA into eukaryotic cells. Transfection of *FYN* wildtype, *FYN* mutant and *GFP* plasmids were performed. This was performed using the Amaxa nucleofactor system and the nucleofection kit for mice neurons (Amaxa® Mouse Neuron Nucleofactor Kit VAPG-1001) as below. As a control cells were also processed for transfection, but in absence of the plasmid.

1. Add 200 μ l feeding media to number of well required of a 24-well cell culture plate and equilibrate to 37°C in a 5% CO₂ incubator. Place a poly-lysine coated cover slide into each well.
2. Add the nucleofector solution to the nucleofector supplement and mix by pipetting. This can be stored for up to three months at 4°C. Equilibrate to room temperature before use.
3. For each transfection add 3 μ g plasmid into a sterile 1.5ml Eppendorf tube.
4. Once the number of mouse prefrontal cortex neuronal cells per sample is known, take the volume needed to have 6-7 10^6 cells.
5. Spin the cells at 80g for 5 mins. Carefully remove as much of the supernatant as possible.
6. Resuspend the pellet in 100 μ l of pre-warmed nucleofector solution/supplement mixture.
7. Transfer the neuronal suspension into the 1.5ml Eppendorf tube containing the plasmid and mix by gently flicking the tube for approximately 20 secs.
8. Transfer 100 μ l of cell/DNA mixture into a nucleofection cuvette and place it into the cuvette holder avoiding air bubbles.
9. Place into the Amaxa device and select programme 0-005 (specific for neurons).
10. Immediately after nucleofection, carefully add 300 μ l pre-warmed DMEM (Sigma D5796)-10%FBS (section 2.7.6.5) to minimize shearing forces when retrieving the cells from the cuvette.

11. Use a plastic pipette supplied in the nucleofection kit to collect the cells from the cuvette and seed them into an equilibrated 24 well plate containing feeding media.
12. Incubate in a 5% CO² incubator at 37°C.
13. After 2-4 hours carefully replace medium with 500µl fresh feeding media to remove cellular debris.
14. Gene expression is often detectable after 6-8 hours and can be observed up to 12-14 days after Nucleofection.
15. After 24 hours, remove half the volume of media and add the same volume as that removed of 7 µM AraC feeding media.
16. Change media once or twice a week (use 3.5 µM AraC Feeding Media).

2.7.4 Immunohistochemistry

Immunohistochemistry is the process of detecting a known antigen within cells or tissues by taking advantage of the specificity of antibody-antigen interactions. The antigen is the protein of interest, in this case *FYN* protein, and other biological markers. Fluorescently labelled antibodies can be produced that are specific to the antigen/protein of interest. In order to increase specificity, an indirect two-step process is employed. An unlabelled primary antibody binds to the antigen. Then a secondary fluorescently labelled antibody is employed which is specific to the primary antibody. Upon binding of the secondary antibody to the primary antibody, fluorescence at a particular wavelength is emitted, which can be detected by confocal microscopy. This allows determination of the distribution and localization of the wildtype and mutant versions of *FYN*, and also for markers of dendrites and of synaptic function. The level of non-specific background staining is low by this approach. Also, it is a very sensitive method as the signal is amplified due to the binding of many secondary antibodies to each primary antibody. A number of antibodies were used; *FYN* (marker of *FYN* protein), *GFP* (marker of *GFP* protein, an indicator of successful transfection), *MAP2* (microtubule associated protein 2; acts as a dendritic marker) and synapsin (associates with synaptic vesicles; acts as a synaptic marker). DAPI is also used which stains nuclei. Only four colours are detectable by confocal microscopy at one time, thus as DAPI (stains blue) and *GFP* (stains green) must be used in each slide to indicate the cellular location and type, and *GFP* must be used to indicate

transfection, a combination of any other two antibodies can be used provided the colours are different.

2.7.4.1 Primary Antibody Staining

1. Prepare heat inactivated goat serum (Sigma G9023) in 1ml aliquots by heating to 56°C for 30mins in a water bath, flicking every 5mins. Store unused aliquots at -20°C.
2. Wash each well of the 24-well culture plate containing transfected cells with PBS (Phosphate-Buffered Saline) 3 times with a 5 incubation at 250 rpm on a shaker in between each wash.
3. Block the non-specific binding by adding 400µl 10% Goat serum in 0.1% PBS-T (Phosphate-Buffered Saline-Triton; 2.7.6.6) to each well and incubate for 1hr at room temperature at 250 rpm.
4. Prepare dilutions of the primary antibodies, in 1% Goat serum in 0.1% PBS-T, in the ratio as specified beside below.
 - a. Rabbit anti-FYN antibody, Millipore, 06-133, 1:1000
 - b. Goat anti-GFP antibody, ABCAM, ab5450, 1:500
 - c. Mouse anti-MAP2 antibody, Millipore, MAB3418, 1:200
 - d. Guinea Pig anti-Synapsin, Millipore, 5454 106 004, 1:1000
5. Add 350µl to each well and incubate overnight.

2.7.4.2 Secondary Antibody Staining

Wash each well of the 24-well culture plate containing transfected cells with PBS 3 times with a 5 incubation at 250 rpm on a shaker in between each wash. Prepare dilutions of the secondary antibodies, in 1% Goat serum in 0.1% PBS-T, according to the ratio specified beside each one (a-e below corresponds to a-e in section 2.7.4.1):

- a. Goat anti-rabbit Cy3, Invitrogen A10520, 1:500.
- b. Donkey anti-goat Dylight 488, Jackson Immuno Research, 705-485-003, 1:500.
- c. Goat anti-mouse Dylight 649, Jackson Immuno Research, 115-495-003, 1:500.
- d. Goat anti-rabbit Dylight 488, Jackson Immuno Research, 115-485-003, 1:500.

- e. Goat antiguinea pig Cy3, Jackson Immuno Research, 106-165-003, 1:500
2. Add 350µl of the secondary antibody in 0.05% PBS-T to each well and incubate for 2 hrs, covered with aluminium foil.
3. Wash each well 3 times with PBS for 5 mins each with incubation at 250 rpm.
4. Mount the slides with Vectashield Hard set with DAPI (Vector H-1500). Place the slides in a slide tray and store at 4°C until the mounting media hardens. Once the mounting media is hard, seal the coverslips with nail polish.

2.7.5 Confocal Microscopy

A Zeiss LSM 510 laser module confocal microscope is used for analysis with the LSM 5 software (Zeiss). Image analysis is performed using Image J (rsb.info.nih.gov/ij/).

2.7.6 Recipes for Broths and Media

2.7.6.1 Super Optimal broth with Catabolite Repression (SOC) Media

To make 1 litre of SOC media add:

- 900ml distilled water
- 20g tryptone
- 5g yeast extract
- 10ml 1M NaCl
- 2.5ml 1M KCl
- 10ml 1M MgCl₂
- 10ml 2M Glucose
- Adjust to 1 litre using distilled water.
- Autoclave and allow to cool to ~55°C.
- Add 1ml of 100µg/ml Carbenicillin per litre of media and store at 4°C.

2.7.6.2 Lysogeny Broth (LB) Agar

To make 500ml of LB agar add:

- 450ml distilled water
- 5g tryptone

- 2.5g yeast extract
- 5g NaCl
- 7.5g Bacteriological agar

Adjust to 500ml with distilled water and autoclave. Allow to cool to ~55°C and add 1ml of 50µg/ml Carbenicillin per 500ml of media and pour into 10cm plates. Allow to cool and harden, then invert and store at 4°C.

2.7.6.3 Feeding Media

The media contains:

- 47.5 ml Neurobasal media (Invitrogen, 21103-049).
- 0.5 ml Penicillin-Streptomycin (Invitrogen, 15140-163).
- 0.5 ml L-Glutamine (Invitrogen, 25030-024).
- 0.5 ml 45% D-Glucose (Sigma, G8769).
- 1 ml B27 (Invitrogen, 17504-044).
- Media is filter sterilised

2.7.6.4 Ara-C Feeding Media

- Ara-C (5mM) stock is prepared with 12.1 mg Ara C (Sigma C1768) and 10 ml autoclaved H₂O and stored at -20°C.
- To make AraC feeding media, make feeding media as per section 2.7.6.3 and add 70µl AraC stock in 50ml Neurobasal media for 3.5µM AraC feeding media.
- Media is filter sterilised

2.7.6.5 DMEM/ 10% FBS Media

- Add 5ml FBS (Analab/PAA A14-101) to 45ml DMEM (Sigma D5796).
- Media is filter sterilised

2.7.6.6 PBS-Triton (PBS-T)

- Make 0.1% PBS-Triton X100 (PBS-T) by mixing 20ml 2% PBS-T and 30ml PBS (PBS Sigma P4417); Triton X100 Sigma T8787).
- Make 0.05% PBS-T by mixing 1.25ml 2% PBS-T and 48.75ml PBS.

3 Selection of Candidate Gene and Optimisation of Multiplex Target Enrichment Method Including Application to Test Pool of Samples

3.1 Aims

The aims of the research described in this chapter were to select candidate genes for sequencing, to design the target enrichment array for this set of genes and to optimise the multiplex target enrichment protocol and apply it to a test pool of samples. The provisional plan was to sequence the exons and flanking sequences of approximately 200 genes using pools of approximately 20 samples. This estimation was primarily based on a previous study carried out by my research group. They developed an advanced method of multiplexing DNA samples prior to target enrichment and next generation sequencing (Kenny et al. 2011). DNA samples are indexed using 6bp indexes or barcodes, pooled in equimolar quantities, enriched in a SureSelect in-solution target enrichment assay and sequenced in a single reaction. Multiplexing of 3 and 9 samples prior to target enrichment and sequencing was assessed. The amount of sequence data generated in this study and knowledge of the typical amount of data generated from an 80bp paired-end sequencing run, allowed estimation that if pooling approximately 20 DNA samples it would be possible to target approximately 1Mb of sequence and yield sufficient sequence coverage for the exons of approximately 200 genes.

The main focus of gene selection was on neurodevelopment and in particular on synaptic functioning and glutamatergic signalling. With over 20,000 genes in the human genome, it was necessary to implement a very systematic approach to initially select candidate genes and then to prioritise them in order to have just 200 genes for inclusion in the study. The next section of this chapter (section 3.2) describes this process in detail.

Preparation of DNA samples for sequencing is an expensive, time consuming and laborious task, especially when large sample numbers are involved. Therefore, a major component of this chapter deals with attempting to make this method more cost efficient and amenable to high sample throughput. This

work is described in section 3.4 and its application to a test pool of samples is described in section 3.5.

3.2 Gene Selection

To begin with, I put together an overview of the genes identified in recent CNV studies of ASD and SZ. I carried out a review of all previous publications that had identified CNVs in ASD and/or SZ. Using information obtained from each study, I compiled a list of genes affected by CNVs, which includes genes that have been shown to be affected by CNVs in cases exclusively or in statistically more cases than controls. In some studies the gene ID's were provided, in others only genomic coordinates were provided. In the latter case, I uploaded the coordinates to UCSC table browser (<http://genome.ucsc.edu/cgi-bin/hgTables>), selected refseq genes track and sent the output to Galaxy (<https://main.g2.bx.psu.edu/>). A file is then downloadable from Galaxy, containing a list of genes within the inputted genomic coordinates. This information was used during gene selection as outlined below.

Overall, I selected 215 genes with a potential role in SZ and/or ASD pathogenesis for inclusion in the study. Support for the role of each gene is based on prior evidence for involvement in neurodevelopment disorders and/or in normal synaptic functioning or glutamatergic signalling. These genes can be grouped into five categories that define how they were selected. Some genes are members of multiple categories but are listed based on method of first selection. A full list of all 215 genes along with their genomic location and reasons for selection can be found in Tables B-F in the appendix.

3.2.1 Neurexin (NRXN1, Interactors and Related Proteins)

Neurexins (*NRXN*) are a family of neuronal cell adhesion molecules (CAMs) that play a major role in synaptic function. They are predominantly presynaptic CAMs that bind to members of the postsynaptic family of CAMs; neuroligins (*NLGNs*), forming a trans-synaptic complex (Sudhof 2008; Betancur, Sakurai, and Buxbaum 2009). This complex serves as a platform for the recruitment of scaffold proteins (e.g. *PSD95* and *SHANK2*), cytoskeletal proteins and signalling molecules to the synapse, which are essential for the morphology, function and plasticity of synapses (Betancur, Sakurai, and

Buxbaum 2009). *NRXN1* mediates *NLGN*-induced presynaptic differentiation in contacting axons and induces postsynaptic differentiation in glutamatergic synapses and GABAergic neurons via interactions with specific members of the *NLGN* family (Geddes, Huang, and Newell 2011).

A wealth of *in vitro*, *in vivo* and association studies support a role for *NRXN1* in neurodevelopmental disorders (as reviewed in (Sudhof 2008; Rujescu et al. 2009)). Furthermore, multiple studies have reported that rare deletions at *NRXN1* are risk factors for SZ and ASD (Autism Genome Project et al. 2007; Kim et al. 2008; Kirov et al. 2008; Marshall et al. 2008; Vrijenhoek et al. 2008; Walsh et al. 2008; Glessner et al. 2009; Rujescu et al. 2009). At the time of gene selection, *NRXN1* was the only instance in which multiple studies have identified CNVs that directly impact just one gene in ASD and SZ (Autism Genome Project et al. 2007; Kirov et al. 2008; Marshall et al. 2008; Vrijenhoek et al. 2008; Walsh et al. 2008; Glessner et al. 2009; Rujescu et al. 2009). Thus, *NRXN1* is a risk gene for both ASD and SZ.

On this basis I included *NRXN1* in the selection and I compiled a list of its protein interactors and related proteins, which may also have a role in SZ and/or ASD pathogenesis. I identified these protein interactors using the online protein interaction databases; HPRD (<http://www.hprd.org/>), BioGRID (<http://thebiogrid.org/>), IntAct (<http://www.ebi.ac.uk/intact/>) and String (<http://string-db.org/>) and by an extensive literature search (Biederer and Sudhof 2000; Sheng and Kim 2000; Kirov et al. 2008; Mukherjee et al. 2008; Sudhof 2008; Konrad et al. 2009; Gill, Donohoe, and Corvin 2010). Upon entering the gene or protein ID into these databases a list of interactors are returned, along with information on the type of evidence for each interaction and whether it is direct or indirect. Default settings were used for each database with the exception of setting them to only returning results with experimental evidence. The total number of genes that encoding proteins that are related to or interact with *NRXN1* was 42. As outlined at the beginning of section 3.2, I compiled a list of genes affected by CNVs in ASD and/or SZ using data published in previous CNV studies. Of the 42 genes that encode *NRXN1* interactors or related proteins, 5 had been reported to be affected by CNVs; *APBA2* (Sebat et al. 2007; Kirov et al. 2008; Marshall et al. 2008), *ERBB4* (Walsh et al. 2008), *CHRNA7* (Marshall et al. 2008), *NLGN1* (Glessner

et al. 2009) and *SHANK3* (Sebat *et al.* 2007; Marshall *et al.* 2008). I also entered these 5 genes into the databases mentioned above to determine their protein interactors, which I classed as *NRXN1* interactors with two degrees of separation (2ds). A further 42 genes were identified, bringing the total in the category to 84. For each gene, I determined if it had a plausible role in synaptic functioning using the refseq description (<http://www.ncbi.nlm.nih.gov/gene/>) and any genes that did not, were removed from the selection. The final number of genes in this category is 46.

3.2.2 Disrupted in Schizophrenia 1 (DISC1) and Interactors

The identification of a balanced t(1:11) chromosomal translocation in a large Scottish family that segregated with a wide spectrum of psychiatric phenotypes identified *DISC1* (Disrupted in Schizophrenia 1) as an important gene for neuropsychiatric disorders (St Clair *et al.* 1990; Hennah *et al.* 2009). Subsequent work has established multiple roles for *DISC1* and its 'Interactome' in neurodevelopment and synapse regulation (Camargo *et al.* 2007 ; Brandon and Sawa 2011). It has been implicated in multiple cellular functions in both the adult and developing brain including; cAMP signalling, neuronal migration, axon guidance and neurite extension, axonal transport and neurogenesis (Chubb *et al.* 2008; Bradshaw *et al.* 2008). Therefore, using online interaction databases and an extensive *DISC1* literature review, I identified 116 genes that encode known interacting proteins of *DISC1* (Biederer and Sudhof 2000; Camargo *et al.* 2007; Bradshaw *et al.* 2008; Chubb *et al.* 2008; Song *et al.* 2008; Kim *et al.* 2009; Mao *et al.* 2009; Porteous and Millar 2009). These were shortlisted by only including genes with a plausible role in synaptic functioning and thus neurodevelopmental disorders, in the final selection (n = 23).

3.2.3 Post-Synaptic Glutamate Receptor Complexes

Glutamate is the most abundant excitatory neurotransmitter and it plays a critical role in synaptic plasticity and cognitive processes. Thus, disruption of glutamate function may contribute to neurodevelopmental disorders (Grant *et al.* 2005). A number of X-linked genes within the NMDA receptor complex are known to be involved in psychiatric disorders (Laumonnier, Cuthbert, and Grant 2007). Using meta-analysis Laumonnier *et al.* devised a comprehensive

list of proteins located in the post-synaptic proteome (PSP) and the distinct structural and functional complexes to which they belong, totalling 1,180 genes. Availing of this data I compiled a list of genes within three multi-protein post-synaptic glutamate receptor complexes: N-methyl-D-aspartic acid (NMDA), metabotropic glutamate receptor 5 (mGluR5) and α -amino-3-hydroxyl-5-methyl-4-isoxazole-propionate (AMPA). I selected genes within these receptor complexes that encode structural proteins: channels, receptors, membrane-associated guanylate kinases (MAGUKs)/adaptors/scaffolders, cytoskeletal molecules and CAMs (n = 58).

3.2.4 Neural Cell Adhesion Molecules

Cell adhesion molecules (CAMs) other than *NRXN1* also have an important role in synapse structure and function, cell signalling and axon extension. I used data from the post-synaptic proteome along with KEGG pathway annotation (<http://www.genome.jp/kegg/pathway.html>) to identify genes that encode neural CAMs for inclusion in the selection (n = 61).

3.2.5 Functional and Positional Candidates

Finally, I selected additional candidate genes based on recent GWAS data including *ZNF804A* and *NRGN* for SZ and *MACROD2* and *SEMA5A* for ASD (Williams, Norton, et al. 2011; Stefansson et al. 2009; Anney et al. 2010; Weiss et al. 2009). I also included candidate genes based on CNV data e.g. *FOXP2* and *CYFIP1*, and data from severe neurodevelopmental disorders (e.g. *MECP2*) (Amir et al. 1999) and data from neurodevelopmental biology (e.g. *SEMA6A*) (Runker et al. 2011) (n = 27).

Genes with high levels of structural variation in the population can cause problems in target enrichment and any sequence reads generated are difficult to align. Thus, to avoid including genes that may not produce usable sequence data, I attempted to identify genes that may have high levels of common structural variation. As outlined at the beginning of section 3.2, I compiled a list of genes affected by CNVs in studies of ASD and/or SZ. Using this list I identified any genes that were affected by CNVs in ≥ 5 individuals in at least 2 different studies. I referred the DGV database (projects.tcag.ca/variation/) to determine the amount of structural variation identified by previous studies of

control individuals in these genes. This was a crude yet simple way to reduce the number of genes assessed in DGV to those most likely to contain high levels of common structural variation in the general population because investigating 200+ genes in this way would have been very time consuming and impractical. I removed the *NRXN1* interactor *CHRNA7* and the cell adhesion molecule *CNTN4* due to many reports of structural variants, bringing the total number of genes included in the study to 215.

3.3 Target Enrichment Array Design

The array to target the exons of the 215 genes included in the study was designed using Agilent's online tool eArray. Using default settings with the exception of small exon expansion as outlined in section 2.2, it was possible to generate baits for 3,672 of 3,709 target exons (99%). The total number of baits was 14,091, which targeted 1,064,238bp of baited sequence of which 693,593bp was exonic sequence. The remaining regions were excluded to avoid extension into repeat masked regions as sequence reads within repetitive regions are very difficult to align accurately. The total occupancy of the array was 57,750 probes, so in order to reach occupancy the bait library was repeated 4 times in the array design.

3.4 Sample Preparation Optimisation

Preparation of DNA samples for sequencing is a time consuming and laborious task so steps were undertaken to make this procedure more amenable for a high throughput of samples as described in section 2.4.

3.4.1 Ligation Test

The quick ligase enzyme used in ligation of the indexed adapters comprises a large proportion of the sample preparation costs. As such, I attempted to reduce costs by testing an alternative to the quick ligase enzyme. This involved using a lower concentration enzyme with a longer incubation. I applied the sample prep protocol to four samples; samples 1 and 2 were fragmented by dsDNA fragmentase and samples 3 and 4 were fragmented by sonication (Table 3.1). I carried out quick ligation as per the sample prep

protocol on samples 1 and 3 and an overnight ligation with a lower concentration of enzyme on samples 2 and 4, as detailed in section 2.4.3.

For each sample, the DNA was quantified pre- and post-PCR and the fold increase is calculated as the increase in DNA quantity post-PCR (Table 3.1). The fold increase in DNA is similar for the sonicated sample and the fragmentase sample that underwent quick ligation; 21.9 and 29 respectively. However, the fold increase in DNA is significantly lower in the samples that underwent overnight ligation; 1.2 for the fragmentase sample and 11.2 for the sonicated sample. The adapters are necessary for the PCR primers to bind and a low PCR yield indicates that there is an insufficient quantity of adapter-ligated DNA fragments present. I attempted to increase the efficiency by increasing the concentration of the enzyme in a two-hour incubation in a sonicated sample but there was no improvement in yield; 0.87 fold increase post PCR (Table 3.2). As neither approach produced adequate yields, they were not adopted.

Table 3.1: Overnight ligation test			
DNA Sample Treatment	Pre-PCR DNA (µg)	Post-PCR Column Purification DNA (µg)	Fold Increase
Sonication/Quick Ligation	0.016	0.351	21.9
Fragmentase/Quick Ligation	0.003	0.087	29
Sonication/Overnight Ligation	0.010	0.012	1.2
Fragmentase/Overnight Ligation	0.00143	0.016	11.2

Table 3.2: Two hour ligation test				
DNA Sample Treatment	Pre-PCR DNA (µg)	Post-PCR DNA (µg)	Post-PCR column Purification DNA (µg)	Fold Increase
Sonication/2hr Ligation	0.00897	0.1025	0.008	0.87

3.4.2 Fragmentase vs. Sonication

The first step of sample preparation is fragmentation of DNA to approximately 200bp and the fragmentase enzyme was tested as an alternative to the currently used method of sonication (section 2.4.4). Four reactions were set up, one with a 37°C incubation for 20 minutes, another for 30 minutes and two controls; one with no enzyme with a 20 minute incubation and the other with enzyme but without a 37°C incubation. DNA samples were separated on an agarose gel following treatment to determine the level of fragmentation achieved (Figure 3.1). The DNA in the two controls did not undergo detectable fragmentation. The 37°C incubation for 20 minutes yielded DNA fragments ranging from approximately 25bp to 400bp with a high level of 200bp fragments. The 30 minute incubation yielded fragments ranging from approximately 25bp to 200bp with a much lower level of 200bp DNA fragments compared to the shorter incubation. Thus, the 37°C incubation for 20 minutes was deemed the most suitable option.

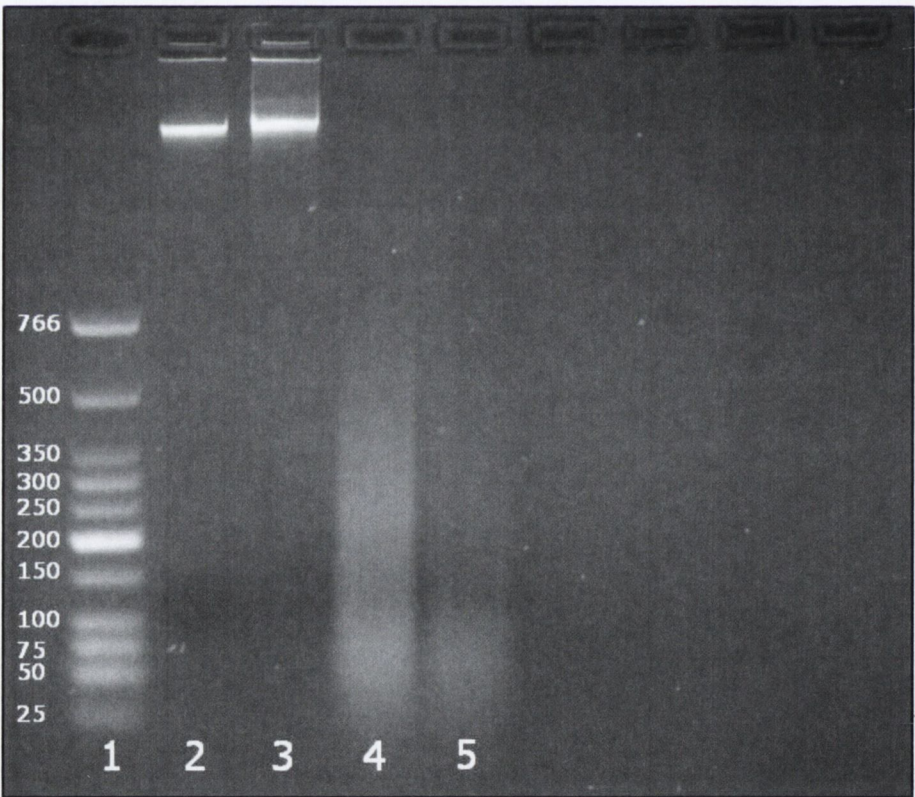


Figure 3.1: Agarose gel of dsDNA fragmentase test.

Lane 1 contains LMW marker with the size of each band marked in bps. Lane 2 contains 1µg DNA that underwent no fragmentase treatment and no incubation. Lane 3 contains 1µg DNA that underwent no fragmentase treatment and 37°C incubation for 20 minutes. Lanes 4 and 5 contain 1µg DNA that underwent fragmentase treatment with 37°C incubation for 20 and 30 minutes respectively.

I proceeded to compare the yield from the fragmentase treatment to sonication as outlined in section 2.4.4. The sample preparation protocol was applied to a sample fragmented by sonication and another by fragmentase to compare the overall efficiency of the two methods. DNA was quantified after each purification step and the yield is calculated as a percentage of the post fragmentation DNA quantity (Table 3.3). The DNA yield post fragmentation purification was similar for the two methods; 0.327 μ g for the sonicated sample versus 0.315 μ g for the enzyme-fragmented sample. Not all DNA is recovered during purification and there is an especially dramatic reduction in DNA quantity in fragmented DNA samples as column purification removes fragments of approximately 100bp or less.

The sonication approach gives rise to the greatest quantity of DNA at 0.257 μ g post PCR purification compared to 0.181 μ g in the fragmentase treated sample. Nonetheless, the quantity of the enzyme-fragmented sample is well above the minimum required DNA quantity of 50ng and this approach is amenable to a 96-well format. In comparison, just six samples can be fragmented in parallel using sonication. Therefore, the dsDNA fragmentase method of fragmentation was employed for future sample preparation.

Table 3.3: Fragmentase vs. sonication sample preparation								
	Fragmentation	Fragmentation Column Purification	End Repair Purification	A'base Purification	Ligation Purification	Gel Excision Purification	PCR Amplification Purification	Yield (%)
Sonication (1µg DNA)	0.629µg	0.327µg	0.249µg	0.196µg	0.182µg	0.0121µg	0.257µg	40.9
Fragmentase (1µg DNA)	0.684µg	0.315µg	0.235µg	0.112µg	0.12µg	0.011µg	0.181µg	26.5

3.4.3 Purification Method Optimisation

As outlined in section 2.4.5, there are six purification steps within the sample preparation protocol, which is not amenable to large sample numbers using the current method of column purification. Therefore, I tested two magnetic bead-based purification methods that can be used in a 96 well format; Agencourt Ampure XP SPRI beads and Chemagen chemagic PCR Pure magnetic beads. To compare DNA recovery following purification with these magnetic bead-based methods to the standard Qiagen column purification method, I fragmented 3 DNA samples with dsDNA framagentase and purified each with a separate method. I quantified each sample pre- and post-purification as per Table 3.4. Purification with the Qiagen column kit gave the highest recovery of 73%. The two magnetic bead-based approaches gave very similar results; 50% with Agencourt and 54% with Chemagen. I proceeded to apply the full sample prep protocol to the three samples and continued to purify by the same method. I quantified the DNA of each sample pre and post each purification step (Table 3.5). The DNA yield is calculated as the percentage of the post-fragmentase purification DNA quantity. The Chemagen and Agencourt methods were very similar up to the post ligation step where there was a dramatic reduction in the recovery from the Chemagen purification. I attempted this numerous times but each time the result was similar. Despite contacting Chemagen regarding this issue, no solution was found. Thus, as the Agencourt method gave a sufficient yield and is comparable in cost to Qiagen but is more suitable for larger sample numbers, this method was employed for sample preparation.

Table 3.4: Purification method comparison			
	DNA quantity pre-purification (µg)	DNA quantity post-purification (µg)	Recovery
Column purification	0.5µg	0.36µg	73%
Agencourt Magnetic Bead purification	0.62µg	0.311µg	50%
Chemagen Magnetic Bead purification	0.67µg	0.361µg	54%

Table 3.5: Sample prep purification method comparison							
Purification Step		DNA quantity pre-purification (µg)	Post Qiagen Column purification (µg) (Recovery)	DNA quantity pre-purification (µg)	Post Agencourt purification (µg) (Recovery)	DNA quantity pre-purification (µg)	Post Chemagen purification (µg) (Recovery)
	Starting Amount	1		1		1	
1	Post Fragmentation	0.5	0.36 (73%)	0.62	0.311 (50%)	0.67	0.361 (54%)
2	Post End Repair	0.4	0.358 (90%)	0.37	0.263 (71%)	0.428	0.193 (45%)
3	Post Add A' base	0.23	0.23 (100%)	0.23	0.162 (70%)	0.181	0.11 (61%)
4	Post Ligation	0.242	0.189 (78%)	0.17	0.125 (74%)	0.13	0.072
5	Post gel excision-Qiagen gel Purification	N/A	0.0131		0.0105		0.0078
6	Post PCR	0.323	0.189	0.31	0.165	0.275	0.09
Yield (%)			37.8		26.7		11.6

Thus, the outcome of my protocol optimisations was to prep samples in a 96-well format with dsDNA fragmentase enzyme used to fragment the DNA and Agencourt AMPure XP magnetic beads used for purification.

3.5 Analysis of Pilot Samples

Once the sample preparation method was optimised I applied it to a pilot set of 20 samples which were sequenced in order to determine the optimum number of samples per pool and to test the efficiency of the target enrichment system. As described at the beginning of the chapter, it was estimated that 20 samples per pool would give adequate sequence coverage but the optimum number of samples may be higher or lower. A greater number of samples per pool reduces project costs and allows more samples to be sequenced overall, giving rise to greater study power. However, too many samples sequenced per lane can cause insufficient sequence coverage per sample reducing the accuracy at which variants can be called. This pilot set of samples consisted of 1 pool of 20 control samples.

In this section, I will provide details on the pool of 20 indexed controls sequenced in a single lane. Integrative Genomics Viewer (IGV; www.broadinstitute.org/igv/) is a tool for visualising sequencing data (Wass 2011) and example images generated from these 20 control samples is provided in Figure 3.2. Figure 3.2A is an example plot of how reads map back to a target exon. Figure 3.2B is magnified image demonstrating detection of C/A SNV.

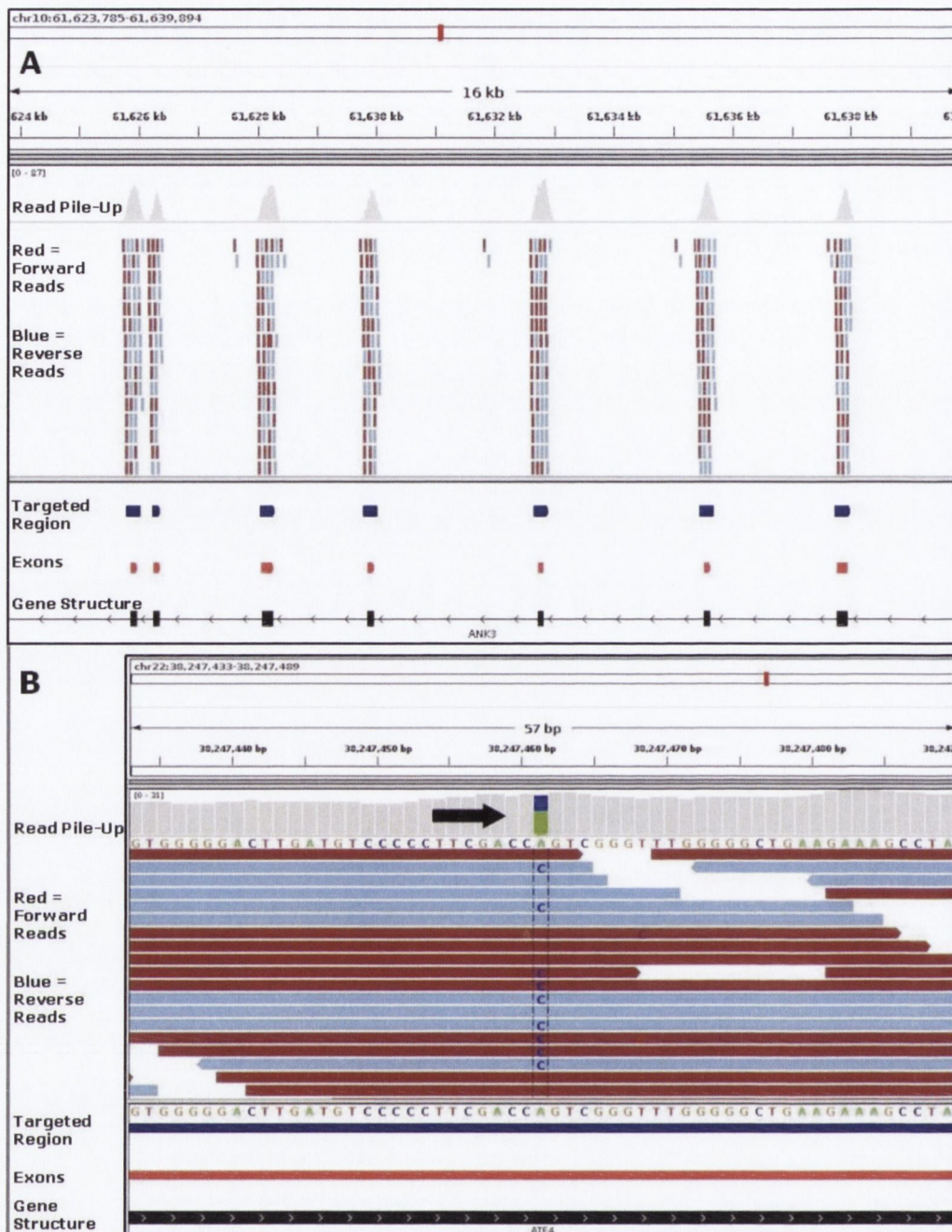


Figure 3.2: A) Example of pile-up of sequence reads on target exons following alignment and B) Magnified image representing detection of a SNV at a target exon.

Images generated using IGV. A) The total pile up of reads can be visualised at the top of the image and individual forward (red bars) and reverse (light blue bars) reads are below. Target regions, exons and the overall gene structure are marked below the reads. B) Magnified image demonstrating detection of SNV in a target exon. The coloured pile up above alignment at this SNV position represents the proportion of reads with each allele at that position. Blue represents the C allele and green represents the A allele.

The number of reads per sample varied from ~3-6% of the total sequence, with an average of 5% (Figure 3.3). Sequence data is presented for the 20 indexed control samples as the total count (either reads or bases) in the pool as an average per sample by a flowchart in Figure 3.4. This pilot pool is denoted as “Pre-modifications” on the left of the image. The total number of sequence reads is 44,814K with an average of 2,241K per sample. There is a high proportion of duplicate reads (45%) which must be removed prior to analysis. Duplicate sequence reads are identical sequence reads that arise due to PCR cycling. The amount of unique reads that aligned to the reference genome is 24,038K (54% of the total number of reads) and the mean number per sample is 1,202K. The mean number of actual bases within unique reads that could be aligned per sample is 88,614K. The mean number that aligned to baited regions is 24,563K (28%) and to target bases is 17,211K (19%). The percentage of target sequence covered by at least two reads is 96%, whilst 85% of target regions have $\geq 10X$ coverage, 61% have $\geq 20X$ coverage and 35% have $\geq 30X$ coverage (Figure 3.5).

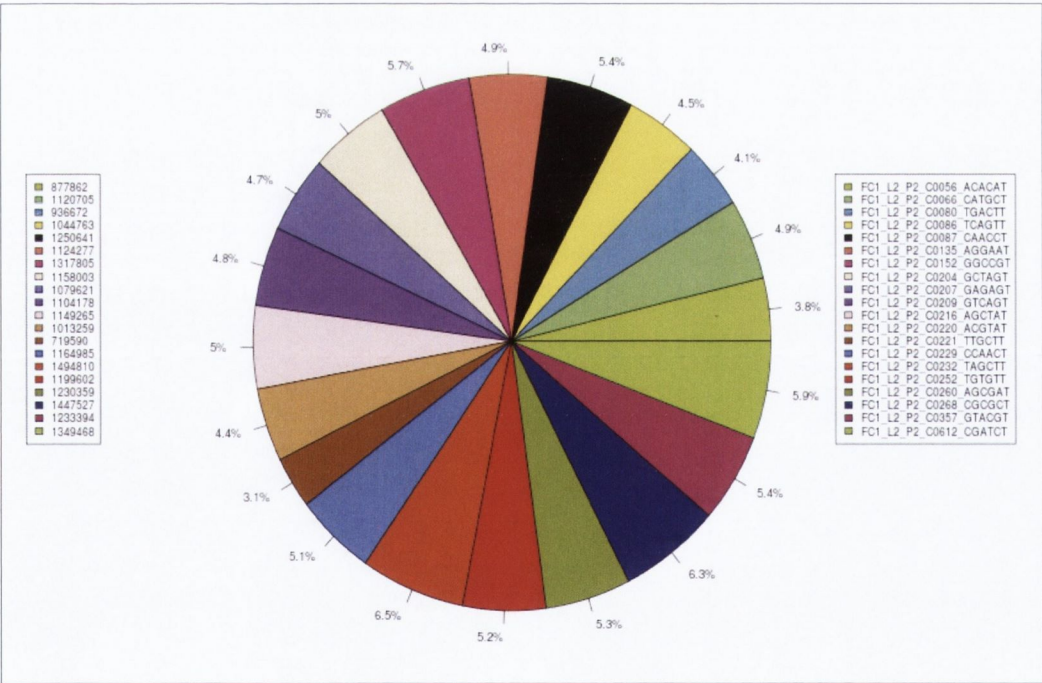


Figure 3.3: Distribution of sequence reads amongst the 20 indexed control samples in pilot study.

Each segment of the pie chart represents the proportion of sequence reads per sample which ranges from 3.1-6.5% with an average of 5%. The box on the left indicates the number of reads per colour coded sample and the box on the right indicates the sample IDs and indexes.

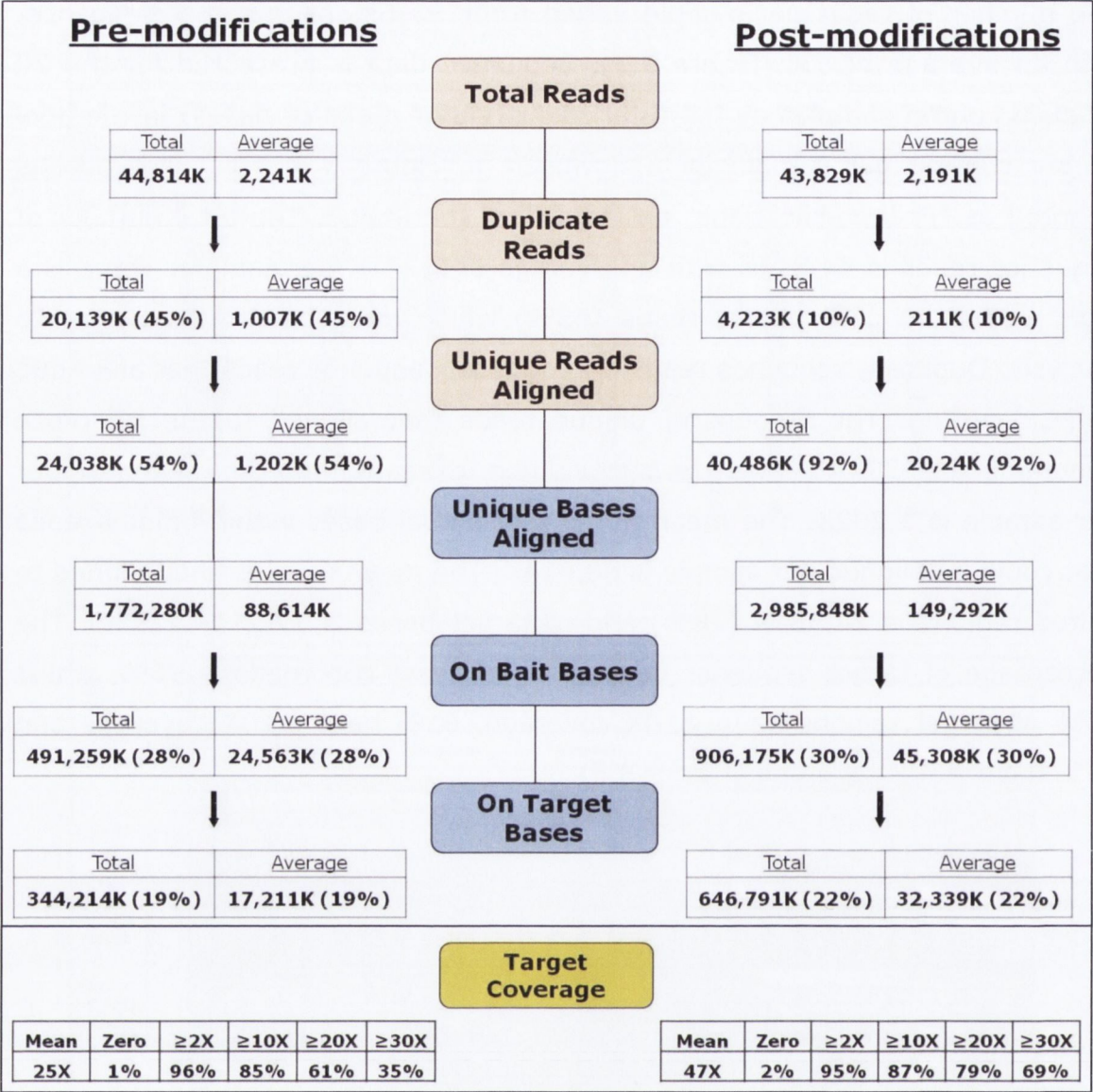


Figure 3.4: Flowchart of sequence information in two test sets of 20 indexed control samples based on the Pre-modifications and Post-modifications sample prep protocol.

Data is presented as the total count (either reads or bases) per pool of 20 samples and as an average per sample. The proportion of Duplicate Reads and the proportion of Unique Reads Aligned are calculated as a percentage of the Total Reads. The proportion of On Bait Bases and the proportion On Target Bases are calculated as a percentage of the Unique Bases Aligned. Also, the mean Target Coverage is provided and a breakdown of the percentage of target regions with zero to ≥30X coverage.

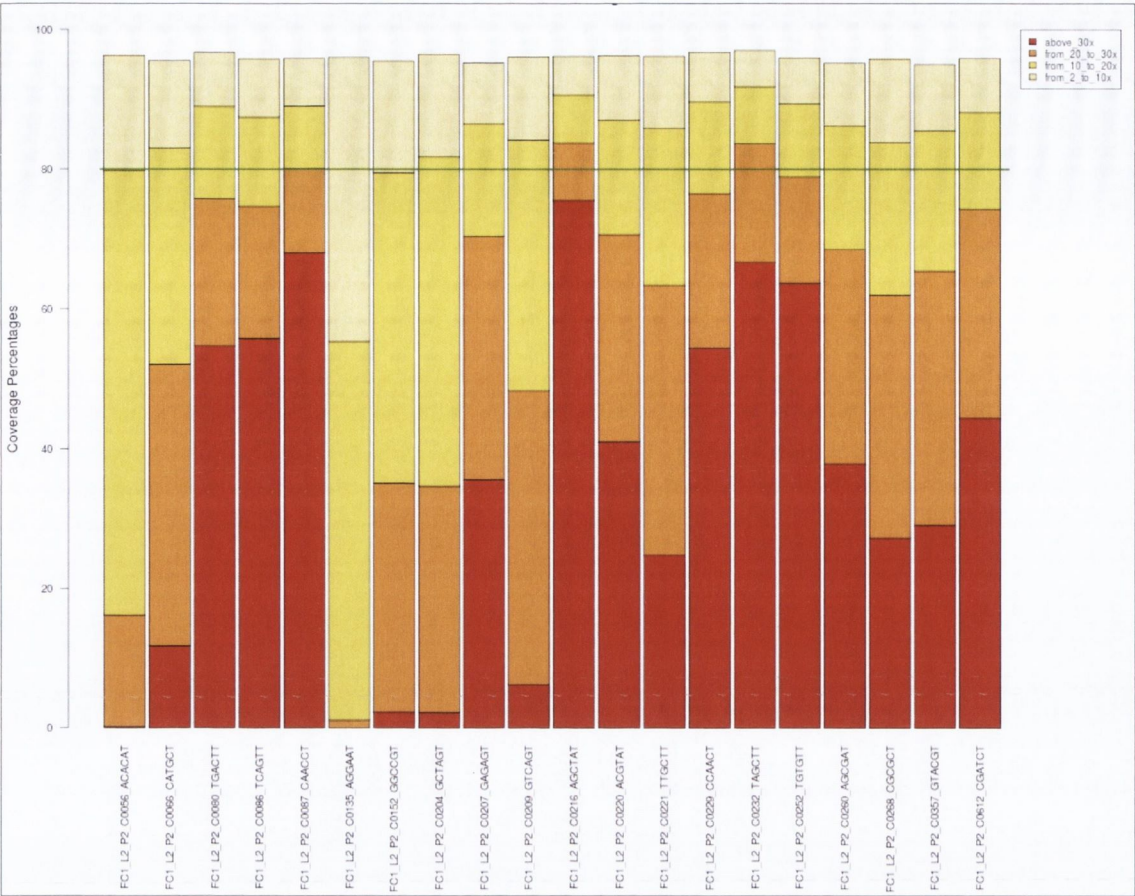


Figure 3.5: Percentage target coverage per sample in pool of 20 indexed controls in pilot study.

Each bar represents one sample and the target coverage is depicted in 4 ranges as illustrated by colours outlined in upper right hand box; light yellow is 2-10X coverage, yellow is 10-20X coverage, orange is 20-30X coverage and red is 30X coverage and above.

As a large proportion of the reads are duplicates in the pilot study, I amended the protocol to combat this issue. I increased the DNA input at the beginning of the protocol from 500ng to 1µg, increased DNA quantities in the PCR steps and reduced the PCR cycle number from 11 cycles to 5 cycles in the pre-enrichment PCR and from 11 to 10 cycles in the post-enrichment PCR. I applied this amended protocol to a further 20 indexed control samples to determine if duplicate rates would be decreased and to establish optimum sample number per pool.

