LEABHARLANN CHOLÁISTE NA TRÍONÓIDE, BAILE ÁTHA CLIATH | TRINITY COLLEGE LIBRARY DUBLIN
Ollscoil Átha Cliath | The University of Dublin

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 glutamatergic synapse genes in schizophrenia susceptibility:

from linkage to genome-wide association studies

William Gilks

Department of Psychiatry, Trinity College Dublin

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

Corrected edition 9th February 2011



Declaration

I hereby declare that this thesis has not been submitted as an exercise for examination to this or any other university. The work described herein has been conducted by the author alone, except where stated otherwise. This thesis may be lent or copied by the library of Trinity College Dublin for study purposes only, subject to normal conditions and acknowledgement.

Signed: _____

Date: 1/3/2011

TRINITY COLLEGE
27 JUL 2011
LIBRARY DUBLIN
THESIS
9256

Summary

Schizophrenia is a common and severe mental illness which is highly heritable. Susceptibility to schizophrenia is likely to be caused by multiple genetic variants in combination with environmental risk factors. The identification of such genetic variants would greatly improve diagnosis and treatment.

The first aim of this study was to identify common susceptibility variants for schizophrenia using a candidate gene case-control association study design. Nine genes were selected based upon positional (linkage data) and functional criteria (*ARHGEF11*, *CNDP1*, *GGCX*, *GLUL*, *HOMER2*, *HTRA2*, *MAPT*, *PEA15* and *SYNGAP1*). Sixty-six tagging SNPs were tested in 375 schizophrenia cases and 812 controls from Ireland. Significant associations were identified at five SNPs in three of the genes (*CNDP1*, *HOMER2*, *SYNGAP1*).

The second aim was to follow-up significant association findings at *HOMER2* with fine-mapping, functional studies and independent replication. The strongest associated SNP at *HOMER2* (rs2306428; p=0.006, OR=0.61) which was predicted to effect splicing. Replication analysis in the International Schizophrenia Consortium (ISC) sample (1,287 cases and 1,129 controls) was supportive of a role for *rs2306428* in schizophrenia susceptibility (p=0.019, OR=0.77). No relationship was identified between genotype of rs2306428 and transcript levels in lymphoblast cells, nor on indices of neurocognitive function which were tested under the endophenotype hypothesis.

The third aim was to investigate HOMER2 interactants in schizophrenia susceptibility using results of the ISC genome-wide association study (GWAS). A SNP at the *CDC42* gene, rs2473227 was highly-associated with schizophrenia ($p=2.00\times10^{-6}$, OR=0.85), in high linkage disequilibrium with a predicted regulatory SNP rs2473307, and correlated with CDC42 mRNA level in two independent studies. Gene reporter assay of rs2473307 indicated that the susceptibility allele reduced expression in a neuronal cell lines. However, this could not be confirmed using an electrophoretic mobility shift assay because binding of transcription factors to either allele was not observed. The heterozygous genotype at rs2473277 was associated with indices of working and episodic memory ($p<0.05$). Further available GWAS data did not replicate the association with schizophrenia at this SNP.

The fourth aim was to investigate the ability of next-generation sequencing to detect copy-number variants (CNV). Sixteen test CNVs were selected from publicly-available data which were verified by population genetic parameters. The true break-points of the selected CNVs

were determined by Sanger sequencing. The CNV detection algorithm developed in-house accurately predicted the existence and break-points of many of the CNVs in targeted resequencing data although its resolution appeared to be affected by the size of the CNV and GC-content of the target region.

Since this study was started there have been massive changes in the way in which the genetics of common disease such as schizophrenia. Techniques such as genome-wide association studies, next-generation sequencing and high-throughput high-content analysis are generating vast amounts of data. Such experiments must be designed and analysed both imaginatively and with strict rigour if current and future investments are to deliver a better outcome for patients.

Acknowledgments

This work was primarily funded by an EMBARK scholarship from the Irish Research Council for Science, Technology and Engineering. The primary supervisor for this study was Dr Derek Morris whose wisdom and patience was inspirational throughout every aspect of the work. It was also an honour to have Dr Aiden Corvin as secondary supervisor whose enthusiasm and insight into the field of schizophrenia research was exemplary.

Many other members of Neuropsychiatric Genetics laboratory contributed to this work and where they have done so directly, they are acknowledged within the main text. Of the remainder, Dr Matthew Hill was of outstanding support in the experimental investigation of SNP function. In addition, the endophenotype analyses were conducted with under the venerable supervision of Dr Gary Donohoe. Contributions to this study were also made through assistance in laboratory experiments, with analysis methods, by proof-reading or through inspiring and illuminating conversations by (in no particular order) Dr Carlos Pinto, Dr Richard Anney, Dr Ricardo Seguardo, Dr Ziara Hawi, Dr Colm O'Dushlaine, Dr Elaine Kenny, Ms Emma Quinn, Ms Karen Sheehan (now Dr), Ms Lynne Cochrane (now Dr), Ms Katherine Tansey (now Dr), Ms Amy Gates, Ms Alison Merikangas, Ms Elif Dagdan, Mr Graham Kenny, Ms Sarah Furlong, Dr Paul Cormican, and Dr Elizabeth Heron.

All those listed above were (or still are) members of the Neuropsychiatric Genetics group, which under Professor Micheal Gill has established superb resources, capabilities and achievements whilst being an honest, friendly and encouraging work-place. Outside of this department, other members of the Institute of Molecular Medicine were also helpful (Dr Anthony Mitchell-Davis for use of a pipetting robot when ours finally died and Dr Steven Smith for use of the CCD camera for imaging the results of the electrophoretic mobility shift assays). Outside of Ireland useful proteomic data were kindly provided Dr Sabine Schrimpf (University of Zürich). Re-prints and illuminating discussions were obtained from Dr Peter Kalivas (Medical University of South Carolina), Dr Elizabeth Thomas (The Scripps Institute) and by Professor Paul Worley (Johns Hopkins University).

Finally, outside of science, family and friends were constant source of encouragement for this work and support for life in general.

To all of these individuals I am very grateful and wish the greatest of luck and success in the future.

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.

Publications arising from this work

Gilks WP, Allott EH, Donohoe G, Cummings E, The International Schizophrenia Consortium, Gill M, Corvin AP, Morris DW (2010) Replicated genetic evidence supports a role for *HOMER2* in schizophrenia. *Neuroscience Letters* **468**(3):229-33

Kenny EM, Cormican P, Gilks WP, Gates A, O' Dushlaine CT, Pinto C, Corvin A, Gill M, Morris DW (2010) Multiplex target enrichment using DNA indexing for ultra-high throughput SNP detection. *DNA Research*(manscript accepted)

Abbreviations

ADHD Attention Deficit Hyperactivity Disorder
CEPH Centre d'Etude du Polymorphisme Humain
CEU Caucasian European from Utah
Chr Chromosome
CNS Central Nervous System
CNV Copy Number Variant
COMT Catechol-O-Methyl Transferase
CDC42 Cell division/cycle 42
DAO D-amino acid oxidase
ddNTPs Dideoxy nucleotide triphosphate
DGV Database of Genomic Variants
DISC1 Disrupted-in-Schizophrenia 1
DNA Deoxyribonucleic Acid
DSM-IV Diagnostic Statistical Manual IV
DTNBP1 Dystrobrevin-binding protein 1
DZ Dizygotic
EMSA Electrophoretic Mobility Shift Assay
GAPDH Glyceraldehyde 3-phosphate dehydrogenase
GASP Genetic Association study of Schizophrenia and related Psychoses
HLA Human lymphocyte antigen
ICD-10 International Classification of Diseases 10
ISC The International Schizophrenia Consortium
KO Knock-Out
LD Linkage Disequilibrium
LOD Log of Odds
LTP Long Term Potentiation
MAF Minor Allele Frequency
MHC Major histocompatability complex
NA Not Available
NCBI The National Center for Biotechnology Information
NMDA N-methyl-D-aspartic acid
NS Not Significant
NRG1 Neuregulin 1
OMIM Online Mendelian Inheritance in Man
OR Odds Ratio
PCA Principle components analysis
PCR Polymerase Chain Reaction
PET Positron Emission Tomography
RNA Ribonucleic Acid
RPGI Resource for Psychoses Genomics Ireland
RR Relative Risk
SAP Shrimp Alkaline Phosphatase
SNP Single Nucleotide Polymorphism
TF Transcription factor
TFBS Transcription factor binding-site
UCSC University College, Santa Cruz
WTCCC Wellcome Trust case-control consortium
ZNF804A Zinc finger protein 804A

Contents

Title page i
Declaration ii
Summary iii
Acknowledgments v
Statement of Work vi
Publications arising from this work vii
Abbreviations vii
Contents ix
Chapter 1: General Introduction 1
1.1 Schizophrenia 1
1.1.1 Clinical aspects of schizophrenia 1
1.2 Schizophrenia Epidemiology 5
1.2.1 Demographic prevalence of schizophrenia 5
1.2.2 Risk factors for schizophrenia 5
1.2.3 Heritability of schizophrenia 6
1.3 The identification of schizophrenia susceptibility genes 8
1.3.1 Incentive for the investigation of schizophrenia genetics 8
1.3.2 Background on linkage mapping 9
1.3.3 Linkage studies of schizophrenia 9
1.3.4 Background on genetic association studies 10
1.3.5 Candidate-gene association studies of schizophrenia 11
1.3.6 The Rare Variant Hypothesis of Schizophrenia 13
1.3.7 Functional investigation of genetic susceptibility variants 14
1.3.8 Copy-number variants and schizophrenia 15
1.3.9 Genome-wide association studies of schizophrenia 16
1.4 Aetiological hypotheses of schizophrenia 17
1.4.1 Schizophrenia pathology: Dysconnectivity and hypofrontality 17
1.4.2 Neurodevelopment in schizophrenia: synaptic plasticity and pruning 19
1.4.3 Schizophrenia drug targets and the dopamine hypothesis 20
1.4.4 NDMA receptor hypofunction and the Glutamate hypothesis 21
1.4.5 The endophenotype hypothesis of psychiatric disorders 22
1.5 Project aims 24
Chapter 2: A candidate gene study of schizophrenia 26
2.1 Introduction 26
2.1.1 Chapter overview 26
2.1.2 Schizophrenia linkage data 26
2.1.3 Glutamate and the synapse in schizophrenia 27
2.1.4 Interactors of known schizophrenia genes 27
2.2 Methods 31
2.2.1 Selection of candidate genes 31
2.2.1.1 Identification of genes within linkage bins 31
2.2.1.2 Identification of glutamatergic synapse genes 31
2.2.1.3 Identification of interactors of schizophrenia genes 32

2.2.2 The GASP schizophrenia case-control sample 33
2.2.3 Novel SNP selection and genotyping 34
2.2.3.1 Basis for genotyping novel SNPs 34
2.2.3.2 The SNaPshot™ genotyping method 34
2.2.3.3 Design and optimisation of PCR reactions for Snapshot 35
2.2.3.4 Investigative genotyping of functional SNPs 35
2.2.4 Association study methods: Tagging SNP selection, genotyping and data analysis 39
2.2.4.1 Selection of tagging-SNPs 39
2.2.4.2 Genotyping of tagging SNPs in the GASP case-control sample 39
2.2.4.3 Quality control and analysis of genotyping results 40
2.3 Results 41
2.3.1 Results of candidate gene selection 42
2.3.1.1 Genes with evidence of linkage from Lewis *et al* (2003) study 42
2.3.1.2 Genes fulfilling glutamatergic and synaptic search criteria 42
2.3.1.3 Identified interactors of putative schizophrenia genes 44
2.3.1.4 Final selection of genes for association analysis 46
2.1.3.5 Background on genes selected for association analysis 47
2.3.2 Results of Novel SNP Investigation 49
2.3.3 Results of association analysis 52
2.4 Discussion 57
2.4.1 Review of candidate gene study design and results 57
2.4.2 Review of candidate gene selection method 57
2.4.3 The association signal at *CNDP1* 58
2.4.4 The association signal at *SYNGAP1* 59
2.4.5 The association signals at *HOMER2* 59
Chapter 3: Further studies of the *HOMER2* gene 60
3.1 Introduction 60
3.1.1 Background on HOMER proteins 60
3.1.2 *Cis*-acting variation and schizophrenia 65
3.1.3 Endophenotype studies of schizophrenia 66
3.1.4 Study Aims 67
3.2 Methods 68
3.2.1 Defining the region of strongest association 68
3.2.2 Identification of functional sites within the associated region 69
3.2.3 Polymorphism detection in functional sites 71
3.2.4 Novel SNP genotyping in the CEU sample 71
3.2.5 Additional SNP genotyping of *HOMER2* in the GASP case-control sample 72
3.2.5.1 Selection of additional tagging-SNPs 72
3.2.5.2 Genotyping of tagging SNPs 72
3.2.6 Replication analyses of *HOMER2* 73
3.2.6.1 Replication analysis in the RPGI sample 73
3.2.6.2 Replication analysis in the ISC sample 73
3.2.7 Investigation of *HOMER2* expression 74
3.2.7.1 Quantitative PCR of *HOMER2* mRNA 74

3.2.7.2 Comparison of *HOMER2* expression with schizophrenia-associated SNPs 75
3.2.8 Neuropsychology analysis methods 74
3.3 Results 76
3.3.1 Novel SNPs in the associated interval of *HOMER2* 76
3.3.1.1 Discovery of novel SNPs 75
3.3.1.2 Functional investigation *in silico* of novel SNPs 76
3.3.1.2 LD relationships of novel SNPs 76
3.3.2 Results of further association analysis in the GASP case-control sample 82
3.3.3 Haplotype association results in the GASP and RPGI samples 82
3.3.4 Replication results at *HOMER2* 86
3.3.4.1 Replication results in a second Irish sample (RPGI) 86
3.3.4.2 Replication results in the ISC GWAS sample 86
3.3.5 Predicted funtional properties of associated SNPs 89
3.3.6 Comparison of genotype at rs2306428 with *HOMER2* expression 90
3.3.7 Comparison of genotype at rs2306428 with neuropsychological variables 92
3.5 Discussion 93
Chapter 4: Association and functional analysis of the *CDC42* gene 96
4.1 Introduction 96
4.1.1 Interactors of the *HOMER2* protein in the International Schizophrenia Consortium (ISC) GWAS data 96
4.1.2 Summary of chapter 96
4.1.3 Background on *CDC42* 96
4.1.3.1 The *CDC42* gene 97
4.1.3.2 The CDC42 protein 97
4.1.3.1 CDC42 and HOMER2 98
4.1.3.2 CDC42 and DISC1 98
4.1.3.3 CDC42 in *post mortem* schizophrenia brain 101
4.1.3.4 CDC42 and aging 102
4.1.4 Aims and summary of methods of investigation 102
4.2 Methods 103
4.2.1 Using PLINK to analyse sub-sample results at *CDC42* 103
4.2.2 Investigation of the 'mRNA-by-SNP Browser' database for effects of rs2473277 on *CDC42* expression 103
4.2.3 Analysis of *CDC42* expression in HapMap samples for effects of rs2473277 104
4.2.4 Bioinformatic selection of a functional variant 104
4.2.5 Reporter Gene Assays 105
4.2.5.1 Background on gene reporter assays 105
4.2.5.2 Cloning into the the pGL3-promoter vector 106
4.2.5.3 Preparation of cells for gene reporter assay 107
4.2.5.4 Quantification and comparison of reporter gene activity 108
4.2.6 Electrophoretic Mobility Shift Assays (EMSA) 109
4.2.6.1 Background on Electrophoretic Mobility Shift Assays 109
4.2.6.2 Preparation of nuclear proteins 109

4.2.6.3 Probe preparation 109
4.2.6.4 Protein-DNA binding reactions 110
4.2.6.5 Visualisation of protein-DNA binding 110
4.2.7 Neuropsychological correlates 110
4.2.8 Replication analysis methods in the GAIN and WTCCC2 samples 111
4.3 Results 112
4.3.1 ISC results across the *CDC42* gene 112
4.3.2 ISC results at rs2473277 114
4.3.3 Effect of rs2473277 on *CDC42* expression in the Dixon *et al* study 116
4.3.4 Effect of rs2473277 on *CDC42* expression in HapMap samples using Genevar data 118
4.3.5 Identification of a putative functional SNP in high LD with rs2473277 120
4.3.6 Gene reporter assay results 121
4.3.7 EMSA results 123
4.3.8 Neuropsychological results 124
4.3.9 Replication results in the GAIN and WTCCC samples 126
4.4 Discussion 128
Chapter 5: Detection of copy-number variation using next generation sequencing 134
5.1 Introduction 134
5.1.1 Defining copy-number variation 134
5.1.2 Molecular biology of copy-number variation 135
5.1.3 Population genetics of copy-number variation 135
5.1.4 Phenotypic effects of copy-number variation 136
5.1.5 Copy-number variation in schizophrenia 137
5.1.6 Detection methods for copy-number variation 138
5.1.7 Next-generation sequencing methods 139
5.1.8 Copy-number detection by next generation sequencing – whole genome resequencing 140
5.1.9 Targeted resequencing methods 144
5.1.10 Relevance of copy-number detection using next-generation sequencing to schizophrenia 145
5.1.11 Aims and experimental overview 145
5.2 Methods 147
5.2.1 Selection of copy-number variable regions for sequencing 147
5.2.2 Bait design for target enrichment sequencing of CNV and flanking regions 149
5.2.3 Sequencing of targeted sites using the Illumina Genome Analyser II 151
5.2.4 Raw data processing 152
5.2.5 Normalisation of sample read count and GC-content, and detection of CNVs 152
5.2.6 Determination of exact breakpoints by dideoxy terminator sequencing 153
5.3 Results 155
5.3.1 Total number of reads mapping for each individual 155
5.3.2 Relationship between GC-content and read count 156
5.3.3 Results of breakpoint determination using Sanger sequencing 158
5.3.4 Comparison of SNP-array breakpoints with true breakpoints 158

5.3.5 Results of CNV detection by NGS analysis 159
5.3.5 CNVs with gains of copy number 175
5.3.6 Summary of results of CNV detection by the NGS analysis method 177
5.4 Discussion 180
Appendix to Chapter 5: CNV detection by Next-generation sequencing 184
A5.1 Normalisation for Sample and GC-Content 184
A5.2 CNV Detection 185
A5.3. Diagrams detection of CNVs by analysis of NGS data 185
Chapter 6: General Discussion 218
6.1 Recent developments in schizophrenia genetics research 218
6.1.1 Genome-wide association studies 218
6.1.2 Copy-number variation 218
6.1.3 Next-generation sequencing 219
6.2 Summary of the aims of the project 219
6.3 Rationale for the identification of candidate susceptibility loci 220
6.3.1 Use of linkage meta-analysis data 220
6.3.2 Functional prioritisation of candidate genes 221
6.4 Candidate gene studies of schizophrenia 222
6.4.1 Candidate gene association study in the GASP schizophrenia sample 222
6.4.2 Association studies of the HOMER2 gene 222
6.4.3 Association analysis of CDC42 gene 223
6.4.4 Summary of association studies 223
6.5 Bioinformatic assessment of disease-associated variants 224
6.5.1 Potential effect of splicing on a schizophrenia-associated SNP at *HOMER2* (rs2306428) 224
6.5.2 Bioinformatic assessment of rs2473277 at the *CDC42* gene 225
6.6 Laboratory testing of associated SNPs with predicted functionality 226
6.6.1 Correlation of *HOMER2* expression with genotype at rs2306428 226
6.6.2 Reporter gene and electrophoretic mobility shift assays with *CDC42* SNP rs2473307 227
6.7 Endophenotype investigation of the associated SNPs at HOMER2 and CDC42 228
6.9 Detection of copy-number variation by next-generation sequencing 229
6.10 Future directions for schizophrenia genetics research 229
6.10.1 Mapping of causative loci from genome-wide association signals 229
6.10.2 Interpretation of results from exome sequencing 230
6.10.3 Functional studies of candidate susceptibility variants 232
6.11 Conclusion 233
On-line resources 234
References 235

Chapter 1

General Introduction

1.1 Schizophrenia

1.1.1 Clinical aspects of schizophrenia

Schizophrenia is a severe and enduring mental illness. It is common in populations across the world with a prevalence of 0.7-1.0% (Jablensky 2000). It has a significant effect on quality of life of both the affected individual and their families or carers.

A diagnosis of schizophrenia is made by psychiatric clinical examination according to criteria in the Diagnostic and Statistical Manual of Mental Disorders-4 (American Psychiatric Association 1994, table 1.1) or alternatively the International Classification of Diseases-10 (WHO 1992, table 1.2). These diagnostic criteria are the result of progressive modifications to Morel's démence précoce which was defined as cognitive impairment starting in late adolescence or early adulthood (Morel 1856, table 1.3). Although cognition impairment is common in schizophrenia, psychosis in the form of delusions and hallucinations and a chronic illness course were identified by Kraepelin as the key features of illness (Kræpelin 1919). Most tellingly these perceptual disturbances are characterized by a lack of insight, which leads the patient to believe them to be real (Schwartz 1998 for review). In modern diagnosis, these so-called 'positive' symptoms are recognized to be only one aspect of illness. In addition, patients often have so-called 'negative' symptoms which include poverty of thought and speech content, lack of motivation and social withdrawal (Andreasen *et al* 1982).

Schizophrenia is clinically heterogeneous and two patients may not necessarily share key symptoms. The course is fluctuating but in many cases progressive, with a substantial deterioration in normal function. Similar illnesses such as schizo-affective, schizo-typal disorder and bipolar disorders are classified and treated separately, but share common symptoms and a possibly shared biology (Craddock & Owen 2005, 2010). Psychosis may be a presenting feature of other disorders including neurological conditions e.g. epilepsy and substance-misuse (Semple *et al* 2006).

The 'positive' symptoms of schizophrenia are treatable with regular administration of antipsychotic drugs (Tandon *et al* 2010) and to a lesser extent cognitive behaviour therapy

(Tarrier 2010). However many patients have a chronic course of illness and require intensive care for long periods of time. Despite this, the condition is still associated with substantial mortality and reduces average life expectancy in those affected by more than a decade. The leading cause of premature death in people with schizophrenia is suicide (10-38%) with many others dying from un-natural causes (Mortensen & Juel 1993, Brown 1997). Additionally, the illness may cause significant distress among companions and stigmatisation by society. Finally, schizophrenia renders many individuals unable to maintain employment, requires significant medical support and thus has profound economic consequences. There is therefore a requirement for further investment into the prevention, management and treatment of schizophrenia. Through continuing investigations into the biological characteristics of individuals with schizophrenia and into its epidemiology, the processes which lead to this devastating disease are gradually becoming clearer.

Table 1.1: Criteria for the diagnosis of schizophrenia according to the Diagnostic and Statistical Manual of Mental Disorder (DSM-IV)

A. Characteristic symptoms: At least two of the following predominant throughout a 1 month period	▪ Delusions ▪ Hallucinations ▪ Disorganised speech ▪ Grossly disorganised or catatonic behaviour ▪ Negative symptoms i.e. Affective flattening, alogia or avolition
Only one of the above symptoms is required if delusions are bizarre or if hallucinations consist of a voice of a running commentary on the person's behaviour or two voices conversing with each other.	
B. Social/occupational	▪ Social/occupational dysfunction
C. Duration	▪ Continuous signs of the disturbance persist for 6 months including one month that meets criterion A.

Table 1.2: Criteria for the diagnosis of schizophrenia according to the International Classification of Disease, 10th revision (ICD10).

1. At least one of:	▪ Thought echo, insertion, withdrawal or broadcasting ▪ Delusions of control, influence or passivity ▪ Hallucinatory voices ▪ Persistent delusions of other kinds which are culturally inappropriate or implausible.
2. Or at least two of:	▪ Persistent hallucinations, accompanied by delusions without clear affective content, persistent over-valued ideas ▪ Breaks of interpolations in the train of thought resulting in incoherence or irrelevant speech or neologisms ▪ Catatonic behaviour such as excitement, posturing, negativism, mutism and stupor ▪ Negative symptoms such as marked apathy, paucity of speech and blunting of emotional responses ▪ Change in the overall quality of aspects of personal behaviour manifest as loss of interest, aimlessness, idleness, self-absorbed attitude and social withdrawal
3. Duration	▪ Duration of at least one month.

Table 1.3: Landmarks in the understanding of schizophrenia

Year	Author	Event
1856	Morel	*Démence précoce* is used to describe a young-onset dementia which is clinically distinct from senile dementia.
1911	Bleuler	The term 'schizophrenia' is coined (Greek, *σχιζ*=split, *ψρεν*=mind). Core symptoms are defined as: thought disorder, autism, affective blunting, attentional impairment, avolition and ambivalence.
1919	Kræpelin	Promoted schizophrenia as an organic, brain disease Distinguished schizophrenia from manic psychosis as being progressive and terminal.
1959	Schneider	Set out the 'symptoms of first rank': auditory hallucinations, thought interference and delusional perception
1966, 1970	Heston, Ketty	Adoption studies exclude the environmental impact on familial aggregation, thus implicating a strong genetic influence.
1975	Seeman & Lee	Antipsychotic drugs are shown to be strong dopamine receptor antagonist implicating neurotransmission in disease pathogenesis
1976	Johnstone	Increased ventricle size first identified as characteristic feature of the schizophrenia brain
1983	Fisher, Kendler	Twin concordance studies estimate heritability to be 0.6-0.8
1983	Anis	The schizogenic drug, phencyclidine is shown to be an antagonist of NMDA-type glutamate receptors sparking the glutamate hypothesis of schizophrenia.
1995	Lewis	2,945 schizophrenia cases in 1,208 pedigrees are analysed in a meta-analysis of 20 genome-wide linkage studies. Although no genes have been identified through this study, it marks the start of multi-national, genome-wide investigations.
2000	Millar	Partial translocation of Disrupted-in-schizophrenia-1 causes heterogeneous schizophrenia in a large pedigree and is later found to be crucial in neurodevelopment (Kamiya *et al* 2005).
2002	Stefansson	Through linkage studies in case-control samples, Neuregulin-1, is identified as a susceptibility gene and later shown to be important in NMDA receptor function (Hahn *et al* 2006).
2008	ISC, Walsh, Xu	The first genome-wide scans of copy-number variation show a role for rare structural variants in schizophrenia susceptibility.
2008, 2009	O'Donovan, ISC, S-Genes	The first genome-wide association studies of schizophrenia implicate novel genes in schizophrenia susceptibility.

1.2 Schizophrenia Epidemiology

1.2.1 Demographic prevalence of schizophrenia

Schizophrenia has an average global prevalence of 0.7% (95% C.I. 0.3-2.7%, Saha *et al* 2005) affecting people of every country and ethnicity (Jablensky *et al* 1993, Saha *et al* 2005, WHO 2007) There is no difference in the prevalence rates between males and females although males tend to have a lower age of onset and greater risk of recurrence (Loranger 1984, Aleman *et al* 2003).

1.2.2 Risk factors for schizophrenia

There is some geographic variation with higher incidence rates in cities (Faris & Dunham 1939, Pedersen & Mortensen 2001) and in immigrant populations (including first and second generation offspring, Veling *et al* 2008). Many peri-natal factors increase the risk of developing schizophrenia (Cannon *et al* 2003). The leading peri-natal risk factors include infections (notably *Rubella*), malnutrition and obstetric CNS damage (figure 1.1). However, the largest risk factor is having an affected first degree relative which increases the odds ratio to 10 times that of baseline (Cardno & Gottesman 2000).

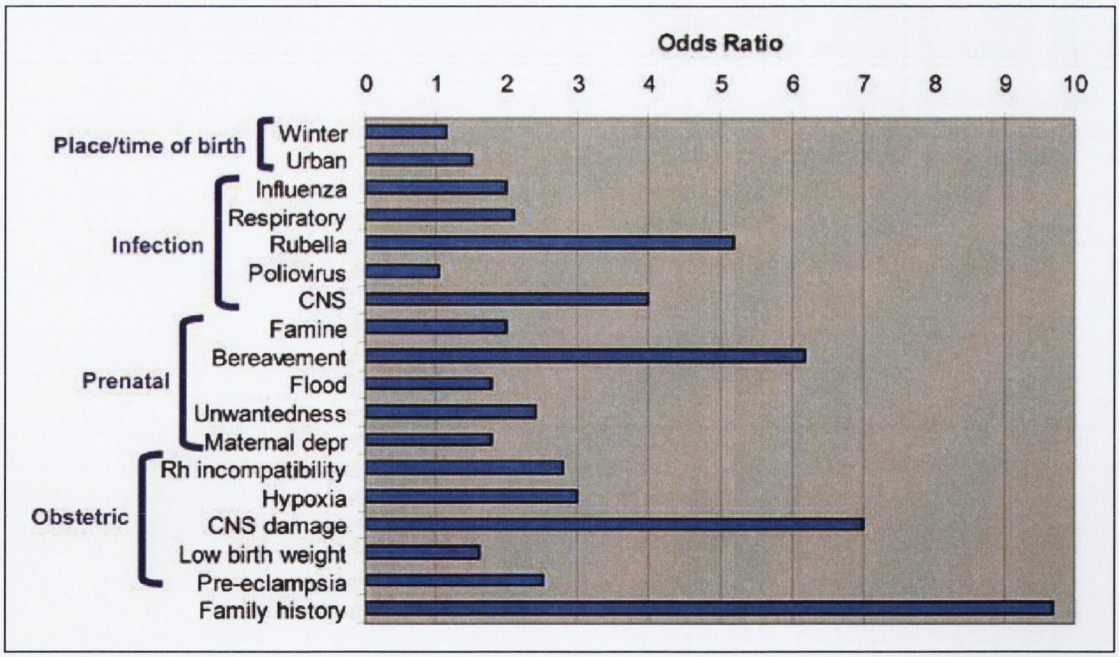


Figure 1.1 Bar-chart of risk factors for schizophrenia with focus on peri-natal factors (Murray *et al* 1992)

1.2.3 Heritability of schizophrenia

The aggregation of schizophrenia in families is well-known with approximately one-third of all cases having an affected first degree relative (Rudin 1916, Tandon *et al* 2008). Additionally, the increase in risk of schizophrenia increases with familial relatedness (figure 1.2, Kendler & Diehl 1993). This could potentially be due to shared environmental factors; however adoption and twin concordance studies have shown an overwhelming effect of genes over environment. Here, schizophrenia risk in adopted individuals significantly matches those of the biological parents compared to foster parents (Heston 1966, Ketty *et al* 1968, reviewed in Shih *et al* 2004). To complement this observation, monozygotic twin concordance rates are 40-50% compared to that for dizygotic twins and non-twin siblings which is only 5-15% (Fischer *et al* 1973, Kendler *et al* 1983, Cardno *et al* 1999). Heritability is defined as the proportion of variation in a disorder that is due to variation in genetic influences. Using data from the twin concordance studies, the heritability, h^2 of schizophrenia is 0.6-0.8 (Farmer *et al* 1987, Cardno *et al* 1999). Large studies using a liberal schizophrenia diagnosis found very similar heritability estimates to those with typical schizophrenia, indicating that both mild and severe forms share genetic susceptibility loci and thus may be classified as the same disease (h^2=0.62 compared to 0.68, McGue *et al* 1983). Modelling of schizophrenia inheritance patterns indicates that susceptibility is conferred by either the interaction of common variants in multiple genes and environmental/external factors or, in a small proportion of cases by rare, highly penetrant mutations (Gottesman & Shields 1982, Risch & Baron 1984). The former susceptibility model closely resembles the common-disease/common-variant hypothesis which is now accepted for the majority of common, mostly sporadic diseases (Risch & Merikangas 1996). Schizophrenia may be genetically heterogeneous as two individuals of similar phenotype may have different susceptibility loci (Fanous & Kendler 2005). Estimates of the number of schizophrenia genes have ranged considerably from five to thousands. It is likely that several of these genes have already been identified (Sullivan & Gejman 2010).

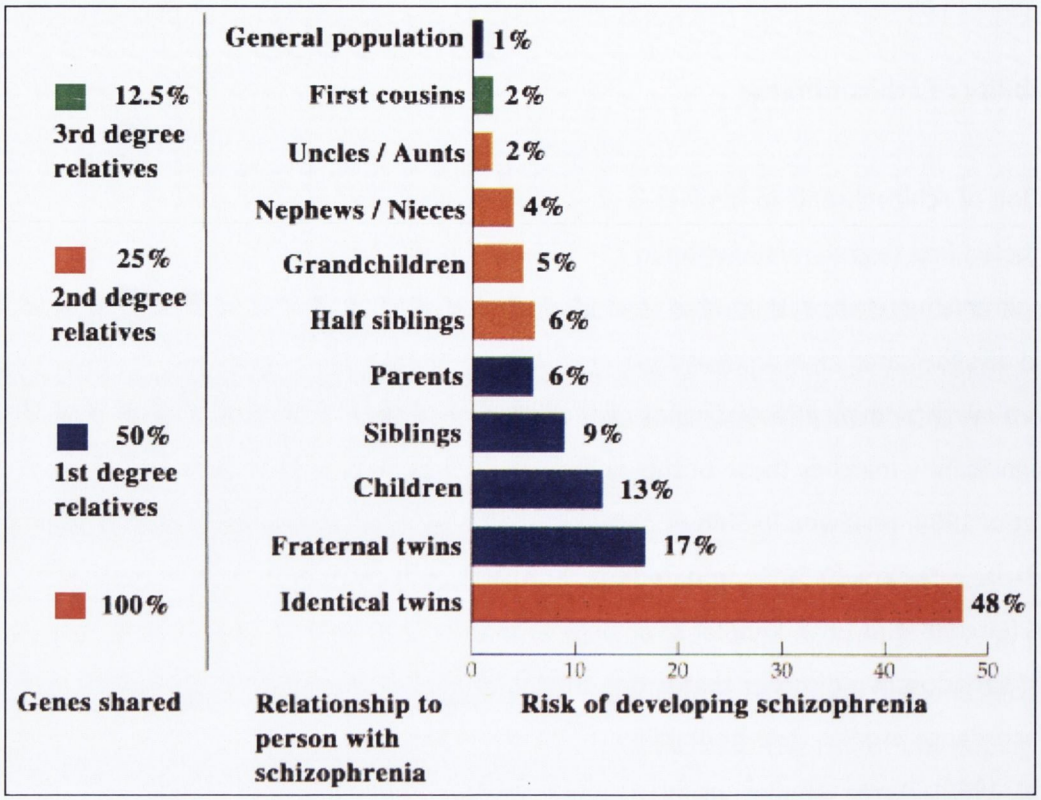


Figure 1.2 Bar-chart showing risk of schizophrenia compared with affection of family members (after Gottesman 1991)

1.3 The identification of schizophrenia susceptibility genes

1.3.1 Incentive for the investigation of schizophrenia genetics

A causative link between genetic variation and phenotypic variation was first established in fruit flies *Drosophilia melanogaster* (Morgan 1911). Since then, the cause of many human phenotypes, diseases and syndromes has been traced to both rare mutations and common polymorphisms. This has been greatly facilitated by the determination of the sequence of the human genome (Lander *et al* 2001), construction of recombination maps (Renwick 1971), and the positioning and functional annotation of genes (Schuler *et al* 1996, Deloukas *et al* 1998). Furthermore, much of this information is publicly and freely-available via the internet through bodies such as the National Centre for Biotechnology Information (NCBI, Jenuth 2000, http://www.ncbi.nlm.nih.gov/) and the University College Santa Cruz (UCSC) Genome Browser (Kent *et al* 2002, http://genome.ucsc.edu/).

Schizophrenia is highly heritable and thus the identification of genetic susceptibility variants not only elucidates disease pathogenesis but is methodologically tractable. In many other diseases, identification of genetic susceptibility variants has greatly contributed to understanding of disease pathogenesis. Such variants may often support existing pathology-based hypothesis, implicate completely novel genes and biochemical pathways in disease pathogenesis or even identify common genetic origins of distinctly different diseases. An additional benefit of identifying genetic susceptibility variants is they provide insight into the processes which lead to disease rather than the pathological investigations which are inevitably based on the effects of disease.

The identification of genetic susceptibility variants is particularly useful for a disease such as schizophrenia for which the underlying pathogenesis is poorly understood. In addition, genes and biochemical pathways which harbour these variants are a valid target for chemotherapeutic intervention. Finally, susceptibility variants may be used as a diagnostic tool (Philibert & Gershelfeld 2007). Such tests are common for rare and penetrant disorders and are an issue of substantial debate in the common genetic disorder community. Such diagnostic tests would be very useful for clinical management of schizophrenia particularly those which could predict disease course and drug response. By definition, a true susceptibility variant should have a quantifiable affect on gene function which can then be extended to a cellular phenotype. Mapping and functional confirmation of susceptibility alleles for common diseases including schizophrenia is a major on-going challenge of population and molecular genetics.

1.3.2 Background on linkage mapping

Linkage disequilibrium (LD) is the non-random segregation of genetic variants caused by a lack of recombination between the two loci (Ohta & Kimura 1969, Collins 2009 for review). In a population sample, LD between two markers is commonly measured as a correlation, r^2, in which an r^2 of 1 indicates that the variants are in complete LD (Mueller 2004, Maniatis 2007 for review). Mutations which cause disease may be identified through positional cloning by linkage mapping (Renwick 1971, Lander & Kruglyak 1995). This process involves the genotyping of thousands of highly-polymorphic microsatellite markers across the genome and comparing their transmission in affected and unaffected family members (Donis-Keller *et al* 1987, Kruglyak & Lander 1995). The sensitivity is primarily dependent on a causative variant being in linkage disequilibrium with a genotyped microsatellite marker. Linkage mapping in genetically complex diseases is feasible with variants of strong effect (Clark 2003) and has successfully identified common risk variants in genes such as CFTR1 (cystic fibrosis, Riordan *et al* 1989), ApoE (Alzheimer's disease, Corder *et al* 1993) and BRCA1 (Breast cancer, Friedman *et al* 1994).

1.3.3 Linkage studies of schizophrenia

Individual, population-based linkage studies of schizophrenia have lead to the identification of susceptibility genes such as Neuregulin-1 (Stefansson *et al* 2002), Dysbindin (Straub *et al* 2002a) and G72/G30 (D-amino acid oxidase activator) (Chumakov *et al* 2002). However replication of these findings in other samples is inconsistent. This may be due to the small effect size of the susceptibility variants combined with the small sample sizes used in the individual linkage studies.

To account for the possible lack of power of individual genome-wide linkage studies of schizophrenia a meta-analysis of schizophrenia linkage data has been performed (Levinson *et al* 2003, Lewis *et al* 2003). Here, data from a total of 2,945 schizophrenia cases in 1,208 pedigrees were analysed. As different studies used different sets of microsatellite markers, results from each were pooled into 120, 10cM bins. The study identified significant evidence of linkage to schizophrenia at nine of the bins, the most significant being at chromosome 2p12-q22.1. A further ten bins had evidence suggestive of linkage to schizophrenia (using criteria defined by the authors). The results of this meta-analysis are in theory extremely useful in the identification of susceptibility variants that are common to many different populations. Additionally the linkage results can be assumed to be similar for the many samples for which no genome-wide data is available. Despite these facts, candidate gene studies have favoured a functional prioritisation of target genes over linkage

mapping. Thus one of the aims of the present study is to use the linkage results from the meta-analysis to increase the likelihood of identifying susceptibility variants.

1.3.4 Background on genetic association studies

An association study is a powerful and commonly-used method of testing the impact of genetic variants on disease susceptibility (Abou-Sleiman *et al* 2006 for review). It is designed to detect small differences in the allele frequencies between case and control population at a gene hypothesised to have a role in disease pathogenesis. A significant difference between allele frequencies, as typically determined by a chi-squared test is indicative of a direct effect on disease susceptibility, albeit without proving causality. Thus, with a single marker of small effect, the ability to detect true association is dependent on its minor allele frequency (MAF) and the sample size (Freimer & Sabatti 2004).

Early pioneering association studies genotyped only the few known variants at any one gene and thus their coverage was extremely limited. Extensive genome re-sequencing has since identified millions of SNPs and other genomic variants, information on which is available online e.g. the dbSNP web-site (http://www.ncbi.nlm.nih.gov/projects/SNP/). Genotype data is available for ~3 million SNPs in the Caucasian Europeans from Utah (CEU) sample via the International HapMap Project (http://hapmap.ncbi.nlm.nih.gov/, HapMap 2003, 2005, 2007). This globally-utilised reference sample consists of thirty parent-off-spring trios and accurately reflects the LD structure of most Caucasian populations (Zeggini *et al* 2005). As a result, in an association study, prior knowledge of high LD between SNPs means that only a small proportion of the total number of SNPs at a gene need be genotyped in order to capture most of common variation across a gene. Therefore, association alone is not necessarily indicative of causation: many other variants may be in high LD with the associated SNP and thus, by proxy also be associated. Furthermore, the potential to detect true association at a gene of interest is dependent not only on the SNP allele frequencies and samples size but on the coverage of other genetic variation by the genotyped SNPs. Association studies have also been aided by the progressive growth of sample collections, high-throughput genotyping methods (Whitcombe *et al* 1998, Chen & Sullivan 2003) and freely-available data analysis software e.g. 'Haploview' (Barrett *et al* 2005) and 'PLINK' (Purcell *et al* 2007).

A major confounding effect in genetic association studies is population stratification (Tian *et al* 2008). Geographically-separated populations have limited gene flow between them and through genetic drift, natural selection and the generation of novel variants develop significantly different

allele frequencies (e.g. Campbell *et al* 2005). Thus, inclusion of individuals from different populations into either the cases or controls causes stratification in the sample which can potentially lead to false associations (Cardon & Palmer 2003, Tian *et al* 2008).

Population stratification is commonly corrected for using genomic control. This method estimates the overall inflation of significant associations due to stratification and uses principle component analysis (PCA) for correction (Price *et al* 2010). The inclusion of related individuals or 'cryptic relatedness' also causes stratification and can generally be detected following 'identity-by-descent' analysis. In the International Schizophrenia Consortium genome-wide association study (described below) multi-dimensional scaling was performed on the seven contributing population samples based-on SNP genotype calls (ISC 2009). This showed that the samples from the British Isles (including Ireland) form a single group whereas the others (from Sweden, the Portugese Islands and Bulgaria) form individual groups indicating that they are genetically separate populations (ISC 2009). Comparison of the schizophrenia association signals between the clusters indicated that allele frequency differences did not significantly influence p-values. Anomalous associations caused by differences in allele frequencies were further reduced by use of the Cochrane-Mantel-Haenzel test for association. This test groups case and control individuals of similar genetic background regardless of individual origin and produces association statistics based-on the average of all groups tested.

An alternative to the case-control test for association is the transmission-disequilibrium test (TDT) which measures the preferential transmission of alleles from heterozygous (unaffected) parents to affected offspring (Spielman *et al* 1993). It is considered to be less prone to false positives due to population stratification but requires more extensive sample collection. The gold standard in genetic association studies is consistent replication in independent and sufficiently-powered samples. The sample used for association analysis in the present study is a case-control sample from Ireland which is a putatively homogenous population.

1.3.5 Candidate-gene association studies of schizophrenia

Linkage studies in schizophrenia population samples identified Neuregulin-1 (Stefansson *et al* 2002), Dysbindin (Straub *et al* 2002a), G72 (Chumakov *et al* 2002) as candidate genes for harbouring susceptibility variants. Common SNPs at these genes were subsequently found to be associated with schizophrenia in the original study samples. Replication of these association signals in independent samples has been more successful than many other genes but still inconsistent. This may be due to small sample size, small effect size of the susceptibility variants, allelic heterogeneity (in which

different SNPs at a gene infer susceptibility to schizophrenia in different populations) or genetic heterogeneity (in which schizophrenia in different populations is caused by different genes). Nevertheless, all of the genes listed above have important functions in neuronal biology which further support their role in schizophrenia susceptibility (Tang *et al* 2009, Verrall *et al* 2009). For example, in the brain, Neuregulin-1 regulates neuronal migration (Anton *et al* 1997), dendrite development (Rieff and Corfas 2006) and receptor localisation (Buonanno & Fischbach 2001, Hahn *et al* 2006). Neuronal dysbindin-1 naturally binds synaptic and microtubule-associated proteins and in doing so links it with Disrupted-in-Schizophrenia-1(DISC1), mutations in which are known to cause rare, monogenic forms of psychosis (Talbot *et al* 2004, Camargo *et al* 2007).

Candidate gene association studies have been performed on many genes fulfilling topical hypothesis on schizophrenia aetiology without evidence of linkage. They have also been performed on genes in defined biological pathways (e.g. genes central to dopaminergic neurotransmission, Xu *et al* 2003). Early association studies tested the contribution of common variation at the MHC complex (Jönsson *et al* 1998, Hawi *et al* 1999) and dopamine receptor genes (Jönsson *et al* 1993). Functional properties of some common, associated variants have been demonstrated. For example schizophrenia-associated SNPs in the dysbindin-1 gene are consistently associated with specific reductions in cognitive performance (Fanous *et al* 2005, Burdick *et al* 2006, Fallgatter *et al* 2006) as well as in early visual processing (Donohoe *et al* 2008) and manic symptoms (Corvin *et al* 2008). Taken together, these findings implicate dysbindin in the pathogenesis of a negative, 'trait-like' schizophrenia (see section 1.4.4). Some schizophrenia-associated SNPs have been correlated with changes in gene expression e.g. in catechol-O-methyltransferase (*COMT*, Bray *et al* 2003) and nitric oxide synthase-1 (*NOS1*, Cui *et al* 2010). A schizophrenia-associated intronic SNP of the nitric oxide synthase 1 adaptor protein (*NOS1AP*) gene is not only correlated with expression but has been shown to affect reporter gene expression and transcription factor binding when studied in isolation (Wratten *et al* 2009, see section 1.3.7). Finally, schizophrenia-associated SNPs in the *NRG1* gene are also associated with alterations in isoform expression levels (Law *et al* 2006, Nicodemus *et al* 2009). Association studies of schizophrenia are recorded in the 'Schizophrenia Gene' on-line database which also hosts discussion forums on schizophrenia research (Allen *et al* 2008, http://www.schizophreniaforum.org/res/sczgene/). The process of candidate gene discovery and biological effects of susceptibility variants is regularly reviewed (Straub & Weinberger 2006, Owen *et al* 2007, Gill *et al* 2010).

1.3.6 The Rare Variant Hypothesis of Schizophrenia

Rare, causative mutations are a genetic variant which is only found in the presence of disease with the exception of delayed onset and incomplete penetrance. By definition they are rare with a population frequency <0.01 but may also be 'private' in which they are only identified in a single pedigree. Many rare diseases are caused by rare mutations which often (although not exclusively) alter the amino acid sequence of the encoded protein. Additionally, there are Mendelian forms of common, sporadic diseases which are caused by rare, highly-penetrant mutations in multi-generational families e.g. osteoarthritis (Palotie *et al* 1989), Parkinson's disease (Polymeropoulos *et al* 1997), high-density lipoprotein deficiency (Bodziok *et al* 1999) and prostate cancer (Carpten *et al* 2002).

The contribution of rare, highly-penetrant variants to schizophrenia was proposed in the early days of schizophrenia genetics research (Böök 1953) but a dearth of multi-generation families and modelling of transmission dynamics makes this unlikely to account for the vast majority of cases (Risch & Baron 1986). Nevertheless the identification of Disrupted-in-Schizophrenia-1 (DISC1) in a single extended pedigree has been of great benefit to schizophrenia genetics (St Clair *et al* 1990, Millar *et al* 2000). Here, linkage analysis combined with cytogenetic techniques identified a balanced translocation (1;11)(q42.1;q14.3) which leads to a truncated protein at the DISC1 locus on chromosome 1 in multiple family members with a spectrum of schizophrenia and other psychiatric phenotypes. The DISC1 protein has subsequently been found to control glutamatergic dendrite growth (Hayashi-Takagi *et al* 2010) with reduced expression inducing behavioural abnormalities in mice (Niwa *et al* 2010). No individuals outside of this family have been identified with DISC1 mutations with the exception of a pair of twins with broad psychiatric symptoms who harbour a large duplication which encompasses seven genes including DISC1 (Crepel *et al* 2010).

Interestingly, one of the leading risk factors for schizophrenia is velo-cardiofacial syndrome (VCFS) or DiGeorge syndrome (MIM 188400, 192430) in which approximately 25% of patients also develop schizophrenia (Karayiorgou *et al* 1995). VCFS/DiGeorge syndrome is characterised by velopharyngeal insufficiency, congenital cardiac defects and learning difficulties and is caused by deletions at 22q11.2. The numerical contribution of 22q11.2 deletions to schizophrenia is very low (~0.5%, Murphy 2002, ISC 2008) but nevertheless implicates a developmental origin of schizophrenia and, based-on linkage studies the region may harbour common susceptibility variants (MAF>0.1).

In addition, rare coding mutations in the ABCA13 gene have recently been identified in multiple, unrelated schizophrenia families (Knight *et al* 2009). Although there is a limited numerical contribution of rare variants to schizophrenia, they nevertheless exist and have the potential to

elucidate pathways to disease pathogenesis and ultimately improve treatment (Mitchell & Porteous 2010) (see section 1.3.8 for the role of copy-number variant in schizophrenia).

1.3.7 Functional investigation of genetic susceptibility variants

Although replication of association is a crucial factor in confirming that a genetic variant contributes to disease susceptibility, it must also been shown to exert an effect on gene function as this is a key step in understanding disease pathogenesis. In general, susceptibility variants in coding regions of genes will affect the behaviour of the protein. In Mendelian disorders, rare non-synonymous mutations cause protein alterations such as mis-folding and ecotopic cellular localisation (e.g. Valente *et al* 1999). Non-synonymous variants also exist at high frequency in human populations and may lead to common phenotypes. A common non-synonymous polymorphism in the COMT gene significantly alters the break-down of neuronal dopamine and has been associated with schizophrenia and other psychiatric disorders (albeit with limited evidence of replication).

Genetic susceptibility variants may also be located in non-coding regions of a gene and are thus likely to exert an effect by altering expression levels. Gene expression is finely regulated by the complex action of polymerases, transcription factors, enhancers and repressors which bind to specific sequences on the DNA (Maston *et al* 2006). Alteration of these binding motifs by genetic variation may thus affect gene expression levels. Such variation is termed cis-acting as the effects are exerted locally. Because gene expression is affected by the activity of many other proteins, variants which affect the expression or activity of these proteins may have a secondary effect of expression on a target gene. This is termed a *trans*-acting effect. *Cis* and *trans*-acting genetic variation is a major source of individual differences in gene expression and common phenotypes in all organisms. Indeed, expression levels of approximately one-third of all human genes expressed in lymphocytes are correlated with common *cis* or *trans*-acting variation (Cheung *et al* 2005, Dixon *et al* 2008). Similarly expression levels of 20% of genes expressed in the human prefrontal cortex are affected by common variants with the majority of *cis*-acting variants being located with 100Kb of the transcription start sites (Myers *et al* 2007). Such variants may lead to subtle cellular phenotypes and are currently under intense investigation for their contribution to a range of common diseases (Cookson *et al* 2009, Epstein 2009, Kleinjan & Coutinho 2009). Common variants which alter non-coding, regulatory sites confer susceptibility to a range of diseases including osteoporosis (Grant *et al* 1996) and cleft-lip (Rahimov *et al* 2008).

Finally, common, disease-associated SNPs may affect splicing. Splicing is the process during gene expression in which the consecutive exons of a gene are joined together following removal of introns from the primary mRNA transcript. Differential splicing causing the inclusion or exclusion of exons is a normal method of generating proteins of different function. Like gene expression, splicing is tightly-controlled (Chisa & Burke 2007), regulated by many other proteins and is affected by common SNPs (Kwan *et al* 2007, Kwan *et al* 2008). There are examples of Mendelian disease caused by rare mutations which affect splicing but few, if any know for complex diseases, probably because this has not yet been fully investigated. SNPs in the ERBB4 gene associated with schizophrenia have been correlated with alterations in isoform balance in the brain indicting a possible route to disease via disrupted interactions with its natural agonist, neuregulin-1 (Law *et al* 2007).

1.3.8 Copy-number variants and schizophrenia

A copy-number variant (CNV) is defined as the deletion, duplication or inversion of genomic DNA sequence between 1000 and 1,000,000bp long. Like single-nucleotide polymorphisms, many CNVs are common in the human population and exhibit standard Mendelian inheritance (MAF>0.1, McCarroll *et al* 2008). The direct biological consequences of CNVs are not well-characterised. Many common CNVs have no observable phenotype (e.g. Egan *et al* 2007) although some are correlated with changes in gene expression levels both in *cis* and in *trans* (Stranger *et al* 2007, Orozco *et al* 2009). Larger CNVs which disrupt many functional sites notably, whole or partial genes and extra-genic regulatory sequences, are likely to have profound biological effects. Indeed, gene 'knock-outs' in model organisms often display extreme phenotypes. Finally, many rare diseases are known to be caused by small CNVs ('InDels') usually in coding sequence (e.g. young-onset Parkinson's disease, Kitado *et al* 1998, OMIM 602544) as well as large chromosomal aberrations detected by fluorescence microscopy (e.g. Charcot-Marie-Tooth disease, Lupski *et al* 1991, OMIM 118220).

Large-scale and sensitive detection of copy-number variation has recently been made possible following technological advancements, culminating in genome-wide copy-number detection chips combined with automated calling algorithms. Large schizophrenia case-control samples have recently been subjected to a genome-wide CNV analysis with two major findings: deletions at chr1q21.1, chr15q11.1 and chr15q13.3 are rare, highly-penetrant risk factors for schizophrenia (ISC 2008, Walsh *et al* 2008, Kirov 2010 for review). Furthermore, schizophrenia patients harbor an excess of rare or unique CNVs which occur most often in neurodevelopmental genes (Walsh *et al* 2008, ISC 2008, Guilmatre *et al* 2009, Tam *et al* 2010). The numerical contribution of both common

and rare CNVs to schizophrenia susceptibility is becoming clearer and should effectively supplement results from SNP-based studies.

1.3.9 Genome-wide association studies of schizophrenia

The genome-wide association study (GWAS) is a recently developed method for the discovery of common disease susceptibility variants. It was intended to replace linkage-based methods which potentially lack the resolution required to identify multiple causative variants with high population frequency and small effect (Risch & Merikangas 1996, Collins *et al* 1997, Altshuler *et al* 2008). By comparison to a candidate gene study, a GWAS is hypothesis-free and thus allows the discovery of novel pathways to disease. Successful genome-wide association studies of Crohn's disease (Rioux *et al* 2007), breast cancer (Gold *et al* 2008), prostate cancer (Eeles *et al* 2008, Thomas *et al* 2008) and type-2 diabetes (Zeggini *et al* 2008) have been carried out by genotying 500,000-1,000,000 SNPs per individual. Two factors have made this possible: i) The discovery and cataloguing of millions of common SNPs across the human genome (Wang *et al* 1998, The International HapMap Consortium 2003, 2005), and ii) Technological improvements in high-throughput genotyping of SNPs (Fan *et al* 2003). Genome-wide association studies test many markers most of which are LD-independent and thus, the correction of significance levels for multiple-testing is reflective of this. Causative alleles of small effect are potentially lost through such a correction. To compensate for this, many sample collections from independent research groups have been combined to increase sample size and thus power to detect true associations.