In this second set of 20 control samples, the percentage of total reads per samples again varied from ~3-6% (Figure 3.6). Sequence data is presented as a flowchart in Figure 3.4 with this pool of samples denoted as “Post-modification”. The total number of sequence reads is 42,829K and the rate of

duplicate reads is 10%. The amount of unique reads that aligned to the reference genome is 40,486K (92% of the total number of reads) and the mean number per sample is 2,024K. The number of unique bases on target regions is 646,791K (22%). The percentage of target sequence with $\geq 2X$ coverage is 95%, $\geq 10X$ coverage is 87%, $\geq 20X$ coverage is 79% and $\geq 30X$ coverage is 69% (Figure 3.7).

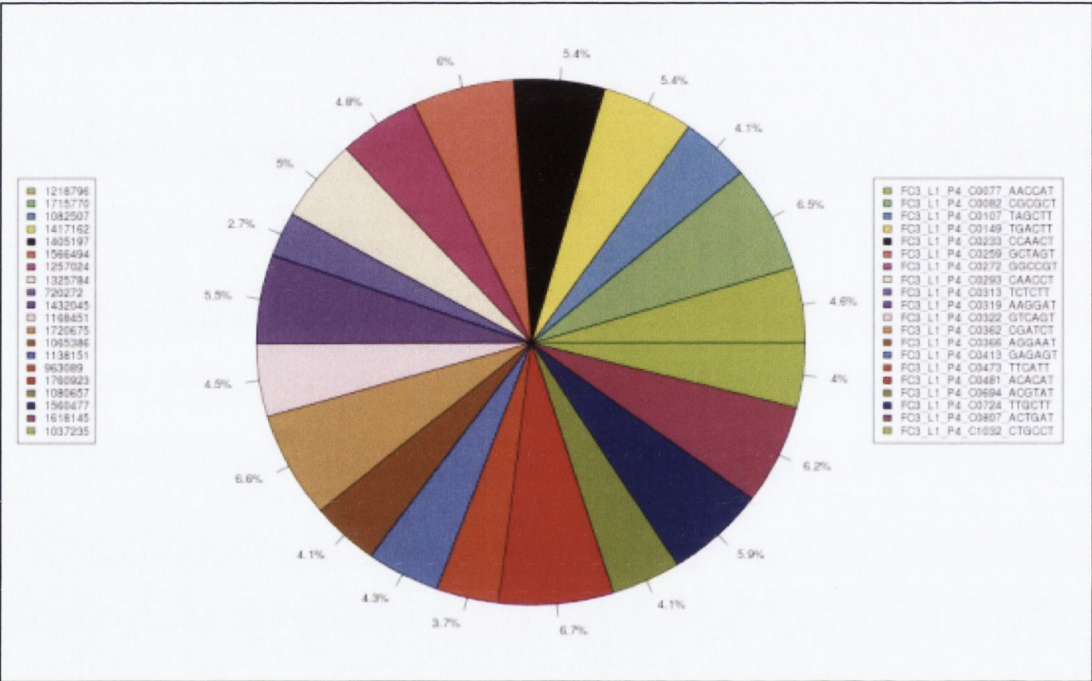


Figure 3.6: Distribution of sequence reads amongst the 20 indexed control samples post protocol modifications. Each segment of the pie chart represents the proportion of sequence reads per sample which ranges from 2.7-6.7% with an average of 5%. The box on the left indicates the number of reads per colour coded sample and the box on the right indicates the sample IDs.

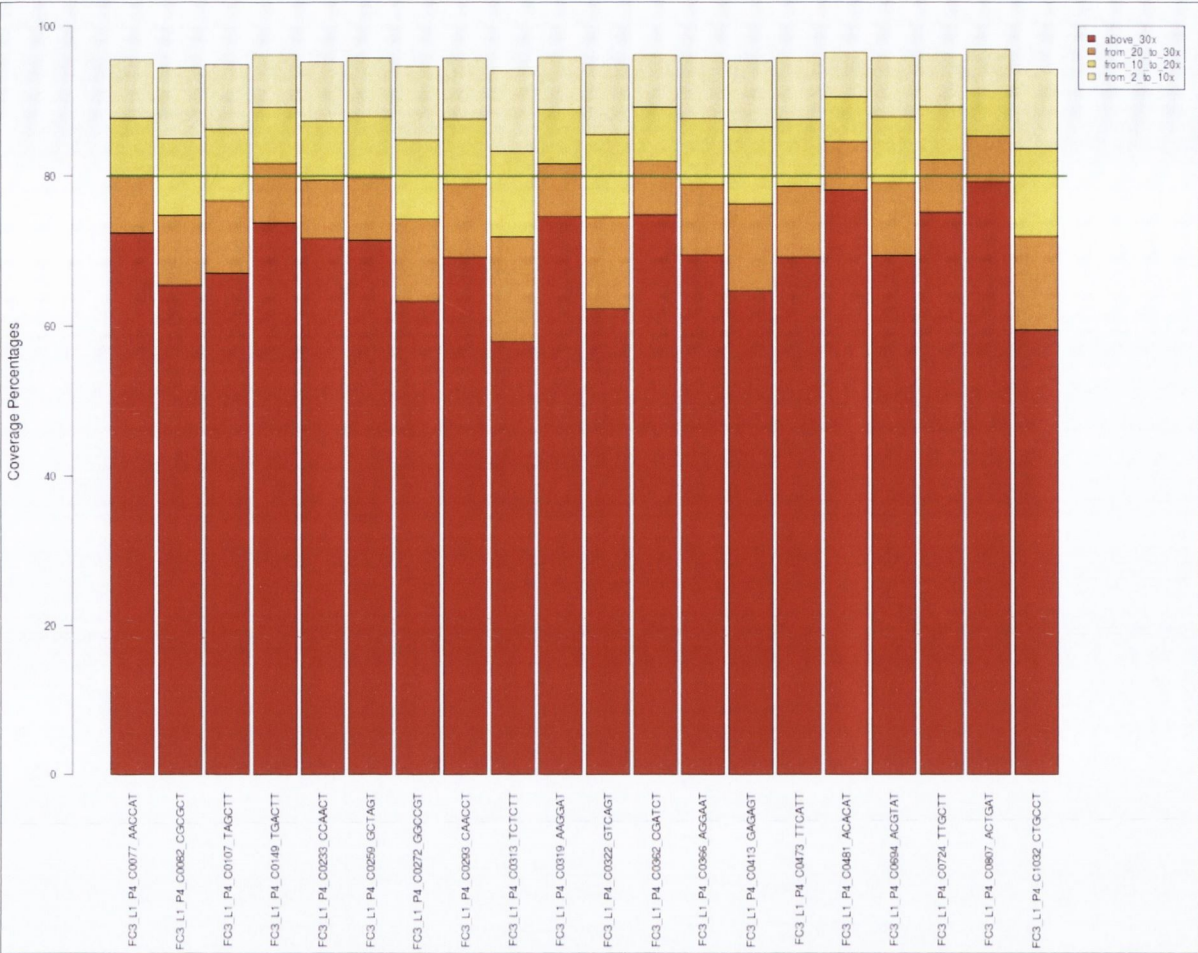


Figure 3.7: Percentage target coverage per sample in pool of 20 indexed controls after protocol amendment.

Each bar represents one sample and the target coverage is depicted in 4 ranges as illustrated by colours outlined in upper right hand box; light yellow is 2-10X coverage, yellow is 10-20X coverage, orange is 20-30X coverage and red is 30X coverage and above.

As the protocol modifications lead to a significant improvement in the number of duplicate reads and in uniformity of sequence coverage across samples, the approach was applied to all remaining samples.

3.6 Conclusion

A systematic approach was undertaken to select candidate genes for inclusion in the study. This is a very subjective process and there are almost certainly risk genes that were not included. The list of potential candidates was far greater than 215 and categories 2, 3 and 4 in particular needed to be reduced by prioritising genes based on supporting evidence.

Modification to the sample prep protocol including changing from fragmentation by sonication to an enzymatic approach and from purification with columns to magnetic bead technology, has allowed it to be performed in a 96-well format and for 24 samples to be processed in parallel. This has made the overall method far less laborious and time consuming.

The rate of duplicate sequence reads was 45% in the pilot set of 20 samples. Implementation of protocol modification in a second set of 20 samples not only reduced this rate to just 10% but gave a greater uniformity of sequence coverage also. As a result there is an increase in the level of usable sequence data generated making it possible to increase the number of samples per pool to 24. This has made the process of sample prep much more cost efficient allowing greater numbers of samples to be sequenced, increasing the overall study power to detect sequence variants.

In the next chapter I will discuss findings from the full sequencing study where my methods were applied to >700 DNA samples.

4 Investigation of Identified Sequence Variants including Validation, *In Silico* analysis and Association Analysis

4.1 Aim

The aim of this chapter is to explore findings from sequencing of the full dataset of SZ, ASD and control samples. The primary objective is to investigate rare (minor allele frequency (MAF) <1%) variants. The variants will be categorized based on likely functional impact on protein sequences. The first subset of variants to be prioritised are those most likely to directly impact protein function and thus, potentially contribute to disease pathogenesis. Such variants are termed loss-of-function (LoF) variants and will be subjected to validation by capillary sequencing and *in silico* analysis to determine the likely impact on the amino acid sequence. The next group of likely functional variants for investigation will be rare missense variants. The study of these variants includes the use of annotation tools for prioritisation of missense variants based on amino acid changes caused in the protein sequence. Association analysis will be performed on both LoF variants and missense variants to determine if there is a significant difference between the amount of these variants detected in cases compared to controls.

4.2 Full Dataset Analysis

This project involved numerous members of the Neuropsychiatric Genetics Research Group. The work described in this section 4.2, was primarily performed by Dr. Paul Cormican and Dr. Elaine Kenny. I include this information here in order to explain how the final dataset of sequence variants was generated and made available to me for further study (4.3 onwards).

From a total sample of 743 individuals, 199.5Gb of sequence data was generated. Samples were split according to index and aligned to a reference human genome using BWA (<http://bio-bwa.sourceforge.net/>). Read recalibration, realignment and calling of both single nucleotide variants (SNVs) and indels performed using GATK (v1.0.5506; (DePristo et al. 2011)). The majority of samples were analyzed in previous GWAS; 277 of 297 SZ cases (Affy 6.0 data), all controls (Affy 6.0 data) and 135 of 152 ASD cases

(Illumina Human 1M-single Infinium BeadChip data). Therefore, it was possible to compare the GWAS SNV data with SNVs identified from the sequence data. For SZ samples the average matching between GWAS and sequence data variant calls was >99%. As a means to determine if the rate of variant detection was even across the sample groups, the level of silent SNVs per sample was assessed. Silent variants are not thought to be under selective pressure and as such, should be similar in all individuals within the same population regardless of disease status. There is an even average of 167 silent variants per SZ sample (s.d.=12.6), 168 per ASD sample (s.d.=12.3) and 167 per control sample (s.d.=12.8). Therefore, despite varying levels of sequence data being available for different sample groups (Figure 4.1), there is the same power to detect variants in each group.

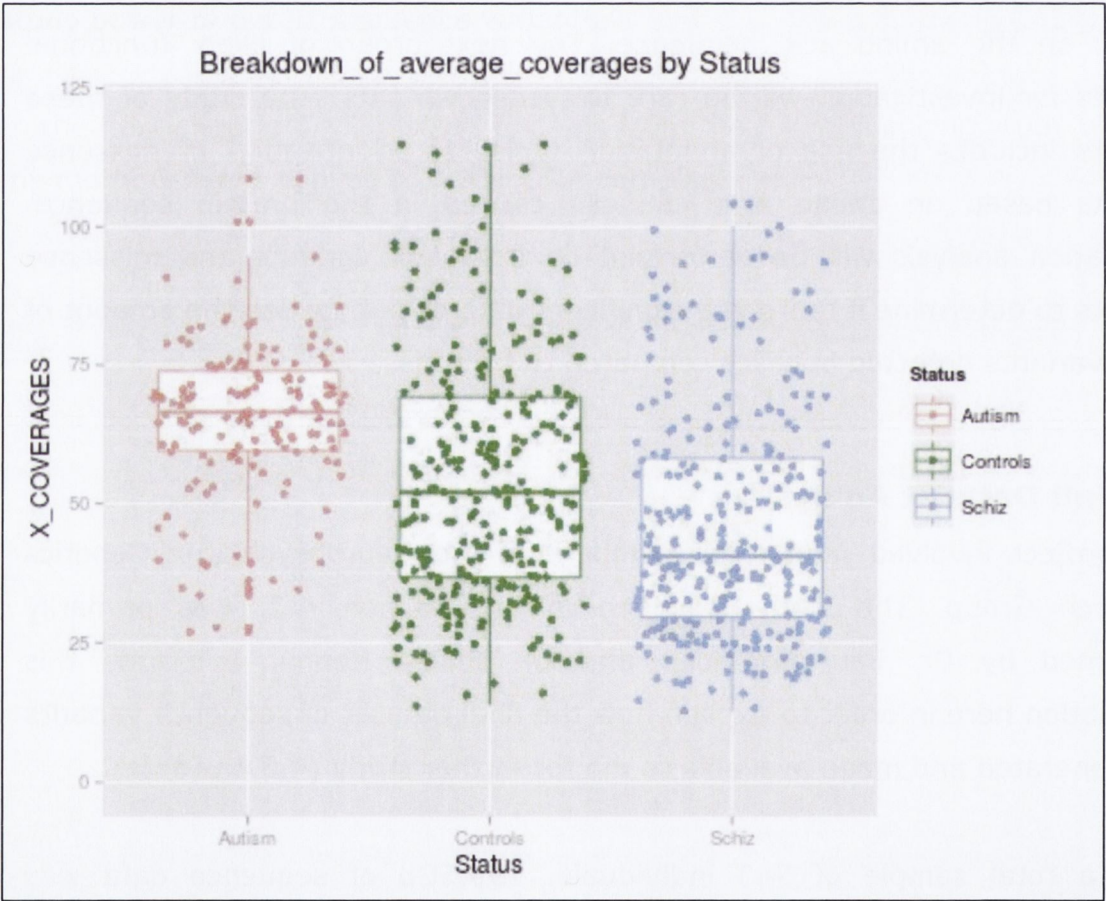


Figure 4.1: Box plots of the sequence coverage for each of the three phenotype groups in the study.
The median coverage for all samples included in the final analysis was 66X for ASD, 52X for controls and 41X for SZ.

Figure 4.2 provides a flowchart of the number of variants detected across all samples and how that number was reduced to a set of variants for inclusion in my analysis following QC. The total number of variants in target exons (+/- target 25bp) was 5,850 (5,389 SNVs and 461 indels). All variants were functionally annotated using variant effect predictor (VEP) (http://.ensembl.org/Homo_sapiens/UserData/UploadVariations). After applying default GATK filters, 830 SNVs and 320 indels were removed leaving 4,700 variants. The dataset was further reduced by applying quality control (QC) filters to both samples and variants. Samples with concordance of <99% with previous genotype calls and/or with missing data at >10% of variants were removed. Variants with genotype quality score <20, read depth <8X or had missing data in >10% of samples were removed. The removal of samples also rendered some monomorphic variants which were removed. This left 3,708 (3,542 SNVs and 90 indels) variants in 273 SZ cases, 147 ASD cases and 287 controls.

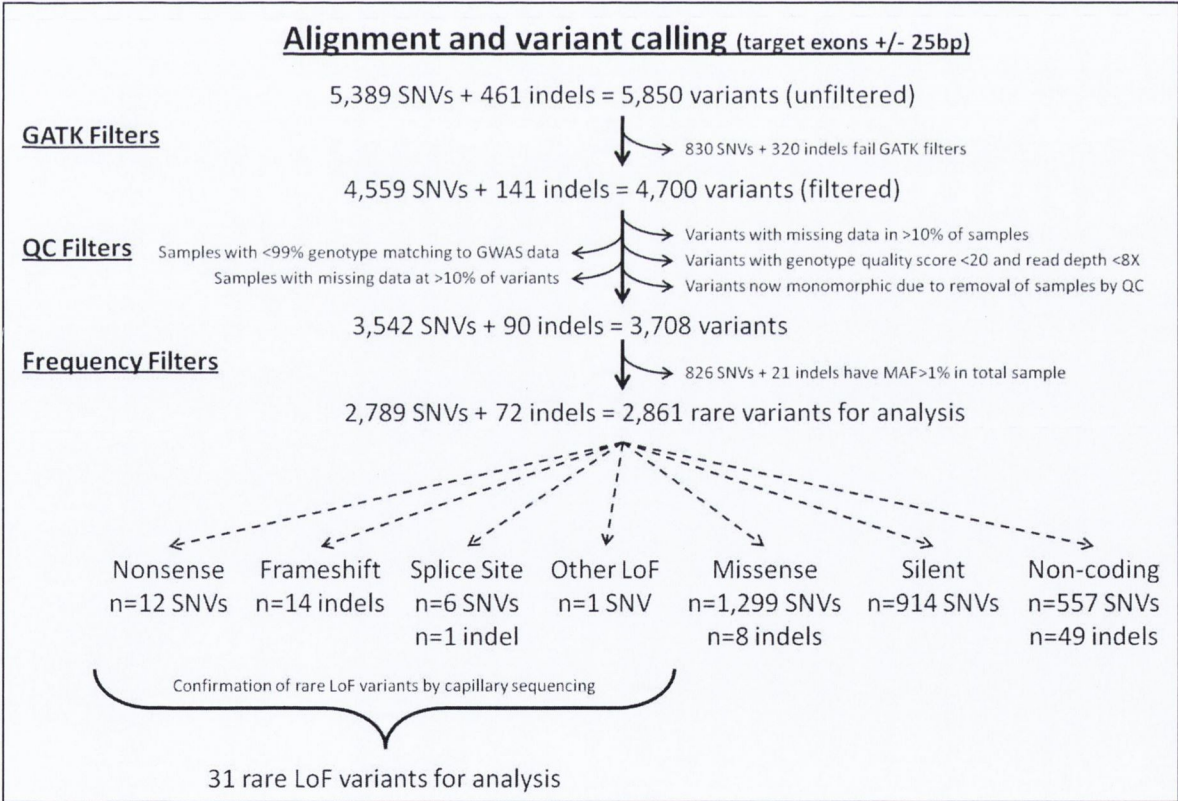


Figure 4.2: Flowchart of breakdown of variants in full dataset. Each step represents a filtering step of the variants.

As the primary focus of this study is the investigation of rare variants, a frequency cut-off of MAF <1% in the combined case and control samples was applied to all variants. As per Figure 4.2, the number of variants reduced from 3,708 post QC to 2,861 rare variants (2,789 SNVs and 72 indels), which were categorised based on potential function based on VEP data. This dataset (in Plink format) and the initial VEP data were supplied to me and I took it forward for further analysis.

4.3 Loss-of-Function Variants

Loss-of-Function (LoF) variants are predicted to severely disrupt protein-coding sequence and my primary analysis was to investigate such variants. The definition of LoF variants was taken from a recent, comprehensive study of LoF variation (MacArthur et al. 2012). LoF comprise nonsense SNVs that introduce stop codons, SNVs that disrupt splice sites and indels that disrupt a transcripts open reading frame or a splice site. Figure 4.3 illustrates such variants. In addition, I detected one other variant type that was considered as potential LoF and included it in my analysis. This was a stop loss SNV that change the predicted stop codon into an amino acid codon and extended the reading frame into the 3' UTR. In total, I found 33 rare LoF variants in the sample.

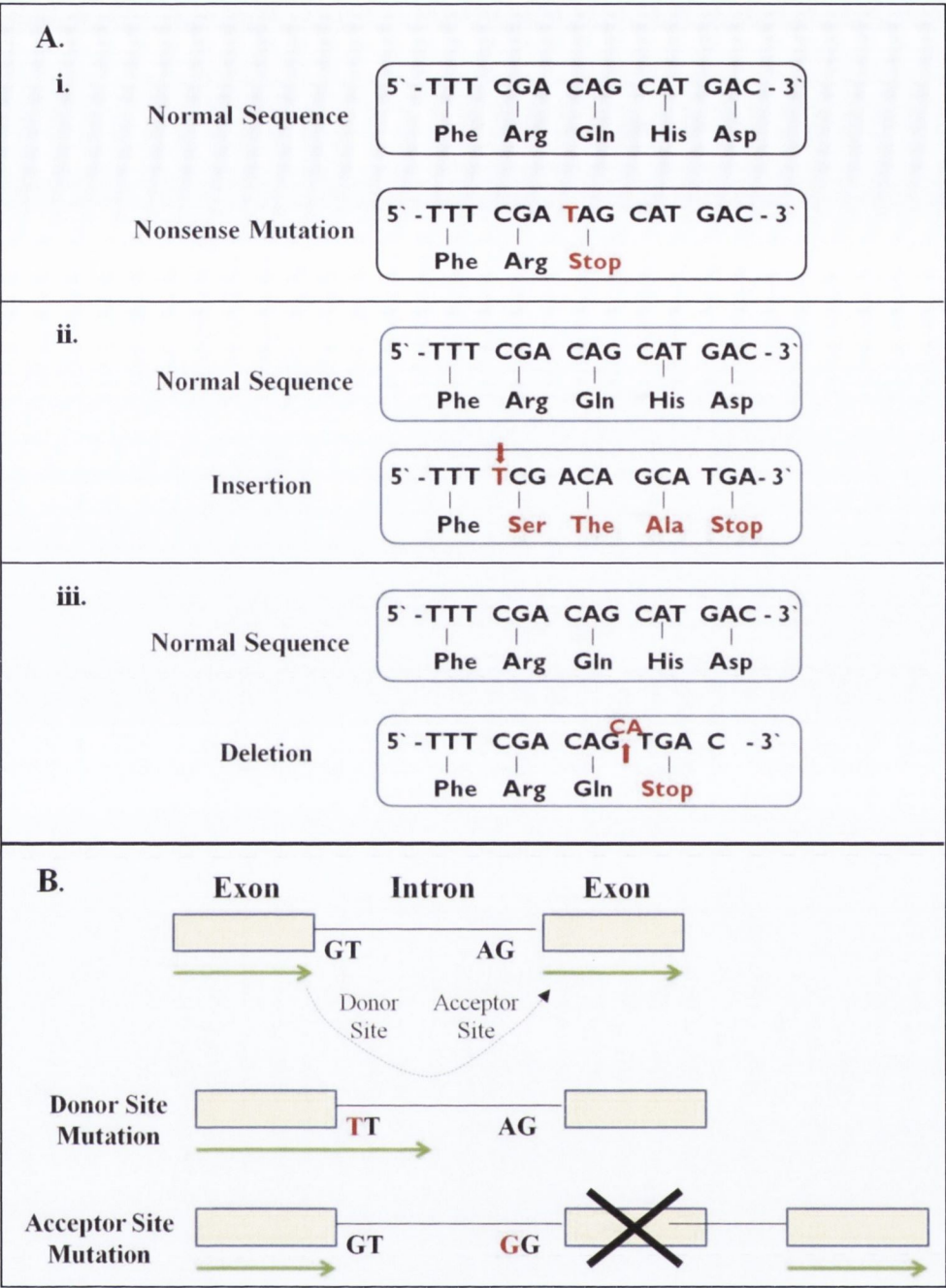


Figure 4.3: Illustration of the affect of LoF variants

A. An example of a section of normal DNA sequence and it's corresponding amino acid sequence is provided, along with examples LoF variants (marked in red). i. The replacement of a C allele with a T allele at this locus would result in the replacement of the amino acid glutamine with a stop codon, truncating the protein. ii) The insertion of a T allele at this locus would result in a frameshift i.e. each amino acid after this point would be different eventually resulting in the introduction of a stop codon. iii) A CA deletion at this locus would result in a frameshift. B. Splice sites at exon-intron boundaries are sites that are recognised by splicing machinery. The donor site is usually GT and the acceptor site is usually an AG. Mutations at donor sites can result in introns not being spliced out and transcription continuing into the intronic sequence. Mutations at acceptor sites can result in exons not being recognised by splicing machinery and thus not being included in the mRNA sequence.

4.3.1 LoF variant Validation

As outlined in section 2.6.2, all LoF variants were subjected to Sanger sequencing, and 31 of 33 were confirmed by this method; 11 nonsense SNVs, 12 frameshift indels, 6 splice site SNVs, 1 splice site indel, and 1 stop loss SNV (Table 4.1). The two variants that did not validate were a 8bp insertion in *SYNGAP1* and a 1bp deletion in *GJB6* as marked in red in Table 4.1. As parental DNA was available for the ASD samples, these were also analysed to determine if variants in the ASD samples were inherited or *de novo*. All ASD variants were inherited except those marked by an e in Table 4.1.

Table 4.1: Breakdown of all LoF variants											
Chr	Position	Gene	Gene Category ^a	Ref allele	Alt allele	Type	SZ ^b	ASD ^b	CON ^b	LoF? ^c	Singleton? ^d
1	39788293	MACF1	1	AAC	A	Coding Indel	1			Yes	Yes
1	62321741	INADL	1	TC	T	Coding Indel		1		Yes	Yes
1	62349979	INADL	1	GC	G	Coding Indel	1			Yes	Yes
1	62456007	INADL	1	C	T	Nonsense SNV	1			Yes	Yes
1	208216512	PLXNA2	5	GT	G	Coding Indel	1			Yes	Yes
1	232144803	DISC1	2	CT	C	Coding Indel ^e		1		No	Yes ^e
2	187519413	ITGAV	5	A	AG	Coding Indel	1			Yes	Yes
2	239257490	TRAF3IP1	2	G	T	Splice Site SNV		1		Yes	Yes
3	57282220	APPL1	3	G	T	Splice Site SNV		1		Yes	Yes
4	72433527	SLC4A4	3	G	GT	Coding Indel			1	No	Yes
4	187628509	FAT1	3	C	A	Nonsense SNV	1			Yes	Yes
6	33400481	SYNGAP1	5	G	GGCTAA AAA	Coding Indel		1		Unconfirmed	
6	56358939	DST	4	TA	T	Coding Indel	1			Yes	Yes
6	56472474	DST	4	G	A	Nonsense SNV	1			Yes	Yes
6	56479284	DST	4	T	C	Splice Site SNV		1		Yes	Yes
6	56482783	DST	4	C	CCT	Splice Site Indel		1		Yes	Yes
6	56483170	DST	4	C	A	Nonsense SNV	1			Yes	Yes
6	56483389	DST	4	G	A	Nonsense SNV			1	Yes	Yes
6	56507564	DST	4	TA	T	Coding Indel	2	2	2	Yes	No
6	112025283	FYN	1	G	A	Nonsense SNV	1			Yes	Yes
8	27463990	CLU	4	CTG	C	Coding Indel		1	1	Yes	No
10	79584235	DLG5	1	C	G	Splice Site SNV			1	Yes	Yes
10	79614016	DLG5	1	C	A	Nonsense SNV	1			Yes	Yes
12	13724778	GRIN2B	3	G	A	Nonsense SNV		1		Yes	Yes ^e
12	66765472	GRIP1	3	A	T	Splice Site SNV	1			Yes	Yes
12	66923668	GRIP1	3	G	A	Nonsense SNV		1		Yes	Yes
13	20797176	GJB6	4	TG	T	Coding Indel	1			Unconfirmed	
13	20797556	GJB6	4	TC	T	Coding Indel		1	3	Yes	No
13	109610055	MYO16	1	C	T	Nonsense SNV	1			Yes	Yes
17	40844654	CNTNAP1	1	C	T	Nonsense SNV		1		Yes	Yes
18	74728772	MBP	3	A	G	Stop Loss SNV	1	1		No	No
X	32429867	DMD	2	G	A	Splice site SNV		1		Yes	Yes
X	70367905	NLGN3	1	TC	T	Coding Indel	1			Yes	Yes

^a 1) NRXN1 and interactors, 2) DISC1 and interactors, 3) Neural cell adhesion molecules, 4) Post-synaptic Glutamate Receptor Complexes (NMDA, mGluR5 and AMPA) and 5) Functional and Positional Candidates

^b Number of variants detected in these phenotype categories

^c Yes = LoF variant located in first 95% of coding sequence. No = LoF variant located in last 5% of coding sequence.

^d Singleton variants are those found in just sample in the study.

^e De novo variant

Example electropherogram images are given below of: a nonsense mutation in a SZ sample in the *FYN* gene (Figure 4.4), a *de novo* nonsense mutation in an ASD sample in *GRIN2B* (Figure 4.5), a *de novo* frameshift deletion in an ASD sample in *DISC1* (Figure 4.6), a frameshift insertion in a control sample in *ITGAV* (Figure 4.7), a donor splice site mutation in a SZ sample in *GRIP1* (Figure 4.8) and an inherited stop loss mutation in an ASD sample in *MBP* (Figure 4.9).

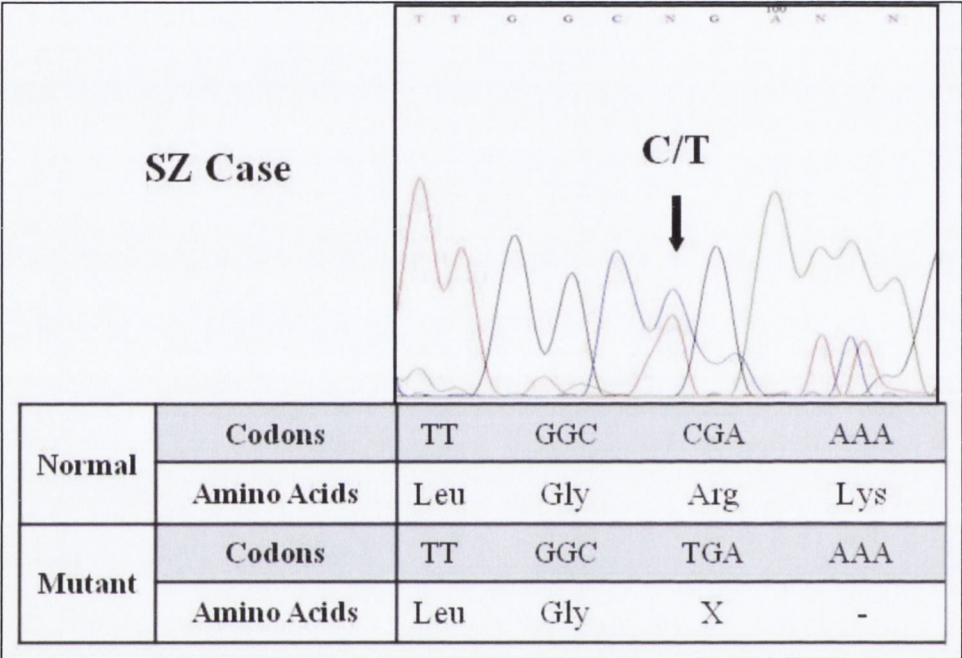


Figure 4.4: Electropherogram image of a nonsense mutation in *FYN*. The C to T mutation at chr6: 112025283 results in the introduction of a stop codon in a SZ sample. The codons corresponding to amino acid sequence are given for the reference sequence and the mutated sequence.

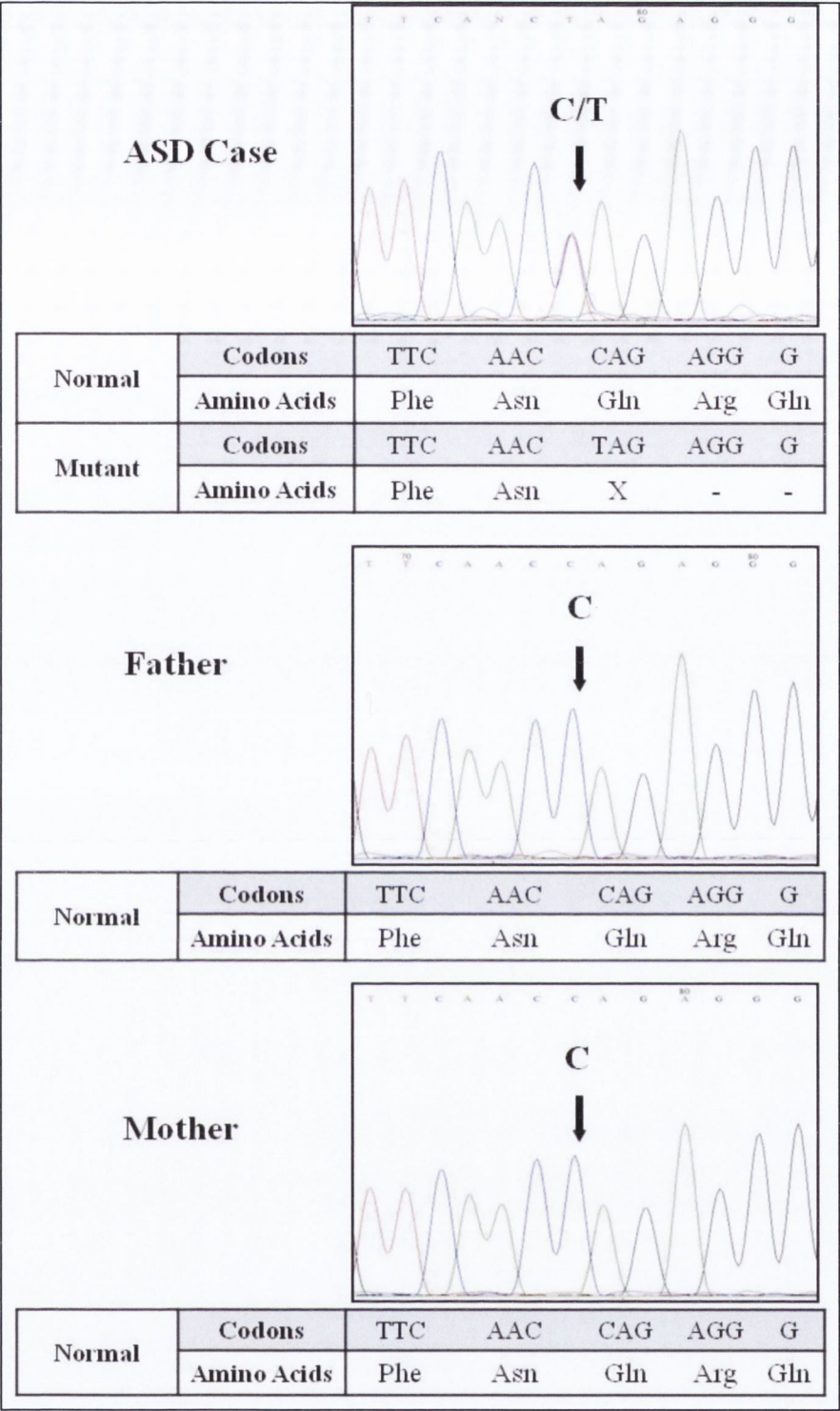


Figure 4.5: Electropherogram image of a *de novo* nonsense mutation in *GRIN2B*.

The C to T mutation at chr12: 13724778 results in the introduction of a stop codon in an ASD sample. The parental DNA does not contain the mutation, indicating that it is *de novo*.

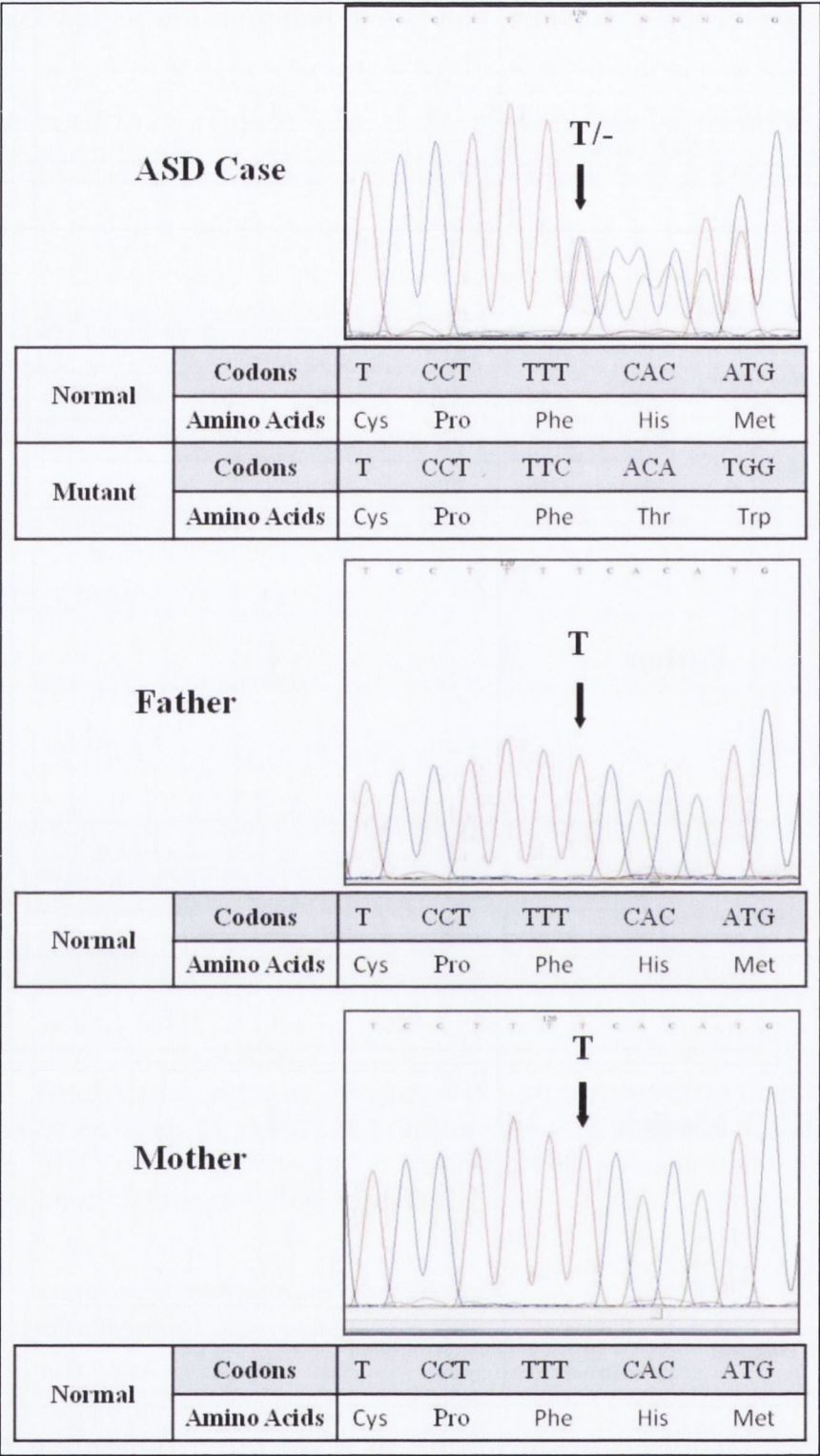


Figure 4.6: Electropherogram image of a *de novo* 1bp deletion in *DISC1*.

The *de novo* deletion of a T allele at chr1:232144803 in an ASD sample results in a frameshift in the coding sequence that introduces a stop codon 44 codons downstream. The parental DNA does not contain the mutation indicating that it is *de novo*.

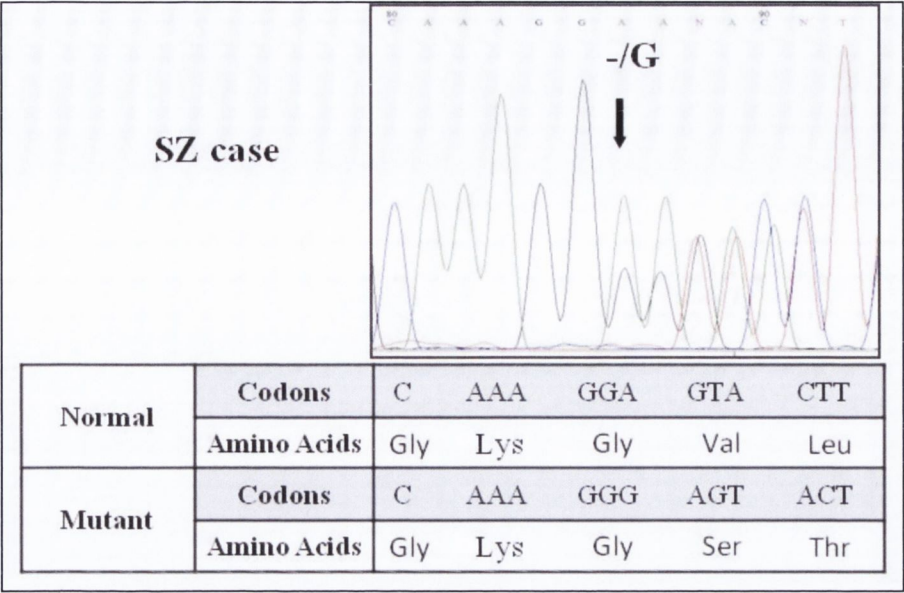


Figure 4.7: Electropherogram image of a 1bp insertion in *ITGAV*.
The insertion of a G allele in a SZ sample at chr2: 187519413 results in a frameshift in the coding sequence that introduces a stop codon 44 codons downstream.

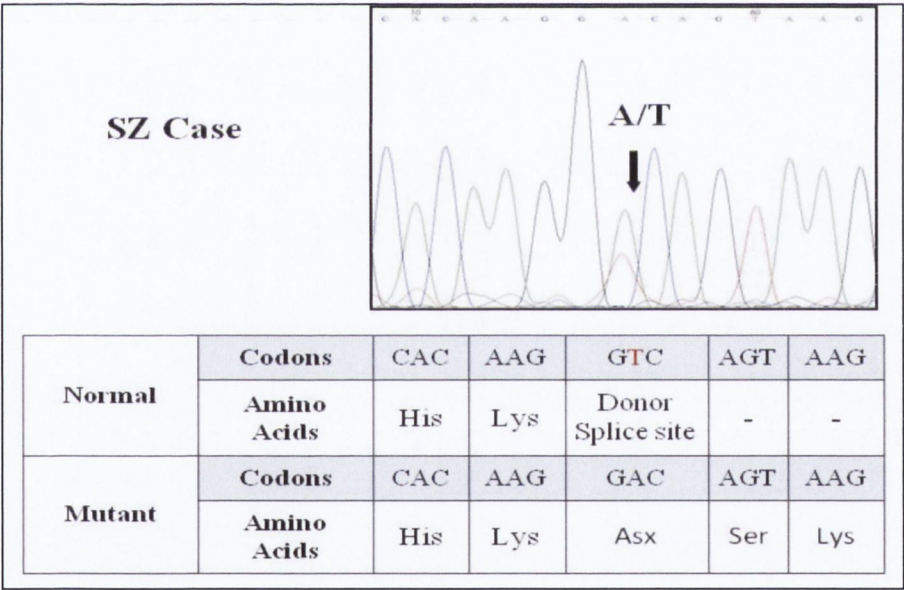


Figure 4.8: Electropherogram image of a donor splice site mutation in *GRIP1*.
An A to T change at a donor splice site in a SZ sample at chr12:66765472. Failure of the donor site to be recognised would result in the reading frame continuing into the intronic sequence.

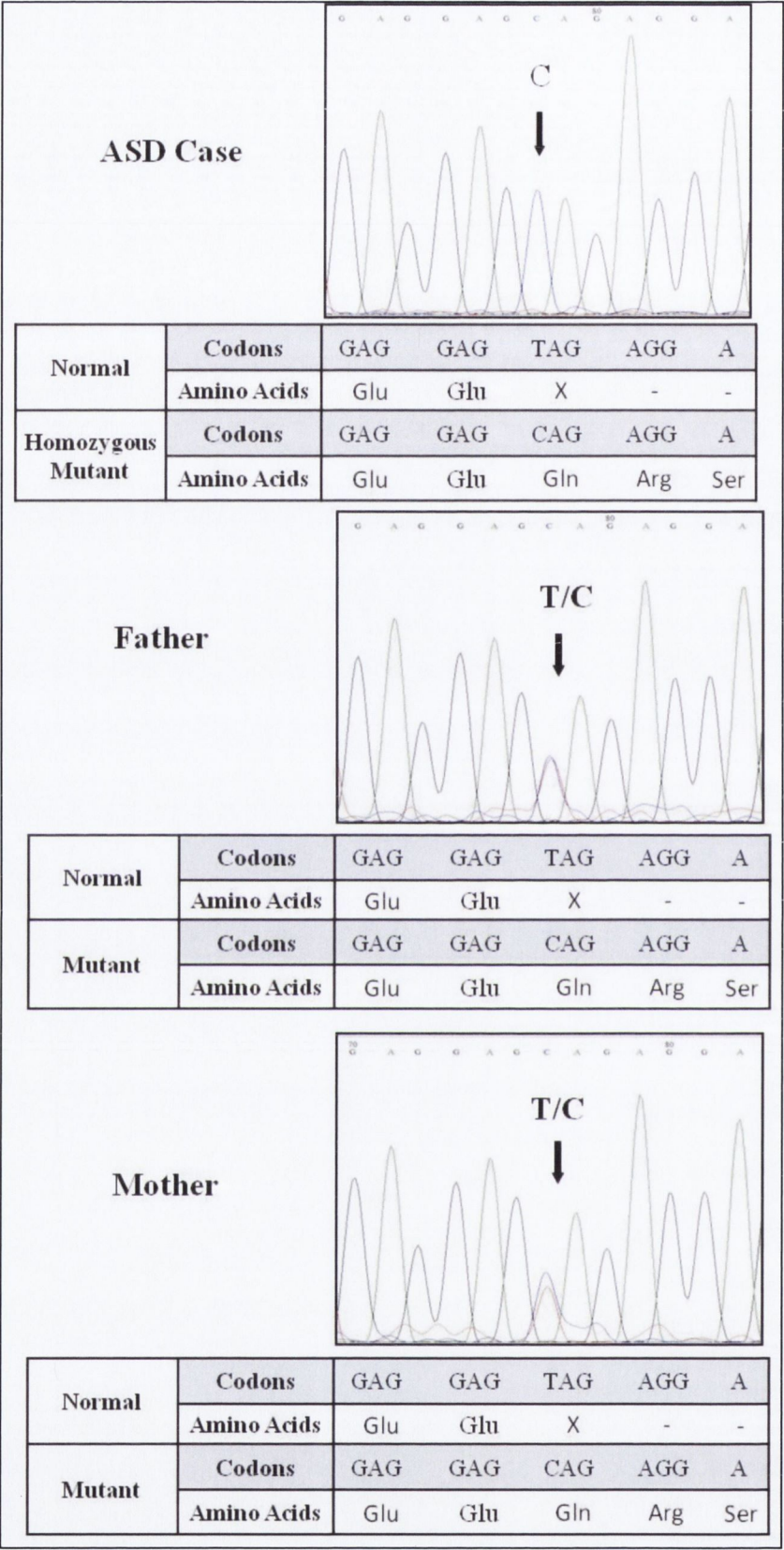


Figure 4.9: Electropherogram image of a stop loss mutation in *MBP*.
The T to C mutation at chr18: 74728772 results in the loss of a stop codon in an ASD sample. The mutation is homozygous in the case and is both maternally and paternally inherited as both parents are heterozygous for the mutation.

4.3.2 Determination of Parent-of-Origin of *De Novo* Variants

As parental DNA was available for the ASD samples, it was possible to determine parental origin of LoF variants identified in these samples. In this analysis, two variants were revealed to be *de novo*; a 1bp deletion at chr1:232144803 (*DISC1*) in sample AS084C1 and a nonsense mutation at chr12:13724778 (*GRIN2B*) in sample AS155C1, as identified in Table 4.1. Electropherograms of these findings can be found in Figure 4.5 for *GRIN2B* and Figure 4.6 for *DISC1*. Determination of the parent-of-origin of these *de novo* LoF variants was achieved using allele-specific (AS) PCR (Edghill et al. 2007). A description of this method can be found in Figure 4.10. First of all, a heterozygous SNV had to be identified in the proband that was near enough to the *de novo* SNV to allow both variants to be amplified in a single PCR product and sequencing. This neighbouring SNV also had to have different genotypes in the parental samples. As outlined in detail in section 2.6.3, SNVs meeting these requirements were identified for both *DISC1* (rs821616) and *GRIN2B* (rs7312845). Table 4.2 and Table 4.3 provide the genotypes of the probands and parents for the *de novo* LoFs and the nearby SNVs for *GRIN2B* and *DISC1*, respectively.