Genome-wide association studies using large schizophrenia case-control samples have detected reproducible associations at Zinc-finger 804A (ZNF804A, MIM 612282, Williams *et al* 2010), Neurogranin (MIM 602350 Steffanson *et al* 2009) and the Major Histocompatibility Complex region (ISC 2009, Shi *et al* 2009, Stefansson *et al* 2009). Investigation into the role of these susceptibility variants is on-going. For ZNF804A, the established schizophrenia risk allele is correlated with reduced connectivity within the dorso-lateral prefrontal cortices and increased connectivity within the hippocampal formation (Esslinger *et al* 2009). It is furthermore associated with a schizophrenia-specific deficit in cognitive performance (Walters *et al* 2010) and with mania (Cummings *et al* 2010). The Psychiatric GWAS Consortium (PGC) has been formed in order to combine the results of multiple GWAS from common psychiatric diseases (PGC 2009). These are schizophrenia, bipolar disorder, major depression, autism, attention deficit hyperactivity disorder. A total of 31,659 cases and 26,945 controls will be analysed (PGC 2009). There is increasing evidence for shared genetic loci in these

disorders and so the pending 'mega-analysis' could confirm this as well as providing results on larger samples for individual diseases.

1.4 Aetiological hypotheses of schizophrenia

1.4.1 Schizophrenia pathology: Dysconnectivity and hypofrontality

There are no outwardly physical features of schizophrenia. As the disease is effectively behavioural, the brain is implicated as the organ in which the biological processes which lead to disease are located. Pathological connectivity between different brain regions is the earliest hypothesis as to the biological cause of schizophrenia and based upon neuro-psychiatric observations (Wernicke 1906, Bleuler 1911). The prefrontal cortex has been long considered as the region of the brain likely to be defective in schizophrenia because it controls complex cognitive behaviours and personality expression (see Robbins 1990).

These early hypothesis have been supported by studies of schizophrenia pathology in *post mortem* brain tissue which have consistently identified reduction in the volume of the prefrontal cortex (Lewis *et al* 2003). In addition, the neurons of the prefrontal cortex, whilst otherwise plentiful and healthy, have reductions in volume and number of dendritic spines (especially in deep layer 3, figure 1.4, Glantz & Lewis 2000). The hippocampus, which controls long-term memory, displays similar abnormalities which may also explain cognitive deficits in schizophrenia (Harrison *et al* 2004). There are many other reported abnormalities in schizophrenia brains *post mortem* but they are often based-upon insufficient sample sizes. Furthermore, the subtle changes in schizophrenia brain pathology may more often represent disease end-points and thus the causes of disease are inseparable from consequences (Lewis & González-Burgos 2008). Nevertheless, investigations into schizophrenia pathology have shown it to be quite different from common neurodegenerative disorders such as Alzheimer's and Parkinson's disease (Harrison 1999, 2004). Thus it is far for more likely to be caused by the dysfunction of higher-level neuronal systems which is initiated in adolescence and maintained throughout life.

Techniques such as florescence magnetic resonance imaging (fMRI) and positron emission topography (PET) have provided a basis for the study of schizophrenia brain function *in vivo* during functional paradigms (Agarwal *et al* 2010, Patel *et al* 2010). In support of dysconnectivity and hypofrontality as causes of schizophrenia, multiple functional imaging studies show imbalanced

activity of different regions of the brain with consistent deficit in prefrontal cortex activity (Weinberger *et al* 1992, Friston *et al* 1996, Davidson & Heinrichs 2003, Hill *et al* 2004). Interestingly such results often anticipate the onset of schizophrenia (Fusar-Poli *et al* 2007). Finally, measurements of 'Gamma activity' which is caused by co-ordinated oscillations of neuronal activity in the brain and measured using electro-encephalograms (EEG) are also consistent with irregular activation of specific brain regions (Lee *et al* 2003, Uhlhaas & Singer 2010).

Studies using EEG have also shown reductions in the P300 and P50 event-related potentials (Bramon *et al* 2004), pre-pulse inhibition (Bender *et al* 1999, Braf & Light 2005) and mis-match negativity (Umbricht & Krljes 2005). Deficits in these characteristics can be traced to multiple brain regions, notably the prefrontal cortex and hippocampus.

Microarray analysis is an advanced technique which quantifies global gene expression levels in tissue samples and are thus unbiased by any *a priori* hypothesis on disease pathogenesis (Bowtell 1999, Luo & Geshwind 2001). Broadly consistent findings in studies of the schizophrenia prefrontal cortex and hippocampus have been abnormal expression of genes involved in neuronal development, mitochondria and oxidative stress, oligodendrocytes, myelination, vesicle transport and neuronal glutamate processing (Kim & Webster 2010). Other than cross-experiment inconsistency, the interpretation of such results is compromised by small sample sizes and compounded by heterogeneity in drugs, disease duration and *post mortem* delay. Nevertheless they provide a huge amount of data which is being increasingly made public and thus open to secondary analysis through the Stanley Neuropathology Consortium (http://www.stanleyresearch.org, Kim & Webster 2010).

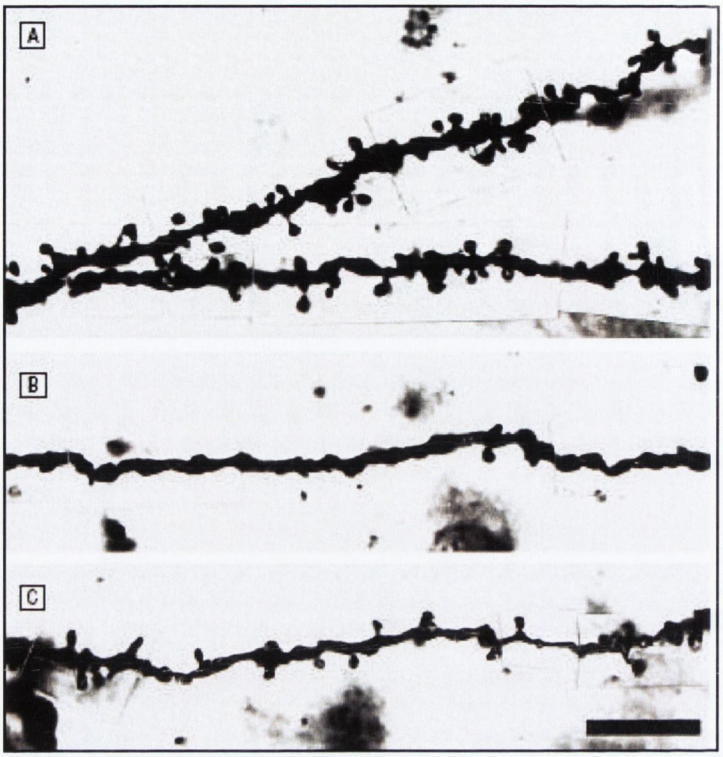


Figure 1.4 Dendritic spine loss in schizophrenia. A: normal control subject. B and C: unrelated schizophrenia cases. Image created as brightfield microscopy of Golgi-stained neurons from prefrontal cortex deep layer 3. Scale bar is 10μm (Glantz & Lewis 2000).

1.4.2 Neurodevelopment in schizophrenia: synaptic plasticity and pruning

The malfunction of connections between neurons, notably those connecting different brain regions, is considered to have a role in schizophrenia pathogenesis. Such connections or 'synapses' (Ramón y Cajal 1899, Loewi 1924) are established throughout childhood but some of these connections are lost during adolescence (Purves & Lichtman 1980, Sisk & Foster 2004). In keeping with the dysconnection hypothesis of schizophrenia, synapses in the prefrontal cortex and hippocampus have reduced activity as indicated by such markers as the Snap-25 and synapsin proteins (Eastwood & Harrison 1995, Thompson *et al* 1998). Furthermore, excessive synaptic 'pruning' during adolescence has been strongly postulated as a cause of schizophrenia (Feinberg 1983, Feinberg 1990, Kershavan *et al* 1994 for review). This is biologically plausible as the symptoms of schizophrenia often arise in late adolescence. In addition, stable synapse formation results in long-term potentiation. This underlies the biological basis of working memory and cognition, both of which are reduced in schizophrenia (Bliss & Collingridge 1993, Kullman & Lamsa 2007 for review). Synapses are sites of intense and complex protein activity which are assembled and dis-assembled in response to external

stimuli (Colón-Ramos 2009 for review). This dynamic behaviour of synapse formation is termed 'synaptic plasticity' and allows subtle changes in higher-order brain function (Bruner & Tauc 1965, Wurtz *et al* 1967). Synaptic plasticity was first identified in the hippocampus and has since been identified in the prefrontal cortex (Otani *et al* 2003).

Thus, defective synaptic plasticity is implicated in schizophrenia pathogenesis not only because it is most active in the period of life shortly before normal disease onset but also in regions of the brain already implicated. Synaptic plasticity thus has a direct effect on the higher-order neuropsychological functions which are most disturbed in schizophrenia. Although the close occurrence of adolescent synaptic pruning and the onset of schizophrenia symptoms has not been experimentally verified, the vulnerability of the developing brain to environmental damage means that deficits in synaptic plasticity which lead on to dysconnectivity are a plausible route to disease manifestation. Finally, the majority of the susceptibility genes identified through linkage and association studies of schizophrenia have primary functions in synaptic transmission, formation and integrity (Harrison & Weinberger 2003).

1.4.3 Schizophrenia drug targets and the dopamine hypothesis

Dopamine is a common neurotransmitter in the brain. Its activity controls cognitive ability, voluntary movement, motivation, and reward. Nearly all of the drugs used in the effective treatment of schizophrenia are strong dopamine receptor antagonists and subsequently reduce dopaminergic activity (Carlsson & Lindqvist 1962, Seeman *et al* 1975, Snyder 1981). In addition psychotic symptoms of schizophrenia are induced by drugs which stimulate dopaminergic activity (Laruelle *et al* 2003). Thus, it was initially proposed that decreased dopamine activity in the brain could lead to schizophrenia. Subsequently, reduced activity of mesocortical dopamine projects in schizophrenia was demonstrated (Weinberger *et al* 1988). Conversely dopamine activity is increased in other regions of the brain, notably the striatum (Laruelle *et al* 1996) and may cause the 'positive' symptoms in schizophrenia (Abi-Dargham & Moore 2003).

Dopamine is not the only neurotransmitter in the brain with GABA, serotonin and glutamate also forming major functional systems. Of these, glutamate is the most common in the prefrontal cortex and hippocampus. Its activity should thus determine those behaviours which are dysfunctional in schizophrenia.

1.4.4 NDMA receptor hypofunction and the Glutamate hypothesis

Glutamate is the most abundant excitatory neurotransmitter in the vertebrate nervous system. In stimulated neurons, glutamate is released from vesicles at the pre-synaptic cell membrane and binds to receptors on the post-synaptic cell membrane. Activation of NMDA-type glutamate receptors stimulates the influx of calcium ion transport and so completes the transmission of the nervous impulse from one cell to the next. Synaptogeneis is directed by this stimulus and thus the activities of all proteins involved are crucial for synaptic remodelling, plasticity and the establishment of long-term potentiation (Collingridge *et al* 2004). Ultimately this determines the activity of cognitive functions such as learning and memory (McEntee & Crook 1993).

The 'glutamate hypothesis of schizophrenia' (Coyle 1996, 2006) is founded on the observations that phencyclidine hydrochloride (PCP) induces symptoms of schizophrenia in both stable patients and healthy volunteers (Luby *et al* 1959, Davies & Beech 1960) and also is a potent antagonist of NMDA-receptors (Anis *et al* 1983). Similar NMDA receptor antagonists such as ketamine and MK-801 also induce relapse of schizophrenia or increase the severity and can induce psychotic symptoms in healthy volunteers (Rosenbaum *et al* 1959, Cohen *et al* 1962). Additionally these drugs induce schizophrenia-like symptoms in rodents and specifically induce neuronal degeneration across the brain, markedly in the pre-frontal cortex (Olney & Faber 1995). Thus it was proposed that schizophrenia was caused by hypofunction of NMDA-receptors leading to compensatory and excessive glutamate release which would in turn induce neuro-degeneration. As a supplement to the previously postulated dopamine hypothesis of schizophrenia, glutamate dysfunction may explain the preponderance of treatment-resistant cases, structural brain changes and cognitive deficits (Stone *et al* 2007). Additionally, as glutamatergic neurotransmission is more important for synaptic plasticity in early life, the glutamate hypothesis may explain the distinctive age at onset of schizophrenia (du Bois & Huang 2007). It remains to be determined though whether schizophrenia is glutamate toxicity is a cause of schizophrenia or whether it is brought about by hypofunction of glutamatergic neurotransmission.

Additional evidence from pathological and neuro-imaging studies of schizophrenia supports the general theory of glutamatergic dysfunction. There is altered distribution of NMDA-receptors brains of schizophrenia patients (Coyle & Tsai 2004), altered levels of glutamatergic metabolites (N-acetylaspartate, glutamate and glutamine) in the brain at disease onset (as determined by single-voxel proton magnetic resonance spectroscopy, Bustillo *et al* 2010) and expression changes genes involved in neuronal glutamate processing (from a microarray study of the pre-frontal cortex,

Tkachev *et al* 2007). In further support of the glutamate hypothesis is the fact that through in the search for genetic susceptibility variants for schizophrenia, unbiased linkage studies have identified causative genes which have central roles in the glutamatergic neurotransmission, neuronal plasticity, learning and memory (Harrison & Weinberger 2005). For example the Dysbindin protein controls NMDA receptor surface expression and in doing so affects long-term potentiation (LTP) in the hippocampus (Tang *et al* 2009). Thus, it is likely that other genes harbouring schizophrenia susceptibility variants also have a profound influence on glutamatergic neurotransmission and on synaptic plasticity.

1.4.5 The endophenotype hypothesis of psychiatric disorders

Many of the subtle differences in schizophrenia patients described above anticipate the onset of disease, may occur in unaffected family members and also in individuals with a similar, non-schizophrenia diagnosis. Such observations are the basis of the 'endophenotype' hypothesis (Gottesman & Gould 2003, Cannon 2005). It was initially proposed that different susceptibility variants lead to distinct sub-clinical, neurological deficits i.e. endophenotypes. An individual harbouring an endophenotype would thus not only be more likely to develop schizophrenia but in doing so, maintain that endophenotype. Family members of schizophrenia patients would thus share and pass on risk endophenotypes without necessarily developing the disorder. This partially explains the observation that risk of developing schizophrenia is the same for the offspring of both members of a set of a discordant monozygotic twin pair (Gottesman & Bertelsen 1989). The following criteria classically define an endophenotype (Gottesman & Gould 2003):

1. Co-occurrence with schizophrenia in the population
2. Heritability
3. State-independent (exists in an individual prior to the onset of symptoms)
4. Co-segregates with schizophrenia in a family
5. Exists in members of the family with and without schizophrenia at a higher prevalence than the population base-line
6. In practical terms the phenotype also needs to be easily and accurately measurable.

The three core psychological deficits of schizophrenia – cognition, attention and working memory – are regarded as the most common endophenotypes. These core endophenotypes may each be

broken-down further into endophenotypes which are quantified specifically in a battery of verbal or manual tasks. For example, cognition may be quantified as performance IQ, verbal IQ and full-scale IQ (Burdick *et al* 2009). In doing this, the different neronal systems which control these individual endophenotypes may be tested against a schizophrenia susceptibility genotype or allele. Endophenotype studies have also investigated genetic effects on clinical features, neurological 'soft signs' e.g. impaired motor co-ordination (Chan *et al* 2010) and neuropathology which is usually determined using fMRI or PET (Kaymaz & van Os 2009).

Schizophrenia susceptibility genes which were identified through candidate gene association studies have been tested for their impact on endophenotypes. Among the most significant and reproducible results come from *DTNBP1* (with negative symptomology (Fanous *et al* 2005, Corvin *et al* 2008), *COMT* (with 'N-Back' performance, Flint & Munafò 2007 for meta-analysis). Many of postive findings from endophenotype studies are supported by animal models which exhibit the equivalent endophenotype behaviours (e.g. Babovic *et al* 2007). The use of animal models of gene disruption here is of great advantage because they are more amenable to pathological, cellular and biochemical studies than human subjects (Amann *et al* 2010 for review).

Genome-wide association studies have also contributed to understanding of endophenotypes. For example, the schizophrenia risk allele at *ZNF804A* gene (which shows reproducible genome-wide association in large schizophrenia samples; O'Donovan *et al* 2008, ISC 2009, Williams *et al* 2010) is associated with enhanced connectivity in the hippocampus (Esslinger *et al* 2010) and with a reduction in cognitive impairment in schizophrenia (Walters *et al* 2010). Conversely the risk allele is also associated with reduced connectivity in the dorso-lateral prefrontal cortex (DLPFC) which may explain preliminary indications that it is also associated with manic episodes in schizophrenia (Cummings *et al* 2010).

Generation of data for the study of endophenotypes is labour-intensive. However, once assimilated it has the advantage of being amenable to rapid testing against genetic variants with a role in schizophrenia susceptibility. Furthermore, the alliance of a specific neuropsychological system to a susceptibility variant generates a hypothesis for exploration with functional neuroscience. A final advantage is that susceptibility variants should have a stronger effect size for an endophenotype than the do for the clinical spectrum of schizophrenia and thus be more amenable to detection by association studies.

1.5 Project aims

Schizophrenia is a disorder for which there is a great requirement for diagnostic tools and new therapeutic interventions. Studies on the biology of schizophrenia are largely inconclusive as to the causes of disease with the exception that the observed dysfunction of the glutamatergic neurons may explain many of the other characteristics. As schizophrenia is highly-heritable and many identified genes are have some involvement in the activity or development of the glutamatergic system it is likely that similar genes also harbour schizophrenia susceptibility variants.

The meta-analysis of genome-wide linkage studies of schizophrenia has provided information on the likely locations of schizophrenia susceptibility genes. Genes located within these regions and which are intimately involved in the functioning of the glutamatergic nervous system are thus excellent candidates for harbouring susceptibility variants for schizophrenia. Thus, the first broad goal of this project is to identify such genes and test common variants within them for association with schizophrenia.

Mid-way through this project the International Schizophrenia Consortium completed and published a genome-wide association study of 3,380 schizophrenia cases and 3,593 controls. The Neuropsychiatric Genetics group within which the current project is based is a member of the ISC and has unlimited access to the complete data-set. Because each individual was genotyped for hundreds of thousands of common SNPs, potentially causative variants can be rapidly identified and tested for functional effects. Thus, a second goal of this study is to identify variants which are not only associated with schizophrenia (in the ISC GWAS) but which have an effect on the function of genes which are involved in glutamate function.

Although a vast number of common SNPs have been reported and are tested in association studies, there are likely to be many more which are unknown. In addition, non-SNP variation such as Indels and CNV's are not well-catalogued and may be important sources of functional genetic variation. Furthermore, the role of single, 'private', highly-penetrant mutations in schizophrenia is still unconfirmed and requires further investigation. The introduction of next-generation sequencing technology means that candidate genes identified through genome-wide association studies can now be studied in great detail in a large number of individuals at a fraction of the cost and time of traditional dideoxy-terminator methods. This method of sequencing was developed with the intention of identifying single-nucleotide changes but is also likely to be able to detect indels and copy-number variants by changes in the depth of coverage. The last section of this project investigates how well sequencing-by-hybridisation identifies copy-number variations with a view to developing an algorithm for standardised, automated detection when re-sequencing candidate genes.

Aim 1: To identify candidate genes for schizophrenia by combining linkage and functional information (related to glutamate neurotransmission) for genetic association studies using a case-control study design (Chapter 2).

Aim 2: Follow-up of significant associations from this analysis, by replication in the International Schizophrenia Consortium (ISC) GWAS dataset. From these studies the gene HOMER2 was taken for additional functional assessment (Chapter 3)

Aim 3: Investigation of potential HOMER2 interacting genes across the ISC sample: using both association and functional analyses (Chapter 4)

Aim 4: Determining performance of next-generation sequencing for the identification of copy number variation (Chapter 5).

Chapter 2

A candidate gene study of schizophrenia

2.1 Introduction

2.1.1 Chapter overview

The previous chapter introduced current theories on schizophrenia pathogenesis and the contribution of molecular genetics to the disorder. This chapter describes the selection of candidate genes and their subsequent examination for association with schizophrenia. The strategy used for candidate gene selection is based upon existing schizophrenia linkage data and prior knowledge of gene function with relevance to schizophrenia pathogenesis and previously reported putative schizophrenia susceptibility genes.

2.1.2 Schizophrenia linkage data

Linkage analysis provides hypothesis-free positional information on susceptibility loci but requires large samples and can have limited positional resolution. Prior to genome-wide association studies (and at the time this project was started) linkage studies provided the only unbiased genome-wide evidence as to the location of susceptibility genes. The most extensive source of linkage data available are the results of a meta-analysis of schizophrenia linkage studies (Levinson *et al* 2003, Lewis *et al* 2003). Here, data was pooled from twenty independent studies, thus creating a statistically more powerful sample than that of any individual study. The meta-analysis results were for 120 genomic bins each assigned with a score based upon evidence of linkage from each contributing study and sample size for each study. The size of each bin was set at 30cM which in physical terms gave an average bin size of 28.9Mb (range 13.6-49.0 Mb). Significant evidence of linkage was found at ten bins and a further nine bins were considered to be suggestive of linkage. It may be possible to fine-map causative genes in these regions, especially in the sample population from which the linkage signal originates. However, with increasing annotation of gene function and better understanding of schizophrenia pathogenesis, an alternative approach is to test individual candidate genes for association following functional prioritisation.

2.1.3 Glutamate and the synapse in schizophrenia

The observation of glutamatergic neuro-transmission dysfunction in schizophrenia supports its central involvement in disease pathogenesis (Coyle 1999, Sodhi *et al* 2008). Genes encoding proteins which are involved in glutamatergic neuro-transmission are logical candidates for harbouring schizophrenia susceptibility variants. On-line databases such as the NCBI Entrez-Gene (Maglot *et al* 2005) and UCSC Genome Browser (Kent *et al* 2002) contain records of all documented genes in the human genome. They also contain some information on gene function, including lists of relevant publications and curated biochemical pathways. Such databases are amenable to key-word searches and so were used here to identify functional candidates.

The neuronal synapse is implicated as the site of pathogenesis in the glutamate hypothesis of schizophrenia (Coyle 1999, Owen *et al* 2005). Thus, a specific aim was to compile a list of genes whose products localise to the neuronal synapse and to subsequently use this to filter the list of genes which met the glutamate search criterion. A large number of potentially uncharacterised proteins are believed to localise to the synapse so data was also obtained from a proteomic survey of the mouse synapse (Shrimpf *et al* 2005) in addition to the key word searches of NCBI Entrez-Gene and UCSC Genome Browser. By considering only the overlap between synaptic and glutamatergic gene lists, functional candidates would be more specific to the glutamate hypothesis of schizophrenia as followed in this study.

2.1.4 Interactors of known schizophrenia genes

Replicated association signals and functional validations have lead to the discovery of a number of putative schizophrenia genes (Owen *et al* 2005, Harrison & Weinberger 2005). Although the contribution of SNPs at these loci to schizophrenia susceptibility is often of small effect, they have identified biochemical pathways that may be at the origin of schizophrenia pathology. Thus, functional alteration of other genes involved in these pathways by SNPs could also increase the risk of developing schizophrenia. In addition, there is evidence for epistatic risk factors in schizophrenia between genes whose proteins directly interact in normal cell function (Norton *et al* 2006, Nicodemus *et al* 2007, Morris *et al* 2008).

For these reasons a screen was conducted for genes whose protein products have a reported biochemical interaction with the proteins of putative schizophrenia susceptibility genes. Studies of protein-protein interactions commonly use co-immunoprecipitation and yeast-2-hybrid methods, although array-based approaches, e.g. to identify post-translational modification targets, are increasingly in use (Kang *et al* 2009, Wolf-Yadin *et al* 2009). Results of such studies are compiled and accessed in on-line databases such as STRING (http://string.embl.de/) and BIOGrid (http://www.thebiogrid.org/). Interacting genes were considered as potential candidates for association analysis if they were located within regions of the genome which had evidence of linkage to schizophrenia.

The flow-chart in figure 2.1 summarises the strategy of gene selection, SNP analysis and association study. In the first stage of candidate gene selection, lists of genes fulfilling positional and functional criteria were generated and the overlap compared. Genes considered to have an important role in glutamatergic neurotransmission and with evidence of linkage were considered as primary candidates. Secondary candidates were prioritised from genes with evidence of linkage and which were found to interact with putative schizophrenia genes and/or fulfilled either of the glutamatergic or synaptic search terms alone. Additional weighting was placed on genes if they were the only one in a linked region which fulfilled functional criteria and if they had functional evidence of potential involvement in neuropsychiatric phenotypes. Genes were excluded if they were very large, of limited or unknown functional relevance or had previously been studied with no evidence of association.

In the second stage of the project, following selection of the candidate genes, an extended tagging-SNP selection method was applied to each gene to identify common variation at each locus. Tagging-SNPs selected from patterns of LD in the HapMap CEU reference panel are an effective and efficient method of mapping independent common variation in a given region (Carlson *et al* 2003, The International HapMap Consortium 2005, http://hapmap.ncbi.nlm.nih.gov/). This has been demonstrated in an Irish sample similar to the one used in the current study (O'Dushlaine *et al* 2008). However, there are many more SNPs reported in dbSNP than have not been genotyped in the HapMap CEU samples (Sherry *et al* 2001). Some of these SNPs may be common and have an impact on gene function but could be missed if a study relied on HapMap CEU information alone. To compensate for this, selected genes were screened *in silico* for predicted functional SNPs for which genotype data was not available in HapMap. These SNPs were tested for heterogeneity, genotyped in-house in the CEU samples and the resultant data was integrated into existing HapMap data prior to selection of tagging-SNPs.

Finally, by genotyping the tagging SNPs in the case-control sample and comparing the allele frequencies between case and control groups, it was hoped to identify schizophrenia susceptibility variants which could then be validated through replication in independent samples and investigated further at the level of molecular function and endophenotype.

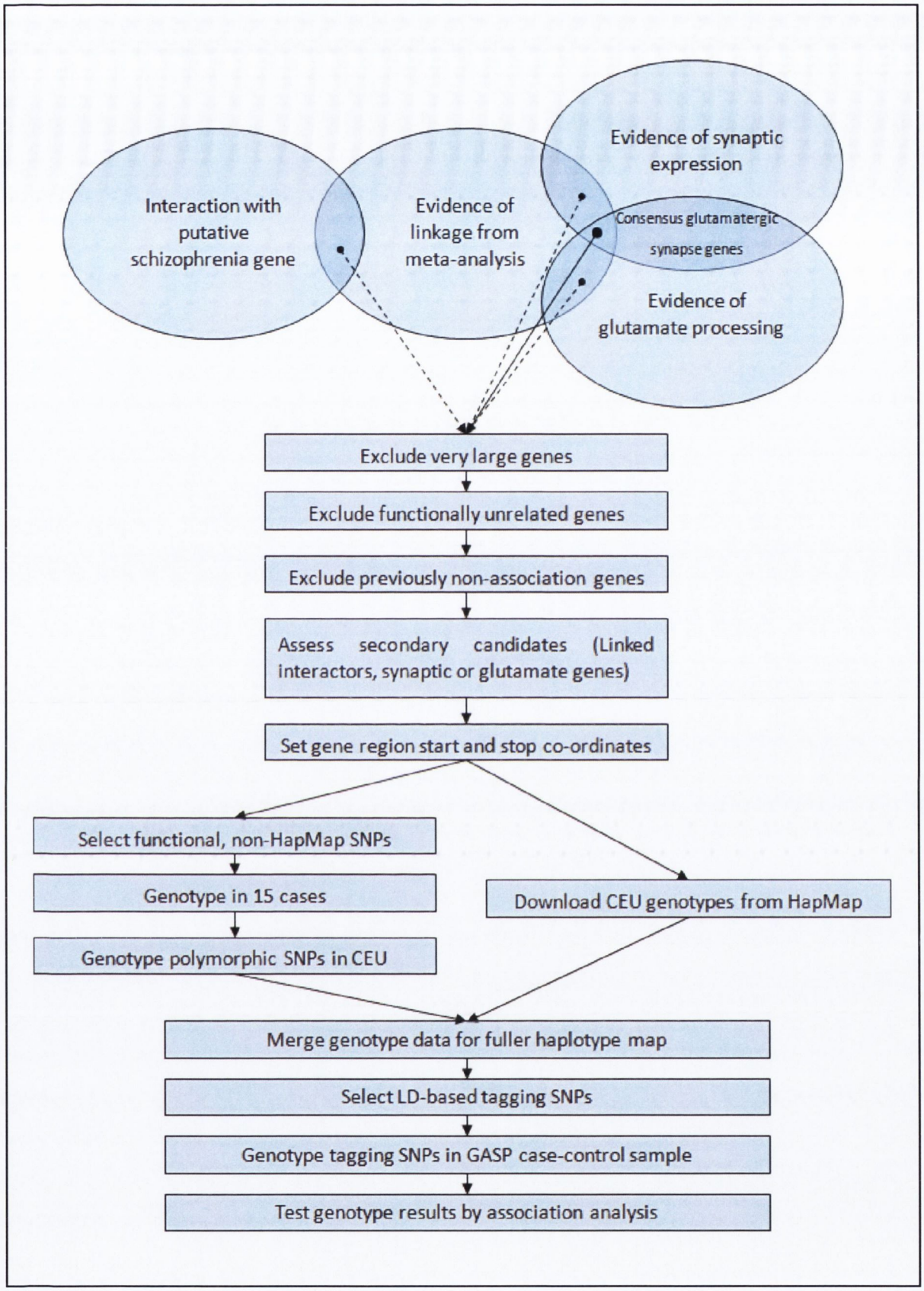


Figure 2.1: Flow-chart of experimental approach for candidate gene selection, functional SNP investigation and association study.

2.2 Methods

2.2.1 Selection of candidate genes

2.2.1.1 Identification of genes within linkage bins

The meta-analysis of genomic-wide linkage studies of schizophrenia identified ten bins with significant evidence of linkage and a further nine bins with suggestive evidence of linkage (Lewis *et al* 2003). All nineteen bins were considered here as potentially harbouring candidate genes. To obtain a list of genes within the nineteen highest-scoring bins from the linkage analysis by Lewis and colleagues, the proximal and distal microsatellite markers which defined the limits of each bin were entered into the UCSC Genome Browser, Tables function with the output set to 'Genes and Gene Prediction Tracks, Known Genes' (See Figure 2.2). The output was filtered for duplicates in Excel and stored as a single list of gene symbols in a text file.

2.2.1.2 Identification of glutamatergic synapse genes

Functional candidates were obtained in two separate searches which were then combined to form a list of consensus glutamatergic synapse genes. To search for genes with any involvement in glutamate processing or activity at glutamatergic synapses, the search-term "glutamat*" was used to query the NCBI Entrez-Gene and UCSC Genome Browser public databases. These two semi-independent databases were searched to maximise coverage. All genes identified by searching these two databases were considered to have some involvement in glutamate processing at this stage.

The same two databases were used to search for genes whose protein products localise to the neuronal synapse using "synap*" as the search-term. To supplement this, a list of proteins detected in the proteomic survey of the mouse synapse was obtained from Dr Sabine Schrimpf (University of Zurich, Switzerland). As the records were in a murine protein-coded format, they were converted to the equivalent human gene symbol using the ID-converter tool on the 'DAVID Bioinformatics' web-site (Dennis *et al* 2003, Huang *et al* 2009). Output from all three data-sources was merged, filtered and stored as before. Genes fulfilling both synaptic and glutamatergic search criteria were isolated using the *make common elements.pl* Perl script (written by Dr Colm O'Dushlaine of Neuropsychiatric Genetics, TCD) to obtain a consensus list of genes considered to have a role in glutamatergic neurotransmission at the synapse and thus likely to harbour schizophrenia susceptibility variants.

The consensus glutamatergic synapse genes were then screened against those genes with evidence of linkage, again using the *make common elements.pl* Perl script. Genes fulfilling both these criteria were considered as primary candidates for association analysis. These primary candidates were screened for the number of tagging-SNPs required for comprehensive association study, previous (well-powered) association studies and literature-based functional relevance to schizophrenia.

2.2.1.3 Identification of interactors of schizophrenia genes

Putative schizophrenia genes were identified through a search of literature reviews of schizophrenia genetics (Owen *et al* 2005, Harrison & Weinberger 2005) and of recent publications which may not have been considered in the reviews (e.g. Mah *et al* 2006). Also considered as putative schizophrenia genes (and thus investigated for interactors) were three genes with important functions in glutamatergic neuro-transmission which were disrupted by rare copy-number variants in the Stanley schizophrenia brain sample (Wilson *et al* 2006). Interactors of proteins encoded by the schizophrenia genes were identified through publications (notably Millar *et al* 2003 and Camargo *et al* 2007) and the on-line protein interaction databases, STRING (Snell *et al* 2000, http://string.embl.de/) and BIOGrid (Stark *et al* 2006, http://www.thebiogrid.org/). A complete list of all genes whose proteins form a direct, physical interaction with proteins of putative schizophrenia genes was compiled and stored as above.

Primary candidates were considered as those genes which were considered to be involved in glutamatergic synapse function and show evidence of linkage. Secondary candidates were obtained by screening individual lists of glutamatergic, synaptic or interactors of schizophrenia genes for evidence of linkage. Priority was given to genes which fulfilled two of these criteria, interacted with more than one schizophrenia gene and which were located in a linked region with no other functional candidates (thus increasing the chances that it was a true susceptibility gene given the aetiological hypotheses followed in this study). Secondary candidates were also checked for required number of tagging-SNPs and relevant publications as described above.

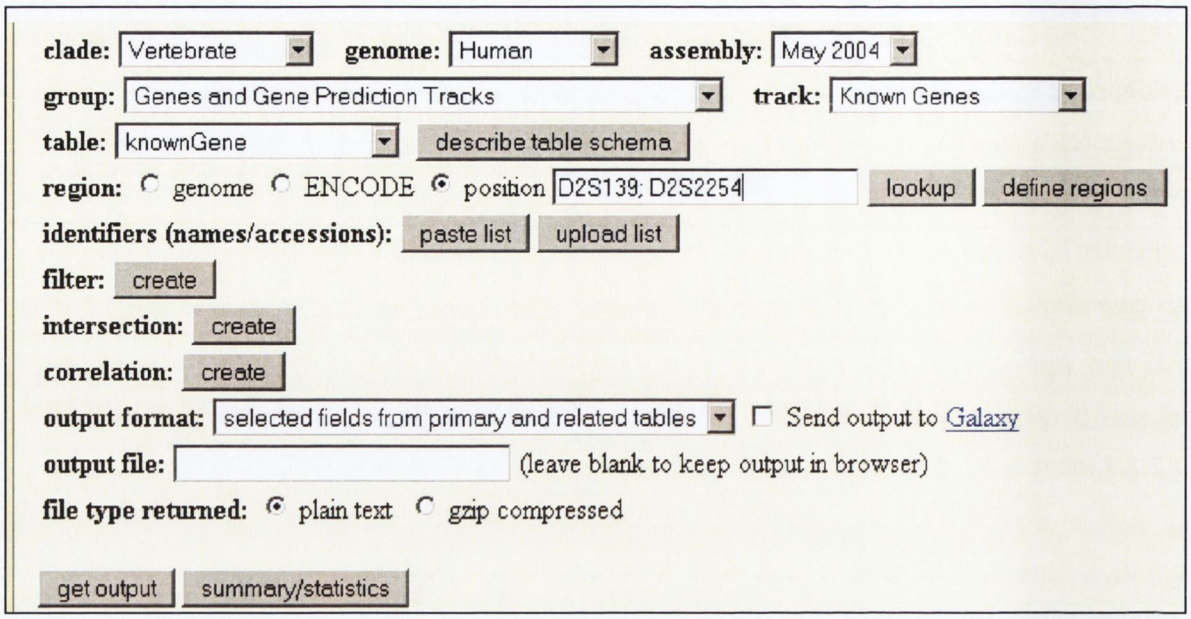


Figure 2.2: Screen-shot from the UCSC Genome Browser, Tables. This demonstrates how a list of genes between two markers specified in the Lewis study was generated.

2.2.2 The GASP (Genetic Association study of Schizophrenia and related Psychoses) schizophrenia case-control sample

Case and control DNA was available for association studies at the start of this project. The schizophrenia sample consisted of 375 cases which had been interviewed with a psychiatrist or psychiatric nurse trained to use the Structured Clinical Interview for DSM (SCID-P) and Positive and Negative Symptom Scale (PANSS). All cases provided written informed consent and were over 18 years old. Diagnosis was made by the consensus lifetime best estimate method with DSM-IV criteria using all available information. Ethics Committee approval was obtained from all participating centres. The control sample was drawn from 812 blood donors in Ireland who are unlikely to have schizophrenia as law excludes those on medication from donating blood. Extraction of genomic DNA from whole blood samples was performed by the phenol-chloroform method (Sambrook *et al* 1982), QIAamp DNA Blood Maxi Kit (Qiagen, Crawley, UK) or a semi-automated 'Gentra' method (Qiagen) at the Trinity Biobank. Stock DNA was dissolved in standard Tris-EDTA buffer and stored at -20°C.

Stock DNA quantified by a Nanodropper (Nanodrop Ltd., Wilmington, DE, USA) was used to create working stock of 10ng/µl. For ease of genotyping 96-well plates of DNA samples were made-up. A number of quality control measures were implemented at this stage to minimise the risk of systemic errors biasing the overall outcome. These were: mixing case and control DNA samples

across each plate, use of non-template (negative) controls, known genotype controls (HapMap CEU samples) and duplicate controls. Genomic DNA for the HapMap CEU sample was obtained from the Coriell Institute for Medical Research, USA.

2.2.3 Novel SNP selection and genotyping

2.2.3.1 Basis for genotyping novel SNPs

The selected genes were screened for predicted functional SNPs using the Intersection function on the UCSC Genome Browser Tables section. This was used to exclude HapMap SNPs and overlay information on the functional regions of each gene with the remaining non-HapMap SNPs from dbSNP. Such information included proximity of SNPs to transcripts, miRNA-binding sites, CpG islands, conservation, transcription factor binding sites and Tajima's D' (a measure of positive selection indicative of function (Tajima 1989)).

Selected SNPs were genotyped to test for heterozygosity in a sample of fifteen schizophrenia cases from Ireland (taken from the GASP sample). A sample of this size has 95% power to detect a SNP with a minor allele frequency greater than 10%. A SNP with a frequency lower than this has minimal power to detect true association in a sample of the size used in this study.

2.2.3.2 The SNaPshot™ genotyping method

The SNaPshot™ method (Applied Biosystems, Warrington, UK) was used to genotype non-HapMap SNPs in the fifteen cases and HapMap CEU trios. The principle of the SNaPshot™ assay is the binding of an oligonucleotide primer (15-35bp long) to a single strand of PCR-amplified DNA adjacent (5') to a target SNP followed by the polymerase-induced addition of a single complementary dideoxy nucleotide (ddNTP) base at the SNP site. The ddNTPs are fluorescently labelled (one colour for each base) and prevent addition of further bases. The result of this mini-sequencing reaction, following purification and denaturation, is a short single-stranded fluorescently-labelled DNA molecule. These products are subjected to electrophoresis and laser-assisted fluorescence detection on the Applied Biosystems ABI3100/3130 in order to resolve their genotypes. The SNaPshot™ method is thus appropriate for genotyping many SNPs in a small number of samples and can be processed in multiplex. (See Figure 2.3 for results of a multiplex SNaPshot™ reaction.)

2.2.3.3 Design and optimisation of PCR reactions for Snapshot

Primers for the PCR of target sites were designed on 200-500bp sequence surrounding the SNP of interest using Primer3 (http://primer3.sourceforge.net/). Primer sequences and conditions for PCR are described in Table 2.1. Where two uncharacterised SNPs were close enough together, these were targeted with the same PCR primers. A standard PCR contained 2.5pmol each primer (Invitrogen, UK), 100nmol dNTP (Sigma-Genosys), 1X reaction buffer, using 1 unit of *Taq* polymerase (Qiagen, UK) to amplify 20ng template DNA in a final volume of 10µl. PCR conditions were: 95°C for 30 seconds followed by 40 cycles of: 95°C for 10 seconds, 54/57°C for 10 seconds and 72°C for 45 seconds, followed by 1 minute at 72°C. All PCR reactions were conducted in a Peltier PTC-225 Tetrad thermal cycler (MJ Research, USA). Primer pairs were initially tested at two annealing temperatures (54°C or 57°C). Successful amplification was determined following standard gel electrophoresis with visualisation of ethidium bromide-stained DNA with UV light. Failed reactions were subjected to modified PCR conditions and Fail-Safe™ buffers (Boehringer Mannheim, Germany). SNaPshot™ extension primer sequences were designed from the more heterogeneous 20bp sequence either side of the target SNP. Where there was more than one target SNP per PCR product, primers for SnaPshot reactions differed by 5-10 nucleotide bases in length to allow for multiplexing of reactions. (Table 2.2 details primer sequences used for SnaPshot reactions.)

2.2.3.4 Investigative genotyping of functional SNPs

Purification of the PCR products was achieved following 30 minute incubation at 37°C with Exonuclease I (*Exo*I) and shrimp alkaline phosphatase (SAP) according to manufacturer's instructions (Amersham Biosciences, UK). The reaction volumes were: 5µl PCR product, 0.5µl SAP, 0.05µl *Exo*I, 1µl 10X reaction buffer in a final volume of 10µl. SNaPshot™ reactions contained 1.5µl purified PCR product, 1µl SNaPshot™ reaction mix, 1X reaction buffer and 3pmol extension primer in a final volume of 5µl. Reaction conditions were according to manufacturer's instructions. Products were purified by SAP and resolved on the ABI3100 with LIZ120 size standard (Applied Biosystems). Allelic discrimination was performed using GeneMapper software. Genotype results were formatted for import into the group's BC-SNPMax data storage facility (Biocomputing Platforms Ltd., Espoo, Finland) using a suite of Perl scripts (written by Drs Carlos Pinto and Colm O'Dushlaine, Neuropsychiatric Genetics, TCD, and freely-available on the group web-site,

http://www.medicine.tcd.ie/neuropsychiatric-genetics/). Those SNPs found to be polymorphic in the fifteen case samples, and therefore likely to be common, were genotyped in the HapMap CEU sample (See Figure 2.1).

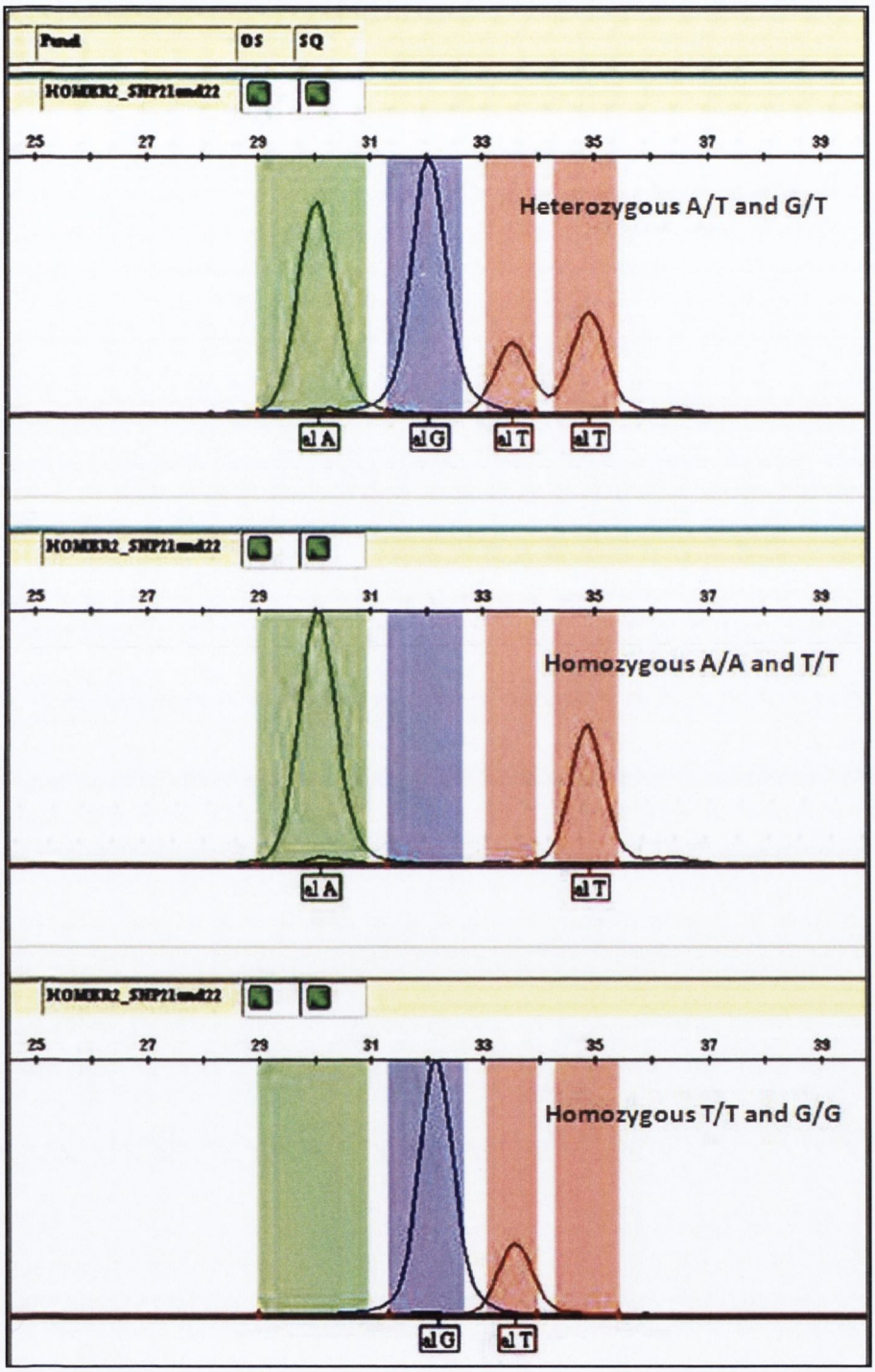


Figure 2.3: Screen-shot from the GeneMapper program showing an example of Snapshot genotyping results. In this instance, three individuals have been genotyped for two SNPs in multiplex. The top individual is heterozygous at both SNPs whereas the other two individuals have homozygous genotypes.

Table 2.1: PCR conditions used to amplify across non-HapMap SNPs for Snapshot

Gene	Amplicon name	Forward primer sequence	Reverse primer sequence	Program
ARHGEF11	intron1	CCAGTTAGGATGTGTGCAGT	CTCAGAGTCTGAACACATGAC	A54, X45
	intron21	GTTCTATCTGCTGCTGGATTC	CCTAGAAGCAGAGCAGGATGT	A57, X45
	intron39	TGCTCGAGTCCTTCTATTCCT	AGGCAACGATGATCCAAGA	A57, X45
	promoter1	CACATCCTTCTCGTGCACT	AACACAGTGGTCATCATGAAC	A57, X45
	promoter2	CTAGGATATCCGCCTGGTT	GCATCTTGTTCTGCTGCTC	A54, X45
CNDP1	D18S880	AGGCAGCTGTGTGAGGTAAC	GGGTGAGGAGAACATGCC	A57, X45
	intron8	GCACTGGCTTCTGGCACT	AGGTCTTGGAGCTGGAATCA	A57, X10+,D
GGCX	Exon11	AGAGGATTTTTGACCCTCGT	ATCTGCAATGAAGACCACCT	A57, X45
	Exon15	ATTAAATTGCTTTTCCCCAAG	AGGTTTAAAGTGGCTTTTTGG	A57, X45
GLUL	Exon3	TCTCACCTTCCACACACTTG	TCTTCTCCAGAACACCTTCC	A57, X45
	Exon5	TTCACCTCTAACCCAAGGAG	TAGATCCCTGATGGCGTATT	A57, X45
HOMER2	exon4a	CCAGAAGTGACAAGGAACAA	GAGAAATTCCAGGAGGTGAA	A57, X45
	exon4b	CCCACTGGGATTTACTTCAT	ATACCAGCCATGAGTGAGGT	A54, X45
	exon5	CAATTTCCTGAAACCCTGTC	CGTGAAACCCCATCTTCTAC	A54, X45
	exon6	GCTCGATCTCCCACTTCTT	AGCTTACTGTGGGGAGAGG	A57, X45
HTRA2	5'UTR	TGGAAGCACAGAATCTCTTGA	TTTGCTTACCACCTGTAGCAC	A54, X45
MAPT	exon1	GACTTCTGTAGGAGAGGACA	TCGGTTTGAAGCAAGTGTG	A54, X60,D
	intron9	GATCTTGGATCTTGGTGACT	ACCTCCTGTAGTTGGAGTCT	A60, X60,D
PEA15	5'UTR	CATCCTAGGAGTCTTCCAGAAC	TGGGTGACACAGTGAGACT	A57, X45,J
	exon3	TCTCCTACATTGAGCACATCT	GTGCAGCTAGTTAAAGGAGTG	A54, X20
	exon4a	CTGTCCCTGGACACATACC	ACATATAGACAAACAGCCAGA	A61, X10
	exon4b	CTAGCATAGGTATGAGCCAG	AGAAGGGTTTCCTAAGTAAGTC	A57, X45
SYNGAP1	Exon14	AGCCACAGCTACAGTGATGA	GCTGACTGTCAACTGCTGAG	A57, X45
	Exon1	CTGCCGTTGGCTCTTATT	GCTCGAGACCTGCTCATC	A57, X45
	3'UTR	GTAACCAGAGCTGACAATCT	TCAGAGTAGGACAGATCAAC	A54, X45

In the Programme column, 'A' refers to the annealing temperature and 'X' the extension time. A '+' symbol is used to indicate the addition of 1 second per cycle. The letters D and J at the end of each programme refer to the use of fail-safe buffers.

Table 2.2: Information on Snapshot assays for investigation of non-HapMap SNPs.

Gene	Amplicon	SNP ID	Snapshot primer sequence	Strand*	Alleles
ARHGEF11	Intron1	rs12137049	CCCACAGCAAGGTTCTCA	+	AT
		rs822571	CTGGGCAGACTCTTGGGG	+	CG
	Intron21	rs2210860	GGAGGTGGGGTGACCATC	-	AC
		rs12128868	CTGTTGGGCTGGTTTTCTG	+	AG
	Intron39	rs2275198	TGTGCTGGGCCTTGGG	+	AG
		rs868187	ATTTAAGGGTGGCTGGG	+	CT
		rs4570419	GCAGGGACCCAAGCCCTG	+	CT
		rs868188	ACGTCCCTGAGGGCCAGGC	+	CT
	Promoter1	rs11264607	TCCCCTACACAGTAGATCC	+	AG
	Promoter2	rs2316334	TGCCTCCACTCCTGCCT	+	CT
		rs11554485	ATAGACAAGGTACGGTTC	+	CT
CNDP1	D18S880	D18S880	n/a	+	162-172
	Intron8	rs11661606	GTTCATTAACCTAAATCCCAC	+	AG
GGCX	Exon11	rs28928872	ACCCTTTCAGCGCACATCCT	-	CG
GLUL	Exon3	rs11541591	GGGACATGTACACCTGCTTG	+	CT
		rs17852995	CATGACCACCTCAGCAAGTT	-	GT
	Exon5	rs17855559	AGACTTACCCTGGGGCCCTG	+	GT
HOMER2	Exon4a	rs11629943	GAATGATTACTTGAGGTCTC	+	AG
	Exon4b	rs8041066	ATAGCCAAAGACAAGACGCA	-	AG
	Exon5	rs1051936	TTGTCATTCTCAGACTTCAG	+	GT
		rs11541402	TCTCACGCCGGTCCAGCCAA	-	AC
	Exon6	rs28550214	CTCCCACTTCTTCACGTTGG	+	CG
HTRA2	5'UTR	rs28381985	GATCCTCTACCCAACATGT	+	A-
PEA15	5'UTR	rs12731466	CTGGCCGGTGATGTCAGA	-	AG
		rs8192600	GCTCGGGCTCCGGCTCC	+	AG
	Exon3	rs11552719	CTGGTGGTTGACTACAGAACC	+	AG
		rs17852096	GCTCATCCTCCTCAGAGAT	-	CT
	Exon4a	rs11552717	TCCAACCTTCCTCCTCTTC	-	GT
		rs11552712	CCACTCCCTTCCCCCATCCT	+	CG
		rs11552714	CAACCCACCATCTACCCACTT	+	AG
	Exon4b	rs17847310	CACCCACCTGTACTCTGG	+	AG
		rs11552711	TACTCAGACTGTTGCACTC	+	AT
		rs17847313	TTTTCCTTTGCACCGATCC	+	CG
MAPT	Exon1	rs2664007	TGGAGTCTTTGTGTCGTTG	-	AG
	Intron9	rs2471737	TCGCACACATGCCGTGC	+	AG
SYNGAP1	Exon 15	rs11756959	CCAGAGGCCAGCCCCCTCA	+	GT
	3'UTR	rs13214818	CTCCCTCCCCACCCAGAACCCA	+	GT

*Strand refers to which direction the SNP was genotyped on. Alleles refer to the genotype calls relative to the strand genotyped. Information sourced from dbSNP as not reported alleles were identified in all SNPs investigated.

2.2.4 Association study methods: Tagging SNP selection, genotyping and data analysis

2.2.4.1 Selection of tagging-SNPs

The genotypes generated in-house on the HapMap CEU samples were merged with the publicly-available HapMap data and analysed to select SNPs that tag the majority of common SNPs at the gene. The region of each candidate gene for study was set at 100kb 5' and 20kb 3' from the boundaries of the longest transcript. This region was likely to include most extra-genic *cis*-acting variants (Myers *et al* 2007).

The SNP selection strategy for case-control analysis was designed to select the minimum number of SNPs required to capture the maximum amount of common variation using LD information from the HapMap CEU sample. The genotype data generated in-house on the CEU sample was merged with the publicly-available data (release #20/phase II Jan06) using the *merge hapmap data.pl* Perl script (written by Dr Carlos Pinto) and up-loaded into HaploView v3.32 (Barrett *et al.* 2004). Tagging SNPs were selected by implementation of the Tagger algorithm with an r^2 cut-off at 0.8. SNPs were only assessed as tagging SNPs if they met the following criteria: MAF>10%, Hardy-Weinberg p-value>0.001, >75% call rate, <1 Mendelian error and minor haplotype frequency >5%. SNPs capturing only themselves were excluded unless they were putatively functional or formed part of a common (>5%) two- or three-marker haplotype (LOD > 3.0). Numerical summaries of the SNPs genotyped in the GASP case-control sample are given in Table 2.3.

2.2.4.2 Genotyping of tagging SNPs in the GASP case-control sample

The majority of the genotyping for the case-control study was outsourced to KBiosciences (Hoddesdon, UK) who use the KASPar assay. Genotyping in-house was by the TaqMan™ allelic discrimination method (Applied Biosystems). Standard reactions were set-up using the Roboseq® 4204 robotic liquid handler (MWG Biotech AG, Germany) and contained of 2.5µl TaqMan Universal PCR Master Mix, 0.125µl 40X TaqMan™ assay mix (which includes PCR primers and allele-specific probes) and 20ng DNA in a total volume of 5µl. PCR and allelic discrimination for TaqMan™ were performed on an ABI PRISM 7900HT (Applied Biosystems) according to manufacturer's instructions.