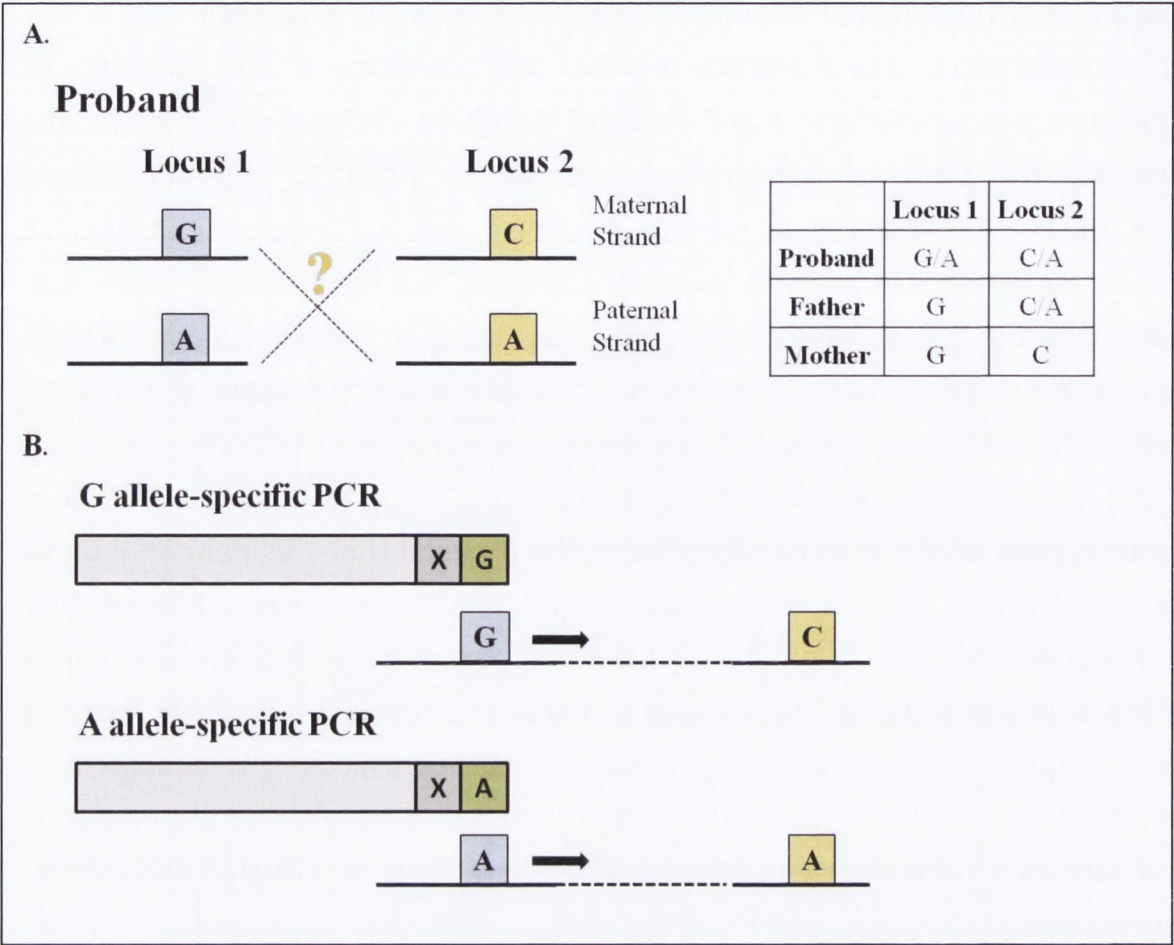


Figure 4.10: Determining parent-of-origin of *de novo* variants by allele-specific PCR

A. In this example scenario, a proband has a G/A genotype at locus 1. The A allele is a *de novo* mutation i.e. both parents are homozygous for the reference G allele. One wants to determine if this mutation arose on the maternal or the paternal DNA strand. At a downstream locus (locus 2), the proband has a C/A genotype. At this locus, the father is heterozygous and the mother is homozygous for the C allele. Thus, the proband must have inherited the C allele at locus 2 from the mother and the A allele from the father. (as per the table). B. AS-PCR can be designed from locus 1 to reveal which of the alleles are on the same strand. The last base of the AS-PCR primers are positioned at locus 1. One primer is G allele-specific and the other is A allele-specific. Each primer has a penultimate base mismatch to increase specificity at the final base. PCR and sequencing of the proband with these primers will reveal which allele is at locus 2. In this scenario, sequencing with the G-allele-specific primer reveals a C allele at locus 2, indicating that they reside on the same strand and thus, the *de novo* mutation is inherited on the maternal strand.

4.3.2.1 **GRIN2B De Novo LoF Variant Parent-of-Origin**

The proband is heterozygous for rs7312845 (C/A), the father is heterozygous (C/A) and the mother is homozygous for the C allele (Table 4.2). AS155C1 must have inherited the A allele from the father and the C allele from the mother. The AS-PCR primers were designed for the *de novo* SNV due to

difficultly with primer design around rs7312845. Primers were designed with either a G or an A base at the 3' end of the forward primer with and without a mismatch at the penultimate base. The primer sequences are provided in Table 2.19. PCR using the primers with mismatches yielded more specific products as determined by gel analysis, where single bands of the correct size could be visualised. The primers without the mismatches yielded many unspecific bands on the gel and as such were not used. Proband and parents all underwent AS-PCR and if a PCR product was attained, sequencing using both allele-specific primers was performed.

Table 4.2: Genotypes of *GRIN2B* SNVs used in AS-PCR.

Samples	Status	Genotype at LoF variant site chr12: 13724778	Genotype at rs7312845
AS155C1	Case	G/A	C/A
AS155F	Father	G	C/A
AS155M	Mother	G	C

Results of the AS-PCRs for *GRIN2B* can be found in Figure 4.11 and Table 4.3. PCR of AS155C1 with the G allele (reference) specific primer revealed an A allele at rs7312845 and PCR with the A allele (mutant) primer revealed a C allele at this site. As the C allele must have been inherited from the mother, this indicates that the *de novo* SNV allele (A) arose on the maternal strand. The reliability of the finding is supported by PCR and sequencing of the parents. PCR with the mutant specific primer did not yield any product for either parent. PCR and sequencing with the G specific primer revealed a C/A allele for the father and a C allele for the mother. This is consistent with neither parent carrying the mutant allele as both are homozygous for the reference allele. As such two strands are amplified and sequenced resulting in both a C and an A allele for the father and a C allele for the mother.

Table 4.3: Alleles found at rs7312845 by AS-PCR from *GRIN2B* *de novo* SNV

Sample	Status	PCR with G allele-specific primer	PCR with A allele-specific primer
AS155C1	Case	A	C
AS155F	Father	C/A	(no product)
AS155M	Mother	C	(no product)

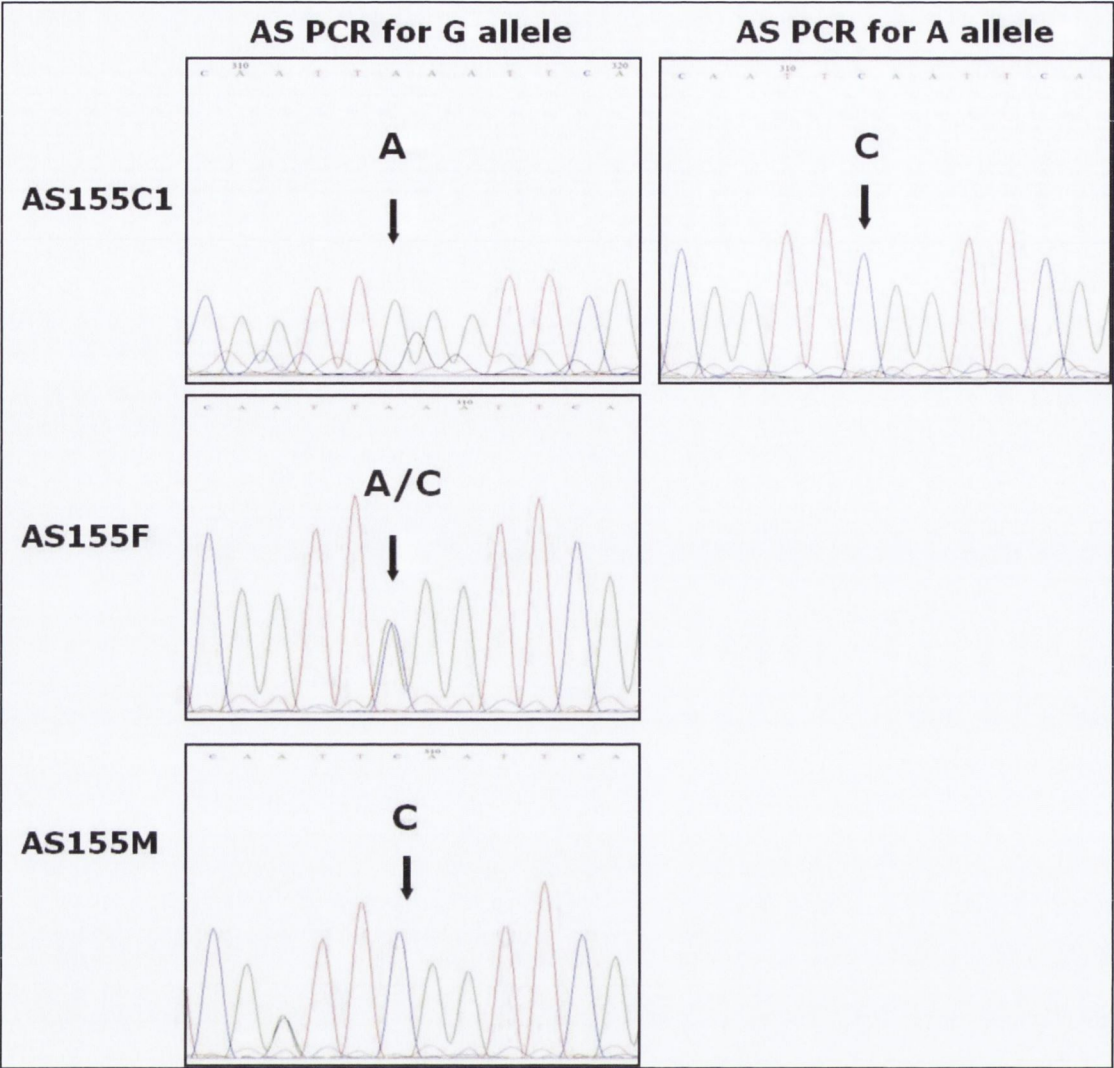


Figure 4.11: Electropherogram images of AS-PCR of *GRIN2B* *de novo* carrier and parents at the informative SNV Site.

Images on the left are the result of PCR and sequencing with the G allele (reference allele at *de novo* site) specific primer and images on the right are for the A allele (mutant) specific primer. In AS155C1, an A allele (paternally inherited) can be seen in AS-PCR and sequencing with the reference G specific primer and a C allele (maternally inherited) can be seen with the mutant A allele-specific primer. PCR with the G specific allele shows both A and C allele for AS155F and a C allele for AS155M. PCR with the mutant specific allele did not yield any results for AS155F and AS155M as they do not contain this allele.

4.3.2.2 *DISC1* De Novo LoF Variant Parent-of-Origin

At rs821616 the proband is heterozygous (T/A), the father is heterozygous (T/A) and the mother is homozygous for the A allele (Table 4.4). Therefore, AS084C1 must have inherited the T allele from the father and the A allele from the mother.

The AS-PCR primers were designed at rs821616 (A/T) (Table 2.19). Primers were designed with either a T or an A base at the 3' end of the forward primer, both with and without a mismatch at the penultimate base.

The results for *DISC1* can be found in Figure 4.12 and Table 4.5. PCR with the T specific primer in the proband, which amplified only the strand inherited from the father, showed the T deletion. PCR with the A allele-specific primer in the proband, which amplified only the strand inherited from the mother showed no deletion at the *de novo* site. This indicates that the *de novo* indel is co-inherited with the paternally inherited T allele at rs821616 and therefore, the mutation is derived from the paternally inherited strand. The reliability of the finding is supported by results from PCR and sequencing of the parental samples.

PCR of the mother's DNA with the T-allele-specific primer did not yield any product and PCR with the A allele gave rise to the reference T allele at the *de novo* site. PCR of the father's DNA with the A allele-specific primer gave rise to the reference T allele. However, sequencing with the T allele-specific primer gave rise to a very low level of the deletion. To ensure that this wasn't due to sample contamination, I made aliquots of DNA from raw sample stocks and repeated the PCR and sequencing. The same result was attained. This indicates possible mosaicism in the father and the LoF is not, in fact, *de novo*. Instead a small proportion of the father's cells may contain the mutation and AS084C1 arose from one of the father's cells that contain the mutation.

Table 4.4: Genotypes of <i>DISC1</i> SNVs used in AS-PCR.			
Samples	Status	Genotype at LoF variant site chr1: 232144803	Genotype at rs821616
AS084C1	Case	C/-	T/A
AS084F	Father	C/T	T/A
AS084M	Mother	C/T	A

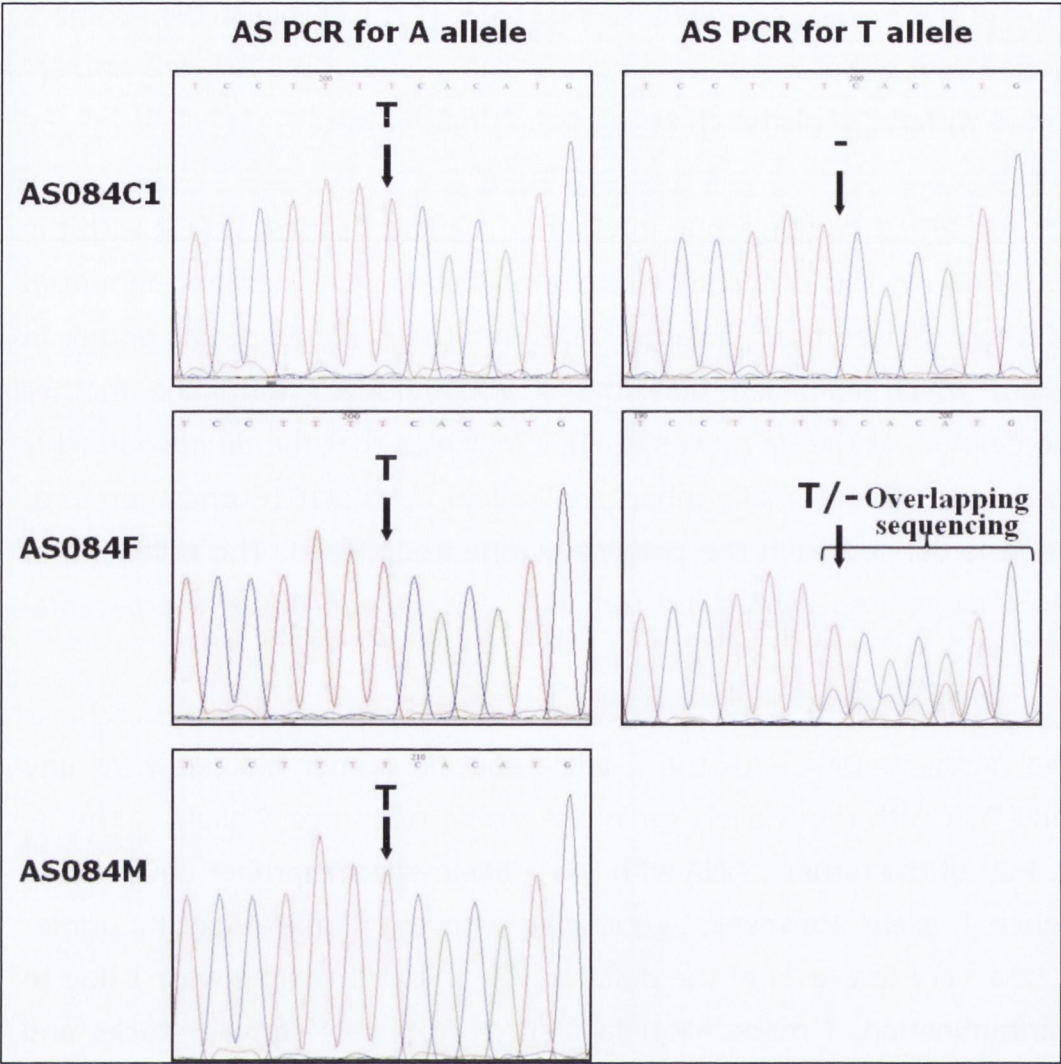


Figure 4.12: Electropherogram images of AS-PCR of *DISC1* *de novo* carrier and parents

Images on the left are the result of PCR and sequencing with the A allele-specific primer and images on the right with the T allele-specific primer. In AS084C1, the 1bp deletion can be seen in AS PCR with the paternally inherited T allele and not with the A allele. PCR with the A specific allele does not show the deletion in AS084F and AS084M. In PCR with the T allele, there is a low level of the deletion in AS084F. There is no image for AS084M with the T allele as the mother does not contain a T allele at the *de novo* site site.

Table 4.5: Alleles found at *DISC1* *de novo* SNV site by AS-PCR from rs821616

Samples	Status	PCR with T allele-specific primer	PCR with A allele-specific primer
AS084C1	Case	- (T deletion)	T
AS084F	Father	T/-	T
AS084M	Mother	T	(No product)

4.3.3 *In silico* Analysis of LoF variants

As per section 2.6.4, I annotated each of the LoF variants using variant effect predictor (http://.ensembl.org/Homo_sapiens/UserData/UploadVariations). This tool provides a list of protein coding transcripts containing the variant and the impact of the variant in each of these transcripts. To expand on this information, for each of the LoF variants I determined the total number of transcripts in the affected gene, the number of protein coding transcripts, the number containing the variant and the number containing the variants with a putative LoF affect (Table L in the Appendix). Most variants do not have a putative LoF affect in all protein transcripts of the gene. For example, *DLG5* has 12 transcripts, of which 4 are protein coding. Figure 4.13 provides a screenshot of the location of the variant in UCSC genome browser. The *DLG5* mutation, listed in Table 4.1, occurs in 3 of the 4 protein coding transcripts and has a putative LoF effect in just two. This variant is upstream relative to other transcripts (Figure 4.13). Of the transcripts containing the variant as a putative LoF variant, I selected the one corresponding to the largest protein and used this transcript to do a detailed *in silico* analysis, specifically outlining the impact of the variant on the protein (Table L in the Appendix).



Figure 4.13: Location of the nonsense LoF variant in DLG5 visualised in UCSC genome browser.

The black vertical line indicates the position of the nonsense mutation at chr10: 79614016 across the Ensembl transcripts of DLG5 (as depicted in red). A) The variant is located in just three transcripts. B) Magnified image: the variant is located within an exon in two transcripts and in an intron in one transcript. (genome.ucsc.edu/).

As per the MacArthur et al. study, “full” LoF variants are present in all protein-coding transcripts and “partial” LoF variants are present in only a proportion of protein coding transcripts (MacArthur et al. 2012). Seven variants were classed as full LoF and 22 as partial LoF variants. For example, *CNTNAP1* is a full LoF variant and *DLG5* is a partial LoF variant (Table L in the Appendix). Variants were not considered putative LoF variants if they were located in the last 5% of coding sequence due to the likelihood that they can be tolerated. There are three such variants; an indel in *DISC1*, an indel in *SLC4A4* and a stop loss mutation in *MBP* (Table 4.1).

A total of 28 variants are considered putative LoF variants from the entire study. All of these were novel and 25 were singleton events, occurring in just one individual and 3 occurred in multiple samples (Table 4.1).

4.4 LoF variant Association Analysis

A breakdown of the 28 LoF variants across all samples is given in Table 4.6. A carrier-based analysis of LoF variants was performed by comparing the number of samples that carried at least one LoF variant in the combined case sample versus controls. No individual sample carried more than 1 LoF variant. This was achieved using a 2x2 contingency table and a two-tailed Fisher’s exact test (<http://vassarstats.net/tab2x2.html>). There is an excess of LoF variants in the combined SZ and ASD cases compared to controls (29 in 420 cases v 8 in 287 controls; $p=0.016$, odds ratio (OR)=2.59, 95% confidence intervals (CI)= 1.10, 6.24) (Table 4.7), with the effect stronger for ASD (13 in 147 cases; $p=0.005$, OR=3.38, 95% CI=1.27, 9.17) (Table 4.9) than for SZ (16 in 273 cases; $p=0.074$, OR=2.17, 95% CI=0.86, 5.64) (Table 4.8).

Table 4.6: Summary of LoF variants.			
	SZ (n=273)	ASD (n=147)	Controls (n=287)
Total LoF Variants	16	13	8
Total Singleton Variants	14	9	2

Table 4.7: Carrier-based analysis of LoF variants in the combined case sample versus controls.

	Cases (n=420)	Controls (n=287)	P = 0.016 OR = 2.59 95% CI = 1.10, 6.24
Carriers	29	8	
Non-Carriers	391	279	

Table 4.8: Carrier-based analysis of LoF carriers in SZ cases versus controls

	SZ (n=273)	Controls (n=287)	P = 0.074 OR = 2.17 95% CI = 0.86, 5.64
Carriers	16	8	
Non-Carriers	257	279	

Table 4.9: Carrier-based analysis of LoF carriers in ASD cases versus controls

	ASD (n=147)	Controls (n=287)	P = 0.005 OR = 3.38 95% CI = 1.27, 9.17
Carriers	13	8	
Non-Carriers	134	279	

Of the 9 LoF variants in controls, 6 are low-frequency variants found in both case and control samples and may represent benign variants. The analysis was therefore restricted to singleton variants that occur in only one individual, which may have a greater functional impact as they are very rare. When the analysis is limited to singleton variants, the data show a more significant excess of LoF variants in the combined case sample versus controls (23 in 420 cases v 2 in 287 controls; $p=0.0007$, $OR=8.26$, 95% $CI=1.87, 51$) (Table 4.10) and the effect is similar for both SZ (14 in 273 cases; $p=0.002$, $OR=7.7$, 95% $CI=1.65, 49.5$) (Table 4.11) and ASD (9 in 147 cases; $p=0.001$, $OR=9.29$, 95% $CI=1.85, 63.2$) (Table 4.12).

Table 4.10: Carrier-based analysis of singleton LoF variants in the combined case sample versus controls.			
	Cases (n=420)	Controls (n=287)	P = 0.0007
Carriers	23	2	OR = 8.26
Non-Carriers	397	285	95% CI = 1.87, 51

Table 4.11: Carrier-based analysis of LoF singleton variants in SZ cases versus controls.			
	SZ (n=273)	Controls(n=287)	P = 0.002
Carriers	14	2	OR = 7.7
Non-Carriers	259	285	95% CI = 1.65, 49.5

Table 4.12: Carrier-based analysis of LoF singleton variants in ASD cases versus controls.			
	ASD (n=147)	Controls (n=287)	P = 0.001
Carriers	9	2	OR = 9.29
Non-Carriers	138	285	95% CI = 1.85, 63.2

I tested if there was a significant excess of rare LoF variants in the individual gene categories (Table 4.13). The Neurexins and Interacting Proteins grouping had the highest number of these variants (9 in cases (7xSZ and 2xASD) and 1 in controls, p=0.05 for SZ and ASD; p=0.03 for SZ; p=0.27 for ASD (all Fisher exact tests)). Results for all other gene categories were non-significant but the number of observations are small, e.g. for Post-synaptic Glutamate Receptor Complexes there was 5 LoF variants in SZ and ASD cases and 0 in controls (p=0.08; Fisher exact test).

Table 4.13: Carrier-based analysis of LoF variants in the individual gene categories in the combined case sample versus controls.			
Category	Case LoF carriers	Control LoF carriers	
1) NRXN1, interactors and related proteins	9	1	P = 0.05
2) DISC1 and interactors	2	0	P = 0.52
3) Glutamate Receptor Complexes	5	0	P = 0.08
4) Neural Cell Adhesion Molecules	12	7	P = 0.82
5) Functional and Positional Candidates	1	0	N/A*

*An association was not calculated for this category as genes within this category are not functionally related.

4.5 Functional Annotation of Missense Variants

This next section details my study of the missense variants identified in the sequencing study. With the enormous amount of missense SNVs identified, it is necessary to prioritise these in order to determine which SNVs are functionally relevant. Various methods have been developed to annotate SNVs according to how likely they are to be harmful to the protein. Some tools focus on sequence conservation and assume that sequences conserved throughout evolution are functionally important. Such tools search for sequence homologs, perform multiple alignments and score SNVs based on how highly conserved the region in which they reside is, e.g. the tool SIFT (Kumar, Henikoff, and Ng 2009). Others use structure based features to predict effect. Example methods include investigating the differences between the reference amino acid and that introduced by a SNV, or by mapping the location of a SNV onto the 3D structure of the protein to determine if it affects functionally significant regions, e.g. PolyPhen-2 (Adzhubei et al. 2010). I searched for tools to functionally annotate sequence variants using literature and internet searches. I assessed each tool identified based on ease of use (e.g. web-based, download required, Linux or Perl required), data input required (genomic coordinates or protein information), year released, number of citations, usefulness of output and underlying method employed. Six freely available annotation tools were selected for use. A list of these tools is provided in Table 4.14 along with the type of algorithm each tool employs to compute annotation scores.

Table 4.14: Tools used for annotation of SNVs				
Tool	Website	Usage	Research Article	Algorithm
ANNOVAR	http://www.openbioinformatics.org/annovar/	Perl required	(Wang, Li, and Hakonarson 2010)	Functional Elements
Massessor	http://mutationassessor.org/	Web-based	(Reva, Antipin, and Sander 2007)	Sequence conservation
MU2A	http://krauthammerlab.med.yale.edu/mu2a/	Web-based	(Garla et al. 2011)	Sequence conservation & structure
PolyPhen-2	http://genetics.bwh.harvard.edu/pph/	Web-based	(Adzhubei et al. 2010)	Sequence conservation & structure
SeattleSeq	http://gvs.gs.washington.edu/SeattleSeqAnnotation/	Web-based	No referenced publication	Combination of other tools
SIFT	http://sift.jcvi.org	Web-based	(Kumar, Henikoff, and Ng 2009)	Sequence conservation

I undertook my study of the annotation tools at a stage in the sequencing study where not all samples had been sequenced and analyzed. I had data from 219 SZ cases and 206 controls that included 2,085 SNVs. My plan was to use these data to understand the workings of the different analysis tools and then have an analysis pipeline in place for the final set of missense variants in the full sequencing dataset. The following sections detail the annotation tools that I investigated.

4.5.1 Annovar

Annovar is a command line run tool which requires PERL (Wang, Li, and Hakonarson 2010). A gene annotation database is downloaded from UCSC and stored locally. I downloaded the refseq annotation system. Annovar scans this gene annotation database and identifies genomics regions i.e. exon, UTR, introns etc. It scans annotated mRNA sequences to determine if exonic SNVs are synonymous or non-synonymous. For multiple alignment and score development, a dataset needs to be downloaded from UCSC. I downloaded "mce44way" which is a multiple alignment of 44 vertebrate species and includes measurements of evolutionary conservation from two methods: PhastCons and pyloP. These are based on multiple alignments of phylogenetically related species. Both scores have been calculated for different groups of organisms such as primates alone, placental mammals, and vertebrates. The number of species within the groups varies between genome builds.

PhastCons is based on a statistical model of sequence evolution called a phylogenetic hidden Markov model (phylo-HMM) that takes account of neighbouring bases, estimating the probability that each base belongs to a conserved element and it is sensitive to runs of conserved sites. It assigns each nucleotide position a score which represents the degree of conservation at that position. The PhastCons score is a probability score so it ranges from zero to one and most of the genome will have a score of zero or no score at all. PhastCons treats gaps in alignment as missing data, which can result in high conservation scores in regions of the alignment with many gaps (Pollard et al. 2010).

PhyloP calculates scores independently for each site and captures conservation as well as acceleration (faster than neutral rates of evolution under assumption of neutral drift). The overall phyloP score is the $-\log(\text{p-value})$ under a null hypothesis of neutral evolution and it can range from -14.08 to 6.42. A negative score indicates acceleration whereas a positive value indicates conservation (Pollard et al. 2010).

4.5.2 Mutation Assessor (Massessor)

Massessor computes a Functional Impact Score (FIS) based on information from protein family alignment grouped into aligned sets (family and subfamily) and takes advantage of 3D structure of homologous proteins (Reva, Antipin, and Sander 2007). It assumes that many different mutations have been tried at each position to a sufficient extent that the distribution of residues in homologs at that site reflects the constraints on these sites (Reva, Antipin, and Sander 2011). This assumes that evolutionary unfavourable residues are observed less often than neutral residues or critical residues while critical residues are highly conserved. It is on this basis that the observed frequencies are converted into estimates of a mutations functional impact. Each position is treated independently and a conservation score is generated for each residue. Variants are broken down into four different categories based on this score as per Table 4.15.

Table 4.15: Massessor Functional Impact Score categories				
Prediction	Neutral Impact	Low impact	Medium Impact	High impact
Functional Impact Score	<0.8	0.8 - 1.9	1.9 - 3.5	> 3.5

4.5.3 MU2A

MU2A reconciles the differences between the genome and transcriptome to predict effects of sequence variants (Garla et al. 2011). Variants are annotated with information on sequence conservation, gene function and effects on domain integrity. MU2A annotates mutations with respect to three types of information; 1) Conservation: conservation score from the 17 or 46-way phylogenetic alignment track of the UCSC genome browser (phyloP score); 2) Gene Function: includes GO annotations, Panther Pathways, Cancer Gene Census, Pfam Domain and 3) Missense mutation effects: includes Catalogue of

Somatic Mutations in Cancer (COSMIC), BLOSUM45, Panther cSNP analysis and LogR.E Score.

The BLOSUM (BLOcks of Amino Acid SUBstitution Matrix) matrix is a substitution matrix used for sequence alignment of proteins (Henikoff and Henikoff 1992). The BLOCKS database (<http://blocks.fhcrc.org/>; publicly available database that contains multiple alignments of conserved regions in protein families) is scanned for conserved regions of protein families and the relative frequencies of amino acids and calculates their substitution probabilities. A log odds score is then calculated for each of the possible 210 substitutions of the 20 amino acids. A more negative value implies a less likely and therefore, potentially deleterious substitution. BLOSUM45 is used for more divergent alignments and was used here.

The LogR.E score takes advantage of Pfam protein motif models and hidden Markov model (HMM) alignments to predict whether amino acid changes in conserved domains affect protein function. It is restricted to missense mutations and proteins characterised in Pfam. Values > 0.7 are predicted to be deleterious (Clifford et al. 2004).

Panther estimates the likelihood that an amino acid substitution (AAS) will affect protein's function in two ways. Firstly, by alignment of homologous protein sequences and statistics from HMMs to derive AAS scores (SubPSEC scores). Secondly it uses Ka/Ks ratios (ratio of the number of non-synonymous substitutions per non-synonymous site to the number of synonymous substitutions per synonymous site), which is a measure of selective pressure on a gene, to make comparisons at the gene level (Mi et al. 2007). The score is a probability that the mutation is deleterious and thus, more positive values indicate a more damaging variant. PhyloP has been described in section 4.5.1.

4.5.4 PolyPhen-2

PolyPhen-2 (PP2) uses 8 sequence based features and three structure based features (selected by an algorithm rather than manually), which mainly focus on comparing the wild type variant to the mutant. A multiple alignment is

carried out and the most informative features are based on; 1) how likely the substitution base was to occur at that site given the pattern of amino acids in homologous proteins; 2) how distantly related is the species that harbours the first deviation from the wild type human allele and 3) whether the substitution occurred at a hypermutable site. The functional importance of the AAS is predicted from its individual characteristics by a Bayes classifier. Variants are broken down into three functional categories based on the probability range that its PP2 scores falls within (Table 4.16).

Table 4.16: PolyPhen-2 probability score categories

Prediction	Benign	Possibly damaging	Probably damaging
PP2 probability	0 - 0.431	0.432 - 0.952	0.953 - 1

4.5.5 Seattle Seq

Seattle Seq does not compute a score of its own but provides others including a PhastCons score, a GERP score and a Grantham score. The genomic evolutionary rate profile (GERP) identifies constrained elements in multiple alignments by quantifying substitution deficits which are referred to as rejected substitutions (Cooper et al. 2005). These deficits represent substitutions that would have occurred if these DNA sequences were under neutral drift but didn't because the sites were under functional constraint. GERP estimates constraint for each alignment column; and if there is an aggregation of constrained columns, these are known as elements. Each element is scored based on the level of substitution deficit. The Grantham matrix predicts the likely effect of AAS based on differences between AA physiochemical properties such as composition, polarity and molecular volume (Grantham 1974). AAS are categorised into classes of increase physiochemical dissimilarity (Table 4.17). PhastCons was described in section 4.5.1 above.

Table 4.17: Grantham score categories

Prediction	Conservative	Moderately conservative	Moderately radical	Radical
Grantham Score	0-50	51-100	101-150	≥151

4.5.6 SIFT

Scale-invariant feature transform (SIFT) is an web accessible algorithm to predict how likely an AAS is to affect function (Kumar, Henikoff, and Ng 2009). The principle of SIFT is based on conservation of amino acids within protein families and assumes that sequences conserved throughout evolution are functionally important. There are two steps to SIFT. Firstly, a dataset of functionally related protein sequence are compiled using PSI-BLAST, which searches protein databases. An alignment is then built from homologous sequences along with the given sequence. In the next step, every position in the alignment is scanned and a probability calculated for all 20 amino acids at each site. The scores are normalised by the probability of the most common AA and recorded in a scaled probability matrix. SIFT predicts that AAS will be harmful if the scaled probability or SIFT score falls below a certain threshold. Highly conserved sites are intolerant to most substitutions and non-conserved sites are tolerant. A confidence value is also provided that is low if there is too little sequence diversity, i.e. if there are not many available homologs. Positions with normalized probabilities less than 0.5 are predicted to be deleterious and those greater than or equal to 0.5 are predicted to be tolerated (Table 4.18).

Table 4.18: SIFT score categories		
Prediction	Deleterious	Tolerated
SIFT score	≤0.5	≥0.5

4.5.7 Use of Annotation Tools

Each tool accepts text based input files, where each line corresponds to one variant. The specific input requirements are provided on the website and/or published article for each individual tool as listed in Table 4.19. Table 4.20 details the score(s) generated by each tool and type of information employed to formulate these scores. I inputted all 2,085 variants identified in this preliminary dataset into each of these tools.

Table 4.19: Annotation tool scores and information used to generate each

Tool	Score	Conservation		Protein Structure	Sequence annotation	Sequence prediction	Physiochemical properties
		Protein Sequence Alignment	DNA sequence alignment				
ANNOVAR	Annovar		YES		YES		
Massessor	FIS	YES		YES			
MU2A	Blosum45	YES					
MU2A	LogR.E	YES			YES		
MU2A	Panther	YES					
MU2A	PhyloP		YES				
PolyPhen-2	PP2	YES		YES	YES	YES	
Seattle Seq	GERP		YES				
Seattle Seq	PhastCons		YES				
Seattle Seq	Grantham						YES
SIFT	SIFT	YES					

Seattle Seq is a tool that not only provides annotation scores for variants but also classifies every single variant based on function, i.e. missense, silent etc. The other tools only provide this information on the variants for which scores were calculated. Therefore, all variants were assigned to functional categories based on the Seattle Seq analysis (Table 4.20). The vast majority of variants are silent or missense and the number is similar between the two categories, 47% and 51.6% respectively. After inputting my SNV dataset to all tools, a major concern was the high level of variants that were missing score data in the output files. I determined the rate of missing data within each functional category for each annotation score (Table 4.20).

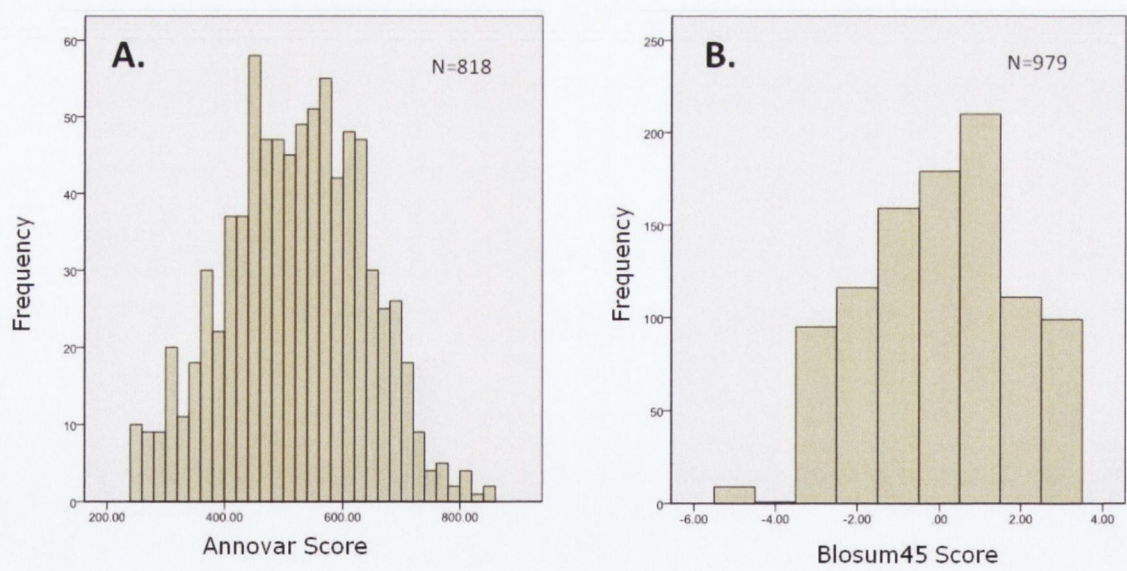
There is a lot of variability in terms of the rates of missing data across the different annotation scores. Two tools (LogR.E and Panther) had very high rates of missing data across all functional categories. This is because these tools generate annotation scores based on protein information and if the proteins containing the variants have not been annotated, scores cannot be provided. These were dropped from further analysis. For some tools, e.g. Grantham, Blosom45 and PP2, the rates of missing data are high because scores are only calculated for missense variants. However, as the primary objective here is the study of missense variants, data from these scores were retained for further analysis. A total of 9 annotation scores were investigated further.

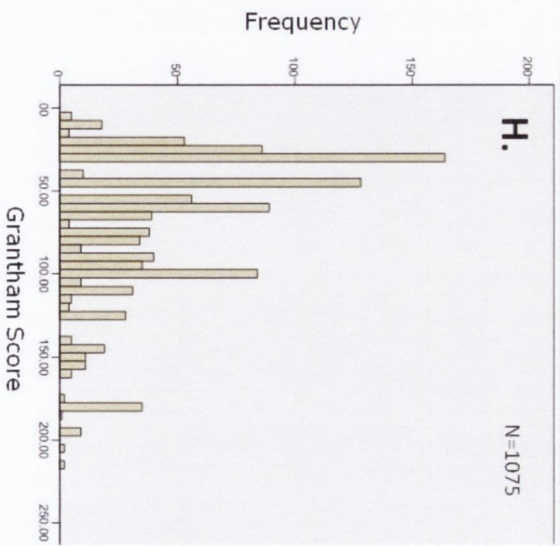
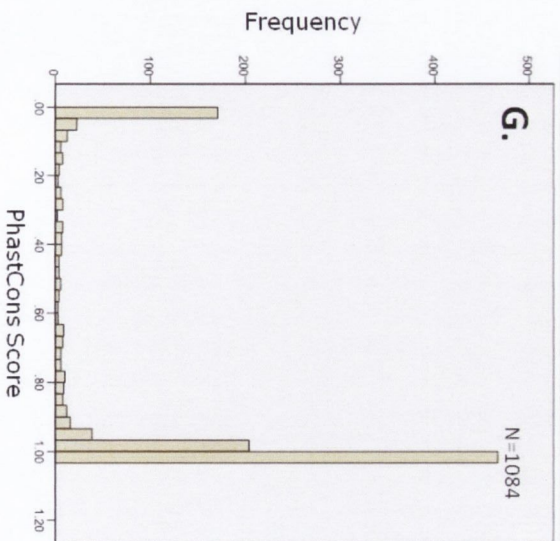
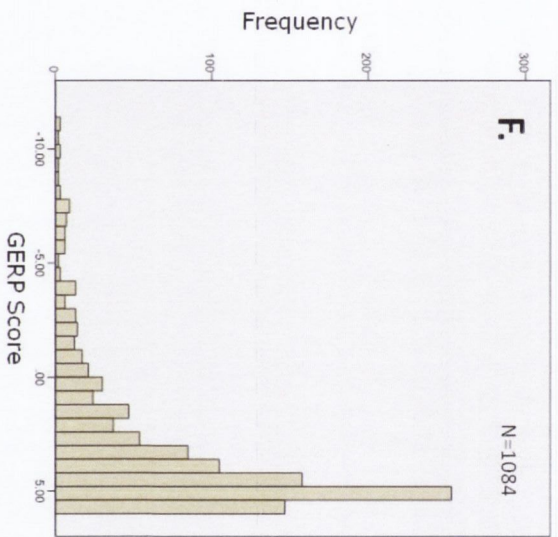
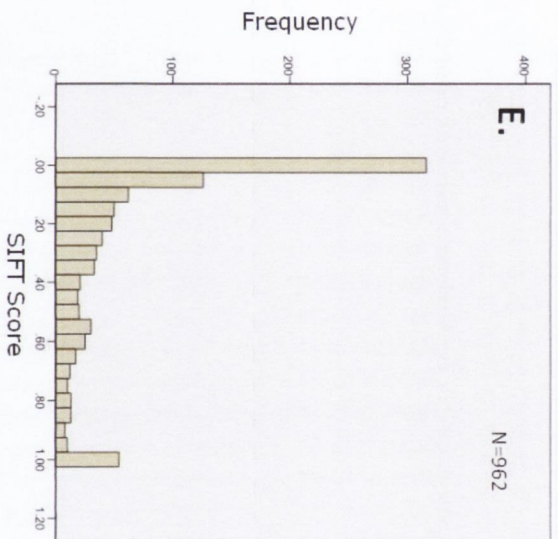
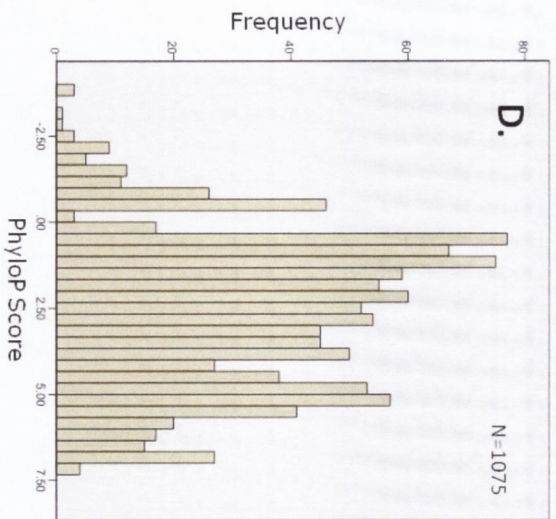
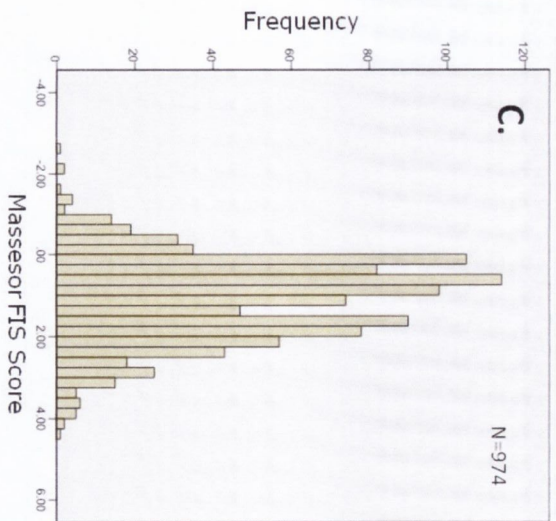
Table 4.20: Breakdown SNVs into functional categories and the rate of missing data in individual categories for each annotation score.

Tool	Score	Total	Non-coding/UTR	Silent	Missense	Nonsense
Overall		2,085	25	979	1075	9
Annovar	Annovar	585 (28.1%)	309 (31.6%)	10 (40%)	263 (24.5%)	3 (33.3%)
MASS	FIS	234 (11.2%)	25 (100%)	99 (10.1%)	101 (9.4%)	9 (100%)
Seattle Seq	GERP	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
	PhastCons	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
	Grantham	992 (47.6%)	16 (64%)	967 (98.8%)	0 (0%)	9 (100%)
MU2A	PhyloP	13 (0.6%)	0 (0%)	4 (0.4%)	9 (0.8%)	0 (0%)
	LogR.E	1,964 (94.2%)	24 (96%)	979 (100%)	952 (88.6%)	9 (100%)
	Panther	1,479 (70.9%)	20 (80%)	976 (99.7%)	474 (44.1%)	9 (100%)
	Blosum45	1,093 (52.4%)	13 (52%)	975 (99.6%)	103 (9.6%)	2 (22.2%)
SIFT	SIFT	230 (11%)	11 (44%)	97 (9.9%)	113 (10.5%)	9 (100%)
PolyPhen-2	PP2	1,081 (51.8%)	15 (60%)	976 (99.7%)	81 (7.5%)	9 (100%)

4.5.8 Investigation of Annotation Scores using SPSS

The purpose of this exercise was to select annotation tools to apply the final set of missense variants in the full sequencing dataset. Two scores were excluded due to high rates of missing data, leaving 9 scores. As it is not practical to employ 9 separate annotation scores, it was necessary to minimize these to a usable subset. One approach for doing this was to determine if some of the scores are effectively calculating the same thing. For example, if two tools, perhaps due to employing similar algorithms, assign variants very similar scores, it could render one of these tools redundant. To achieve this, I used the statistical software package SPSS to determine the level of correlation between scores (SPSS 19). In order to select the correct type of correlation method to employ, it was necessary to determine the distribution of variants across the different annotation scores. The data was plotted in SPSS for each annotation score, with the scores plotted on the X-axis and the frequency of variants on the Y-axis (Figure 4.14). Data is described as normally distributed if it forms a bell-shaped symmetrical curve when plotted. The data was normally distributed for Annovar scores and MASS FIS but the remainder were not normally distributed e.g. the majority of variants had very high or very low scores (i.e. skewed).





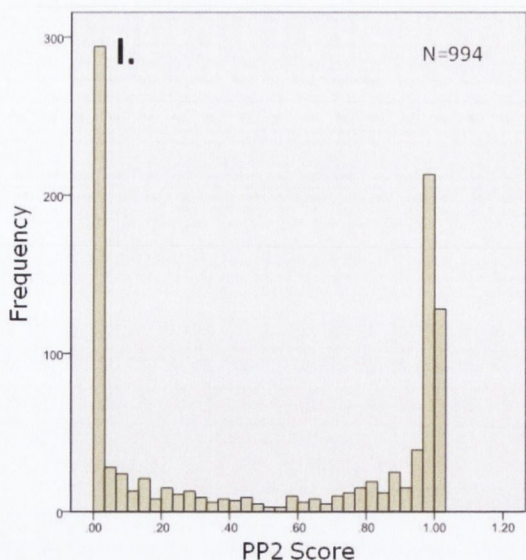


Figure 4.14: Plots of the distribution of missense variants across annotation scores

The data was plotted in SPSS for each annotation score with the annotation scores plotted on the X-axis and the frequency of variants on the Y-axis. A) = Annovar, B) = Blosum, C) = Massesor FIS, D) = PhyloP, E) = SIFT, F) = GERP score, G) = PhastCons, H) = Grantham Score and I) = PP2. N = number of variants scored.