2.2.4.3 Quality control and analysis of genotyping results

Raw genotype data in text format from either KBiosciences or in-house TaqMan™ results were processed using Perl scripts and uploaded into BC-SNPmax for storage and basic analysis. The Perl script, *hapmap_checks.pl* was used to check CEU genotype results against those in HapMap online. Individual SNPs were tested for association with the phenotype (disease status) using a 2x2 contingency table to calculate a χ^2 statistic. Pairwise measures of LD (D' and r^2) and Hardy-Weinberg Equilibrium (HWE) were calculated using Haploview.

Table 2.3: SNP information on genes studied

Gene name	Co-ordinates studied	HapMap SNPs MAF>0.1	SNPs genotyped in case-control sample	Coverage (%) at r2>0.8
ARHGEF11	chr1:153,700,000-153,860,000	113	9	99
CNDP1	chr18:70,341,000-70,410,000	58	12	74
GGCX	chr2:85,688,000-85,700,000	17	4	94
GLUL1	chr1:179,084,000-179,093,000	25	3	72
HOMER2	chr15:81,315,000-81,413,000	85	13	67
HTRA2	chr2:74,736,000-74,766,000	15	3	67
MAPT	chr17:41,265,000-41,475,000	246	16	95
PEA15	chr1:156,984,000-156,999,000	9	2	100
SYNGAP1	chr6: 33,490,000-33,531,000	16	4	50*
Total	644,000 bp	584	66	50-100

*This value was calculated using HapMap Build 36, which may contain more SNPs than Build 35 that had been used to select the tagging SNPs and calculate the coverage estimations for the other genes.

2.3 Results

2.3.1 Results of candidate gene selection

2.3.1.1 Genes with evidence of linkage from Lewis *et al* (2003) study

A gene was considered here to show evidence of linkage to schizophrenia if it was located within one the nineteen highest-scoring bins from the meta-analysis of genome-wide linkage studies. Using the UCSC Genome Browser Tables function and filtering duplicate names in Excel, 3,915 genes were identified as having evidence of linkage to schizophrenia (from 23,606 genes in all 120 bins). The nineteen bins varied in size (15.4-47.7Mb), their total number of genes (85-1003) and density of genes (7-21 genes per Mb).

Table 2.4: Properties of the nineteen highest-ranked bins from the linkage meta-analysis study

Bin rank	Genomic region	Proximal marker	Distal Marker	Length (Mb)	Number of genes
1	2p12-q22.1	D2S139	D2S2254	40.09	231
2	5q23.2-q34	D5S2098	D5S422	38.31	331
3	3p25.3-p22.1	D3S3589	D3S3521	28.02	121
4	11q22.3-q24.1	D11S4161	D11S1316	19.25	161
5	6pter-p22.3	6p-ter	D6S274	16.85	96
6	2q22.1-q23.3	D2S2254	D2S2275	31.82	113
7	1p13.3-q23.3	D1S221	D1S2705	49.02	464
8	22pter-q12.3	22p-ter	D22S424	34.03	278
9	8p22-p21.1	D8S1790	D8S1809	15.14	102
10	6p22.3-p21.1	D6S274	D6S1582	26.45	430
11	20p12.3-p11	D20S115	D20S190	13.65	56
12	14pter-q13.1	14p-ter	D14S70	33.53	156
13	16p13-q12.2	D16S3103	D16S415	34.85	258
14	18q22.1-qter	D18S55	18q-ter	16.09	40
15	10pter-p14	D10S547	D10S1768	25.64	122
16	1q23.3-q31.1	D1S2705	D1S202	26.11	217
17	15q21.3-q26.1	D15S1033	D15S116	31.27	274
18	6q15-q23.2	D6S462	D6S472	41.60	166
19	17q21.33-q24.3	D17S1860	D17S1350	26.87	298

2.3.1.2 Genes fulfilling glutamatergic and synaptic search criteria

Key word searches of NCBI Entrez-Gene and UCSC Genome Browser using the search-term "glutamat*" produced 300 and 144 genes respectively. Only 114 genes were common to both databases. The total number of unique gene names fulfilling the glutamatergic search criteria was 288. This combined list was used for further candidate gene analysis and selection. Of these genes, 32 were found to be located within one of the nineteen bins showing evidence of linkage. No glutamatergic genes were identified in bins ranked 9th (8p22-p21), 11th (20p12.3-p11) and 12th (14pter-q13.1).

Key word searches of NCBI Entrez-Gene and UCSC Genome Browser for "synap*" yielded 828 and 188 genes respectively. From the proteomic study of the mouse synapse, 764 genes were identified. The total number of unique gene names fulfilling the synaptic search criteria was 1,321. All of the top nineteen bins with evidence of linkage harboured synaptic genes with the exception of the bin ranked 11th (20p12.3-p11) in which no glutamatergic genes were identified either.

The glutamatergic and synaptic gene lists were screened for genes common to both categories resulting in 112 consensus glutamatergic synapse genes. Fourteen were located within the nineteen genomic bins with evidence of linkage to schizophrenia (Figure 2.5, Table 2.5). No glutamatergic synapse genes were identified in ten of the nineteen linked bins. These fourteen genes were investigated for the number of tagging SNPs required for reasonable coverage, and for previous publications on genetic association studies with schizophrenia.

Table 2.5: Consensus glutamatergic synapse genes with evidence of linkage

Bin rank	Genomic region	Genes fulfilling both glutamate and synaptic search criteria
1	2p12-q22.1	-
2	5q23.2-q34	*GRIA1, CAMK2A*
3	3p25.3-p22.1	-
4	11q22.3-q24.1	*GRIA4, GRIK4*
5	6pter-p22.3	-
6	2q22.1-q23.3	-
7	1p13.3-q23.3	-
8	22pter-q12.3	*TUBA8*
9	8p22-p21.1	-
10	6p22.3-p21.1	*GRM4, GABBR1, SYNGAP1*
11	20p12.3-p11	-
12	14pter-q13.1	-
13	16p13-q12.2	*SIAH1*
14	18q22.1-qter	-
15	10pter-p14	*GAD2, GPR158*
16	1q23.3-q31.1	*GLUL*
17	15q21.3-q26.1	*HOMER2*
18	6q15-q23.2	*GRIK2*
19	17q21.33-q24.3	-

2.3.1.3 Identified interactors of putative schizophrenia genes

Literature searches yielded 21 genes putatively associated with schizophrenia which were subsequently investigated for interactors. These genes were *AKAP5*, *AKT1*, *CACNG2*, *COMT*, *CPLX1*, *CPLX2*, *DAAO*, *DISC1*, *DTNBP1*, *G72*, *GDA*, *GRIK3*, *KPNA3*, *KPNB3*, *MUTED*, *NRG1*, *OLIG2*, *PLXNA2*, *PRKCA*, *RGS4*, *SYN2* (Owen *et al* 2005, Harrison & Weinberger 2005, Wilson *et al* 2006). A total of 428 interactors were found from searches of NCBI Entrez-Gene reports, the 'String', functional protein association network (http://string.embl.de/) and publications. Of these 428 interacting genes, 72 were located within the nineteen intervals (See Table 2.6). *PEA15*, *MAPT* and *HTRA2* were selected for association analysis: all have published protein interactions with the putative schizophrenia gene *AKT1* as well as relevant neuronal functions. *ARHGEF11* was selected for association analysis because of its protein interaction with *DISC1* (Millar *et al* 2003) as well as fulfilling the glutamatergic search criterion.

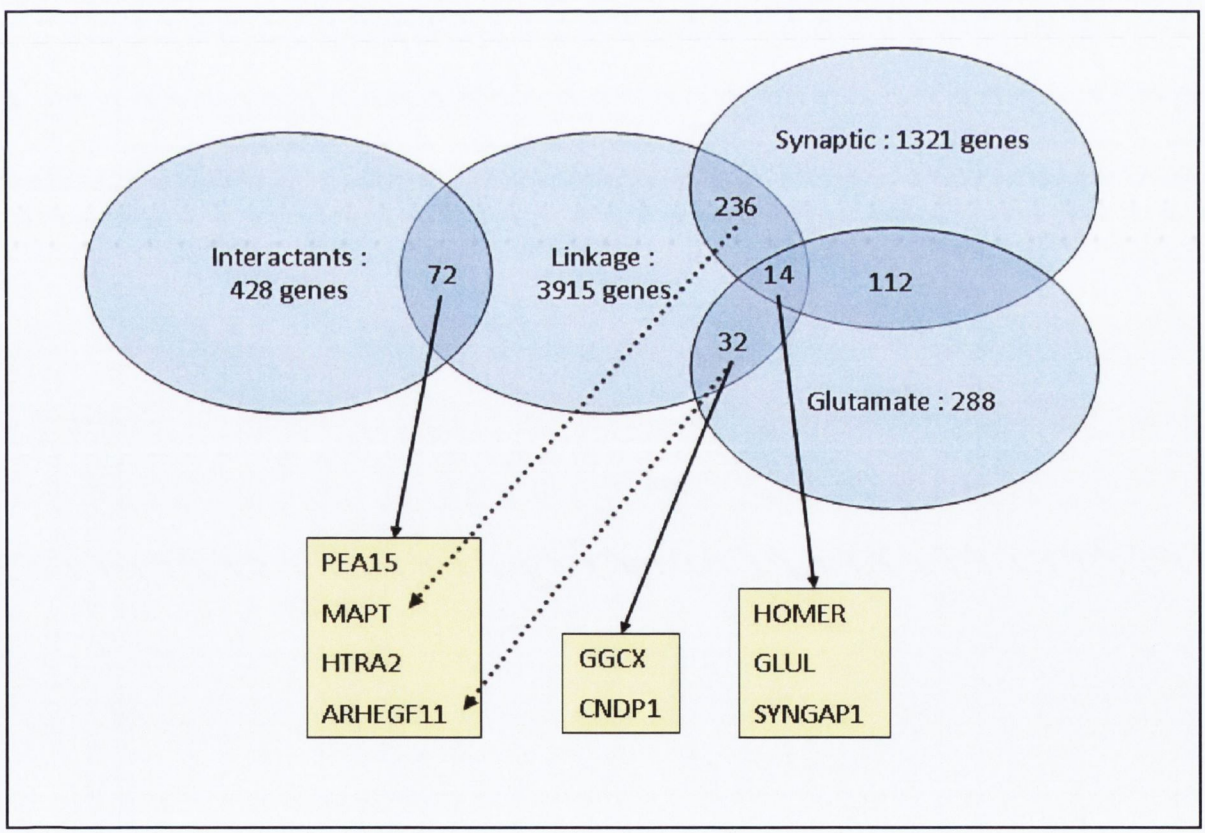


Figure 2.5: Diagram summarising candidate gene search strategy including numbers of genes found within each search criteria and numbers of genes fulfilling multiple criteria.

Table 2.6: Interactors of putative schizophrenia genes which were found within the 19 highest-scoring bins.

Bin rank	Genomic region	*AKT1*	*COMT*	*DISC1*	*DTNBP1*	*GDA*	*MUTED*	*PLXNA2*	*PRKCA*	*RGS4*	*SYN2*
1	2p12-q22.1	*UXS1, HTRA2*	-	*IMMT*	-	-	-	-	-	-	-
2	5q23.2-q34	-	-	*ANKHD1, DPYSL3, KIF3A, MATR3*	-	-	-	-	*LMNB1*	-	-
3	3p25.3-p22.1	*RAF1*	-	*SH3BP5*	*ZFYVE20*	-	-	-	*RAF1*	-	-
4	11q22.3-q24.1	-	-	-	-	-	-	-	*TRIM29*	-	*HSPA8*
5	6pter-p22.3	*HSPCB*	-	*RANBP9, TUBB2*	*MUTED*	-	-	-	-	-	-
6	2q22.1-q23.3	-	-	-	-	-	-	-	-	-	-
7	1p13.3-q23.3	*PEA15, CTMP*	-	*ARHGEF11*	*SNAPAP*	*SNAPAP*	-	-	*PEA15, SELL*	-	*NOS1AP*
8	22pter-q12.3	-	-	*TFIP11*	*DGCR6L*	-	-	-	-	-	*SYN3*
9	8p22-p21.1	-	-	*EPB49, CLU, DPYSL2*	-	-	-	-	-	-	-
10	6p22.3-p21.1	-	-	-	-	-	*DTNBP1*	-	-	-	-
11	20p12.3-p11	-	*XRN2*	*CRNKL1, XRN2*	-	-	-	-	*PLCB1*	*PLCB1*	-
12	14pter-q13.1	-	-	-	*AKAP6*	-	-	-	*TEP1*	-	-
13	16p13-q12.2	*FTS*	-	-	-	-	*BCKDK*	-	-	-	-
14	18q22.1-qter	-	-	-	-	-	-	-	*MBP*	-	-
15	10pter-p14	*MAP3K8*	-	-	-	-	-	-	*ITGB1*	-	-
16	1q23.3-q31.1	-	-	*KIFAP3*	-	-	-	-	-	-	-
17	15q21.3-q26.1	-	-	*MAP1A*	*RAB11A*	-	-	-	*AKAP13*	-	-
18	6q15-q23.2	-	-	*KCNQ5, SYNE1, C12orf182*	*KIAA0408*	-	-	*HACE1*	*GABRR2, GJA1*	-	-
19	17q21.33-q24.3	*RPS6KB1, MAPT*	-	*CDK5RAP3, PPM1E*	-	-	-	-	*PDK2*	-	-
TOTALS		10	1	24	8	1	2	1	13	1	3

2.3.1.4 Final selection of genes for association analysis

The primary gene selection criteria for association analysis were a specific role in glutamatergic synapse function and localisation to a region suggestive of linkage to schizophrenia as determined from the meta-analysis of schizophrenia linkage studies. Fourteen genes fulfilled these criteria. However, it was found that many of these genes would require an excessive number of tagging SNPs for a comprehensive association analysis and this was not practically possible (*CAMKII*, *GRIA1*, *GRIK1*, *GRIK4*, *GRIK2* or *GRM4*). Other genes fulfilling the primary selection criteria had been studied before with no evidence of association in well-powered samples (*GAD2*, Zhao *et al* 2007) or had no evidence of obvious functional relevance to schizophrenia following literature review (*GPR158*, *TUBA8* and *SIAH*). Thus, of the fourteen primary candidates, only three were brought forward for association analysis: *GLUL*, *HOMER2* and *SYNGAP1*.

Secondary candidates were selected from genes in linked regions and which were identified in either glutamate or synaptic searches or interactors of putative schizophrenia genes. Two genes were selected which fulfilled the glutamatergic search criterion without immediate evidence of synaptic expression (*CNDP1* and *GGCX*) although neuronal expression was confirmed following literature and database searches. Two genes brought forward for association analysis (*ARHGEF11* and *MAPT*) were identified in synaptic and glutamatergic searches respectively and also interacted with putative schizophrenia genes. The last two selected genes (*HTRA2* and *PEA15*) interacted with at least one schizophrenia susceptibility gene and were located in high-ranking linkage bins in which no other glutamatergic synapse genes were identified. Table 2.7 gives the full names and genomic locations of the genes selected for association analysis. A more detailed description of the function of each gene and its basis for inclusion is given below.

Table 2.7: Summary of genes selected for association analysis

Gene Symbol	Description	MIM	Location	Lewis bin rank
ARHGEF11	A Rho Guanine Nucleotide Exchange Factor 11	605708	1q21	7
CNDP1	Carnosine dipeptidase 1	609064	18q22.3	14
GGCX	Gamma-glutamyl carboxylase	137167	2p12	1
GLUL	Glutamate-ammonia ligase	138290	1q.31	16
HOMER2	Homer homolog 2 (*Drosophila*)	604799	15q25.2	17
HTRA2	Omi/HtrA serine peptidase 2	606441	2p12	1
MAPT	Microtubule-Associated Protein Tau	157140	17q21.31	19
PEA15	Phosphatase-Enriched in Astrocytes-15kDa	603434	1q23.2	7
SYNGAP1	Synaptic GTPase-activating protein 1	603384	6p21.3	10

2.1.3.5 Background on genes selected for association analysis

ARHGEF11 (A Rho GTPase Exchange Factor 11, MIM 605708) was identified as a candidate for association analysis from its interaction with DISC1 (Millar *et al* 2003). It was also fulfilled the glutamatergic search criterion. The protein was first discovered as a binding-partner of the neuronal glutamate transporter EAAT4 which facilitates clearance of excess glutamate from the synapse (Jackson *et al* 2001). This implicates *ARHGEF11* in synaptic plasticity and the prevention of glutamatergic excitotoxicity. Expression is highest in the cerebellum, hippocampus, prefrontal cortex and thalamus (Huerta *et al* 2006). The ARHGEF11 protein contributes to a neuro-developmental signalling pathway alongside plexins and semaphorins which have recently shown functional and genetic evidence for involvement in schizophrenia (Hirotani *et al* 2002, Perrot *et al* 2002, Mah *et al* 2006).

CNDP1 (Carnosine dipeptidase 1, MIM 609064) was the only glutamatergic candidate found in the bin ranked 14th in the linkage meta-analysis study and is highly expressed in the brain (Teufel *et al* 2003). Its neuronal function is to cleave homocarnosine into histidine and γ-aminobutyric acid (GABA). GABA is an inhibitory neurotransmitter which is closely linked to glutamatergic neuro-transmission and with good biochemical evidence for involvement in schizophrenia (Roberts 1972,

Lewis & Hashimoto 2007). Serum levels of CNDP1 are reduced in non-Alzheimer's dementia and in mental and developmental disorders (Cohen *et al* 1985, Balion *et al* 2007).

GGCX (γ-glutamyl carboxylase, MIM 137167) was one of two genes which met the glutamatergic search criterion located in the highest-ranking bin (at 2p12-q22.1, the other being IL1B). In neurons its primary function is in the conversion of excess glutamate to γ-carboxyglutamate in the tricarboxylic acid (TCA) cycle (Hassel 2001). In this way, GGCX activity contributes to control glutamate homeostasis at the synapse, perturbations of which can lead to excito-toxicity or under-activity.

GLUL (Glutamate-ammonia ligase, MIM 138290) was the only glutamatergic gene identified in the bin ranked 16th in the linkage meta-analysis study (1q23.3-q31.1, Lewis *et al* 2003) and was also identified in the screen for synaptic genes. The Glul protein is involved in the cycle of synaptic glutamate processing, predominantly in astrocytes (Norenburg & Martinez 1979). By converting excess synaptic glutamate to glutamine it prevents excito-toxicity and forms an important step in replenishing the glutamate supply at the pre-synapse (Laake et al. 1995).

HOMER2 (MIM 604799) met both synaptic and glutamatergic search criteria. Homer proteins are central in regulating post-synaptic signal transduction (Brakeman *et al* 1997, Shiraishi-Yamaguchi & Furuichi 2007). Expression of Homer genes (*HOMER-1*, *-2* and *-3*) is region- and age-dependent with *HOMER2* expressed mainly in the hippocampus and prefrontal cortex (Shiraishi *et al* 2004). There are several studies of Homer protein dysfunction in model organisms which show behavioural and neurological phenotypes relevant to schizophrenia (Szumlinski *et al* 2006).

HTRA2 (Omi/Htra2 serine peptidase, MIM 606441) is located in the highest-ranked linkage bin from the meta-analysis study and was identified as a candidate through its interaction with the schizophrenia susceptibility gene *AKT1*. Here, Akt-1 phosphorylates HtrA2 to inhibit its pro-apoptotic activity in response to biochemical stress including neuronal glutamate (Shalbuyeva *et al* 2006, Yang *et al* 2007). The HtrA2 protein was first identified through its interaction with Alzheimer's gene presenilin-1 (Gray *et al* 2000). Rare, non-synonymous polymorphisms in HTRA2 have been shown to reduce neuronal cell durability through mitochondrial dysfunction which has been tentatively linked to schizophrenia pathogenesis (Kung & Roberts 1999, Millar *et al* 2005, Kvajo *et al* 2008).

MAPT (Microtuble-Associated Protein Tau, MIM 157140) was identified as a candidate through its interaction with the schizophrenia susceptibility gene *AKT1*. It was also found in the initial search for synaptic genes (NCBI Entrez-Genes and Schrimpf *et al* 2004) but not for glutamatergic genes. In a mouse strain harbouring an amino acid mutation in *MAPT* expression of neuronal glutamate transporters was reduced and lead to neurological and psychological abnormalities (Dabir et al. 2006). Genetic variation in *MAPT* is already associated with Alzheimer's disease, Pick's disease, fronto-temporal dementia, cortico-basal degeneration and progressive supranuclear palsy (reviewed by Pittman et al. 2006).

PEA15 (Phosphoprotein-Enriched in Astrocytes-15KDa, MIM 603434) was identified as a candidate gene through its protein interactions with products of the putative schizophrenia susceptibility genes, *AKT1* and *PRKCA*. The recruitment of Pea-15 to either anti-apoptotic or glucose metabolising pathways is exclusively dependent on phosphorylation by Akt-1 or protein kinase C α (Trencia *et al* 2003, Condorelli *et al* 2001). The Pea-15 protein is highly expressed in the brain and controls cell fate in astrocytes, which are critical in synaptic glutamate cycling (Renault *et al* 2003, Danzinger *et al* 1995). In addition, PEA15 was identified as one of three 'nodal point proteins' which are disregulated in both degenerative and non-degenerative neurological disorders including a model of impaired synaptic transmission (Zabel et al. 2006).

SYNGAP1 (Synaptic GTPase-Activating Protein 1, MIM 603384) met both synaptic and glutamatergic search criteria. It is highly expressed in the brain at hippocampal excitatory synapses where it binds to NMDA- and AMPA- receptors regulating Ca^{2+}-based neuronal transmission (Kim et al. 1998, Chen et al. 1998) It is important for synaptic plasticity, spine and synapse development (Vasquez et al. 2004).

2.3.2 Results of Novel SNP Investigation

Predicted functional SNPs were tested for heterozygosity in a small sample of fifteen schizophrenia cases and, if found to be polymorphic, genotyped in the CEU sample for integration into the tagging-SNP selection process. A total of 36 predicted functional SNPs, and SNPs in the same PCR amplicon as functional SNPs, were investigated in this way. Due to technical reasons (PCR failure), a number of SNPs originally selected were not assessed. A summary of the number of SNPs investigated, found to be polymorphic and found to be untagged by currently-available HapMap genotype data is

presented in Table 2.8. Of the 36 SNPs investigated, 10 were found to be polymorphic and 4 were not tagged by existing HapMap SNPs. Nevertheless, this process identified predicted functional SNPs which were subsequently genotyped in the case-control sample which would have otherwise been omitted. The results of each SNP genotyped in the fifteen case samples are described in Table 2.9.

Table 2.8 Summary of results of investigation of functional, non-HapMap SNPs

Gene name	Novel SNPs investigated	Polymorphic	Untagged in HapMap
ARHGEF11	11	6	2
CNDP1	2	1	1
GGCX	1	0	0
GLUL1	3	0	0
HOMER2	4	0	0
HTRA2	1	1	0
MAPT	2	1	0
PEA15	10	1	1
SYNGAP1	2	0	0
Total	36	10	4

Table 2.9 Results of non-HapMap SNP investigation

Gene	SNP ID	Region/amplicon	Functionality	MAF in 15 cases	Genotyped in case-controls
ARHGEF11	rs2275198	Intron/exon 39	-	0.20	No
	rs868187		-	0.13	No
	rs4570419		-	0.14	Yes
	rs868188		Non-synonymous	0.27	Yes
	rs2210860	Intron 21	Conserved	0	-
	rs12128868		-	0.18	No
	rs12137049	Intron 1	-	0	-
	rs822571		Conserved	0	-
	rs11264607	5'region(1)	Conserved	0.17	No[2]
	rs2316334	5'region(2)	Conserved	0	-
	rs11554485		-	0	-
CNDP1	rs11661606	Intron 8	Conserved	0.4	Yes
	D18S880	Exon 2	Conserved	0	-
GGCX	rs28928872	Exon 11	Non-synonymous	0	-
GLUL	rs11541591	Exon 3	Non-synonymous	0	-
	rs17852995		Non-synonymous	0	-
	rs17855559	Exon 5	Non-synonymous	0	-
MAPT	rs2664007	Exon 1	Non-synonymous	0	-
	rs2471737	Intron 9	Tajima's D'	0.5	No[2]
PEA15	rs12731466	5'region	Promoter	0	-
	rs8192600			0.06	No[2]
	rs11552719	Exon 3	Non-synonymous	0	-
	rs17852096		-	0	-
	rs11552717	Exon 4a	-	0	-
	rs11552712		miRNA BS[1]	0	-
	rs11552714		-	0	-
	rs17847310	Exon 4b	-	0	-
	rs11552711		miRNA BS	0	-
	rs17847313		-	0	-
HOMER2	rs8041066	Exon 4	Non-synonymous	0	-
	rs28550214	Exon 4	Non-synonymous	0	-
	rs1051936	Exon 5	Non-synonymous	0	-
	rs11541402	Exon 5	Non-synonymous	0	-
HTRA2	rs28381985	5'UTR	Conserved	0.47	No[2]
SYNGAP1	rs11756959	Exon 15	Non-synonymous	0	-
	rs13214818	3'UTR	Conserved	0	-

1. miRNA BS microRNA binding-site. All information from the functionality column obtained through the UCSC Genome Browser, March 2004 Build.

2. Some predicted functional SNPs were found to be polymorphic but not genotyped in the case-control sample. This was because the Kaspar and TaqMan assay failed at the design, manufacture or genotyping stage.

2.3.3 Results of association analysis

Genotyping quality was determined using non-template controls and reference samples of known genotype from the CEU HapMap trios. There were very rare instances of genotype calls being generated from non-template control reactions and genotypes which did not match those provided from the HapMap web-site. These were considered insignificant. In the case-control sample genotype call rate was good with a minimum call rate of 94% per SNP.

Across the eight genes studied, 58 SNPs were selected for genotyping in the GASP case-control sample (see tables 2.7a-l for detailed results). Significant allelic associations with the phenotype were found at five SNPs in three genes: *CNDP1*, *HOMER2* and *SYNGAP1* (See Table 2.10). The associated SNP at *CNDP1* is located in intron 1 and captures a number of other SNPs, none of which were in predicted functional sites. The three associated SNPs in the *HOMER2* gene are located in intron 1 (see Figure 2.6). These three SNPs are almost completely independent from one another in terms of LD, the highest r^2 being 0.02 in the CEU and with similar values in case and control samples. The three SNPs de-mark several sites of potential functional significance including exon 2 and a highly conserved 100bp sequence in intron 1. The associated SNP at the *SYNGAP1* gene, rs4231, is located ~4Kb upstream from transcription start and is in the 3' UTR of another gene, *PHF1*.

Table 2.10 Association statistics of significantly associated SNPs in the candidate gene study

Gene name	SNP ID	Case MAF (375)	Control MAF (812)	χ^2	p-value	OR	95% C.I.	Case HWEp*	Control HWEp
CNDP1	rs11661606	0.172	0.209	4.09	0.043	0.79	0.62-0.99	0.01	0.37
HOMER2	rs7174726	0.066	0.095	4.97	0.026	0.68	0.48-0.96	0.63	0.98
	rs12913501	0.109	0.143	4.84	0.028	0.74	0.56-0.97	0.22	0.08
	rs869498	0.133	0.100	6.09	0.014	1.39	1.07-1.81	0.95	0.65
SYNGAP1	rs4231	0.202	0.153	8.24	0.004	1.39	1.10-1.75	0.92	0.45

*HWEp, Hardy-Weinberg p-value

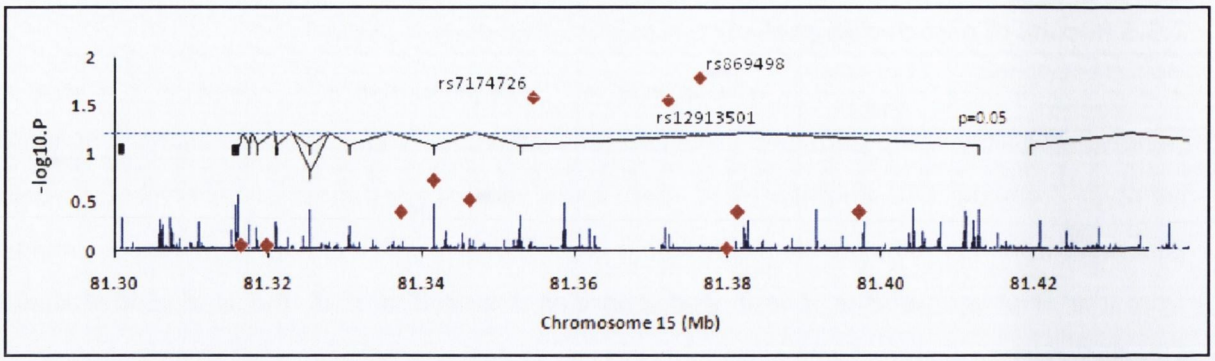


Figure 2.6: Association results at the *HOMER2* gene. Each red diamond represents a SNP which was genotyped in the GASP case-control sample. The level of each SNP on the y-axis indicates the level of association (determined by transforming the p-value by -log10.) Also shown is the intron-exon structure and conservation, in blue (both from UCSC Genome Browser, March 2006 Build).

Table 2.11a Results of association analysis of *ARHGEF11*

SNP ID	Alleles	Case MAF	Controls MAF	χ^2	P	OR	95% CI
rs12143768	AG	0.160	0.152	0.196	0.658	1.059	0.822-1.363
rs1572407	TC	0.465	0.468	0.015	0.904	0.989	0.822-1.189
rs883425	AT	0.127	0.149	1.841	0.175	0.829	0.633-1.087
rs1572411	GT	0.244	0.268	1.323	0.250	0.883	0.714-1.092
rs17412099	GA	0.169	0.155	0.644	0.422	1.107	0.864-1.419
rs866069	CT	0.479	0.505	1.223	0.269	0.901	0.750-1.083
rs703111	TA	0.235	0.206	2.243	0.134	1.183	0.949-1.474
rs12760229	TC	0.111	0.109	0.039	0.845	1.030	0.768-1.380

Table 2.11b Results of association study for *CNDP1*

SNP ID	Alleles	Case MAF	Controls MAF	χ^2	P	OR	95% CI
rs11661606	AG	0.172	0.2089	4.088	***0.043***	0.79	0.62-0.99
rs17089378	GA	0.2298	0.213	0.785	0.376	1.10	0.89-1.37
rs17238585	GA	0.0722	0.0867	1.324	0.250	0.82	0.58-1.15
rs17816861	TG	0.1792	0.1836	0.061	0.805	0.97	0.77-1.23
rs17817077	AG	0.4263	0.4139	0.300	0.584	1.05	0.88-1.26
rs7244647	TC	0.3439	0.3458	0.007	0.932	0.99	0.82-1.20
rs8087768	GT	0.2847	0.267	0.744	0.388	1.09	0.89-1.34
rs9319909	GC	0.263	0.2577	0.070	0.791	1.03	0.84-1.26
rs9967490	AG	0.3078	0.2997	0.146	0.702	1.04	0.85-1.26

Table 2.11c: Results of association study for *GGCX*

SNP ID	Alleles	Case MAF	Control MAF	χ^2	P	OR	95% CI
rs10179904	AG	0.125	0.114	0.526	0.469	1.107	0.842-1.455
rs699664	AG	0.333	0.341	0.129	0.720	0.966	0.799-1.167
rs6757263	CT	0.428	0.421	0.114	0.735	1.032	0.861-1.236

Table 2.11d: Results of association study for *GLUL*

SNP ID	Alleles	Case MAF	Control MAF	χ^2	P	OR	95% CI
rs2224994	TC	0.175	0.191	0.886	0.347	0.895	0.711-1.127
rs2251767	AG	0.247	0.250	0.025	0.873	0.984	0.801-1.207
rs1058111	CT	0.469	0.483	0.350	0.554	0.948	0.794-1.132

Table 2.11e: Results of association study for *HOMER2*

SNP ID	Alleles	Case MAF	Control MAF	χ^2	P	OR	95% CI
rs1256430	AG	0.165	0.162	0.02	0.878	1.02	0.80-1.29
rs12442119	TC	0.269	0.272	0.02	0.883	0.99	0.81-1.20
rs1256444	CT	0.462	0.443	0.71	0.398	1.08	0.90-1.29
rs1256455	TC	0.251	0.226	1.73	0.189	1.15	0.93-1.41
rs2621226	AG	0.441	0.429	0.27	0.602	1.05	0.88-1.26
rs17359494	CT	0.224	0.243	1.08	0.300	0.89	0.73-1.10
rs7174726	CT	0.066	0.095	4.97	***0.026***	0.68	0.48-0.96
rs12913501	TC	0.109	0.143	4.84	***0.028***	0.74	0.56-0.97
rs869498	TC	0.133	0.100	6.09	***0.014***	1.39	1.07-1.81
rs12148275	GA	0.486	0.484	0.01	0.918	1.01	0.85-1.21
rs12050725	AT	0.201	0.186	0.71	0.401	1.10	0.88-1.37
rs2061947	TC	0.467	0.448	0.71	0.400	1.08	0.90-1.29

Table 2.11f: Results of association study for *HTRA2*

SNP ID	Alleles	Case MAF	Control MAF	χ^2	P	OR	95% CI
rs1183739	CG	0.176	0.170	0.06	0.8	1.05	072-1.52
rs10779958	AC	0.841	0.859	0.54	0.5	1.15	0.77-1.71
rs6546909	AT	0.167	0.150	0.45	0.5	0.88	0.60-1.30

Table 2.11g: Results of association study for *PEA15*

SNP ID	Alleles	MAF cases	MAF controls	χ^2	P	OR	95% CI
rs686015	TA	0.432	0.428	0.032	0.859	1.016	0.849-1.217
rs4656252	AG	0.059	0.047	1.648	0.199	1.291	0.873-1.910

Table 2.11h Results of association study for *MAPT*

SNP ID	Alleles	Case MAF	Control MAF	χ^2	P	OR	95% CI
rs242944	GA	0.385	0.385	<0.001	0.982	0.998	0.830-1.200
rs4616318	GA	0.464	0.467	0.034	0.854	0.983	0.821-1.177
rs962885	TC	0.351	0.359	0.128	0.721	0.966	0.800-1.168
rs2055797	TC	0.208	0.216	0.200	0.655	0.951	0.763-1.185
rs8078967	TC	0.440	0.437	0.200	0.888	1.013	0.845-1.215
rs11867549	GA	0.277	0.273	0.034	0.853	1.019	0.833-1.247
rs242562	AG	0.358	0.366	0.113	0.737	0.968	0.803-1.168
rs16940758	TC	0.162	0.143	1.358	0.244	1.159	0.904-1.486
rs17651093	AG	0.207	0.217	0.290	0.590	0.941	0.755-1.174
rs2435206	TC	0.317	0.305	0.337	0.561	1.059	0.873-1.285
rs3785885	AG	0.283	0.298	0.462	0.497	0.933	0.765-1.139
rs2435212	AG	0.189	0.174	0.716	0.398	1.105	0.877-1.394
rs2471738	TC	0.243	0.243	<0.001	0.997	0.999	0.809-1.235
rs7521	AG	0.483	0.466	0.538	0.463	1.07	0.894-1.280
rs1078997	GA	0.109	0.131	2.181	0.140	0.809	0.610-1.072

Table 2.11i Results of association study for *SYNGAP1*

SNP ID	Alleles	Case MAF	Control MAF	χ^2	P	OR	95% CI
rs4231	GT	0.202	0.153	8.241	0.0041	1.39	1.10-1.75
rs2274730	CG	0.677	0.641	2.369	0.1238	0.84	0.69-1.04
rs10807124	AG	0.749	0.737	0.356	0.5509	0.94	0.76-1.15
rs3106196	CG	0.122	0.121	0.007	0.9338	1.02	0.78-1.34

2.4 Discussion

2.4.1 Review of candidate gene study design and results

The main aim of this project was to identify schizophrenia susceptibility variants using positional and functional criteria to maximise the possibility of selecting a true risk gene. All genes within published linkage peaks (using evidence from Lewis *et al* 2003) were considered for relevant function. Of fourteen genes which met the primary functional criteria (involvement in glutamatergic neurotransmission), three were brought forward to association analysis. Six genes which fulfilled secondary search criteria were also tested for association. This study identified significant associations with schizophrenia at three genes.

2.4.2 Review of candidate gene selection method

Nine genes were selected for association analysis although only three were selected from the primary criteria. This was due to practical constraints and the prioritisation of genes fulfilling secondary search criteria (interactors of schizophrenia genes and synaptic or glutamatergic function). Nevertheless, two of these three genes showed significant association with schizophrenia (*HOMER2* and *SYNGAP1*). The third gene showing association with schizophrenia, *CNDP1*, was selected on the basis of its role in glutamate metabolism and evidence of neuronal expression. One of the genes which met the primary search criteria but was not selected for further analysis due to lack of further relevant evidence was *GPR158* which has since been significantly associated with schizophrenia in four independent genome-wide association studies (ISC 2009, O'Donovan *et al* 2009, Shi *et al* 2009, Stefansson *et al* 2009).

The gene selection process was dependent on mining on-line data using key-words. Whilst a number of databases dedicated to biochemical pathways could have been used to select candidate genes, this information is also stated in NCBI Entrez-Gene and UCSC Genome Browser. Of the 388 'glutamatergic' genes identified by searching these databases only 181 have been assigned to KEGG pathways. The proteomic screen of the mouse synapse was initially considered to be an extensive source of information although following conversion of the results from mouse protein-coded format to human gene symbol only 764 genes were identified compared with the estimated 1,500 genes whose products localise to neuronal synapse (Klemmer *et al* 2009). There have been several

more proteomic surveys of the synapse published, the data of which could be useful in prioritising further schizophrenia candidates (e.g. Cheng *et al* 2006, Trinidad *et al* 2008). One of the limitations of the gene selection method used in this study was that although synaptic and glutamatergic genes may have been correctly identified, they may not have been unique to these processes. Indeed, many genes involved in synaptic glutamate signalling may also be involved in a wide range of cellular activities. An alternative method of prioritising candidate genes e.g. within linkage peaks, could be to use microarray gene expression data, especially from schizophrenia brain tissue. However, microarray results on such samples are compromised by small sample sizes, lack of reproducibility and the study of a disease end-point.

2.4.3 The association signal at *CNDP1*

The associated SNP at *CNDP1*, rs11661606 (p=0.043, OR=0.79) is located in a conserved region of intron 6 (UCSC Genome Browser, hg17, chr18:70,395,440-70,395,540). It was initially investigated in-house as genotype data was not previously available in the CEU (using Build 35.1). The Cndp1 protein is involved in neuronal glutamate homeostasis so feasibly perturbation of this pathway by rs11661606 or a variant on the haplotype could be a route of schizophrenia susceptibility. The localised conservation is tentatively indicative of a splice-enhancer binding-site (Fouser & Friesen 1986). Disruption of such features has recently been linked to Duchenne muscular dystrophy and eye-colour (Sturm *et al* 2008, Gladman & Chandler 2009). In the ISC GWAS, 28 SNPs were genotyped across *CNDP1* (compared to the fourteen genotyped in the present study) one was significant (rs9949990, p=0.025, OR 2.1). SNP rs9949990 is monomorphic in the CEU samples and thus is unlikely to be in high LD with rs11661606.

2.4.4 The association signal at *SYNGAP1*

The strongest level of association at any gene studied in the project (by lowest p-value) was found at the *SYNGAP1* gene (rs4231, p=0.0041, OR=1.39). Investigation of this gene for novel, functional SNPs and association with schizophrenia was performed by Mr Darragh Crowley, IMM/TCD and a report submitted as part of the PhD program in Molecular Medicine. The associated SNP at *SNYGAP1* is located in the 3'UTR of a neighbouring gene, *PHF1* which is a transcriptional repressor of homeotic gene expression. It is also in high LD (r^2>0.8) with rs9278021 which is predicted to cause an arginine-

to-tryptophan substitution in the *LYPLA2P1* processed pseudogene. The fact that the association signal at SYNGAP1 was located in a neighbouring gene and within a region of complex long-range LD meant that further investigation would be problematic. Thus this signal was not follow-up with any further experimental studies.

Since this work was undertaken, three genome-wide association studies of schizophrenia in large samples (>3,000 cases and controls each) have identified multiple, strong associations across the HLA/MHC region which borders the SYNGAP1 gene (ISC 2009, Shi *et al* 2009, Stefansson *et al* 2009). However, rs4231 was not significantly associated with schizophrenia in the ISC study (p=0.133, r^2=1 with proxy rs9278036).

2.4.5 The association signals at *HOMER2*

Three SNPs at the *HOMER2* gene were found to be significantly associated with schizophrenia. All were associated at a similar nominal level (p=0.014-0.028) but were not in LD (maximum r^2=0.04 in the cases). All three of the SNPs are located in intron 1. No predicted functional sites could be identified at the position of any of the associated SNPs or their proxies. Homer proteins are important modulators of glutamatergic neurotransmission via their physical interaction with post-synaptic glutamate receptors (Brakeman *et al* 1996) and so are highly relevant to the glutamate hypothesis of schizophrenia. There are three *HOMER* genes although only *HOMER2* is located within a region suggestive of linkage from the meta-analysis study at 15q21.3-q26.1. This linkage peak was primarily supported by studies on samples from Finland (Paunio *et al* 2001), Wales/Ireland (Williams *et al* 1999) and Ireland (Straub *et al* 2002b). The *HOMER2* gene was selected for further study because it harboured multiple associated SNPs, is of great functional relevance to schizophrenia and is located within a linkage peak that was identified in families samples that share a similar ancestry to the association sample used in this study.

Chapter 3

Further studies of the *HOMER2* gene

3.1 Introduction

In chapter 2, asssociation studies were performed on nine genes. Significant allelic associations with disease status were found at three genes: *CNDP1*, *HOMER2* and *SYNGAP1*. Of these three genes, *HOMER2* harboured three LD-independent association signals and therefore warranted further investigation.

3.1.1 Background on HOMER proteins

There are three *HOMER* genes in vertebrates (*HOMER1*, *-2* and *-3*), all of which encode structurally-conserved proteins characterised by an N-terminal target-binding domain and a C-terminal coiled-coil (see Shiraishi-Yamaguchi & Furuichi 2007 for review). They were first identified in a screen for synaptic 'immediate early genes' which are expressed rapidly following neuronal stimulation (Brakeman *et al* 1997). HOMER proteins are highly expressed in the brain (figure 3.1, Shiraishi *et al* 1999), constitutively expressed in glutamatergic neurons (figure 3.2) and have a central role in synaptic function (figure 3.3). At the post-synapse HOMER proteins bind metabotropic glutamate receptors to secure their position in the post-synaptic membrane and transmit stimulation to downstream effectors (Kato *et al* 1998). Alternatively they may bind adaptor proteins such as shank (Tu *et al* 1999) and drebrin (Shiraishi *et al* 2003) which are in turn bound to NDMA receptors (Naisbitt *et al* 1999) and the cytoskeleton (Ishikawa *et al* 1994). They may also bind IP3 receptors which control internal calcium ion release (Tu *et al* 1998). Via their coiled-coil domain HOMER proteins bind one another, eventually forming tetramers (figure 3.4, Hayashi *et al* 2006). In this way a HOMER protein complex allows for cross-talk between synaptic receptors and down-stream effectors (Mao *et al* 2005). In response to glutamatergic stimulation, truncated HOMER proteins are expressed which, lacking the ability to polymerise, inhibit further neuronal stimulation (Shiraishi *et al* 2003, Kammermeier & Worley 2007). Thus, dynamic splicing of HOMER proteins acts as a natural dominant-negative to prevent against excitotoxicity.

The possibility that *HOMER* family genes harbour schizophrenia susceptibility variants fits with the glutamate hypothesis of schizophrenia for a number of reasons: Firstly, the post-synaptic density where HOMER proteins localise is strongly implicated in schizophrenia pathogenesis (Ross *et al* 2006). Secondly mGluRs and NMDARs, both of which are regulated by HOMER activity, show abnormal expression in the brains of schizophrenia patients post-mortem (Mueller *et al* 2004). Thirdly, HOMER proteins transduce signals from stimulated NMDARs, which when antagonised produce a phenotype closely resembling schizophrenia (Szumlinski *et al* 2004). Additionally communication between HOMER proteins and the cytoskeleton facilitates synaptic plasticity (Inoue *et al* 2007), defects in which explain many symptoms of schizophrenia (Port & Seybold 1995, Stephan *et al* 2006). Furthermore, model organisms with altered *HOMER* expression exhibit neurological and behavioural abnormalities similar to schizophrenia. For example, mice lacking the *HOMER1* gene exhibit reduced learning, impaired prepulse inhibition, enhanced sensory reactivity, enhanced motor behaviour when stimulated by NDMA receptor antagonists (MK-801 and methamphetamine) and altered glutamate distribution in the brain (Szumlinski *et al* 2005, Jaubert *et al* 2007). Although *HOMER2* KO mice do not exhibit all of these characteristics they importantly display exacerbated response to glutamate receptor antagonists and abnormal glutamate distribution (Kalivas *et al* 2002).

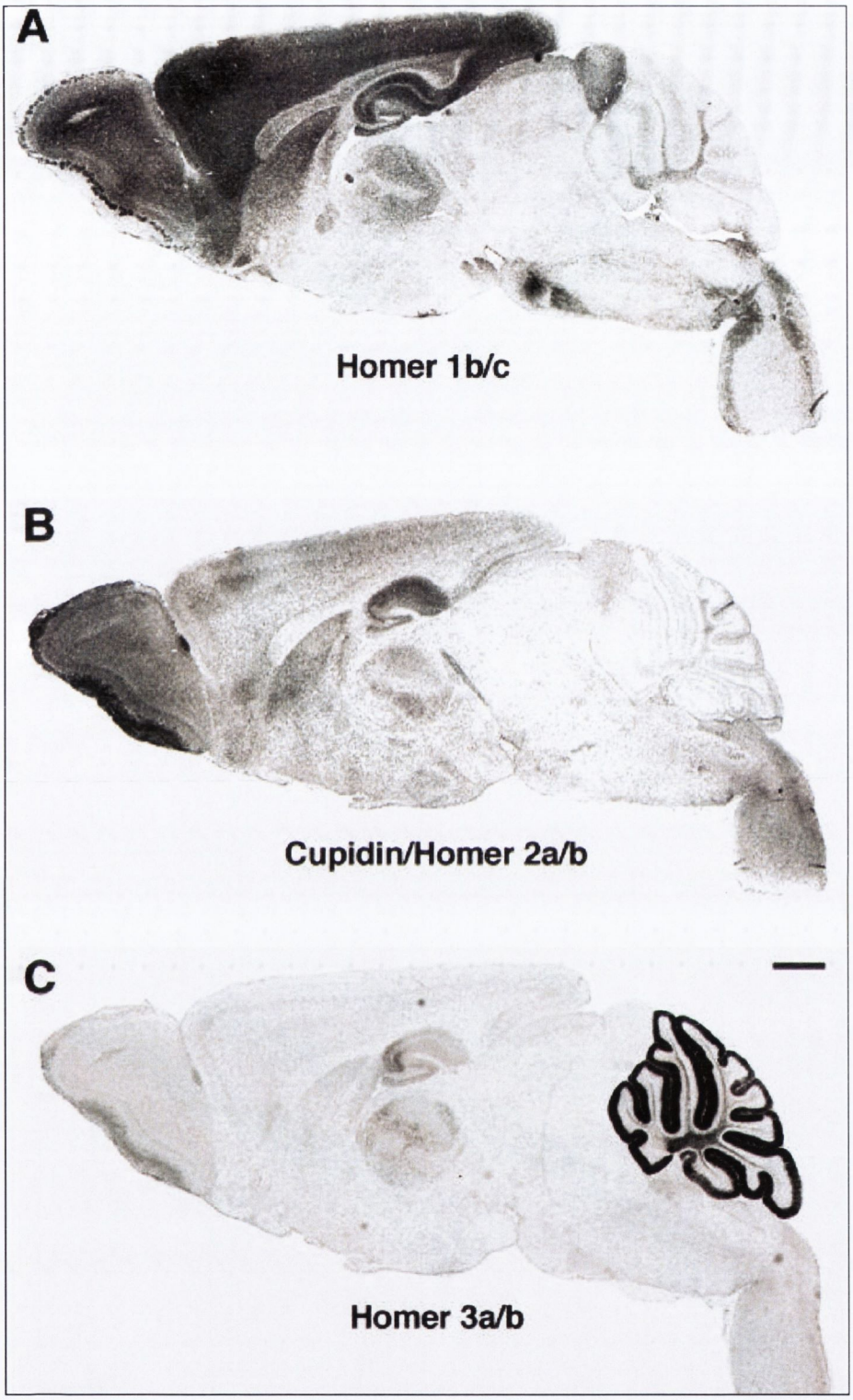


Figure 3.1: Parasagittal section through mouse brains (P14) stained with antibodies to different HOMER proteins. Scale bar 1mm (Shiraishi *et al* 2004). These images show how different HOMER family proteins localise to specific parts of the brain. HOMER1 is expressed across the cortex (A). HOMER2 is expressed pre-dominantly in the pre-frontal cortex (B). HOMER3 expression is dominant in the cerebellum (C). All HOMER family proteins are expressed in the hippocampus.

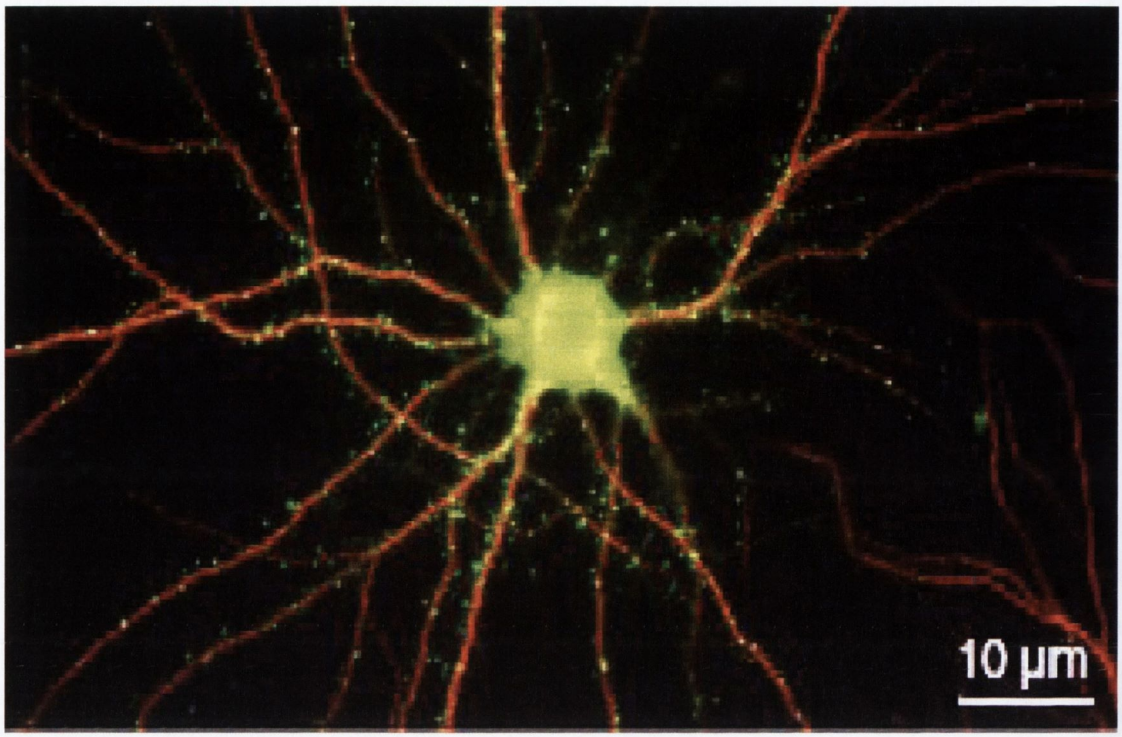


Figure 3.2: Hippocampal neuron in culture. HOMER2 proteins are expressed with 'green floresent protein' (GFP). The tetramers of HOMER proteins show as distinct 'puntata' in the dendrites (red actin stain). (Shirashi-Yamaguchi & Furuichi 2007)

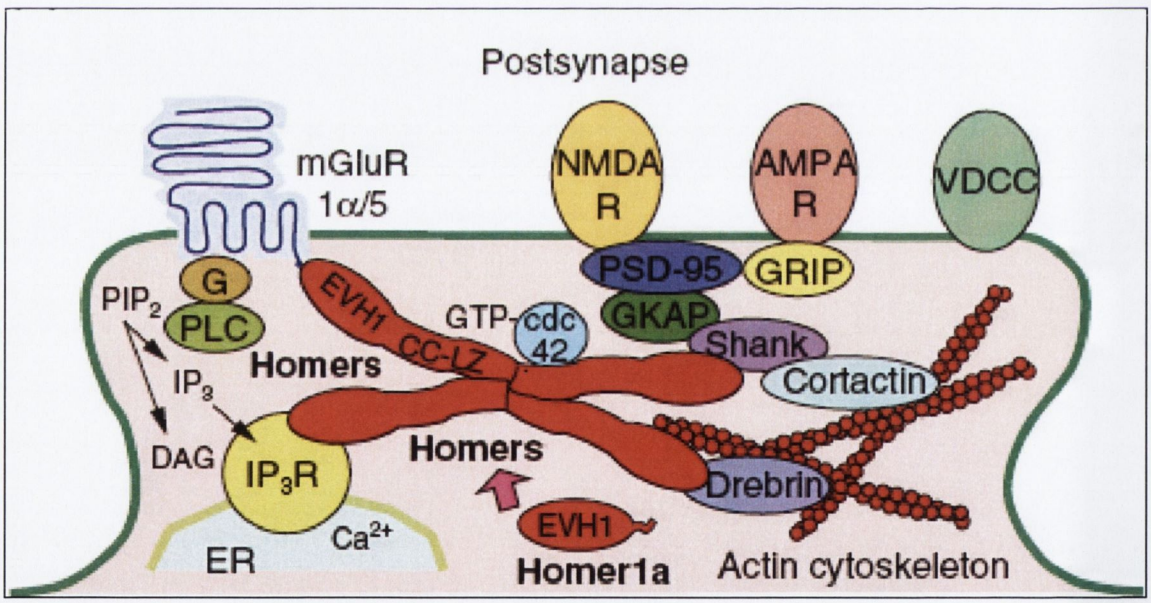


Figure 3.3: Diagram of how HOMER proteins polymerise and cross-link multiple proteins at the post-synaptic density (Shiraishi-Yamaguchi & Furuichi 2007). By these interactions HOMER proteins form a physical link from metabotropic glutamate, NMDA and AMPA receptors to calcium signalling and synapse remodelling via the actin cytoskeleton.



Figure 3.4: Structure of a HOMER protein tetramer based-on crystallographic and homology data. Scale bar, 5nm (Hayashi *et al* 2006).

3.1.2 *Cis*-acting variation and schizophrenia

Levels of gene expression are naturally-regulated and may have profound biological consequences if perturbed. Gene expression levels are controlled by the binding of proteins such as transcription factors, repressors and enhancers to specific upstream and downstream sites on the DNA (Tomkins *et al* 1969, Gillies *et al* 1983, Kleinjan *et al* 2006). Binding-sites on the DNA are sequence-specific (Pelham 1982). Thus SNPs may affect protein-binding affinity and indirectly alter gene expression. Such SNPs are thus termed *cis*-acting variants. The effect of *cis*-acting variants on distinct phenotypes in model organisms has been studied for some time (e.g. Hjorth 1979, Rabinow & Dickinson 1986). Genome-wide association studies using microrray expression data as a phenotype indicate that approximately 30% of human genes are influenced by common *cis*-acting variation (Morely *et al* 2004, Dixon *et al* 2007, Ge *et al* 2009). More recently SNPs which are associated with common diseases have been found to be *cis*-acting (Knight 2005, Manolio *et al* 2008 for reviews). For example, common variants associated with childhood asthma (Moffatt *et al* 2007) were subsequently associated with expression levels of the adjacent ORMDL3 gene (Dixon *et al* 2009). This approach also identifies inherited, rather than acquired alterations in gene expression (Bray 2008). The most effective method of interrogating a SNP for *cis*-acting approach is quantification of 'allelic expression imbalance' in which differential expression of transcribed SNPs in a single heterozygous individual is indicative of a cis-acting haplotype and importantly controls for other (*trans* and environmental) influences (Heighway *et al* 2005). Alternatively the expression level of a particular gene may be correlated with genotype through an appropriate statistical method. One aim of the present study is to take putative susceptibility variants at the *HOMER2* gene forward for analysis as potential *cis*-acting variants.

3.1.3 Endophenotype studies of schizophrenia

The term 'endophenotype' was first used in schizophrenia to describe an "internal phenotype discoverable by biochemical test or microscopic examination" (Gottesman & Sheilds 1972). This definition has since been expanded upon and is now defined as cognitive marker which displays a high heritability and is independent of the onset of schizophrenia (Gottesman & Gould 2003). Examples of schizophrenia endophenotypes include deficits in working memory, pre-pulse inhibition and sensory gating. In a similar way to gene expression, schizophrenia endophenotypes are likely to have a less complex genetic contribution than the disease itself although share susceptibility variants. In the present study, SNPs which showed good evidence for association with schizophrenia were tested for correlation with neuropsychiatric variables with the aim of identifying which endophenotype of schizophrenia the associated variant might be contributing to. Extensive neuropsychological data representing distinct processes in the brain are available in-house on a large proportion of the group's Irish schizophrenia sample collection. They have already been used to show an association between a specific genetic variant at the *NOS1* gene and working memory and verbal intelligence (Donohoe *et al* 2009).

3.1.4 Study Aims

The aim of this chapter was firstly to identify common variation at the *HOMER2* gene that was associated with schizophrenia and secondly to determine if this variant(s) also had a functional effect either on gene expression or on a schizophrenia endophenotype.

The identification of a potential causative variation was undertaken by novel SNP detection in functional domains of the *HOMER2* gene followed by extended genotyping in the GASP sample and replication analyses in both a second Irish schizophrenia sample and in the International Schizophrenia Consortium sample. The International Schizophrenia Consortium sample consists of schizophrenia and control samples of European origin, from seven different research groups including the Neuropsychiatric Genetics group at Trinity College Dublin. All of these samples have been genotyped on the Affymetrix 5.0 or 6.0 SNP arrays (ISC 2009).

The effect of associated alleles was investigated on a functional basis for two reasons. Firstly, it provides independent and additional evidence other than statistical association with the phenotype that the variant has an effect on gene function and possibly on neuronal function. Secondly, an associated variant may contribute toward a specific and quantifiable endophenotype of schizophrenia such as changes in neuropsychological indices. Endophenotypes of schizophrenia are often heritable in the general population and therefore may have a more tractable genetic component than schizophrenia itself. They also inform as to what biological factors (whether caused by genetics are not) may pre-dispose an individual to schizophrenia

3.2 Methods

3.2.1 Defining the region of strongest association

Variation at the *HOMER2* gene is likely to exert susceptibility to schizophrenia by disrupting gene function. However, none of the associated SNPs had obvious functional effects, nor did they capture any putatively functional SNPs by LD. The three SNPs at which significant association with schizophrenia was identified were all located within a 15Kb section of intron 1 and were not in high LD with each other (GASP sample case and controls, max r^2<0.02). The HapMap SNPs captured by the three associated SNPs (r^2>0.8) spanned 36,179bp from intron 1 across exon 2 and into intron 2 of the gene (see table 3. 1 and figure 3.6). This implicated a specific region of the *HOMER2* gene in schizophrenia susceptibility.

Table 3.1 SNPs 'tagged' by schizophrenia-associated SNPs

SNP	Tagging SNP	r^2 (CEU)	Position (Chr15)*	Region
rs3784366	rs7174726	1.00	81,351,052	Intron 2
rs7174726	rs7174726	-	81,354,763	Intron 1
rs17158168	rs7174726	1.00	81,356,715	Intron 1
rs12913501	rs12913501	-	81,372,191	Intron 1
rs869498	rs869498	-	81,376,406	Intron 1
rs17294476	rs12913501	0.81	81,377,930	Intron 1
rs17158157	rs869498	1.00	81,381,221	Intron 1
rs17158155	rs869498	0.85	81,387,231	Intron 1

*NCBI Build 36/hg18

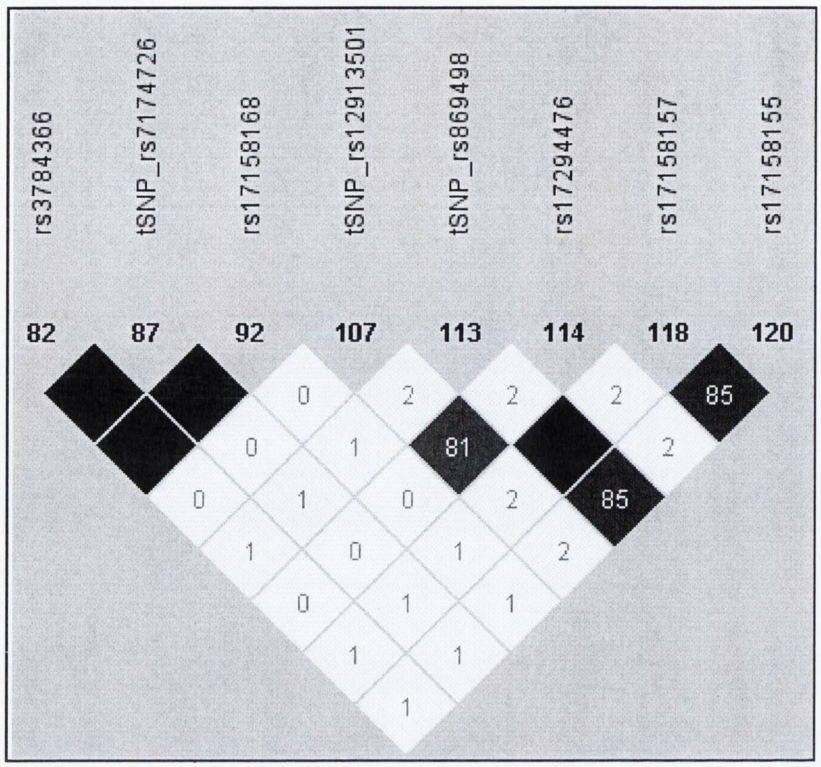


Figure 3.6: Linkage disequilibrium (r^2) relationships between schizophrenia-associated tagging SNPs (tSNPs) and captured SNPs at the *HOMER2* gene based-on HapMap CEU genotypes. Image generated using Haploview.

3.2.2 Identification of functional sites within the associated region

Investigation of this region *in silico* identified four sites of potential functional importance (see table 3.2 and figure 3.8). This was performed by visual inspection of the functional tracks on the UCSC Genome Browser (Kent *et al* 2002) and by using the Cis-ter (*Cis*-element cluster finder) on-line database which scans for recognition sites of common transcription factors (Frith *et al* 2001). The feature most strongly predicted by Cis-ter was a Sp1 transcription factor binding site (ACCTCCGCCTCCC, score=0.96 out of a maximum of 1, chr15: 81366270-81366283, Build 35/hg17). The next best prediction was an oestrogen-response element binding-site (score=0.92, 114 bp 5' of the Sp1 site) which are known to co-occur with Sp1 sites (Porter *et al* 1997).

Within the four identified functional regions, six SNPs were reported in dbSNP for which genotype data was not available in HapMap. Thus, they were potentially un-tagged and not effectively captured in the initial association study. Rather than genotype these SNPs alone e.g. by SnapShot, it was decided to conduct a more comprehensive mutational screen of each of the four

functional sites by direct sequencing. This would allow for the discovery of previously unreported genetic variants.

Table 3.2 Functional motifs in the associated region of the *HOMER2* gene

Feature	Co-ordinates (Build 35.1, hg17)	Length (bp)	Comments
Sp1 binding-site	81,366,270-81,366,283	13	Reduced expession of Sp1 in schizophrenia cortex[1]
DNaseHI site[1]	81,363,448-81,363,661	213	Feature of active *cis*-regulation[2]
High conservation	81,357,945-81,358,624	679	May affect splicing or transcription[3]
Exon 2	81,352,440-81,352,597	157	EVH1 domain of HOMER2

References: [1]Ben-Shachar & Karry 2007, [2]Crawford *et al* 2006, [3]Cech 1988

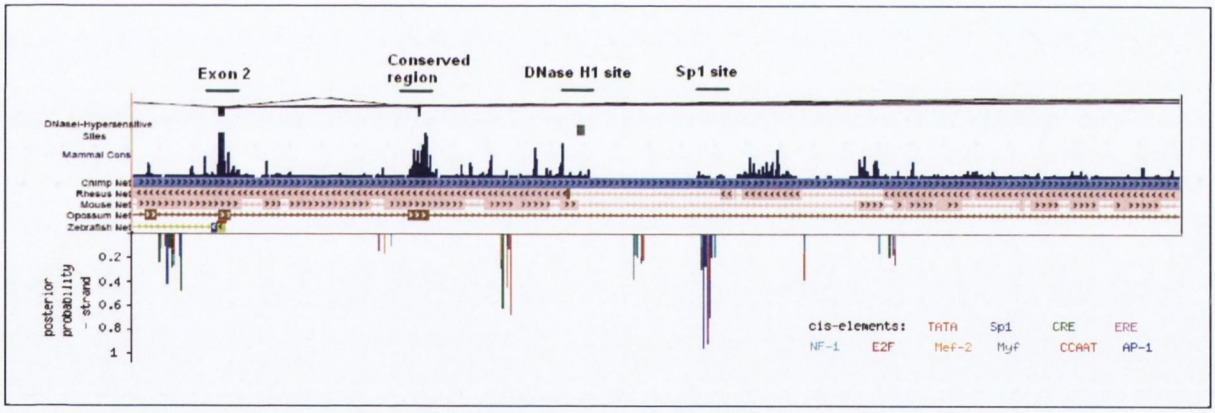


Figure 3.8: Putative functional sites within the region of association at the *HOMER2* gene. Image taken from the UCSC Genome Browser and result from the Cis-ter scan. Note that *HOMER2* is transcribed on the reverse DNA strand. Note that there is evidence from a single mRNA clone that an additional exon is present at the conserved region.

3.2.3 Polymorphism detection in functional sites

These four predicted function sites in the region of association were investigated for novel polymorphisms by dideoxynucleotide sequencing in fifteen randomly-selected schizophrenia cases (fifteen individuals were studied because this gives a 95% power to detect an allele with a frequency of 0.1). Primer design and PCR optimization were performed as described in Chapter 2. Primer sequences and PCR conditions for these reactions are presented in Table 3.3. The PCR products were purified using ExoSAP as previously described. Sequencing reactions were performed according to manufacturer's instructions. Each reaction contained 2.5ul purified PCR product, 50pmol forward or reverse PCR primer and 2ul BigDye 3.1 (Applied Biosystems, Warrington, UK) and were purified using Dye Terminator removal plates (Abgene, Epsom, UK). The sequence was resolved using an ABI3130xl machine (Applied Biosytems) with automated base-calling. Trace files were aligned by implementation of the ClustalX algorithm on the BioEdit sequence manipulation platform which also highlights sequence differences between individuals. Electropherograms were also inspected visually to reduce the potential that the base-calling algorithm would miss a heterozygous polymorphism.

Table 3. 3. Primer sequences and PCR conditions for investigative sequencing of *HOMER2*

Feature	Forward primer	Reverse primer	Conditions[1]
Sp1 binding-site	CTGCAAACTCTATGGCAGCC	TTATAACTTTCCTTGCTCTTAA	A55, X60, E
DnaseHI sensitive site	TTCCAGTTTCCAGTTCAGCA	AGTCTTCAGCAATTCCCCAA	A60, X60+
Conserved site	CTGACTTTGAAAGAGTCTTCTG	TGGAGCTAAACATTGCAAGT	A55, X60
Exon 2	ACGAACCCTTCTGGAAAACA	AATCTTTTGGGGCAGTAGGTG	A60, X60+

1. An 'A' refers to the annealing temperature and 'X' the extension time. A '+' symbol is used to indicate the addition of 1 second per cycle. The letter 'E' refers to the use of a 'Fail-Safe™' buffer (Boerhinger Manhaime, Germany).

3.2.4 Novel SNP genotyping in the CEU sample

Those SNPs that were identified as being polymorphic from re-sequencing of functional sites in fifteen schizophrenia cases were genotyped in the CEU sample unless genotype data was already available from HapMap. This genotyping was performed using the SnaPshot method and the results integrated into the publicly-available data as described in Chapter 2.

3.2.5 Additional SNP genotyping of HOMER2 in the GASP case-control sample

3.2.5.1 Selection of additional tagging-SNPs

The initial association analysis of the *HOMER2* gene described in Chapter 2 used tagging SNPs to cover 67% of the common variation. Thus, although the association results implicated a region from intron 1 to intron 2, 33% of common variants had not yet been investigated. The HapMap CEU data across the *HOMER2* gene (100Kb upstream and 20Kb downstream) was supplemented with genotypes generated in-house which had been discovered by sequencing of predicted functional regions. Additional tagging SNPs were selected this combined genotype data-set using the Tagger function in Haploview. In this process, SNPs which had been previously genotyped in the GASP sample (directly or by proxy) were omitted using the 'Force exclude' option.

3.2.5.2 Genotyping of tagging SNPs

Genotyping of additional tagging SNPs in the GASP case-control sample was primarily out-sourced to Kbiosciences although some SNPs were genotyped in-house using TaqMan® technology. The TaqMan® genotyping method is depended on allele-specific fluorophore release (FAM or VIC) from a complementary probe during a PCR reaction. Oligonucleotides and probes specific to each SNP were supplied as a single mixture which was either readily-available or designed by the manufacturers from a sequence indicating the SNP position and allele. Each reaction consisted of 20ng genomic DNA, 0.125µl assay mix (containing probes and primers) and 2.5µl Universal PCR master-mix (containing Taq polymerase, dNTPs and reaction buffer, all Applied Biosystems, Warrington, UK) made-up to a final volume of 6ul using double-processed tissue-culture grade water (Sigma). A robotic liquid-handler was used to facilitate reaction set-up in 384-well plates (RoboSeq®4204SE,

MWG Biotech AG, Ebersberg, Germany). A number of quality control measures were implemented at this stage to minimise the risk of systemic errors biasing the overall outcome. These were: positioning case and control DNA samples randomly across each plate, use of non-template (negative) controls, use of positive controls with known genotype (HapMap CEU samples) and duplicate controls. Reactions were performed in an ABI PRISM® 7900HT Sequence Detection System (Applied Biosystems) starting with 15 minutes at 95°C (for polymerase activation) followed by 40 cycles of 95°C for 15 seconds and 65°C for 60 seconds. Genomic DNA for the HapMap CEU sample was obtained from the Coriell Institute for Medical Research, USA. Following amplification, individual genotypes were determined automatically based on the wavelength of fluorescent signal emitted from each reaction. Reactions with low intensities or ambiguous genotypes are omitted. Failed reactions were repeated.

3.2.6 Replication analyses of HOMER2

3.2.6.1 Replication analysis in the 'Resource for Psychiatric Genetics in Ireland' (RPGI) sample

The RPGI sample was used in this study for primary replication analysis. The sample consisted of 408 cases and 840 controls that were independent of the GASP sample described in chapter 2. Patients were clinically stable with a diagnosis of DSM-IV schizophrenia (n=338) or schizoaffective disorder (n=70), gave written informed consent and were recruited from five sites across the Republic of Ireland and Northern Ireland using the same inclusion/exclusion criteria as for GASP sample. The RPGI case sample was 66.3% male. The control sample, which was 29.6% male, was collected by the Irish Blood Transfusion Service and was made available for this study by the Trinity College Dublin Biobank. Participating individuals in this control collection gave written informed consent and met the same ethnicity criteria as cases. These controls were not specifically screened for psychiatric illness but were not taking medications, were not financially remunerated and unlikely to be over-represented for schizophrenia.

3.2.6.2 Replication analysis in the International Schizophrenia Consortium (ISC) sample

This sample was used for secondary replication analysis. The International Schizophrenia Consortium (ISC) sample consisted of 3,322 cases of DSM-IV schizophrenia and 3,587 controls from five European countries (ISC 2009). Individuals were genotyped using the Affymetrix Genome-Wide Human SNP array, version 5.0 or 6.0 (Affymetrix, Santa Clara, California, USA). Genotype and phenotype information files were obtained from the ISC. The SNP of interest at *HOMER2* was only genotyped by the Affymetrix 6.0 array. Thus, the ISC sample used in this study consisted of 1,287 cases and 1,128 controls from Sweden, Bulgaria and Scotland. A Cochran-Mantel-Haenszel test was used to test for association. This method was used in the combined analysis of the ISC data and compensates for heterogeneous samples. It works by generating clusters of genetically-similar individuals regardless of origin. The analysis was performed using PLINK on the combined data-set with the options to only test for association at the specified SNP and to exclude a specified list of Irish samples which were also part of the GASP sample.

3.2.7 Investigation of HOMER2 expression

3.2.7.1 Quantitative PCR of *HOMER2* mRNA

Lymphoblastoid cell lines from the CEU samples (sourced from Corriell Institute for Medical Research, New Jersey, USA) were grown to log-linear phase and total RNA extracted using TRIzol® according to manufacturer's instructions (Invitrogen, Carlsbad, California, USA). Generation of cDNA by reverse transcription was performed on 1µg RNA in a final volume of 20µl according to manufacturer's instructions (Qiagen, UK). Levels of *HOMER2* mRNA were measured by quantitative PCR using an ABI PRISM® 7900HT Sequence Detection System. Pre-designed gene expression assays targeting exons 1-2 (Assay ID Hs01561580_g1) and 8-9 (Assay ID Hs01561579_mH) of the *HOMER2* gene were used (Applied Biosystems). Exons 1-2 are common to all *HOMER2* isoforms whereas exons 8-9 are present only in longer isoforms. Quantitative PCR reactions were performed as standard, according to manufacturer's instructions. Each reaction was performed in triplicate and the average Ct used to calculate the ΔΔCt relative to GAPDH gene expression (Assay ID Hs99999905_m1). Experimental measurements with a 2% variance from the mean of the triplicate samples were repeated or omitted. Isoform imbalance was quantified by calculating the ΔΔCt for exons 1-2 compared to the ct for exons 8-9 rather than GAPDH.

3.2.7.2 Comparison of *HOMER2* expression with schizophrenia-associated SNPs

The effect of genotype on expression levels of *HOMER2* was investigated only in SNPs which were significantly associated with schizophrenia following replication. Expression data for *HOMER2* was obtained in-house by quantitative PCR (described above) and from expression array data on the CEU cell lines publicly-available from the Genevar Project web-site (Illumina WG-6 expression array, probe ID GI_4758547-S, http://www.sanger.ac.uk/humgen/genevar/). Data was grouped by genotype. In instances where a SNP had a low MAF, individuals homozygous for the minor allele were grouped with the heterozygotes. A normal distribution with equal variance within each group was assumed. As an alternative data analysis method *HOMER2* expression levels were subjected to quantitative trait analysis using PLINK.

3.2.8 Neuropsychology analysis methods

Detailed neuropsychological assessment was obtained on a subset of Irish cases (n=335) who satisfied the following additional criteria (based on interview and chart review) of being aged between 18 and 60 years, showed no history of head injury, had no substance abuse in the preceding six months, were outpatients and clinically stable in the view of their treating team. Current general cognitive function was assessed using the vocabulary and block design subtests from the Wechsler Adult Intelligence Test, 3rd edition (Wechsler 1999), and premorbid IQ was assessed using the Wechsler Test of Adult Reading (Wechsler 2001). Verbal episodic and working memory was assessed using the Logical Memory and Letter Number Sequencing subtests from the Wechsler Memory Scales, 3rd edition (Wechsler 1998). Spatial memory recall and working memory was assessed using the Paired Associate Learning task (PAL) and the Spatial Working Memory task (SWM) from the Cambridge Automated Test Battery (CANTAB) (Cambridge Cognition 2003). Attentional control was assessed using the distractibility version of the Continuous Performance Test (CPT) and the CANTAB Intra-Extra Dimensional Set Shifting task (IED) (both Cambridge Cognition 2003).

3.3 Results

3.3.1 Novel SNPs in the associated interval of *HOMER2*

3.3.1.1 Discovery of novel SNPs

Four predicted or known functional sites were investigated for novel SNPs by direct nucleotide sequencing in fifteen randomly-selected schizophrenia cases. None of these sites contained SNPs for which genotype data in the CEU was available from HapMap. Seven SNPs were identified as being polymorphic including four which were not reported in dbSNP. Four previously unreported SNPs were also identified (table 3.4). Two of these novel SNPs, herein termed 'A' (figure 3.9a) and 'B' (figure 3.9b), were found close to the Sp1 TFBS. The other two, 'C' (figure 3.9d) and 'D' (figure 3.9e) were found in non-coding sequence either side of exon 2. We also confirmed three SNPs for which frequency data were not previously available: rs7183030, rs7168760 (both figure 3.9c) and rs880289 (figure 3.9f).

Table 3.4: Information of SNPs identified by re-sequencing

Region	SNP identifier	Alleles[1]	MAF (15 SZ)[2]	MAF (CEU)	Position[3]
Sp1 binding-site	Novel A	T/C	0.04	0.15	81,366,190
	Novel B	A/T	0.05	0.01	81,366,498
Conserved site	rs7183030	C/A	0.40	0.38	81,358,031
	rs7168760	T/A	0.40	0.38	81,358,032
Exon 2	Novel C	T/A	0.03	0.005	81,352,610
	Novel D	G/A	0.33	0.13	81,352,332
	rs880289	T/C	0.33	0.46	81,352,146

1. The minor allele is stated first.

2. This is based-on the fifteen schizophrenia samples in which the SNP was initially detected.

3. Chromosome 15, Build 35.1/hg17)

3.3.1.2 Functional investigation *in silico* of novel SNPs

All SNPs discovered or confirmed as being polymorphic were located close to functional or predicted functional motifs. None of them were located within any of the motifs themselves (including exon 2) nor were any at highly conserved sites. Novel SNP 'B' is located 229bp 3' of the predicted Sp1 transcription factor binding-site in intron 1 but with no alteration to the strength of the prediction in the presence of the rare allele. The rare allele at novel SNP 'B' is weakly predicted to introduce a Mef-2 transcription factor binding site (using Cis-ter, score=0.13) and thus is unlikely to be functional. Thus, whilst novel and rare SNPs were identified at predicted functional sites, no direct effect on the function of HOMER2 could be inferred.

3.3.1.2 LD relationships of novel SNPs

Seven SNPs which were identified by re-sequencing the four predicted functional sites at the *HOMER2* gene in fifteen schizophrenia cases. These seven SNPs were genotyped in the HapMap CEU sample using SNaPshot™ and the results integrated into the existing HapMap data. Two novel SNPs, 'A' and 'C' were found to be poorly captured by the initial tagging SNPs. The remainder were in high LD with initial tagging SNPS, including novel SNP 'B' (located close to the Sp1 binding site) which was in high LD with an associated SNP (rs12913501, r^2=0.87 in HapMap CEU data). In order to further investigate these novel SNPs and achieve more complete coverage of the gene, three novel SNPs and an additional ten tagging SNPs were genotyped in the GASP sample as part of a secondary, more detailed scan of *HOMER2*. This increased the coverage of common variants (MAF>0.1) in the study region from 67% to 84% at r^2>0.8.

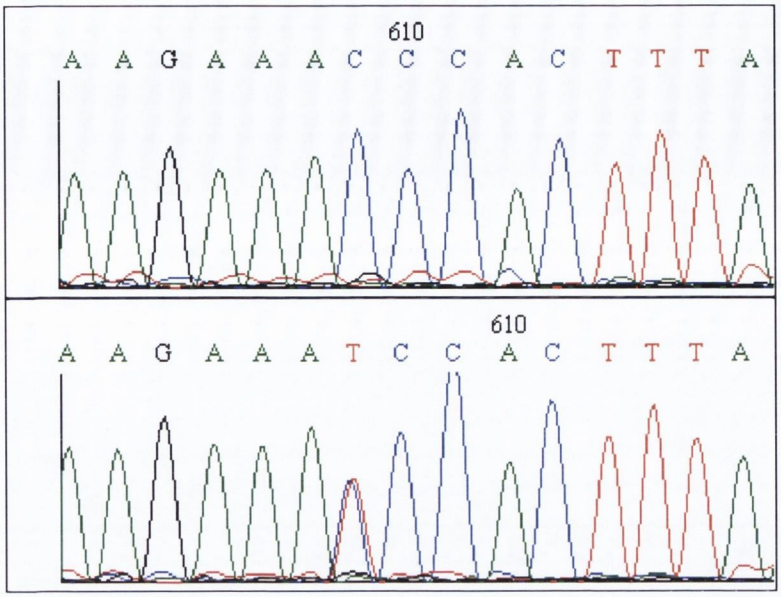


Figure 3.9a: Electropherograms showing novel SNP 'A' identified close to the predicted Sp1 binding-site in intron 1 of *HOMER2*. Top image C/C common homozygote. Bottom image T/C, rare heterozygote. No individuals homozygous for the rare allele were sequenced.

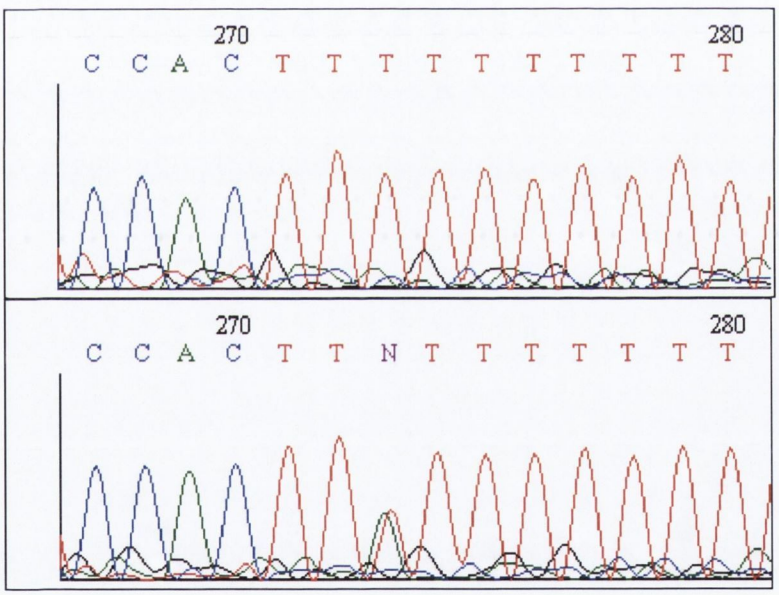


Figure 3.9b: Electropherograms showing novel SNP 'B' identified close to the predicted Sp1 binding-site. Top image T/T homozygote (common alleles), bottom image rare A/T heterozygote. No individuals homozygous for the rare allele were sequenced.

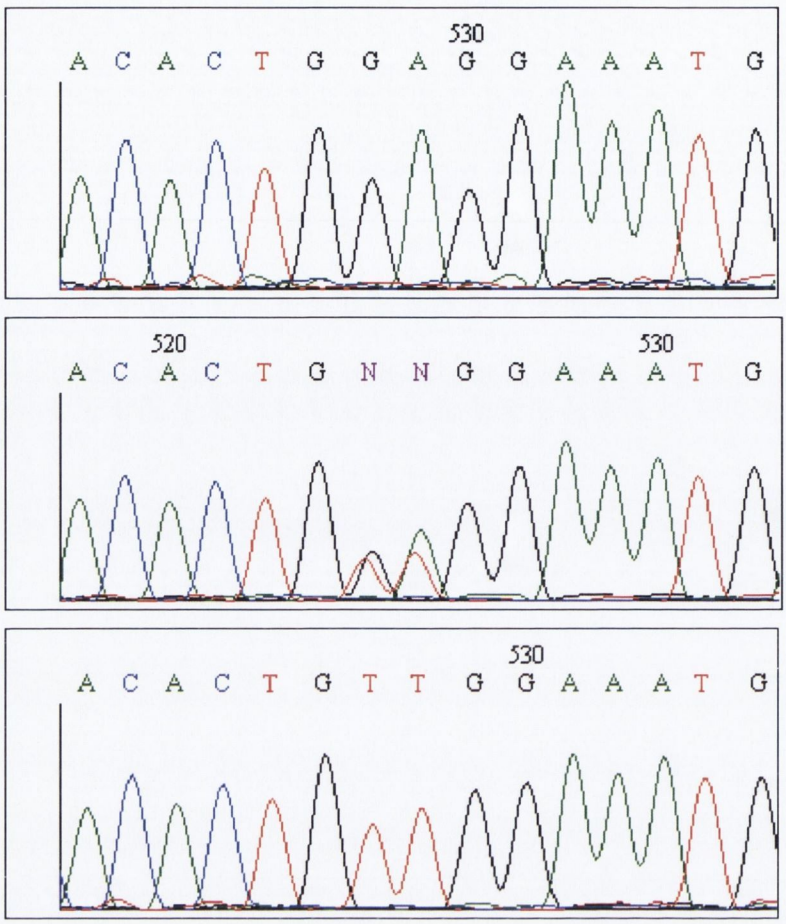


Figure 3.9c: Electropherograms from re-sequencing of highly conserved region in intron 1 of *HOMER2*. This confirms the heterogeneity of two adjacent SNPs, rs7168760 and rs7183030. Top image: GA homozygous, middle image: GA/TT heterozygous, bottom image TT homozygous. These SNPs were genotyped in the CEU sample and found to be in complete LD.

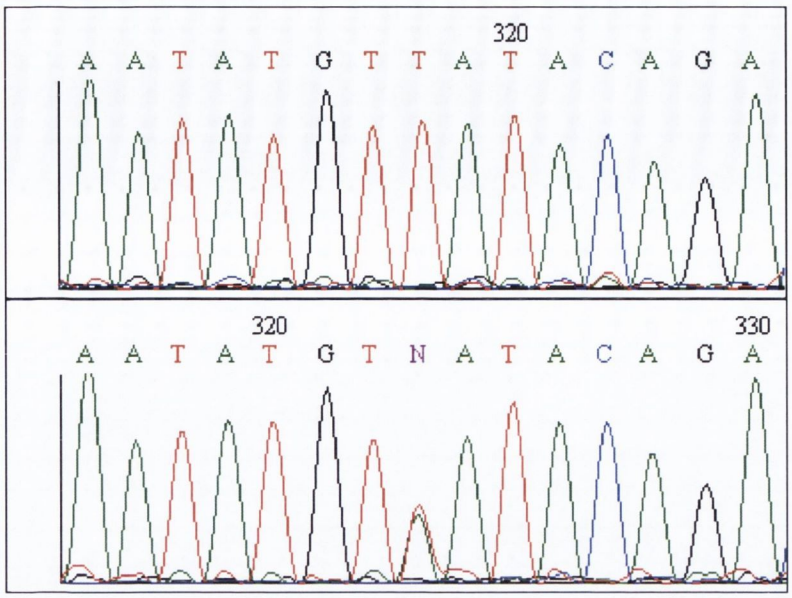


Figure 3.9d: Electropherograms showing novel SNP 'C' identified in intron 2 of *HOMER2*. Top image TT common homozygote. Bottom image C/T heterozygote. No individuals homozygous for the rare allele were sequenced.

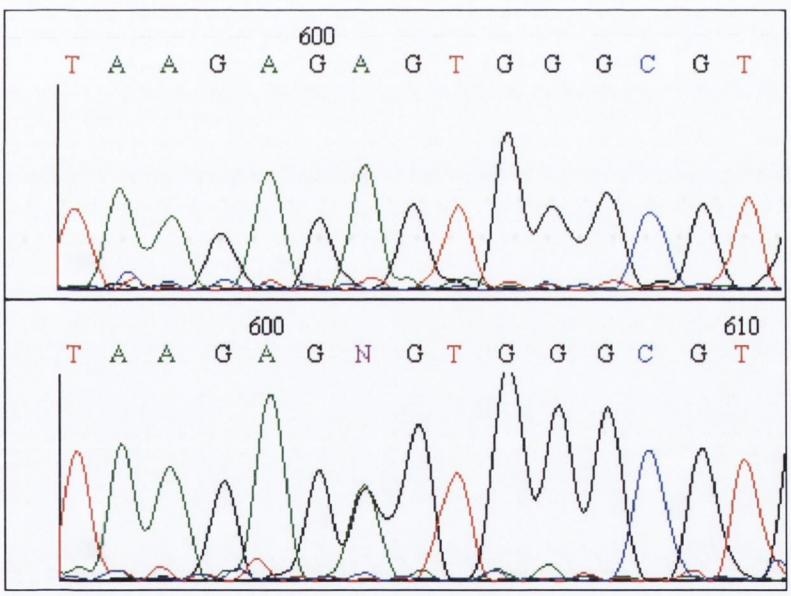


Figure 3.9e: Electropherograms showing novel SNP 'D' identified in intron 1 of *HOMER2*. Top image A/A, common homozygote. Bottom image G/A rare heterozygote. No individuals homozygous for the rare allele were sequenced.

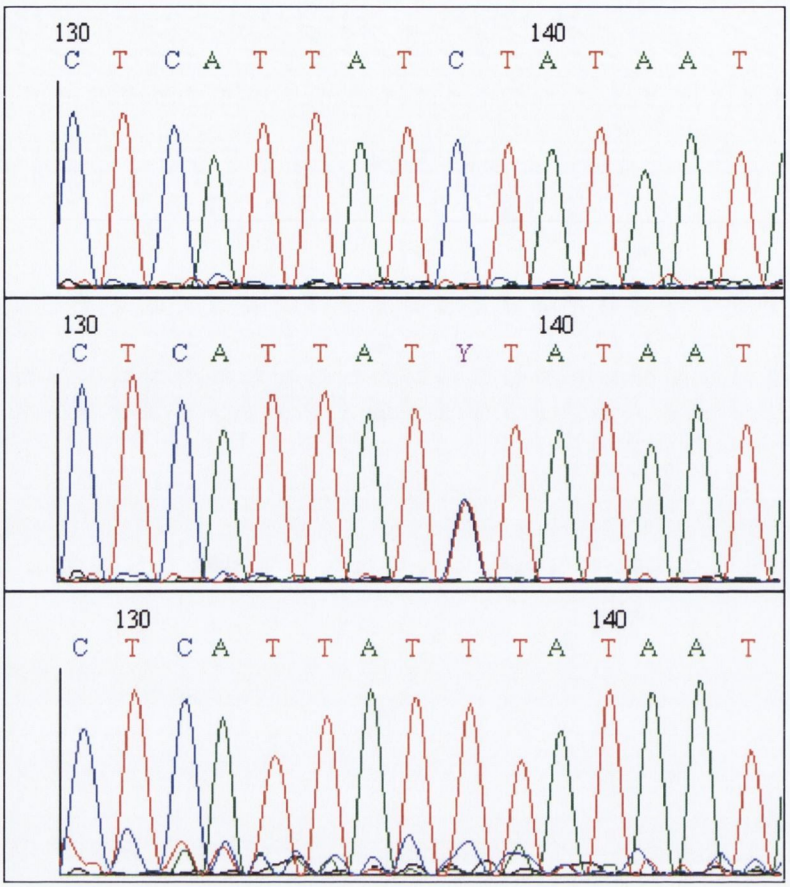


Figure 3.9f: Electropherograms confirming the heterogeneity of rs880289. Top image, C/C common heterozygote. Middle image C/T heterozygote. Bottom image T/T 'rare' homozygote.

3.3.2 Results of further association analysis in the GASP case-control sample

In Chapter 2, twelve SNPs were genotyped across the *HOMER2* region in the GASP case-control sample (Stage 1). In order to improve coverage of common variants and potentially refine the existing association signal, ten supplementary tagging SNPs and three novel, potentially-functional SNPs were genotyped in the GASP sample (Stage 2). Of the fourteen additional SNPs tested for association two of them, rs2306428 and novel SNP 'B' were found to be significantly associated with schizophrenia (table 3.5 and figure 3.10). SNP rs2306428 is in high LD with rs7174726 (CEU r2=0.74) at which significant association was previously identified (see Chapter 2) and located 34,322bp away in intron 6 and 8bp 3' of exon 6. The associated allele at rs2306428 was on the same haplotype as that for rs7174726. No other SNPs in high LD with rs2306428 were close to potential functional sites. This finding widens greatly the region of association. Novel SNP 'B' is also in high LD with a previously associated SNP (rs12913501, CEU r^2=0.87) and located close to the predicted Sp1 binding-site in intron 1. The associated allele at novel SNP 'B' was on the same haplotype as that for rs12913501

3.3.3 Haplotype association results in the GASP and RPGI samples

Of the five SNPs at *HOMER2* associated with schizophrenia in the GASP sample, three were selected for haplotype analysis: rs2306428, Novel SNP 'B' and rs869498. These three were on different haplotypes and prioritised over other associated SNPs on the same haplotype because of predicted functionality. The strongest result was with rs2306428-rs869498 (CC haplotype, frequency of 0.833 in cases and 0.773 in controls, $p=6.0\times10^{-4}$, permuted p=0.002) although other haplotypes were also significant (see table 3.6).

#[1]	SNP ID	Stage	Position[2]	Alleles[3]	Case MAF	Control MAF	χ^2	P	OR	95% CI
1	rs1256430	1	81,316,526	A/G	0.165	0.162	0.02	0.878	1.02	0.80-1.29
2	rs12442119	1	81,319,799	T/C	0.269	0.272	0.02	0.883	0.99	0.81-1.20
3	rs1256426	2	81,320,422	G/C	0.473	0.454	0.71	0.400	1.08	0.90-1.29
4	**rs2306428**	**2**	**81,320,441**	**T/C**	**0.054**	**0.085**	**7.43**	**0.006**	**0.61**	**0.43-0.87**
5	rs1256444	1	81,337,307	C/T	0.462	0.443	0.71	0.398	1.08	0.90-1.29
6	rs1256455	1	81,341,576	T/C	0.251	0.226	1.73	0.189	1.15	0.93-1.41
7	rs2621226	1	81,342,479	A/G	0.441	0.429	0.27	0.602	1.05	0.88-1.26
8	rs17359494	1	81,346,341	C/T	0.224	0.243	1.08	0.300	0.89	0.73-1.10
9	rs2621227	2	81,349,516	A/G	0.493	0.498	0.04	0.843	0.98	0.82-1.17
10	SNP 'C'	2	81,352,332	A/T	0.010	0.014	0.71	0.399	0.69	0.29-1.63
11	**rs7174726**	**1**	**81,354,763**	**C/T**	**0.066**	**0.095**	**4.97**	**0.026**	**0.68**	**0.48-0.96**
12	rs7496832	2	81,355,797	C/A	0.109	0.138	3.71	0.054	0.76	0.58-1.01
13	SNP 'A'	2	81,366,190	A/T	0.026	0.017	2.14	0.144	1.55	0.86-2.81
14	**SNP 'B'**	**2**	**81,366,498**	**T/C**	**0.112**	**0.142**	**4.18**	**0.041**	**0.76**	**0.59-0.99**
15	**rs12913501**	**1**	**81,372,191**	**T/C**	**0.109**	**0.143**	**4.84**	**0.028**	**0.74**	**0.56-0.97**
16	**rs869498**	**1**	**81,376,406**	**T/C**	**0.133**	**0.100**	**6.09**	**0.014**	**1.39**	**1.07-1.81**
17	rs12148275	1	81,379,899	G/A	0.486	0.484	0.01	0.918	1.01	0.85-1.21
18	rs12050725	1	81,381,349	A/T	0.201	0.186	0.71	0.401	1.10	0.88-1.37
19	rs12443081	2	81,394,567	A/G	0.101	0.099	0.01	0.930	1.01	0.76-1.36
20	rs2061947	1	81,397,236	T/C	0.467	0.448	0.71	0.400	1.08	0.90-1.29
21	rs7170046	2	81,402,099	T/C	0.102	0.113	0.63	0.429	0.89	0.67-1.19
22	rs1010820	2	81,405,958	A/T	0.194	0.219	1.84	0.175	0.86	0.69-1.07
23	rs11857990	2	81,414,427	G/A	0.295	0.328	2.52	0.112	0.86	0.71-1.04
24	rs4842928	2	81,417,560	A/T	0.172	0.170	0.01	0.918	1.01	0.80-1.28
25	rs1871658	2	81,428,253	G/C	0.433	0.396	2.99	0.084	1.17	0.98-1.39
26	rs6603038	2	81,437,335	T/C	0.510	0.478	2.03	0.155	1.14	0.95-1.36

[1] Number is reference to SNP position in figure 3.10

[2] Chromosome 15, NCBI Build 36.1/hg18.

[3]The minor allele is stated first. All alleles are called on the forward strand.

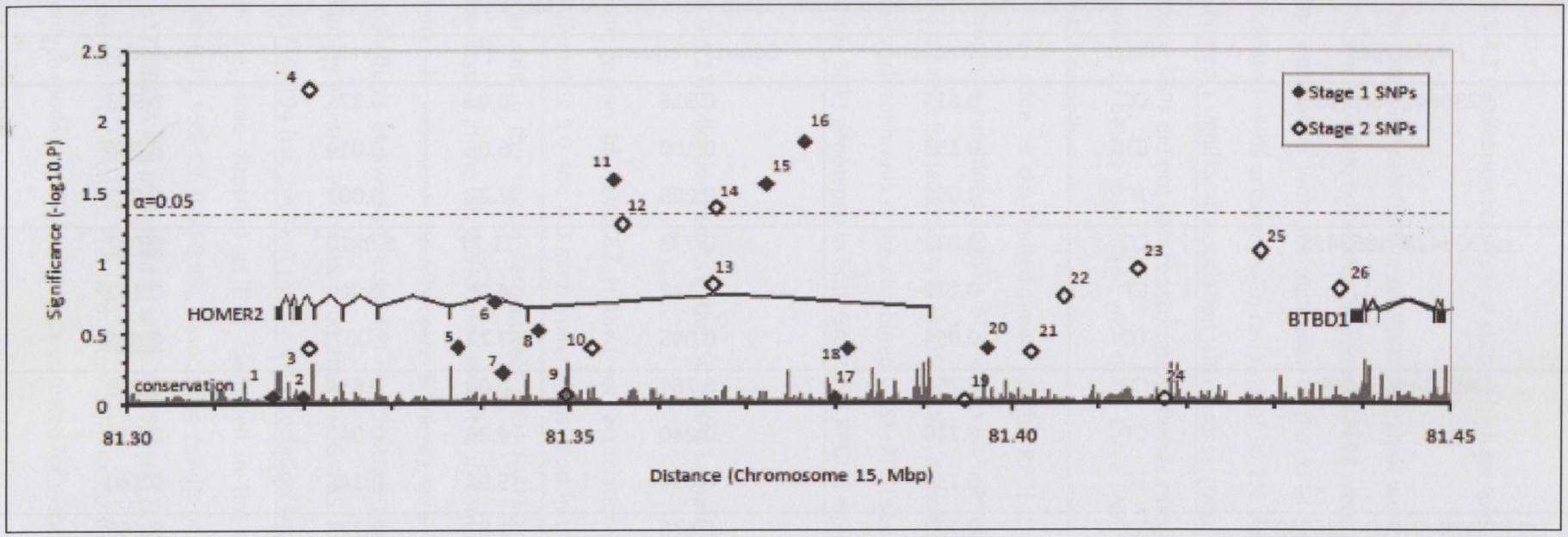


Figure 3.10: Diagram of the *HOMER2* gene region show position of SNPs genotyped (x-axis) and level of association with schizophrenia in the GASP sample. The number next to each SNP marker is a reference to further association statistics in Table 3.5. SNPs genotyped in Stage 1 are indicated by filled diamonds. SNPs genotyped in Stage 2 are indicated by unfilled diamonds.

Table 3.6: Haplotype association results from the GASP sample

Haplotypes	Alleles	Case frequency	Control frequency	χ^2	p-value	Permuted p-value[1]
rs2306428-SNP20	CC	0.813	0.816	0.03	0.871	0.998
	CT	0.133	0.100	6.06	0.014	***0.046***
	TC	0.054	0.085	7.27	0.007	***0.023***
rs2306428-rs869498	CC	0.833	0.773	11.77	6.0×10^{-4}	***0.002***
	CT	0.112	0.142	4.06	0.044	0.142
	TC	0.054	0.085	7.24	0.0071	***0.028***
SNP20-rs869498	CC	0.757	0.761	0.03	0.854	0.999
	CT	0.110	0.240	4.20	0.040	0.052
	TC	0.130	0.097	5.98	0.145	0.140
rs2306428-SNP20-rs869498	CCC	0.703	0.676	1.82	0.178	0.466
	CCT	0.110	0.140	4.23	0.040	0.122
	CTC	0.131	0.098	6.04	0.014	***0.041***
	TCC	0.054	0.085	7.27	0.007	***0.022***

[1] n=10,000 permutations

3.3.4 Replication results at *HOMER2*

3.3.4.1 Replication results in a second Irish sample (RPGI)

The three SNP selected for replication analysis were the same as those selected for haplotype analysis. Thus, rs2306428, novel SNP 'B' and rs869498 were genotyped using the TaqMan® method and tested for association with schizophrenia in the RPGI sample. The genotyping quality control measures indicated >99% accuracy with call rate of 95.2%. All SNPs were in HWE (p>0.01). None of these SNPs showed a significant association with disease status (see table 3.7). Similarly, haplotype analyses with these SNPs did not detect association. Power calculations indicate that the RPGI sample had 28-45% power to detect association with these three SNPs based on their observed odds ratios (ORs) in the GASP sample.

Due to the limited power of the RPGI to replicate the association results from the GASP sample, a combined analysis of the two Irish samples was performed with replication testing of any nominally associated SNPs in the ISC GWAS dataset. One SNP, rs2306428, was significantly associated with schizophrenia in the combined GASP+RPGI sample (p=0.008, OR=0.73). Association results of the three SNPs genotyped in both GASP and RPGI samples are detailed in table 3.7.

3.3.4.2 Replication results in the ISC GWAS sample

The region of *HOMER2* studied for association was chr15: 81,310,000-81,425,000 (hg18). The Affymetrix 5.0 and 6.0 genotyping chips used in the ISC GWAS captured 53% and 85% of common variants in this region respectively (CEU MAF>0.1, r^2 threshold of 0.8). SNP rs2306428 was not genotyped in the ISC GWAS. However it is in complete LD with rs17158194 (r^2=1.0 in HapMap CEU) which is present on the Affymetrix 6.0 array only. Thus 1,287 schizophrenia cases and 1,128 control individuals of Swedish, Bulgarian and Scottish ancestry were tested for association at rs1715894. Quality control information on the ISC GWAS has been reported previously and was good for rs17158194 (ISC 2009). A significant association with schizophrenia was identified at rs17158194 and in the same direction as that identified in the combined GASP+RPGI sample (χ^2=5.53, p=0.019, OR=0.77, 95% C.I. 0.60-0.96). Association results of the individual and combined samples in the ISC at rs17158194 are presented in table 3.8. SNP rs17158194 is located in intron 3 of *HOMER2* and has no identifiable function. The best proxy SNP for rs2306428 on the Affymetrix 5.0 genotyping array was only in moderate LD and so was not tested for association (rs1002881, CEU r^2=0.54).

Table 3.7: Replication analysis of *HOMER2* and association with schizophrenia

Sample:	GASP sample					RPGI sample					GASP+RPGI					ISC GWAS Sample[1] (excluding all Irish samples)				
SNP ID	MAF		P	OR	95% CI	MAF		P	OR	95% CI	MAF		P	OR	95% CI	MAF		P	OR	95% CI
	Cases (375)	Controls (812)				Cases (408)	Controls (840)				Cases (783)	Controls (1651)				Cases (1287)	Controls (1128)			
rs2304628	0.054	0.085	0.006	0.61	0.43-0.87	0.078	0.090	0.310	0.85	0.622-1.20	0.065	0.088	0.008	0.73	0.58-0.94	0.054	0.070	0.019	0.75	0.59-0.95
Novel SNP 'B'	0.112	0.142	0.041	0.76	0.59-0.99	0.144	0.127	0.232	1.16	0.91-1.48	0.128	0.134	0.556	0.95	0.79-1.13	-	-	-	-	-
rs869498	0.133	0.100	0.014	1.39	1.07-1.81	0.113	0.120	0.590	0.93	0.71-1.21	0.123	0.110	0.186	1.13	0.94-1.37	-	-	-	-	-

[1] Values presented for p-value and Odds Ratio are from a Cochran-Mantel-Haenszel test for association. MAF was calculated independently. SNP rs2304628 was not genotyped in the ISC sample so results from proxy SNP rs17158194 (r^2=1.0 in CEU) are presented here. Note that the values in brackets correspond to the number of cases and controls genotyped.

Table 3.8: Association analysis of rs17158194 in individual ISC samples[1]

Study site	Ethnicity	Sample size		MAF		p-value	OR	95% C.I.
		Cases	Controls	Cases	Controls			
Edinburgh	Scottish	388	283	0.074	0.078	0.821	0.953	0.63-1.44
Cardiff	Bulgarian	482	647	0.048	0.060	0.211	0.785	0.54-1.46
Sweden II	Swedish	425	242	0.042	0.074	0.018	0.572	0.35-0.90
Combined		1295	1172	0.054	0.070	0.019	0.754	0.60-0.96

[1] The p-value and odds ratios presented here were calculated with the Cochran-Mantel-Haenszel test for association. Minor allele frequencies were calculated separately.

3.3.5 Predicted funtional properties of associated SNPs

Although located in intron 6, rs2306428 is located only 8bp 3' of exon 6. To investigate the possible effects on splicing, a 50bp sequence centred on rs2306428 was screened for splice-enhancer binding-sites using the ESRsearch Tool (http://ast.bioinfo.tau.ac.il/, after Zhang & Chasin 2004). This identified a splice enhancer binding-site motif which was abolished in the presence of the rare allele (AC**G**GTGGA→AC**A**GTGGA, chr15: 81,320,436-81,320,444, reverse stand, see figure 3.11).

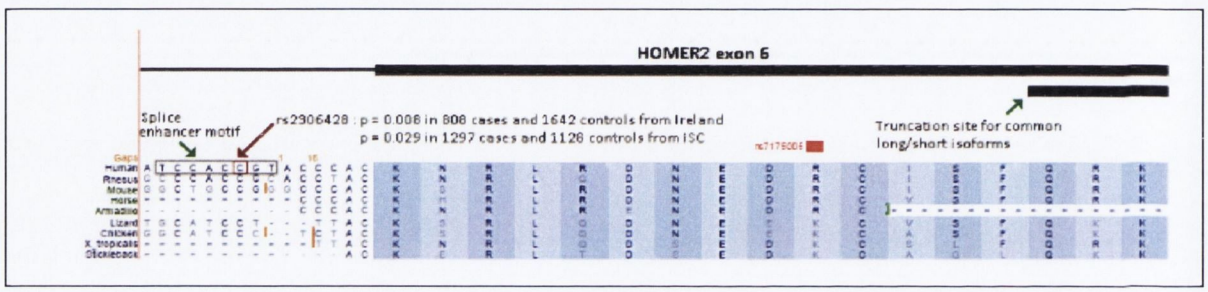


Figure 3.11: Diagram showing position of rs2306428 in relation to exon 6 and the predicted splice-enhancer binding site. Note that *HOMER2* is transcribed on the reverse strand. Also shown is the sequencing other organisms demonstrating conservation in higher vertebrates. The SNP in exon 6 (rs7175005) was previously investigated and found to be monomorphic in fifteen schizophrenia cases (see Chapter 2). This is in agreement with recent data from the HapMap although the rare variant is present up to MAF 0.2 in non-Caucasian populations (HapMap, http://hapmap.ncbi.nlm.nih.gov/).

3.3.6 Comparison of genotype at rs2306428 with *HOMER2* expression

The rare allele at rs2306428 was predicted to reduce natural truncation of HOMER2 at exon 6 which may lead to an excess of long isoforms of the gene. *HOMER2* mRNA was quantified by PCR targeting both exons 1 and 2, and 8 and 9, using GAPDH mRNA as a control. The difference in the cycle threshold (ΔΔct) between exon 1-2 and exon 8-9 should be an accurate refection of the relative quantities of short and long HOMER2 isoforms. Thus, the ΔΔct for exon 1-2 compared to that for exons 8-9 (rather than GAPDH) was used as an indicator of splicing alterations. As only a single individual at rs2306428 was homozygous for the rare allele, only two genotype groups were compared: 'AA+AG' verses 'GG'. There was no significant difference in this value between the two genotype groups indicating that variation at rs2306428 did not affect splicing (Table 3.9).

The effect of genotype at rs2306428 on overall *HOMER2* expression was also investigated. Expression data on the same CEU cell lines were also used from the Genvar Project. There was general agreement between data generated in-house by quantitative PCR and that obtained from the Genvar Project (figure 3.12, note that high ΔΔct and expression levels are inversely correlated as expected). Data from single probe targeting long and short isoforms was available. No significant differences were found between the two genotype groups in *HOMER2* expression from any data source discounting any possible *cis*-acting effect of rs2306428 (Table 3.9).

No significant difference was identified between the two genotypes on *HOMER2* isoform difference or overall expression when the data was analyzed using the quantitate trait analysis function in PLINK (best result exons 8-9, β=0.32, SE=0.50, r^2=0.006, T=0.69, p=0.494).

Table 3.9: *HOMER2* expression in CEU lymphoblastoid cells grouped by genotype at rs2306428

Experiment type	Target region	Units	Expression level (average±SD)		Student's p-value[2]	QT p-value[3]
			AA+AG (n=16)	GG (n=74)		
Microarray[1]	3'UTR	RIU	06.40 (0.26)	06.37 (0.25)	0.343	0.674
Q-PCR[4]	Exons 1-2	ΔΔct	16.86 (2.13)	16.71 (2.11)	0.402	0.718
	Exons 8-9	ΔΔct	12.00 (2.24)	11.72 (1.76)	0.294	0.494
	Ex1-2 vs Ex8-9	ΔΔct	04.86 (1.20)	04.99 (1.26)	0.349	0.684

[1]Illumina WG-6 expression array, probe ID GI_4758547-S. RIU: Relative Intensity Units.

[2]One-tailed Student's t-test, unpaired with equal variance.

[3]Calculated using the quantitative trait function in PLINK

[4]The ΔΔct for exons 1-2 and 8-9 was calculated using GAPDH as an internal control. The ΔΔct Ex1-2 vs Ex8-9 was calculated for exons 1-2 using exons 8-9 as an internal control. Note that ΔΔct is always inversely proportional to actual gene expression levels.