4.5.8.1 Correlation

The statistics package SPSS was used to assess the level correlation between the sets of annotation scores. As the data is not normally distributed or does not have a defined distribution, i.e. is non-parametric, the Spearman’s correlation test was used. The level of correlation is denoted as the correlation coefficient (R), with a score of 1 indicating perfect correlation and a score of zero indicating no correlation. A list of the annotations scores for the missense variants was entered into SPSS. This data was used to generate a correlation matrix for all pairs of scores (Table 4.21). The numerical values are the level of correlation as denoted by the correlation coefficient R. The R values are colour coded based on the scheme in Table 4.22. Visualisation of the relationship between score pairs is necessary as some pairs may have a high R value but do not have an obvious linear relationship. Pairs of tools with an R value >0.5 in Table 4.21 were plotted in a simple scatter dot plot as shown in Figure 4.15. Blosum45 and Grantham do have a linear relationship (R = 0.823). GERP and phyloP are also well correlated (R = 0.801). Despite having a quite high R value of 0.536, phyloP and PhastCons do not have a linear relationship: the variants are concentrated at either extreme of the scores. This is also the case for GERP and PhastCons. Overall there is little correlation

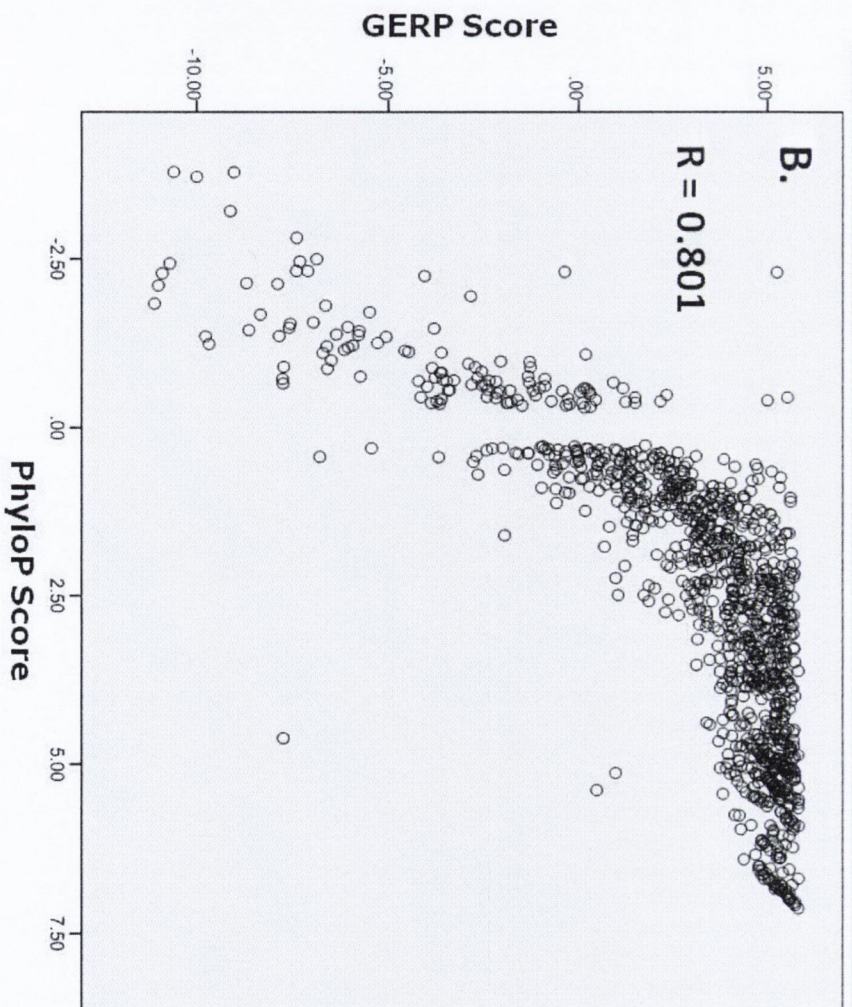
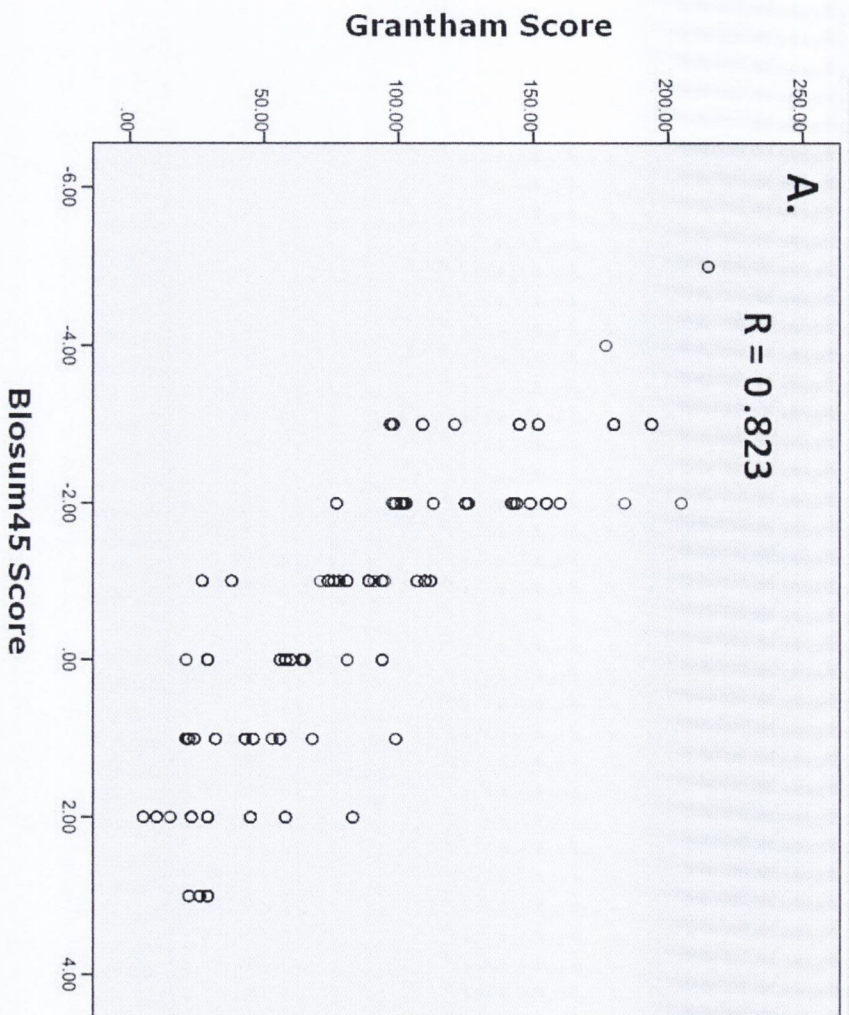
between tools making it viable to use any of these tools for annotation of SNVs. As it is impractical to use all annotation scores, I decided to proceed with the two tools that are most often mentioned in current literature as applied to rare variant analysis; SIFT and PolyPhen-2.

Table 4.21: Correlation between pairs of annotation scores calculated using Spearmans' rho in SPSS. Values are R

	Annovar	Blosum45	Mass FIS	MU2A PhyloP	SIFT	Sseq GERP	Sseq PhastCons	Sseq Grantham	PP2
Annovar									
Blosum45	0.075								
Mass FIS	0.079	-0.122							
MU2A PhyloP	0.128	-0.100	0.462						
SIFT	0.035	0.341	-0.422	-0.220					
Sseq GERP	0.181	-0.098	0.316	0.801	-0.157				
Sseq PhastCons	0.446	0.057	0.289	0.536	-0.062	0.501			
Sseq Grantham	-0.068	-0.823	0.113	0.096	-0.293	0.096	-0.013		
PP2	0.014	-0.293	0.492	0.483	-0.499	0.376	0.267	0.282	

Table 4.22: Heat map shading scheme for table 4.21

Colour	R value
	(-0.25, 0.25)
	(-0.5, -0.25) or (0.25, 0.5)
	(-0.75, -0.5) or (0.5, 0.75)
	(-1, -0.75) or (0.75, 1)



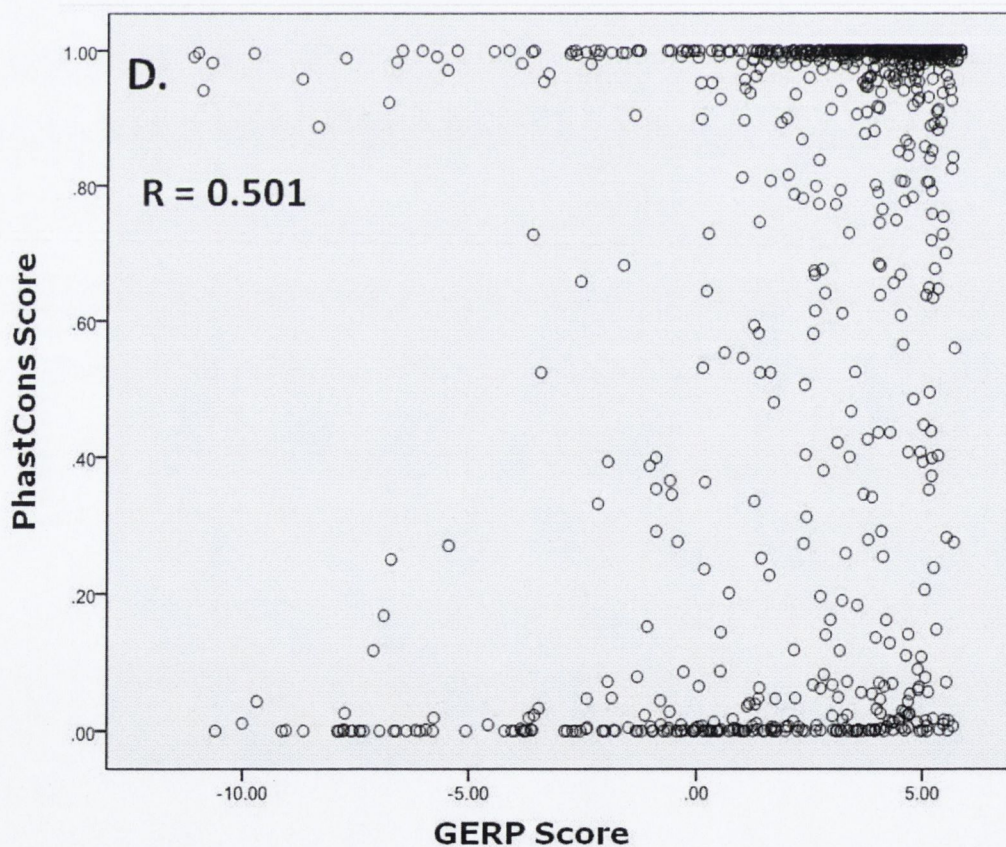
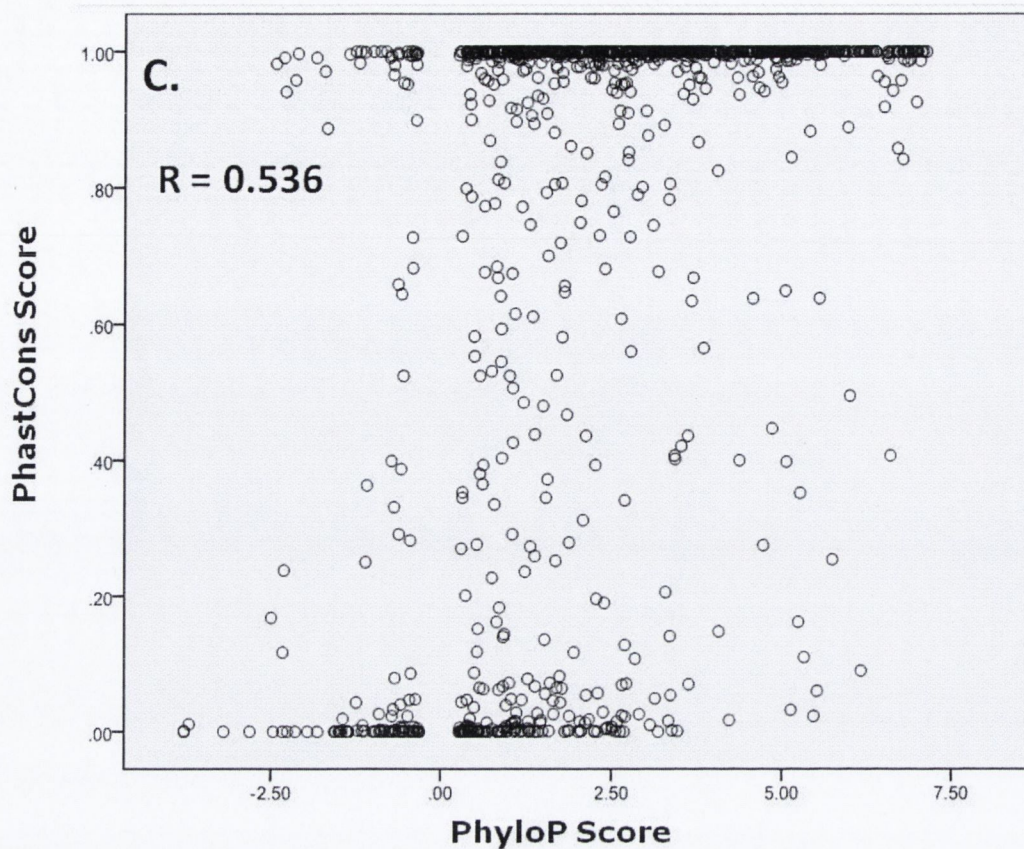


Figure 4.15: Correlation plots in annotation tools sets with R values ≥ 5

A) = Grantham vs. Blosom, B) = GERP vs. PhyloP, C) = PhastCons vs. PhyloP, and D) = PhastCons vs. GERP. R = correlation coefficient.

4.6 Missense Variant Association Analysis

At the time I undertook this analysis, various methods of analysing missense variants had been proposed in the literature but there was little consensus on the best method to use. One of the few published rare variants association studies was that by Johansen et al. (2010), which performed a sequencing study of candidate genes for hypertriglyceridemia (HTG), following the analysis method using that paper, I performed genotype based association analysis on the final dataset to determine if there is a significant excess of rare missense variants in cases versus controls. This was carried out on variants with >90% of calls in all individuals and an overall MAF of <1% in the combined case-control sample (Lemire 2011). The genotype based association analysis involved counting the number of non-reference and reference genotypes in cases and controls for a set of genes and inputting these numbers into a two-by-two contingency table for chi-squared analysis.

I performed association analysis on all 215 genes combined, and each of the four functional categories of genes based on the original gene selection. This was performed for all rare missense variants (Table 4.23), on rare missense variants that were classified as damaging by either SIFT or PP2 (termed one-hit damaging) and rare missense variants classified as damaging by both SIFT and PP2 (termed two-hit damaging). These analyses did not identify any significant excess of rare missense variants in any of the sets of genes tested.

Table 4.23: Genotype based association analysis of all rare variants using chi-squared test

	All rare missense	One-hit Damaging	Two-hit damaging
Test	P value		
215 Genes	0.7219	0.7502	0.6713
1) NRXN1, interactors and related proteins	1	0.1589	0.7454
2) DISC1 and interactors	0.2515	0.5115	0.903
3) Glutamate Receptor Complexes	0.9679	0.9599	0.8934
4) Neural Cell Adhesion Molecules	0.4606	0.5681	0.4286

4.7 Conclusion

The objective of this chapter was to investigate rare variants identified following sequencing of the SZ, ASD and control samples. Initially, I focussed on LoF variants that are predicted to severely disrupt protein-coding sequence. A total of 33 of these variants were identified, 31 of which validated. *In silico* analysis revealed that 3 of these variants were unlikely to be LoF as they are located in the last 5% of the coding sequence. There is a significant excess of cases in this dataset carrying these variants compared to the controls and this finding was stronger when limited to just singleton variants ($p=0.0007$, $OR=8.26$, $95\% CI=1.87, 51$).

As parental DNA was available for the ASD samples, it was possible to determine the origin of the LoF variants in these cases. Two variants were found to be *de novo*; in *GRIN2B* and *DISC1*. Parent-of-origin analysis revealed that the *GRIN2B* variant arose on the maternal chromosome. However, analysis of the *DISC1* variant identified possible paternal mosaicism and the variant was not *de novo*.

A preliminary dataset was used to investigate the tools available for annotation of rare missense variants. There was little consensus between these tools which is indicative of the fact that different scoring algorithms are employed and the difficulty in predicting the functional impact of SNVs. I

applied SIFT and PP2 to my association analysis of the final dataset of rare missense variants. This analysis did not identify a significant excess of rare missense variants in cases versus controls.

In summary, the most interesting results related to the novel LoF variants that have been identified in interesting genes in this study. But, understanding the impact of these variants requires further functional investigation at a cellular and molecular level. As it was impractical for me to investigate each variant, I investigated four genes for potential follow-up studies. These genes were; *INADL*, *GRIP1*, *FYN* and *GRIN2B*. The next chapter will focus on the reasons for selecting these genes and I will detail in full my functional studies.

5 Functional Investigation of a Loss-of-Function Variant in *FYN*

5.1 Aim

A number of LoF variants were identified in this study and elucidation of the impact of these variants on neuronal function requires further investigation. As it is impractical to investigate each variant, I selected a single variant for analysis following assessment of four of the LoF variant containing genes. Custom plasmids were designed to express green fluorescent protein (*GFP*) tagged wild type and mutant versions of the LoF gene. The plasmids were first transformed into *E.coli* to generate a sufficient amount of plasmid for transfection. Primary cortical neurons were transfected in suspension using the Amaxa system and after seven days in culture (7DIV) the cultures were fixed and stained. Immunohistochemistry was performed using antibodies to select for the LoF containing proteins and for markers of dendrites and of synaptic function. Expression and localisation of the wildtype and mutant proteins was analysed by confocal microscopy in order to study the impact of the LoF variant on dendrite projections and on the expression of markers for synaptic function.

5.2 LoF Variant Selection

In order to select variants for functional investigation, I first shortlisted the variants to singletons that were found in case samples only. I then excluded splice variants as the impact of these is more difficult to predict than frameshift indels and nonsense mutations that truncate the protein. I carried out a literature review of genes containing protein truncating variants to determine those with the best described roles in synaptic function. This produced a shortlist of; *INADL*, *GRIN2B*, *GRIP1* and *FYN*. Detail of this literary evidence supporting these genes is described in my discussion in chapter 6.

In order to determine the final size for the transfection plasmids, I calculated the cDNA size (excluding UTR) of these genes using Ensembl. For this analysis I used the same transcripts as those used in the previous *in silico* analysis, i.e. the largest transcript containing the LoF variant (Table 5.1).

5.2.1 Determination of Homology Between Human LoF Containing Transcripts with Mouse Transcripts

As these human genes are being transfected into mouse neuronal cells, it is necessary to determine if the mouse contains homologous genes to the LoF containing genes. The human gene and its corresponding protein must be similar enough to that of the mouse to have the same function in the murine cell. I obtained the corresponding protein sequences for the transcripts in Table 5.1.

Homologous mouse transcripts were identified in Ensembl based on the size of the transcript and of the size of the coding region (Table 5.1). The corresponding mouse protein sequences were also determined. The cDNA sequences (minus the UTR) and protein sequences were all obtained from Ensembl. Determination of the homology between mouse and human versions of the cDNA sequences and the protein sequences were achieved using the sequence alignment software T-Coffee (www.tcoffee.org/). Homology scores ranged from 98-100% indicating near identical human and mouse cDNA and protein sequences (Table 5.1).

As all four shortlisted genes had homologous genes and proteins in the mouse, this process did not eliminate any of the LoF genes. Transfection of large plasmids into neurons can be very difficult so I decided to use this as an additional factor in selection. Thus, I selected the LoF variant in *FYN* for functional investigation for two reasons; 1) for prior evidence of its role in synaptic function, as discussed below, and 2) for its small size, which would provide a higher chance of successful transfection of into mouse neurons. This nonsense variant truncates the protein from 537 to 155 amino acids.

FYN is a receptor tyrosine kinase located mainly in the post synapse that regulates the phosphorylation and trafficking of NMDARs, acting as a point of convergence for many signalling pathways regulating NMDARs (Hennah et al. 2009). Although *FYN* has been shown to phosphorylate both NR2A and NR2B receptor subunits (Suzuki and Okumura-Noji 1995), several lines of evidence suggest that *FYN* selectively modulates the function of NR2B-containing NMDARs (Hennah et al. 2009). *FYN*-mutant mice exhibit blunting of long term potentiation and impaired spatial learning, plus other neurological defects

including uncoordinated hippocampal architecture and reduced neural CAM-dependent neurite outgrowth (Grant et al. 1992; Beggs, Soriano, and Maness 1994). Also, *FYN* knockout mice have reduced number of dendrites compared to controls (Babus et al. 2011) and it may regulate spine formation via its interaction with SEMA3A (Morita et al. 2006). Use of a *FYN* transgene indicates that the gene is involved in the regulation of synaptic plasticity in a manner that is independent of its role in development (Kojima et al. 1997).

Increased expression of *FYN* has been reported in the prefrontal cortex of SZ patients compared to controls (McCarthy et al. 2009; Funk et al. 2012) and the latter study also reported increased phosphorylation of NR2B in the same SZ and control sample. However, Hattori et al. reported reduced expression of *FYN* in SZ in a study of platelet samples from SZ patients, first-degree relatives and controls (Hattori et al. 2009). Altered NRG1 (Neuregulin 1)-erbB4 signalling has been suggested to contribute to NMDA hypofunction in SZ (Hahn et al. 2006) and NRG1-erbB4 signalling stimulates NR2B phosphorylation through the activation of *FYN* (Bjarnadottir et al. 2007). Studies using *FYN*-deficient mice support a role for *FYN* in the induction of epilepsy (Cain et al. 1995) and data from *FYN*-transgenic mice indicates that this may be mediated via phosphorylation of NR2B (Kojima et al. 1997).

Table 5.1: Information on LoF variant transcripts shortlisted for potential functional investigation												
Gene	Type	Human Transcript	Wild Type Size		Mutant Size		Mouse Transcript	Wild Type Size		Homology Score ^c		Mouse Transcript Brain Expressed? ^d
			cDNA ^a (bp)	Protein (AA) ^b	cDNA ^a (bp)	Protein (AA) ^b		cDNA ^a (bp)	Protein (AA) ^b	cDNA	Protein	
INADL	Nonsense SNV	ENST00000371158	5406	1801	3840	1279	ENSMUST00000041284	5505	1834	99	99	Yes
INADL	Frameshift Indel	ENST00000371158	5406	1801	2160	719	ENSMUST00000041284	5505	1834	99	99	Yes
INADL	Nonsense SNV	ENST00000371158	5406	1801	3840	1279	ENSMUST00000041284	5505	1834	99	99	Yes
FYN	Nonsense SNV	ENST00000354650 ^e	1614	537	468	155	ENSMUST00000099967	1614	537	100	100	Yes
GRIN2B	Nonsense SNV	ENST00000279593	4455	1484	2133	710	ENSMUST00000111905	4449	1482	99	99	Yes
GRIP1	Nonsense SNV	ENST00000359742	3387	1128	447	148	ENSMUST00000138410	3384	1127	98	99	Yes
^a Excluding UTRs												
^b Amino acids												
^c As per T-Coffee (www.tcoffee.org/)												
^d Determined by "A Transcriptomic Atlas of Mouse Neocortical Layers" (http://genserv.anat.ox.ac.uk/layers)												
^e Transcripts used in plasmid design												

5.3 Plasmid Design

Plasmids containing the wildtype (WT) and mutant (MUT) versions of *FYN* were ordered from Genecopoeia (www.genecopoeia.com/). I provided the company with cDNA sequences of the WT and MUT versions of the human transcript. For the WT plasmid, I provided the sequence up to, but not including the stop codon. For the MUT plasmid, I provided the sequence up to, but not including the mutant stop codon. The plasmids were designed with *GFP* located upstream of the coding DNA inserts so that the resultant WT and MUT proteins would be coexpressed with *GFP*. The expression of *GFP* is used as a marker of successful transfection of the plasmid. The plasmids contain a very strong promoter, human cytomegalovirus (CMV) promoter, which maximises the expression of genes within the plasmids. Each plasmid has an ampicillin (AMP) gene, allowing selection of transformed *E.coli* cells in carbenicillin containing media (Figure 5.1). A third plasmid with just *GFP* was also ordered. This plasmid acts as a control to determine the affect of transfection on neurons. A schematic of all three plasmids along with their corresponding sizes can be found in Figure 5.2.

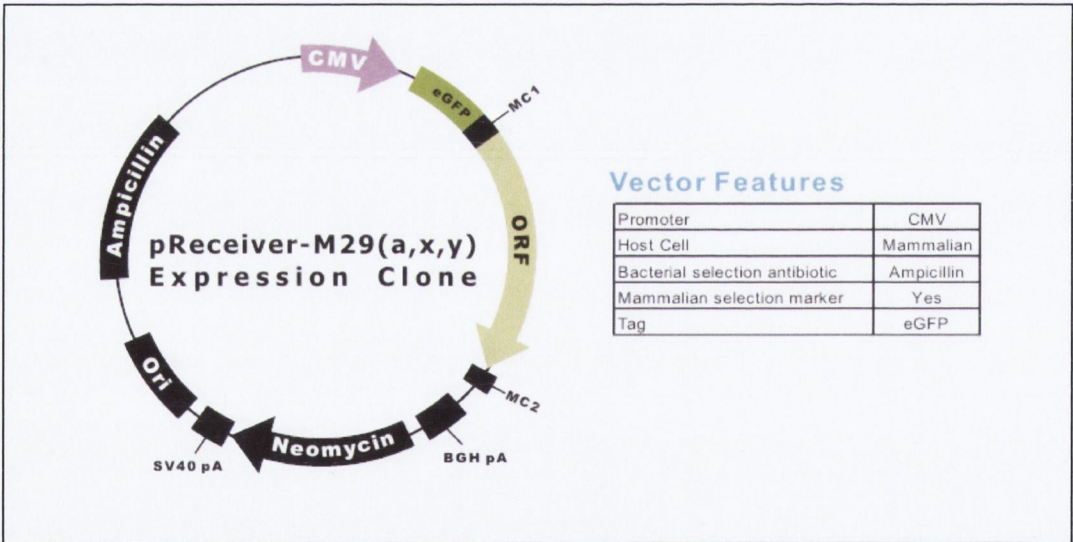


Figure 5.1: Expression clone with CMV promoter

Diagram of the plasmids supplied by Genecopoeia. Plasmids contain a CMV promoter. The insert is located at site marked ORF (open reading frame) in green and is *GFP* tagged. There are two selection markers; neomycin and ampicillin.

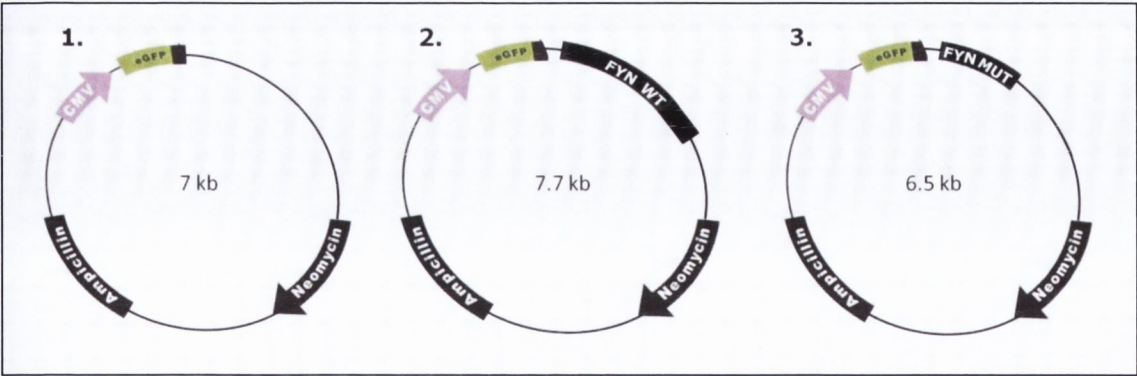


Figure 5.2: Schematic of the plasmids used in transfection

1. *GFP* plasmid 2. *FYN* WT (wildtype) plasmid tagged by *GFP*. 3. *FYN* MUT (mutant) plasmid tagged by *GFP*. The plasmid sizes are provided.

5.4 Determination of the Correct Plasmid Insert

Sequencing of plasmid inserts was performed to ensure the correct insert was present. As detailed in section 2.7.1, PCR primers were designed to amplify overlapping fragments of the inserts, because the *FYN* WT insert was too large to sequence in a single reaction. As the flanking insert sequence was not provided it was necessary to design primers within the insert sequence. For *FYN* WT, 3 primer sets were designed, one of which amplified the entire *FYN* MUT. Post-transformation purified plasmids underwent PCR and sequencing. Sequence was successfully generated for 1559bp of the 1611bp *FYN* WT insert and for 439bp of the 465bp *FYN* MUT insert. T-coffee software was used to determine homology between plasmid insert PCR sequence and the expected cDNA sequence. There were no discrepancies between the sequences for *FYN* WT or MUT inserts and those expected.

5.5 Impact on *FYN* Expression and Dendrite Length

The 3 plasmids were transfected into cortical neuronal cells from newborn mice as per section 2.7.3 and immunohistochemistry was performed as per section 2.7.4. As a control, cells were also processed for transfection, but in the absence of the plasmid. Four markers were used in parallel in the first set of analyses: DAPI, *GFP*, *FYN* and *MAP2*. An example confocal image is provided in Figure 5.3. DAPI stains nucleic acid and allows detection of both neuronal cells and glial cells (Figure 5.3A). The *GFP* signal acts as an indicator of transfected neurons as each plasmid contains a *GFP* gene (Figure 5.3B).

FYN protein expression and localisation can be determined with a *FYN* antibody (Figure 5.3C). *MAP2* acts as a marker of dendritic projections from neuronal cells (Figure 5.3D). It is primarily expressed in neurons and thus can be used to distinguish glia from neurons. Dendrites are the primary site of excitatory synaptic transmission and their number and morphology play an important role in synaptic plasticity (Verpelli et al. 2010).

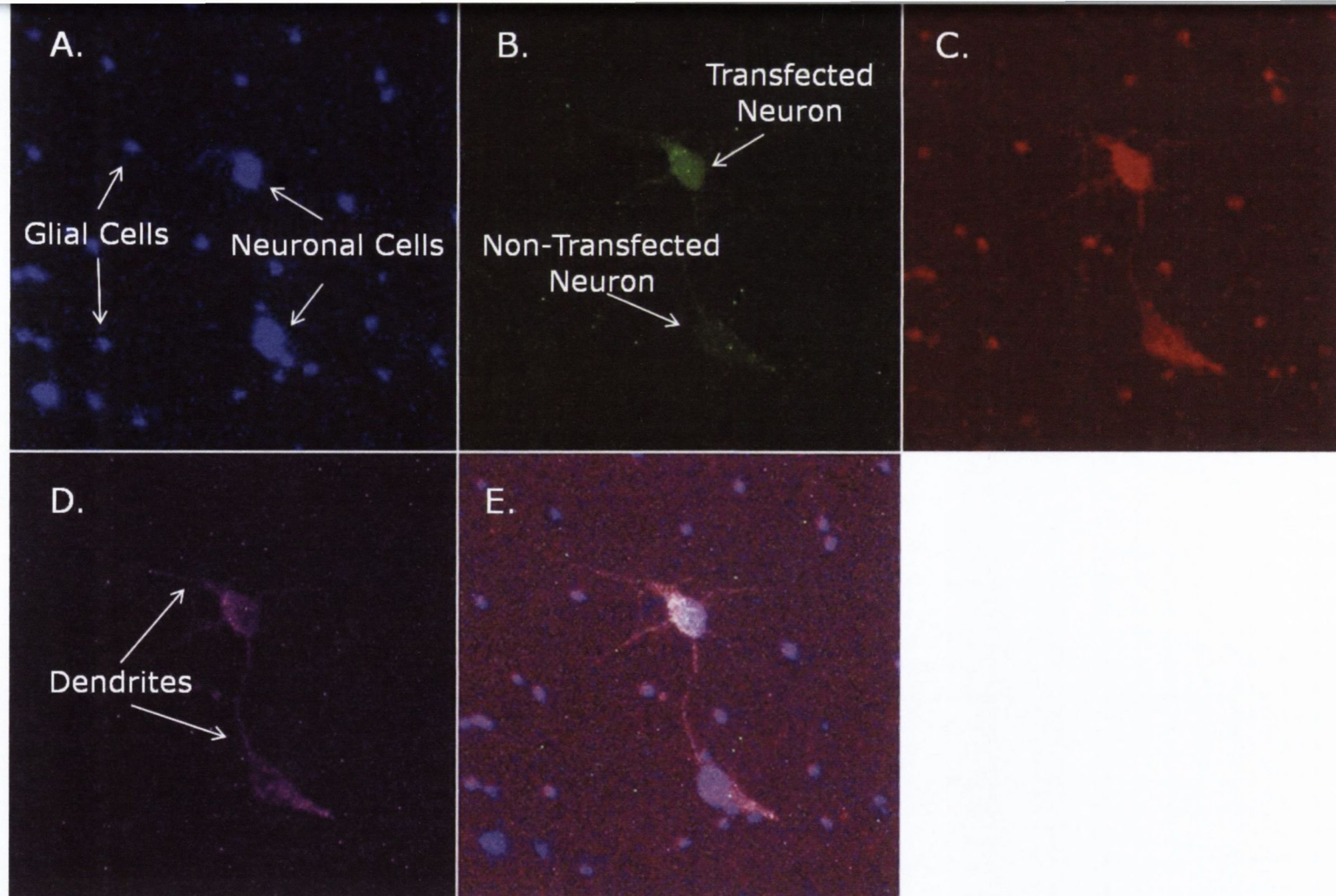


Figure 5.3: Confocal Image of mouse cortical cells stained with A. DAPI, B. GFP C. FYN and D. MAP2 and E. an overlap of all four stains.

A) Neuronal and glial cell can be visualised with DAPI staining. B. Transfected neuronal cells can be distinguished from non-transfected cells upon staining with *GFP*. C. FYN expression within neurons can be visualised using an antibody that binds to *FYN*. D. Dendrites can be detected upon staining with MAP2. E. All four stains can be observed together.

Numerous confocal images, as in Figure 5.3, were captured following transfections of the mouse neuronal cells with each plasmid. Analysis of these images was performed using Image J software (as outlined in section 2.7.5). A number of measures were taken for each neuron identified. The intensity of the green *GFP* signal was measured in each neuron along with a measure of the background signal. Also, the intensity of the red *FYN* signal was measured in each neuron along with a measure of the background signal. The length of the dendrites was measured using the purple *MAP2* signal.

As *GFP* is not naturally expressed in mouse, it can act as an indicator of successful transfection. Transfection can occur at a very low level in neurons, in which the case the *GFP* signal would be very low. Also, low *GFP* signal can sometimes be detected in non-transfected cells due to interference with a very intense *FYN* signal. In order to determine the level of *GFP* signal noise, a control group of cells were processed for transfection in the absence of a plasmid. The *GFP* signal from these cells was used to set the laser parameters at a base level of *GFP*, so that any higher *GFP* signal would correspond to real transfection signal. In addition, the background signal of *GFP* was also taken into account. In order to distinguish transfected versus non transfected cells in every field, both the intensity values of the cellular staining and the background were acquired, and a neuron was classified as transfected if it had a *GFP* intensity of at least 3 times greater than that of the background.

All measures were taken in transfected and non-transfected neurons. The non-transfected neurons are informative as they indicate the endogenous or basal levels of the proteins of interest in neurons. The intensity of *FYN* in both transfected and non-transfected neurons for the *GFP*, the *FYN* WT and *FYN* MUT plasmids was determined. The *GFP* plasmid also acts as a control for the affect of transfection on neurons.

The purpose of the analysis was to determine; 1) if the *FYN* WT or MUT plasmids result in an increased expression of *FYN* compared to the *GFP* plasmid, 2) if *FYN* MUT plasmid results in a different level of *FYN* expression to *FYN* WT and 3) if *FYN* WT or MUT have an impact on dendrite length.

I calculated the average level of *FYN* expression for both transfected and non-transfected neurons for each plasmid group (Figure 5.4). In order to determine if there is a significant difference in *FYN* expression I performed statistical analysis. As it cannot be determined if the data is normally distributed due to low sample number, I used the Kruskal-Wallis non-parametric test. This test compares between the medians of two or more samples to determine if the samples originate from the same distribution. Firstly, I compared the non-transfected neurons from each of the three groups and attained a p-value of 0.693. This indicates that there is no difference between these sets of neurons as expected, as all should have the same basal levels of *FYN*.

Upon comparison of *GFP* transfected (*GFP-T*), *FYN*-WT transfected (*FYN* WT-T) and *FYN* MUT transfected (*FYN* MUT-T), a near significant p-value of 0.055 was attained. Comparison of *GFP-T* to *FYN* WT-T does not give a significant result ($p=0.079$), despite what appears to be an obvious difference on the bar chart. This may be due to very small sample numbers for *FYN* WT-T, where data was only available for two neurons. Since plasmid size affects the efficiency of transfection, there are more neurons transfected for *GFP* and mutant *FYN* than for *FYN*-WT. There is a significant difference between *GFP-T* and *FYN* MUT-T ($p\text{-value} = 0.035$). Thus, there is an increased expression of *FYN* following transfection with the MUT version of the gene. However, there is no significant difference in *FYN* expression between *FYN*-WT and *FYN*-MUT ($p = 0.77$). This indicates that the mutation does not affect expression of the gene. Again, a caveat to remember is the small number of neurons that were successfully transfected for the *FYN*-WT plasmid. Figure 5.5 provides example confocal images of *FYN* expression in neurons transfected with the 3 different plasmids.

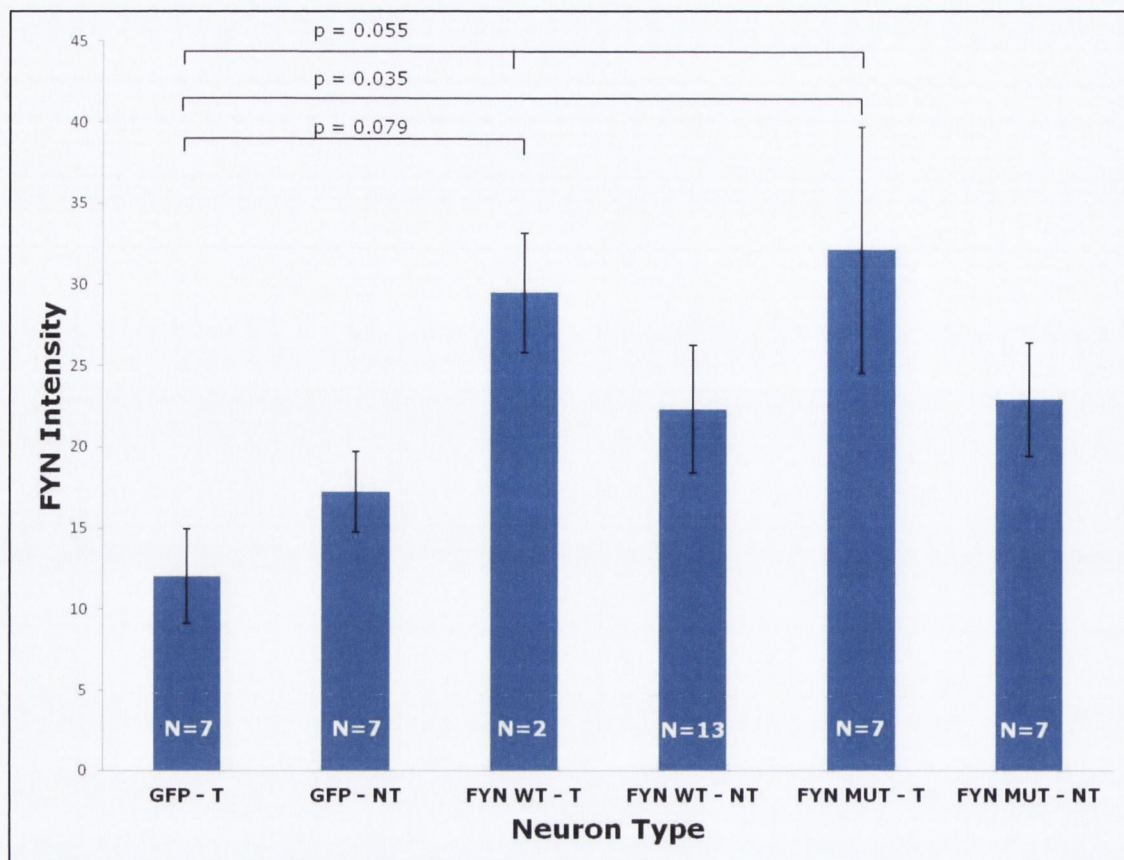


Figure 5.4: Expression in *FYN* in transfected and non transfected neurons following attempted transfection with *GFP*, *FYN* WT and *FYN* MUT plasmids.

T = transfected neurons, NT = non-transfected neurons. N = number of neurons used to calculate the average *FYN* intensity. P values were attained using Kruskal Wallis test. Significant and near significant observations only are shown.

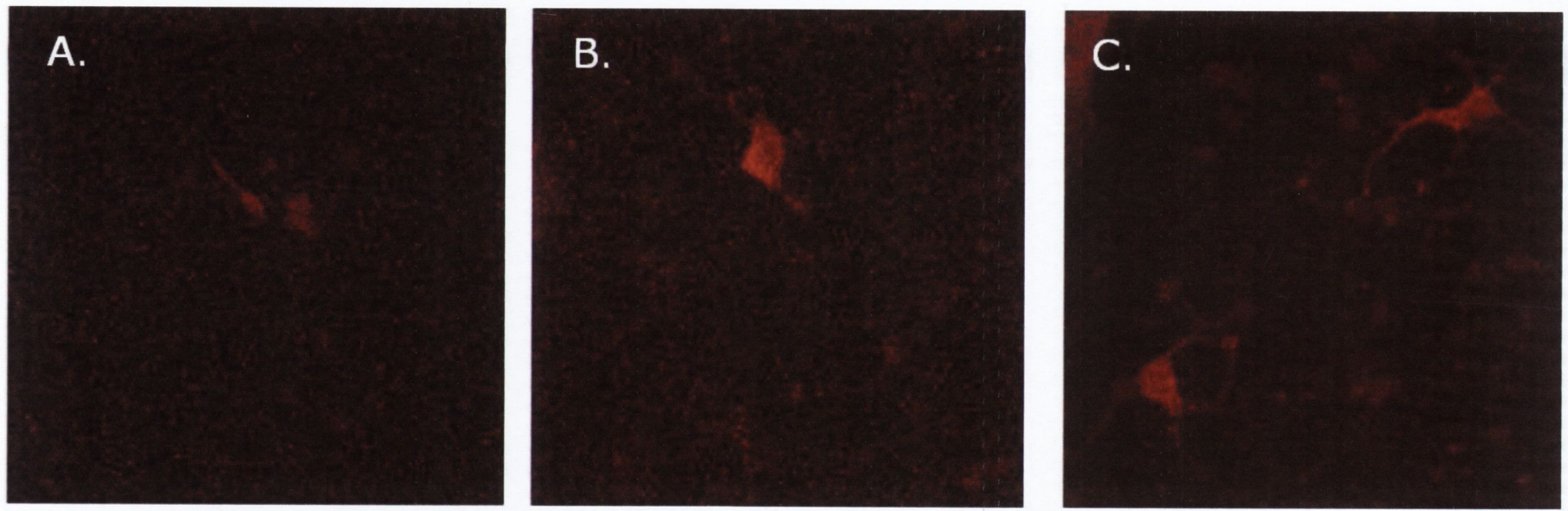


Figure 5.5: *FYN* Expression in neurons transfected with A. *GFP* B. *FYN* WT and C. *FYN* MUT

I also investigated the impact of *FYN*-WT and *FYN*-MUT on dendrite length. For any neurons that contained dendrites, I measured the lengths using the *MAP2* intensity (Figure 5.6). Firstly, I compared the non-transfected neurons in each group and a significant difference was observed ($p = 0.044$). This result was surprising as these neurons should have comparable dendrite length as all are non-transfected. The difference is driven by the long dendrites in *GFP*-NT but is it worth noting that this group has substantially fewer NT neurons than the other groups. In general, there are shorter dendrites in the transfected compared to the non-transfected cells for each group, indicating that transfection may impact neurite length. However, the only significant finding for this is for *FYN* MUT ($p = 0.005$). This result suggests that expression of mutant *FYN* may impact on dendrite length in the cells. Figure 5.7 provides example confocal images of dendrite length in neurons transfected with the 3 different plasmids.

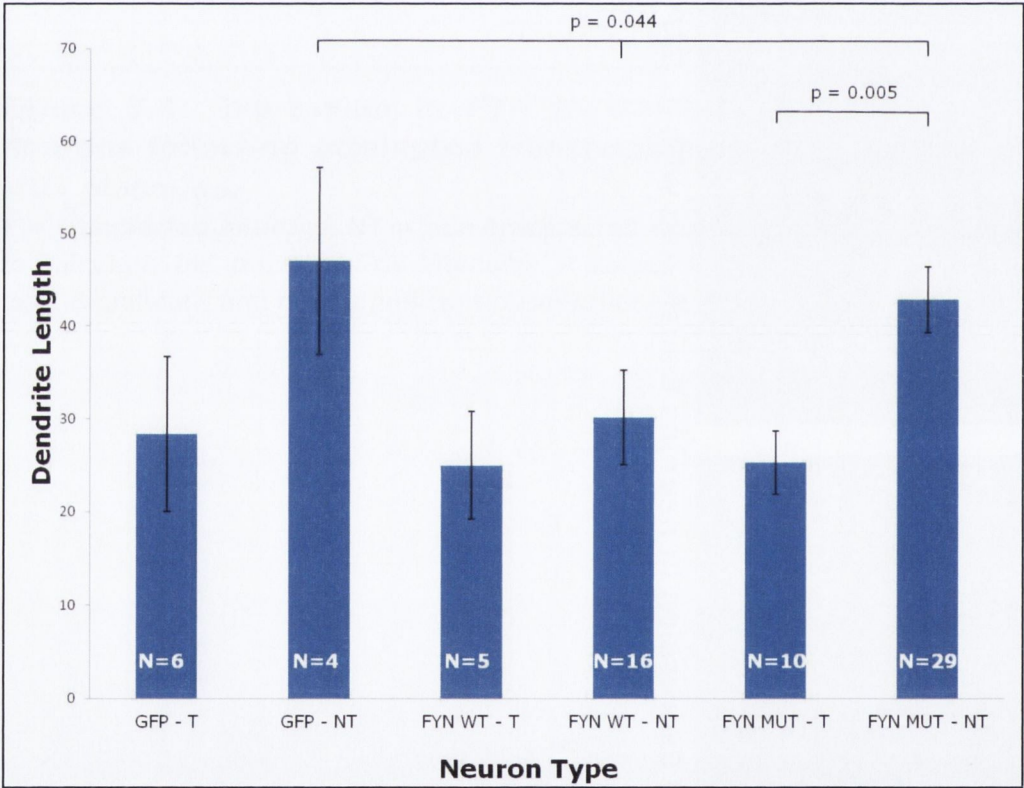


Figure 5.6: Dendrite length in transfected and non transfected neurons following attempted transfection with *GFP*, *FYN* WT and *FYN* MUT plasmids.