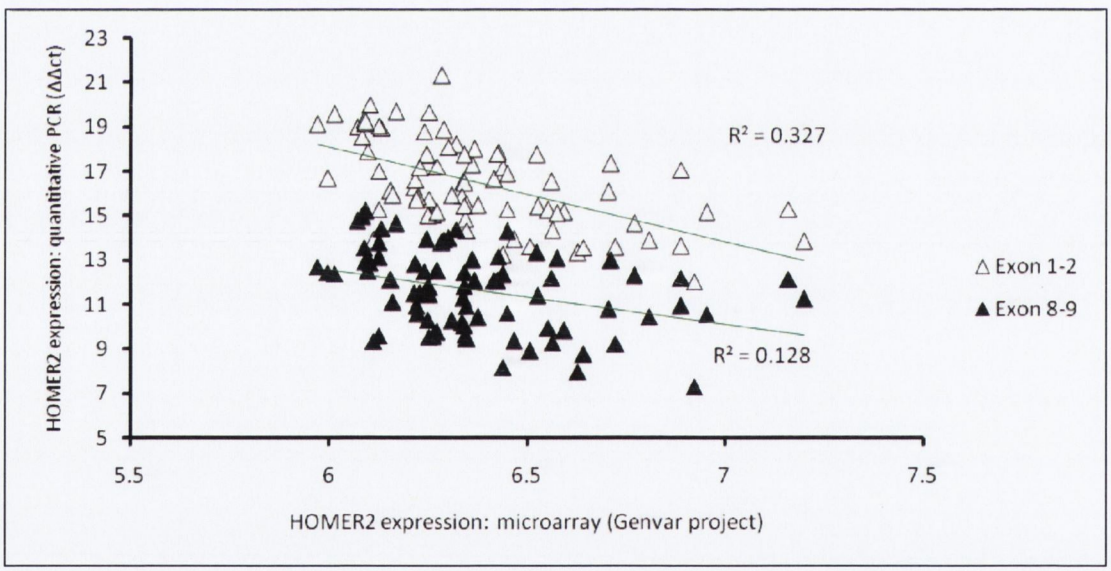


Figure 3.12: Comparison of expression levels of *HOMER2* determined by either microarray (Genvar Project) or quantitative PCR (in-house) in the same CEU lymphoblastoid cells. Note that ΔΔct is always inversely proportional to the gene expression levels.

3.3.7 Comparison of genotype at rs2306428 with neuropsychological variables

The effect of genotype at rs2306428 on neuropsychological endophenotypes was investigated in 335 schizophrenia cases. Given the low frequency of the 'A' allele, analysis of neuropsychological variables was undertaken as a two groups comparison, i.e. carriers of at least one copy of the T allele (n=40) versus carriers of two copies of the 'G' allele (n=295). Given the differences in sample size between genotype groups, comparisons were made using a series of Mann-Whitney-U tests. *HOMER2* rs2306428 genotypes were not associated with any neuropsychological performance differences (table 3.10).

Table 3.10: Neuropsychological performance of controls grouped by genotype at rs2306428

Neuropsychological tasks	AA+AG (n=64)	GG (n=369)	Mann-Whitney U	p-value
Cognition				
WTAR estimated IQ score	96.6 (11.2)	96.2 (11.9)	12048.5	0.933
Full scale IQ	88.1 (17.9)	89.9 (17.2)	10663.5	0.358
Verbal IQ	91.7 (19.7)	92.8 (16.6)	5344	0.356
Performance IQ	86.8 (16.9)	90.5 (19.1)	5262	0.307
Episodic memory				
Logical memory 1 total recall scaled score	7.0 (3.4)	6.7 (3.4)	10830.5	0.461
Logical memory 2 total recall scaled score	8.1 (3.3)	7.5 (3.1)	9774	0.17
CANTAB PAL (standard score)	-2.5 (3.5)	-2.3 (4.8)	10020	0.486
Working memory				
Letter-number sequencing scaled score	7.8 (3.2)	7.7 (3.4)	10728.5	0.748
CANTAB SWM (Standard scores)	-0.88 (1.35)	-0.85 (2.23)	6297	0.664
Attention				
CANTAB IDED ED errors	221 (427)	165 (375)	706.5	0.719
CANTAB IDED ID errors	222 (427)	155 (363)	671.5	0.473
CANTAB IDED total errors	200 (35)	189 (41.8)	9233.5	0.419
CPT D' (2 numbers)	2.6 (1.2)	2.7 (1.1))	4831	0.96
CPT D' (3 numbers)	1.9 (1.0)	1.8 (1.0)	4368.5	0.831
CPT D' (4 numbers)	1.0 (0.8)	1.0 (0.8)	4332.5	0.899

3.5 Discussion

This is the first comprehensive genetic association study of *HOMER2* as a putative schizophrenia susceptibility gene. In a multi-stage study, data was utilized from a total of 2,096 cases and 2,780 controls. Altogether, 26 SNPs were analysed in GASP sample, effectively capturing 84% of common SNPs at this locus based on HapMap CEU LD data (r^2>0.8). Five SNPs were nominally associated (p<0.05) with schizophrenia, representing three independent signals. Three SNPs representing these signals were taken forward for analysis but did not show evidence of association in the RPGI sample which was relatively under-powered (<45%). One SNP, rs2306428, was associated in the combined Irish samples (p=0.008) and we sought to replicate this finding in the larger European ISC dataset. This SNP was not genotyped directly in the ISC GWAS but the association was replicated by a nearby SNP, rs17158194, in complete LD with rs2306428 (p=0.029). A combined analysis of all the Irish and ISC samples has not been undertaken as data is not available on the same SNP in all samples. However, rs17158194 is the most strongly associated SNP at *HOMER2* in the ISC GWAS (p=0.00067) when the Irish samples within ISC are included.

The SNP, rs2306428 is 8bp away from the 3' end of exon 6 of *HOMER2*. Because of its proximity to the splice-junction, we investigated *in silico* the potential effect of variation at this site on splicing. The common, susceptibility allele was found to be located in a predicted exon splicing enhancer motif, which is abolished by the rare, protective allele. Splicing at exon 6 is an important part of the normal function of *HOMER2* as this is the point at which the mRNA is truncated in response to neuronal stimulation (Soloviev *et al* 2000b, Shiraishi *et al* 2003b). Truncation of HOMER proteins prevents polymerization. This in turn, causes disassociation of glutamate receptors from one another and from downstream effectors and is thus crucial to the regulation of glutamatergic neurotransmission (Brakeman *et al* 1997, Kammermeier & Worley 2007). Whether the rare allele at rs2306428 affects truncation of *HOMER2* or not remains to be determined. If so, it is likely to have a phenotypic effect as over-expression of the long HOMER2 isoform in the nucleus accumbens of mice increases ethanol-induced neuroplasticity by prolonged activation of metabotropic and NMDA glutamate receptors (Szumlinski et al. 2008). In addition, overexpression of short HOMER1 isoforms in the hippocampus of mice specifically impairs spatial working memory and abolishes long-term potentiation in the CA3-CA1 region (Celikel *et al* 2007).

The direction of association at rs2306428 indicates that the rare allele is protective. Thus, altered sensitivity of HOMER2 truncation in response to neuronal stimulation may protect an otherwise susceptible individual against schizophrenia. Alternatively, even if rs2306428 is a

functional polymorphism, it cannot be discounted that rs2306428 is LD with a true protective or susceptibility variant. We found an initial association at another potentially-functional SNP (SNP20), close to a strongly predicted transcription factor binding site in intron 1. However, we were unable to replicate the association in the second Irish sample or by proxy in the ISC sample. The number of SNPs initially found to be associated at the *HOMER2* gene was higher than expected by chance. Thus *HOMER2* may harbour multiple common, susceptibility variants to schizophrenia which each act independently. This is biologically feasible as a functional variant anywhere in the *HOMER2* gene should have a significant effect on glutamatergic transmission starting from the post-synapse and notably in brain regions dominated by expression of HOMER2 (rather than HOMER1 or 3). In addition such variants may interact to confer a scale of susceptibility.

Although no *single* study has found significant evidence of linkage to the *HOMER2* region (15q25.2), studies which contributed to the finding in the meta-analysis study by Lewis and collegues (2003) used samples from Wales/UK (Williams *et al* 1999), Ireland (Straub *et al* 2002b) and Finland (Paunio *et al* 2001). The first two samples are genetically-similar to the sample used in the present study and thus may also harbour significant associations at *HOMER2*. However no associations with schizophrenia were identified at four SNPs genotyped in *HOMER2* using the sample from Wales/UK, notably from rs2306428 (Norton *et al* 2003). Alternatively another gene in 15q25 may harbour susceptibility variants. *HOMER2* is the only one in this region known to be directly involved in glutamatergic neurotransmission but there are others whose products localise to the synapse and thus are functionally-relevant to schizophrenia (Schrimpf *et al* 2006).

Among the twenty-two genes whose products are known to interact with HOMER proteins (see Shiraishi-Yamaguchi & Furuichi 2007 for review), Cell division/cycle 42 (*CDC42*, OMIM 116952) is located 15Kb downstream from the eighth most significant association signal identified in the ISC GWAS. *CDC42* is located at chromosome 1p36.12 and encodes a Rho GTPase. In mammals expression is particularly high in the hippocampus, cerebellum, neocortex and thalamus (Olenik *et al* 1997). The CDC42 protein has a central neurodevelopmental role both in establishing neuronal polarity and dendrite outgrowth (Threadgill *et al* 1997, Schwamborn & Püschel 2004, Pertz *et al* 2008). Activated CDC42 binds to the C-terminal of HOMER2 in developing cerebellar cells and may be important in determining cellular localisation and binding to other proteins (Shiraishi *et al* 1999). Two lines of evidence link CDC42 to schizophrenia pathophysiology. Firstly, levels of *CDC42* mRNA are significantly decreased in schizophrenia brains and correlated with reduction in dendritic spine density in cortical deep layer 3 neurons (Hill *et al* 2006). Secondly, the DISC-1 interactor, NUDEL is known to facilitate cell migration by sequestering CDC42GAP at the leading edge of mouse fibroblasts to stabilize CDC42 activity in response to extracellular stimuli (Shen *et al* 2008). This

suggests a potential mechanism whereby scaffolding proteins involved in the DISC-1 complex could stabilize cell migration, polarity and dendrite outgrowth with direct down-stream effects (through HOMER2) on NMDA-related synaptic plasticity. The most significant signal at *CDC42* in the ISC GWAS was at rs2473277 ($p=2.00\times10^{-6}$, OR=0.85) in the promoter region of the gene (ISC 2009).

In conclusion, we have investigated common variation at *HOMER2*, a functionally relevant gene with prior evidence of linkage, and found modest evidence to support association with schizophrenia. The SNP most strongly associated with disease is located close to the known point of isoform diversification and predicted to change a relevant sequence motif. Two interactors of the HOMER2 protein were among the most significant gene findings from the ISC GWAS study.Further studies are required to fully assess the functional impact of this SNP on *HOMER2* splicing and elucidate the putative role of *HOMER2* in schizophrenia aetiology, in particular involving interaction with proteins such as MYO18B and CDC42.

Chapter 4

Association and functional analysis of the *CDC42* gene

4.1 Introduction

4.1.1 Interactors of the HOMER2 protein in the International Schizophrenia Consortium (ISC) GWAS data

In Chapter 3, an association of a common genetic polymorphism at the *HOMER2* gene and schizophrenia was identified. Homer family proteins are crucial regulators of post-synaptic activity which is putatively implicated in schizophrenia pathogenesis. Thus other genes encoding regulators of post-synaptic activity, for example HOMER2-binding partners, may also harbour susceptibility variants. Twenty genes encode proteins which exhibit a physical interaction with the HOMER2 protein and have been tested for association in the ISC GWAS (table 4.1). Of these genes, the greatest level of association in the ISC GWAS (by smallest p-value) was identified at the *CDC42* gene (rs2473277, $p_{CMH}=2.003\times10^{-6}$, OR=0.845). This association signal was ranked 8th by LD-independent p-value in the ISC GWAS and is located 15Kb upstream from *CDC42* transcription start (figures 4.1 and 4.2). Reduced expression of *CDC42* in the brain has been demonstrated in schizophrenia subjects (Hill *et al* 2006). For these reasons, the *CDC42* gene was selected for further investigation as a schizophrenia susceptibility candidate.

4.1.2 Summary of chapter

In the ISC GWAS the strongest association at any of the twenty known interactors of HOMER2 was at the *CDC42* gene at SNP rs2473277. This SNP is located upstream of *CDC42* and so was investigated for effects on gene expression level. Two publicly-available resources were mined for correlations of rs2473277 with *CDC42* expression level. This approach identified evidence of a reduction in expression with the susceptibility allele. SNP rs2473277 is not located in a predicted functional site but was found to be in high LD with another SNP in which the allele associated with reduced expression removed a predicted transcription factor binding-site. A gene reporter assay and an

electrophoretic mobility shift assay were undertaken in order to test whether the allelic status of rs2473307 had a direct functional effect on gene expression. To investigate whether the associated variant conferred susceptibility to a specific endophenotype of schizophrenia, quantitative neuropsychological variables were also tested against genotype at rs2473277. Finally, replication analyses of the association signal were performed using data from two independent genome-wide association studies of schizophrenia.

4.1.3 Background on *CDC42*

4.1.3.1 The *CDC42* gene

The *CDC42* gene (OMIM 116952) is located at chromosome 1p36.12 (Jensen *et al* 1997). No linkage to schizophrenia has been reported at this locus. Rare CNVs in a non-disease sample have been reported to overlap *CDC42* (Perry *et al* 2008) although none were detected in the International Schizophrenia Consortium (ISC) copy-number analysis (ISC 2008). The 2,182bp transcript is made up of eight exons with 5' and 3' UTRs spanning 11,111bp genomic sequence (chr1:22,251,707-22,292,023 Build 36.1, March 2006).

4.1.3.2 The CDC42 protein

The human CDC42 protein (first named G_P and then G25K) was first identified as a GTPase in the placental membrane (Evans *et al* 1986). CDC42 was also identified independently in mutated yeast strains exhibiting cell-division/cycle phenotypes (Johnson & Pringle 1990, Shinjo *et al* 1990). The GTPase family consists of many different proteins which hydrolyse GTP, a major source of energy for a range of cellular functions. They behave as a molecular switch for particular pathways such as DNA replication, protein synthesis, vesicle trafficking and signal transduction (Heasman & Ridley 2008). Interestingly other brain-related disorders such as mental retardation and William's syndrome can be caused by mutations in GTPase-encoding genes (Negishi & Kato 2005). In stimulated neurons, CDC42 promotes formation of the N-WASP–Arp2/3 complex which in turn activates actin microtubule assembly (Nobes & Hall 1995, Rohatgi *et al* 1999, Shen *et al* 2006). Through microtubule assembly CDC42 regulates the formation and morphology of dendritic spines and synapses

(Threadgill *et al* 1997, Scott *et al* 2003). Furthermore, the cdc42 protein was found to be central in the control of growth cone establishment and neurite extension in neuronal proteome study (Pertz *et al* 2008). In the most extreme demonstration of the importance of CDC42 to neuronal development, mice lacking the *cdc42* gene in neuronal progenitor cells fail to establish polarity in the telencephalon. This leads to holoprosencephaly, a brain malformation characterised by incomplete separation of the cerebral hemispheres (Chen *et al* 2006). Expression of *CDC42* in the brain is particularly high in the hippocampus, cerebellum, neocortex and thalamus, all regions of interest in schizophrenia (Olenik *et al* 1997, Harrison 1999 for review). In the mRNA, exons 7 and 8 are mutually-exclusive and highly-conserved indicating that they a crucial feature of Cdc42 protein function (Marks & Kwiatkowski 1996, Nicole *et al* 1999).

4.1.3.1 CDC42 and HOMER2

The interaction between CDC42 and HOMER2 was first identified in murine cerebellar cells (Shiraishi *et al* 1999). Binding of HOMER2 and CDC42 inhibits filopodia formation in HeLa cells and is likely to be the same in neuronal cells in the attenuation of dendrite outgrowth (Shiraishi-Yamaguchi *et al* 2007).

4.1.3.2 CDC42 and DISC1

There is substantial evidence for the role of Disrupted-in-Schizophrenia-1 (DISC1) gene in schizophrenia pathogenesis (Porteous *et al* 2006 for review). The CDC42 protein is biochemically linked with the DISC1 protein through their intermediate binding partners, CDC42GAP and NUDEL. Here, CDC42GAP binds and inactivates CDC42 until sequestered by NUDEL (Shen *et al* 2008). The binding of NUDEL to DISC1 is also essential for neurite outgrowth, defects of which may lead to schizophrenia (Ozeki *et al* 2003).

Table 4.1: Summary association results for interactors of the HOMER2 protein in the ISC GWAS

Protein symbol[1]	Reference	Co-ordinates[2]	Best p-value	SNP ID	Genotyping density (SNPs per Kb)[3]
CDC42	Shiraishi *et al* 1999	chr1:22,251,707-22,292,023	2.00×10^{-6}	rs2473277	0.25
DNM3	Gray *et al* 2003	chr1:170,077,261-170,648,480	0.003	rs6678749	0.32
ITPR1	Tu *et al* 1998	chr3:4,700,074-4,864,286	0.013	rs874141	0.56
GRM2	Kato *et al* 1998	chr3:51,716,128-51,727,665	0.029	rs4687592	0.08
DBR1	Shiraishi *et al* 2003	chr3:139,362,542-139,365,735	0.054	rs774803	0.08
TRPC1	Yuan *et al* 2003	chr3:143,925,956-144,009,419	0.027	rs7430429	0.18
HOMER1	Shiraishi *et al* 2003	chr5:78,705,542-78,845,456	0.009	rs2404149	0.19
ITPR3	Tu *et al* 1998	chr6:33,697,139-33,772,326	0.003	rs12205634	0.36
GRM1	Kato *et al* 1998	chr6:146,390,475-146,800,424	0.014	rs979775	0.15
UBXN1	Ishibashi *et al* 2005	chr11:62,200,548-62,203,103	0.188	rs12808085	0.07
SHANK2	Tu *et al* 1999	chr11:70,026,549-70,420,323	0.104	rs11237157	0.24
GRM5	Kato *et al* 1998	chr11:88,357,521-88,438,761	0.105	rs7122757	0.10
IQSEC3	Inaba *et al* 2004	chr12:117,694-156,583	0.035	rs3888294	0.31
AGAP2	Rong *et al* 2003	chr12:56,405,261-56,422,207	0.019	rs2277326	0.15
HOMER3	Shiraishi *et al* 2003	chr19:18,901,012-18,912,113	0.12	rs3177137	0.14
RYR1	Tu *et al* 1998	chr19:43,616,180-43,770,044	0.003	rs8112211	0.18
SHANK1	Tu *et al* 1999	chr19:55,856,896-55,912,007	1.62×10^{-4}	rs6509496	0.15
APP	Parisiadou *et al* 2008	chr21:26,174,732-26,465,003	0.002	rs2829955	0.35
MYO18B	Ajima *et al* 2007	chr22:24,752,411-24,757,007	0.028	rs2105945	0.59
SHANK3	Tu *et al* 1999	chr22:49,459,936-49,518,507	0.027	rs739365	0.16

1. Protein symbol is based upon the official HUGO nomenclature. Thus different names may be used in publications.
2. The stated co-ordinates refer to the longest transcript start and stop positions. Those used in the analysis were for 100Kb upstream and 20Kb downstream, taking into account which strand (+ or -) the gene is transcribed from.
3. Genotyping density (SNPs per Kb) is also based on 100Kb upstream and 20Kb downstream. Genotyping density on the Affymetrix 5.0 may be lower in some cases. This value does not take LD into account.

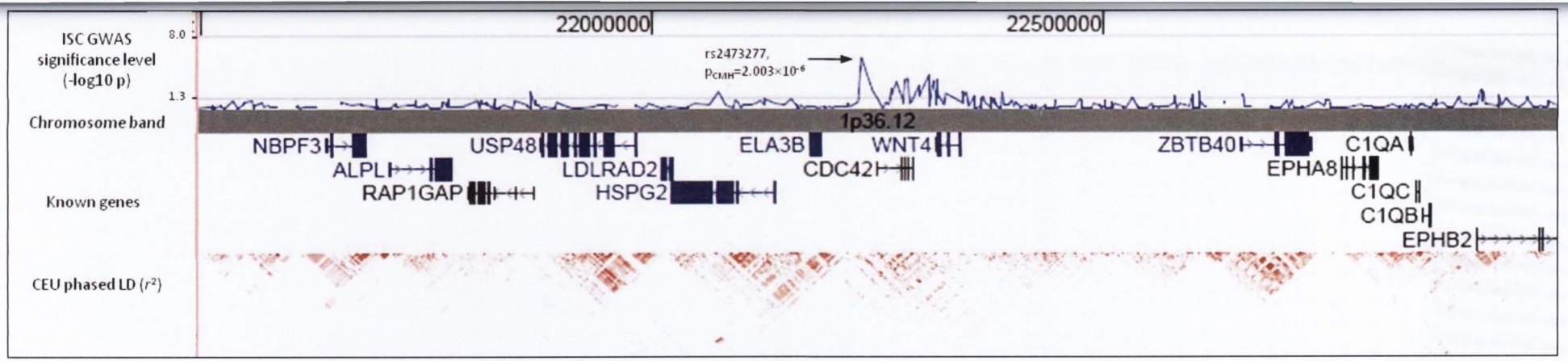


Figure 4.1: Diagram showing significance levels from the ISC GWAS in a 1.5Mb region centred on the *CDC42* gene (chr1:21,500,000-23,000,000, hg18). The highest level of association with schizophrenia here is indicated by the arrow at rs2473277, p_{CMH}=2.003×10^{-6}, OR=0.8453. SNPs genotyped in the ISC are connected with light blue lines unless separated by >20Kb. Also shown are the positions on chromosome 1, chromosome band, neighbouring genes and linkage disequilibrium patterns in the CEU. (Diagram generated using the UCSC Genome Browser.)

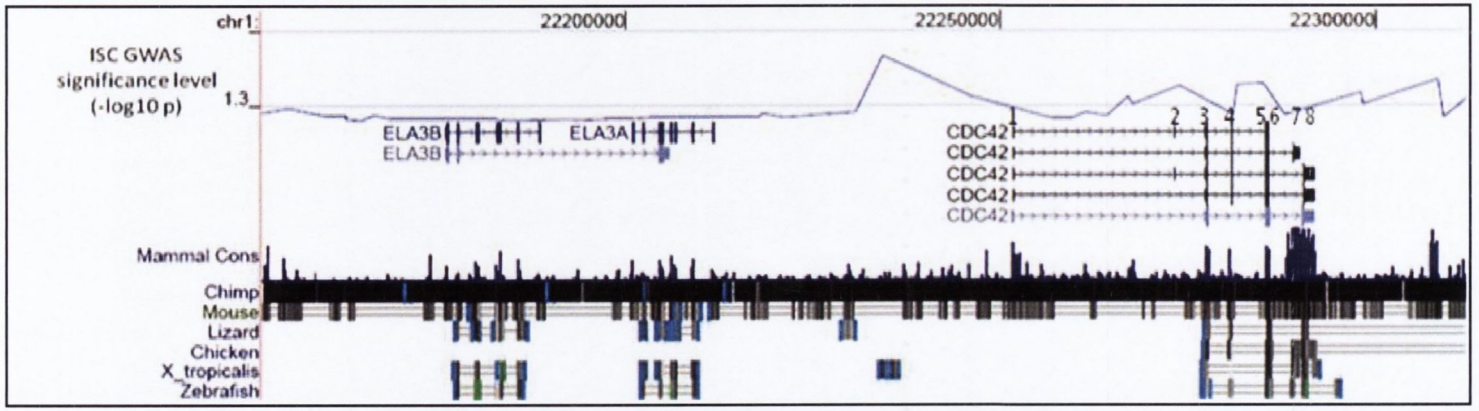


Figure 4.2: Diagram showing significance levels from the ISC GWAS at the *CDC42* gene including 100Kb upstream and 20kb downstream. Also shown are splice variants and cross-species conservation. Consensus exons of *CDC42* are labelled 1-8. Note the very high conservation at exons 7 and 8.

4.1.3.3 CDC42 in *post mortem* schizophrenia brain

CDC42 promotes dendritic spine growth (Luo *et al* 1996) which has been reported as being reduced in *post mortem* schizophrenia brain samples (Harrison *et al* 1999) and has thus been studied at a neuropathological level in schizophrenia. By using *in situ* hybridisation, Hill *et al* (2006) quantified mRNA from *CDC42* and other RhoGTPase proteins from the prefrontal cortex. These authors identified a significant reduction in CDC42 expression in schizophrenia which was not due to the effects of antipsychotic medication. Furthermore the CDC42 levels were correlated with dendritic spine density in cortical deep layer 3 where spine loss was at its most pronounced. This led the authors to suggest that reduction of CDC42 in this region, in part contributes to either loss of spine generation or loss of spine integrity observed in schizophrenia. A follow-up study of CDC42 effector proteins showed increased expression of CDC42EP3 which is preferentially expressed in deep layer 3 (Ide & Lewis 2010). This imbalance of CDC42 and CDC42EP3 would inhibit the influx of post-synaptic density proteins required for normal spine plasticity and may explain some of the symptoms observed in schizophrenia. Interestingly, mutations introduced into the brain-dominant exon 7 of CDC42 inhibit a post-translational modification known as palmitylation and lead to greatly reduced dendrite density comparable to that observed in schizophrenia brains post mortem (figure 4.3, Kang *et al* 2009).

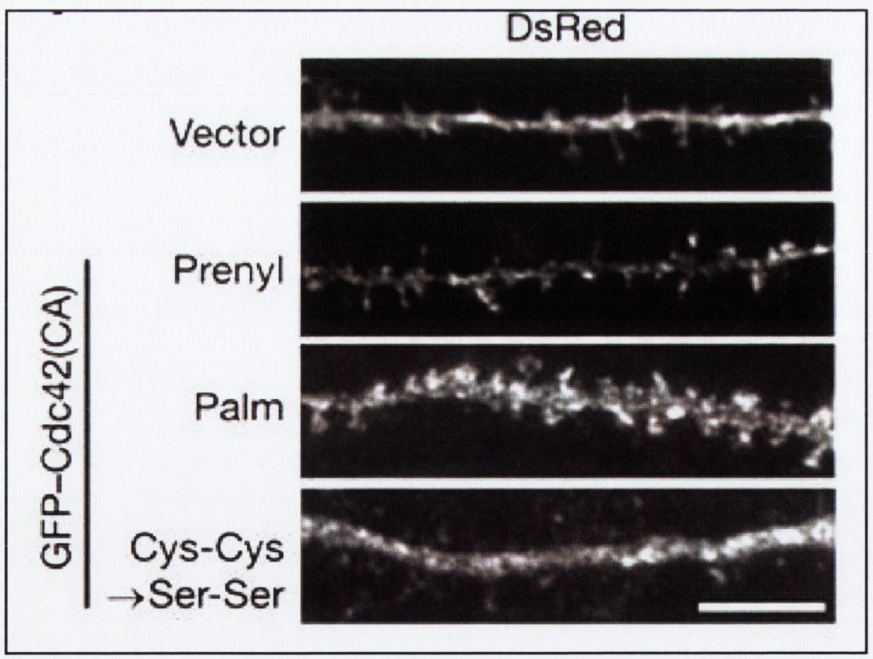


Figure 4.3: Reduction in dendritic spine density with over-expression of GFP-CDC42 harbouring mutations in the brain-dominant exon 7 (Kang *et al* 2009). The mutation prevents palmitylation of CDC42 which is required for dendrite development. Images show electron micrographs of axons from cultured hippocampus neurons and transfected with different vectors. Scale bar is 5μm.

4.1.3.4 CDC42 and aging

In mice expression levels of *CDC42* increase continuously with age throughout the body (Wang *et al* 2007). This is supported by a study of global gene expression in human lymphoblasts which found that increases in *CDC42* expression were correlated higher than any other gene with mortality rate and aging (Kerber *et al* 2009). This process has been elucidated in mice which, lacking the CDC42-regulator CDC42GAP, accumulate DNA damage and exhibit a grossly accelerated aging phenotype (Wang *et al* 2007). Although there is little biological overlap known between schizophrenia and senescence, this is an interesting characteristic of CDC42 biology and may warrant further investigation.

4.1.4 Aims and summary of methods of investigation

The highly-associated SNP at the *CDC42* gene is 17Kbp 5' of the transcription start site. If the association with schizophrenia is driven by a functional effect of the susceptibility allele, it is plausible that this is a *cis*-acting variant that alters expression levels of *CDC42*. Given the extreme effect of *CDC42* deletion on brain development in mice (Chen *et al* 2006) and the observation of decreased *CDC42* expression in schizophrenia brains (Hill *et al* 2006), the susceptibility allele could reduce expression of *CDC42*.

To test this hypothesis, *CDC42* gene expression data in human lymphoblastoid cell lines was obtained from publicly-available sources. Expression levels were tested for association with rs2473277 using PLINK. To obtain experimental support for the observed correlation, a second SNP in high LD with rs2473277 and located within a predicted transcription factor binding site was selected for study using both a gene reporter assay and an electrophoretic mobility shift assay (EMSA).

Given that *CDC42* is such an important gene for the development of the brain it may be possible that even small functional variation has an effect on endophenotypes of schizophrenia. Thus measures of neuro-cognition in controls and schizophrenia cases were compared with genotype at rs2473277. Finally additional case-control samples were investigated for further independent evidence of association of this *CDC42* SNP with schizophrenia.

4.2 Methods

4.2.1 Using PLINK to analyse sub-sample results at *CDC42*

Genotype results from the ISC GWAS were obtained from the consortium web-site (http://pngu.mgh.harvard.edu/isc/). To investigate whether the association signal at rs2473277 was due to a particular sub-sample, each sub-sample was tested individually for association with a standard χ^2 test for association using PLINK (Purcell *et al* 2007). Hardy-Weinberg equilibrium was also tested. All genotype and phenotype information came from the final ISC dataset (ISC 2009).

4.2.2 Investigation of the 'mRNA-by-SNP Browser' database for effects of rs2473277 on *CDC42* expression

The 'mRNA-by-SNP Browser 1.0' database was downloaded and installed from the Michigan University QTL software web-site (http://www.sph.umich.edu/csg/liang/). This database provided access to the results of a genome-wide association study of gene expression in lymphoblasts collected from 206 trios where the offspring had asthma. The database can be queried for results of a particular gene or SNP (Dixon *et al* 2007). Here it was queried for the effects of genotype at rs2473277 on *CDC42* mRNA levels. Expression levels of *CDC42* were determined using six independent probes in the Dixon *et al* study (U133 Plus 2.0 GeneChip, Affymetrix). The effect of a SNP on gene expression was presented as effect (relative magnitude of difference between the two alleles), heritability (H^2), a LOD score and a p-value. An effect was considered genome-wide significant after correction for multiple testing if the LOD score was greater than 6 (Dixon *et al* 2007). To identify the binding-sites of the probes used in the quantification of CDC42 mRNA, sequences were obtained from the Affymetrix web-site (http://www.affymetrix.com/Auth/analysis/downloads/data/) and screened against the human genome using Basic Local Alignment Tool (BLAT) in the UCSC Genome Browser (http://genome.ucsc.edu/cgi-bin/hgBlat). This was also used to check for potential binding to other, non-CDC42 transcripts and polymorphisms under the binding-sites which could affect binding efficiency.

4.2.3 Analysis of CDC42 expression in HapMap samples for effects of rs2473277

A second data set was examined for the effect of rs2473277 on expression of *CDC42*. Genome-wide expression data on the four principle HapMap reference samples (CEU, YRI, CHB and JPT) was generated on the Illumina WG-1, version 1 expression array and made publicly-available through the Genevar project (Stranger *et al* 2007, http://www.sanger.ac.uk/humgen/genevar/). The population samples include 30 Caucasian trios of northern and western European origin (CEU), 45 unrelated Chinese individuals from Beijing University (CHB), 45 unrelated Japanese individuals from Tokyo (JPT), and 30 Yoruba trios from Ibadan, Nigeria (YRI). Each population sample was processed independently by the original authors so normalised expression data were used allowing for cross-population comparison. Expression data generated using probe GI_16357471-A was used initially because it binds the multiple exons of *CDC42* and is thus not likely to be influenced by non-specific binding or genetic variation in splicing. Genotypes of these individuals across the *CDC42* including 100Kb upstream and 20Kb downstream (chr1:22,151,707-22,312,023, hg18) was obtained from HapMap (http://hapmap.ncbi.nlm.nih.gov/).

Expression levels of *CDC42* were tested for association with rs2473277 in the HapMap samples using a standard test for association in PLINK. Here, when the phenotype of an individual is anything other than 0, 1, 2 or 'missing' i.e. contains gene expression values, it is assumed to be a quantitative trait and the analysis method adjusted accordingly. A combined analysis of all unrelated individuals was performed to maximise the power of detecting an effect. Probe sequences used to quantify *CDC42* mRNA were obtained from the Genevar project web-site and checked for binding-sites and confounding effects as described above (section 4.2.2).

4.2.4 Bioinformatic selection of a functional variant

SNP rs2473277 was investigated for linkage disequilibrium (LD) with other SNPs that had not been genotyped in the ISC GWAS and which were potentially functional. LD was inferred with HapMap/CEU genotype data using Haploview. The UCSC Genome Browser (Kent *et al* 2002) and TFSEARCH v1.3 (http://www.cbrc.jp/research/db/TFSEARCH, Heinemeyer *et al* 1998) were used for functional inference of linked SNPs. This approach identified rs2473307 which was in high LD in with rs2473277 (r^2: CEU 1.0, YRI 0.69, CHB 1.0, JPT 1.0) and in a predicted functional site upstream of *CDC42*.

4.2.5 Reporter Gene Assays

4.2.5.1 Background on gene reporter assays

Gene reporter assays are normally used to test the effect of a DNA sequence of interest on gene expression but have been increasingly used to look at the effects of different alleles. They first involve the cloning a DNA sequence of interest into a bacterial plasmid vector. This vector harbours a firefly luciferase gene with an SV40 promoter in front of which the test DNA sequence is inserted (figure 4.4). Cell cultures are transfected with the plasmid construct and the luciferase activity measured by florescence. In this study two test plasmids were created, each harbouring a different allele.

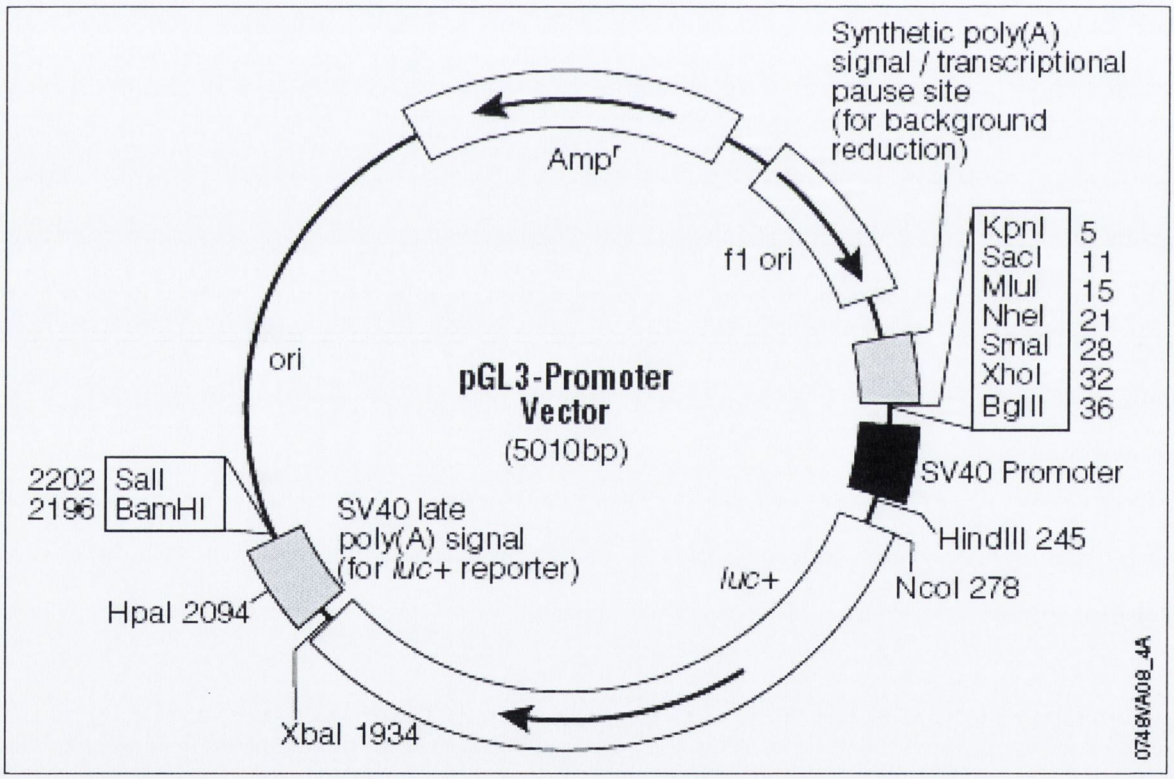


Figure 4.4: Diagram of the pGL3-Promoter vector (Promega, USA) used in this study.

4.2.5.2 Cloning into the the pGL3-promoter vector

Oligonucleotide primers for PCR amplification of rs2473307 were designed on a surrounding 500bp region. The 5' end of each primer harboured restriction sites for *Mlu*I and *Blg*II respectively which allowed for cloning into the pGL3-promoter vector (Invitrogen, UK, figure 4.4) and an additional four-base overhang (to facilitate restriction enzyme digestion). Primers used for the amplification of the region surrounding rs2473307 were CTCTACGCGTTCACGGTCATTCTGGTAGG and CTCTAGATCTCATGATCTGGTTTGCGG (MWG Biotech, Germany). The DNA from two CEU individuals homozygous for each genotype at rs2473307 was selected for cloning. Polymerase chain reactions (PCR) were used to amplify the target sequence from 30ng genomic DNA using 50pmol each primer, 100mmol dNTP, 1X reaction buffer and 1 unit *Taq* polymerase (Qiagen, UK) in a final volume of 50µl. PCR conditions were 95°C for 2 minutes followed by 40 cycles of 95°C for 15 seconds and 65°C for 1 minute + 1 second per cycle. The PCR products and 1µg pGL3-promoter vector were cut using 1µl each of *Mlu*I and *Blg*II according to manufacturer's instructions (New England Biolabs, USA) with 37°C incubation for 3 hours followed by 68°C for 20 minutes to de-activate the restriction enzymes. Digested PCR products were electrophoresed in standard 1% agarose gel stained with ethidium bromide. Single bands of the expected size were carefully excised over UV light using a disposal scalpel.

Purification of DNA from gel blocks was performed using the standard Qiagen methods according to manufacturer's instructions (Qiagen, UK) and the eluted DNA was quantified using a NanoDrop™ 8000 spectrophotometer (Thermo Scientific, Wilmington, DE, USA). Ligations of PCR product and vector at a ratio of 1:1 were performed overnight at 4°C according to manufacturer's instructions (Promega, Wisconsin USA). Approximately 1ng DNA was added to 50µl *E.coli*, strain JM109 (Promega) which were transformed by 1 minute heat-shock at 42°C and grown overnight on agar plates supplemented with antibiotic to which the plasmid confers resistance. Twenty colonies were seeded into 20ml LB broth and grown for a further twelve hours. Plasmid DNA was purified by standard Qiagen methods according to manufacturer's instructions. The presence of the insert was tested by restriction digestion with *Mlu*II and *Blg*I (figure 4.5). Clones containing the insert were sequenced as described previously to test for the presence of the correct allele (Section 3.2.3 for sequencing method). Primers used for sequencing clones to were RVprimer3 CTAGCAAAATAGGCTGTCCC and RVprimer4 GACGATAGTCATGCCCCGCG (MWG Biotech).

Five plasmids were used for the gene reporter assays. The first, pGL3-basic contains the firefly luciferase gene but no promoter, and was used to determine the background level of gene activity. To determine the base-line luciferase activity an 'empty' pGL3-promoter vector was used.

One plasmid for each allele of rs2473307 was used harbouring a 500bp surrounding sequence located immediately 5' to the luciferase gene SV40 promoter. The fifth plasmid encodes *Renilla* luciferase which acts as an internal control when measuring luciferase activity.

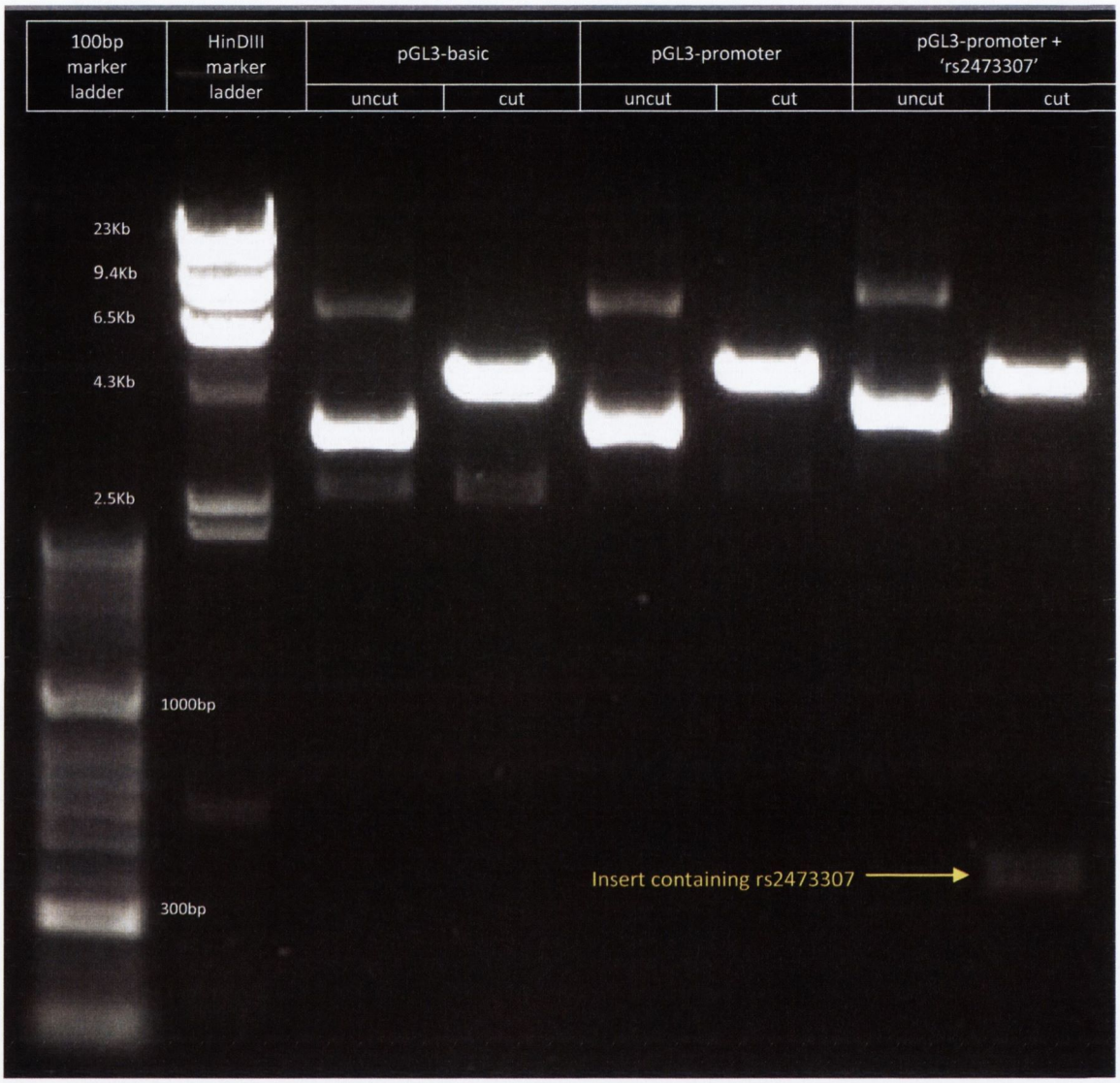


Figure 4.5: Gel electrophoresis of plasmids used in the gene reporter assay (300ng per lane), uncut and cut with MluI and BlgII. Marker ladders were obtained from Bioline, UK.

4.2.5.3 Preparation of cells for gene reporter assay

Two different cell lines were used. SHSY-5Y cells are neuroblasts derived from a four-year old female and are commonly used for neuronal cell functional studies (Biedler *et al* 1973). Chinese hamster

ovarian cells (CHO) are the most commonly used cell-line because of both high growth rate and protein production. Here they serve as a non-neuronal and non-human comparison. Both cell lines were cultured at 37°C and 5% CO_2 in RPMI-1640 media supplemented with 10% foetal bovine serum and 0.4mM/ml L-glutamine (all Sigma-Aldrich, UK). At 60% confluency, cells were plated and transfected with 50nM plasmid for each well, using 'Lipofectamine 2000®' reagent according to manufacturer's instructions (Invitrogen, UK). Six replicate wells were used for each plasmid. SHSY-5Y and CHO cells were transfected with the same plasmid preparations to provide accurate comparison between the cell lines. Each experiment was conducted at least twice. Within experiment and global significance levels in the difference between allelic expression was determined using a two-tailed unpaired Student's t-test.

4.2.5.4 Quantification and comparison of reporter gene activity

Two days following transfection, cells were lysed and the luciferase protein activity determined. This was performed using the Dual-Luciferase Reporter Assay System according to manufacturer's instructions (Promega). Activity of firefly luciferase and *Renilla* luciferase for each well was measured using a Victor™ 1420 multi-label plate-reader (Wallac/Perkin-Elmer, USA). The background relative firefly luciferase activity (as determined from the pGL3-basic vector) was subtracted from the other wells in the experiment. The relative luciferase activity for each of the 'allelic' vectors was expressed relative to that for the 'empty' pGL3-promoter vector.

4.2.6 Electrophoretic Mobility Shift Assays (EMSA)

4.2.6.1 Background on Electrophoretic Mobility Shift Assays

The EMSA has historically been used to explore transcription factor binding-sites (Hellman & Fried 2007). It is being surpassed by array technologies but is still in use for testing the effects of different alleles in upstream regulatory sites on protein binding affinity. The principle is that during gel electropheresis, DNA with proteins bound to it will migrate slower than un-bound DNA. The binding reactions are performed using artificially synthesised double-stranded DNA incubated with protein extracted from the nuclei of the cells of interest and are thus enriched with transcription factors. Prior labelling of the target or probe DNA with biotin allows for chemiluminescent detection. The electrophoretic mobility shift assay has recently been used to show differential binding of nuclear proteins to a schizophrenia-associated SNP in the *NOS1AP* gene (Wratten *et al* 2009).

4.2.6.2 Preparation of nuclear proteins

Nuclear extracts were prepared from cultured SHSY-5Y cells using the NE-PER Nuclear Extraction Kit (Pierce/Thermo Scientific, USA). Protein concentration and purity was determined using a spectrophotometer and compared to commercially-supplied BSA of known concentrations (New England Biolabs, USA).

4.2.6.3 Probe preparation

Complementary biotin-labelled and unlabelled pairs of oligonucleotides corresponding to a 50bp sequence incorporating rs2473307 were synthesised by MWG biotech. Labelled and unlabelled oligonucleotide probes were annealed at a final concentration of 1pmol/ul and 10pmol/ul respectively to form double-stranded DNA. The annealing buffer consisted of 10mM Tris, 1mM EDTA, 50mM NaCl. The annealing temperature conditions were 95°C for 5 minutes followed by a cooling regime of 70, 1 minute cycles with the temperature being reduced by 1°C each cycle. The probe sequence was 5' Bio-GCAGTGGCAATGTTGACAAAGGGCA[T/C]ATCTTCCTTTCCCCAAGAACATTG.

4.2.6.4 Protein-DNA binding reactions

Binding of probe DNA to nuclear proteins was in a final volume of 20ul containing 1X binding buffer, 1ug nuclear proteins, poly.d(I-C), BSA, glycerol. For competition assays, 100-fold molar excess unlabeled probe was added. All binding reactions were pre-incubated for 30 minutes prior to the addition of biotin-labelled probes to reduce non-specific binding. Incubations were carried out at 20°C in a thermal cycler to ensure consistency.

4.2.6.5 Visualisation of protein-DNA binding

DNA and DNA-protein complexes were separated on a non-denaturing 6% acrylamide gel in 0.5X TBE buffer at 150V for approximately 45 minutes. They were then dry-transferred to a nylon membrane at 100V for approximately 45 minutes and cross-linked over UV light for 15 minutes. Membranes were washed and stained with strepavidin which binds to biotin using the Chemiluminescent Nucleic Acid Detection Module (Pierce, USA). Streptavidin fluorescence was visualised at -30°C using a Fuji-LAS CCD camera (Fujifilm, Japan). Comparative mobility of DNA corresponding to each allele was determined visually.

4.2.7 Neuropsychological correlates

Detailed neuropsychological assessment was obtained on 172 control individuals from the RPGI sample (section 3.2.7) and a subset of 506 Irish schizophrenia cases from GASP and RPGI samples (section 3.2.11). SNP rs2473277 (at which the association was initially identified in the ISC GWAS) was genotyped in-house using a pre-designed TaqMan® assay on an ABI PRISM® 7900HT Sequence Detection System (Applied Biosystems, Warrington, UK) as previously described (section 3.2.6). Three domains of cognition were tested (IQ, memory and attention) which relate to deficits commonly observed in schizophrenia (Section 1.4.4). Neuropsychological variables were tested for distribution and grouped by genotype at rs2473277. Significant differences between groups for each variable were identified using either ANCOVA (Analysis of co-variance, for parametric data) or a Kruskal-Wallis test (non-parametric data).

4.2.8 Replication analysis methods in the GAIN and WTCCC2 samples

For the purposes of replication we accessed The Wellcome Trust Case-Control Consortium 2 (WTCCC2) Schizophrenia sample which included 1,606 schizophrenia cases and 1,794 controls sampled in Ireland and genotyped on the Affymetrix 6.0 SNP genotyping array. There is overlap between the subjects in the WTCCC2 study and the Irish subset of the ISC study. To facilitate an independent comparison of data from both studies, the overlap samples were removed from the WTCCC2 study and association analysis was performed on a dataset of 1,310 cases and 1,023 controls. A second replication sample was also available from the Genetic Association Information Network (GAIN) who performed a GWAS on a large schizophrenia case-control sample using the Affymetrix 6.0 SNP genotyping array. The dataset available at the time was 1,172 schizophrenia cases and 1,378 controls. This data is publicly available via dbGaP subject to the permissions detailed in the support documentation for the GAIN project (http://www.ncbi.nlm.nih.gov/gap).

Association statistics in the CDC42 region (chr1:22,151,707-22,312,023, hg18) were generated using PLINK. The power of GAIN and WTCCC2 samples to replicate the ISC association at rs2473277 was calculated using the on-line Genetic Power Calculator (Case-control for discrete traits, http://pngu.mgh.harvard.edu/~purcell/gpc/cc2.html, Purcell *et al* 2003). The relative risk values for homozygous and heterozygous risk allele carriers used in the power calculations were calculated using genotype counts.

4.3 Results

4.3.1 ISC results across the *CDC42* gene

The *CDC42* gene is located at chr1:22,251,707-22,292,023 (hg18). To include potential extragenic regulatory polymorphisms, 100Kb proximal and 20Kb distal of these co-ordinates were considered for association analysis (chr1:22,151,707-22,312,023). Coverage of common SNPs in this region by the Affymetrix 5.0 and 6.0 genotyping arrays was good although not complete (see Table 4.2). Of the uncaptured SNPs, six were in high LD (r2>0.75), relatively rare (MAF ~0.1) and formed a 92Kb haplotype across the ELA3A and ELA3B genes 5' of *CDC42*.

Table 4.2: Coverage of the *CDC42* gene in the ISC GWAS

Affymetrix platform	Assayed SNPs	Common CEU SNPs[1]	Captured at r^2>0.8[2]	Captured at r^2>0.6[2]
5.0	20	63	43 (68%)	52 (82%)
6.0	40	63	48 (76%)	52 (82%)

[1]Common CEU SNPs in this instance have a MAF>0.1, with <1 Mendelian error, greater than 75% call rate and Hardy-Weinberg p-value >0.001. [2]Number and percentage of common SNPs captured was determined using Haploview 4.2 using aggresive 2-and 3-haplotype marker tagging.

In the combined analysis, there was a significant association with disease status at ten SNPs which were all located in non-coding regions across the gene (α=0.05, Table 4.3). The highest level of association by lowest p-value was at rs2473277 which is 17Kb from transcription start of *CDC42*. At a low LD threshold (r^2>0.5) there were six independent association signals. A diagram of the LD between these SNPs in the CEU sample is displayed in Figure 4.6.

Table 4.3: Significantly associated SNPs at *CDC42* in the ISC GWAS

SNP	Position (bp)[1]	Region	Case MAF	Control MAF	p-value	OR	95% C.I.	LD block[2]
rs2473277	22,234,432	5'	0.458	0.493	2.00×10^{-6}	0.845	0.79-0.91	-
rs2501276	22,246,211	5'	0.103	0.086	0.0077	1.192	1.05-1.36	1
rs17837965	22,267,212	Intron 1	0.046	0.055	0.0067	0.806	0.69-0.94	-
rs2473317	22,267,838	Intron 1	0.138	0.124	0.0313	1.135	1.01-1.27	1
rs10917139	22,274,125	Intron 2	0.131	0.149	8.35×10^{-4}	0.869	0.76-0.93	2
rs2056974	22,281,681	Intron 4	0.230	0.257	6.13×10^{-4}	0.868	0.80-0.94	2
rs10917145	22,285,232	Intron 4	0.131	0.149	5.41×10^{-4}	0.840	0.76-0.93	2
rs10917152	22,298,229	3'	0.126	0.141	0.0030	0.858	0.77-0.95	2
rs16826595	22,298,449	3'	0.014	0.018	0.0342	0.663	0.45-0.97	-
rs2092322	22,308,310	3'	0.455	0.483	2.39×10^{-4}	0.878	0.82-0.94	-

[1]Position is from hg18/UCSC Genome Browser Build 36.1, March 2007.

Region analysed is chr1:22,151,707-22,312,023, hg18

[2]LD was determined from the CEU sample, $r^2>0.5$.

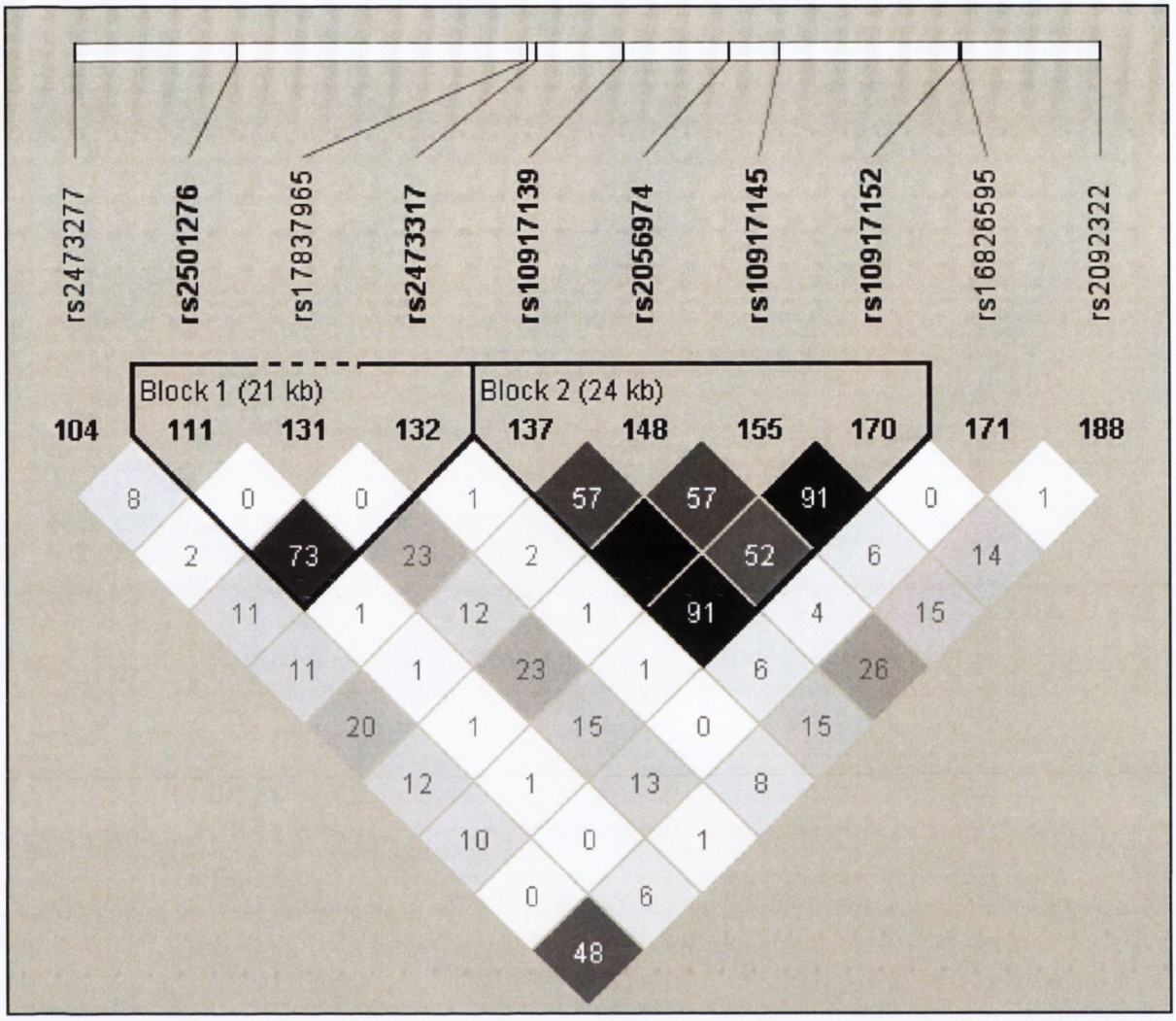


Figure 4.6: Linkage disequilibrium between significantly associated SNPs in the combined ISC analysis. The r^2 values are based upon CEU genotype data. Numbers below the SNP identifiers refer to HapMap SNPs across the region studied (chr1:22,151,707-22,312,023, hg18).

4.3.2 ISC results at rs2473277

The population and association statistics at rs2473277 were investigated in the sub-samples used in the combined ISC GWAS sample (See Table 4.4). The control samples had a 'G' allele frequency in the range of 0.476-0.559. All case and control sub-samples were in Hardy-Weinberg equilibrium at rs2473277 with the exception of Aberdeen controls (n=696, p=0.028). A significant association with schizophrenia was identified at rs2473277 in samples from Aberdeen, Dublin, the Portuguese Islands and London and in each case the association was in the same direction (OR<1, A is the risk allele).

Table 4.4: Association analysis of rs2473277 in individual ISC samples[1]

Study site	Sample origin	Sample size		MAF[2]		HWE p-value		p-value	OR	95% C.I.
		Cases	Controls	Cases	Controls	Cases	Controls			
Aberdeen	Aberdeen	727	696	0.441	0.482	0.547	0.028	***0.027***	0.857	0.74-0.99
Cardiff	Bulgaria	482	647	0.464	0.481	0.715	0.431	0.422	0.923	0.78-1.09
Dublin	Dublin	268	874	0.418	0.476	0.900	0.378	***0.018***	0.782	0.64-0.95
Edinburgh	Edinburgh	388	283	0.485	0.498	0.186	0.191	0.620	0.954	0.77-1.19
Portugal	Portuguese Islands	339	202	0.419	0.512	1.000	0.575	***0.003***	0.675	0.52-0.87
Sweden I	Sweden	203	195	0.515	0.559	0.779	0.772	0.211	0.838	0.63-1.11
Sweden II	Sweden	425	242	0.514	0.529	0.628	0.248	0.603	0.938	0.75-1.18
UCL	London	548	454	0.439	0.502	0.298	0.398	***0.005***	0.778	0.65-0.93
Combined		3,380	3,593	0.458	0.493	0.944	0.367	***2.00×10^{-6}***	0.845	0.79-0.91

[1] The p-value and odds ratios presented here were calculated using a standard χ^2 trend test for association.

[2] The minor allele here is the 'G' allele and thus the 'A' allele is the susceptibility/risk allele in the combined sample.

4.3.3 Effect of rs2473277 on *CDC42* expression in the Dixon *et al* study

The 'mRNA-by-SNP Browser 1.0' displayed results of six different probes from the U133-Plus 2.0 expression array (Affymetrix) to quantify *CDC42* expression levels (Table 4.5). The authors considered a LOD score >2 as a nominal level of association and a LOD score >6 as significant genome-wide after correction for multiple testing. Expression levels of *CDC42* as determined using probe 208727_s_at were significantly reduced in the presence of the 'A' allele (LOD 14.28). Additionally, probe 208728_s_at showed a nominally significant reduction in expression in the presence of the 'A' allele (LOD 3.12). As the 'A' allele is associated with schizophrenia, these results agree with previous reports that *CDC42* expression levels are known to be reduced in schizophrenia (Hill *et al* 2006). However results were not consistent for all probes used in the microarray experiment. Conversely, *CDC42* expression as determined using probe 226400_at showed a nominally-significant increase in the presence of the 'A' allele (LOD 5.77). None of the results from three other probes were significant.

Table 4.5: Effect of rs2473277 on *CDC42* gene expression from the Dixon *et al* study

Probe ID	Binding sites	Allele	Effect	H^2	LOD	p-value
1556931_at	Intron 1	-	-	-	-	NS
208727_s_at	Exon 8	A	-0.609	18.42	14.28	5.1×10^{-16}
208728_s_at	3'UTR (exon 8)	A	-0.283	3.99	3.12	0.00015
210232_at	Exon 7	-	-	-	-	NS
214230_at	3' UTR (exon 7)	-	-	-	-	NS
226400_at	3' UTR (exon 8)	A	0.386	7.41	5.77	2.5×10^{-7}
Average	-	-	-	-	-	NS

Because the results for the impact of rs2473277 alleles on *CDC42* expression were not definitive, the probe binding-sites were investigated. Probe sequences were obtained from the Affymetrix web-site (http://www.affymetrix.com/Auth/analysis/downloads/data/) and screened against the human genome using Basic Local Alignment Tool (BLAT) in the UCSC Genome Browser (figure 4.7).

The three probes which identified an effect of rs2473277 on *CDC42* expression, all bound to exon 8 or the 3'UTR of exon 8 which is known to be expressed at high levels in lymphoblasts. Probes 208727_s_at and 226400_at (which gave contradicting results) have 98% homology to coding sequence and UTR of a hypothetical gene, NM_001089595 which is located within an intron of GBA3 and thus the results at CDC42 could potentially be affected by non-specific binding here. Probe

208728_s_at, which provided evidence of reduced expression with the 'A' allele was not found to bind to any other coding or transcribed sequence. Although 32 SNPs have been reported in exon 8 and its associated 3'UTR, only one is common and is not situated in a probe binding-site (rs1064240). The remainder have a low MAF in the Asian population in which they were identified or have no frequency information. Furthermore none of them are reported in HapMap. This information indicates that common variation at the probe binding-sites is unlikely to be cause of the discrepancy in direction of effect of the 'A' allele at rs2474277.

The probes which did not provide significant evidence of effect of rs2473277 either bound to unconfirmed coding sequence in intron 1 (1556931_at) or to exon 7 (210232_at, 214230_at) which is expressed at low levels in lymphoblasts (Nicole *et al* 1999).

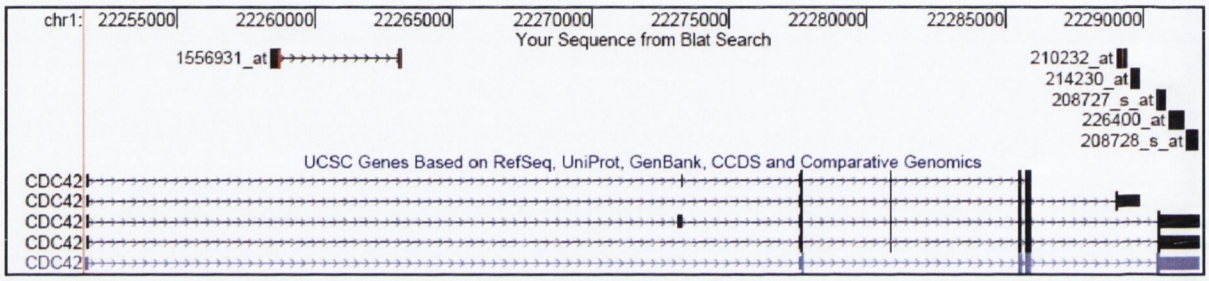


Figure 4.7. Diagram of the *CDC42* gene showing binding-sites for the difference probes on the U133 2.0 Plus expression array.

4.3.4 Effect of rs2473277 on CDC42 expression in HapMap samples using Genevar data

The impact of rs2473277 on expression levels of *CDC42* was studied in 210 unrelated adults from four HapMap samples. Expression results from GI_16357471-A which binds to both common isoforms of *CDC42* (see figure 4.9, Nicole *et al* 1999) were considered to be best able to detect a *cis*-acting affect. The mean relative expression level (± standard deviation) of individuals carrying two copies of the 'A' allele was 10.19±0.30 (n=63) compared with 10.40±0.27 for heterozygotes (n=96) and 10.46±0.30 for 'GG' homozygotes (n=50) (figure 4.8).

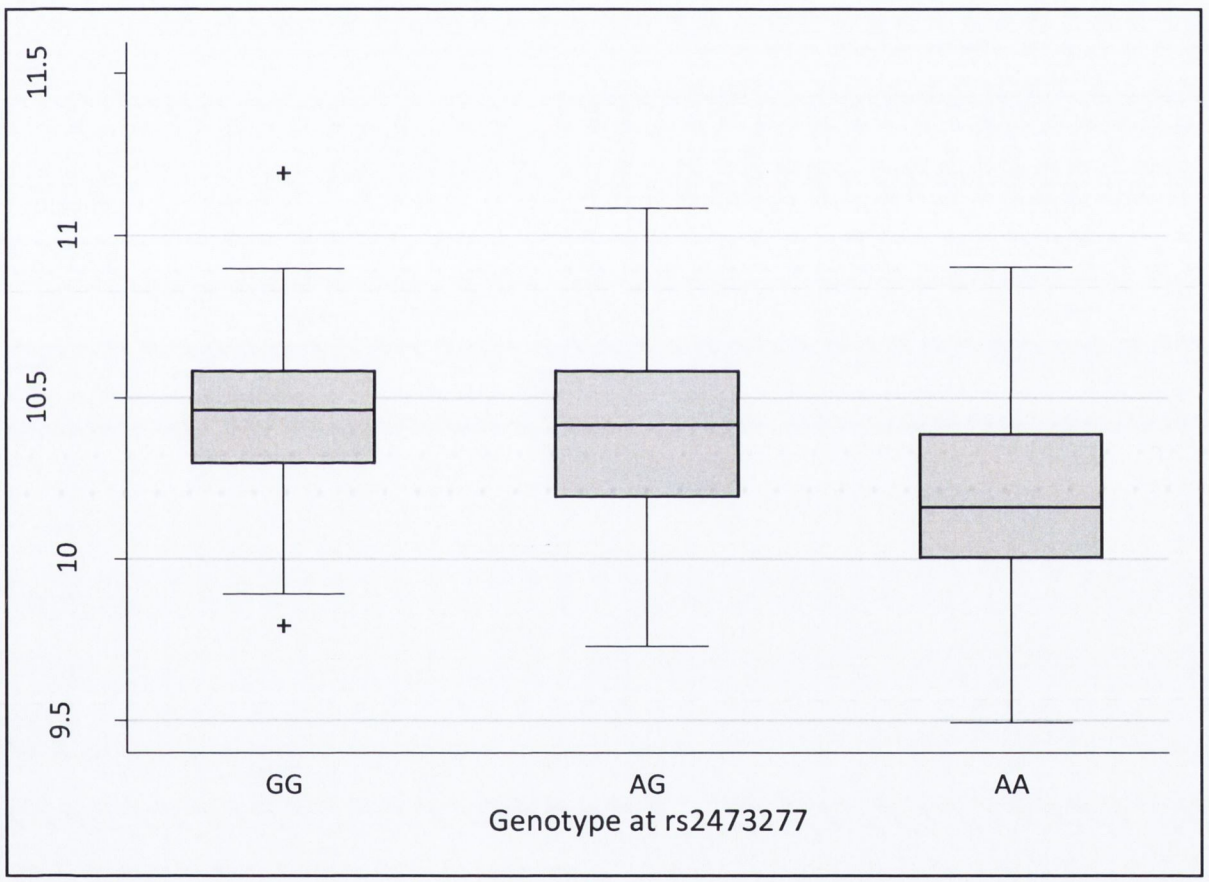


Figure 4.8: Box-plot showing level of *CDC42* mRNA in 210 unrelated adults grouped by genotype at rs2473277. Boxes indicate 25% and 50% percentiles, separated by the median. Whiskers indicate data range, excluding outliers (indicated by '+') which are >2.5 SD from the mean. Outliers were included in the quantitative association analysis.

Expression levels of *CDC42* as determined by probe GI_16357471-A were tested against rs2473277 using the quantitative trait association function in PLINK. A significant reduction in *CDC42* expression in the presence of the 'A' allele was identified indicating that rs2473277 is a quantitative trait locus for *CDC42* expression (β=0.13, SE=0.02, r^2=0.12, T_{Wald}=5.24, p=3.85×10^{-7}). Expression levels of *CDC42* were also determined in the HapMap samples using two other probes: GI_16357471-I which binds to the 3'UTR of exon 7 which is predominant expressed in the brain, and GI_16357470-I which binds to the 3'UTR of exon 8 which is expressed ubiquitously (figure 4.9). Quantitative association analysis of CDC42 expression determined by these probes did not show a significant effect of rs2473277 (Table 4.6). Although only GI_16357471-A provided a significant result, the direction of effect of rs2473277 was the same in all three instances.

Table 4.6: Effect of rs2473277 on *CDC42* gene expression in 210 unrelated HapMap adults

Probe ID	Binding sites	β	SE	r^2	T_{Wald}	p-value
GI_16357471-A	Exon 6	0.139	0.02	0.117	5.24	3.85×10^{-7}
GI_16357471-I	3'UTR (Exon 7)	0.035	0.03	0.004	0.94	0.34
GI_16357470-I	3' UTR (Exon 8)	0.007	0.02	0.0005	0.32	0.74

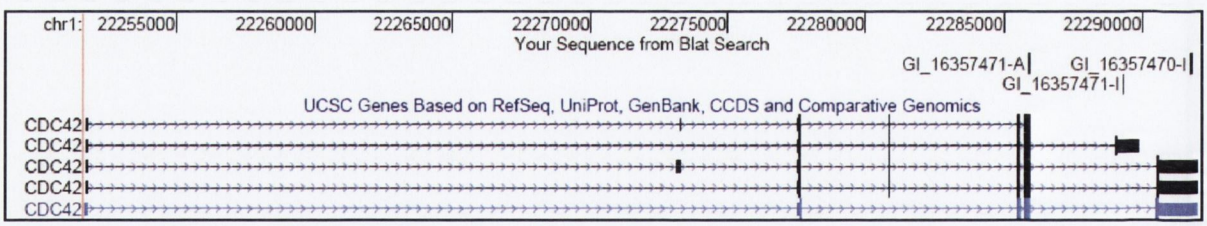


Figure 4.9: Diagram of the *CDC42* gene showing binding-sites for the difference probes on the Illumina WG-1, version 1 expression array

4.3.5 Identification of a putative functional SNP in high LD with rs2473277

Although rs2473277 is in the 5' region of *CDC42* and found to be correlated with expression levels of *CDC42* in the Dixon *et al* study and in analysis of the Genvar data, it was not in a predicted functional region of gene regulation. The coverage of common variants at *CDC42* in the ISC GWAS was high so any causative and functional SNP was likely to be in very high LD with SNPs analyzed in this region. SNP rs2473277 captured three SNPs at r^2>0.5 which had not been genotyped in the ISC GWAS: rs2501279 (0.60), rs1534949 (0.63), rs2473307 (0.93). SNP rs2473307 is a common C/T polymorphism (CEU MAF 0.49) in which the 'C' allele is on the same haplotype as the 'A' allele at rs2473277 which was associated with schizophrenia in the ISC GWAS (figure 4.10). A 500bp sequence centred on rs2473307 was screened for differences in predicted transcription factor binding sites between 'C' and 'T' alleles using the TFSEARCH tool. The sequence harbouring the 'T' allele had predicted binding sites for GATA-X, GATA-2 or c-Ets transcription factors on the sequence 5'-'AGCAGTGGCAATGTTGACAAAGGGCA[T]ATCTTCCTTTCCCCAAGAACATT'. These were not predicted for the sequence harbouring the 'C' allele.

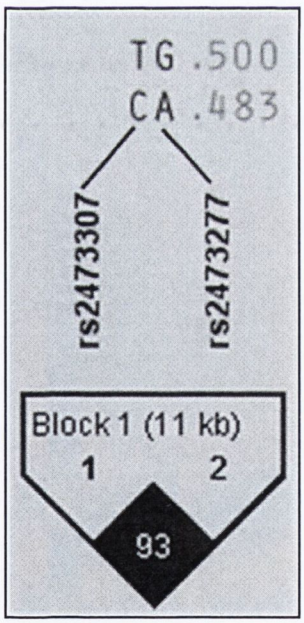


Figure 4.10: Linkage disequilibrium and haplotype relationship between rs2473277 and rs2473307 in the CEU sample.

4.3.6 Gene reporter assay results

A gene reporter system was used to test and compare the effects of rs2473307 on gene expression in both human neuronal (SHSY-5Y) and non-human, non-neuronal (CHO) cells. The expression levels of the firefly luciferase reporter gene in the pGL3-promoter vector were determined for each allele relative to the *Renilla* vector and the vector alone. Each experiment was performed four times in double triplicate. The presented results are the means of the four experiments.

In the SHSY-5Y cells, relative luciferase activity of the construct harbouring the 'T' allele was significantly greater than that of harbouring the 'C' allele (see figure 4.8). Overall, the activity of the promoter vector harbouring the 'T' allele was 1.53 times greater than the vector alone (SD 0.18) compared to the same vector with the 'C' allele which only 1.07 times greater (SD 0.21) (figure 4.9). Thus the effect of the 'T' allele on upstream gene expression in this system conferred a significant 64% increase in comparison to the 'C' allele ($p=0.03$). In the CHO cells, the relative luciferase assay of the 'T' and 'C' alleles was not significantly different ('T' allele 0.929±0.079 vs 'C' allele 0.916±0.206, $p=0.91$), see figure 4.9).

These results indicate that the 'C' allele at rs2473307 confers a loss of transcriptional activity in human neuronal SHSY-5Y cells but not in non-human, non-neuronal CHO cells. This is in agreement with the correlation studies of *CDC42* gene expression and SNP genotype in terms of the direction of effect. Notably though, the association of rs2473307 with *CDC42* expression level was identified in human lymphoblastoid cell lines.

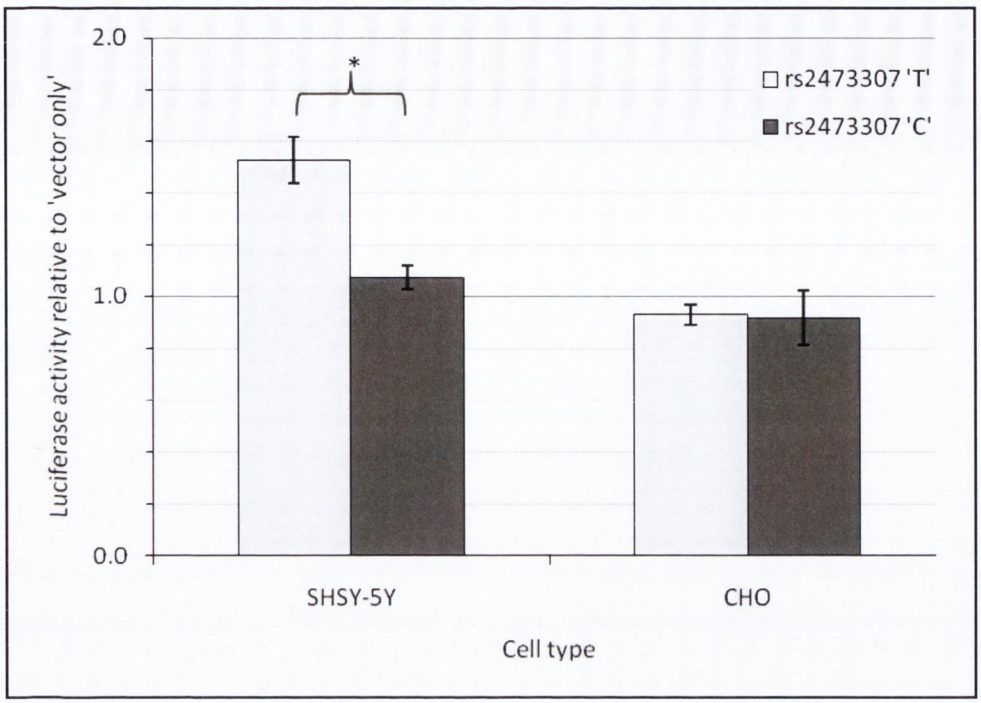


Figure 4.12: Reporter gene activity in SHSY-5Y and CHO cells transfected with pGL3-promoter vector harbouring an insert with either the 'T' or 'C' allele at rs2473307. Error bars show standard deviation. '*' indicates $p<0.05$.

4.3.7 EMSA results

An electrophoretic mobility shift assay was performed on the two alleles of rs2473307 in order to test whether the observed difference in reporter gene expression was caused by differential binding of transcription factors. Evidence of binding of protein to rs2473307 was indicated by a faint band of DNA which migrated slower than unbound DNA. This was more pronounced in the presence of the 'T' allele than the 'C' allele (figure 4.13, lanes 5 and 10). However this band was still visible with 1000-fold excess of unlabelled probe (lanes 6 and 11). In an attempt to determine the origin of this band, addition reactions were also tested which contained nuclear protein extract and unlabelled probe (lanes 7 and 12) or labelled and unlabelled probes (without protein) (lanes 8 and 13). Faint, slow-migrating bands were observed for the 'C' allele reaction with protein extract and unlabelled probe only. Alterations of probe and protein concentrations in the binding reactions did not lead to any visible binding of proteins to either allele.Thus, whilst there was some evidence of protein binding, it could not be confirmed that this is due to specific binding of transcription factors to rs2473307, nor was a difference between the two alleles identified.

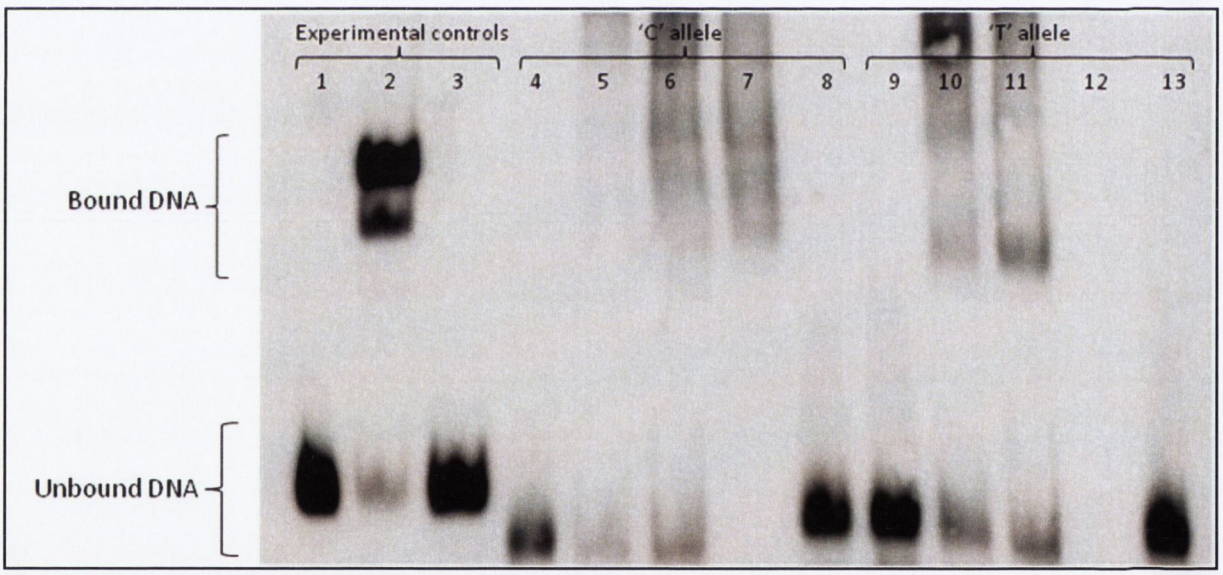


Figure 4.13: EMSA for rs2473307 incubated with 1µg SHSY-5Y nuclear extract. Experimental control reactions used Epstein-Barr virus antigen and corresponding target DNA.

Lanes 1, 4 and 9: No protein, labelled DNA only.

Lanes 2, 5 and 10: Protein with labelled DNA.

Lanes 3, 6 and 11: Protein with labelled and unlabelled DNA.

Lanes 7 and 12: Protein and unlabelled probe (no labelled probe).

Lanes 8 and 13: Labelled and unlabelled probes (no protein).

4.3.8 Neuropsychological results

Genotype at SNP rs2473277 was tested for effects on fifteen neuropsychological variables covering general cognitive ability (IQ), memory and attention. Not all individuals used in the analysis completed all the tasks and thus each statistical test is based upon slightly different sample sizes (mean 451, range 225-555). Significant effects were identified in Working Memory and Episodic Memory ($p<0.05$, table 4.7).

In the verbal working memory task individuals with the 'protective' GG genotype performed worse than (mean rank 239.35) those with the risk genotype AA (mean rank 261.79). However individuals with the heterozygous genotype performed better than both other groups (mean rank 288.06) indicating a possible heterozygous advantage. Neither of the two indices for spatial working memory was significantly affected by genotype at rs2473277.

Episodic memory measured by paired-associates learning and adjusted following the six-shapes method was also affected by genotype at rs2473277. Again, individuals with the heterozygous genotype performed significantly better in the other two groups (mean rank 219.42). However, in this instance individuals homozygous for the risk genotype performed worse than those with the protective homozygous genotype (mean ranks, 185.79 vs 211.28). Genotype at rs2473277 was not found to have another other significant effect on any other indices of episodic memory as tested through any of the logical memory, faces recognition or paired-associates learning methods.

Table 4.7: Comparison of neuropsychological variables with genotype at rs2473277

Task	Sample size	AA[1]	AG[1]	GG[1]	p-value
General cognitive ability					
Verbal IQ[2]	406	103.1	103.1	102.4	0.191
Performance IQ	416	204.0	219.4	183.4	0.077
Full-scale IQ	478	237.9	245.7	223.0	0.425
Working memory					
Verbal working memory (Letter-number sequencing)	542	261.79	288.06	239.35	***0.019***
Spatial working memory (Errors)	531	257.07	265.99	280.22	0.508
Spatial working memory (Strategy)	530	267.18	268.72	253.57	0.694
Episodic memory					
Logical memory 1	555	278.48	275.88	283.53	0.919
Logical memory 2	551	275.27	273.87	283.53	0.871
Faces recognition 1	530	274.66	257.31	274.52	0.426
Faces recognition 2	528	263.60	262.68	271.38	0.838
Paired associates learning: stages completed	510	244.20	263.48	251.84	0.307
Paired associates learning: total errors (adjusted) standard score	523	253.67	269.99	252.46	0.446
Paired associates learning: standard score (six shapes adjusted)	415	185.79	219.42	211.28	***0.044***
Attention					
D' sensitivity score (two-number condition)	269	138.16	129.92	144.52	0.480
D' sensitivity score (three-number condition)	257	133.42	125.30	133.10	0.691
D' sensitivity score (four-number condition)	255	130.48	127.36	126.20	0.943

1. The values presented for each genotype group are the group means. Values for all IQ tests are measured in IQ units. Values for all other tests are the mean of the 'rank' score for each individual.
2. Verbal IQ was tested using an ANCOVA. All other tests were performed with a Kruskal-Wallis test.

4.3.9 Replication results in the GAIN and WTCCC samples

Replication of the significant association signal at rs2473277 was sought in two large, independent samples: GAIN and WTCCC2. However there was no evidence of association at rs2473277 with schizophrenia in these two samples (Table 4.8). Notably, case MAF in both replication samples was similar to that in the ISC sample, whereas the control MAF was lower than in the ISC sample. Power analyses using calculated values for genotype relative risk indicated that the replication GAIN and WTCCC2 samples had 25% and 23% power respectively to replicate the association identified in the ISC study. The remaining SNPs which were significantly associated in the ISC study were also tested for replication in the GAIN and WTCCC2 studies. Here, rs17837965, located in intron 1 of *CDC42* was associated both in the ISC sample (p=0.0067, OR 0.81) and in the GAIN sample (p=0.0031, OR 0.77). In addition rs2092322 was associated in the ISC sample ($p=2.39\times10^{-4}$, OR 0.88) and in the WTCCC sample (p=0.044, OR 0.89) (table 4.9).

Table 4.8: Association results for rs2473277 in ISC, GAIN and WTCCC2 schizophrenia samples

Study	Platform	Cases	Controls	Case MAF	Control MAF	χ^2	p-value	OR	Replication power
ISC	Affymetrix 5.0 and 6.0	3,380	3,593	0.458	0.493	22.59	2.00×10^{-6}	0.845	-
GAIN	Affymetrix 6.0	1,172	1,378	0.464	0.464	4.8×10^{-5}	0.9944	1.00	0.25
WTCCC2	Affymetrix 6.0	1,310	1,023	0.447	0.461	0.908	0.34	0.945	0.23

Table 4.9 Replication analysis of associated ISC GWAS SNPs in the GAIN and WTCCC samples for *CDC42*

SNP	Region	LD block[1]	ISC				GAIN				WTCCC			
			Case MAF	Control MAF	p-value	OR	Case MAF	Control MAF	p-value	OR	Case MAF	Control MAF	p-value	OR
rs2473277	5′	-	0.458	0.493	2.00×10^{-6}	0.85	0.464	0.464	0.994	1.00	0.447	0.461	0.341	0.95
rs2501276	5′	1	0.103	0.086	0.0077	1.19	0.098	0.101	0.737	0.97	0.109	0.097	0.182	1.14
rs17837965	Intron 1	-	0.046	0.055	0.0067	0.81	0.051	0.065	***0.031***	0.77	0.045	0.043	0.674	1.06
rs2473317	Intron 1	1	0.138	0.124	0.0313	1.14	0.147	0.139	0.428	1.07	0.152	0.136	0.118	1.14
rs10917139	Intron 2	2	0.131	0.149	8.35×10^{-4}	0.87	0.138	0.142	0.667	0.97	0.126	0.127	0.874	0.99
rs2056974	Intron 4	2	0.230	0.257	6.13×10^{-4}	0.87	0.235	0.244	0.426	0.92	0.231	0.243	0.311	0.93
rs10917145	Intron 4	2	0.131	0.149	5.41×10^{-4}	0.84	0.136	0.145	0.336	0.92	0.126	0.127	0.938	0.99
rs10917152	3′	2	0.126	0.141	0.0030	0.86	0.127	0.138	0.278	0.91	0.120	0.123	0.790	0.98
rs16826595	3′	-	0.014	0.018	0.0342	0.66	-[2]	-	-	-	-[2]	-	-	-
rs2092322	3′	-	0.455	0.483	2.39×10^{-4}	0.88	0.457	0.467	0.46	0.96	0.451	0.480	***0.044***	0.89

[1]LD was determined from the CEU sample, r^2>0.5.

[2]Association data was not available for rs16826595 in the GAIN or WTCCC study. The best genotyped proxy SNP in both of these studies was rs17837965, CEU r^2=0.14

4.4 Discussion

The primary reason for investigating the role of common genetic variation at *CDC42* in susceptibility to schizophrenia was due to its interaction with HOMER2 at which association with schizophrenia was previously identified (Chapter 3). Binding of HOMER2 and CDC42 proteins in developing dendritic spines prevents further elongation but promotes widening thereby increasing the area for glutamate receptors (Shiraishi-Yamaguchi *et al* 2009).

The ISC GWAS was investigated for evidence of association at interactors of HOMER2. The strongest association signal from this group was identified at a SNP 15Kb from transcription start of *CDC42* (at rs2473277, $p=2.00\times10^{-6}$, OR 0.85). This was the eighth strongest LD-independent association signal (by lowest p-value) in the ISC GWAS and the strongest signal of any HOMER2 interactant. The association p-value would not be significant using a Bonferroni correction for multiple testing. However, this considers that all SNPs tested in the ISC GWAS are independent and also the sample size is still relatively under-powered.

In addition to its interaction with HOMER2 and strong association with schizophrenia in the ISC GWAS, CDC42 is functionally highly-relevant because it promotes dendritic spine formation. In schizophrenia, the commonly observed reduction in spine density (Harrison 1999 for review) is correlated with reduction in CDC42 expression (Hill *et al* 2007). In a resting dendritic spine CDC42EP3 binds septin to form a matrix which inhibits entry of post-synaptic molecules such as Homer and Shank proteins. Upon stimulation, CDC42 sequesters CDC42EP3 and the matrix disintegrates to allow entry of the post-synaptic molecules required for further activity (Joberty *et al* 1999, Kinoshita *et al* 2002). Based upon this pathway, Ide & Lewis (2010) have proposed that the reduction in dendritic spine density observed in schizophrenia is due partially to a deficit of CDC42 and subsequently the incomplete entry of post-synaptic proteins to the spine head.

Of further relevance to schizophrenia is the interactome proximity of CDC42 to DISC1. Here, normal neurite outgrowth is stimulated by binding of DISC1 to NUDEL (Ozeki *et al* 2003). However NUDEL may also bind CDC42GAP which leads to the release of CDC42 and thus also promotes dendritic spine growth (Shen *et al* 2008). Although both DISC1 and CDC42 promote spine growth they do so via different mechanisms and thus functional variants in each might produce subtly different phenotypes.

Common variants in distant 5' regulatory sites often affect gene expression levels (Epstein 2009, Webster *et al* 2009) and furthermore expression of *CDC42* is reduced in schizophrenia brain

(Hill *et al* 2006). Thus the effect of rs2473277 on *CDC42* expression was investigated using two publicly-available gene expression datasets and then, following a positive result, investigated using two experimental paradigms of gene expression regulation.

In the first dataset, the results of a genome-wide linkage study of gene expression in lymphoblastoid cell lines (Dixon *et al* 2008)showed a significant reduction of *CDC42* expression levels in the presence of the schizophrenia risk allele at rs2473277 (A). This was identified in the only microarray probe which bound unambiguously to coding sequence of *CDC42* (208728_s_at). Furthermore, results from a second probe (208727_s_at) at rs2473277 provided a LOD of 14.28 and a p-value of 5.1×10^{-16} which is one of strongest SNP effects on any gene in the study by Dixon *et al* (2008). Results from other probes differed in magnitude and direction of effect although when the average of all seven probes was taken, the result for rs2473277 was reduced *CDC42* expression with the susceptibility allele.

A second test of the relationship between rs2473277 and *CDC42* expression was performed using a quantitative trait association analysis. Here, the risk 'A' allele was also significantly associated with decreased CDC42 expression. Expression in individuals with an 'A/G' genotype was similar to those with a 'G/G' genotype indicating a possible recessive effect on *CDC42* expression. However using a recessive model for the association analysis of the ISC GWAS data at rs2473277 did not indicate such an effect on schizophrenia susceptibility ($p=2.1\times10^{-4}$ compared to standard allelic test, $p=4.00\times10^{-5}$ and $p=3.00\times10^{-6}$ for Cochrane-Mantel-Hanzel test for allelic association).

Both of these two studies of rs2473277 and *CDC42* expression showed a reduction with the risk 'A' allele which was in agreement with *post mortem* observations in schizophrenia brain tissue (Hill *et al* 2006). Both studies used expression data from different microarray platforms and different individuals but quantified *CDC42* mRNA in the same tissue type (lymphoblastoid cell lines). Validation of this result using quantitative PCR in a large number of brain tissue samples would further strengthen this finding.

The biological mechanism by which rs2473277 might alter *CDC42* expression was investigated. This SNP is not located in any known or predicted regulatory site in which the allelic status might alter binding of transcription factors, enhancers or insulator proteins. A scan of local variation identified a SNP in high LD, rs2473307, which was more likely to be functional. SNP rs2473307 is 10.7Kbp proximal of rs2473277 and 28.1Kbp proximal of the transcription start of CDC42. The r^2 between these two SNPs in the CEU is 0.93.

The risk 'C' allele at rs2473307 removes a predicted binding site for GATA-X, GATA-2 and c-Ets transcription factors which are all expressed in human neurons. The effect of this SNP on gene expression was investigated using a gene reporter assay and an EMSA. The gene reporter assay in

the neuronal SHSY-5Y cell line indicated that the protective 'T' allele significantly enhanced gene expression whereas the susceptibility 'C' allele had no effect. In the CHO cell line there was no difference in activity between the two alleles nor was the activity significantly greater than that of the vector alone. A classical explanation for this difference in result between the two cell lines might be that the effect on gene expression is tissue-specific. Alternatively, given that the CHO cell line is derived from a hamster, the effect might be specific to humans or primates. Indeed, althought the 100bp sequence either side of rs2473307 is conserved in primates, it does not exist anywhere in the genome of rodents.

There are however, two potential caveats to consider before this interpretation can be finalised. Firstly, the initial correlations of rs2473277 with *CDC42* expression were identified in non-neuronal, lymphoblastoid cell lines. This reduces the likelihood that the effect is tissue specific. Secondly, the overall protein synthesis rate in CHO cells is known to be far greater than in the SHSY-5Y cells. This could mean that an allelic effect on reporter gene expression would be more defined in the CHO cells and that the SHSY-5Y cells would be more sensitive to confounding effects such as the concentration and purity of the transfected plasmids. Thus, further studies using the firefly luciferase gene reporter assays should use cell lines of similar activity and independent experiments should use plasmid constructs which have been independently cloned and purified.

An EMSA was used to investigate whether the difference in reporter gene activity between the two alleles of rs2473307 was due to differential binding of transcription factors. Initial experiments showed increased binding of a nuclear protein to the 'T' allele although this also occurred in the presence of 100-fold excess of unlabelled DNA probe indicating that the effect may have been non-specific. Subsequent experiments did not show any protein binding to probes for either allele. If transcription factors such as those predicted bioinformatically (GATA-X, GATA-2 or c-Ets) do bind to rs2473307 it is possible that binding was not observed for a number of reasons. Firstly, the 50-mer probe may not harbour all the necessary recognition and binding sites for a transcription factor. Secondly, the EMSA protein-DNA binding reaction was performed at room temperature (22°C, according to manufacture's instructions and typically for all EMSAs) and thus may not be sufficient for protein binding which normally occurs at 37°C. Thirdly, the purified protein extracts are sensitive to degradation. Whilst they were stored at minus 80°C and without repeated freeze-thawing, it remains possible that those proteins which bind to rs2473307 significantly degraded through the process of purification or in the early stages of the EMSA prior to the transfer from gel to nylon membrane.

Correlations of rs2473277 with a battery of neuropsychological indices were also investigated under the endophenotype hypothesis of schizophrenia. These identified two significant

effects, on verbal working memory and on paired-association learning, a paradigm of episodic memory. In both indices, heterozygous individuals performed the best. Given the allelic association of rs2473277 with *CDC42* expression levels, it may be possible that an intermediate level of expression is required for optimum episodic and verbal working memory as mediated through regulation of dendritic spine morphology. In verbal working memory individuals with the protective GG genotype performed worse than those with the risk AA genotype. This was the opposite direction of effect for episodic memory. Thus, the present study indicates that rs2473277 may confer susceptibility to an endophenotype of schizophrenia with enhanced verbal working memory, possibly at the expense of reduced episodic memory. Given the number of tests that were performed, these results would require replication in a larger sample before any definitive conclusions can be made.

The functional and neuropsychological correlates of rs2473277 with CDC42 activity have, in summary, identified a modest effect. Futhermore this was only partially supported in the gene reporter and EMSA study of variation at rs2473307. The results from the genome-wide linkage study of gene expression (Dixon *et al* 2008) indicate that other SNPs, not in LD with rs2473277 have a greater effect on *CDC42* expression although this varies depending on which expression probe is considered (Table 4.8). Given the central role of CDC42 in dendrite formation and in senescence, the identification of true *cis*-acting variants on expression levels could be very interesting. In conclusion it is likely that susceptibility to schizophrenia is not manifest through individual differences in *CDC42* expression levels and that SNPs other than rs2473307 have a greater, direct *cis*-acting affect on *CDC42* expression.

SNPs which directly affect CDC42 expression may also act in *trans* as suggested by some of the results from the genome-wide linkage study of global gene expression (Dixon *et al* 2008). Furthermore, preliminary results using data from a genome-wide association study of whole gene expression in the human cortex (Webster *et al* 2009, with secondary analysis specifically for CDC42 by Dr Lynn Cochran and Dr Richard Anney, both of the Neuropsychiatric Genetics group, TCD) identified an effect from chromosome 9, in the gene encoding protein tyrosine phosphatase receptor, type D (*PTPRD*, OMIM 601598) (Best genome-wide association result at rs10958892, 'C' allele, n=189, β=-0.1428, Stat=-5.791, $p=2.91\times10^{-8}$). Interestingly PTPRD has strong expression levels in the hippocampus (Mizuno *et al* 1993) where it promote neurites outgrowth (Wang & Bixby 1999). In addition, mice with targeted knock-out of the *PTPRD* gene have grossly normal hippocampi but exhibit pathogenic food avoidance, reduced spatial learning and elevated long-term potentiation (Uetani *et al* 2000). Epistatic analysis has not been performed in the ISC sample but would be an interesting future study especially if directed by knowledge of common *trans*-acting variants.

In a final analysis of the role of *CDC42* in schizophrenia susceptibility, two independent association studies were queried for association at rs2473277. No significant association was identified either in the GAIN or WTCCC2 studies. Notably the MAF of rs2473277 in the ISC sample controls was higher than that in case or control samples from either replication study whereas the case MAF was similar in all three studies. The power to detect association at rs2473277 was however only 0.23-0.25. Thus the association at rs2473277 with schizophrenia can not be completely discounted on this evidence. A replication study with a power of 0.80 to detect an association would require ~6,000 cases and a similar number of controls.

In summary, of the genes whose protein products interact with HOMER2, *CDC42* showed the strongest level of association in the ISC GWAS. The SNP with the strongest level of association was also associated with changes in CDC42 expression. This may partly explain previously published *post mortem* findings of reduced CDC42 levels in the schizophrenia brain (Hill *et al* 2008). In addition, a reporter gene assay generated supportive results in neuronal but not non-neuronal cells. However, this was not supported by the results of an EMSA in which specific binding of transcription factors to either allele could not be demonstrated. Furthermore, other, non-associated SNPs at *CDC42* had a greater effect on expression levels.

Thus, the role of common genetic variation at *CDC42* in schizophrenia susceptibility remains to be tested in replication samples of sufficient size. Further functional studies should employ fine mapping of association signals and have evidence from multiple, complementary functional experiments. The data relating to CDC42 and other interactors of HOMER2 e.g. SHANK1 thus warrants further investigation.

Table 4.8: Best results from each probe on *CDC42* gene expression from the Dixon *et al* (2008) study

Probe ID	Binding sites	Best SNP	Chr	Position	Gene	Region	Allele	Effect	H^2	LOD	p-value	ISC p-vlaue	ISC OR
1556931_at	Intron 1	rs2056974	1	22,154,399	*CDC42*	Intron 4	T	0.44	7.61	6.55	4×10^{-8}	6.1×10^{-4}	0.87
208727_s_at	Exon 8	rs2268177	1	22,160,716	*CDC42*	Intron 5	A	-1.21	40.33	35.00	6.3×10^{-37}	0.30	1.07
208728_s_at	3'UTR (exon 8)	rs2815432	1	69,439,984	*LRRC7*[1]	5'	G	-0.67	8.31	5.846	2.1×10^{-7}	0.51	1.04
210232_at	Exon 7	rs12660382	6	31,551,301	*HCP5*[2]	3'	C	-0.52	7.19	5.20	9.8×10^{-7}	0.15	0.91
214230_at	3' UTR (exon 7)	rs6472528	8	71,499,000	*GRIP2*[3]	5'	A	0.41	6.03	4.42	6.3×10^{-6}	0.019	1.15
226400_at	3' UTR (exon 8)	rs4655026	1	22,218,963	*WNT4*	3'	C	0.442	9.65	7.49	4.3×10^{-9}	0.013	1.13
Average	-	rs2268177	1	22,160,716	*CDC42*	Intron 5	A	-0.49	6.56	5.811	2.3×10^{-7}	0.30	1.07

[1] The closest gene to rs2815432 is *LRRC7*, 0.4Mb distal. LRRC7 encodes a post-synaptic protein involved in dendritic morphology (Quitsch *et al* 2005).

[2] The closest gene to rs12660382 is *HCP5*, 10Kb proximal. Although the biological relevance is unknown, multiple strong associations with schizophrenia were identified in the MHC region here by the ISC GWAS.

[3] The closest gene to rs6472528 is *NCOA2/GRIP2*, expression of which is reduced in the schizophrenia prefrontal cortex (Choi *et al* 2009).

Chapter 5

Detection of copy-number variation using next generation sequencing technology

5.1 Introduction

This chapter reports on my involvement in a study within the lab to develop a method for the detection of copy-number variation (CNV) using next-generation sequencing (NGS) technology. In recent years, copy-number variation (large deletions, duplications or inversions of genomic sequence) has been identified as a common form of human genomic variation (Sebat *et al* 2004, McCarroll *et al* 2008) and has been implicated in susceptibility to a number of phenotypes including schizophrenia (ISC 2008, Stefansson et al, 2008, Xu et al, 2008, Walsh *et al* 2008, Tam *et al* 2009 for review). It is detectable using NGS technology (Yoon *et al* 2009, Alkan *et al* 2009) which is now in routine use for whole-genome and targeting sequencing for the identification of disease susceptibility variants. This chapter reviews current knowledge on CNVs and describes my role in the CNV detection experiment using NGS.

5.1.1 Defining copy-number variation

The term 'copy-number variation' was originally used in microbial genetics and was subsequently adopted for use in plant and mammalian genetics to describe small genomic sequences, which were deleted or duplicated. Sub-microscopic losses and gains of genomic material are common in cancer cells and it was through these studies that the technique of comparative genome hybridization (CGH) was developed for high-resolution, genome-wide detection of copy-number variation in humans (du Manoir *et al* 1993, Kallioniemi *et al* 1994, Pinkel *et al* 1998). The length of common CNVs is typically 1kb to 1Mb (Sebat *et al* 2009). Very large CNVs (>5Mb) that are detectable using

cytological techniques and can cause, often developmental, clinical phenotypes (Lupski 1998 for review). Examples of these include Wolf-Hirschhorn syndrome (Gusella *et al* 1985, OMIM 194190) and Charcot-Marie-Tooth disease (Lupski *et al* 1993, OMIM 118220)

5.1.2 Molecular biology of copy-number variation

Across the human genome CNV distribution is not generally influenced by the position of genes: a CNV may completely encompass many genes, it may just disrupt the coding sequence of one gene, it may be located within an intron of a gene or it may be distant from any known transcripts. CNVs may occur within the boundaries of other CNVs in which case they are termed 'complex'. Small CNVs (1-1,000bp in length) known as 'InDels' (Insertion/Deletion) are detectable by dideoxy terminator sequencing and are common, a recent estimate suggests that an individual human genome may contain up to 400,000 (Mills *et al* 2006). Their frequency within human populations is difficult to estimate accurately because, unlike SNPs, they are difficult to assay with high-throughput techniques.

5.1.3 Population genetics of copy-number variation

One of the most comprehensive studies of common CNVs in population samples is by McCarroll *et al* (2008). Here, 270 HapMap individuals (CEU, YRI, JPT, CHB) were screened using a hybrid SNP-and-CNV genotyping array (Affymetrix SNP 6.0 SV, see section 5.1.6 for CNV detection methods). The study identified 1,320 CNVs with a minor allele frequency greater than 0.01. The reported lengths of these CNVs ranged from 73bp to 1,134,231bp (average 26,288bp, median 7,383bp) although these figures are dependent on the genomic distances between probes on the array. The first major finding was that large, common CNVs (>5Kb) were much smaller than previously estimated by aCGH methods and thus only affect 5% of the genome rather than the previously estimated 12%. Secondly, they confirmed that CNVs behave in a very similar way to SNPs with regards to allele frequencies, linkage disequilibrium patterns, transmission dynamics and population uniqueness with less than 1% generated *de novo*. The genotype calls are publicly available via the Database of Genomic Variants (http://projects.tcag.ca/variation/). These data show that of the 1,320 CNVs reported in all population samples analysed, only 653 and 643 are present in the CEU and JPT+CHB samples respectively compared with 913 in the YRI. Analysis of the CNVs in the HapMap sample shows that

there are 309 bi-allelic, autosomal CNVs with population characteristics (transmission between generations) which would suggest that they are not artefactual. The vast majority of these CNVs range from 1,000bp to 100,000bp in length and have a MAF less than 0.1. In addition, the MAF of a CNV does appears to be independent of its length (figure 5.1).

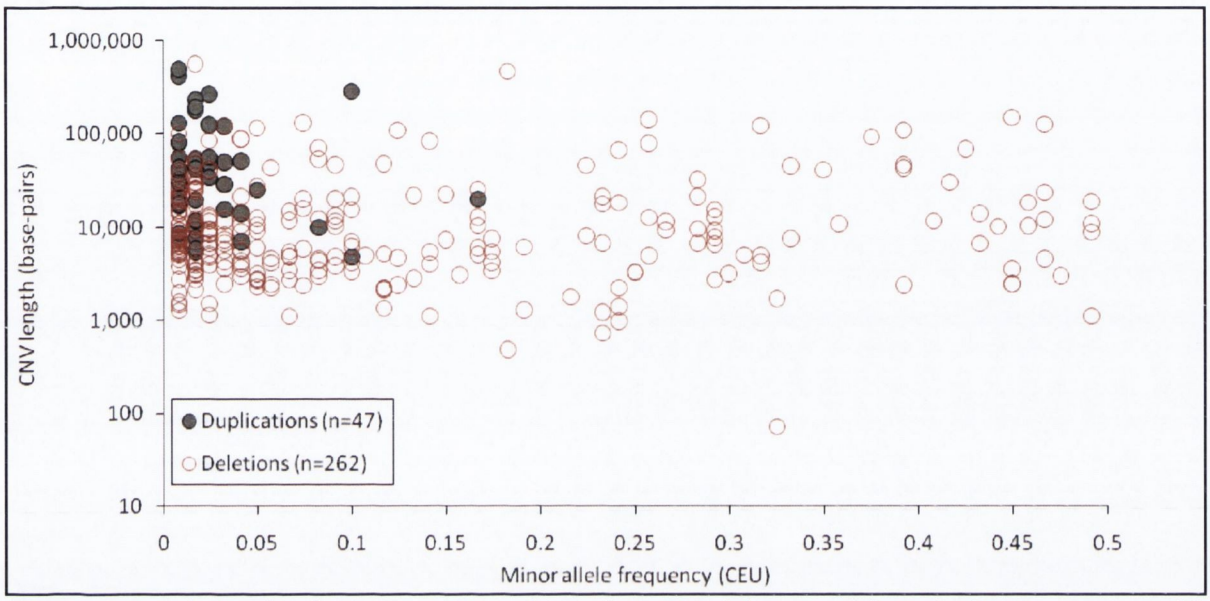


Figure 5.1: Relationship between MAF and length in 262 deletions and 47 duplications reported by McCarroll *et al* (2008). All CNVs here are from the CEU sample and are bi-allelic, autosomal, with good call rates, perfect Mendelisation and are in HWE.

5.1.4 Phenotypic effects of copy-number variation

The function of many genes has been elucidated by 'knock-outs' (which are comparable to deletion CNVs) in model organisms and often lead to extreme or otherwise distinctive phenotypes. Genomic rearrangements which disrupt the coding sequence of specific genes can cause Mendelian forms of common diseases. For example, deletion of multiple exons of the *PARK2* gene implicated in early-onset Parkinson's disease (Kitada *et al* 1998) and deletion of multiple genes at 16p11.2 causing syndromic obesity (Walters *et al* 2010). Common CNVs can be strongly correlated with *cis* and *trans* effects on gene expression (Stranger *et al* 2007, Orozco *et al* 2009). A CNV affects a far larger genomic region than that of a SNP. Therefore, it is more likely to have an effect on gene function and in principle be of more relevance to disease phenotypes.

5.1.5 Copy-number variation in schizophrenia

Large schizophrenia case-control samples have recently been subjected to genome-wide CNV analyses with two major findings:

Firstly, there are large rare CNVs that are associated with schizophrenia. These are CNVs (deletions and duplications) at 1q21.1, 15q13.3, 16p11.2 and 22q11.2 that are present at a frequency 0.2-0.5% in schizophrenia cases (odds ratios of 6-30; ISC 2008, Stefansson *et al* 2008, Walsh *et al* 2008, McCarthy *et al* 2009). At each locus all CNVs were of different sizes, indicating a rare, 'private' mutation. Neither the parental genotypes nor phenotypes were reported so it is unknown whether independent de novo events occur at each locus. These CNV loci have been found to be associated with other phenotypes, both psychiatric e.g. autism (1q21.1; Mefford *et al* 2008) and non-psychiatric e.g. epilepsy (15q13.3; Helbig *et al* 2009, Dibbens *et al* 2009). Thus, CNV analyses have identified overlap in genetic aetiology between clinically distinct disorders (Sebat *et al* 2009).

Secondly, there is an increase in total CNV burden in individuals with schizophrenia (average of 0.99 large (>100kb) rare (MAF<0.01) CNVs per control compared to an average of 1.15 per schizophrenia subject, $p=3\times10^{-5}$, ISC 2008). Many of the risk CNVs found in schizophrenia patients disrupt genes which encode neuronal and synaptic proteins. Notable genes affected by private CNVs in schizophrenia patients include NRXN1, PDE4B, NPAS3, CNTNAP2 and APBA2 (Sebat *et al* 2009). Pathway analyses of rare CNVs in schizophrenia patients have identified enrichment of genes in 'synaptic long-term potentiation' and 'nitric oxide-signalling' amongst others (Walsh *et al* 2008). Such discoveries generally support existing hypotheses on schizophrenia aetiology but importantly inform on novel genetic lesions which exist long before symptom onset. Two potential caveats are that although the burden of rare CNVs is increased in schizophrenia the causative ones cannot be readily identified. Furthermore, bias may be introduced into pathway analysis because many neuronal genes are large and therefore more likely to harbour CNVs.

By contrast, a recent genome-wide association study of eight common disorders using common CNVs greater than 500bp in length only identified five significant associations; with diabetes (types I and II), rheumatoid arthritis and Crohn's disease (two different CNVs) but not with bipolar disorder (WTCCC 2010). Overall the study concluded that common CNVs are unlikely to have a major role in risk for common disorders.