T = transfected neurons, NT = non-transfected neurons. N= number of neurons used to calculate the average dendrite length. P values were attained using Kruskal Wallis test. Significant observations only are shown.

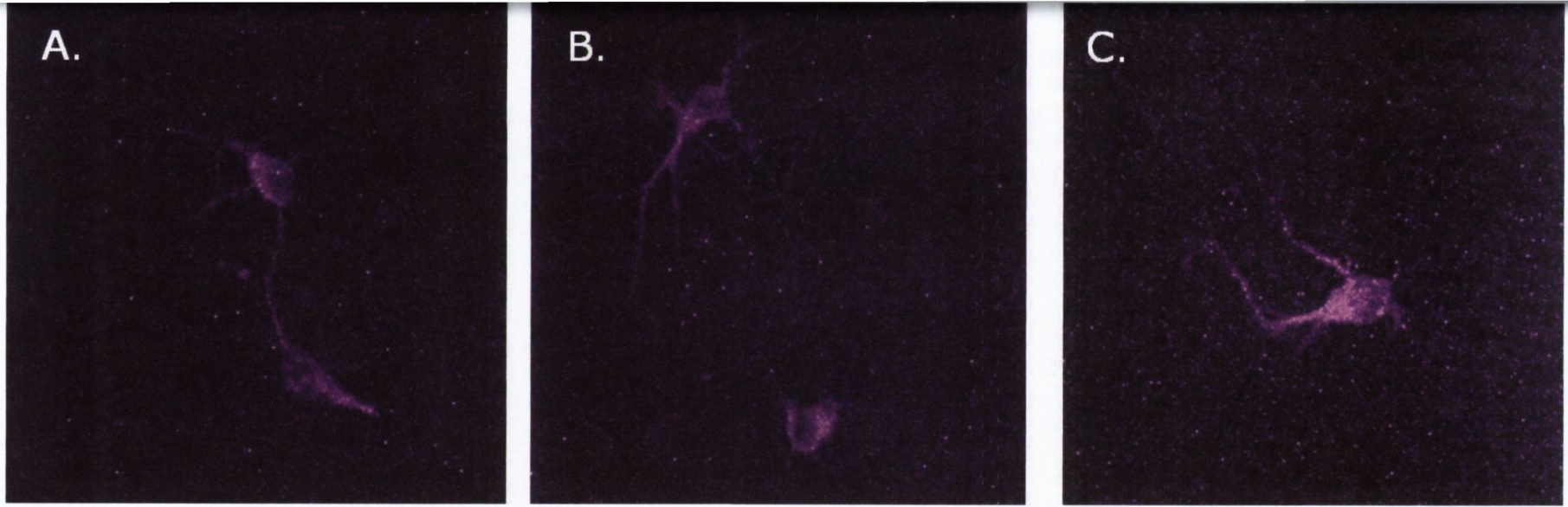


Figure 5.7: Dendrite Length in neurons transfected with A. *GFP* B. *FYN* WT and C. *FYN* MUT

5.6 Impact on Synapse Function

Data from the previous section indicates that *FYN* WT and MUT plasmids result in a higher rate of *FYN* expression and that the MUT *FYN* potentially impacts on dendrite length. I proceeded to investigate the affect of transfection on synapse function. Immunohistochemistry was performed with DAPI, *GFP* and synapsin following transfection with *FYN* WT and *FYN* MUT. Synapsin is a phosphoprotein that coats synaptic vesicles and is involved in the binding between these vesicles and the cytoskeleton (Nomura et al. 2000). It acts as a marker for synapses. DAPI is also used and it stains nuclei. There was no *GFP* plasmid assessed in this analysis. An example image is provided in Figure 5.8. Again, *GFP* intensity comparison to the background was used to distinguish transfected from non-transfected neurons. I initially set the threshold of $\geq 3X$ background as before.

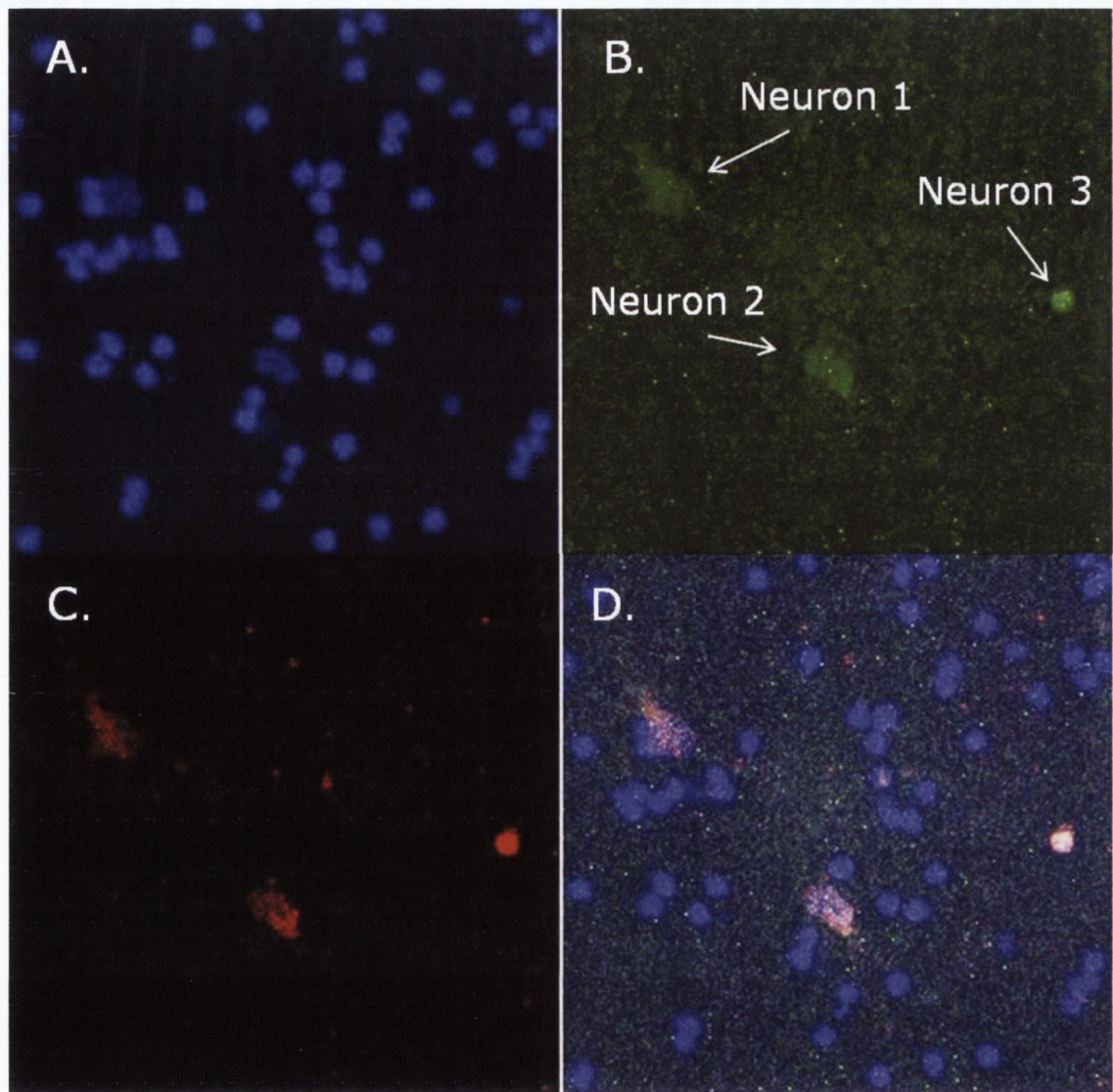


Figure 5.8: Confocal image of mouse prefrontal cortex cells stained with A. DAPI, B. GFP C. Synapsin and D. an overlap of all four stains.

A) Neuronal and glial cell can be visualised with DAPI staining. B. Transfected neurons can be distinguished from non-transfected neurons upon observing staining with *GFP*. C. Synapsin staining indicates the presence of synapses. D. All three stains can be observed together.

Analysis was carried out on the intensity of synapsin in transfected and non-transfected neurons in *FYN*-WT and *FYN*-MUT transfected neurons (Figure 5.9). In this analysis just one *FYN*-WT and two *FYN*-MUT neurons are classed as transfected. It is possible that these thresholds are too stringent as the background signal for *GFP* in these sets of slides is very high (Figure 5.8B). As described above there is an initial adjustment of the lasers to control for background *GFP*, so this is a second level of control. The threshold is arbitrary and may differ across different set of experiments. Upon visual inspection of the neurons in this set of images it was decided that this threshold may be too stringent. For example, in Figure 5.8B only neuron 3 is considered transfected. Therefore, I decided to set a lower threshold of *GFP* intensity of $\geq 2X$ the background for transfection. For example, in Figure 5.8B all three neurons are now classified as transfected with this lower threshold. I also carried out the analysis of synapsin expression with the lower threshold (Figure 5.10). The pattern is very similar with both thresholds (Figure 5.9 and Figure 5.10). The *FYN* WT-T and *FYN* WT-NT, and *FYN* MUT-NT are very similar, with a much higher intensity of synapsin in *FYN* MUT-T, although no significant results were attained due to low sample numbers.

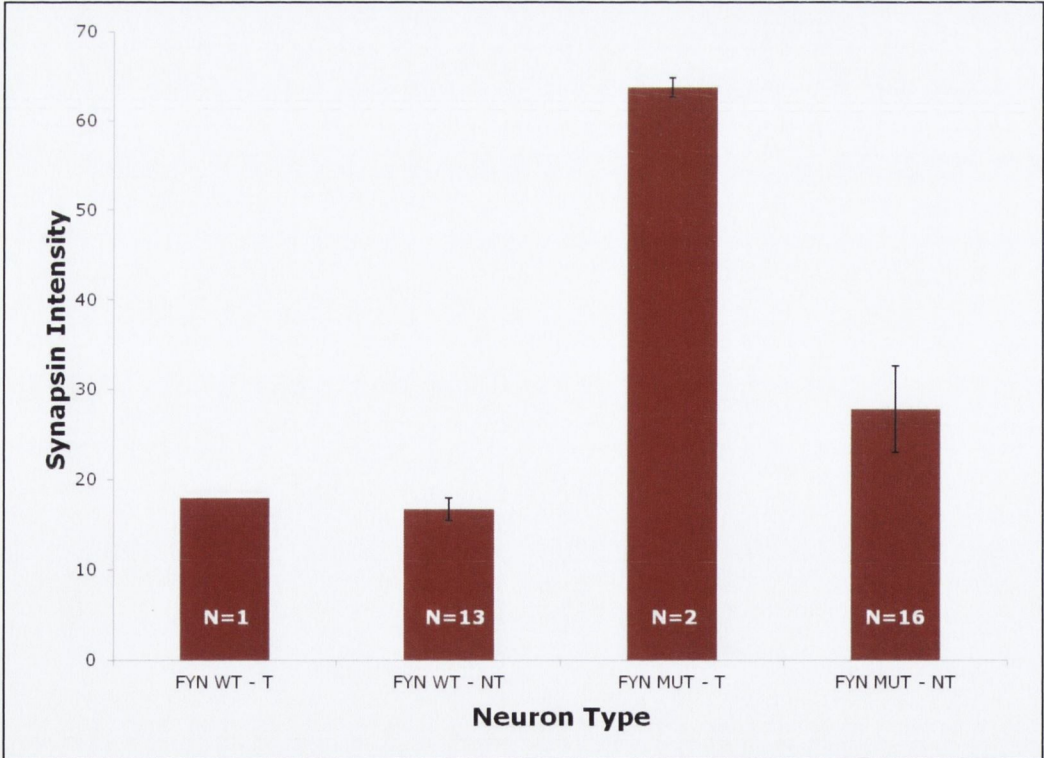


Figure 5.9: Synapsin intensity with more stringent (*GFP* intensity in neuron $\geq 3X$ background) transfection threshold.
No significant findings were attained using Kruskal Wallis test.

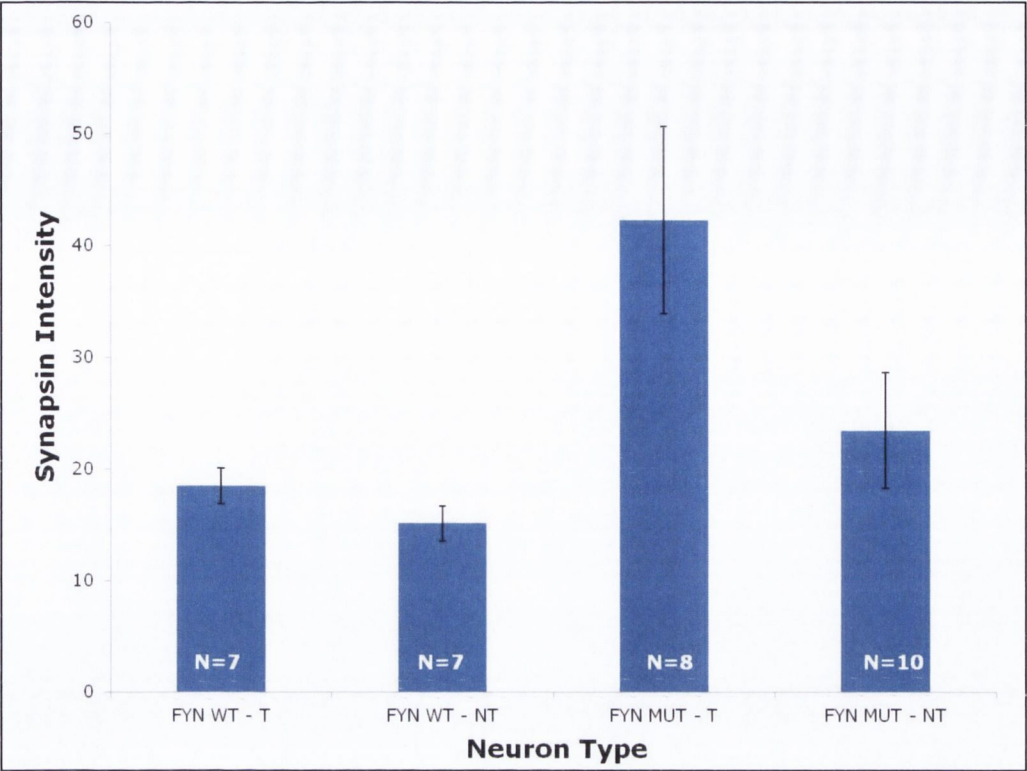


Figure 5.10: Synapsin intensity with less stringent (*GFP* intensity in neuron $\geq 2\times$ background) transfection threshold.

No significant findings were attained using Kruskal Wallis test.

The localisation of synapsin was determined by measuring its intensity within the dendrites. With the more stringent criteria for transfection, the there are just two *FYN* WT-T and no *FYN* MUT-T dendrites (Figure 5.11). Therefore, I just focused on the less stringent transfection threshold where more data was available (Figure 5.12). There is a difference between *FYN* WT-NT and *FYN* MUT-NT ($p = 0.037$), which should be comparable. There is also a significant difference between *FYN* WT-T and *FYN* MUT-T ($p = 0.045$), but as the non-transfected neurons are not comparable no firm conclusion can be drawn.

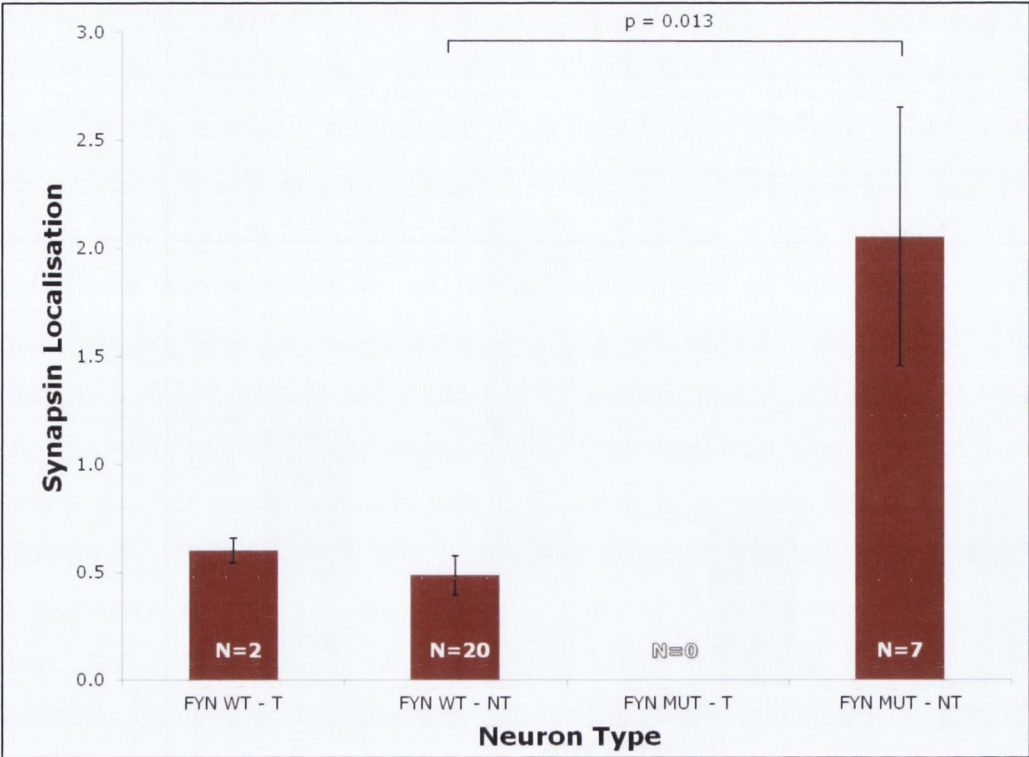


Figure 5.11: Localisation of synapsin measured by determining intensity of synapse signal in dendrites with more stringent (*GFP* intensity in neuron $\geq 3X$ background) transfection threshold.
P values were attained using Kruskal Wallis test. Significant observations only are shown.

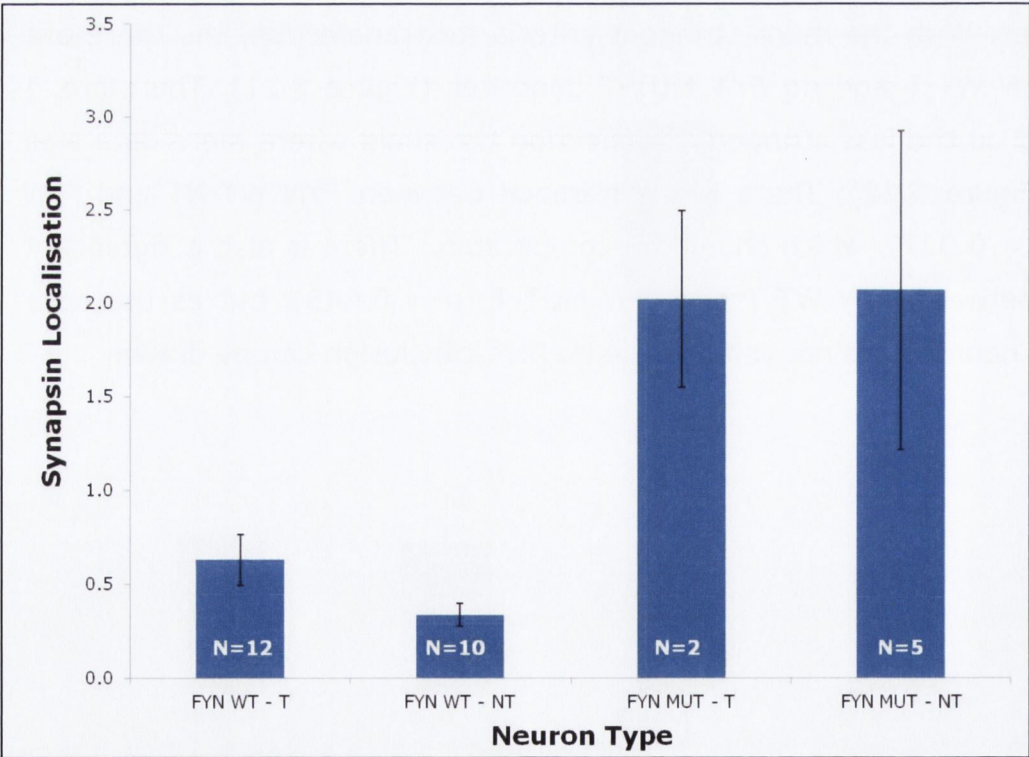


Figure 5.12: Localisation of synapsin measured by determining intensity of synapse signal in dendrites with less stringent (*GFP* intensity in neuron $\geq 2X$ background) transfection threshold.
P values were only calculated where greater 2 sample numbers were involved. And were calculated using Kruskal Wallis test, yielding no significant findings.

Upon inspection of the images it was noted that there were a high number of small round-shaped neurons with a high intensity of *GFP* and synapsin in the *FYN* MUT (Figure 5.13). These neurons resemble glial cells in morphology but as synapsin is only expressed in neurons under normal conditions, these cells are most likely neurons. However, to be certain these experiments need to be repeated with a *MAP2* marker. It may be that the very intense *FYN* signal is interfering with and exaggerating the *GFP* signal. It is also possible that the intense synapsin signal is due to restriction of synapsin to the cell bodies as it may not be transported along the dendrites.

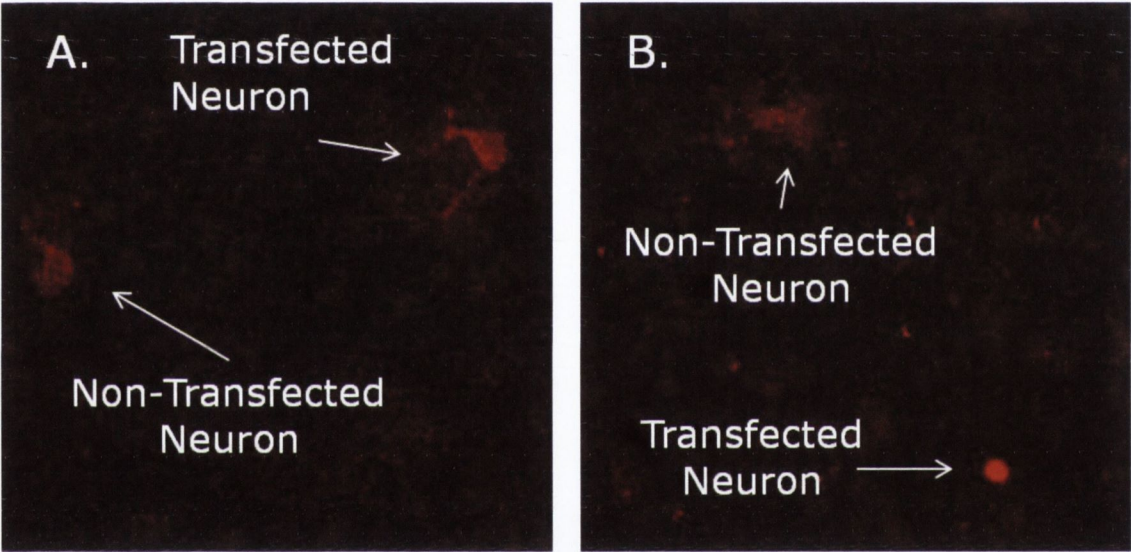


Figure 5.13: Synapsin localisation in neurons transfected with A. *FYN* WT and B. *FYN* MUT

A transfected and non-transfected neurons can be observed in A. and B. The transfected neurons in *FYN* MUT have a different morphology to the other neurons. Neurons were classed as transfected based on the lower threshold of 2X background *GFP* intensity.

5.7 Conclusion

Elucidation of the impact of LoF variants requires functional investigation. I selected a LoF variant for investigation based on 1) strong prior evidence of its role in synaptic function and 2) its small size which should make the task of transfecting the gene into mouse neurons a less difficult task. I selected a nonsense mutation in the *FYN* gene that results in truncation of more than half of the protein.

Confocal microscopy image analysis following transfection with the WT and MUT version of *FYN* and with a control *GFP* plasmid, revealed an increased expression of *FYN* following transfection with both the WT and MUT versions of the gene. However, there was no significant difference in *FYN* expression between *FYN*-WT and *FYN*-MUT. This indicates that the LoF variant does not affect expression of the gene.

Nonetheless, there is an observable decrease in dendrite length in the transfected neurons compared to the non-transfected cells for each group and this was significant for *FYN*-MUT ($p = 0.005$). This suggests that expression of mutant *FYN* may impact on dendrite length in the cells.

Investigation of the affect of the variant on synaptic function did not reveal a significant finding related to synapsin intensity or localisation. However, it is worth noting that there a far fewer dendrites with synapsin expression in *FYN*-MUT ($n = 7$) compared to *FYN*-WT ($n = 22$), despite a similar number of neurons (*FYN*-MUT $n = 18$; *FYN*-WT $n = 14$). In *FYN*-MUT neurons that do contain synapsin within dendrites, it tends to have travelled further in the dendrites, although this is based on very small numbers.

Inspection of the images revealed *FYN*-MUT neurons with an unusual morphology. It is possible that synapsin is not transported properly along the dendrites and is instead restricted to the cell bodies.

This study is preliminary and in addition to the requirement for many more sample numbers to perform accurate statistical analyses, a number of control measures are required to ensure that these results are reliable. In particular, verification of both the plasmids and the antibodies should be performed. The entire insert sequences should be sequenced along with a part of the plasmid sequence flanking the insert, if possible. Plasmid length and identity could be verified by restriction enzyme digest. The plasmids contain many restriction sites and both a single and double digest could be performed, followed by analysis of the restriction site pattern by agarose gel electrophoresis. The specificity of the *FYN* antibody should be validated and this could be achieved by applying the antibody to human cell lines followed by Western blot analysis to determine if the *FYN* protein is specifically targeted. Also, it is necessary to

confirm that the plasmids are expressing *FYN* by transfecting human cell lines and again performing a Western Blot.

6 Discussion

6.1 Review of the Field at the Beginning of the Project

SZ and ASD are complex neurodevelopmental disorders and the underlying pathogenesis of each is poorly understood. Both have a very strong genetic component but the genetic architecture of them remains to be elucidated. At the point this project began neither linkage studies nor candidate genes studies had identified many robust findings in either disorder, and GWAS had come to the forefront of genetic studies of complex disorders. GWAS examine million of loci across the genome to identify common variants of small effect associated with disease. They rely on the premise that at least a proportion of the underlying heritability of the disorder under study is due to common variants of small effect. For SZ, a study published by the ISC suggested that a substantial proportion of variance in genetic susceptibility was polygenic and due to many variants of small effect (International Schizophrenia Consortium et al. 2009). A few loci with genome-wide level significance and evidence of independent replication had been identified for SZ including *ZNF804A* (O'Donovan et al. 2008), *TCF4* (Stefansson et al. 2009) and the Major Histocompatibility Complex (MHC) region on chromosome 6 (Stefansson et al. 2009; Shi et al. 2011; International Schizophrenia Consortium et al. 2009). Although some GWAS hits were identified for ASD, few stood up to replication. The lack of success of GWAS left a huge challenge to explain the missing heritability of these disorders and put a dent in the common disease common variant (CDCV) hypothesis. It was suggested that either sample sizes were underpowered to detect the small effect sizes of the many common risk alleles in the population given stringent significance levels of GWAS, or that the causal variants were too rare to be captured by GWAS arrays either directly or via LD with any directly genotyped variants. Rare causal variants are an attractive explanation as the low fecundity of individuals affected with ASD and SZ puts very strong selection pressure on causal variants (Doherty, O'Donovan, and Owen 2012).

The unexpected ability to use GWAS data to genotype CNVs led to a surge in the investigation of these variants in complex disorders genetics, specifically neuropsychiatric disorders. Large but rare chromosomal deletions (100Kb – 3Mb) were found to occur more frequently in SZ cases compared to healthy

controls (International Schizophrenia Consortium 2008) and specific loci known as deletion regions were consistently identified e.g. 1q21.1, 15q11.2, 15q13.3 and 22q11.2 (Stefansson et al. 2008; International Schizophrenia Consortium 2008). This was in parallel with the discovery of an increased rate of CNVs in individuals with autism (Sebat et al. 2007), including a number of recurrent CNVs (e.g. 1q21.1, 15q11-13 and 16p11.2) (Weiss et al. 2008; Glessner et al. 2009; Pinto et al. 2010). Discovery of rare CNVs in these disorders along with the modest findings from GWAS focussed attention on rare variants with a large effect. This coincided with the emergence of next generation sequencing (NGS) technologies which permitted the routine study of rare variants. Together this prompted a new area of research in the field of psychiatric genetics that concentrated on the search for rare high penetrant risk variants.

As described in detail in the introduction (Section 1.3 and 1.7), there is substantial phenotypic and genetic overlap between SZ and ASD, indicating that they may share a common pathogenic mechanism. Common clinical features include impairment and deficits in social functioning, whilst imaging studies support neural dysconnectivity in both disorders. The strongest evidence of shared aetiology comes from genetic studies, especially the identification of rare CNVs found in both disorders, e.g. 1q21.1 (International Schizophrenia Consortium 2008; Mefford et al. 2008), 15q11.2 (Hogart et al. 2010; Stefansson et al. 2008), 15q13.3 (International Schizophrenia Consortium 2008; Miller et al. 2009), 16p11.2 (McCarthy et al. 2009; Weiss et al. 2008) and 17q12 (Kirov et al. 2009). In particular, a large number of risk variants related to synaptic function and glutamate signalling have been identified e.g. CNVs affecting *NRXN1*.

This study has combined SZ and ASD and moved beyond GWAS and CNVs studies to investigate rare pathogenic mutations in the form of small indels and SNVs detected using NGS. My focus was on genes involved in neurodevelopment and in particular on genes involved in synaptic functioning and glutamatergic signalling.

6.2 Review of this Study

6.2.1 Challenges for Genes Selection

I put a systematic approach in place to select genes for inclusion in the study. There was a very specific focus for gene selection that included five separate biological categories; 1) *NRXN1*, interactors and related proteins (n = 46), 2) *DISC1* and interactors (n = 23), 3) Postsynaptic Glutamate Receptor Complexes (n = 58), 4) Neural cell adhesion molecules (n = 61), and 5) Functional and Positional Candidates (n = 27). Despite this narrow focus, there were still a huge number of genes that qualified for selection. For example, investigation of the *NRXN1* interactors and related proteins highlighted more than 80 genes, yet only half of those were included in the study. Also, *DISC1* is a hub protein for numerous cellular processes, so unsurprisingly 116 of its interactors were identified, but this was limited to 23 genes for inclusion in the study. Prioritisation was achieved by focussing on genes with evidence for involvement in neurodevelopment. I availed of information from a study by Laumonnier et al. that devised a comprehensive list of proteins located in the postsynaptic proteome (PSP), and the structural and functional complex to which they belong, totalling approximately 1,200 genes. To select genes within the glutamate receptor complexes, I only included the 58 genes encoding structural proteins. Thus, selection of the 215 genes was a very subjective process and there were undoubtedly risk genes that were not included.

6.2.2 Protocol Modifications

This was a large scale project and the protocol that was in place for sample prep at the start of the project was laborious and expensive. Thus, it was necessary for me to improve it. A number of modifications were successfully introduced. The first was the replacement of the single-tube processing format to a 96-well format, which dramatically reduced the amount of time and labour required. Secondly, post target enrichment PCR modifications following analysis of a pilot set of samples lead to a significant reduction in the level of duplicate sequence reads. This allowed a greater number of samples to be processed, significantly improving study power and reducing project costs.

6.2.3 Rare LoF Variant Analysis

Sequencing the exons of 215 genes in the combined dataset of SZ cases, ASD cases and controls identified 3,708 variants that passed quality control filters. As the majority of samples were part of previous GWA studies, it was possible to compare GWAS SNV calls with SNVs calls from this study. The concordance rate was greater than 99%, indicating very a high specificity for variant calls in the sequence data. A large proportion of these variant were rare variants (2,861) of which 33 were novel LoF variants, which are predicted to severely disrupt protein-coding sequence. All LoF variants were subject to validation by Sanger sequencing and all but two indels were successfully validated. Indels are difficult to call in NGS data and a very high proportion was removed from this dataset during quality control (QC). These variants still warrant attention as they have a large propensity to be LoF when they arise in coding sequence. Of the 31 validated LoF variants identified in this study, 12 were indels. The QC measures put in place here were very conservative and there may be more LoF indels that did not progress to validation using Sanger sequencing. It may be possible to identify these variants at a later date using new and improved indel calling software.

A significant excess of SZ and ASD cases in this dataset carry rare LoF variants compared to the controls ($p=0.02$), with the excess greater for ASD cases ($p=0.005$) than for SZ ($p=0.07$). The selection of a $MAF<0.1$ as a frequency cut-off for rare variants is arbitrary; not all variants above this threshold will be benign and not all variants below this threshold will be pathogenic. But highly pathogenic variants are likely to be rare or even unique. Therefore, to focus on variants that may be most deleterious, the association analysis was also performed on singleton variants. There was an even greater excess of these singleton LoF variants in the combined case sample ($p=0.0007$) and the excess was present in both ASD ($p=0.001$) and SZ ($p=0.002$).

Upon testing the individual gene categories I observed a significant excess ($p=0.05$) of variants in the *NRXN1*-related and interacting proteins category (category 1), where 7 of 9 LoF variants in cases are in SZ samples. This lends further support for disruption of *NRXN1* pathways in the pathogenesis of SZ and ASD. The results for the other gene categories were non-significant, e.g.

for Post-synaptic Glutamate Receptor Complexes there was 5 LoF variants in SZ+ASD cases and 0 in controls ($p=0.08$). However, with such few observations and small sample numbers, this study was underpowered to detect association.

Support for these findings could be attained by performing replication of this association analysis in independent datasets. Whole exome case-control datasets for SZ and ASD would be required as there are no other studies that would have performed target enrichment on the same set of genes. From the exome data, the exons of these 215 genes could be specifically investigated for the presence of LoF variants. Members of my research group are currently in collaboration with other international researchers to perform such replication analysis.

During the time period of my studies, new CNV studies have lent further support for abnormal synapse formation and maintenance in the pathogenesis of SZ (Malhotra and Sebat 2012; Kirov et al. 2012) and ASD (Gilman et al. 2011). Kirov et al. noted that *de novo* CNVs were enriched for components of the NMDAR and for components of the PSD (Kirov et al. 2012). Interestingly, they identified multiple *de novo* CNVs affecting genes that encode members of the DLG (discs large) family of membrane-associated guanylate kinases (MAGUKs) and this study identified a nonsense mutation in *DLG5* in a SZ case. Further evidence that synapse biology is important for these neurodevelopmental disorders is provided by SNP-based group/pathway analysis in SZ (Lips et al. 2012; O'Dushlaine et al. 2011), transcriptomic analysis of the brain in SZ (Mudge et al. 2008) and ASD (Voineagu et al. 2011), and protein interactome analysis in ASD (Sakai et al. 2011). In section 6.2.5, I will discuss recent exome sequencing studies in ASD that identify an important role for *de novo* mutation. Of relevance to my study is the report that protein-protein interaction network analysis of genes carrying severe *de novo* mutations indicates that a high proportion of the genes have a function in neuronal development (O'Roak et al, 2012).

6.2.4 Analysis of Individual LoF Variants

LoF variants have been identified in a number of interesting genes and some of these genes contained more than one LoF variant. I will provide detail on some of the genes in this section. *DST* (Dystonin) contains seven LoF variants, the most of any gene in my study. This very large and transcriptionally complex gene encodes multiple isoforms. It is a member of the plakin family of cytolinker proteins, which link cytoskeletal networks to each other and to junctional complexes. *DST* is expressed throughout mouse development and loss of its function results in neuromuscular dysfunction and early death in the mouse mutant *dystonia musculorum* (Guo et al. 1995). Deleterious recessive mutations in *DST* have been identified as the likely cause of a lethal autonomic sensory neuropathy (Edvardson et al. 2012). A frameshift indel in exon 1 (of 84) of the longest transcript was present in 4 cases (2xSZ and 2xASD) and in 2 controls. The remaining variants were singleton events.

Three LoF variants were found in *INADL*, a member of category 1. These variants in exons 18, 22 and 28 (of 43) are all in case samples (2xSZ and 1xASD) and are predicted to produce truncated forms of the protein. Figure 6.1 provides an illustration of the *INADL* protein along with its functional domains, as determined in UniProt (<http://www.uniprot.org/uniprot/>). *INADL* functions to help anchor transmembrane proteins to the cytoskeleton and organize signalling complexes. It interacts with neurexins and neuroligins and is important for cell polarity, migration and may play a role in neurite extension (Rujescu et al. 2009; Glessner et al. 2009).

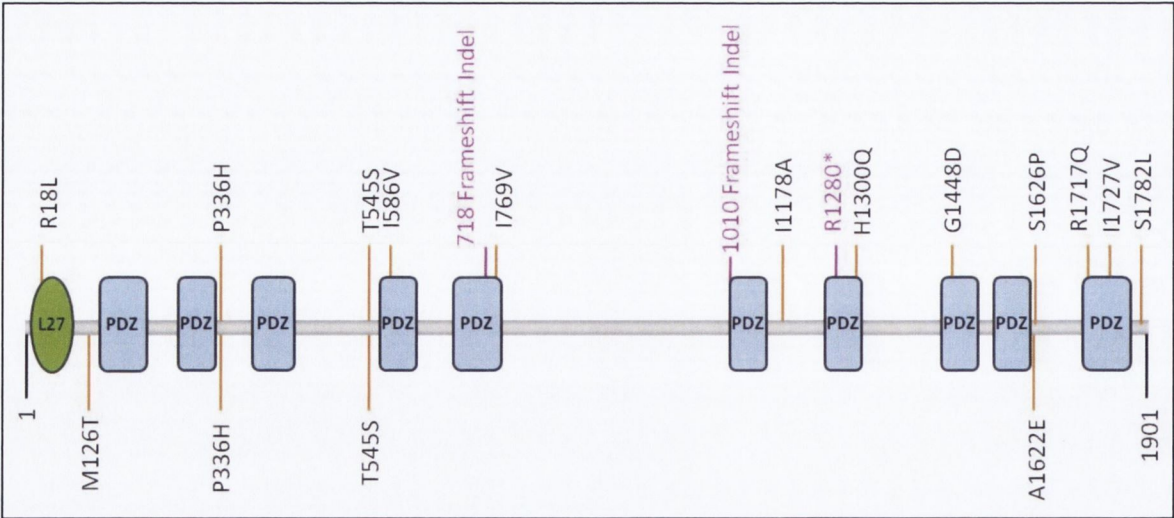


Figure 6.1: Diagram of *INADL* protein with functional domains and sites of variants identified in this study

LoF variants are represented by purple bars and missense variants are represented with orange bars. Variants in cases are shown above the image and variants in controls are shown below the image.

Two LoF variants were detected in *GRIP1* (1xSZ and 1xASD). The ASD variant is a nonsense SNV located in exon 5 (of 25). The SZ variant is a splice site SNV that is predicted to produce a truncated form of the protein after exon 23 (Figure 6.2). *GRIP1* is a member of the glutamate receptor interacting protein family and plays a role in receptor trafficking, synaptic organization, and transmission in glutamatergic and GABAergic synapses (Bradshaw et al. 2008). Following a SNP association between *GRIP1* and ASD, sequencing identified 5 rare missense variants in highly conserved regions of the gene in ASD cases only (Mao et al. 2009). These variants were shown to be associated with altered *GRIP1* interaction with glutamate receptors, faster recycling and increased surface distribution of GluA2 in neurons *in vitro*, which supports a gain of *GRIP1* function in these variants. Knockout mouse studies demonstrated that *GRIP1* is essential for embryonic development and deficits in the protein lead to increased prepulse inhibition (Mao et al. 2009).

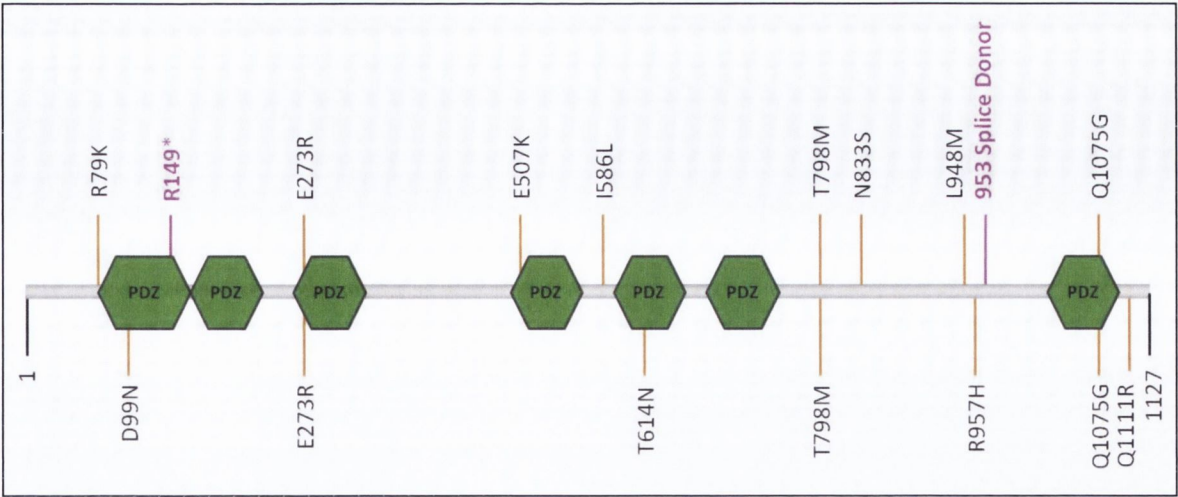


Figure 6.2: Diagram of *GRIP1* protein with functional domains and sites of variants.

LoF variant are represented by purple bars and missense variants are represented with orange bars. Variants in cases are shown above the image and variants in controls are shown below the image.

6.2.5 Importance of *De Novo* Variants

The advent of NGS has revolutionised the field of complex disease genetics. The first NGS studies for SZ and ASD took the form of small numbers of trios samples to investigate *de novo* mutation (Girard et al. 2011; Xu et al. 2011; O’Roak et al. 2011), family-based exome sequencing in ASD (Chahrour et al. 2012) or targeted association studies in SZ of small numbers of candidate genes in pooled DNA samples (Moens et al. 2011). These studies indicate a role for rare sequence variation in risk of SZ and ASD. More recently, this has been extended to larger exome sequencing studies in ASD (Neale et al. 2012; O’Roak et al. 2012; Sanders et al. 2012) and in SZ (Xu et al. 2012). These studies were carried out on very well phenotyped datasets and highlighted the importance of *de novo* mutation and paternal age. At the same time, further studies of CNVs in SZ (International Schizophrenia Consortium 2008; Malhotra et al. 2011) and ASD (O’Donovan et al. 2009; Conrad et al. 2010) have also highlighted the importance of *de novo* mutation.

The discovery of a *de novo* nonsense mutation in an ASD case at *GRIN2B* in this study adds to the three recently reported *de novo* LoF mutations in other ASD samples from one of these large exome datasets and supports *GRIN2B* as a risk gene for the disorder (Figure 6.3). Other data indicates that mutation at *GRIN2B* can contribute to various neurodevelopmental disorders (O’Roak et al.

2012). Ende et al. identified *de novo* translocations with breakpoints disrupting *GRIN2B* in two individuals, one with mild mental retardation (MR) (46,XY,t(9;12)(p23;p13.1)) and one with severe MR (46,XY,t(10;12)(q21.1;p13)) (Ende et al. 2010). Further screening of *GRIN2B* for mutations in 468 individuals with MR and/or epilepsy identified four individuals with moderate MR and behavioural anomalies who had *de novo* *GRIN2B* mutations; a missense SNV, splice donor SNV, splice acceptor SNV and a 2bp frameshift deletion. Talkowski et al. characterized balanced chromosomal abnormalities in 38 subjects with neurodevelopmental abnormalities and identified a *de novo* translocation in an ASD case (46,XY,inv(12)(p13.1q21.31)dn) that disrupted *GRIN2B* (Talkowski et al. 2012).

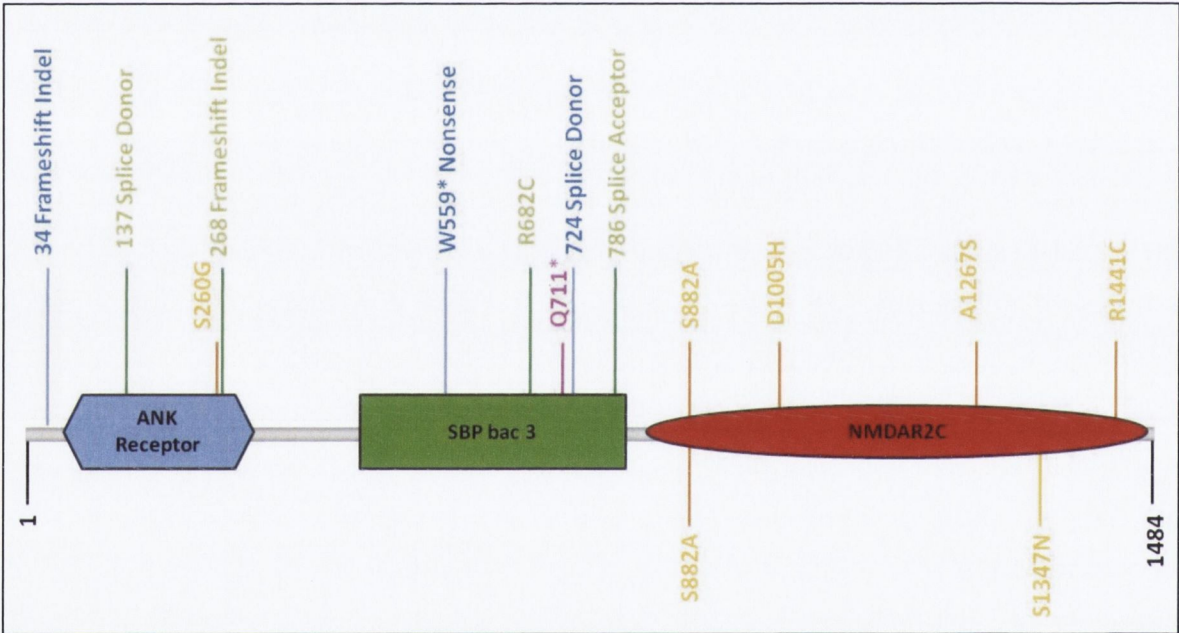


Figure 6.3: Diagram of *GRIN2B* protein with functional domains and sites of variants

From my study the LoF variant is represented by a purple bar and missense variants are represented with orange bars. Variants in cases are shown above the image and variants in controls are shown below the image. *De novo* variants identified by O’Roak et al. (2012) are marked in blue and *de novo* variants from Ende et al. (2010) are marked in green.

GRIN2B, which encodes the glutamate-binding NR2B subunit of the NMDA receptor (NMDAR), is important for channel function, organization of postsynaptic macromolecular complexes, dendritic spine formation or maintenance, and regulation of the actin cytoskeleton (Akashi et al. 2009). Overexpression of the gene in animal models is associated with improved

performance in learning and memory tasks (Tang et al. 1999; Wang, Cui, et al. 2009). *GRIN2B* mutations in humans may affect brain function and cognition by disturbing the electrophysiological balance of the receptor during neurodevelopment (Endele et al. 2010). The *de novo* nonsense SNV (Q711*) identified in an ASD case *GRIN2B* is located in exon 10 truncating a large proportion of the protein (Figure 6.3).