In summary, the detection of rare CNVs in schizophrenia has provided a dramatic new insight into the genetic aetiology of the disorder, which differs somewhat from the finding from other

common disorders. The resolution of the CNV analysis methods has been limited and thus little is known about the role of CNVs under 40Kb in disease. NGS should provide an excellent method through which such smaller CNVs can be detected.

5.1.6 Detection methods for copy-number variation

Chromosomal rearrangements are generally considered to be large-scale CNVs. Their effective identification in individuals has been possible for some time using standardised cytogenetic techniques such fluorescence *in situ* hybridisation (FISH) followed by visualisation using a microscope (Carter 1994 for review). The detection of smaller CNVs, down to 5Mb in length was made possible through the development of comparative genome hybridisation (du Manoir *et al* 1993). In this process, subject and reference genomic DNA is labelled with different fluorophores and, in array format, then hybridised to thousands of unlabelled probes which cover the genome. A CNV is indicated by a difference in fluorescence and the genomic location corresponds to which probe the signal is detected at. One evolution of CGH is representational oligonucleotide microarray analysis (ROMA) which uses restriction enzyme digestion to reduce genomic complexity and thus improve sensitivity down to CNVs of 100Kb in length (Lucito *et al* 2003). Currently, available probe-based CNV detection methods (coming under the umbrella term array-CHG) are capable of detecting CNVs down to 1Kb long (e.g. Perry *et al* 2008).

Detection of CNVs is also possible and commonly practiced using SNP genotyping arrays (Winchester *et al* 2009 for review). This is possible because the PCR reaction generates intensity data, designed for detection of either or both alleles at a SNP but which can also be analysed as a proxy for copy-number. In the analysis of a single SNP in multiple individuals, the formation of clusters of different intensities is indicative of a CNV. Most analysis platforms use hidden Markov models to identify such clusters although there are a range of methods available (Dellinger *et al* 2010 for review). CNV detection using SNP genotyping arrays has the potential to detect smaller CNVs than aCHG, especially at high SNP densities. However, there is a high signal-to-noise ratio and thus sensitivity and specificity can be compromised. The method has been used to detect and genotype CNVs in the HapMap reference samples (CEU, YRI, HCB and JPT, McCarroll *et al* 2008, see section 5.1.3). CNVs which are detected using array-CGH or SNP genotyping platforms can be confirmed using real-time PCR in which the amplification rate of the duplicated or deleted region will differ from that of a copy-number normal region. Finally, individual CNVs can be confirmed by PCR followed by dideoxy terminator sequencing in which the primers for PCR bind outside the CNV. This

approach has been used to map the exact breakpoints of a schizophrenia-associated deletion within the ERBB4 gene (Walsh *et al* 2008).

5.1.7 Next-generation sequencing methods

Nucleotide sequencing provides a map of genetic variation in an individual at the resolution of a single base-pair and thus is an extremely powerful tool for the investigation of disease genetics. Traditional dideoxy terminator sequencing has recently been surpassed by 'next-generation' sequencing technologies which have vastly higher throughput and reduced operating costs. This greatly facilitates whole-genome sequencing for the detection of novel variants in humans and model organisms as well as *de novo* sequencing of novel genomes. Several companies now offer NGS machines, reagents and protocols, each of which has subtly different methods (table 5.1, Mardis 2008, Metzker 2010).

The NGS platform in use at the Neuropsychiatric Genetics laboratory in TCD is the Illumina Genome Analyzer II which is based upon the sequencing-by-synthesis. The core procedure starts with fragmentation of genomic DNA (5µg) and the ligation of short adaptor sequences. The fragment-adaptor constructs are immobilised to a solid surface (flow cell) and subsequently amplified using bridge-PCR (Fedurco *et al* 2006). Here, priming of the universal adaptor sequences is the critical process which allows for the simultaneous amplification of thousands of different target sequences. Prior to sequencing, one end of each 'bridge' is cleaved from the solid surface to create up to 200 million short strands on the flow cell. These strands appear as spots when viewed from above and its colour corresponds to the first nucleotide of a DNA fragment. Through a cycle of image capture and individual nucleotide addition, the sequence of millions of overlapping DNA fragments can thus be determined. Sequencing-by-synthesis has been highly successful at detecting SNPs (Bentley *et al* 2008, Li *et al* 2009) and large CNVs (Yoon *et al* 2009, Alkan *et al* 2009, section 5.1.9). However the power of the Genome Analyser II to detect smaller CNVs, at the level of a single exon, is currently unknown.

Table 5.1. Comparative performance of different NGS platforms

	Platform		
Manufacturer	Roche	Illumina	Applied Biosystems
Instrument name	454	Genome Analyser II	SOLiD
Sequencing chemistry	Pyro-sequencing	Polymerase-synthesis	Ligation-based
Amplification approach	Emulsion PCR	Bridge amplification	Emulsion PCR
Paired ends/separation	Yes/3 kb	Yes/200 bp	Yes/3 kb
Mb/run	100 Mb	1,300 Mb	3,000 Mb
Time/run (paired ends)	7 hours	4 days	5 days
Read length (bp)	250	32–40	35
Cost per run[1] (US $)	8,439	8,950	17,447
Cost per Mb[1] (US $)	84.39	5.97	5.81

1. This value includes the cost of reagents, consumables, labour, instrument and data storage. Information for this table taken from Mardis *et al* (2008). Running costs have decreased since and are still doing so. Thus the information is useful for comparative purposes.

5.1.8 Copy-number detection by next generation sequencing – whole genome resequencing

The total number of reads generated from a sequencing run is dependent on the amount of starting DNA and thus is dose-dependent. This has implications for the detection of CNVs: less DNA in a given region, as caused by a deletion, would be reflected by less reads and with the opposite effect for a duplication or other gain of copy-number. The number of reads in any given region is called read depth of coverage. Large CNVs have been identified in whole genome sequence data by comparing read depth of coverage between individuals (Alkan *et al* 2009). Because whole genome data was used in this example, read depth of coverage was determined in comparatively large 5Kb-long windows. In doing this stochastic effects on read depth of coverage were reduced at the expense of being insensitive to smaller CNVs. Furthermore, the effect of GC-content on read depth of coverage was characterised and corrected for. The benefit of this correction is unknown when performing a pair-wise comparison between read depth in two individuals for CNV detection. In theory, regions with extreme GC content are harder to detect CNVs in because they will have a lower read depth of coverage and therefore less data. These authors (Alkan *et al* 2009) found that read depth of coverage in normal sequence peaks at a GC-content of approximately 40%. Although read depth of coverage clearly decayed either side of 40% GC (figure 5.2), the authors corrected the read depth using a linear model which would reduce sensitivity of detecting CNVs in sequence <40% GC.

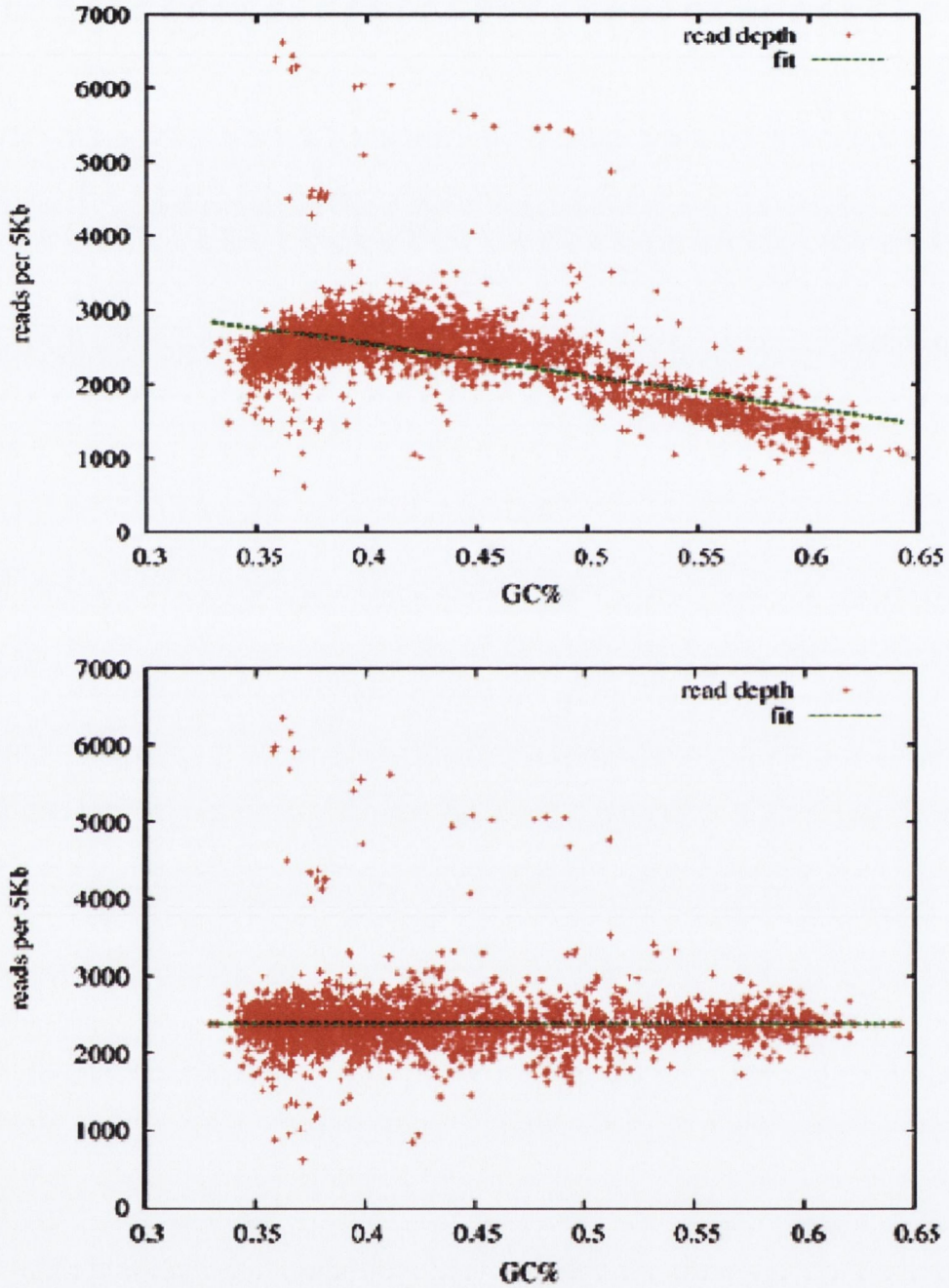


Figure 5.2: Variation in read depth of coverage with GC-content using the Illumina NGS method for whole genome sequencing (left) and after normalisation (right) using a 'locally-estimated scatterplot smoothing' (LOESS) method (Alkan *et al* 2009, supplementary information).

An alternative approach for screening whole genome sequence for CNVs has adopted an 'event-wise detection' algorithm (figure 5.3, Yoon *et al* 2009). This effectively treats read depth of coverage as series data and looks for significant departures from the expected read depth in consecutive, non-overlapping windows, each of 100bp in length. In this second example, pre-analysis adjustment of read depth of coverage was based upon the average read depth of coverage for windows of a similar GC-content and thus is a more refined method than that used by Alkan *et al* (2009). The sensitivity of this overall method was 99.9% for known CNVs >1000bp and 77% for those <1000bp, and thus this makes NGS a more sensitive method for CNV detection than aCHG.

The two examples reviewed above used reads generated from whole-genome sequencing which produces a vast amount of data. This is advantageous for the detection of CNVs because the relationship between GC-content and normal read depth of coverage can be accurately determined and also because comparatively small CNVs, <1000bp in length, will still be detected by reductions in read depth of coverage in multiple neighbouring windows even if stochastic effects mean that an 'event' is not detected in some of the windows. The disadvantages of whole-genome resequencing are that it is requires considerable bioinformatic and financial resources to be performed at the scale of sample numbers required for the discovery of rare disease-associated variants.

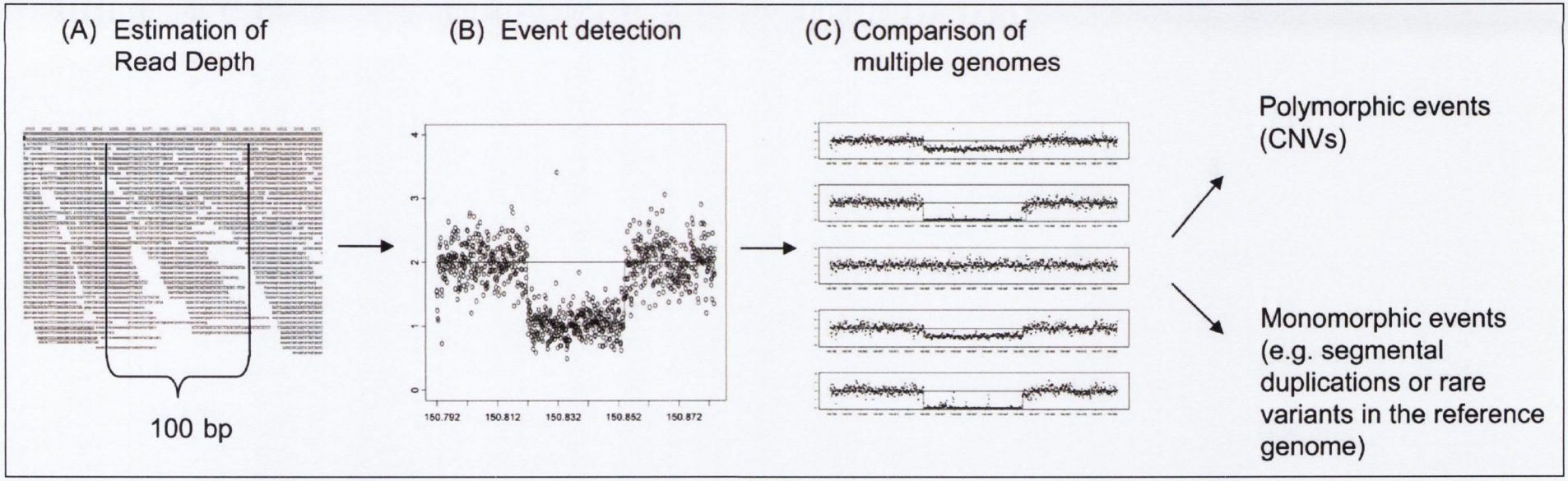


Figure 5.3: Pipeline for the detection of CNVs based on analysis on read depth (RD) as applied by Yoon *et al* (2009).

A: RD was determined by counting the start position of reads in non-overlapping windows of 100bp.

B: Events were detected using a custom CNV-calling algorithm, event-wise testing.

C: Each event was examined in multiple genomes in order to distinguish polymorphic events (CNVs) from the majority of events that were found to show a similar copy number change in all five genomes in this study (monomorphic events).

5.1.9 Targeted resequencing methods

Whilst whole genome sequencing remains the most comprehensive means of variant detection, it is still costly and not yet feasible in a large number of samples. When candidate genes can be prioritised using functional and association data, it is more efficient to sequence only the most important functional sites such as exons and UTRs. This is based on the premise that the majority of disease related mutations are located within coding sequences (approximately 1% of the genome). From Mendelian disease we know that mutations causing amino acid changes account for ~60% of disease mutations (Botstein & Risch 2003). Thus, exon sequence must be isolated from the DNA sample prior to sequencing. This process is known as sequence capture, pull-down or target enrichment and can be applied to most regions in the genome and not just exons (Mamanova *et al* 2010 for review). The use of conventional PCR methods for this process is inefficient because the amount of sequence that can be targeted per reaction is limiting. As an alternative 'genome partitioning' methods are being developed for the enrichment of targets of interest. Of these, the hybrid selection method provides the greatest combination of coverage and accuracy and is currently in use in the Neuropsychiatric Genetics laboratory, TCD (Gnirke *et al* 2009).

In the hybrid selection method, generation of a library of genomic DNA which contains only the regions of interest is performed by hybridizing total fragmented genomic DNA to synthetic biotin-labelled, 120-mer 'bait' cRNA sequences which correspond to the target DNA (Bhattacharjee *et al* 2009). Strepdavidin-coated beads allow for the capture of the bait-target complexes. The target sequences are then eluted-off and ligated into universal adaptor sequences which allow for the preparation of sequencing libraries (section 5.1.8).

An additional technological advancement is 'barcode' sequencing which allows for DNA from multiple individuals to be sequenced in single lane of the Genome Analyzer II. Here, a unique sequence of six bases is ligated to the fragmented DNA and prior to the ligation of universal PCR adaptors (Craig *et al* 2008). When the sequence data is generated, reads are parsed using the six-base barcode allowing sequence analysis of individual samples.

Genome-wide exon or 'exome' resequencing has already identified novel coding variants that represent causal mutations for rare Mendelian diseases (Ng *et al* 2010a, 2010b). However exome data has not been analysed for CNVs and thus the ability of targeted NGS to detect CNVs remains to be investigated.

5.1.10 Relevance of copy-number detection using next-generation sequencing to schizophrenia

The detection of copy-number variation using NGS is important to schizophrenia for two reasons:

1) Genome-wide association studies are generating strong and reproducible association signals. In order to fine-map the region of association, re-sequencing using NGS is being routinely performed. If the causative variant is a CNV it is crucial that it can be detected in the sequence data.

2) Genome-wide association studies assaying common variants have not been able to account for a large proportion of the heritability of common diseases including schizophrenia, although this may reflect a lack of power to identify small effects (Manolio *et al* 2009). Large rare CNVs have been identified in schizophrenia patients and potentially contribute to susceptibility. The size limit for microarray and SNP-based detection of CNVs means that many smaller CNVs are not detected by these methods. Post-GWAS, a number of rare variant resequencing studies are already underway at different scales (whole genome, whole exome, candidate genes). NGS has the potential to detect small CNVs whilst simultaneously detecting SNPs in these studies. Thus, there is a requirement for investigation of changes in read depth of coverage in the presence of a CNV and development of an algorithm for their automated detection in massive sequence data sets.

5.1.11 Aims and experimental overview

A major on-going project in the laboratory is the identification of rare variants which cause schizophrenia. This is being performed using NGS to target resequence the exons of 215 genes. Because rare causative variants can be CNVs, it is imperative that they can be detected using NGS. However, the amount of data generated by NGS means that it is not amenable to detailed manual inspection for anomalies which might indicate the presence of a CNV. Thus the aim of this work is to examine the effect of copy-number variation on read depth and develop an algorithm whereby changes in read depth can be used to detect a CNV.

Sixteen common CNVs were sequenced in HapMap CEU individuals for whom the genotype and approximate breakpoints had been previously determined (McCarroll *et al* 2008). Two flanking regions for each CNV were also sequenced to generate data on read depth variation in the absence of a CNV. The major factor influencing read depth other than copy-number variation was expected

to be GC-content. Using data from this experiment, a model was developed with the aim of removing fluctuations in the read depth caused by extremes of GC-content.

The work described in this chapter involved significant input from other individuals, all in the Neuropsychiatric Genetics laboratory. Generation of raw data using the Illumina Genome Analyzer II was undertaken by Dr Elaine Kenny. Determination of the read depth from the raw data was performed by Dr Paul Cormican. Development of a high-throughput method for the detection of CNVs based-on read depth corrected for GC-content was undertaken by Dr Elizabeth Heron. Project co-ordination was under the supervision of Dr Derek Morris. My role was to select known CNVs for analysis by NGS, design baits to target the copy number variable regions and flanking regions at each CNV for the enrichment assay pre-NGS, determine the exact breakpoints of each CNV by traditional dideoxyterminator/Sanger sequencing and assist in the interpretation and presentation of results. The overall aim of the study was to combine my data on the breakpoint sequencing and the genotypes from the HapMap samples to produce 'positive control' data that could be used to test the method for detecting CNVs in the targeted sequencing data.

5.2 Materials and methods

5.2.1 Selection of copy-number variable regions for sequencing

In order to generate empirical data on the changes in NGS read depth across CNVs it was necessary to sequence across known CNVs in available DNA samples. Genomic DNA from the HapMap CEU sample, which comprises thirty parent-offspring trios, was obtained from the Corriell Institute for Medical Research (New Jersey, USA). The genotypes of CNVs for these individuals were generated previously (McCarroll *et al* 2008) and made publicly-available on the Database of Genomic Variants (DGV; http://projects.tcag.ca/variation/; Zhang *et al* 2006). The CEU sample was found to have 653 polymorphic CNVs. Because this represented far more data than was required for the study, a number of criteria were used to select CNVs for sequencing, with the objective of selecting the CNVs that were most likely to be real and also from a range of sizes.

The genotypes of CNVs in the CEU had been generated through analysis of SNP genotyping arrays. Thus, to reduce the possibility of obtaining inaccurate genotype data, only CNVs which were deletions and had genotypes of zero, one or two copies in the CEU samples were considered for further investigation. These CNVs were imported into HaploView (v4.1, Barrett *et al* 2006) by converting genotype calls for the CEU samples into a SNP format for the purposes of analysis. Only CNVs with a MAF>0.2, perfect Mendelisation, 100% call rate and were in Hardy-Weinberg equilibrium ($p>0.05$) were considered for further investigation. CNVs located on sex-chromosomes and greater than 10Kbp long were also excluded. The remaining CNVs were further investigated using the DGV track in the UCSC Genome Browser to check if they had detected by an independent study even if the reported breakpoints differed slightly. Fourteen CNVs that met the above criteria, were of a wide range of sizes (below 10kbp) and were spread across the autosomal genome were selected for sequencing (table 5.2). Finally, two CNVs with reported gains of copy-number, including one with a maximum copy-number of six were also selected for analysis using NGS.

Table 5.2: Overview of CNVs selected for analysis by NGS

CNV ID[1]	Co-ordinates (hg18)[2]	Length (bp)[2]	MAF (CEU)	Observed copy numbers	HWE p-value	Also Reported by
84	chr1:108535758-108539019	3,261	0.25	0,1,2	1.00	3, 4
337	chr2:208063423-208067581	4,158	0.20	0,1,2	0.26	3, 5
435	chr3:82955469-82955733	264	0.10	0,1,2	0.13	6
451	chr3:100427148-100432111	4,963	0.26	0,1,2	0.81	4,5
468	chr3:129878180-129895021	16,841	0.11	2,3,4	0.99	3,7
682	chr4:138311645-138312864	1,219	0.23	0,1,2	0.79	4, 8
703	chr4:156022455-156022528	73	0.33	0,1,2	0.68	6
724	chr4:172611459-172614508	3,049	0.27	0,1,2	0.84	4, 9
1343	chr8:112363455-112364436	981	0.24	0,1,2	1.0	6, 10
1602	chr10:56353096-56362605	9,509	0.27	0,1,2	0.37	11
1610	chr10:58880511-58880997	486	0.18	0,1,2	0.60	4, 12
2215	chr16:86671164-86672892	1,728	0.23	0,1,2	1.0	4, 6
2326	chr18:36514418-36519387	4,969	0.23	0,1,2	1.0	4, 5
2529	chr21:43794765-43797240	2,475	0.45	0,1,2	1.0	9, 10
2546	chr22:20055998-20175294	119,296	-	2,3,4,5,6	-	13
2575	chr22:37625201-37626850	1,649	0.40	0,1,2	0.41	4, 9

1: CNV ID is taken from the CNVs identified by McCarroll *et al* (2008), data taken from the Database of Genomic Variants.

2: Stated co-ordinates are derived from a genotyping array (Affymetrix 6.0 SV, McCarroll *et al* 2008) and thus the precision is influenced by probe locations. Only CNVs reported by a different group and using a different detection method were used. These were identified through the UCSC Genome Browser. References for CNVs independently identified are: 3: Cooper *et al* (2008), 4: Perry *et al* (2008), 5: Wang *et al* (2007), 6: Mills *et al* (2006), 7: de Smith *et al* (2007), 8: Conrad *et al* (2005), 9: Wheeler *et al* (2008), 10: Korbel *et al* (2007), 11: Pinto *et al* (2007), 12: Hinds *et al* (2005), 13: Locke *et al* (2006).

5.2.2 Bait design for target enrichment sequencing of CNV and flanking regions

The method of target enrichment used in this experiment was the Agilent SureSelect Target Enrichment System (see section 5.1.10). The bait library used in the sequencing reaction was generated by Agilent Technologies earray system (Santa Clara, CA, USA) following submission of the genomic co-ordinates for analysis (https://earray.chem.agilent.com/earray/). The co-ordinates entered for the CNV regions were expanded from the breakpoints stated by McCarroll *et al* (2008), firstly to the maximum co-ordinates for a CNV reported by a second study (all accessed via the UCSC Genome Browser, see table 5.2 for references) and secondly by another 4Kbp to potentially sequence across the true breakpoints of the CNV. For each selected CNV, two 2Kb-long flanking regions were also selected. Each flanking region was located 50-100Kb from the CNV in a region in which no CNVs were reported. The purpose of sequencing the flanking regions was to generate sequence data at copy-number normal loci (see figure 5.5 for overview of bait-design locations).

The co-ordinates for the fourteen selected bi-allelic deletion targets, the two gain-of-copy-number targets and their corresponding flanking sites were submitted to the Agilent eArray bait design which generates sequences for overlapping probes 120 bases in length each. These probe sequences were tested for potential non-specific binding to other parts of the human genome which could be a serious cause of experimental error. Non-specific probes were determined as those having greater than 80% homology to another site. Stretches of target sequence were excluded on the basis of non-unique probe homology. Stretches containing excessive amounts of non-unique probes were actively excluded from a second round of bait generation using the eArray system.

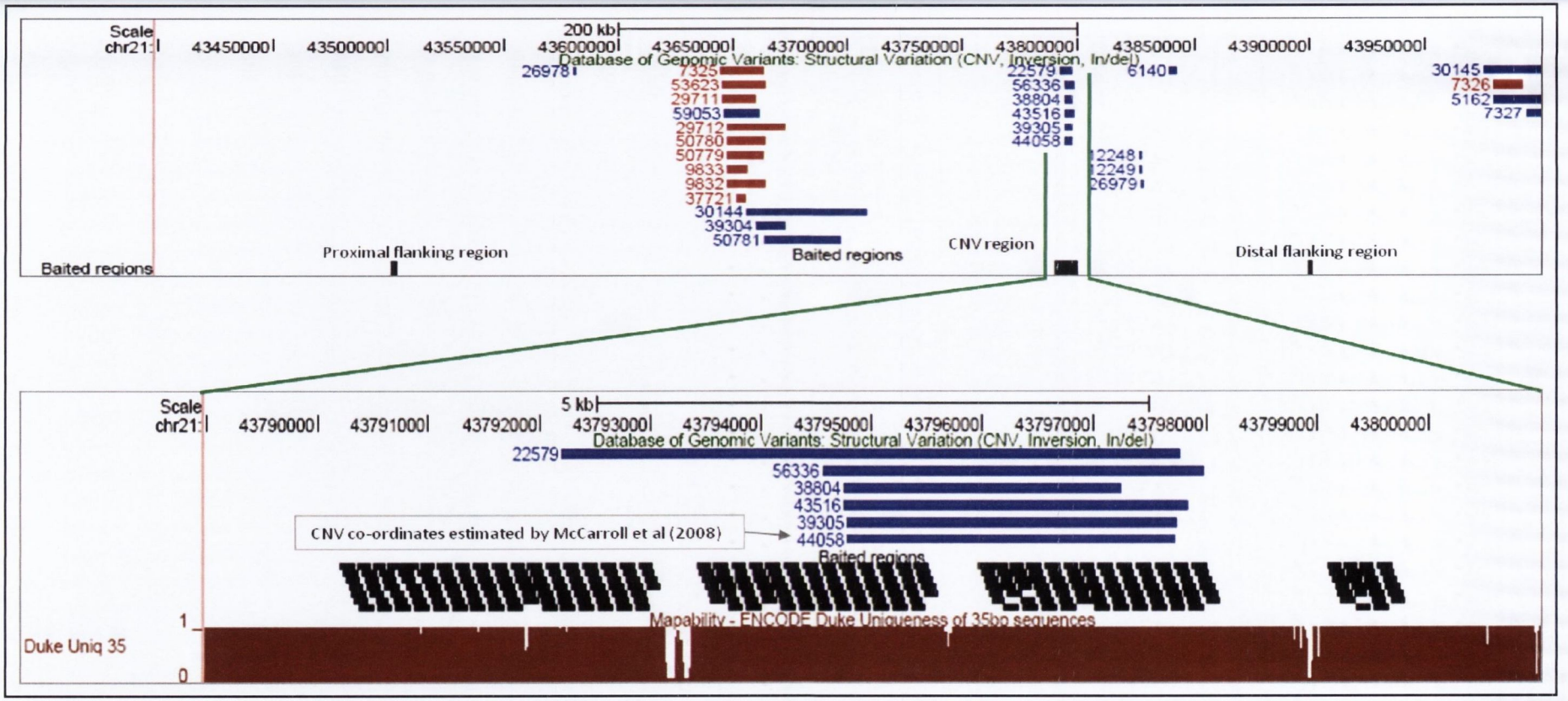


Figure 5.4: Diagram showing position of baited regions at an example CNV (CNV ID 2529). The upper panel shows 400Kb of chromosome 21 in which CNVs reported by different studies and submitted to the Database of Genomic Variants are indicated in blue (deletions) or red (duplications). The baited flanking regions (proximal and distal) are located away from any potential CNVs. The lower panel is an enlargement of the targeted CNV region. It shows how overlapping baits (short black bars) were designed to target the widest possible breakpoints (CNV 2529 is reported as #44058 in the USCSC Genome Brower but baits were also designed to cover the largest reported overlapping in this region, #22579). Gaps in the baits are due to repetitive sequence that is not targeted because it is likely to be un-mappable sequence.

5.2.3 Sequencing of targeted sites using the Illumina Genome Analyser II

The work performed in this section was undertaken primarily by Dr Elaine Kenny and with the assistance of Ms Amy Gates. Genomic DNA from each of nine CEU HapMap samples (three trios) was used. The individuals used are summarised in table 5.3.

Table 5.3: Individuals from the CEU sequenced in this study

Individual ID[1]	Family ID[1]	Gender	Status
NA10846	1334a	Male	Offspring
NA12144	1334a	Male	Parent
NA12145	1334a	Female	Parent
NA10847	1334b	Female	Offspring
NA12146	1334b	Male	Parent
NA12239	1334b	Female	Parent
NA10859	1347	Female	Offspring
NA11881	1347	Male	Parent
NA11882	1347	Female	Parent

1. Individual and family ID correspond to those routinely used for this sample

The indexing methodology was taken from the study by Craig *et al* (2009). Each unique 6bp index was designed so that it can tolerate at least 1 base change as a result of a sequencing error without mutating into one of the other indices used. This helps guard against mix-up of individual sample data. The preparation of each sample is a 2-step process; in the first step the DNA is prepared as an Illumina sequencing library and in the second step the sequencing library is enriched for the desired target using the Agilent SureSelect enrichment protocol. The library preparation and enrichment methods were followed according to the Agilent Illumina Single-End Sequencing Platform Library Prep protocol (v1.2 April 2009) with the following modifications: 1. Instead of shearing the DNA with a Covaris system, a biorupter (Diagenode) was used. The samples were sonicated on high for 30secs and off for 30secs for a total 30 mins with addition of ice after every 10 mins to keep the samples cool. 2. Illumina adapters were replaced with custom made indexed adapters according to methods described above and supplied by IDT DNA. 3. For the pre-capture enrichment PCR 1.3µl of DNA was used as input with 11 cycles of PCR instead of 1µl input DNA and 14 cycles of PCR. 4. Post-enriched library DNA from the 9 HapMap samples was combined to a total quantity of 500ng. This

solution was allowed to evaporate off overnight instead of using a vacuum concentrator and resuspended in 3.4μl of elution buffer to give a final concentration of approximately 147ng/μl. This pooled DNA sample was each enriched using one custom SureSelect enrichment reaction (library design ELID: 0236181). Target enriched libraries were stored at a stock concentration of 10nM ready for sequencing. 6pM of the target enriched library was sequenced on the Illumina Genome Analyzer II using 40bp reads following the manufacturers protocol.

5.2.4 Raw data processing

The work performed in this section was undertaken by Dr Paul Cormican. The base sequence data was called from the image files with the Illumina Bustard.py script and the illumina GA pipeline version 1.4. The RunInfo.xml file was edited to allow for the 6bp index to be called as a separate index read in the analysis, with matrix and phasing estimated from the *Phi*X control lane. This allowed the 6bp index to be ignored in the alignment of the sequences to the reference genome (hg18) to facilitate basic quality control. The effective read length was 34bp. Reads were subsequently parsed by the index sequence using an in-house Perl script. The generation of raw coverage data was performed using MAQ (Li *et al* 2008) on the individual sample sequence data. A SNP-masked reference genome was used for alignment of the reads.

Read density was measured in 100bp windows from the start of each sequenced region. The number of reads which started in each window was termed the absolute read density because the read density was later adjusted by GC-content prior to analysis for CNV detection. Duplicate reads i.e. those with the same starting location, were included as it is not possible to distinguish a random duplicate read from one which has been generated as a PCR artefact prior to sequencing. Because each sequenced region was not a perfect multiple of 100bp in length, data from windows less than 100bp was omitted.

5.2.5 Normalisation of sample read count and GC-content, and detection of CNVs

The work detailed in this section was undertaken by Dr. Elizabeth Heron. A more detailed account of these methods is provided in the Appendix A. Two variables that could confound methods to detect CNVs in sequencing data are variability in read count from sample to sample and variability in read count due to GC content. Sequence reads were binned based on their GC content (0-10%,

10-20%...etc) and a Bayesian method was used to normalize the data accounting for both sample variability and variability due to GC content. Normalised read counts in 100bp windows were assessed for copy-number variation using the 'DNAcopy' analysis package (Venkatraman & Olshen 2007; http://www.watson.nci.nih.gov/ bioc_mirror/packages/2.3/bioc/html/DNAcopy.html) which is implemented in R (http://www.r-project.org/). DNAcopy is designed for detecting copy number variation from array CGH data. Here, consecutive 100bp windows that have approximately the same normalized read count are considered to have the same copy number. Runs of windows with either an increase or decrease in normalized read count are considered as potential copy number variable regions. These potential CNV sites are then investigated using a Kruskal-Wallis non-parametric test to test if at least one sample has a significantly different mean normalised read count, indicating the presence of a CNV.

5.2.6 Determination of exact breakpoints by dideoxy terminator sequencing

This was work undertaken by the author. In order to provide an indication as to how accurate SNP-array data and NGS data are at mapping CNV breakpoints, the exact breakpoints of the CNVs were determined using dideoxy terminator sequencing. Primers for PCR were designed to amplify across the deleted sequence of targeted CNVs in CEU individuals. This meant identifying primer binding-sites which were firstly far enough away from the breakpoints that the binding-site would not be within the deleted sequence and secondly were close enough together that the breakpoint could be covered from a single read on the ABI 3130 sequencer (~1kp). In addition, the amplicons were designed to be of a suitable size for PCR and confirmation by standard gel electrophoresis. Because the true breakpoint positions were unknown for the purposes of primer design, a 'best guess' was used with data from the UCSC Genome Browser and from the NGS data. For the two CNVs, which were identified as duplications by McCarroll *et al* (2008), primers were designed so that the forward primer would bind to the 3' end of the first copy and the reverse primer would bind to the 5' end of the second copy. The 'Web Primer' tool (http://www.yeastgenome.org/cgi-bin/web-primer) was used to assist in the design of primers which were manufactured by MWG Biotech (Ebersberg, Germany; see table 5.4 for details). The PCR conditions and dideoxy terminator sequencing were performed as described previously (see section 3.2.3). Generated sequences were subjected to a BLAT search (via the UCSC Genome Browser, Kent 2002) which clearly identified 'missing' sequence and the exact breakpoints.

Table 5.4: PCR parameters used for identification of CNV breakpoints[1]

CNV ID[2]	Forward primer sequence	Reverse primer sequence	PCR conditions[3]	Product sizes (base-pairs)[4]
84	CTTGTGTATTGACAGAGACTC	GAAGCCTAACTCTGAAACGTG	A65, X30+	4901, 990
435	CTATGCTGTATGTCAGTATGAC	CAACTAGACTATATTTGAGCATC	A62, X30+	4541, 499
451	GTATAGTGTACACTGCTTGG	GCTTATACACTGCTGGTAGG	A60, X30+	5987, 274
703	GTTATGGCATGAAATCTACCT	ACTTGTCTACTGGTAGATTTG	A60, X60+	377, 297
724	GGTATCAAGCCATGGAATAG	GCATATGCTTAGAAAGTTTAT	A55, X60+	5522, 507
1343	GATGAATTATGAATTGATCCTG	GACTGTCCTATTGGATTTCA	A60, X30+	3430, 341
1602	CCAAGTCTCATGTCCTTCTC	CTTAGTAGACTAGATAGCTTC	A55, X60+	2811, 1230
1610	GAGAGACAGCTAATCCAGATG	GTATGACTTGGACAATCTACAG	A60, X30+, E	2351, 284
2215	GAATGCTGTGGAATTGAAAG	GGTGTGATCTTGGCTGATTG	A65, X30+	3465, 527
2326	CAAGCACTTAAGCAGCTGC	TTACTCTCTGTTTGTCATACC	A53, X60+	7971, 1113
2529	CAAGGTAGCTCAATGTTTCC	GACTCTGAATTTATAAGTGG	A60, X60+	3322, 385
2575	GTGTCTATGCTACTAACTTG	GCCATAAGATAGCAGTACAG	A60, X60+	4834, 253

1. PCR parameters stated here are only for those which lead to the identification of CNV breakpoints.

2. The CNV identification number is taken directly from McCarroll *et al* (2008).

3. The letter 'A' refers to the annealing temperature. 'X' refers to the extension time (in seconds), in which the 'plus' symbol indicates the addition of 1 second per cycle. 'E' refers to the use of 'Fail-safe buffer™ E'(Boerhinger Manhaime, Germany).

4. The two values refer to the 'undeleted' and 'deleted' PCR product sizes. Undeleted product size was determined using the UCSC Genome Browser. Deleted product size was determined by subtracting the size of the deletion from the undeleted production size.

5.3 Results

5.3.1 Total number of reads mapping for each individual

Nine individuals were sequenced for known CNVs and distant flanking regions which were considered to be CNV-free. The total number of 100bp-long windows analysed per individual was 1,580 (corresponding to 158,000bp of genomic sequence). The two flanking regions for each CNV provided data on 640 windows. The total number of reads in all 1,580 windows clearly varied (range 107,951-157,047 per individual, table 5.5). Furthermore, the total read counts of the flanking regions only differed also (range 52,758-73,304) and was highly correlated with the total number of reads per individual in CNV regions (r^2=0.93). This indicates that the variation in total read count was not likely to be a result of different CNV burdens between individuals. In the flanking regions the highest number of reads per window was 364 in NA10846. In the same window, NA12145 had only 171 reads – a reduction of over 50%. It is unlikely that the difference in read count in this window was due to a CNV because the two samples had the highest and lowest total read counts respectively. This highlights a potential confounding effect for the detection of CNVs using sequence data, i.e. the quantity of sequence data generated for each sample differs due to experimental variation, before any consideration is given to CNVs being present. In addition, the lowest number of reads in any window in a flanking, copy-number normal region was 0. However, for the same window, the highest read count for any individual was only 4. The window has a GC-content of 0.21 and is located within a 150bp long SINE/AluSg repetitive element. The presence of this repetitive sequence means that it is difficult to map reads at this position and consequently it would be difficult to detect a CNV if it was present.

Table 5.5 Number of reads mapped for each individual sequenced

Individual	Total mapped reads[1]	Total mapped reads in the 640 flanking regions only	Average number of reads per 100bp window in the 640 flanking regions only
NA10846	157,047	73,304	114
NA10847	126,762	61,811	97
NA10859	123,662	58,234	91
NA11881	140,816	65,667	103
NA11882	136,863	64,117	100
NA12144	140,816	65,667	103
NA12145	110,446	52,758	82
NA12146	107,951	54,095	84
NA12239	121,889	59,701	93

1. Total mapped reads is the sum of reads in all windows 100bp long.

5.3.2 Relationship between GC-content and read count

The greatest read count (number of reads per 100bp window) across all samples was observed at approximately 0.55 GC although there was variation in read count between windows with similar GC content. Figure 5.6 contains data for two samples that had the highest (NA10846) and lowest (NA12145) total numbers of reads mapping to the flanking regions. A similar pattern of read count versus GC content was observed for both samples. The read count either side of 0.55 GC gradually reduced with a 50% decrease by 0.3 and 0.7 GC content but there was little data available to study read count in windows with GC content >0.7. If the data is normalized by plotting read counts in each window as a percentage of the read count for the highest count window for that individual, then the GC-read counts graphs are very similar for two different individuals, regardless of the total mapped reads (figure 5.7). These data highlight the variability in read count from sample to sample and variability in read count due to GC content that is corrected for in the CNV detection method described above.

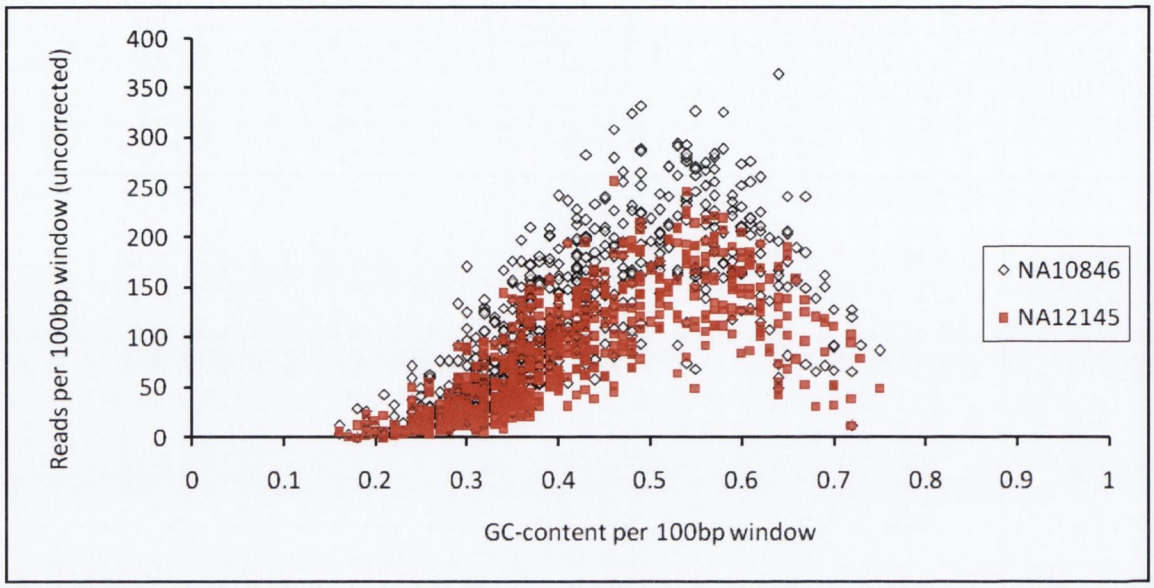


Figure 5.5: Relationship between GC-content and absolute read count in two individuals in non-CNV (flanking) sequence. Each data point represents a read count in a window of specific GC content.

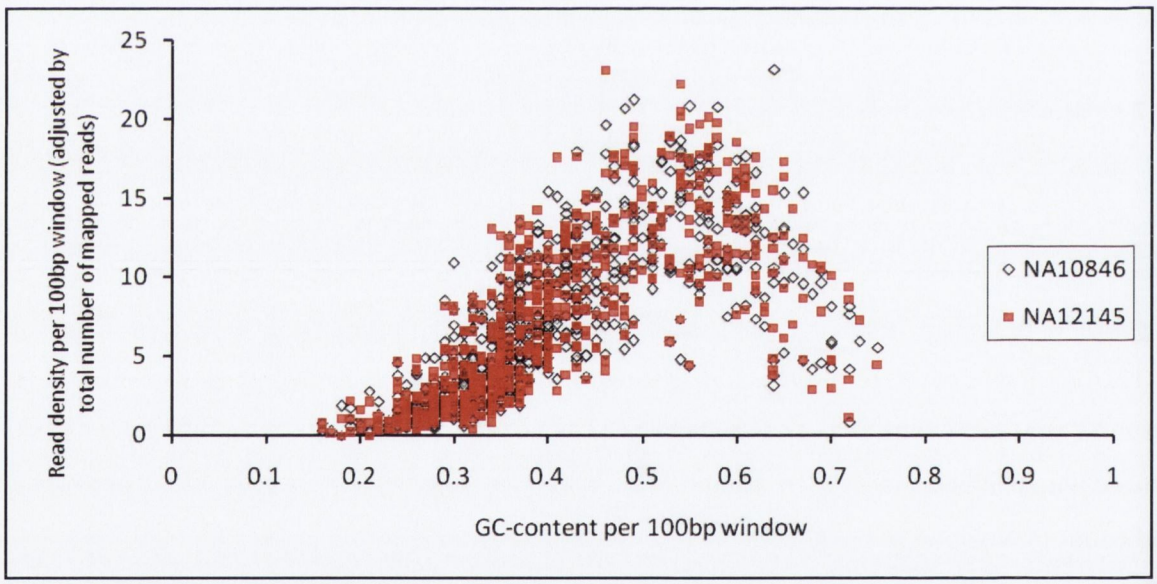


Figure 5.6: Relationship between GC-content and adjusted read count in two individuals in non-CNV sequence. Here the read count has been adjusted for the total number of mapped reads aligning to all 100bp windows analysed.

5.3.3 Results of breakpoint determination using Sanger sequencing

Using PCR and Sanger sequencing, it was possible to determine the breakpoints of twelve of the sixteen targeted CNVs. Here the generated sequence was screened against the human genome using the BLAT function on the UCSC Genome Browser. The breakpoints of a CNV were indicated in the results where there was a break in the homology between submitted and reference sequence. In some instances, the exact location of the breakpoints could not be determined because the surrounding start and stop bases were identical although this only resulting in a loss of resolution to a few bases. It was not possible to determine the breakpoints of four of the targeted sixteen CNVs despite repeated re-design of the PCR primers and different reaction conditions (CNV Ids 337, 468, 682 and 2546).

An unexpected discovery made through the Sanger sequencing across the targeted CNVs was the identification of a short internal sequence in three of the CNVs ranging in length from 5bp to 18bp. In two CNVs the internal sequence aligned to sequence in the middle of the deleted region (CNV 84, sequence: AAATAAAATAATTTTATT and CNV 2215, sequence: GATGG). Thus such CNVs could be interpreted as two distinct variants which are in high LD. For a third CNV (1610), the internal sequence (GTTAAAGATCAATC) was not present in deleted region; rather it aligned to several other distant regions of the genome.

5.3.4 Comparison of SNP-array breakpoints with true breakpoints

The true breakpoints identified for twelve of the sixteen targeted CNVs were compared with those estimated by the study by McCarroll *et al* (2008). These data are presented in table 5.6 and indicate that most of the CNV lengths estimated by McCarroll *et al* (2008) were several thousand bases too small. This is probably a reflection of probe density on the genotyping chip. Conversely the length of one CNV was found to be over-estimated by approximately 7Kb (CNV ID 1602). This may be because there are two CNVs in the same region which were not distinguishable using the SNP-based approach.

Table 5.6: Accuracy of SNP-array based method at identifying CNV breakpoints[1]

CNV ID	CNV co-ordinates	CNV length	Proximal breakpoint difference (bp)[2]	Distal breakpoint difference (bp)[2]
84	chr1:108,534,848-108,539,019	3,928	+910	+210
435	chr3: 82,951,528-82,955,570	4,042	+3,941	+163
451	chr3: 100,426,494-100,432,207	5,713	+645	-65
703	chr4: 156,022,461-156,022,541	80	+6	-13
724	chr4: 172,610,986-172,616,001	5,015	+473	-1,492
1343	chr8: 112,363,227-112,366,316	3,089	+228	-1,880
1602	chr10: 56,352,786-56,354,367	1,581	+310	+8,238
1610	chr10: 58,879,793-58,881,860	2,067	+718	-863
2215	chr16: 86,670,231-86,673,169	2,938	+993	-277
2326	chr18: 36,513,891-36,520,749	6,858	+527	-1,362
2529	chr21: 43,794,795-43,797,732	2,937	-30	-492
2575	chr22: 37,624,053-37,628,634	4,581	+1,148	-1,784

1. Only CNVs for which the exact breakpoints were identified by Sanger sequencing are presented here.

2. The breakpoint difference is stated relative to the genome. Thus for the proximal breakpoints, a '+' prefix indicates that the breakpoint estimated by either McCarroll *et al* (2008) or the NGS was further upstream than the true breakpoint.

5.3.5 Results of CNV detection by NGS analysis

The CNV detection method as described above was applied to NGS data generated at each of the targeted CNV loci. This section describes the results by displaying for an individual CNV: (a) a plot of target region included the position of baits used for the NGS, reported McCarroll CNV breakpoints, the NGS-identified breakpoints and the Sanger sequencing-identified breakpoints, (b) a plot of the Sanger sequencing where available, (c) a plot of the normalized read count in each of the nine HapMap samples and (d) a plot of the mean normalized read count versus the McCarroll reported genotypes for the nine HapMap samples. Data for a subset of CNVs that represent the range of results observed for the CNVs are reported in the chapter. Data plots for the remaining CNVs are in Appendix A.

CNV 84

The 3,261bp CNV reported by McCarroll *et al* (2008) was completely baited including 4,526bp upstream and 2,053bp downstream (figure 5.8a). By Sanger sequencing of an individual who was homozygous for the deletion (NA12146) the CNV was found to be 910bp longer at the proximal end and 210bp shorter at the distal end than reported by McCarroll *et al* (2008). In addition, a sequence of 17bp was identified between the breakpoints of the CNV. This sequence mapped to chr1:108,537,326-108,537,343 which is located within the CNV. Thus CNV 84 could be regarded as two separate CNVs. The distal half only of CNV 84 had been identified in a previous study previously and submitted to dbSNP (Mills *et al* 2006), the position of which was in agreement with that determined by in-house Sanger sequencing. The proximal deletion of CNV 84 may not have been detected by Mills *et al* (2006) because only the distal deletion was present in the individual analysed. However, this is unlikely given that all of the individuals with deletions at CNV 84 had both sections deleted (as determined by NGS analysis). Thus it is more likely that the detection of CNV 84 by Mills *et al* (2006) was incomplete. Using read count data generated from NGS, the CNV detection algorithm identified a single CNV which was a deletion (Kruskal-Wallis $p=3.56\times10^{-36}$, figure 5.8b+c). The breakpoints of CNV 84 as determined by NGS analysis were 116bp longer at the proximal end and 45bp shorter at the distal end than that of the actual CNV (as determined by Sanger sequencing). Because the NGS data was analysed in 100bp windows, this is the limit of resolution for the determination of breakpoints. NGS misidentified the proximal breakpoint by 116bp which is likely because the window immediately before the breakpoint had a GC-content of 0.23 and thus the low number of reads associated with this departure from the optimum GC of 0.55 probably reduced the detection accuracy. The mean normalized read count of CNV 84 for each of the nine individuals analysed was concordant with the genotype calls determined by McCarroll *et al* (2008). The NGS analysis of CNV 84 was thus more accurate than the SNP chip-based method in mapping the exact breakpoints although NGS did not identify the short undeleted internal sequence.

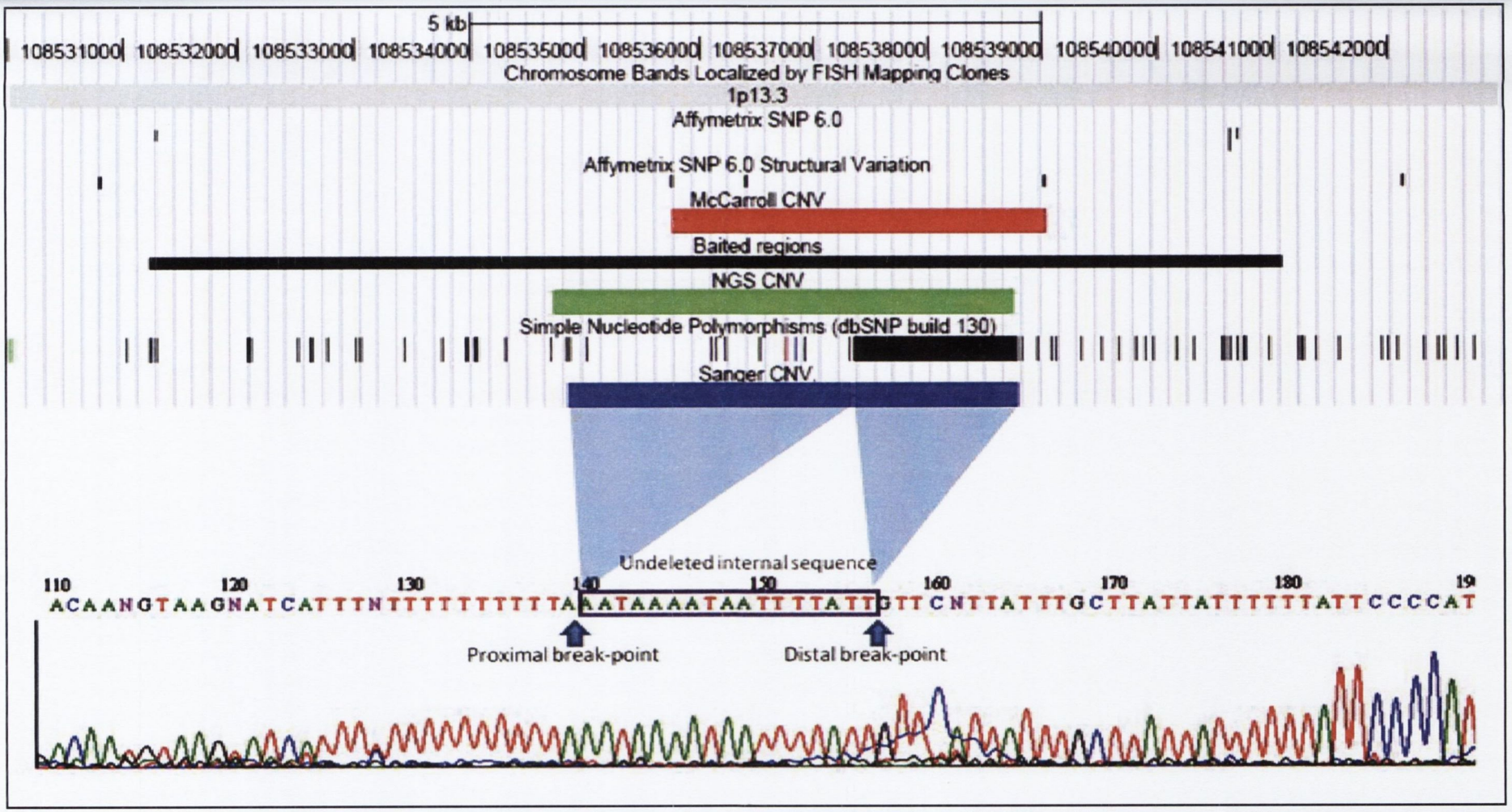


Figure 5.7a: CNV 84 on the UCSC Genome Browser showing locations determined by McCarroll *et al* (2008) and in-house using NGS and Sanger sequencing. Also shown is the region baited and variants in dbSNP (130) including an indel which corresponds to the distal half of CNV 84. The lower panel shows the electropherogram and sequence generated from sequencing across the CNV breakpoints in an individual homozygous for the deletion.