There is robust evidence to support a link between advanced paternal age and the risk of developing SZ and ASD. A Swedish epidemiological study identified that offspring of father aged above 50 years of age had a 2.2 fold increased risk of ASD (Hultman et al. 2011). A similar finding of a 2 fold increase in risk of SZ in offspring of fathers aged above 55 years compared to offspring of fathers aged 20-24 years was reported in an Icelandic population (Frans et al. 2011). It has been suggested that *de novo* mutation occurring in the male germ line are responsible for this association. Spermatogonia undergo cell division approximately every 16 days, accumulating to 200 divisions by the age of 20 and 660 by the age of 40 (Drake et al. 1998). Each time the cell divides, there is a possibility of the introduction of new mutations as the genome is replicated. It has been shown that the sperm of older men have a higher rate of mutations (Glaser et al. 2003). Furthermore, a large scale sequencing study of control trios from an Icelandic population by deCODE Genetics (<http://www.decode.com>), has shown that the rate of *de novo* mutations increases proportionally with father's age (Kong et al. 2012). This indicates that as a man ages, the number of *de novo* mutation in his sperm increases and therefore there is an increased possibility that offspring will carry a deleterious mutation that could lead to SZ or ASD.

I determined the paternal origin of the *GRIN2B de novo* mutation and found it arose on the maternal chromosome. A frameshift indel LoF in *DISC1* was also found to be *de novo*. As this *LoF* variant was located in the last 5% of the protein coding sequence, it was not included in association analysis of these variants. Although, variants in the last 5% of the coding are less likely to be functional, investigation of them would be required to be certain of this. Parent-of-origin analysis indicated that this variant was on the paternal chromosome and closer study of the paternal DNA revealed evidence of the

LoF allele, suggesting mosaicism in the father's blood cells and that the variant is not *de novo* in the proband.

Mosaicism denotes the presence of more than one genetically different cell line derived from a single fertilised egg. This phenomenon is caused when a mutation occurs early in development, resulting in a mixture of cells. The stage of development that mutation occurs determines what tissue(s) proportion of cells will be affected. There are two forms of mosaicism: germline mosaicism or somatic mosaicism. Germline mosaicism is restricted to the gametes, and has been observed in a number of conditions, including Duchenne muscular dystrophy (Passos-Bueno et al. 1992). Somatic mutation can affect any other cells throughout the body, such mutation have been identified in genetic disorders. For example, approximately 2-4% of individuals with Down syndrome are mosaic for cells with trisomy 21 (Hulten et al. 2010).

6.2.6 Functional Genomics

A nonsense mutation was found in *FYN* in a SZ case (Figure 6.4). This Src family protein tyrosine kinase is a key regulator of the NR2B subunit (encoded by *GRIN2B*) of the NMDA receptor (NMDAR) and has previously been linked to both SZ and epilepsy. Interestingly, the SZ case containing this LoF variant has epilepsy. Only two other SZ cases in the study had comorbid epilepsy and both were found to carry LoF variants, in *MACF1* and in *PLXNA2*. These samples were not included in previous SZ GWAS because of the comorbid epilepsy. This highlights the benefit of taking an inclusive approach when selecting phenotype for rare variant studies by potentially identifying some common pathophysiology between the two disorders.

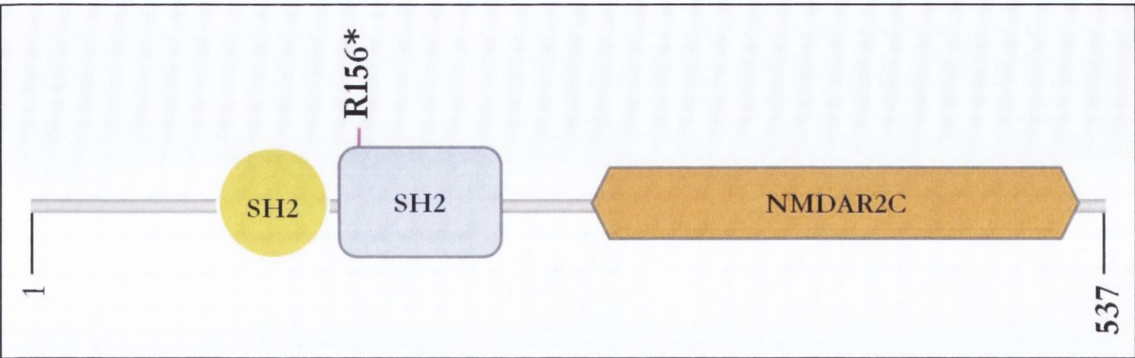


Figure 6.4: Diagram of *FYN* protein with functional domains and sites of variants identified in this study

The nonsense LoF variant identified in a SZ sample is marked. No rare missense variants were identified in this gene.

I selected a LoF variant in *FYN* for functional investigation for two reasons; 1) for strong prior evidence of its role in synaptic function, as discussed in detail in chapter 5, and 2) for its small size, which would provide a higher chance of successful transfection of into mouse neurons. I investigated of the impact of the LoF variant identified in this gene by transfecting mouse prefrontal cortex neurons with a *wildtype* and mutant version of the gene. The purpose of this was to determine the impact of the variant on *FYN* expression and on neuronal function. The introduction of both genes led to a significant increase in the level of *FYN* expression as detected by confocal microscopy. However, there was no significant difference in *FYN* expression between the *wildtype* and *mutant* indicating that the LoF variant does not affect expression of the gene.

There appeared to be a decrease in dendrite length in the transfected neurons compared to the non-transfected cells for each group and this was significant for *FYN mutant* ($p=0.005$). This suggests that expression of mutant *FYN* may impact dendrite length in the cells, supporting the previous findings by Beggs et al. (Beggs, Soriano, and Maness 1994). Investigation of the affect of the variant on synaptic function using synapsin as a marker did not reveal a significant finding. However, it was noted that there are far fewer dendrites with synapsin expression in *FYN mutant* ($n = 7$) compared to *FYN wildtype* ($n = 22$), despite a similar number of neurons (*FYN mutant* $n = 18$; *FYN wildtype* $n = 14$). In addition, inspection of the images revealed *FYN mutant* neurons with an unusual globular morphology and very intense synapsin signal. It is possible that synapsin is not transported properly along the dendrites and is

instead restricted to the cell bodies. This preliminary data suggests that this LoF variant in *FYN* may be impacting on synaptic functioning. Many more observations are required to draw more accurate conclusions. In repeat experiments *MAP2* could also be used in conjunction with synapsin to be able to better distinguish neurons from glial cells and to get a better indication of the expression of synapsin in the dendrites.

I also designed plasmids with *wildtype* and mutant *GRIN2B de novo* LoF variant but transfection of these plasmids using the method employed for *FYN* was unsuccessful and as such was not presented here. The plasmid was very large and transfection of it was most likely unsuccessful due to its size. An alternate approach to transfection could be explored, e.g. using Lipofectamine transfection. Another possibility is to use neuroblastoma cell lines which are easier to transfect. These could subsequently be differentiated into neuron-like cells using a technique involving a combination of retinoic acid and BDNF (Encinas et al. 2000).

Another approach to generating mutant plasmids is site directed mutagenesis, which is a technique for introducing mutations into genes. There are a variety of approaches, one of which involves PCR of whole plasmids. A pair of complementary mutagenic primers are designed to PCR amplify the entire plasmid in a thermocycling reaction using a high-fidelity enzyme. Template DNA is subsequently eliminated by enzymatic digestion with *Dpn1* restriction enzyme, which is specific for methylated DNA (PCR generated DNA will not be methylated). I applied this technique to *FYN* wildtype plasmid and successfully introduced the LoF mutation (as confirmed by PCR and sequencing). This technique could significantly reduce the cost of investigation of variants using this approach of transfecting plasmids containing the gene of interest into neuronal cells because it would only be necessary to purchase a plasmid containing the wildtype version the gene. Then and relatively cheaply, the mutant version could be developed using site directed mutagenesis. Furthermore, it would be possible to introduce other interesting mutations identified in the gene of interest. For example, this study identified many rare missense variants in *GRIN2B* and the impact of those on expression and neuronal functioning could be investigated.

6.2.7 Alternative Investigation of LoF Variants

There are many routes of investigation that were not followed during my studies due to constraints on time and resources. It would be interesting to determine the origin of the LoF variants in the SZ cases and to identify the pattern of inheritance of LoF variants within the families of SZ and ASD patients. Various comparisons could be investigated in relation to carriers versus non-carriers within families. It is well established that unaffected family members of individuals with these disorders can exhibit some of the symptoms of these disorders. Susceptibility to these SZ and ASD is most likely caused by a combination of multiple risk variants and environmental risk factors. Thus, family members can harbour risk variants and may exhibit some of symptoms of the disorders (e.g. cognitive impairments or mildly altered brain connectivity) but not to a sufficient degree to be diagnosed with the disorder. Neurocognitive assessment of family members could be performed e.g. IQ, memory and attention. Furthermore, neuroimaging studies could be employed to investigate structural or connectivity abnormalities e.g. by EEG and fMRI. Within families, the goal would be to determine if the LoF variants co-segregate with abnormalities of brain function.

Dr. Gary Donohoe, Prof. Louise Gallagher and Dr. Eleisa Heron assessed the effects of LoF mutations on cognitive and clinical intermediate phenotypes in SZ and ASD. This was based on a comparison of the scores for symptom severity, general cognitive ability (IQ), memory function, working memory and attentional control in carriers versus non-carriers. Little evidence was found to suggest that the LoF variant carriers differed significantly on clinical and cognitive metrics from non-carriers. However, the number of cases carrying LoF variants is low for performing such analyses. Investigators will want to move away from analysis of individual samples and instead study very large datasets where either multiple samples with variants in the same gene or ideally multiple samples with the same rare variant will be available for study.

Another option would be to generate lymphoblastoid cell lines from the blood of patients by a well-established procedure of transforming B cells derived from patient blood with Epstein-Barr Virus. Expression of the LoF genes in these cell lines could be then investigated *in vitro*. Various splice variants occurring at the consensus splice sites have been identified. *In silico* analysis

can predict the affect of these variants; retained intron, exon skipping or protein truncation. However, the affect of these on the protein requires *in vitro* confirmation. Using these cell lines, this could be achieved by Rapid Amplification of cDNA Ends (RACE), which obtains the full sequence of an RNA transcript by reverse transcription PCR (rtPCR) and sequencing. There are many other splice variants that do not reside within the donor or acceptor splice sites and these variants could have an impact on splicing but are even more difficult to predict. This would be an expensive and time consuming approach that is reliant on the transcription of the genes of interest in lymphoblastoid cells.

Characterisation of disorders on a cellular basis is becoming possible with the advent of induced human derived pluripotent stem cells (iPSC) technology. Cells taken from peripheral tissues of patients and controls can be differentiated into other cell types, such as neurons. This technology offers greater understanding of psychiatric disorders on a molecular and cellular level. These iPSCs contain the complete set of risk variants in these patients so are potentially far more informative than other genetic models of these disorders. Brennand et al. developed iPSCs from fibroblasts in four SZ patients and differentiated them into disorder-specific neurons. These neurons showed reduced neuronal connectivity, decreased neurite number, lower PSD95 within neurites and lower expression of glutamate receptors (Brennand et al. 2011). Furthermore, these cell and molecular defects could be ameliorated following treatment with an antipsychotic drug. In the context of my study, it would be interesting to generate neurons from patient samples using iPSCs that contain specific LoF variants and study the neuronal phenotypes mentioned above.

6.2.8 Missense Variant Analysis

A total of 1,307 rare missense variants were identified in this study. I attempted to prioritise such variants by identifying those most likely to directly impact protein function and thus potentially contribute to disease pathogenesis. To do so, I carried out an assessment of annotation tools available at the time. There are a huge number of tools available and I selected 9 tools that were user friendly, generated interpretable information and had low levels of missing data in their output. Each of these tools is a

valid option for annotating variants but applying each separately would cause a huge increase in the number of tests being performed. I attempted to see if there was correlation between these tools, so I could select a subset that would be representative of the whole set. However, there was little consensus between the tools, which is indicative of the fact that different scoring algorithms are employed. *In silico* prediction of the impact of these variants is very difficult and there is a lack of consensus in the literature regarding the best approach for annotation. The most commonly used tools are PP2 and SIFT and this, rather crudely, is why I selected these tools for application to the final dataset. Condel is a tool that was developed to combine annotation information attained from different tools to generate a weighted average score of the tools (Gonzalez-Perez and Lopez-Bigas 2011). Although, it was originally released in 2011, it is still experiencing problems.

Functional annotation of missense variants is one challenge; the next challenge is deciding which type of statistical analyses that should be performed. One of the few published rare variants association studies at the time was that by Johansen et al. (2010), which performed a sequencing study of candidate genes for hypertriglyceridemia (HTG), and carried out a missense accumulation analysis. A response to this publication pointed out some inherent problems with the analysis performed; rare variants were defined by the frequency in controls which could increase type I error and the allele counting method was flawed (Lemire 2011). Thus, using the recommendations from this article I estimated the allele frequency based on data from the combined case and control sample and performed a genotype based association analysis, to determine if there is a significant excess of rare missense variants in cases versus controls. This analysis did not reveal any interesting findings.

My colleague, Dr. Elaine Kenny performed carrier-based association analysis on this data. This was similar to the LoF analysis where cases and controls were identified as carriers of 0, ≥ 1 , ≥ 2 , ≥ 3 missense variants across all genes together, within gene categories and also at the individual gene level. A thorough analysis of the data using this method, similarly, did not reveal any major differences between cases and controls.

A number of other tools have more recently become available. For example, ATAV (Association Tests for Annotated Variants; www.duke.edu/~minhe/atav/) is a statistical tool designed to detect complex disease-associated rare variants by performing association analysis on annotated variants. Many of the rare variant tests developed take account of multiple rare variants simultaneously, in a form of rare variant accumulation analysis. One of the original approaches for this was Li and Leal's Combined Multivariate and Collapsing (CMC) method, which collapses information on all rare variants within a target region by counting the number of rare alleles (Li and Leal 2008). This type of approach has more power to detect rare variant associations than single-SNP analyses, which was later supported by other similar analysis methods (Morris and Zeggini 2010). These tests are often termed burden tests. GRANVIL (Gene- or Region-based ANalysis of Variants of Intermediate and Low frequency) is a tool that was developed to perform such analysis by measuring the accumulation of minor alleles of rare variants (Magi, Kumar, and Morris 2011).

These accumulation tests or burden tests assume all rare variants in the target region have effects on the phenotype in the same direction and have a similar magnitude of effect. To combat this, the C-alpha test was developed. This tests the variance, rather than the mean in the burden test, maintaining consistent power when the target region contains both risk and protective rare variants (Neale et al. 2011). The SKAT (kernel association test) test is an extension of this method to test for an unusual distribution of rare variants (Wu et al. 2011). These tests are less powerful than burden tests when a large number of variants in the target are risk variants and in the same direction.

This is a challenging task and the type of tests should differ depending on the nature of the relationship between rare variants and phenotype, and perhaps multiple different tests should be performed, but this then imposes the problem of multiple testing (Ladouceur et al. 2012). The development of analytical strategies for maximising the output from sequencing studies, specifically analysis of missense variants, is of critical importance.

6.2.9 Summary of Findings

The aim of this study was to identify rare variants that contribute to pathophysiology of ASD and SZ in genes involved in synaptic functioning and in glutamate signalling. Candidate genes related to these function were investigated in a final sample of 147 cases of ASD, 273 cases of SZ and 287 controls. Within these samples, 31 validated LoF, that are predicted to severely disrupt protein-coding sequence were identified. A significant excess of cases in this dataset carried these variants compared to the controls and this finding was stronger when limited to just singleton variants. Also, as an individual gene category, *NRXN1* and Interactors carried an excess of LoF variants in cases compared to controls. This study adds support to the theory that synaptic function and altered neural connectivity play a role in pathogenesis of neurodevelopmental disorders and this study has identified previously implicated genes (e.g. *GRIN2B*) and novel genes (e.g. *INADL*) with involvement in these processes.

On average each person is predicted to carry approximately 250-300 LoF variants (Genomes Project Consortium 2010). The true impact of LoF variants requires functional investigation. *In vitro* investigation of a novel LoF variant in *FYN* suggests that it may impact dendrite extensions and synaptic function. However, this is preliminary data and many more observations are required to draw concrete conclusions. This is only one study type amongst many that could be followed with these LoF variants, both on a cellular and molecular level and with regard to understanding the implication on brain function in patients.

6.3 Where is the Field of Schizophrenia and Autism Genetics at in 2012?

Since this project began, three more large scale GWAS studies have been performed in SZ. Two studies of Han Chinese populations reported three new loci; 1q24.2 (Shi et al. 2011), 8p12 (Shi et al. 2011) 11p11.2 (Yue et al. 2011) and provided extra support for the MHC region (Yue et al. 2011). The Psychiatric Genomics Consortium (PGC) carried out study of >20,000 individuals and reported hits at two previously implicated sites (MHC and *TCF4*) and five new loci (1p21.3, 2q32.3, 8p23.2, 8q21.3 and 10q24.32-

q24.33) (Schizophrenia Psychiatric Genome-Wide Association Study 2011). One of these findings was for a site in MIR137 (microRNA137), a regulator of neuronal development. It is noteworthy that four of the hits contained predicted target sites of MIR137 suggesting MIR137-mediated dysregulation as a previously unknown etiologic mechanism in SZ. Also, further support has been identified for previous hits of *ZNF804A* (Steinberg et al. 2011; Williams, Norton, et al. 2011), *NRGN* (Williams, Craddock, et al. 2011) and MHC (Williams, Craddock, et al. 2011). The effect sizes for even these robust findings are low ($OR < 1.5$). The most successful finding thus far from a GWAS in psychiatric illness was reported at the World Congress on Psychiatric Genetics in Hamburg, Germany in October 2012. Following a GWAS of more than 25,000 SZ cases and 28,000 controls, the PGC reported 62 independent risk loci for SZ.

For ASD on the other hand, genome-wide evidence for common variation contributing to disease has been very limited, despite some large studies. For example, the AGP have analysed over 2,700 ASD families over a two stage project and identified no robust findings (Anney et al. 2012). Previous hits at *MACROD2*, 5p14.1 and 5p15.2 have failed to replicate (Anney et al. 2012; Curran et al. 2011). After this failure to identify common variants, it has been concluded that common variants of a large effect do not exist for ASD (Devlin and Scherer 2012). There may be variants of intermediate to modest effect, but identification of these would require sample sizes of tens of thousands (Sullivan and Psychiatric Genetics 2012). Sample sizes for ASD studies have not nearly reached those of SZ studies but completion of studies on a similar scale could possibly reveal novel risk loci for this disorder, although the effect sizes of such variants are likely to be small.

CNV studies have revealed more rare variants that are found in SZ and ASD patients, e.g. 3q29 (Ballif et al. 2008; Levinson et al. 2011) and 16p13.11 (Ingason et al. 2011; Pinto et al. 2010). CNV loci in ASD and SZ are often associated with multiple other neuropsychiatric, developmental and neurological phenotypes (Cooper et al. 2011) (Malhotra and Sebat 2012). CNVs associated with SZ and/or ASD often impact multiple genes making the identification of the important risk gene(s) difficult. However, Golzio et al. attempted to determine the driver of mirrored neuroanatomical phenotypes at

the 16p11.2 by overexpressing human transcripts of all 29 genes within this region in *Zebrafish* and discovered that *KCTD13* was the likely causative gene (Golzio et al. 2012). Thus, with the huge advancements in technologies and analysis methods for genetic analysis in recent years, has come the discovery of risk variants for these disorders. The actual impact of these variants is not straightforward but understanding them will be vital to elucidate the underlying pathogenesis of these disorders.

The debate between the CDCV and the CDRV hypothesis is becoming less polarised. The identification of an excess of rare LoF variants in this study and numerous other findings support the rare variant hypothesis. It is now widely believed that disease is caused by a spectrum of common and rare variants, with small and large effects, and that each only contributes a small proportion to heritability (Doherty, O'Donovan, and Owen 2012). Studies will undoubtedly take place that examine both types of variants simultaneously. Full understanding of SZ and ASD pathogenesis will require integration of rare variants, common variants and environmental risk factors. Such studies will be expanded to greater samples numbers as sequencing cost continue to fall and in the near future routine whole genome sequencing in such studies will become feasible.

It is now generally accepted that these SZ and ASD are routed in neurodevelopment. From what we have learned so far from the biological pathways highlighted and the significant level of phenotypic overlap in psychiatric disorders, it might be beneficial to view these disorders as different phenotypic expressions of abnormal neurodevelopment.

Elucidation of the genetic architecture of these SZ and ASD could lead to huge advancements in terms of patient outcome. Biomarkers may allow us to predict those at risk of disease and to determine their prognosis, perhaps allowing preventative measures to be implemented. New therapeutic strategies could be developed with the ultimate goal to benefit patients, their families and society as a whole.

References

- Abrahams, B. S., and D. H. Geschwind. 2008. Advances in autism genetics: on the threshold of a new neurobiology. *Nat Rev Genet* 9 (5):341-55.
- Adzhubei, I. A., S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, and S. R. Sunyaev. 2010. A method and server for predicting damaging missense mutations. *Nat Methods* 7 (4):248-9.
- Akashi, K., T. Kakizaki, H. Kamiya, M. Fukaya, M. Yamasaki, M. Abe, R. Natsume, M. Watanabe, and K. Sakimura. 2009. NMDA receptor GluN2B (GluR epsilon 2/NR2B) subunit is crucial for channel function, postsynaptic macromolecular organization, and actin cytoskeleton at hippocampal CA3 synapses. *J Neurosci* 29 (35):10869-82.
- American Psychiatric Association. 1980. *Diagnostic and statistical manual of mental disorders: 3rd edition (DSM-III)*. Washington, DC: American Psychiatric Press, Inc.
- ———. 1994. *Diagnostic and statistical manual of mental disorders: 4th edition (DSM-IV)*. Washington, DC: American Psychiatric Press, Inc.
- Amir, R. E., I. B. Van den Veyver, M. Wan, C. Q. Tran, U. Francke, and H. Y. Zoghbi. 1999. Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. *Nat Genet* 23 (2):185-8.
- Anney, R., L. Klei, D. Pinto, J. Almeida, E. Bacchelli, G. Baird, N. Bolshakova, S. Bolte, P. F. Bolton, T. Bourgeron, S. Brennan, J. Brian, J. Casey, J. Conroy, C. Correia, C. Corsello, E. L. Crawford, M. de Jonge, R. Delorme, E. Duketis, F. Duque, A. Estes, P. Farrar, B. A. Fernandez, S. E. Folstein, E. Fombonne, J. Gilbert, C. Gillberg, J. T. Glessner, A. Green, J. Green, S. J. Guter, E. A. Heron, R. Holt, J. L. Howe, G. Hughes, V. Hus, R. Iglizzi, S. Jacob, G. P. Kenny, C. Kim, A. Kolevzon, V. Kustanovich, C. M. Lajonchere, J. A. Lamb, M. Law-Smith, M. Leboyer, A. Le Couteur, B. L. Leventhal, X. Q. Liu, F. Lombard, C. Lord, L. Lotspeich, S. C. Lund, T. R. Magalhaes, C. Mantoulan, C. J. McDougle, N. M. Melhem, A. Merikangas, N. J. Minshew, G. K. Mirza, J. Munson, C. Noakes, G. Nygren, K. Papanikolaou, A. T. Pagnamenta, B. Parrini, T. Paton, A. Pickles, D. J. Posey, F. Poustka, J. Ragoussis, R. Regan, W. Roberts, K. Roeder, B. Roge, M. L. Rutter, S. Schlitt, N. Shah, V. C. Sheffield, L. Soorya, I. Sousa, V. Stoppioni, N. Sykes, R. Tancredi, A. P. Thompson, S. Thomson, A. Tryfon, J. Tsiantis, H. Van Engeland, J. B. Vincent, F. Volkmar, J. Vorstman, S. Wallace, K. Wing, K. Wittemeyer, S. Wood, D. Zurawiecki, L. Zwaigenbaum, A. J. Bailey, A. Battaglia, R. M. Cantor, H. Coon, M. L. Cuccaro, G. Dawson, S. Ennis, C. M. Freitag, D. H. Geschwind, J. L. Haines, S. M. Klauck, W. M. McMahon, E. Maestrini, J. Miller, A. P. Monaco, S. F. Nelson, J. I. Nurnberger, Jr., G. Oliveira, J. R. Parr, M. A. Pericak-Vance, J. Piven, G. D. Schellenberg, S. W. Scherer, A. M. Vicente, T. H. Wassink, E. M. Wijsman, C. Betancur, J. D. Buxbaum, E. H. Cook, L. Gallagher, M. Gill, J. Hallmayer, A. D. Paterson, J. S. Sutcliffe, P. Szatmari, V. J. Vieland, H. Hakonarson, and B. Devlin. 2012. Individual common variants exert weak effects on the risk for autism spectrum disorderspi. *Hum Mol Genet* 21 (21):4781-92.
- Anney, R., L. Klei, D. Pinto, R. Regan, J. Conroy, T. R. Magalhaes, C. Correia, B. S. Abrahams, N. Sykes, A. T. Pagnamenta, J. Almeida, E.

Bacchelli, A. J. Bailey, G. Baird, A. Battaglia, T. Berney, N. Bolshakova, S. Bolte, P. F. Bolton, T. Bourgeron, S. Brennan, J. Brian, A. R. Carson, G. Casallo, J. Casey, S. H. Chu, L. Cochrane, C. Corsello, E. L. Crawford, A. Crossett, G. Dawson, M. de Jonge, R. Delorme, I. Drmic, E. Duketis, F. Duque, A. Estes, P. Farrar, B. A. Fernandez, S. E. Folstein, E. Fombonne, C. M. Freitag, J. Gilbert, C. Gillberg, J. T. Glessner, J. Goldberg, J. Green, S. J. Guter, H. Hakonarson, E. A. Heron, M. Hill, R. Holt, J. L. Howe, G. Hughes, V. Hus, R. Iglizzi, C. Kim, S. M. Klauck, A. Kolevzon, O. Korvatska, V. Kustanovich, C. M. Lajonchere, J. A. Lamb, M. Laskawiec, M. Leboyer, A. Le Couteur, B. L. Leventhal, A. C. Lionel, X. Q. Liu, C. Lord, L. Lotspeich, S. C. Lund, E. Maestrini, W. Mahoney, C. Mantoulan, C. R. Marshall, H. McConachie, C. J. McDougale, J. McGrath, W. M. McMahon, N. M. Melhem, A. Merikangas, O. Migita, N. J. Minshew, G. K. Mirza, J. Munson, S. F. Nelson, C. Noakes, A. Noor, G. Nygren, G. Oliveira, K. Papanikolaou, J. R. Parr, B. Parrini, T. Paton, A. Pickles, J. Piven, D. J. Posey, A. Poustka, F. Poustka, A. Prasad, J. Ragoussis, K. Renshaw, J. Rickaby, W. Roberts, K. Roeder, B. Roge, M. L. Rutter, L. J. Bierut, J. P. Rice, J. Salt, K. Sansom, D. Sato, R. Segurado, L. Senman, N. Shah, V. C. Sheffield, L. Soorya, I. Sousa, V. Stoppioni, C. Strawbridge, R. Tancredi, K. Tansey, B. Thiruvahindrapduram, A. P. Thompson, S. Thomson, A. Tryfon, J. Tsiantis, H. Van Engeland, J. B. Vincent, F. Volkmar, S. Wallace, K. Wang, Z. Wang, T. H. Wassink, K. Wing, K. Wittemeyer, S. Wood, B. L. Yaspan, D. Zurawiecki, L. Zwaigenbaum, C. Betancur, J. D. Buxbaum, R. M. Cantor, E. H. Cook, H. Coon, M. L. Cuccaro, L. Gallagher, D. H. Geschwind, M. Gill, J. L. Haines, J. Miller, A. P. Monaco, J. I. Nurnberger, Jr., A. D. Paterson, M. A. Pericak-Vance, G. D. Schellenberg, S. W. Scherer, J. S. Sutcliffe, P. Szatmari, A. M. Vicente, V. J. Vieland, E. M. Wijsman, B. Devlin, S. Ennis, and J. Hallmayer. 2010. A genome-wide scan for common alleles affecting risk for autism. *Hum Mol Genet* 19 (20):4072-82.

- Autism Genome Project, Consortium, P. Szatmari, A. D. Paterson, L. Zwaigenbaum, W. Roberts, J. Brian, X. Q. Liu, J. B. Vincent, J. L. Skaug, A. P. Thompson, L. Senman, L. Feuk, C. Qian, S. E. Bryson, M. B. Jones, C. R. Marshall, S. W. Scherer, V. J. Vieland, C. Bartlett, L. V. Mangin, R. Goedken, A. Segre, M. A. Pericak-Vance, M. L. Cuccaro, J. R. Gilbert, H. H. Wright, R. K. Abramson, C. Betancur, T. Bourgeron, C. Gillberg, M. Leboyer, J. D. Buxbaum, K. L. Davis, E. Hollander, J. M. Silverman, J. Hallmayer, L. Lotspeich, J. S. Sutcliffe, J. L. Haines, S. E. Folstein, J. Piven, T. H. Wassink, V. Sheffield, D. H. Geschwind, M. Bucan, W. T. Brown, R. M. Cantor, J. N. Constantino, T. C. Gilliam, M. Herbert, C. Lajonchere, D. H. Ledbetter, C. Lese-Martin, J. Miller, S. Nelson, C. A. Samango-Sprouse, S. Spence, M. State, R. E. Tanzi, H. Coon, G. Dawson, B. Devlin, A. Estes, P. Flodman, L. Klei, W. M. McMahon, N. Minshew, J. Munson, E. Korvatska, P. M. Rodier, G. D. Schellenberg, M. Smith, M. A. Spence, C. Stodgell, P. G. Tepper, E. M. Wijsman, C. E. Yu, B. Roge, C. Mantoulan, K. Wittemeyer, A. Poustka, B. Felder, S. M. Klauck, C. Schuster, F. Poustka, S. Bolte, S. Feineis-Matthews, E. Herbrecht, G. Schmotzer, J. Tsiantis, K. Papanikolaou, E. Maestrini, E. Bacchelli, F. Blasi, S. Carone, C. Toma, H. Van Engeland, M. de Jonge, C. Kemner, F. Koop, M. Langemeijer, C. Hijmans, W. G. Staal, G. Baird, P. F. Bolton, M. L. Rutter, E. Weisblatt, J. Green, C. Aldred, J. A. Wilkinson, A. Pickles, A. Le Couteur, T. Berney, H. McConachie, A. J. Bailey, K. Francis, G. Honeyman, A. Hutchinson, J. R. Parr, S. Wallace,

- A. P. Monaco, G. Barnby, K. Kobayashi, J. A. Lamb, I. Sousa, N. Sykes, E. H. Cook, S. J. Guter, B. L. Leventhal, J. Salt, C. Lord, C. Corsello, V. Hus, D. E. Weeks, F. Volkmar, M. Tauber, E. Fombonne, A. Shih, and K. J. Meyer. 2007. Mapping autism risk loci using genetic linkage and chromosomal rearrangements. *Nat Genet* 39 (3):319-28.
- Babus, L. W., E. M. Little, K. E. Keenoy, S. S. Minami, E. Chen, J. M. Song, J. Caviness, S. Y. Koo, D. T. Pak, G. W. Rebeck, R. S. Turner, and H. S. Hoe. 2011. Decreased dendritic spine density and abnormal spine morphology in Fyn knockout mice. *Brain Res* 1415:96-102.
 - Bailey, A., A. Le Couteur, I. Gottesman, P. Bolton, E. Simonoff, E. Yuzda, and M. Rutter. 1995. Autism as a strongly genetic disorder: evidence from a British twin study. *Psychol Med* 25 (1):63-77.
 - Ballif, B. C., A. Theisen, J. Coppinger, G. C. Gowans, J. H. Hersh, S. Madan-Khetarpal, K. R. Schmidt, R. Tervo, L. F. Escobar, C. A. Friedrich, M. McDonald, L. Campbell, J. E. Ming, E. H. Zackai, B. A. Bejjani, and L. G. Shaffer. 2008. Expanding the clinical phenotype of the 3q29 microdeletion syndrome and characterization of the reciprocal microduplication. *Mol Cytogenet* 1:8.
 - Baron-Cohen, S., F. J. Scott, C. Allison, J. Williams, P. Bolton, F. E. Matthews, and C. Brayne. 2009. Prevalence of autism-spectrum conditions: UK school-based population study. *Br J Psychiatry* 194 (6):500-9.
 - Bassett, A. S., C. R. Marshall, A. C. Lionel, E. W. Chow, and S. W. Scherer. 2008. Copy number variations and risk for schizophrenia in 22q11.2 deletion syndrome. *Hum Mol Genet* 17 (24):4045-53.
 - Beggs, H. E., P. Soriano, and P. F. Maness. 1994. NCAM-dependent neurite outgrowth is inhibited in neurons from Fyn-minus mice. *J Cell Biol* 127 (3):825-33.
 - Betancur, C., T. Sakurai, and J. D. Buxbaum. 2009. The emerging role of synaptic cell-adhesion pathways in the pathogenesis of autism spectrum disorders. *Trends Neurosci* 32 (7):402-12.
 - Biederer, T., and T. C. Sudhof. 2000. Mints as adaptors. Direct binding to neurexins and recruitment of munc18. *J Biol Chem* 275 (51):39803-6.
 - Bjarnadottir, M., D. L. Misner, S. Haverfield-Gross, S. Bruun, V. G. Helgason, H. Stefansson, A. Sigmundsson, D. R. Firth, B. Nielsen, R. Stefansdottir, T. J. Novak, K. Stefansson, M. E. Gurney, and T. Andreasson. 2007. Neuregulin1 (NRG1) signaling through Fyn modulates NMDA receptor phosphorylation: differential synaptic function in NRG1+/- knock-outs compared with wild-type mice. *J Neurosci* 27 (17):4519-29.
 - Bourgeron, T. 2009. A synaptic trek to autism. *Curr Opin Neurobiol* 19 (2):231-4.
 - Bradshaw, N. J., F. Ogawa, B. Antolin-Fontes, J. E. Chubb, B. C. Carlyle, S. Christie, A. Claessens, D. J. Porteous, and J. K. Millar. 2008. DISC1, PDE4B, and NDE1 at the centrosome and synapse. *Biochem Biophys Res Commun* 377 (4):1091-6.
 - Brandon, N. J., and A. Sawa. 2011. Linking neurodevelopmental and synaptic theories of mental illness through DISC1. *Nat Rev Neurosci* 12 (12):707-22.
 - Brennand, K. J., A. Simone, J. Jou, C. Gelboin-Burkhart, N. Tran, S. Sangar, Y. Li, Y. Mu, G. Chen, D. Yu, S. McCarthy, J. Sebat, and F. H. Gage. 2011. Modelling schizophrenia using human induced pluripotent stem cells. *Nature* 473 (7346):221-5.

- Bromley, R. L., G. Mawer, J. Clayton-Smith, G. A. Baker, Liverpool, and Group Manchester Neurodevelopment. 2008. Autism spectrum disorders following in utero exposure to antiepileptic drugs. *Neurology* 71 (23):1923-4.
- Cain, D. P., S. G. Grant, D. Saucier, E. L. Hargreaves, and E. R. Kandel. 1995. Fyn tyrosine kinase is required for normal amygdala kindling. *Epilepsy Res* 22 (2):107-14.
- Camargo, L. M., V. Collura, J. C. Rain, K. Mizuguchi, H. Hermjakob, S. Kerrien, T. P. Bonnert, P. J. Whiting, and N. J. Brandon. 2007. Disrupted in Schizophrenia 1 Interactome: evidence for the close connectivity of risk genes and a potential synaptic basis for schizophrenia. *Mol Psychiatry* 12 (1):74-86.
- Cardno, A. G., and Gottesman, II. 2000. Twin studies of schizophrenia: from bow-and-arrow concordances to star wars Mx and functional genomics. *Am J Med Genet* 97 (1):12-7.
- Carroll, L. S., and M. J. Owen. 2009. Genetic overlap between autism, schizophrenia and bipolar disorder. *Genome Med* 1 (10):102.
- Chahrour, M. H., T. W. Yu, E. T. Lim, B. Ataman, M. E. Coulter, R. S. Hill, C. R. Stevens, C. R. Schubert, Arra Autism Sequencing Collaboration, M. E. Greenberg, S. B. Gabriel, and C. A. Walsh. 2012. Whole-exome sequencing and homozygosity analysis implicate depolarization-regulated neuronal genes in autism. *PLoS Genet* 8 (4):e1002635.
- Chubb, J. E., N. J. Bradshaw, D. C. Soares, D. J. Porteous, and J. K. Millar. 2008. The DISC locus in psychiatric illness. *Mol Psychiatry* 13 (1):36-64.
- Clifford, R. J., M. N. Edmonson, C. Nguyen, and K. H. Buetow. 2004. Large-scale analysis of non-synonymous coding region single nucleotide polymorphisms. *Bioinformatics* 20 (7):1006-14.
- Conrad, D. F., D. Pinto, R. Redon, L. Feuk, O. Gokcumen, Y. Zhang, J. Aerts, T. D. Andrews, C. Barnes, P. Campbell, T. Fitzgerald, M. Hu, C. H. Ihm, K. Kristiansson, D. G. Macarthur, J. R. Macdonald, I. Onyiah, A. W. Pang, S. Robson, K. Stirrups, A. Valsesia, K. Walter, J. Wei, Consortium Wellcome Trust Case Control, C. Tyler-Smith, N. P. Carter, C. Lee, S. W. Scherer, and M. E. Hurles. 2010. Origins and functional impact of copy number variation in the human genome. *Nature* 464 (7289):704-12.
- Cooper, G. M., B. P. Coe, S. Girirajan, J. A. Rosenfeld, T. H. Vu, C. Baker, C. Williams, H. Stalker, R. Hamid, V. Hannig, H. Abdel-Hamid, P. Bader, E. McCracken, D. Niyazov, K. Leppig, H. Thiese, M. Hummel, N. Alexander, J. Gorski, J. Kussmann, V. Shashi, K. Johnson, C. Rehder, B. C. Ballif, L. G. Shaffer, and E. E. Eichler. 2011. A copy number variation morbidity map of developmental delay. *Nat Genet* 43 (9):838-46.
- Cooper, G. M., E. A. Stone, G. Asimenos, E. D. Green, S. Batzoglu, and A. Sidow. 2005. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* 15 (7):901-13.
- Craig, D. W., J. V. Pearson, S. Szelinger, A. Sekar, M. Redman, J. J. Corneveaux, T. L. Pawlowski, T. Laub, G. Nunn, D. A. Stephan, N. Homer, and M. J. Huentelman. 2008. Identification of genetic variants using bar-coded multiplexed sequencing. *Nat Methods* 5 (10):887-93.
- Curran, S., P. Bolton, K. Rozsnyai, A. Chiocchetti, S. M. Klauck, E. Duketis, F. Poustka, S. Schlitt, C. M. Freitag, I. Lee, P. Muglia, Itan, M. Poot, W. Staal, M. V. de Jonge, R. A. Ophoff, C. Lewis, D. Skuse, W. Mandy, E. Vassos, R. Fossdal, P. Magnusson, S. Hreidarsson, E. Saemundsen, H. Stefansson, K. Stefansson, and D. Collier. 2011. No

- association between a common single nucleotide polymorphism, rs4141463, in the MACROD2 gene and autism spectrum disorder. *Am J Med Genet B Neuropsychiatr Genet* 156B (6):633-9.
- DePristo, M. A., E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, M. Hanna, A. McKenna, T. J. Fennell, A. M. Kernysky, A. Y. Sivachenko, K. Cibulskis, S. B. Gabriel, D. Altshuler, and M. J. Daly. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 43 (5):491-8.
 - Devlin, B., and S. W. Scherer. 2012. Genetic architecture in autism spectrum disorder. *Curr Opin Genet Dev* 22 (3):229-37.
 - Doherty, J. L., M. C. O'Donovan, and M. J. Owen. 2012. Recent genomic advances in schizophrenia. *Clin Genet* 81 (2):103-9.
 - Drake, J. W., B. Charlesworth, D. Charlesworth, and J. F. Crow. 1998. Rates of spontaneous mutation. *Genetics* 148 (4):1667-86.
 - Edghill, E. L., A. L. Gloyn, A. Goriely, L. W. Harries, S. E. Flanagan, J. Rankin, A. T. Hattersley, and S. Ellard. 2007. Origin of de novo KCNJ11 mutations and risk of neonatal diabetes for subsequent siblings. *J Clin Endocrinol Metab* 92 (5):1773-7.
 - Edvardson, S., Y. Cinnamon, C. Jalas, A. Shaag, C. Maayan, F. B. Axelrod, and O. Elpeleg. 2012. Hereditary sensory autonomic neuropathy caused by a mutation in dystonin. *Ann Neurol* 71 (4):569-72.
 - Encinas, M., M. Iglesias, Y. Liu, H. Wang, A. Muhaisen, V. Cena, C. Gallego, and J. X. Comella. 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):991-1003.
 - Ende, S., G. Rosenberger, K. Geider, B. Popp, C. Tamer, I. Stefanova, M. Milh, F. Kortum, A. Fritsch, F. K. Pientka, Y. Hellenbroich, V. M. Kalscheuer, J. Kohlhasse, U. Moog, G. Rappold, A. Rauch, H. H. Ropers, S. von Spiczak, H. Tonnies, N. Villeneuve, L. Villard, B. Zabel, M. Zenker, B. Laube, A. Reis, D. Wieczorek, L. Van Maldergem, and K. Kutsche. 2010. Mutations in GRIN2A and GRIN2B encoding regulatory subunits of NMDA receptors cause variable neurodevelopmental phenotypes. *Nat Genet* 42 (11):1021-6.
 - Enticott, P. G., K. E. Hoy, S. E. Herring, P. J. Johnston, Z. J. Daskalakis, and P. B. Fitzgerald. 2008. Reduced motor facilitation during action observation in schizophrenia: a mirror neuron deficit? *Schizophr Res* 102 (1-3):116-21.
 - Farrer, L. A., L. A. Cupples, J. L. Haines, B. Hyman, W. A. Kukull, R. Mayeux, R. H. Myers, M. A. Pericak-Vance, N. Risch, and C. M. van Duijn. 1997. Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease. A meta-analysis. APOE and Alzheimer Disease Meta Analysis Consortium. *JAMA* 278 (16):1349-56.
 - Fatjo-Vilas, M., D. Gourion, S. Campanera, F. Mouaffak, M. Levy-Rueff, M. E. Navarro, M. Chayet, S. Miret, M. O. Krebs, and L. Fananas. 2008. New evidences of gene and environment interactions affecting prenatal neurodevelopment in schizophrenia-spectrum disorders: a family dermatoglyphic study. *Schizophr Res* 103 (1-3):209-17.
 - Fernell, E., and C. Gillberg. 2010. Autism spectrum disorder diagnoses in Stockholm preschoolers. *Res Dev Disabil* 31 (3):680-5.

- Fombonne, E. 2009. Epidemiology of pervasive developmental disorders. *Pediatr Res* 65 (6):591-8.
- Frans, E. M., J. J. McGrath, S. Sandin, P. Lichtenstein, A. Reichenberg, N. Langstrom, and C. M. Hultman. 2011. Advanced paternal and grandpaternal age and schizophrenia: a three-generation perspective. *Schizophr Res* 133 (1-3):120-4.
- Freitag, C. M., W. Staal, S. M. Klauck, E. Duketis, and R. Waltes. 2010. Genetics of autistic disorders: review and clinical implications. *Eur Child Adolesc Psychiatry* 19 (3):169-78.
- Friedrichs, F., and M. Stoll. 2006. Role of discs large homolog 5. *World J Gastroenterol* 12 (23):3651-6.
- Funk, A. J., R. E. McCullumsmith, V. Haroutunian, and J. H. Meador-Woodruff. 2012. Abnormal activity of the MAPK- and cAMP-associated signaling pathways in frontal cortical areas in postmortem brain in schizophrenia. *Neuropsychopharmacology* 37 (4):896-905.
- Garla, V., Y. Kong, S. Szpakowski, and M. Krauthammer. 2011. MU2A--reconciling the genome and transcriptome to determine the effects of base substitutions. *Bioinformatics* 27 (3):416-8.
- Geddes, A. E., X. F. Huang, and K. A. Newell. 2011. Reciprocal signalling between NR2 subunits of the NMDA receptor and neuregulin1 and their role in schizophrenia. *Prog Neuropsychopharmacol Biol Psychiatry* 35 (4):896-904.
- Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature* 467 (7319):1061-73.
- Geschwind, D. H. 2009. Advances in autism. *Annu Rev Med* 60:367-80.
- Gill, M., G. Donohoe, and A. Corvin. 2010. What have the genomics ever done for the psychoses? *Psychol Med* 40 (4):529-40.
- Gillberg, C., M. Cederlund, K. Lamberg, and L. Zeijlon. 2006. Brief report: "the autism epidemic". The registered prevalence of autism in a Swedish urban area. *J Autism Dev Disord* 36 (3):429-35.
- Gilman, S. R., I. Iossifov, D. Levy, M. Ronemus, M. Wigler, and D. Vitkup. 2011. Rare de novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses. *Neuron* 70 (5):898-907.
- Girard, S. L., J. Gauthier, A. Noreau, L. Xiong, S. Zhou, L. Jouan, A. Dionne-Laporte, D. Spiegelman, E. Henrion, O. Diallo, P. Thibodeau, I. Bachand, J. Y. Bao, A. H. Tong, C. H. Lin, B. Millet, N. Jaafari, R. Joobar, P. A. Dion, S. Lok, M. O. Krebs, and G. A. Rouleau. 2011. Increased exonic de novo mutation rate in individuals with schizophrenia. *Nat Genet* 43 (9):860-3.
- Glaser, R. L., K. W. Broman, R. L. Schulman, B. Eskenazi, A. J. Wyrobek, and E. W. Jabs. 2003. The paternal-age effect in Apert syndrome is due, in part, to the increased frequency of mutations in sperm. *Am J Hum Genet* 73 (4):939-47.
- Glessner, J. T., K. Wang, G. Cai, O. Korvatska, C. E. Kim, S. Wood, H. Zhang, A. Estes, C. W. Brune, J. P. Bradfield, M. Imielinski, E. C. Frackelton, J. Reichert, E. L. Crawford, J. Munson, P. M. Sleiman, R. Chiavacci, K. Annaiah, K. Thomas, C. Hou, W. Glaberson, J. Flory, F. Otieno, M. Garris, L. Soorya, L. Klei, J. Piven, K. J. Meyer, E. Anagnostou, T. Sakurai, R. M. Game, D. S. Rudd, D. Zurawiecki, C. J. McDougale, L. K. Davis, J. Miller, D. J. Posey, S. Michaels, A. Kolevzon, J. M. Silverman, R. Bernier, S. E. Levy, R. T. Schultz, G. Dawson, T. Owley, W. M. McMahon, T. H. Wassink, J. A. Sweeney, J. I. Nurnberger, H. Coon, J. S. Sutcliffe, N. J. Minshew, S. F. Grant, M. Bucan, E. H.

- Cook, J. D. Buxbaum, B. Devlin, G. D. Schellenberg, and H. Hakonarson. 2009. Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. *Nature* 459 (7246):569-73.
- Gnirke, A., A. Melnikov, J. Maguire, P. Rogov, E. M. LeProust, W. Brockman, T. Fennell, G. Giannoukos, S. Fisher, C. Russ, S. Gabriel, D. B. Jaffe, E. S. Lander, and C. Nusbaum. 2009. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol* 27 (2):182-9.
 - Golzio, C., J. Willer, M. E. Talkowski, E. C. Oh, Y. Taniguchi, S. Jacquemont, A. Reymond, M. Sun, A. Sawa, J. F. Gusella, A. Kamiya, J. S. Beckmann, and N. Katsanis. 2012. KCTD13 is a major driver of mirrored neuroanatomical phenotypes of the 16p11.2 copy number variant. *Nature* 485 (7398):363-7.
 - Gonzalez-Perez, A., and N. Lopez-Bigas. 2011. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *Am J Hum Genet* 88 (4):440-9.
 - Gottesman, II, and L. Erlenmeyer-Kimling. 2001. Family and twin strategies as a head start in defining prodromes and endophenotypes for hypothetical early-interventions in schizophrenia. *Schizophr Res* 51 (1):93-102.
 - Grant, S. G., M. C. Marshall, K. L. Page, M. A. Cumiskey, and J. D. Armstrong. 2005. Synapse proteomics of multiprotein complexes: en route from genes to nervous system diseases. *Hum Mol Genet* 14 Spec No. 2:R225-34.
 - Grant, S. G., T. J. O'Dell, K. A. Karl, P. L. Stein, P. Soriano, and E. R. Kandel. 1992. Impaired long-term potentiation, spatial learning, and hippocampal development in fyn mutant mice. *Science* 258 (5090):1903-10.
 - Grantham, R. 1974. Amino acid difference formula to help explain protein evolution. *Science* 185 (4154):862-4.
 - Guilmatre, A., C. Dubourg, A. L. Mosca, S. Legallic, A. Goldenberg, V. Drouin-Garraud, V. Layet, A. Rosier, S. Briault, F. Bonnet-Brilhault, F. Laumonnier, S. Odent, G. Le Vacon, G. Joly-Helas, V. David, C. Bendavid, J. M. Pinoit, C. Henry, C. Impallomeni, E. Germano, G. Tortorella, G. Di Rosa, C. Barthelemy, C. Andres, L. Faivre, T. Frebourg, P. Saugier Veber, and D. Campion. 2009. Recurrent rearrangements in synaptic and neurodevelopmental genes and shared biologic pathways in schizophrenia, autism, and mental retardation. *Arch Gen Psychiatry* 66 (9):947-56.
 - Guo, L., L. Degenstein, J. Dowling, Q. C. Yu, R. Wollmann, B. Perman, and E. Fuchs. 1995. Gene targeting of BPAG1: abnormalities in mechanical strength and cell migration in stratified epithelia and neurologic degeneration. *Cell* 81 (2):233-43.
 - Gusella, J. F., N. S. Wexler, P. M. Conneally, S. L. Naylor, M. A. Anderson, R. E. Tanzi, P. C. Watkins, K. Ottina, M. R. Wallace, A. Y. Sakaguchi, and et al. 1983. A polymorphic DNA marker genetically linked to Huntington's disease. *Nature* 306 (5940):234-8.
 - Hahn, C. G., H. Y. Wang, D. S. Cho, K. Talbot, R. E. Gur, W. H. Berrettini, K. Bakshi, J. Kamins, K. E. Borgmann-Winter, S. J. Siegel, R. J. Gallop, and S. E. Arnold. 2006. Altered neuregulin 1-erbB4 signaling contributes to NMDA receptor hypofunction in schizophrenia. *Nat Med* 12 (7):824-8.
 - Hallmayer, J., S. Cleveland, A. Torres, J. Phillips, B. Cohen, T. Torigoe, J. Miller, A. Fedele, J. Collins, K. Smith, L. Lotspeich, L. A. Croen, S.

Ozonoff, C. Lajonchere, J. K. Grether, and N. Risch. 2011. Genetic Heritability and Shared Environmental Factors Among Twin Pairs With Autism. *Arch Gen Psychiatry*.

- Hamada, S., and T. Yagi. 2001. The cadherin-related neuronal receptor family: a novel diversified cadherin family at the synapse. *Neurosci Res* 41 (3):207-15.
- Hattori, K., H. Fukuzako, T. Hashiguchi, S. Hamada, Y. Murata, T. Isosaka, S. Yuasa, and T. Yagi. 2009. Decreased expression of Fyn protein and disbalanced alternative splicing patterns in platelets from patients with schizophrenia. *Psychiatry Res* 168 (2):119-28.
- Henikoff, S., and J. G. Henikoff. 1992. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* 89 (22):10915-9.
- Hennah, W., P. Thomson, A. McQuillin, N. Bass, A. Loukola, A. Anjorin, D. Blackwood, D. Curtis, I. J. Deary, S. E. Harris, E. T. Isometsa, J. Lawrence, J. Lonnqvist, W. Muir, A. Palotie, T. Partonen, T. Paunio, E. Pylkko, M. Robinson, P. Soronen, K. Suominen, J. Suvisaari, S. Thirumalai, D. St Clair, H. Gurling, L. Peltonen, and D. Porteous. 2009. DISC1 association, heterogeneity and interplay in schizophrenia and bipolar disorder. *Mol Psychiatry* 14 (9):865-73.
- Hogart, A., D. Wu, J. M. LaSalle, and N. C. Schanen. 2010. The comorbidity of autism with the genomic disorders of chromosome 15q11.2-q13. *Neurobiol Dis* 38 (2):181-91.
- Hulten, M. A., S. D. Patel, M. Westgren, N. Papadogiannakis, A. M. Jonsson, J. Jonasson, and E. Iwarsson. 2010. On the paternal origin of trisomy 21 Down syndrome. *Mol Cytogenet* 3:4.
- Hultman, C. M., S. Sandin, S. Z. Levine, P. Lichtenstein, and A. Reichenberg. 2011. Advancing paternal age and risk of autism: new evidence from a population-based study and a meta-analysis of epidemiological studies. *Mol Psychiatry* 16 (12):1203-12.
- Ingason, A., D. Rujescu, S. Cichon, E. Sigurdsson, T. Sigmundsson, O. P. Pietilainen, J. E. Buizer-Voskamp, E. Strengman, C. Francks, P. Muglia, A. Gylfason, O. Gustafsson, P. I. Olason, S. Steinberg, T. Hansen, K. D. Jakobsen, H. B. Rasmussen, I. Giegling, H. J. Moller, A. Hartmann, C. Crombie, G. Fraser, N. Walker, J. Lonnqvist, J. Suvisaari, A. Tuulio-Henriksson, E. Bramon, L. A. Kiemeny, B. Franke, R. Murray, E. Vassos, T. Touloupoulou, T. W. Muhleisen, S. Tosato, M. Ruggeri, S. Djurovic, O. A. Andreassen, Z. Zhang, T. Werge, R. A. Ophoff, Group Investigators, M. Rietschel, M. M. Nothen, H. Petursson, H. Stefansson, L. Peltonen, D. Collier, K. Stefansson, and D. M. St Clair. 2011. Copy number variations of chromosome 16p13.1 region associated with schizophrenia. *Mol Psychiatry* 16 (1):17-25.
- Insel, T. R. 2010. Rethinking schizophrenia. *Nature* 468 (7321):187-93.
- International HapMap Consortium, K. A. Frazer, D. G. Ballinger, D. R. Cox, D. A. Hinds, L. L. Stuve, R. A. Gibbs, J. W. Belmont, A. Boudreau, P. Hardenbol, S. M. Leal, S. Pasternak, D. A. Wheeler, T. D. Willis, F. Yu, H. Yang, C. Zeng, Y. Gao, H. Hu, W. Hu, C. Li, W. Lin, S. Liu, H. Pan, X. Tang, J. Wang, W. Wang, J. Yu, B. Zhang, Q. Zhang, H. Zhao, H. Zhao, J. Zhou, S. B. Gabriel, R. Barry, B. Blumenstiel, A. Camargo, M. Defelice, M. Faggart, M. Goyette, S. Gupta, J. Moore, H. Nguyen, R. C. Onofrio, M. Parkin, J. Roy, E. Stahl, E. Winchester, L. Ziaugra, D. Altshuler, Y. Shen, Z. Yao, W. Huang, X. Chu, Y. He, L. Jin, Y. Liu, Y. Shen, W. Sun, H. Wang, Y. Wang, Y. Wang, X. Xiong, L. Xu, M. M. Waye, S. K. Tsui, H. Xue, J. T. Wong, L. M. Galver, J. B. Fan, K. Gunderson, S. S. Murray, A. R. Oliphant, M. S. Chee, A. Montpetit, F.

- Chagnon, V. Ferretti, M. Leboeuf, J. F. Olivier, M. S. Phillips, S. Roumy, C. Sallee, A. Verner, T. J. Hudson, P. Y. Kwok, D. Cai, D. C. Koboldt, R. D. Miller, L. Pawlikowska, P. Taillon-Miller, M. Xiao, L. C. Tsui, W. Mak, Y. Q. Song, P. K. Tam, Y. Nakamura, T. Kawaguchi, T. Kitamoto, T. Morizono, A. Nagashima, Y. Ohnishi, A. Sekine, T. Tanaka, T. Tsunoda, P. Deloukas, C. P. Bird, M. Delgado, E. T. Dermitzakis, R. Gwilliam, S. Hunt, J. Morrison, D. Powell, B. E. Stranger, P. Whittaker, D. R. Bentley, M. J. Daly, P. I. de Bakker, J. Barrett, Y. R. Chretien, J. Maller, S. McCarroll, N. Patterson, I. Pe'er, A. Price, S. Purcell, D. J. Richter, P. Sabeti, R. Saxena, S. F. Schaffner, P. C. Sham, P. Varilly, D. Altshuler, L. D. Stein, L. Krishnan, A. V. Smith, M. K. Tello-Ruiz, G. A. Thorisson, A. Chakravarti, P. E. Chen, D. J. Cutler, C. S. Kashuk, S. Lin, G. R. Abecasis, W. Guan, Y. Li, H. M. Munro, Z. S. Qin, D. J. Thomas, G. McVean, A. Auton, L. Bottolo, N. Cardin, S. Eyheramendy, C. Freeman, J. Marchini, S. Myers, C. Spencer, M. Stephens, P. Donnelly, L. R. Cardon, G. Clarke, D. M. Evans, A. P. Morris, B. S. Weir, T. Tsunoda, J. C. Mullikin, S. T. Sherry, M. Feolo, A. Skol, H. Zhang, C. Zeng, H. Zhao, I. Matsuda, Y. Fukushima, D. R. Macer, E. Suda, C. N. Rotimi, C. A. Adebamowo, I. Ajayi, T. Aniagwu, P. A. Marshall, C. Nkwodimmah, C. D. Royal, M. F. Leppert, M. Dixon, A. Peiffer, R. Qiu, A. Kent, K. Kato, N. Niikawa, I. F. Adewole, B. M. Knoppers, M. W. Foster, E. W. Clayton, J. Watkin, R. A. Gibbs, J. W. Belmont, D. Muzny, L. Nazareth, E. Sodergren, G. M. Weinstock, D. A. Wheeler, I. Yakub, S. B. Gabriel, R. C. Onofrio, D. J. Richter, L. Ziaugra, B. W. Birren, M. J. Daly, D. Altshuler, R. K. Wilson, L. L. Fulton, J. Rogers, J. Burton, N. P. Carter, C. M. Clee, M. Griffiths, M. C. Jones, K. McLay, R. W. Plumb, M. T. Ross, S. K. Sims, D. L. Willey, Z. Chen, H. Han, L. Kang, M. Godbout, J. C. Wallenburg, P. L'Archeveque, G. Bellemare, K. Saeki, H. Wang, D. An, H. Fu, Q. Li, Z. Wang, R. Wang, A. L. Holden, L. D. Brooks, J. E. McEwen, M. S. Guyer, V. O. Wang, J. L. Peterson, M. Shi, J. Spiegel, L. M. Sung, L. F. Zacharia, F. S. Collins, K. Kennedy, R. Jamieson, and J. Stewart. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449 (7164):851-61.
- International Schizophrenia Consortium. 2008. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* 455 (7210):237-41.
 - International Schizophrenia Consortium, S. M. Purcell, N. R. Wray, J. L. Stone, P. M. Visscher, M. C. O'Donovan, P. F. Sullivan, and P. Sklar. 2009. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* 460 (7256):748-52.
 - Kana, R. K., L. E. Libero, C. P. Hu, H. D. Deshpande, and J. S. Colburn. 2012. Functional Brain Networks and White Matter Underlying Theory-of-Mind in Autism. *Soc Cogn Affect Neurosci*.
 - Kenny, E. M., P. Cormican, W. P. Gilks, A. S. Gates, C. T. O'Dushlaine, C. Pinto, A. P. Corvin, M. Gill, and D. W. Morris. 2011. Multiplex target enrichment using DNA indexing for ultra-high throughput SNP detection. *DNA Res* 18 (1):31-8.
 - Kim, H. G., S. Kishikawa, A. W. Higgins, I. S. Seong, D. J. Donovan, Y. Shen, E. Lally, L. A. Weiss, J. Najm, K. Kutsche, M. Descartes, L. Holt, S. Braddock, R. Troxell, L. Kaplan, F. Volkmar, A. Klin, K. Tsatsanis, D. J. Harris, I. Noens, D. L. Pauls, M. J. Daly, M. E. MacDonald, C. C. Morton, B. J. Quade, and J. F. Gusella. 2008. Disruption of neurexin 1 associated with autism spectrum disorder. *Am J Hum Genet* 82 (1):199-207.

- Kim, J. Y., X. Duan, C. Y. Liu, M. H. Jang, J. U. Guo, N. Pow-anpongkul, E. Kang, H. Song, and G. L. Ming. 2009. DISC1 regulates new neuron development in the adult brain via modulation of AKT-mTOR signaling through KIAA1212. *Neuron* 63 (6):761-73.
- King, B. H., and C. Lord. 2011. Is schizophrenia on the autism spectrum? *Brain Res* 1380:34-41.
- Kirov, G., D. Grozeva, N. Norton, D. Ivanov, K. K. Mantripragada, P. Holmans, Consortium International Schizophrenia, Consortium Wellcome Trust Case Control, N. Craddock, M. J. Owen, and M. C. O'Donovan. 2009. Support for the involvement of large copy number variants in the pathogenesis of schizophrenia. *Hum Mol Genet* 18 (8):1497-503.
- Kirov, G., D. Gumus, W. Chen, N. Norton, L. Georgieva, M. Sari, M. C. O'Donovan, F. Erdogan, M. J. Owen, H. H. Ropers, and R. Ullmann. 2008. Comparative genome hybridization suggests a role for NRXN1 and APBA2 in schizophrenia. *Hum Mol Genet* 17 (3):458-65.
- Kirov, G., A. J. Pocklington, P. Holmans, D. Ivanov, M. Ikeda, D. Ruderfer, J. Moran, K. Chambert, D. Toncheva, L. Georgieva, D. Grozeva, M. Fjodorova, R. Wollerton, E. Rees, I. Nikolov, L. N. van de Lagemaat, A. Bayes, E. Fernandez, P. I. Olason, Y. Bottcher, N. H. Komiyama, M. O. Collins, J. Choudhary, K. Stefansson, H. Stefansson, S. G. Grant, S. Purcell, P. Sklar, M. C. O'Donovan, and M. J. Owen. 2012. De novo CNV analysis implicates specific abnormalities of postsynaptic signalling complexes in the pathogenesis of schizophrenia. *Mol Psychiatry* 17 (2):142-53.
- Kojima, N., J. Wang, I. M. Mansuy, S. G. Grant, M. Mayford, and E. R. Kandel. 1997. Rescuing impairment of long-term potentiation in fyn-deficient mice by introducing Fyn transgene. *Proc Natl Acad Sci U S A* 94 (9):4761-5.
- Kong, A., M. L. Frigge, G. Masson, S. Besenbacher, P. Sulem, G. Magnusson, S. A. Gudjonsson, A. Sigurdsson, A. Jonasdottir, A. Jonasdottir, W. S. Wong, G. Sigurdsson, G. B. Walters, S. Steinberg, H. Helgason, G. Thorleifsson, D. F. Gudbjartsson, A. Helgason, O. T. Magnusson, U. Thorsteinsdottir, and K. Stefansson. 2012. Rate of de novo mutations and the importance of father's age to disease risk. *Nature* 488 (7412):471-5.
- Konrad, A., G. Vucurevic, F. Musso, P. Stoeter, N. Dahmen, and G. Winterer. 2009. ErbB4 genotype predicts left frontotemporal structural connectivity in human brain. *Neuropsychopharmacology* 34 (3):641-50.
- Kumar, P., S. Henikoff, and P. C. Ng. 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* 4 (7):1073-81.
- Laumonnier, F., P. C. Cuthbert, and S. G. Grant. 2007. The role of neuronal complexes in human X-linked brain diseases. *Am J Hum Genet* 80 (2):205-20.
- Lemire, M. 2011. Defining rare variants by their frequencies in controls may increase type I error. *Nat Genet* 43 (5):391-2; author reply 394-5.
- Leung, C. L., M. Zheng, S. M. Prater, and R. K. Liem. 2001. The BPAG1 locus: Alternative splicing produces multiple isoforms with distinct cytoskeletal linker domains, including predominant isoforms in neurons and muscles. *J Cell Biol* 154 (4):691-7.
- Levinson, D. F., J. Duan, S. Oh, K. Wang, A. R. Sanders, J. Shi, N. Zhang, B. J. Mowry, A. Olincy, F. Amin, C. R. Cloninger, J. M. Silverman, N. G. Buccola, W. F. Byerley, D. W. Black, K. S. Kendler, R.

- Freedman, F. Dudbridge, I. Pe'er, H. Hakonarson, S. E. Bergen, A. H. Fanous, P. A. Holmans, and P. V. Gejman. 2011. Copy number variants in schizophrenia: confirmation of five previous findings and new evidence for 3q29 microdeletions and VIPR2 duplications. *Am J Psychiatry* 168 (3):302-16.
- Li, B., and S. M. Leal. 2008. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet* 83 (3):311-21.
 - Lips, E. S., L. N. Cornelisse, R. F. Toonen, J. L. Min, C. M. Hultman, Consortium the International Schizophrenia, P. A. Holmans, M. C. O'Donovan, S. M. Purcell, A. B. Smit, M. Verhage, P. F. Sullivan, P. M. Visscher, and D. Posthuma. 2012. Functional gene group analysis identifies synaptic gene groups as risk factor for schizophrenia. *Mol Psychiatry* 17 (10):996-1006.
 - Lord, C., M. Rutter, S. Goode, J. Heemsbergen, H. Jordan, L. Mawhood, and E. Schopler. 1989. Autism diagnostic observation schedule: a standardized observation of communicative and social behavior. *J Autism Dev Disord* 19 (2):185-212.
 - Lord, C., M. Rutter, and A. Le Couteur. 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):659-85.
 - MacArthur, D. G., S. Balasubramanian, A. Frankish, N. Huang, J. Morris, K. Walter, L. Jostins, L. Habegger, J. K. Pickrell, S. B. Montgomery, C. A. Albers, Z. D. Zhang, D. F. Conrad, G. Lunter, H. Zheng, Q. Ayub, M. A. DePristo, E. Banks, M. Hu, R. E. Handsaker, J. A. Rosenfeld, M. Fromer, M. Jin, X. J. Mu, E. Khurana, K. Ye, M. Kay, G. I. Saunders, M. M. Suner, T. Hunt, I. H. Barnes, C. Amid, D. R. Carvalho-Silva, A. H. Bignell, C. Snow, B. Yngvadottir, S. Bumpstead, D. N. Cooper, Y. Xue, I. G. Romero, Consortium Genomes Project, J. Wang, Y. Li, R. A. Gibbs, S. A. McCarroll, E. T. Dermitzakis, J. K. Pritchard, J. C. Barrett, J. Harrow, M. E. Hurles, M. B. Gerstein, and C. Tyler-Smith. 2012. A systematic survey of loss-of-function variants in human protein-coding genes. *Science* 335 (6070):823-8.
 - Magi, R., A. Kumar, and A. P. Morris. 2011. Assessing the impact of missing genotype data in rare variant association analysis. *BMC Proc* 5 Suppl 9:S107.
 - Malhotra, D., S. McCarthy, J. J. Michaelson, V. Vacic, K. E. Burdick, S. Yoon, S. Cichon, A. Corvin, S. Gary, E. S. Gershon, M. Gill, M. Karayiorgou, J. R. Kelsoe, O. Krastoshevsky, V. Krause, E. Leibenluft, D. L. Levy, V. Makarov, A. Bhandari, A. K. Malhotra, F. J. McMahon, M. M. Nothen, J. B. Potash, M. Rietschel, T. G. Schulze, and J. Sebat. 2011. High frequencies of de novo CNVs in bipolar disorder and schizophrenia. *Neuron* 72 (6):951-63.
 - Malhotra, D., and J. Sebat. 2012. CNVs: harbingers of a rare variant revolution in psychiatric genetics. *Cell* 148 (6):1223-41.
 - Mamanova, L., A. J. Coffey, C. E. Scott, I. Kozarewa, E. H. Turner, A. Kumar, E. Howard, J. Shendure, and D. J. Turner. 2010. Target-enrichment strategies for next-generation sequencing. *Nat Methods* 7 (2):111-8.
 - Manolio, T. A., F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E.

- Eichler, G. Gibson, J. L. Haines, T. F. Mackay, S. A. McCarroll, and P. M. Visscher. 2009. Finding the missing heritability of complex diseases. *Nature* 461 (7265):747-53.
- Mao, Y., X. Ge, C. L. Frank, J. M. Madison, A. N. Koehler, M. K. Doud, C. Tassa, E. M. Berry, T. Soda, K. K. Singh, T. Biechele, T. L. Petryshen, R. T. Moon, S. J. Haggarty, and L. H. Tsai. 2009. Disrupted in schizophrenia 1 regulates neuronal progenitor proliferation via modulation of GSK3 β /beta-catenin signaling. *Cell* 136 (6):1017-31.
 - Marchini, J., and B. Howie. 2010. Genotype imputation for genome-wide association studies. *Nat Rev Genet* 11 (7):499-511.
 - Marshall, C. R., A. Noor, J. B. Vincent, A. C. Lionel, L. Feuk, J. Skaug, M. Shago, R. Moessner, D. Pinto, Y. Ren, B. Thiruvahindrapduram, A. Fiebig, S. Schreiber, J. Friedman, C. E. Ketelaars, Y. J. Vos, C. Ficicioglu, S. Kirkpatrick, R. Nicolson, L. Sloman, A. Summers, C. A. Gibbons, A. Teebi, D. Chitayat, R. Weksberg, A. Thompson, C. Vardy, V. Crosbie, S. Luscombe, R. Baatjes, L. Zwaigenbaum, W. Roberts, B. Fernandez, P. Szatmari, and S. W. Scherer. 2008. Structural variation of chromosomes in autism spectrum disorder. *Am J Hum Genet* 82 (2):477-88.
 - McCarthy, S. E., V. Makarov, G. Kirov, A. M. Addington, J. McClellan, S. Yoon, D. O. Perkins, D. E. Dickel, M. Kusenda, O. Krastoshevsky, V. Krause, R. A. Kumar, D. Grozeva, D. Malhotra, T. Walsh, E. H. Zackai, P. Kaplan, J. Ganesh, I. D. Krantz, N. B. Spinner, P. Roccanova, A. Bhandari, K. Pavon, B. Lakshmi, A. Leotta, J. Kendall, Y. H. Lee, V. Vacic, S. Gary, L. M. Iakoucheva, T. J. Crow, S. L. Christian, J. A. Lieberman, T. S. Stroup, T. Lehtimaki, K. Puura, C. Haldeman-Englert, J. Pearl, M. Goodell, V. L. Willour, P. Derosse, J. Steele, L. Kassem, J. Wolff, N. Chitkara, F. J. McMahon, A. K. Malhotra, J. B. Potash, T. G. Schulze, M. M. Nothen, S. Cichon, M. Rietschel, E. Leibenluft, V. Kustanovich, C. M. Lajonchere, J. S. Sutcliffe, D. Skuse, M. Gill, L. Gallagher, N. R. Mendell, Consortium Wellcome Trust Case Control, N. Craddock, M. J. Owen, M. C. O'Donovan, T. H. Shaikh, E. Susser, L. E. Delisi, P. F. Sullivan, C. K. Deutsch, J. Rapoport, D. L. Levy, M. C. King, and J. Sebat. 2009. Microduplications of 16p11.2 are associated with schizophrenia. *Nat Genet* 41 (11):1223-7.
 - Mefford, H. C., A. J. Sharp, C. Baker, A. Itsara, Z. Jiang, K. Buysse, S. Huang, V. K. Maloney, J. A. Crolla, D. Baralle, A. Collins, C. Mercer, K. Norga, T. de Ravel, K. Devriendt, E. M. Bongers, N. de Leeuw, W. Reardon, S. Gimelli, F. Bena, R. C. Hennekam, A. Male, L. Gaunt, J. Clayton-Smith, I. Simoncic, S. M. Park, S. G. Mehta, S. Nik-Zainal, C. G. Woods, H. V. Firth, G. Parkin, M. Fichera, S. Reitano, M. Lo Giudice, K. E. Li, I. Casuga, A. Broome, B. Conrad, M. Schwerzmann, L. Raber, S. Gallati, P. Striano, A. Coppola, J. L. Tolmie, E. S. Tobias, C. Lilley, L. Armengol, Y. Spysschaert, P. Verloo, A. De Coene, L. Goossens, G. Mortier, F. Speleman, E. van Binsbergen, M. R. Nelen, R. Hochstenbach, M. Poot, L. Gallagher, M. Gill, J. McClellan, M. C. King, R. Regan, C. Skinner, R. E. Stevenson, S. E. Antonarakis, C. Chen, X. Estivill, B. Menten, G. Gimelli, S. Gribble, S. Schwartz, J. S. Sutcliffe, T. Walsh, S. J. Knight, J. Sebat, C. Romano, C. E. Schwartz, J. A. Veltman, B. B. de Vries, J. R. Vermeesch, J. C. Barber, L. Willatt, M. Tassabehji, and E. E. Eichler. 2008. Recurrent rearrangements of chromosome 1q21.1 and variable pediatric phenotypes. *N Engl J Med* 359 (16):1685-99.
 - Metzker, M. L. 2010. Sequencing technologies - the next generation. *Nat Rev Genet* 11 (1):31-46.

- Mi, H., N. Guo, A. Kejariwal, and P. D. Thomas. 2007. PANTHER version 6: protein sequence and function evolution data with expanded representation of biological pathways. *Nucleic Acids Res* 35 (Database issue):D247-52.
- Miles, J. H. 2011. Autism spectrum disorders--a genetics review. *Genet Med* 13 (4):278-94.
- Miller, D. T., Y. Shen, L. A. Weiss, J. Korn, I. Anselm, C. Bridgemohan, G. F. Cox, H. Dickinson, J. Gentile, D. J. Harris, V. Hegde, R. Hundley, O. Khwaja, S. Kothare, C. Luedke, R. Nasir, A. Poduri, K. Prasad, P. Raffalli, A. Reinhard, S. E. Smith, M. M. Sobeih, J. S. Soul, J. Stoler, M. Takeoka, W. H. Tan, J. Thakuria, R. Wolff, R. Yusupov, J. F. Gusella, M. J. Daly, and B. L. Wu. 2009. Microdeletion/duplication at 15q13.2q13.3 among individuals with features of autism and other neuropsychiatric disorders. *J Med Genet* 46 (4):242-8.
- Moens, L. N., P. De Rijk, J. Reumers, M. J. Van den Bossche, W. Glassee, S. De Zutter, A. S. Lenaerts, A. Nordin, L. G. Nilsson, I. Medina Castello, K. F. Norrback, D. Goossens, K. Van Steen, R. Adolfsson, and J. Del-Favero. 2011. Sequencing of DISC1 pathway genes reveals increased burden of rare missense variants in schizophrenia patients from a northern Swedish population. *PLoS One* 6 (8):e23450.
- Morita, A., N. Yamashita, Y. Sasaki, Y. Uchida, O. Nakajima, F. Nakamura, T. Yagi, M. Taniguchi, H. Usui, R. Katoh-Semba, K. Takei, and Y. Goshima. 2006. Regulation of dendritic branching and spine maturation by semaphorin3A-Fyn signaling. *J Neurosci* 26 (11):2971-80.
- Morris, A. P., and E. Zeggini. 2010. An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet Epidemiol* 34 (2):188-93.
- Mudge, J., N. A. Miller, I. Khrebtukova, I. E. Lindquist, G. D. May, J. J. Huntley, S. Luo, L. Zhang, J. C. van Velkinburgh, A. D. Farmer, S. Lewis, W. D. Beavis, F. D. Schilkey, S. M. Virk, C. F. Black, M. K. Myers, L. C. Mader, R. J. Langley, J. P. Utsey, R. W. Kim, R. C. Roberts, S. K. Khalsa, M. Garcia, V. Ambriz-Griffith, R. Harlan, W. Czika, S. Martin, R. D. Wolfinger, N. I. Perrone-Bizzozero, G. P. Schroth, and S. F. Kingsmore. 2008. Genomic convergence analysis of schizophrenia: mRNA sequencing reveals altered synaptic vesicular transport in post-mortem cerebellum. *PLoS One* 3 (11):e3625.
- Mukherjee, K., M. Sharma, H. Urlaub, G. P. Bourenkov, R. Jahn, T. C. Sudhof, and M. C. Wahl. 2008. CASK Functions as a Mg²⁺-independent neurexin kinase. *Cell* 133 (2):328-39.
- Myers, R. A., F. Casals, J. Gauthier, F. F. Hamdan, J. Keebler, A. R. Boyko, C. D. Bustamante, A. M. Piton, D. Spiegelman, E. Henrion, M. Zilversmit, J. Hussin, J. Quinlan, Y. Yang, R. G. Lafreniere, A. R. Griffing, E. A. Stone, G. A. Rouleau, and P. Awadalla. 2011. A population genetic approach to mapping neurological disorder genes using deep resequencing. *PLoS Genet* 7 (2):e1001318.
- Myers, S. M., C. P. Johnson, and American Academy of Pediatrics Council on Children With Disabilities. 2007. Management of children with autism spectrum disorders. *Pediatrics* 120 (5):1162-82.
- Neale, B. M., Y. Kou, L. Liu, A. Ma'ayan, K. E. Samocha, A. Sabo, C. F. Lin, C. Stevens, L. S. Wang, V. Makarov, P. Polak, S. Yoon, J. Maguire, E. L. Crawford, N. G. Campbell, E. T. Geller, O. Valladares, C. Schafer, H. Liu, T. Zhao, G. Cai, J. Lihm, R. Dannenfelser, O. Jabado, Z. Peralta, U. Nagaswamy, D. Muzny, J. G. Reid, I. Newsham, Y. Wu, L. Lewis, Y.

- Han, B. F. Voight, E. Lim, E. Rossin, A. Kirby, J. Flannick, M. Fromer, K. Shakir, T. Fennell, K. Garimella, E. Banks, R. Poplin, S. Gabriel, M. DePristo, J. R. Wimbish, B. E. Boone, S. E. Levy, C. Betancur, S. Sunyaev, E. Boerwinkle, J. D. Buxbaum, E. H. Cook, Jr., B. Devlin, R. A. Gibbs, K. Roeder, G. D. Schellenberg, J. S. Sutcliffe, and M. J. Daly. 2012. Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* 485 (7397):242-5.
- Nechiporuk, T., T. E. Fernandez, and V. Vasioukhin. 2007. Failure of epithelial tube maintenance causes hydrocephalus and renal cysts in Dlg5^{-/-} mice. *Dev Cell* 13 (3):338-50.
 - Ng, S. B., E. H. Turner, P. D. Robertson, S. D. Flygare, A. W. Bigham, C. Lee, T. Shaffer, M. Wong, A. Bhattacharjee, E. E. Eichler, M. Bamshad, D. A. Nickerson, and J. Shendure. 2009. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* 461 (7261):272-6.
 - Nomura, M., Y. Ueta, R. Serino, Y. Yamamoto, I. Shibuya, and H. Yamashita. 2000. Upregulation of synapsin IIa and IIb mRNAs in the paraventricular and supraoptic nuclei in chronic salt loaded and lactating rats. *Neurosci Res* 37 (3):201-10.
 - O'Donovan, M. C., N. Craddock, N. Norton, H. Williams, T. Peirce, V. Moskvina, I. Nikolov, M. Hamshere, L. Carroll, L. Georgieva, S. Dwyer, P. Holmans, J. L. Marchini, C. C. Spencer, B. Howie, H. T. Leung, A. M. Hartmann, H. J. Moller, D. W. Morris, Y. Shi, G. Feng, P. Hoffmann, P. Propping, C. Vasilescu, W. Maier, M. Rietschel, S. Zammit, J. Schumacher, E. M. Quinn, T. G. Schulze, N. M. Williams, I. Giegling, N. Iwata, M. Ikeda, A. Darvasi, S. Shifman, L. He, J. Duan, A. R. Sanders, D. F. Levinson, P. V. Gejman, S. Cichon, M. M. Nothen, M. Gill, A. Corvin, D. Rujescu, G. Kirov, M. J. Owen, N. G. Buccola, B. J. Mowry, R. Freedman, F. Amin, D. W. Black, J. M. Silverman, W. F. Byerley, C. R. Cloninger, and Collaboration Molecular Genetics of Schizophrenia. 2008. Identification of loci associated with schizophrenia by genome-wide association and follow-up. *Nat Genet* 40 (9):1053-5.
 - O'Donovan, M. C., N. Norton, H. Williams, T. Peirce, V. Moskvina, I. Nikolov, M. Hamshere, L. Carroll, L. Georgieva, S. Dwyer, P. Holmans, J. L. Marchini, C. C. Spencer, B. Howie, H. T. Leung, I. Giegling, A. M. Hartmann, H. J. Moller, D. W. Morris, Y. Shi, G. Feng, P. Hoffmann, P. Propping, C. Vasilescu, W. Maier, M. Rietschel, S. Zammit, J. Schumacher, E. M. Quinn, T. G. Schulze, N. Iwata, M. Ikeda, A. Darvasi, S. Shifman, L. He, J. Duan, A. R. Sanders, D. F. Levinson, R. Adolfsson, U. Osby, L. Terenius, E. G. Jonsson, S. Cichon, M. M. Nothen, M. Gill, A. P. Corvin, D. Rujescu, P. V. Gejman, G. Kirov, N. Craddock, N. M. Williams, M. J. Owen, and Collaboration Molecular Genetics of Schizophrenia. 2009. Analysis of 10 independent samples provides evidence for association between schizophrenia and a SNP flanking fibroblast growth factor receptor 2. *Mol Psychiatry* 14 (1):30-6.
 - O'Dushlaine, C., E. Kenny, E. Heron, G. Donohoe, M. Gill, D. Morris, Consortium International Schizophrenia, and A. Corvin. 2011. Molecular pathways involved in neuronal cell adhesion and membrane scaffolding contribute to schizophrenia and bipolar disorder susceptibility. *Mol Psychiatry* 16 (3):286-92.
 - O'Roak, B. J., P. Deriziotis, C. Lee, L. Vives, J. J. Schwartz, S. Girirajan, E. Karakoc, A. P. Mackenzie, S. B. Ng, C. Baker, M. J. Rieder, D. A. Nickerson, R. Bernier, S. E. Fisher, J. Shendure, and E. E. Eichler. 2011.

Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nat Genet* 43 (6):585-9.

- O'Roak, B. J., L. Vives, S. Girirajan, E. Karakoc, N. Krumm, B. P. Coe, R. Levy, A. Ko, C. Lee, J. D. Smith, E. H. Turner, I. B. Stanaway, B. Vernet, M. Malig, C. Baker, B. Reilly, J. M. Akey, E. Borenstein, M. J. Rieder, D. A. Nickerson, R. Bernier, J. Shendure, and E. E. Eichler. 2012. Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature*.
- Passos-Bueno, M. R., E. Bakker, A. L. Kneppers, R. I. Takata, D. Rapaport, J. T. den Dunnen, M. Zatz, and G. J. van Ommen. 1992. Different mosaicism frequencies for proximal and distal Duchenne muscular dystrophy (DMD) mutations indicate difference in etiology and recurrence risk. *Am J Hum Genet* 51 (5):1150-5.
- Pedersen, C. B., and P. B. Mortensen. 2001. Family history, place and season of birth as risk factors for schizophrenia in Denmark: a replication and reanalysis. *Br J Psychiatry* 179:46-52.
- Picchioni, M. M., and R. M. Murray. 2007. Schizophrenia. *BMJ* 335 (7610):91-5.
- Pinkel, D., R. Seagraves, D. Sudar, S. Clark, I. Poole, D. Kowbel, C. Collins, W. L. Kuo, C. Chen, Y. Zhai, S. H. Dairkee, B. M. Ljung, J. W. Gray, and D. G. Albertson. 1998. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat Genet* 20 (2):207-11.
- Pinto, D., A. T. Pagnamenta, L. Klei, R. Anney, D. Merico, R. Regan, J. Conroy, T. R. Magalhaes, C. Correia, B. S. Abrahams, J. Almeida, E. Bacchelli, G. D. Bader, A. J. Bailey, G. Baird, A. Battaglia, T. Berney, N. Bolshakova, S. Bolte, P. F. Bolton, T. Bourgeron, S. Brennan, J. Brian, S. E. Bryson, A. R. Carson, G. Casallo, J. Casey, B. H. Chung, L. Cochrane, C. Corsello, E. L. Crawford, A. Crossett, C. Cytrynbaum, G. Dawson, M. de Jonge, R. Delorme, I. Drmic, E. Duketis, F. Duque, A. Estes, P. Farrar, B. A. Fernandez, S. E. Folstein, E. Fombonne, C. M. Freitag, J. Gilbert, C. Gillberg, J. T. Glessner, J. Goldberg, A. Green, J. Green, S. J. Guter, H. Hakonarson, E. A. Heron, M. Hill, R. Holt, J. L. Howe, G. Hughes, V. Hus, R. Igliozzi, C. Kim, S. M. Klauck, A. Kolevzon, O. Korvatska, V. Kustanovich, C. M. Lajonchere, J. A. Lamb, M. Laskawiec, M. Leboyer, A. Le Couteur, B. L. Leventhal, A. C. Lionel, X. Q. Liu, C. Lord, L. Lotspeich, S. C. Lund, E. Maestrini, W. Mahoney, C. Mantoulan, C. R. Marshall, H. McConachie, C. J. McDougle, J. McGrath, W. M. McMahon, A. Merikangas, O. Migita, N. J. Minshew, G. K. Mirza, J. Munson, S. F. Nelson, C. Noakes, A. Noor, G. Nygren, G. Oliveira, K. Papanikolaou, J. R. Parr, B. Parrini, T. Paton, A. Pickles, M. Pilorge, J. Piven, C. P. Ponting, D. J. Posey, A. Poustka, F. Poustka, A. Prasad, J. Ragoussis, K. Renshaw, J. Rickaby, W. Roberts, K. Roeder, B. Roge, M. L. Rutter, L. J. Bierut, J. P. Rice, J. Salt, K. Sansom, D. Sato, R. Segurado, A. F. Sequeira, L. Senman, N. Shah, V. C. Sheffield, L. Soorya, I. Sousa, O. Stein, N. Sykes, V. Stoppioni, C. Strawbridge, R. Tancredi, K. Tansey, B. Thiruvahindrapduram, A. P. Thompson, S. Thomson, A. Tryfon, J. Tsiantis, H. Van Engeland, J. B. Vincent, F. Volkmar, S. Wallace, K. Wang, Z. Wang, T. H. Wassink, C. Webber, R. Weksberg, K. Wing, K. Wittemeyer, S. Wood, J. Wu, B. L. Yaspan, D. Zurawiecki, L. Zwaigenbaum, J. D. Buxbaum, R. M. Cantor, E. H. Cook, H. Coon, M. L. Cuccaro, B. Devlin, S. Ennis, L. Gallagher, D. H. Geschwind, M. Gill, J. L. Haines, J. Hallmayer, J. Miller, A. P. Monaco, J. I. Nurnberger, Jr., A. D. Paterson, M. A. Pericak-Vance, G. D.

- Schellenberg, P. Szatmari, A. M. Vicente, V. J. Vieland, E. M. Wijsman, S. W. Scherer, J. S. Sutcliffe, and C. Betancur. 2010. Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* 466 (7304):368-72.
- Pollard, K. S., M. J. Hubisz, K. R. Rosenbloom, and A. Siepel. 2010. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* 20 (1):110-21.
 - Porteous, D., and K. Millar. 2009. How DISC1 regulates postnatal brain development: girdin gets in on the AKT. *Neuron* 63 (6):711-3.
 - Rapoport, J. L., J. N. Giedd, and N. Gogtay. 2012. Neurodevelopmental model of schizophrenia: update 2012. *Mol Psychiatry*.
 - Reva, B., Y. Antipin, and C. Sander. 2007. Determinants of protein function revealed by combinatorial entropy optimization. *Genome Biol* 8 (11):R232.
 - ———. 2011. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res* 39 (17):e118.
 - Riordan, J. R., J. M. Rommens, B. Kerem, N. Alon, R. Rozmahel, Z. Grzelczak, J. Zielenski, S. Lok, N. Plavsic, J. L. Chou, and et al. 1989. Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science* 245 (4922):1066-73.
 - Rizzolatti, G., and M. Fabbri-Destro. 2010. Mirror neurons: from discovery to autism. *Exp Brain Res* 200 (3-4):223-37.
 - Rujescu, D., A. Ingason, S. Cichon, O. P. Pietilainen, M. R. Barnes, T. Touloupoulou, M. Picchioni, E. Vassos, U. Ettinger, E. Bramon, R. Murray, M. Ruggeri, S. Tosato, C. Bonetto, S. Steinberg, E. Sigurdsson, T. Sigmundsson, H. Petursson, A. Gylfason, P. I. Olason, G. Hardarsson, G. A. Jonsdottir, O. Gustafsson, R. Fossdal, I. Giegling, H. J. Moller, A. M. Hartmann, P. Hoffmann, C. Crombie, G. Fraser, N. Walker, J. Lonnqvist, J. Suvisaari, A. Tuulio-Henriksson, S. Djurovic, I. Melle, O. A. Andreassen, T. Hansen, T. Werge, L. A. Kiemeney, B. Franke, J. Veltman, J. E. Buizer-Voskamp, Group Investigators, C. Sabatti, R. A. Ophoff, M. Rietschel, M. M. Nothen, K. Stefansson, L. Peltonen, D. St Clair, H. Stefansson, and D. A. Collier. 2009. Disruption of the neurexin 1 gene is associated with schizophrenia. *Hum Mol Genet* 18 (5):988-96.
 - Runker, A. E., C. O'Tuathaigh, M. Dunleavy, D. W. Morris, G. E. Little, A. P. Corvin, M. Gill, D. C. Henshall, J. L. Waddington, and K. J. Mitchell. 2011. Mutation of Semaphorin-6A disrupts limbic and cortical connectivity and models neurodevelopmental psychopathology. *PLoS One* 6 (11):e26488.
 - Sakai, Y., C. A. Shaw, B. C. Dawson, D. V. Dugas, Z. Al-Mohtaseb, D. E. Hill, and H. Y. Zoghbi. 2011. Protein interactome reveals converging molecular pathways among autism disorders. *Sci Transl Med* 3 (86):86ra49.
 - Sanders, S. J., M. T. Murtha, A. R. Gupta, J. D. Murdoch, M. J. Raubeson, A. J. Willsey, A. G. Ercan-Sencicek, N. M. Dilullo, N. N. Parikshak, J. L. Stein, M. F. Walker, G. T. Ober, N. A. Teran, Y. Song, P. El-Fishawy, R. C. Murtha, M. Choi, J. D. Overton, R. D. Bjornson, N. J. Carriero, K. A. Meyer, K. Bilguvar, S. M. Mane, N. Sestan, R. P. Lifton, M. Gunel, K. Roeder, D. H. Geschwind, B. Devlin, and M. W. State. 2012. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature*.
 - Scarr, E., and B. Dean. 2009. Role of the cholinergic system in the pathology and treatment of schizophrenia. *Expert Rev Neurother* 9 (1):73-86.

- Schizophrenia Psychiatric Genome-Wide Association Study, Consortium. 2011. Genome-wide association study identifies five new schizophrenia loci. *Nat Genet* 43 (10):969-76.
- Schmitt, A., A. Hasan, O. Gruber, and P. Falkai. 2011. Schizophrenia as a disorder of disconnectivity. *Eur Arch Psychiatry Clin Neurosci* 261 Suppl 2:S150-4.
- Sebat, J., B. Lakshmi, D. Malhotra, J. Troge, C. Lese-Martin, T. Walsh, B. Yamrom, S. Yoon, A. Krasnitz, J. Kendall, A. Leotta, D. Pai, R. Zhang, Y. H. Lee, J. Hicks, S. J. Spence, A. T. Lee, K. Puura, T. Lehtimäki, D. Ledbetter, P. K. Gregersen, J. Bregman, J. S. Sutcliffe, V. Jobanputra, W. Chung, D. Warburton, M. C. King, D. Skuse, D. H. Geschwind, T. C. Gilliam, K. Ye, and M. Wigler. 2007. Strong association of de novo copy number mutations with autism. *Science* 316 (5823):445-9.
- Sheng, M., and E. Kim. 2000. The Shank family of scaffold proteins. *J Cell Sci* 113 (Pt 11):1851-6.
- Shi, Y., Z. Li, Q. Xu, T. Wang, T. Li, J. Shen, F. Zhang, J. Chen, G. Zhou, W. Ji, B. Li, Y. Xu, D. Liu, P. Wang, P. Yang, B. Liu, W. Sun, C. Wan, S. Qin, G. He, S. Steinberg, S. Cichon, T. Werge, E. Sigurdsson, S. Tosato, A. Palotie, M. M. Nothen, M. Rietschel, R. A. Ophoff, D. A. Collier, D. Rujescu, D. S. Clair, H. Stefansson, K. Stefansson, J. Ji, Q. Wang, W. Li, L. Zheng, H. Zhang, G. Feng, and L. He. 2011. Common variants on 8p12 and 1q24.2 confer risk of schizophrenia. *Nat Genet* 43 (12):1224-7.
- Song, W., W. Li, J. Feng, L. L. Heston, W. A. Scaringe, and S. S. Sommer. 2008. Identification of high risk DISC1 structural variants with a 2% attributable risk for schizophrenia. *Biochem Biophys Res Commun* 367 (3):700-6.
- Sonnenberg, A., and R. K. Liem. 2007. Plakins in development and disease. *Exp Cell Res* 313 (10):2189-203.
- Sorensen, H. J., E. L. Mortensen, J. Schiffman, J. M. Reinisch, J. Maeda, and S. A. Mednick. 2010. Early developmental milestones and risk of schizophrenia: a 45-year follow-up of the Copenhagen Perinatal Cohort. *Schizophr Res* 118 (1-3):41-7.
- Sparrow, S., D. Cicchetti, and D. Balla. 2005. Vineland Adaptive Behavior Scales, Second Edition (Vineland-II).
- St Clair, D., D. Blackwood, W. Muir, A. Carothers, M. Walker, G. Spowart, C. Gosden, and H. J. Evans. 1990. Association within a family of a balanced autosomal translocation with major mental illness. *Lancet* 336 (8706):13-6.
- Stankovic, M., A. Lakic, and N. Ilic. 2012. Autism and autistic spectrum disorders in the context of new DSM-V classification, and clinical and epidemiological data. *Srp Arh Celok Lek* 140 (3-4):236-43.
- Steen, R. G., C. Mull, R. McClure, R. M. Hamer, and J. A. Lieberman. 2006. Brain volume in first-episode schizophrenia: systematic review and meta-analysis of magnetic resonance imaging studies. *Br J Psychiatry* 188:510-8.
- Stefansson, H., R. A. Ophoff, S. Steinberg, O. A. Andreassen, S. Cichon, D. Rujescu, T. Werge, O. P. Pietiläinen, O. Mors, P. B. Mortensen, E. Sigurdsson, O. Gustafsson, M. Nyegaard, A. Tuulio-Henriksson, A. Ingason, T. Hansen, J. Suvisaari, J. Lonnqvist, T. Paunio, A. D. Borglum, A. Hartmann, A. Fink-Jensen, M. Nordentoft, D. Hougaard, B. Norgaard-Pedersen, Y. Bottcher, J. Olesen, R. Breuer, H. J. Moller, I. Giegling, H. B. Rasmussen, S. Timm, M. Mattheisen, I. Bitter, J. M. Rethelyi, B. B. Magnusdottir, T. Sigmundsson, P. Olason, G. Masson, J. R. Gulcher, M.

- Haraldsson, R. Fossdal, T. E. Thorgeirsson, U. Thorsteinsdottir, M. Ruggeri, S. Tosato, B. Franke, E. Strengman, L. A. Kiemeney, Risk Genetic, Psychosis Outcome in, I. Melle, S. Djurovic, L. Abramova, V. Kaleda, J. Sanjuan, R. de Frutos, E. Bramon, E. Vassos, G. Fraser, U. Ettinger, M. Picchioni, N. Walker, T. Touloupoulou, A. C. Need, D. Ge, J. L. Yoon, K. V. Shianna, N. B. Freimer, R. M. Cantor, R. Murray, A. Kong, V. Golimbet, A. Carracedo, C. Arango, J. Costas, E. G. Jonsson, L. Terenius, I. Agartz, H. Petursson, M. M. Nothen, M. Rietschel, P. M. Matthews, P. Muglia, L. Peltonen, D. St Clair, D. B. Goldstein, K. Stefansson, and D. A. Collier. 2009. Common variants conferring risk of schizophrenia. *Nature* 460 (7256):744-7.
- Stefansson, H., D. Rujescu, S. Cichon, O. P. Pietilainen, A. Ingason, S. Steinberg, R. Fossdal, E. Sigurdsson, T. Sigmundsson, J. E. Buizer-Voskamp, T. Hansen, K. D. Jakobsen, P. Muglia, C. Francks, P. M. Matthews, A. Gylfason, B. V. Halldorsson, D. Gudbjartsson, T. E. Thorgeirsson, A. Sigurdsson, A. Jonasdottir, A. Jonasdottir, A. Bjornsson, S. Mattiasdottir, T. Blondal, M. Haraldsson, B. B. Magnusdottir, I. Giegling, H. J. Moller, A. Hartmann, K. V. Shianna, D. Ge, A. C. Need, C. Crombie, G. Fraser, N. Walker, J. Lonnqvist, J. Suvisaari, A. Tuulio-Henriksson, T. Paunio, T. Touloupoulou, E. Bramon, M. Di Forti, R. Murray, M. Ruggeri, E. Vassos, S. Tosato, M. Walshe, T. Li, C. Vasilescu, T. W. Muhleisen, A. G. Wang, H. Ullum, S. Djurovic, I. Melle, J. Olesen, L. A. Kiemeney, B. Franke, Group, C. Sabatti, N. B. Freimer, J. R. Gulcher, U. Thorsteinsdottir, A. Kong, O. A. Andreassen, R. A. Ophoff, A. Georgi, M. Rietschel, T. Werge, H. Petursson, D. B. Goldstein, M. M. Nothen, L. Peltonen, D. A. Collier, D. St Clair, and K. Stefansson. 2008. Large recurrent microdeletions associated with schizophrenia. *Nature* 455 (7210):232-6.
 - Steinberg, S., O. Mors, A. D. Borglum, O. Gustafsson, T. Werge, P. B. Mortensen, O. A. Andreassen, E. Sigurdsson, T. E. Thorgeirsson, Y. Bottcher, P. Olason, R. A. Ophoff, S. Cichon, I. H. Gudjonsdottir, O. P. Pietilainen, M. Nyegaard, A. Tuulio-Henriksson, A. Ingason, T. Hansen, L. Athanasiu, J. Suvisaari, J. Lonnqvist, T. Paunio, A. Hartmann, G. Jurgens, M. Nordentoft, D. Hougaard, B. Norgaard-Pedersen, R. Breuer, H. J. Moller, I. Giegling, B. Glenthoj, H. B. Rasmussen, M. Mattheisen, I. Bitter, J. M. Rethelyi, T. Sigmundsson, R. Fossdal, U. Thorsteinsdottir, M. Ruggeri, S. Tosato, E. Strengman, Risk Genetic, Psychosis Outcome in, L. A. Kiemeney, I. Melle, S. Djurovic, L. Abramova, V. Kaleda, M. Walshe, E. Bramon, E. Vassos, T. Li, G. Fraser, N. Walker, T. Touloupoulou, J. Yoon, N. B. Freimer, R. M. Cantor, R. Murray, A. Kong, V. Golimbet, E. G. Jonsson, L. Terenius, I. Agartz, H. Petursson, M. M. Nothen, M. Rietschel, L. Peltonen, D. Rujescu, D. A. Collier, H. Stefansson, D. St Clair, and K. Stefansson. 2011. Expanding the range of ZNF804A variants conferring risk of psychosis. *Mol Psychiatry* 16 (1):59-66.
 - Sudhof, T. C. 2008. Neuroligins and neurexins link synaptic function to cognitive disease. *Nature* 455 (7215):903-11.
 - Sullivan, P., and Investigators Psychiatric Genetics. 2012. Don't give up on GWAS. *Mol Psychiatry* 17 (1):2-3.
 - Suzuki, T., and K. Okumura-Noji. 1995. NMDA receptor subunits epsilon 1 (NR2A) and epsilon 2 (NR2B) are substrates for Fyn in the postsynaptic density fraction isolated from the rat brain. *Biochem Biophys Res Commun* 216 (2):582-8.

- Talkowski, M. E., J. A. Rosenfeld, I. Blumenthal, V. Pillalamarri, C. Chiang, A. Heilbut, C. Ernst, C. Hanscom, E. Rossin, A. M. Lindgren, S. Pereira, D. Ruderfer, A. Kirby, S. Ripke, D. J. Harris, J. H. Lee, K. Ha, H. G. Kim, B. D. Solomon, A. L. Gropman, D. Lucente, K. Sims, T. K. Ohsumi, M. L. Borowsky, S. Loranger, B. Quade, K. Lage, J. Miles, B. L. Wu, Y. Shen, B. Neale, L. G. Shaffer, M. J. Daly, C. C. Morton, and J. F. Gusella. 2012. Sequencing Chromosomal Abnormalities Reveals Neurodevelopmental Loci that Confer Risk across Diagnostic Boundaries. *Cell* 149 (3):525-37.
- Tang, Y. P., E. Shimizu, G. R. Dube, C. Rampon, G. A. Kerchner, M. Zhuo, G. Liu, and J. Z. Tsien. 1999. Genetic enhancement of learning and memory in mice. *Nature* 401 (6748):63-9.
- Trikalinos, T. A., A. Karvouni, E. Zintzaras, T. Ylisaukko-oja, L. Peltonen, I. Jarvela, and J. P. Ioannidis. 2006. A heterogeneity-based genome search meta-analysis for autism-spectrum disorders. *Mol Psychiatry* 11 (1):29-36.
- Tuchman, R., and I. Rapin. 2002. Epilepsy in autism. *Lancet Neurol* 1 (6):352-8.
- van Os, J., G. Kenis, and B. P. Rutten. 2010. The environment and schizophrenia. *Nature* 468 (7321):203-12.
- Verpelli, C., G. Piccoli, C. Zibetti, A. Zanchi, F. Gardoni, K. Huang, D. Brambilla, M. Di Luca, E. Battaglioli, and C. Sala. 2010. Synaptic activity controls dendritic spine morphology by modulating eEF2-dependent BDNF synthesis. *J Neurosci* 30 (17):5830-42.
- Voineagu, I., X. Wang, P. Johnston, J. K. Lowe, Y. Tian, S. Horvath, J. Mill, R. M. Cantor, B. J. Blencowe, and D. H. Geschwind. 2011. Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* 474 (7351):380-4.
- Vrijenhoek, T., J. E. Buizer-Voskamp, I. van der Stelt, E. Strengman, Risk Genetic Consortium Outcome in Psychosis, C. Sabatti, A. Geurts van Kessel, H. G. Brunner, R. A. Ophoff, and J. A. Veltman. 2008. Recurrent CNVs disrupt three candidate genes in schizophrenia patients. *Am J Hum Genet* 83 (4):504-10.
- Walsh, T., J. M. McClellan, S. E. McCarthy, A. M. Addington, S. B. Pierce, G. M. Cooper, A. S. Nord, M. Kusenda, D. Malhotra, A. Bhandari, S. M. Stray, C. F. Rippey, P. Roccanova, V. Makarov, B. Lakshmi, R. L. Findling, L. Sikich, T. Stromberg, B. Merriman, N. Gogtay, P. Butler, K. Eckstrand, L. Noory, P. Gochman, R. Long, Z. Chen, S. Davis, C. Baker, E. E. Eichler, P. S. Meltzer, S. F. Nelson, A. B. Singleton, M. K. Lee, J. L. Rapoport, M. C. King, and J. Sebat. 2008. Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science* 320 (5875):539-43.
- Wang, D., Z. Cui, Q. Zeng, H. Kuang, L. P. Wang, J. Z. Tsien, and X. Cao. 2009. Genetic enhancement of memory and long-term potentiation but not CA1 long-term depression in NR2B transgenic rats. *PLoS One* 4 (10):e7486.
- Wang, K., M. Li, and H. Hakonarson. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 38 (16):e164.
- Wang, K., H. Zhang, D. Ma, M. Bucan, J. T. Glessner, B. S. Abrahams, D. Salyakina, M. Imielinski, J. P. Bradfield, P. M. Sleiman, C. E. Kim, C. Hou, E. Frackelton, R. Chiavacci, N. Takahashi, T. Sakurai, E. Rappaport, C. M. Lajonchere, J. Munson, A. Estes, O. Korvatska, J. Piven, L. I. Sonnenblick, A. I. Alvarez Retuerto, E. I. Herman, H. Dong,

- T. Hutman, M. Sigman, S. Ozonoff, A. Klin, T. Owley, J. A. Sweeney, C. W. Brune, R. M. Cantor, R. Bernier, J. R. Gilbert, M. L. Cuccaro, W. M. McMahon, J. Miller, M. W. State, T. H. Wassink, H. Coon, S. E. Levy, R. T. Schultz, J. I. Nurnberger, J. L. Haines, J. S. Sutcliffe, E. H. Cook, N. J. Minshew, J. D. Buxbaum, G. Dawson, S. F. Grant, D. H. Geschwind, M. A. Pericak-Vance, G. D. Schellenberg, and H. Hakonarson. 2009. Common genetic variants on 5p14.1 associate with autism spectrum disorders. *Nature* 459 (7246):528-33.
- Wass, S. 2011. Distortions and disconnections: disrupted brain connectivity in autism. *Brain Cogn* 75 (1):18-28.
 - Weiss, L. A., D. E. Arking, Hopkins Gene Discovery Project of Johns, Consortium the Autism, M. J. Daly, and A. Chakravarti. 2009. A genome-wide linkage and association scan reveals novel loci for autism. *Nature* 461 (7265):802-8.
 - Weiss, L. A., Y. Shen, J. M. Korn, D. E. Arking, D. T. Miller, R. Fossdal, E. Saemundsen, H. Stefansson, M. A. Ferreira, T. Green, O. S. Platt, D. M. Ruderfer, C. A. Walsh, D. Altshuler, A. Chakravarti, R. E. Tanzi, K. Stefansson, S. L. Santangelo, J. F. Gusella, P. Sklar, B. L. Wu, M. J. Daly, and Consortium Autism. 2008. Association between microdeletion and microduplication at 16p11.2 and autism. *N Engl J Med* 358 (7):667-75.
 - Williams, H. J., N. Craddock, G. Russo, M. L. Hamshere, V. Moskvina, S. Dwyer, R. L. Smith, E. Green, D. Grozeva, P. Holmans, M. J. Owen, and M. C. O'Donovan. 2011. Most genome-wide significant susceptibility loci for schizophrenia and bipolar disorder reported to date cross-traditional diagnostic boundaries. *Hum Mol Genet* 20 (2):387-91.
 - Williams, H. J., N. Norton, S. Dwyer, V. Moskvina, I. Nikolov, L. Carroll, L. Georgieva, N. M. Williams, D. W. Morris, E. M. Quinn, I. Giegling, M. Ikeda, J. Wood, T. Lencz, C. Hultman, P. Lichtenstein, D. Thiselton, B. S. Maher, Sgene-plus Group Molecular Genetics of Schizophrenia Collaboration International Schizophrenia Consortium, A. K. Malhotra, B. Riley, K. S. Kendler, M. Gill, P. Sullivan, P. Sklar, S. Purcell, V. L. Nimgaonkar, G. Kirov, P. Holmans, A. Corvin, D. Rujescu, N. Craddock, M. J. Owen, and M. C. O'Donovan. 2011. Fine mapping of ZNF804A and genome-wide significant evidence for its involvement in schizophrenia and bipolar disorder. *Mol Psychiatry* 16 (4):429-41.
 - Woo, T. U., A. M. Kim, and E. Viscidi. 2008. Disease-specific alterations in glutamatergic neurotransmission on inhibitory interneurons in the prefrontal cortex in schizophrenia. *Brain Res* 1218:267-77.
 - Wu, M. C., S. Lee, T. Cai, Y. Li, M. Boehnke, and X. Lin. 2011. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* 89 (1):82-93.
 - Xu, B., I. Ionita-Laza, J. L. Roos, B. Boone, S. Woodrick, Y. Sun, S. Levy, J. A. Gogos, and M. Karayiorgou. 2012. De novo gene mutations highlight patterns of genetic and neural complexity in schizophrenia. *Nat Genet*.
 - Xu, B., J. L. Roos, P. Dexheimer, B. Boone, B. Plummer, S. Levy, J. A. Gogos, and M. Karayiorgou. 2011. Exome sequencing supports a de novo mutational paradigm for schizophrenia. *Nat Genet* 43 (9):864-8.
 - Yue, W. H., H. F. Wang, L. D. Sun, F. L. Tang, Z. H. Liu, H. X. Zhang, W. Q. Li, Y. L. Zhang, Y. Zhang, C. C. Ma, B. Du, L. F. Wang, Y. Q. Ren, Y. F. Yang, X. F. Hu, Y. Wang, W. Deng, L. W. Tan, Y. L. Tan, Q. Chen, G. M. Xu, G. G. Yang, X. B. Zuo, H. Yan, Y. Y. Ruan, T. L. Lu, X. Han, X. H. Ma, Y. Wang, L. W. Cai, C. Jin, H. Y. Zhang, J. Yan, W. F. Mi, X. Y.

Yin, W. B. Ma, Q. Liu, L. Kang, W. Sun, C. Y. Pan, M. Shuang, F. D. Yang, C. Y. Wang, J. L. Yang, K. Q. Li, X. Ma, L. J. Li, X. Yu, Q. Z. Li, X. Huang, L. X. Lv, T. Li, G. P. Zhao, W. Huang, X. J. Zhang, and D. Zhang. 2011. Genome-wide association study identifies a susceptibility locus for schizophrenia in Han Chinese at 11p11.2. *Nat Genet* 43 (12):1228-31.

Appendix

Table A: Gene Selection Category 1: *NRXN1* interactors and related proteins (n=46)

Gene	Chromosome	Start	End	Reason for Selection
APBA1	9	72045201	72287222	NRXN interactor
APBA2	15	29213852	29410515	NRXN interactor
CASK	X	41374188	41782287	NRXN interactor
CNTNAP1	17	40834631	40851832	NRXN interactor
DAG1	3	49507564	49573048	NRXN interactor
DLG2	11	83166052	84634465	NRXN interactor
DLG3	X	69664723	69725337	NRXN interactor
DLG4	17	7093209	7123057	NRXN interactor
DLG5	10	79550548	79686348	NRXN interactor - related
DLGAP1	18	3498836	3880135	NRXN interactor
DLGAP2	8	1449568	1656642	NRXN interactor
DLGAP4	20	34995448	35157040	NRXN interactor - related
ERBB4	2	212240441	213403352	NRXN interactor
FYN	6	111982484	112194627	NRXN interactor
INADL	1	62208148	62629591	NRXN interactor
LIN7A	12	81191170	81331694	NRXN interactor
LIN7B	19	49617617	49621717	NRXN interactor
LIN7C	11	27516122	27528303	NRXN interactor
LRP1	12	57522281	57607125	NRXN interactor
LRP2BP	4	186285032	186300152	NRXN interactor
LRRTM1	2	80529002	80531487	NRXN interactor - related
LRRTM2	5	138205078	138211057	NRXN interactor
LRRTM3	10	68685792	68860867	NRXN interactor - related
LRRTM4	2	77743786	77749502	NRXN interactor - related

Gene	Chromosome	Start	End	Reason for Selection
MACF1	1	39547117	39952789	NRXN interactor
MYO16	13	91295239	91906847	NRXN interactor
NLGN1	3	173116243	174001116	NRXN interactor
NLGN2	17	7311501	7323179	NRXN interactor
NLGN3	X	70364710	70391051	NRXN interactor
NLGN4Y	Y	16636453	16955527	NRXN interactor - related
NRXN1	2	50147487	51259674	NRXN interactor - related
NRXN2	11	64373645	64490660	NRXN interactor - related
NRXN3	14	78870092	80330760	NRXN interactor - related
NXPH1	7	8473584	8792593	NRXN interactor
NXPH2	2	139428450	139537811	NRXN interactor
NXPH3	17	47653478	47657718	NRXN interactor
NXPH4	12	57610577	57620226	NRXN interactor - related
PDZD2	5	31798996	32111038	NRXN interactor
RPH3A	12	113229741	113335246	NRXN interactor
SDCBP	8	59465727	59495419	NRXN interactor
SHANK1	19	51165083	51220195	NRXN interactor
SHANK2	11	70313960	70507872	NRXN interactor
SHANK3	22	51113069	51171641	NRXN interactor
SIPA1L1	14	72926289	73136369	NRXN interactor
STXBP1	9	130374567	130454995	NRXN interactor
UBC	12	125396191	125399577	NRXN interactor

Table B: Gene Selection Category 2: DISC1 and Interactors (n=23)

Gene	Chromosome	Start	End	Reason for Selection
AKT1	14	105235686	105262080	DISC1 interactor
ATF4	22	39916568	39918691	DISC1 interactor
ATF5	19	50431973	50437192	DISC1 interactor
AXIN1	16	337439	402464	DISC1 interactor
CCDC88A	2	55514977	55646963	DISC1 interactor
CIT	12	120123597	120315092	DISC1 interactor
CTNNB1	3	41240941	41281939	DISC1 interactor
DISC1	1	231762560	232177018	DISC1
DMD	X	31137344	33357726	DISC1 interactor
EIF3H	8	117657054	117768062	DISC1 interactor
FEZ1	11	125315647	125366123	DISC1 interactor
GRB2	17	73314156	73401790	DISC1 interactor & NRC/MASC MAGUK/adaptor scaffold
GSK3B	3	119545545	119812513	DISC1 interactor
MAP1A	15	43809805	43823818	DISC1 interactor & mGluR5 Cytoskeletal and cell adhesion
NDE1	16	15744103	15818990	DISC1 interactor & ASD & SZ CNV
NDEL1	17	8339178	8371481	DISC1 interactor
PAFAH1B1	17	2496922	2588909	DISC1 interactor
PCNT	21	47744035	47865682	DISC1 interactor
PDE4B	chr1	66258192	66840262	DISC1 interactor
PDE4D	chr5	58264865	59189621	DISC1 interactor
TRAF3IP1	chr2	239229081	239307797	DISC1 interactor
YWHAE	chr17	1247842	1303517	DISC1 interactor & NRC/MASC MAGUK/adaptor scaffold
ZNF365	chr10	64133916	64162217	DISC1 interactor

Table C: Gene Selection Category 3: Glutamate Receptor Complexes (n=58)

Gene	Chromosome	Start	End	Reason for Selection
ACAN	15	91545670	91617581	mGluR5 Cytoskeletal and cell adhesion
AKAP5	14	64932216	64941221	NRC/MASC MAGUK/adaptor scaffold
AKAP9	7	91570191	91739989	NRC/MASC MAGUK/adaptor scaffold
ALDOA	16	30064490	30081735	AUT and SZ CNV and in mGluR5
APPL1	3	57286725	57332459	NRC/MASC MAGUK/adaptor scaffold
ATP1A1	1	116915835	116947396	NRC/MASC Channels and receptors
ATP1A2	1	160085547	160113374	mGluR5 Channels and receptors
ATP1A3	19	42470733	42498382	NRC/MASC Channels and receptors
ATP1B1	1	169075946	169101960	mGluR5 Channels and receptors
ATP2B1	12	89981825	90049844	mGluR5 Channels and receptors
ATP2B4	1	203595927	203713209	NRC/MASC Channels and receptors
CACNA1A	19	13317256	13617274	mGluR5 Channels and receptors
CDH2	18	25530934	25757191	NRC/MASC Cytoskeletal and cell adhesion
DLG1	3	196769430	197025447	NRC/MASC MAGUK/adaptor scaffold
FAT1	4	187508937	187516980	NRC/MASC Cytoskeletal and cell adhesion
GNB2L1	5	180663927	180670906	NRC/MASC MAGUK/adaptor scaffold
GRIA1	5	152870083	153193429	AMPA Channels and receptors: Glutamate receptors
GRIA2	4	158141735	158287226	AMPA Channels and receptors: Glutamate receptors
GRIA3	X	122318095	122624766	Other PSP Channels and receptors: Glutamate receptors supported by rare variant
GRIA4	11	105480799	105852819	AMPA Channels and receptors: Glutamate receptors
GRID1	10	87359311	88126250	Non-PSP Glutamate receptor & AUT CNV
GRID2	4	93225549	94693649	Other PSP Channels and receptors: Glutamate receptors
GRIK2	6	101846904	102517958	NRC/MASC Channels and receptors
GRIN1	9	140033608	140063208	NRC/MASC Channels and receptors
GRIN2A	16	9855421	10276611	NRC/MASC Channels and receptors
GRIN2B	12	13714409	14133022	NRC/MASC Channels and receptors
GRIN2D	19	48898131	48948188	Other PSP Channels and receptors: Glutamate receptors

Gene	Chromosome	Start	End	Reason for Selection
GRIP1	12	66742798	67072753	NRC/MASC MAGUK/adaptor scaffold
GRM1	6	146348781	146758731	NRC/MASC Channels and receptors
GRM3	7	86273229	86494192	Other PSP Channels and receptors: Glutamate receptors
GRM5	11	88240977	88781190	NRC/MASC Channels and receptors
GRM7	3	6902926	7783217	Other PSP Channels and receptors: Glutamate receptors supported by rare variant
HOMER1	5	78669785	78809700	NRC/MASC MAGUK/adaptor scaffold
ITPR1	3	4535033	4889286	mGluR5 Channels and receptors
L1CAM	X	153127386	153141311	NRC/MASC Cytoskeletal and cell adhesion
LGI1	10	95517565	95557916	NRC/MASC Cytoskeletal and cell adhesion
MAP1B	5	71403117	71505397	mGluR5 Cytoskeletal and cell adhesion
MAPK3	16	30125425	30134630	AUT and SZ CNV and in NRC/MASC
MAPK8IP1	11	45907201	45928016	NRC/MASC MAGUK/adaptor scaffold
MBP	18	74690788	74844774	NRC/MASC Cytoskeletal and cell adhesion
MOG	6	29624808	29640149	NRC/MASC Cytoskeletal and cell adhesion
MPP2	17	41952726	41985113	NRC/MASC MAGUK/adaptor scaffold
MPP3	17	41878166	41910538	NRC/MASC MAGUK/adaptor scaffold
PLP1	X	103031438	103047547	NRC/MASC Cytoskeletal and cell adhesion
SEPT5	22	19701986	19712297	AUT and SZ CNV and in mGluR5 complex
SLC12A5	20	44657837	44688789	mGluR5 Channels and receptors
SLC1A2	11	35272751	35441105	NRC/MASC Channels and receptors
SLC4A4	4	72053002	72437799	mGluR5 Channels and receptors
SPTB	14	65213000	65289866	NRC/MASC Cytoskeletal and cell adhesion
TBCA	5	76951239	77036429	NRC/MASC Cytoskeletal and cell adhesion
TJP1	15	29992356	30114706	NRC/MASC MAGUK/adaptor scaffold
VDAC1	5	133307606	133340433	NRC/MASC Channels and receptors
VDAC2	10	76970562	76991206	NRC/MASC Channels and receptors
YWHAB	20	43514343	43537161	mGluR5 MAGUK/adaptor scaffold
YWHAG	7	75956115	75988316	NRC/MASC MAGUK/adaptor scaffold

Gene	Chromosome	Start	End	Reason for Selection
YWHAH	22	32340478	32353590	NRC/MASC MAGUK/adaptor scaffold
YWHAQ	2	9724105	9771106	mGluR5 MAGUK/adaptor scaffold
YWHAZ	8	101930913	101965569	NRC/MASC MAGUK/adaptor scaffold

Table D: Gene Selection Category 4: Neural Cell Adhesion Molecules (n=61)

Gene	Chr	Start	End	Reason for Selection
ARVCF	22	19957418	20004309	PSD Cytoskeletal & cell adhesion: cell-adhesion molecules supported by rare variant
ASTN1	1	178563579	178867400	PSD Cytoskeletal & cell adhesion: cell-adhesion molecules supported by rare variant
ASTN2	9	119187506	120177317	PSD Cytoskeletal & cell adhesion: cell-adhesion molecules supported by rare variant
CACNA1C	12	2162415	2807115	PSD Channels & receptors supported by rare variant
CADM1	11	115044344	115375241	Neural CAMs
CADM3	1	159141376	159172932	Neural CAMs
CASKIN1	16	2227183	2246465	MAGUK/adaptor scaffold: Non-PDZ domain-containing scaffolders
CDH1	16	68771194	68869444	Neural CAMs
CDH10	5	24487209	24644911	PSD Cytoskeletal & cell adhesion: cell-adhesion molecules
CDH13	16	82660577	83830199	PSD Cytoskeletal & cell adhesion: cell-adhesion molecules supported by rare variant
CHL1	3	238650	451097	Neural CAMs
CLDN11	3	170136664	170151885	Neural CAMs
CLU	8	27454450	27472327	PSD Cytoskeletal & cell adhesion: cell-adhesion molecules
CNTN1	12	41086357	41464094	Neural CAMs
CNTN2	1	205012339	205047138	Neural CAMs
CNTNAP2	7	145813452	148118086	Neural CAMs
CORO1A	16	30194925	30200397	PSD Cytoskeletal & cell adhesion: cytoskeletal supported by rare variant
CSPG5	3	47603727	47620359	PSD Cytoskeletal & cell adhesion: cell-adhesion molecules
CTNNA1	5	138089106	138270723	PSD Cytoskeletal & cell adhesion: Catenins
CTNND1	11	57529233	57586652	PSD Cytoskeletal & cell adhesion: Catenins
CTNND2	5	10971951	11904110	PSD Cytoskeletal & cell adhesion: Catenins supported by rare variant
DSP	6	7541869	7586946	PSD Cytoskeletal & cell adhesion: cell-adhesion molecules supported by rare variant
DST	6	56214826	56711454	PSD Cytoskeletal & cell adhesion: cell-adhesion molecules
ERBB2IP	5	65222383	65376850	PSD Cytoskeletal & cell adhesion: cell-adhesion molecules
GJA1	6	121756745	121770873	PSD Cytoskeletal & cell adhesion: cell-adhesion molecules
GJB6	13	20796100	20806534	PSD Cytoskeletal & cell adhesion: cell-adhesion molecules
HAPLN1	5	82898261	82981140	PSD Cytoskeletal & cell adhesion: cell-adhesion molecules

Gene	Chr	Start	End	Reason for Selection
HAPLN2	1	156589085	156595517	PSD Cytoskeletal & cell adhesion: cell-adhesion molecules
ITGA8	10	15559087	15761770	Neural CAMs
ITGAV	2	187454799	187545628	Neural CAMs
ITGB1	10	33189245	33247293	Neural CAMs
ITGB8	7	20370724	20455382	Neural CAMs
LRRC7	1	70453270	70816583	PSD Cytoskeletal & cell adhesion: cell-adhesion molecules
MAG	19	35783037	35804707	Neural CAMs
MAPT	17	43971749	44105699	PSD Cytoskeletal & cell adhesion supported by rare variant
MPZ	1	161274524	161279762	Neural CAMs
MPZL1	1	167691207	167760060	Neural CAMs
NCAM1	11	112831994	113149158	Neural CAMs
NCAM2	21	22370632	22911214	Neural CAMs
NCAN	19	19322782	19363061	PSD Cytoskeletal & cell adhesion: cell-adhesion molecules
NCDN	4	126201549	126211153	PSD Cytoskeletal & cell adhesion: cell-adhesion molecules
NEGR1	1	71868624	72748277	Neural CAMs
NEO1	15	73344874	73597545	Neural CAMs
NFASC	1	204797822	204991949	Neural CAMs
NLGN4X	X	5808082	6146706	PSD Cytoskeletal & cell adhesion: cell-adhesion molecules supported by rare variant
NRCAM	7	107788081	108096826	Neural CAMs
OPCML	11	132284874	133402403	PSD Cytoskeletal & cell adhesion: cell-adhesion molecules supported by rare variant
PICK1	22	38453261	38471708	PSP MAGUK/adaptor scaffold: PDZ domain-containing scaffolders
PTPRF	1	43996546	44089343	Neural CAMs
PVR	19	45147225	45166850	Neural CAMs
PVRL1	11	119508807	119599435	Neural CAMs
PVRL2	19	45349392	45392485	Neural CAMs
PVRL3	3	110790864	110853062	Neural CAMs
RREB1	6	7163189	7306695	PSD Cytoskeletal & cell adhesion: cell-adhesion molecules supported by rare variant
SDC1	2	20400557	20425194	Neural CAMs

Gene	Chr	Start	End	Reason for Selection
SDC2	8	97505881	97624037	Neural CAMs
SDC3	1	31342312	31381480	Neural CAMs
SDCBP2	20	1290620	1309838	Neural CAMs
SFN	1	27189632	27190947	PSD MAGUK/adaptor scaffold: 14-3-3
TSC1	9	135766735	135820020	PSD Cytoskeletal & cell adhesion: cell-adhesion molecule supported by rare variant

Table E: Gene Selection Category 5: Functional and Positional Candidates (n=27)

Gene	Chromosome	Start	End	Reason for Selection
A2BP1	16	6069131	7762499	Good candidate
ANK3	10	61788158	62149488	Bipolar disorder risk gene: contains genome-wide significant SNP
ATXN2	12	1.12E+08	1.12E+08	Good candidate
ATXN2L	16	28834414	28848558	Good candidate
CYFIP1	15	22892683	23003603	AUT and SZ CNV and good candidate
DDX6	11	1.19E+08	1.19E+08	Good candidate
DTNBP1	6	15523039	15663271	SZ risk gene, local interest
FOXP2	7	1.14E+08	1.14E+08	AUT and SZ CNV and good candidate
GAD1	2	1.72E+08	1.72E+08	Good candidate
ITGA4	2	1.82E+08	1.82E+08	AUT risk gene, local interest
MACROD2	20	13976145	16033841	AUT risk gene
MECP2	X	1.53E+08	1.53E+08	Good candidate
NOS1	12	1.19E+08	1.19E+08	SZ risk gene, local interest
NRG1	8	32405753	32622073	SZ risk gene, local interest
NRGN	11	1.25E+08	1.25E+08	SZ risk gene, genome-wide significant (SGENE)
PLXNA2	1	2.08E+08	2.08E+08	SZ candidate gene, local interest
PLXNA4	7	1.32E+08	1.32E+08	SZ candidate gene, local interest
PTCHD1	X	23352984	23414918	Good candidate
RELN	7	1.03E+08	1.04E+08	Good candidate
S100B	21	48018530	48025035	SZ risk gene, local interest
SEMA5A	5	9035137	9546233	AUT risk gene
SEMA6A	5	1.16E+08	1.16E+08	SZ candidate gene, local interest
SEMA6B	19	4542600	4558501	SZ candidate gene, local interest
SYNGAP1	6	33387846	33421466	Good candidate
TCF4	18	52889562	53255860	SZ risk gene, genome-wide significant (SGENE, ISC)
VRK2	2	58273719	58387005	SZ risk gene, genome-wide significant (SGENE)
ZNF804A	2	1.85E+08	1.86E+08	SZ risk gene, genome-wide significant (Cardiff)

Table F: Index Sequences

Index No.	Index Sequence	Index No.	Index Sequence
1	AGCTAT	22	CTATCT
2	CATGCT	23	GCGCGT
3	GTACGT	24	TTCATT
4	TGTGTT	25	AAGGAT
5	AGCGAT	26	CTGCCT
6	CAACCT	27	GACTGT
7	GAGAGT	28	TCTCTT
8	TCAGTT	29	ACTGAT
9	ACGTAT	30	CGTACT
10	CGATCT	35	CTTACT
11	GTCAGT	36	GATCGT
12	TAGCTT	37	GCCGGT
13	AGGAAT	38	GGAAGT
14	CCAACT	39	GTGTGT
15	GGCCGT	40	ACCAAT
16	TGACTT	41	ATTAAT
17	ACACAT	42	TACGTT
18	GCGCGT	43	TATATT
19	GCTAGT	44	TGCATT
20	TTGCTT	45	TTACTT
21	AACCAT	46	GTTGGT

Table G: LoF variant validation PCR primer sequences and programmes

Chr	Position	Gene	FWD primer sequence	REV primer sequence	Programme	Annealing Temperature ^a	CAMBIO Optimisation Buffer ^b
1	62456007	INADL	CTGCCTGGAGAACTGCACATTA	TACAACATCCACTGCACCCATC	Touchdown	58°C	-
1	39788293	MACF1	GTTTGCTAACCTCCAGCATGAGTG	CACCTCTTGAGTGATGGCATCTGA	Regular	54°C	-
1	232144803	DISC1	TGGATGACTTAGAGGGAGCTG	AAGTTTATATGCTCAATGGGAAGC	Regular	58°C	Buffer G
1	62321741	INADL	CACCAGAGTGCCAGACACAT	CACAGCTTCAGCAAGTGAGG	Regular	58°C	Buffer H
1	62349979	INADL	GATCCCGAATGATGTCCAAG	CAAACGTTCTACCAAAATCACG	Regular	56°C	-
1	208216512	PLXNA2	CAGGACTCGCATAGCGTAGG	AAAAGTCCCTCCCCTCGTC	Regular	56°C	-
2	187519413	ITGAV	AGAAAGATGAGATGTGGGTGA	TTTTTAAGTGCTGTGTATGCAGTAT	Regular	60°C	-
2	239257490	TRAF3IP1	GACCTGCTGGCCAAGATAAG	TGAATTCCTTAAATCTCCAAA	Regular	56°C	-
3	57282220	APPL1	CCCGAAACAAATATGCAGGT	TCCAGCCTTTTCGGGTAAAT	Regular	58°C	-
4	187628509	FAT1	CTGTGGCTTCAACCTGGATGATT	CTCCTCTTGACCGTGAAACAACAG	Regular	54°C	-
4	72433527	SLC4A4	GAGAACAAAGAAGGACACATTTGA	TTTCTGTTCCCTTGCTCCTC	Regular	58°C	Buffer A
6	56472474	DST	ATGAGACTGGGAACACACACACTG	AGGATGAATGTACAGCTACACCAAG	Touchdown	56°C	-
6	56483170	DST	GCTGGAATTGGACTTCTTTGGG	ACTTGAGGATGAGCTGATAGCC	Touchdown	56°C	-
6	112025283	FYN	TCACTCTCGCGGATAAGAAAGG	AGGACTAAATTGGAGGAGAGGG	Touchdown	58°C	-
6	56507564	DST	AGGATTCACTGCCGAAGCTA	TGCCACTTTTCACCGTTAGA	Regular	58°C	-
6	33400481	SYNGAP1	GCCTCTGGGAGACTTCCCTTT	CAATCTCATCCCACATGCTC	Regular	58°C	-
6	56358939	DST	GTTTCCACCGATTCCCATTT	GATGAGGAGCTTGGGACTGA	Regular	56°C	-
6	56482783	DST	CAATGAGCCAATCACATTCAA	AGAGACCAAGCCCTGAACAA	Regular	58°C	Buffer H
6	56483389	DST	TCTGGGCTATCAGCTCATCC	CAGGCATTAGGCTTGACTTTG	Regular	58°C	-
6	56479284	DST	GCTTAGAGAAGGGAGGTTTTCC	TTGTTTGCTATGGAGAGGTCAG	Regular	58°C	-
6	56371316	DST	AATCCCCTTACCTGCAGTCC	TGGCTTACTTATCAGAATGGCTTT	Regular	58°C	Buffer E
8	27463990	CLU	CATGTGTCCCCTTTTCACCT	GCAGCTCTTGAGTGCATGA	Regular	58°C	-
10	79614016	DLG5	CCCTGGCTACTAAGTCCAGAGAAA	GGTGTGTGTGGTGCATAGAAGA	Touchdown	56°C	-
10	79584235	DLG5	GCTCTGCCATACTCCTGAC	GACTTTGGCACCTTGCAATT	Regular	58°C	-
12	13724778	GRIN2B	TCAAGGACTCTGAGCTTGGA	AATGACTTCTCACCCCCTTTC	Regular	54°C	-
12	66923668	GRIP1	GATTTTTCTAACCAACACCTACCTC	CAACAGGGATCTGTAATATTTTATGC	Regular	56°C	-
12	66765472	GRIP1	AGCTTCTTTGCTTGGCTGAG	GCCTTCAGATGTGGGTAGGA	Regular	56°C	-
13	109610055	MYO16	CCATGTCCAGGTGATTCCCATCAT	TGAAGCCAACATTCAGGGCTC	Touchdown	58°C	-
13	20797176	GJB6	GGGTCAATCCCACATTTCAA	TTCAGGCGAGGAGAGAAGAG	Regular	62°C	-

13	20797556	GJB6	CGGGAAAAAGTGGTCATAGC	TAGGGATAAACACGCGCAAT	Regular	58°C	-
17	40844654	CNTNAP1	GTGGTCTTCGCCTTTGATGT	CACATAGAGGGGCTGGTCAT	Regular	56°C	-
18	74728772	MBP	GCACGGAACGAGACCTTAGA	CTCCCAAGGCACAGAGACAC	Regular	60°C	Buffer F
X	32429867	DMD	ATCAAAACAACCCCATGGAA	TGCTTGAACAGAGCATCCAG	Regular	56°C	-
X	70367905	NLGN3	GGCCTGTGGACCAATACCT	GTCGGCACATAGACGTTTCAG	Regular	56°C	-
^a Annealing temperature for Touchdown programmes refers to starting annealing temperature							
^b Buffer A-L refers to CAMBIO optimisation buffers, Qiagen buffer was used where not specified							

Table H: Regular and Touchdown PCR programmes

Regular			Touchdown		
Step	Temperature	Time	Step	Temperature	Time
1	98°C	30 secs	1	98°C	30 secs
2	98°C	10 secs	2	98°C	10 secs
3	X°C	30 secs	3	X°C	30 secs
4	72°C	30 secs	4	72°C	30 secs
Repeat step 2 through 4 for a total of 35 times.			Repeat step 2 through 4 with X°C decreasing by 1°C per cycle for 5 cycles, then repeat steps 2 through 5 at X-5°C 30 times.		
X = variable annealing temperature					

Table I: Primers for parent-of-origin informative SNV discovery

Amplicon Name	FWD primer sequence	REV primer sequence	Programme	Annealing Temperature*
GRIN2B_Upstream	GCATGAATCAATATTTAAACCAA	CCAGAGGGGTGTAGATGATG	Touchdown	64°C
GRIN2B_Downstream	TTGCGAATATTTCTCTGTGC	GAATAACAGGTTCTGATGGGTCA	Touchdown	62°C
DISC1_Upstream	GGTGCTTTGGGACCATGC	TGGAGGATTCCACCTGTAGC	Touchdown	60°C
DISC1_Downstream	CCCCAAATGTTAAGGGAGAGA	TCCTTTTATCCTCGGAGTGG	Touchdown	62°C
*Annealing temperature for Touchdown programmes refers to starting annealing temperature				

Table J: Parent-of-origin PCR Primers

Amplicon Name	Primer sequence	Programme	Annealing Temperature
GRIN2B FWD primer for de novo SNV G allele	GCATCATCTACACCCCTCT G	Touchdown	62°C
GRIN2B FWD primer for de novo SNV G allele including mismatch	GCATCATCTACACCCCTC AG	Touchdown	62°C
GRIN2B FWD primer for de novo SNV A allele	AATGCATCATCTACACCCCTCT A	Touchdown	62°C
GRIN2B FWD primer for de novo SNV A allele including mismatch	GCATCATCTACACCCCTC CA	Touchdown	62°C
GRIN2B REV primer	AAAGGTGACTAGCCCACTGC	Touchdown	62°C
DISC1 FWD primer for rs821616 A allele	GCTTGTGCGATTGCTTATCCAG A	Touchdown	64°C
DISC1 FWD primer for rs821616 A allele including mismatch	GCTTGTGCGATTGCTTATCCAT A	Touchdown	64°C
DISC1 FWD primer for rs821616 T allele	GCTTGTGCGATTGCTTATCCAG T	Touchdown	64°C
DISC1 FWD primer for rs821616 T allele including mismatch	GCTTGTGCGATTGCTTATCCAT T	Touchdown	64°C
DISC1 REV primer	AAATATGCCAAAAGTTGGGTTTT	Touchdown	64°C
*Annealing temperature for Touchdown programmes refers to starting annealing temperature. Base pairs in red are mismatches and those in bold are allele-specific sites			

Table K: FYN colony screen primers

Amplicon name	FWD primer sequence	REV primer sequence	Plasmid insert amplified		Programme	Annealing temperature
			FYN wildtype	FYN mutant		
SET1	ATGGGCTGTGTGCAATGTAA	GCCAAGTTTTCCAAAGTACCA	YES	YES	Regular	56°C
SET2	CAGCAATTATGTGGCTCCAG	GCTCGATGTAAGCCATTCT	YES	NO	Regular	56°C
SET3	CATCGTCACCGAGTATATGAACA	GGCTCTGTCGCGGTAAAGTA	YES	NO	Regular	56°C

Table L: *In Silico* Analysis of validated LoF variants

Chr	Position	Gene	Type	Samples			Singleton? ^b	LoF? ^c	Number of transcripts:				Full or Partial LoF?	Affected exon / Total Number Exons ^e	Affect of variant ^e
				SZ ^a	AUT ^a	CON ^a			Total	Protein Coding	Carrying Variant	Carrying variant as LoF			
1	39788293	MACF1	Coding Indel	1			Yes	Yes	39	22	12	10	Partial	32/102	Frameshift introducing stop codon 7 codons downstream
1	62321741	INADL	Coding Indel		1		Yes	Yes	17	8	5	5	Full	18/43	Frameshift introducing stop codon 2 codons downstream
1	62349979	INADL	Coding Indel	1			Yes	Yes	17	8	5	5	Full	22/43	Frameshift introducing stop codon 44 codons downstream
1	62456007	INADL	Nonsense SNV	1			Yes	Yes	17	8	8	6	Partial	28/43	R1280*
1	208216512	PLXNA2	Coding Indel	1			Yes	Yes	6	1	1	1	Full	21/32	Frameshift introducing stop codon 44 codons downstream
1	232144803	DISC1	Coding Indel		1		Yes	No ^d	16	15	7	1	Partial	3/3	Frameshift introducing stop codon 44 codons downstream
2	187519413	ITGAV	Coding Indel	1			Yes	Yes	9	5	4	3	Partial	16/30	Frameshift introducing stop codon 44 codons downstream
2	239257490	TRAF3IP1	Splice Site SNV		1		Yes	Yes	7	4	4	4	Full	11/17	Donor site variant introduces stop codon 30 codons downstream
3	57282220	APPL1	Splice Site		1		Yes	Yes	9	3	1	1	Full	10/22	Acceptor site variant results in complete removal of exon 10

			SNV												
4	72433527	SLC4A4	Coding Indel			1	Yes	No	6	5	4	1	Full	25/25	Frameshift introducing stop codon 44 codons downstream
4	187628509	FAT1	Nonsense SNV	1			Yes	Yes	12	6	2	2	Full	2/27	E825*
6	56358939	DST	Coding Indel	1			Yes	Yes	38	19	8	6	Partial	83/102	Frameshift introducing stop codon immediately
6	56472474	DST	Nonsense SNV	1			Yes	Yes	38	19	9	6	Partial	39/102	Q2285*
6	56479284	DST	Splice Site SNV		1		Yes	Yes	38	19	10	10	Full	36/102	Acceptor site variant results in removal of exon 36
6	56482783	DST	Splice Site Indel		1		Yes	Yes	38	19	12	1	Partial	23/24	Donor site variant introduces stop codon 23 codons downstream
6	56483170	DST	Nonsense SNV	1			Yes	Yes	38	19	12	1	Partial	23/24	E1888*
6	56483389	DST	Nonsense SNV			1	Yes	Yes	38	19	12	1	Partial	23/24	Q1815*
6	56507564	DST	Coding Indel	2	2	2	No	Yes	38	19	13	5	Partial	1/84	Frameshift introducing stop codon 44 codons downstream
6	112025283	FYN	Nonsense SNV	1			Yes	Yes	32	22	12	12	Full	7/14	R156*
8	27463990	CLU	Coding Indel		1	1	No	Yes	32	22	13	11	Partial	4/9	Frameshift introducing stop codon 4 codons downstream
10	79584235	DLG5	Splice Site SNV			1	Yes	Yes	12	4	4	4	Full	14/32	Acceptor site variant results in removal of exon 14 and protein continues as normal

10	79614016	DLG5	Nonsense SNV	1			Yes	Yes	12	4	3	2	Partial	4/32	E217*
12	13724778	GRIN2B	Nonsense SNV		1		Yes	Yes ^d	1	1	1	1	Full	10/13	Q711*
12	66765472	GRIP1	Splice Site SNV	1			Yes	Yes	18	14	6	5	Partial	23/25	Donor site variant introduces stop codon 24 codons downstream
12	66923668	GRIP1	Nonsense SNV		1		Yes	Yes	18	14	11	11	Full	5/25	R149*
13	20797556	GJB6	Coding Indel		1	3	No	Yes	4	4	4	4	Full	5/5	Frameshift introducing stop codon 44 codons downstream
13	109610055	MYO16	Nonsense SNV	1			Yes	Yes	8	6	6	6	Full	16/34	Q627*
17	40844654	CNTNAP1	Nonsense SNV		1		Yes	Yes	1	1	1	1	Full	17/24	R890*
18	74728772	MBP	Stop Loss SNV	1	1		No	No	32	16	15	2	Partial	17/4	Stop codon lost and another introduced 16 codons downstream
X	32429867	DMD	Splice site SNV		1		Yes	Yes	33	24	9	7	Partial	30/79	Donor site variant introduces stop codon 24 codons downstream
X	70367905	NLGN3	Coding Indel	1			Yes	Yes	6	5	5	5	Full	2/8	Frameshift introducing stop codon 42 codons downstream

^a Number of variants detected in these phenotype categories

^b Singleton variants are those found in just sample in the study

^c Yes = LoF variant located in first 95% of coding sequence. No = LoF variant located in last 5% of coding sequence.

^d De novo variant

^e Relevant to the largest transcript that the variant has a LoF effect in.

^e Relevant to the largest transcript that the variant has a LoF effect in.