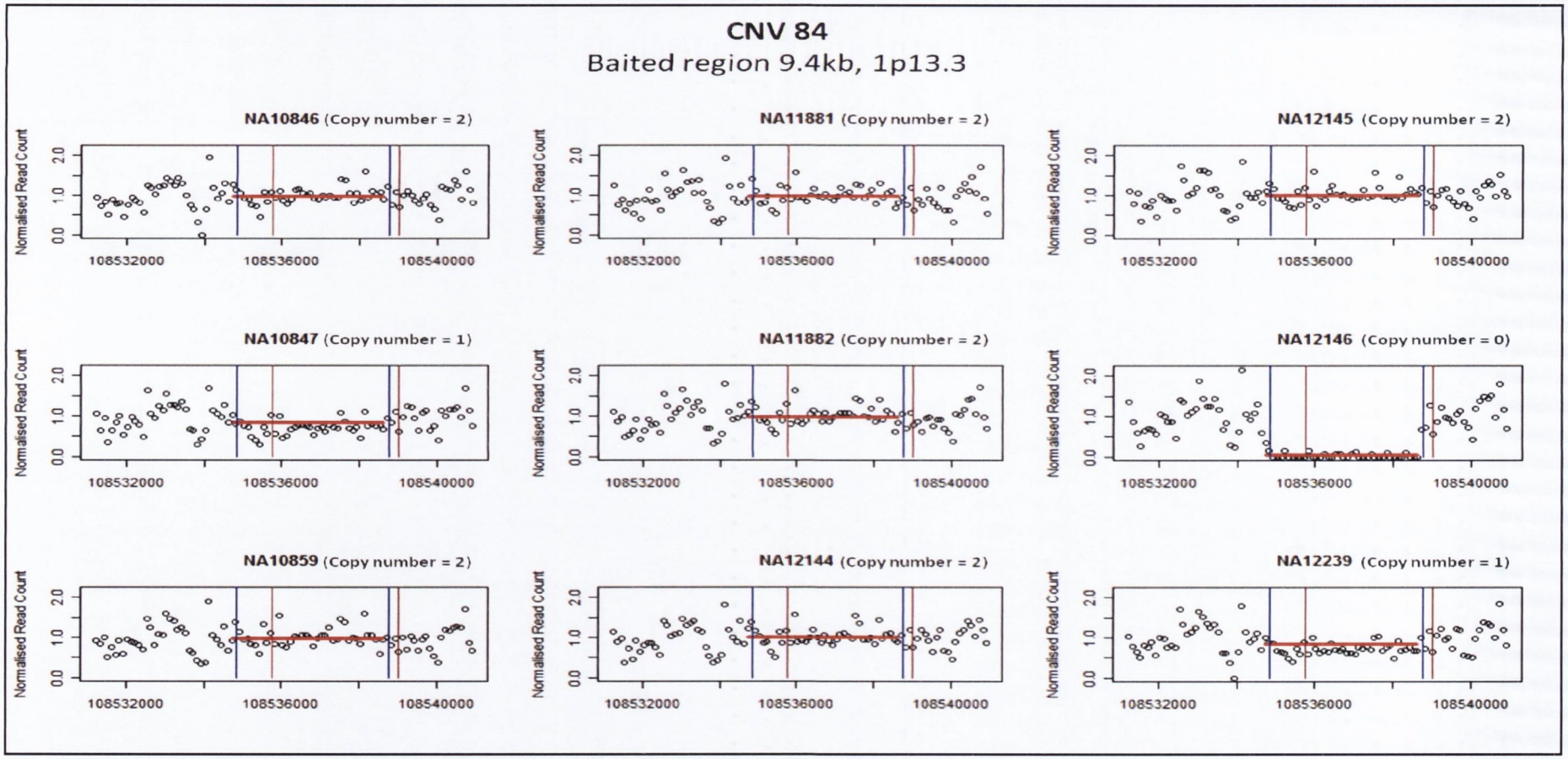


Figure 5.7b: Detection of CNVs from normalised NGS read count data. The x-axis is distance on chromosome 1. The y-axis is the normalised read count. Each data point represents the normalised read count for a 100bp window. Horizontal red bars show the position of the CNV determined by the DNAcopy program using normalised NGS read count data. Vertical pink lines indicate the position of the CNV estimated by McCarroll *et al* (2008) using SNP genotyping. Vertical blue lines indicate the position of the CNV as determined by Sanger sequencing which represents the most accurate of all methods used.

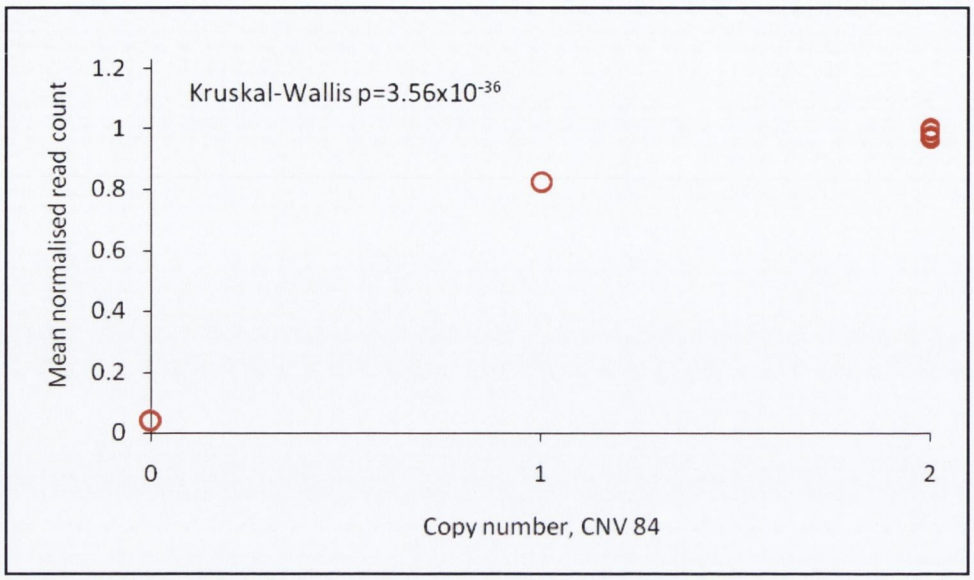


Figure 5.7c: Scatter-plot showing relationship between copy-number at CNV 84 (determined by McCarroll *et al* (2008)) and mean normalised read count across windows in the CNV determined by NGS analysis.

CNV 435

The 264bp CNV reported by McCarroll *et al* (2008) was completely baited including 2,029bp upstream and 2,027bp downstream (figure 5.9a). By Sanger sequencing of an individual who was homozygous for the deletion (NA10859) the CNV was found to be 3,941bp longer at the proximal end and 163bp shorter at the distal end than reported by McCarroll *et al* (2008). No internal undeleted sequence was present. CNV 435 had been identified in a previous study and submitted to dbSNP (Mills *et al* 2006), the position of which was in agreement with that determined by in-house Sanger sequencing. Using read count data generated from NGS, the CNV detection algorithm identified a single CNV which was a deletion (Kruskal-Wallis $p=8.17x10^{-14}$, figure 5.9b). The proximal breakpoint of CNV 435 could not be determined by NGS analysis because baits did not extend far enough although. NGS analysis did however indicate that the CNV extended at least 2,209bp proximally from the position determined by McCarroll *et al* (2008). The distal breakpoint as determined by NGS analysis was 69bp longer than the actual CNV (as determined by Sanger sequencing) which was probably due to the very low GC-content in the final window (0.24) affecting the accuracy of the detection method. The mean normalized read count of CNV 435 for

each of the nine individuals analysed was concordant with the genotype calls determined by McCarroll *et al* (2008). Thus, although there was inaccuracy in the initial position of the CNV by McCarroll *et al* (2008), which lead to the true CNV been incompletely baited, the NGS detection method accurately identified the CNV.

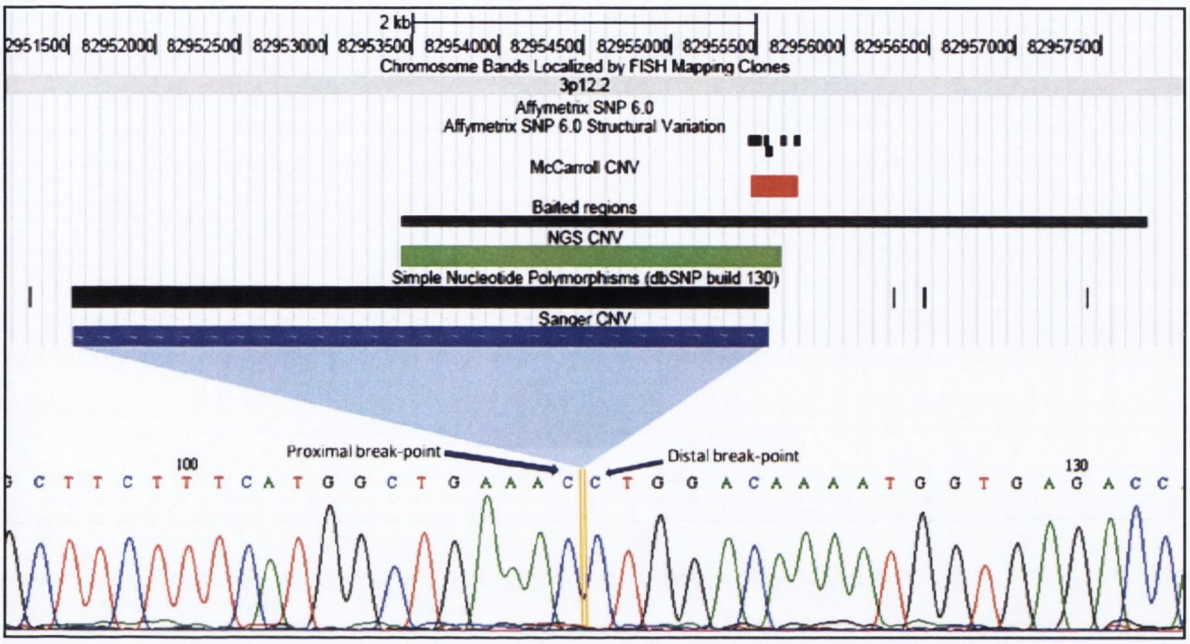


Figure 5.8a: CNV 435 on the UCSC Genome Browser showing locations determined by McCarroll *et al* (2008) and in-house using NGS and Sanger sequencing. Also show is the region baited and variants in dbSNP (130) including an indel which corresponds to CNV 435 as determined by Sanger sequencing in-house.

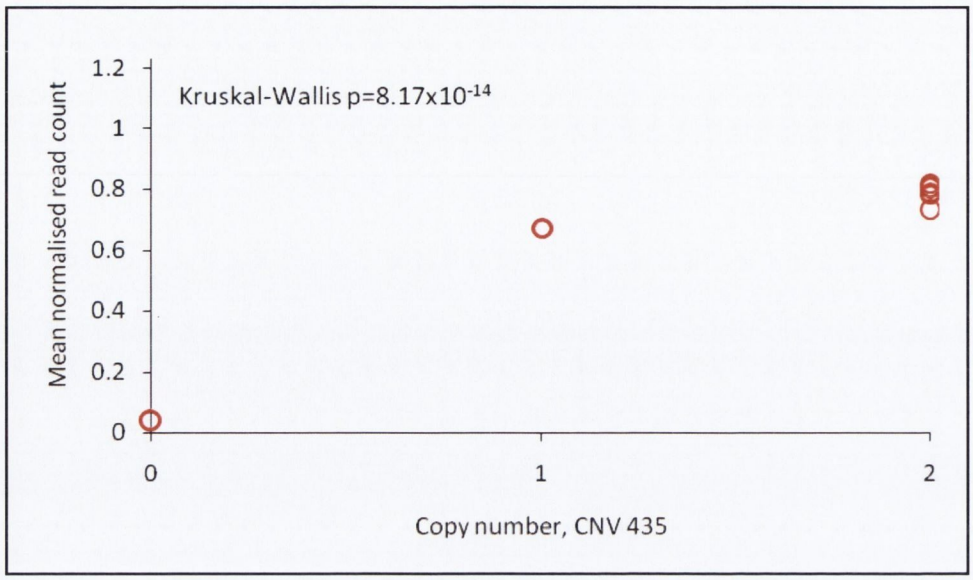


Figure 5.8c: Scatter-plot showing relationship between copy-number at CNV 435 (determined by McCarroll *et al* (2008)) and mean normalised read count across windows in the CNV determined by NGS analysis.

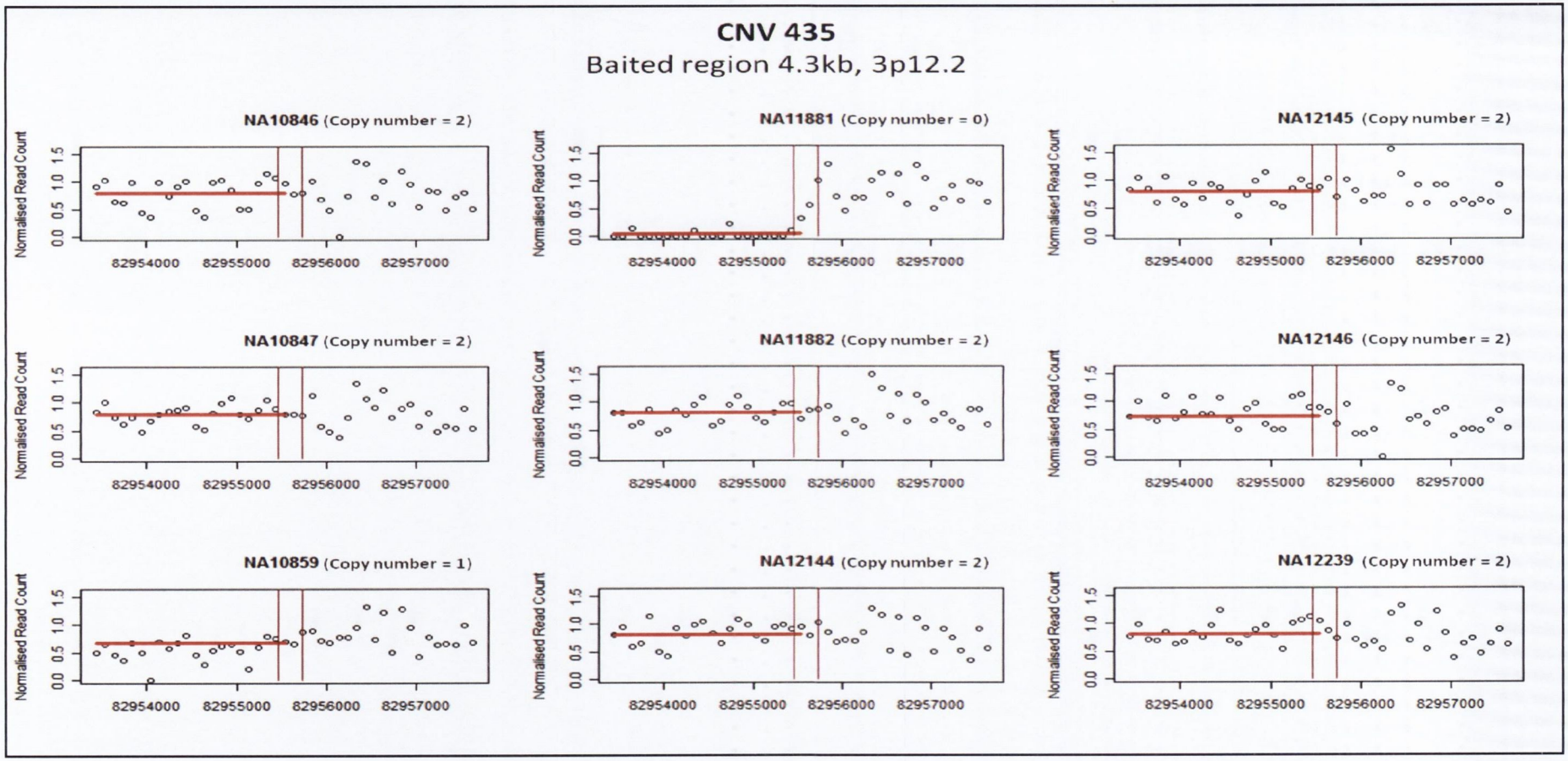


Figure 5.8b: Detection of CNV 435 from normalised NGS read count data. The x-axis is distance on chromosome 3. The y-axis is the normalised read count. Each data point represents the normalised read count for a 100bp window. Horizontal red bars show the position of the CNV determined by the DNAcopy program using normalised NGS read count data. Vertical pink lines indicate the position of the CNV estimated by McCarroll *et al* (2008) using SNP genotyping. It was not possible to establish the exact breakpoints of CNV 435 using Sanger sequencing.

CNV 703

The 73bp CNV reported by McCarroll *et al* (2008) was completely baited including 2,070bp upstream and 2,057bp downstream (figure 5.10). By Sanger sequencing of an individual who was homozygous for the deletion (NA11882) the CNV was found to be 6bp longer at the proximal end and 13bp longer at the distal end than reported by McCarroll *et al* (2008). No internal, undeleted sequence was present. CNV 703 had been identified in a previous study previously and submitted to dbSNP (Mills *et al* 2006), the position of which was in agreement with that determined by in-house Sanger sequencing. Using the NGS read depth data for this region, the detection algorithm was unable to detect a CNV. CNV 703 is only 80bp long which is 20bp smaller than the 100bp analysis window used in this study.

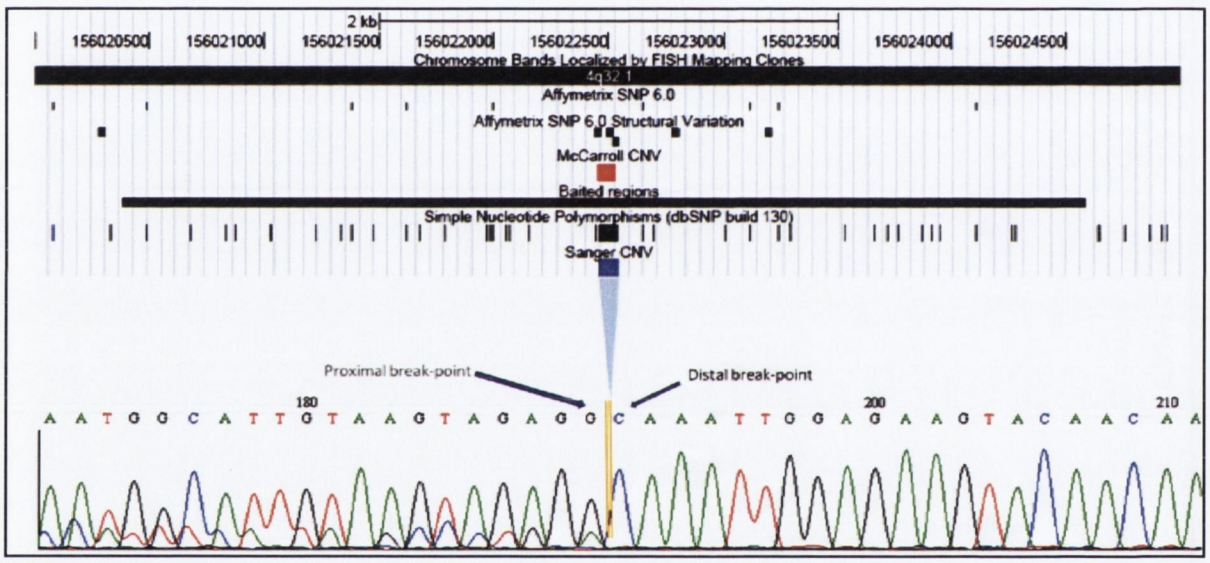


Figure 5.9a: CNV 703 on the UCSC Genome Browser showing locations determined by McCarroll *et al* (2008) and in-house using Sanger sequencing. Also show is the region baited and variants in dbSNP (130) including an indel which corresponds to CNV 703 as determined by Sanger sequencing in-house.

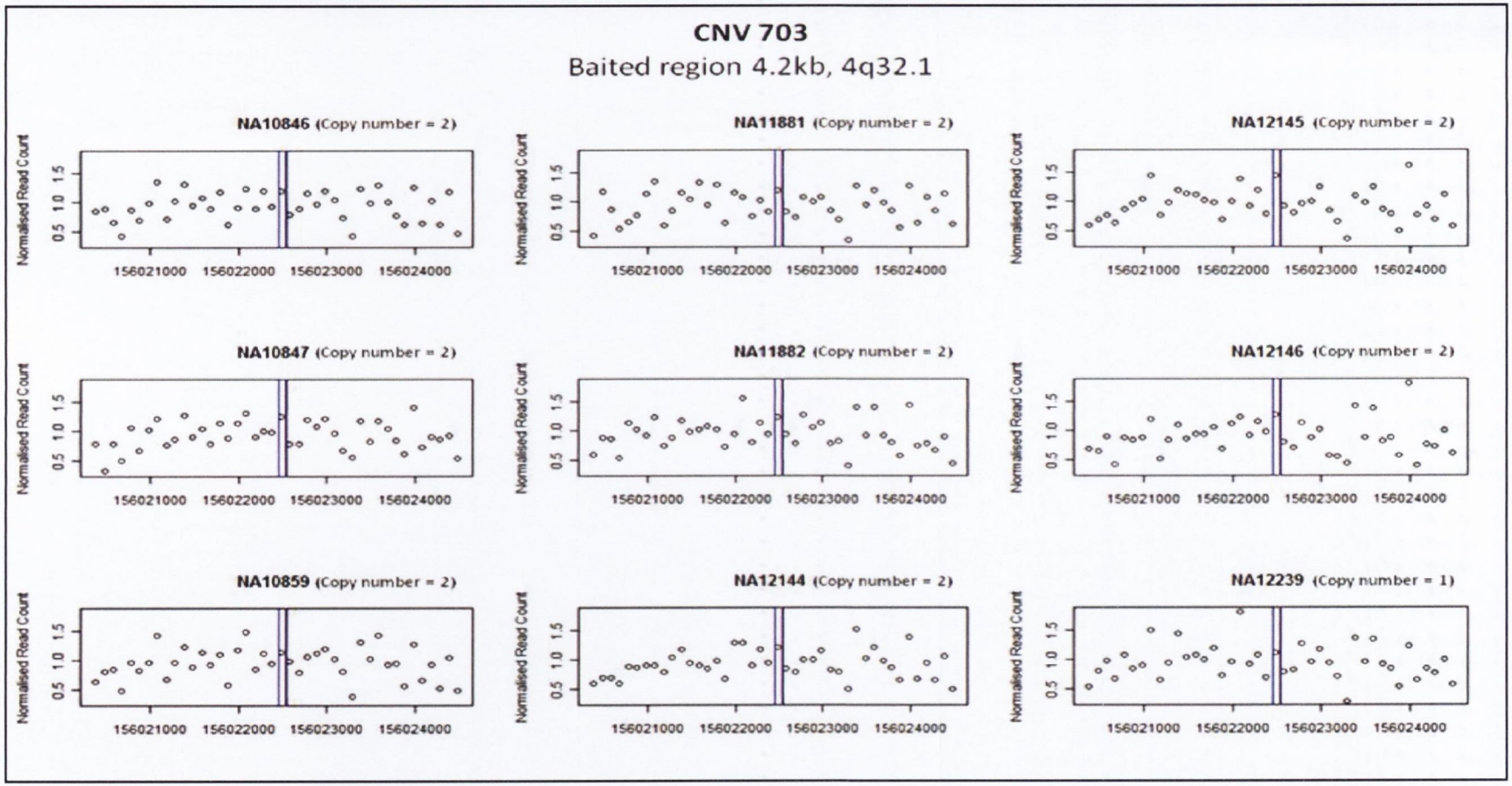


Figure 5.9b: Screening for CNV 703 from normalised NGS read count data. The x-axis is distance on chromosome 4. The y-axis is the normalised read count. Each data point represents the normalised read count for a 100bp window. There are no horizontal red bars in this figure because the DNAcopy software did not detect a CNV in the NGS data. Vertical pink lines indicate the position of the CNV estimated by McCarroll *et al* (2008) using SNP genotyping. Vertical blue lines indicate the position of the CNV as determined by Sanger sequencing which represents the most accurate of all methods used.

CNV 1610

The 486bp CNV reported by McCarroll *et al* (2008) was completely baited including 1,780bp upstream and 2,894bp downstream. By Sanger sequencing of an individual who was homozygous for the deletion (NA12239) the CNV was found to be 718bp longer at the proximal end and 863bp longer at the distal end than reported by McCarroll *et al* (2008). In addition, an internal sequence was identified between the breakpoints of the CNV. The sequence was 14bp long (GTTAAAGATCAATC) and mapped to multiple points in the human genome none of which were within the deleted region of CNV 1610. The CNV had been identified in a previous study and submitted to dbSNP (Mills *et al* 2006), the position of which was in agreement with that determined by in-house Sanger sequencing although there was no report of the internal sequence identified in this study. Using read depth data generated from NGS, the CNV detection algorithm identified two, adjacent and consecutive deletion events which occurred in the same individuals and thus constituted the same CNV (Kruskal-Wallis p=0.019 and 0.007; figure 5.11b). The breakpoints of CNV 1610 as determined by NGS analysis were 62bp longer and 70bp longer at the distal end and proximal ends respectively than that of the actual CNV (as determined by Sanger sequencing). The mean normalized read count of both events at CNV 1610 for each of the nine individuals analysed was concordant with the genotype calls determined by McCarroll *et al* (2008). Thus the NGS CNV detection method was accurate for CNV 1610 even if two separate events were detected in the same individuals.

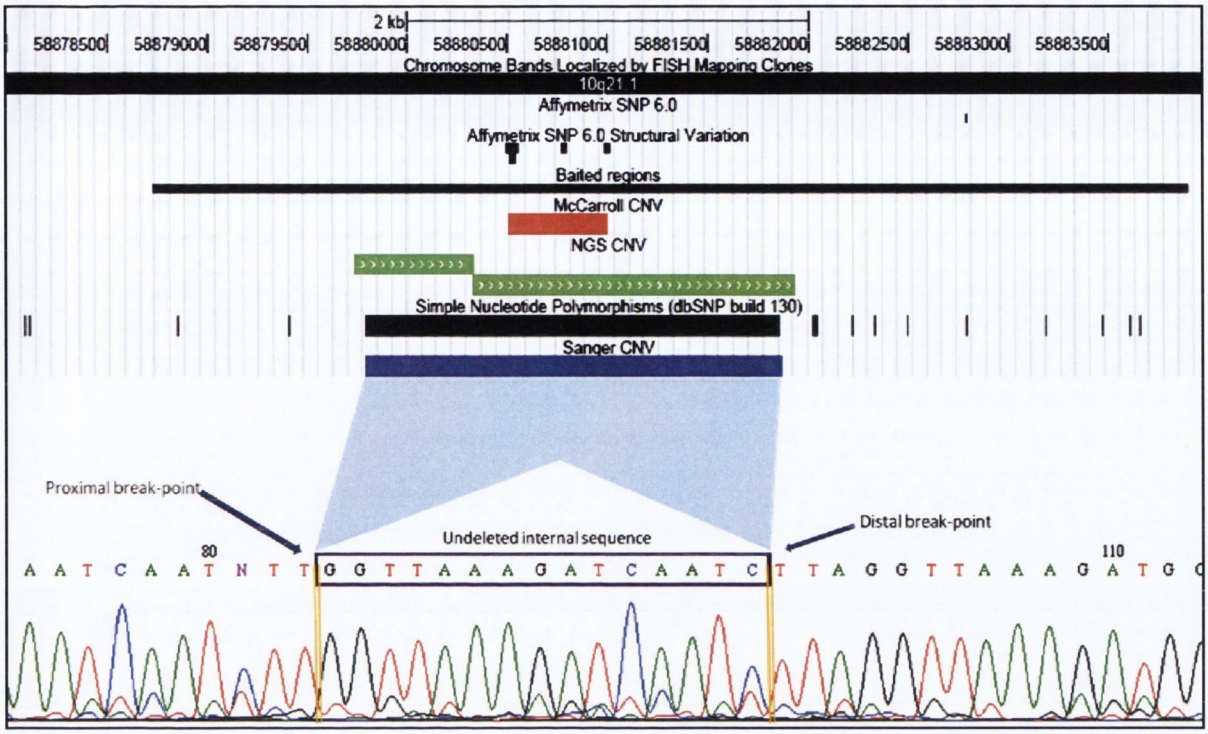


Figure 5.10a: CNV 1610 on the UCSC Genome Browser showing locations determined by McCarroll *et al* (2008) and in-house using NGS and Sanger sequencing. Also show is the region baited and variants in dbSNP (130) including an indel which corresponds to CNV 1610 as determined by Sanger sequencing in-house.

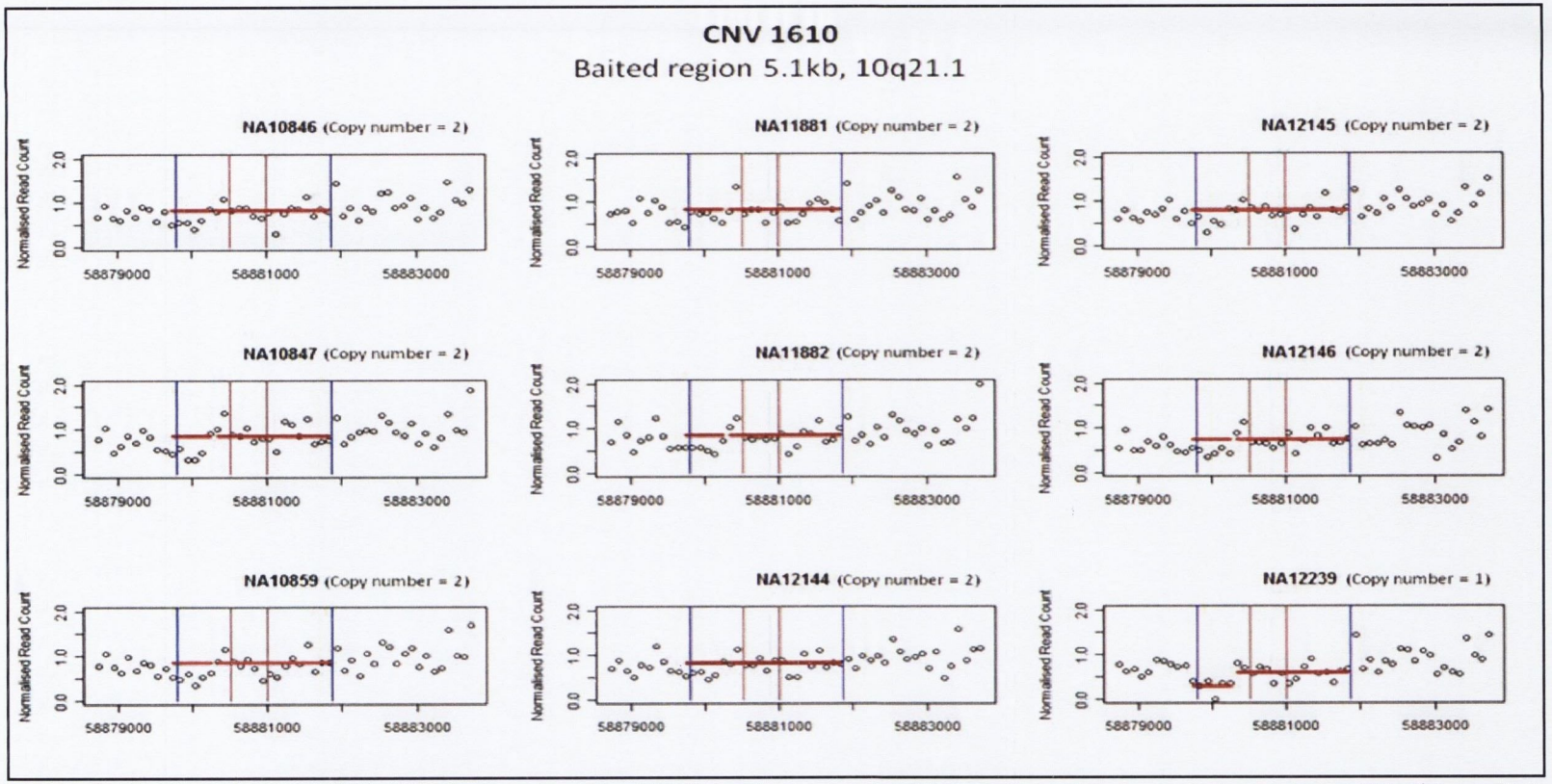


Figure 5.10b: Detection of CNV 1610 from normalised NGS read count data. The x-axis is distance on chromosome 10. The y-axis is the normalised read count. Each data point represents the normalised read count for a 100bp window. Horizontal red bars show the position of the two CNVs determined by the DNAcopy program using normalised NGS read count data. Vertical pink lines indicate the position of the CNV estimated by McCarroll *et al* (2008) using SNP genotyping. Vertical blue lines indicate the position of the CNV as determined by Sanger sequencing which represents the most accurate of all methods used.

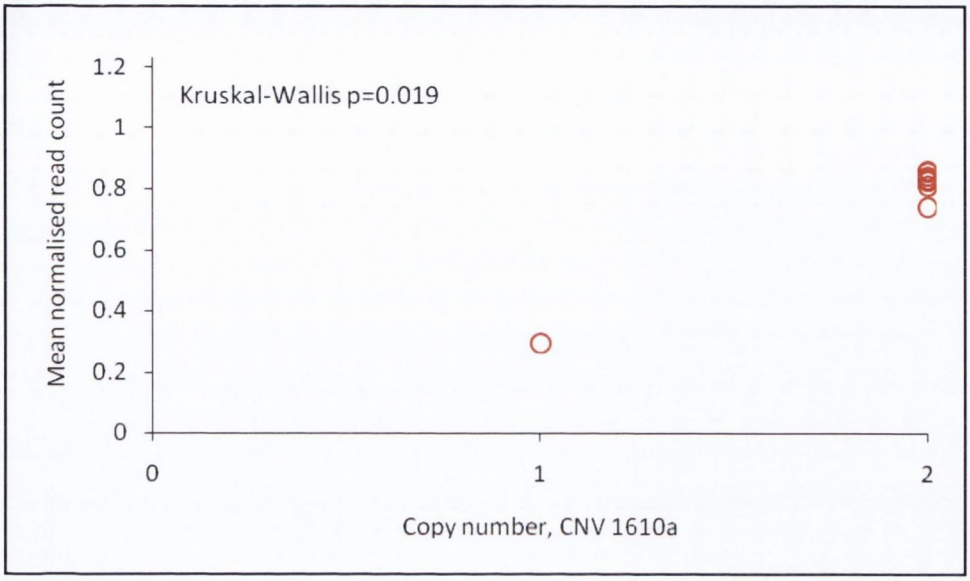


Figure 5.10c: Scatter-plot showing relationship between distal copy-number at CNV 1610 (determined by McCarroll *et al* (2008)) and mean normalised read count across windows in the CNV determined by NGS analysis.

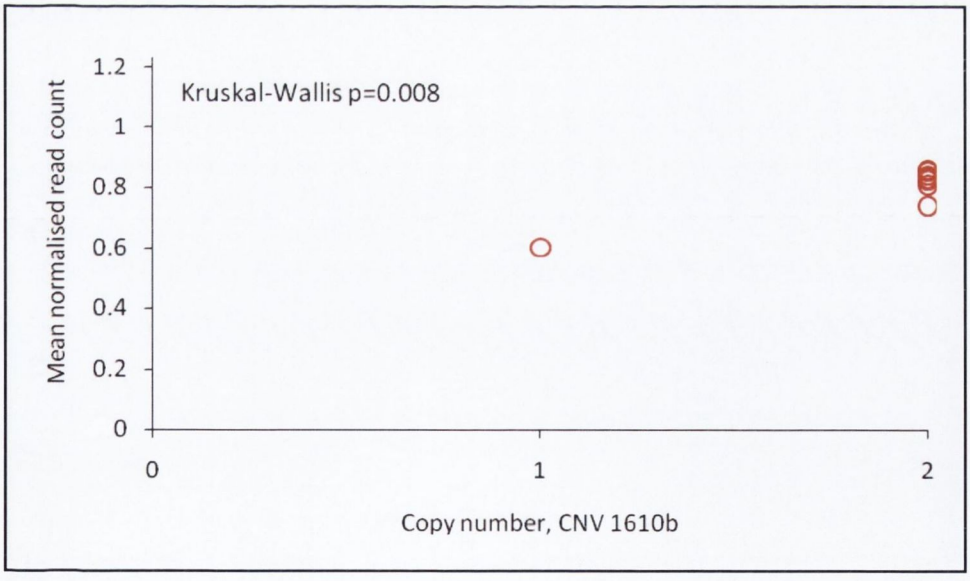


Figure 5.10d: Scatter-plot showing relationship between proximal copy-number at CNV 1610 (determined by McCarroll *et al* (2008)) and mean normalised read count across windows in the CNV determined by NGS analysis.

CNV 2326

Baits covered all but 380bp of the 4,969bp CNV reported by McCarroll *et al* (2008) and included 2,040bp upstream and 3,163bp downstream (figure 5.12A). By Sanger sequencing of an individual who was homozygous for the deletion (NA12145) the CNV was found to be 527bp longer at the proximal end and 1,362bp longer at the distal end than reported by McCarroll *et al* (2008). No internal, undeleted sequence was present. There was no record of CNV 2326 in dbSNP. Although there were two baited regions, separated by 380bp, the read depth data from the NGS was analysed as a single region. A single deletion CNV was identified (Kruskal-Wallis $p=5.48x10^{-16}$, figure 5.12b). The breakpoints of CNV 2326 as determined by NGS analysis were 687bp and 456bp shorter at the proximal end and distal ends respectively than those of the actual CNV (as determined by Sanger sequencing). The mean normalized read count of CNV 2326 for each of the nine individuals analysed was concordant with the genotype calls determined by McCarroll *et al* (2008) (figure 5.13c). Thus for CNV 2326 the NGS method was accurate for the genotype calls with two separate bait groups. However, the estimated breakpoints were 687bp and 456bp shorter at each end that the true breakpoints as determined by Sanger sequencing.

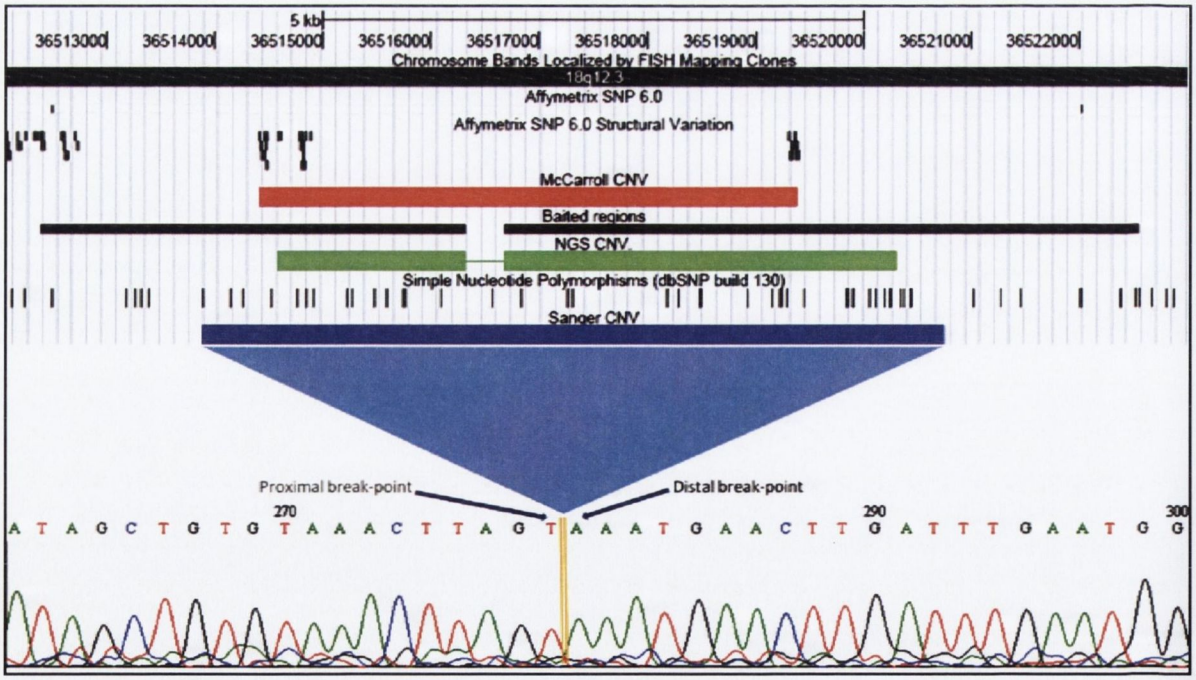


Figure 5.11a: CNV 2326 on the UCSC Genome Browser showing locations determined by McCarroll *et al* (2008) and in-house using NGS and Sanger sequencing. Also shown is the baited region.

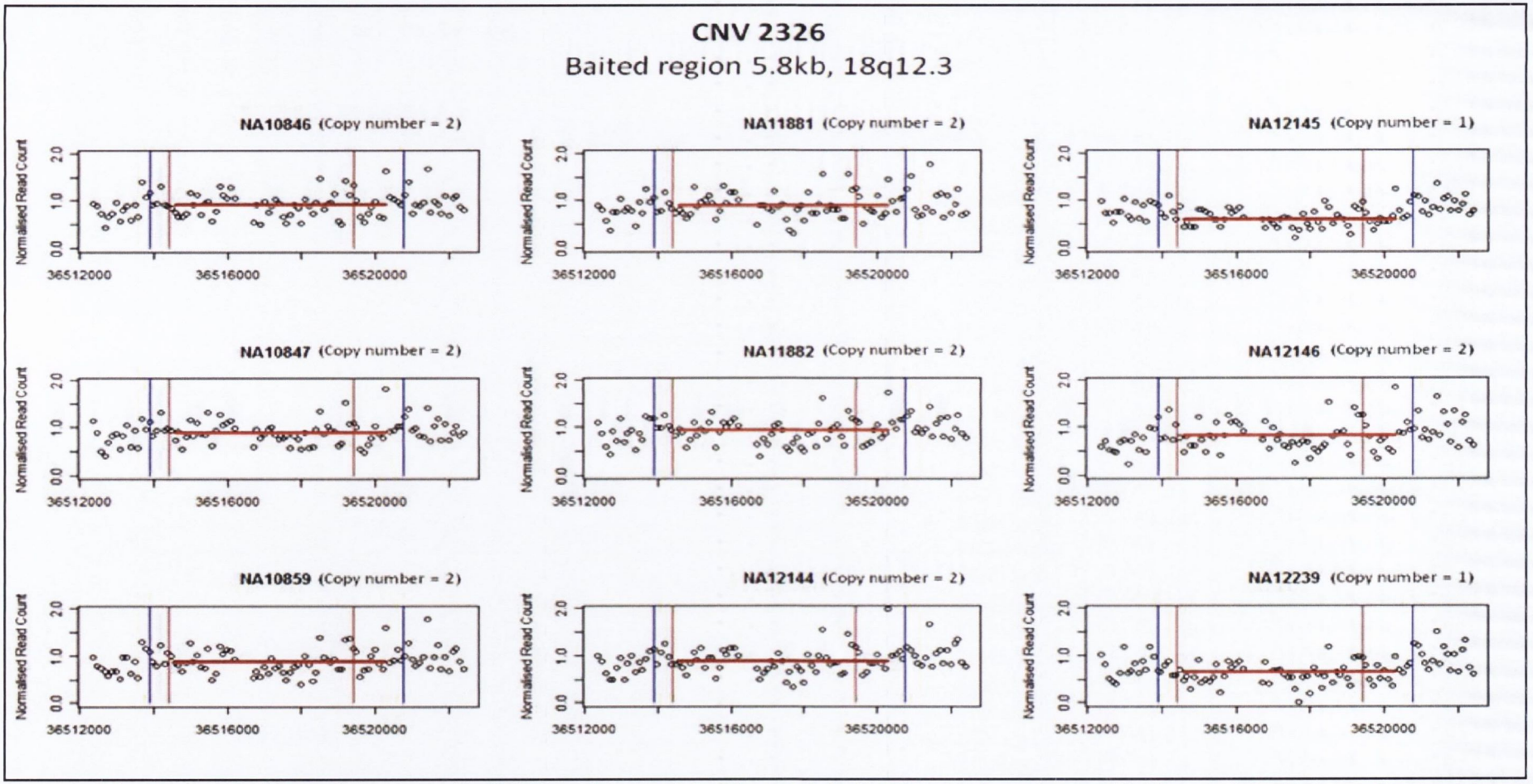


Figure 5.11b: Detection of CNVs from normalised NGS read count data. The x-axis is distance on chromosome 18. The y-axis is the normalised read count. Each data point represents the normalised read count for a 100bp window. Horizontal red bars show the position of the CNV determined by the DNAcopy program using normalised NGS read count data. Vertical pink lines indicate the position of the CNV estimated by McCarroll *et al* (2008) using SNP genotyping.

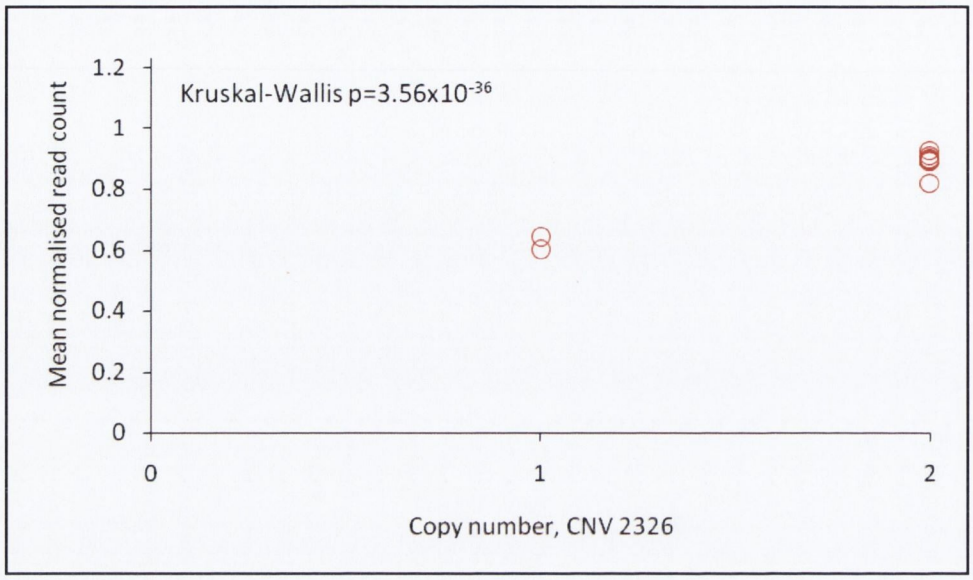


Figure 5.11c: Scatter-plot showing relationship between copy-number at CNV 2326 (determined by McCarroll *et al* (2008)) and mean normalised read count across windows in the CNV determined by NGS analysis.

Data plots for the remaining CNVs are presented in Appendix A.

5.3.5 CNVs with gains of copy number

In addition to the fourteen bi-allelic deletion CNVs, two CNVs which were gains of copy number were also targeted; CNVs 468 and 2546. CNV 468 is 16,841bp long has and copy numbers of 2, 3 and 4 in the CEU (both as determined by McCarroll *et al* (2008)). Independent studies of CNVs have identified both gains and losses of sequence in comparison to a reference genome. No CNV was identified using the NGS CNV detection method, nor was it possible to map any of the breakpoints using Sanger sequencing. CNV 468 is unlikely to be artefactual because it exhibits Mendelian segregation in the CEU sample and because it was detected as a duplication by independent studies. Furthermore, an increase in read depth of coverage across CNV 468 is visible when the read counts are transformed and plotted (figure 5.12). CNV 468 is may not have been detected

using the NGS method because only 840bp were baited. This corresponds to eight 100bp analysis windows, which may not have been sufficient for the DNAcopy program to identify a CNV.

CNV 2546 is 119,296bp long and has copy numbers of 2,3,4,5 and 6 in the CEU (both as determined by McCarroll *et al* (2008)). It was not possible to determine the breakpoints of these using Sanger sequencing, nor did the NGS detection method identify a CNV in the targeted region. Many independent studies identified gains and losses of sequence which overlap CNV 2546. In addition, the only one of these studies to use NGS (following the paired-end method) identified an inversion which overlaps with CNV 2546 (Tuzun *et al* 2005). Thus, CNV 2546 may be an inversion. Because an inversion does not cause any loss or gain of genetic material, it would be hard to detect using the read depth of coverage approach used in this study.

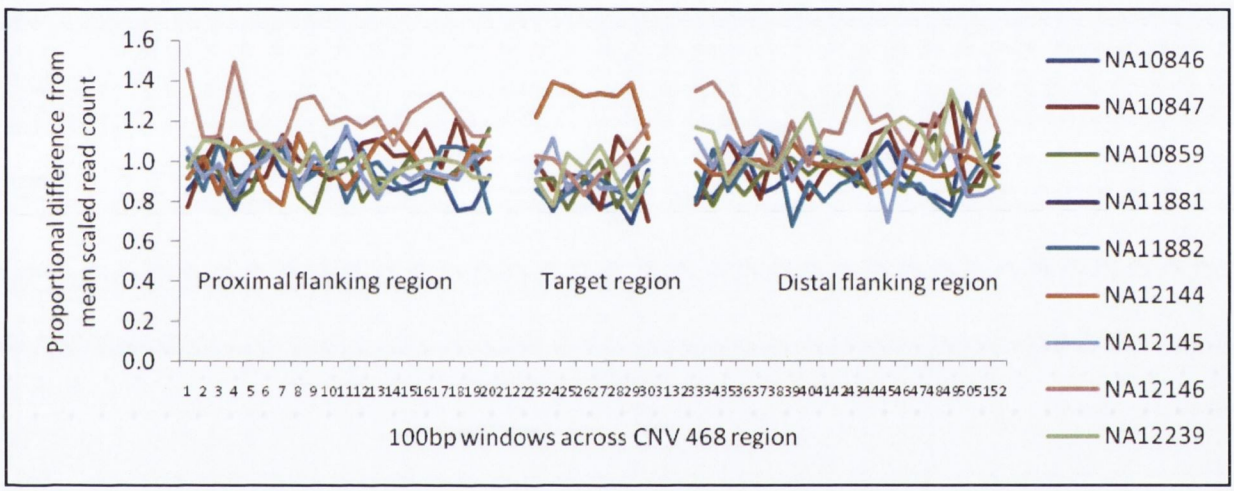


Figure 5.12: Diagram showing change in read depth of coverage in the flanking regions and CNV/target region of CNV 468. Because the flanking regions are so far from the CNV the x-axis has not been plotted on a genomic scale. The values for the y-axis were generated by first scaling the absolute read counts per window by the total number of read counts in all windows (see figure 5.6). Then for each individual the scaled read count was divided by the average scaled read count for all nine individuals. The diagram clearly shows an increased read depth of coverage for individual NA12144 which was the only one reported by McCarroll *et al* (2008) to have a duplication at CNV 468.

5.3.6 Summary of results of CNV detection by the NGS analysis method

Of the sixteen CNVs selected, the NGS detection method identified CNVs at twelve of the sites (Table 5.7). For CNVs 337, 2326, 2529 and 2575 which were targeted by more than one group of baits i.e. which were separated by an 'unbaitable' repetitive region, the read depth was analysed as a single contig in which a single CNV was accurately identified. For CNVs 682, 1343, 1610, 2529 and 2575 multiple events were detected which when combined accurately corresponded to the known position of the CNV. Importantly, where there were two or more events detected for a single CNV, the normalised read count corresponded to the copy-number as previously determined by McCarroll *et al* (2008). The detection of multiple events in a single CNV is likely to be an artefact of the DNAcopy algorithm and does not appear to affect the overall sensitivity of the method. The shortest event detected was at CNV 1602 (Kruskal-Wallis $p=7.99x10^{-7}$) where the deletion was estimated by NGS analysis to be 600bp in length and matched one of the breakpoints perfectly. (The other breakpoint was not baited and therefore no assessment is possible.) At CNV 2575 the position of the CNV was accurately identified in two independent events although a third 'false positive' event was also generated (Kruskal-Wallis p-value 0.002).

The CNVs which were not detected by the NGS method were 468, 703, 2215 and 2546. CNV 468 may not have been detected because it was too small (with only eight 100bp windows to analyse) or because only one individual out of the nine analysed harboured a duplication which in addition was heterozygous. CNV 703 is only 80bp long and thus would not have been detected by the DNAcopy algorithm which requires signal reductions in multiple consecutive windows to detect an event. CNV 2215 was completely missed by baits and so it was impossible to detect a CNV here. CNV 2546 was reported by McCarroll *et al* (2008) as a duplication although it is more likely to be an inversion and thus not detectable using the NGS-based CNV detection method.

The CNV detection method had a resolution of 100bp because reads were counted and analysed in windows of this size. Where baited regions over-lapped a known breakpoint, nine breakpoints were estimated to be within 100bp of the true breakpoints. Seven of the estimated were inaccurate by over 100bp. The biggest difference was for CNV 2326 where the estimated proximal breakpoint was over 600bp upstream from the true breakpoint. A comparison of the breakpoints estimated by NGS analysis with the true CNV breakpoints is shown in table 5.8.

Table 5.7: Overview of results of NGS analysis

CNV ID	Type	CNV identified	Kruskal-Wallis p-value	Comment
84	Deletion	Yes	$3.56x10^{-36}$	
337	Deletion	Yes	$2.32x10^{-41}$	
435	Deletion	Yes	$8.17x10^{-14}$	
451	Deletion	Yes	$1.57x10^{-18}$	
468	Duplication	No	N.S.	CNV possibly not detected because it was short (800bp), because heterozygous duplications are not well detected or because only one individual had the duplication.
682	Deletion	Yes	$7.27x10^{-7}$, 0.0026, $1.26x10^{-9}$	
703	Deletion	No	N.S.	CNV only 80bp so possibly too short to be identified using the NGS method
724	Deletion	Yes	$1.07x10^{-47}$	
1343	Deletion	Yes	$6.43x10^{-15}$, $9.49x10^{-5}$	
1602	Deletion	Yes	$7.99x10^{-7}$	
1610	Deletion	Yes	0.019 and 0.007	
2215	Deletion	No	N.S.	The CNV was completely missed by baits.
2326	Deletion	Yes	$7.18x10^{-11}$, $7.05x10^{-9}$	
2529	Deletion	Yes	$7.18x10^{-11}$, $7.05x10^{-9}$,	
2546	Duplication	No	N.S.	Although reported by McCarroll *et al* (2008) as a duplication, pair-end sequencing has recently identified an inversion overlapping this site.
2575	Deletion	Yes	$2.01x10^{-3}$, $1.37x10^{-15}$ and $4.45x10^{-24}$	The first CNV identified here (p-value=$2.01x10^{-3}$) was a false positive.

Table 5.8: Accuracy of SNP-array based and NGS at identifying CNV breakpoints[1]

CNV ID	CNV co-ordinates	CNV length	Proximal breakpoint difference (bp)[2]		Distal breakpoint difference (bp)[2]	
			McCarroll	NGS	McCarroll	NGS
84	chr1:108,534,848-108,539,019	3,928	+910	-116	+210	-45
435	chr3: 82,951,528-82,955,570	4,042	+3,941	n/a	+163	+69
451	chr3: 100,426,494-100,432,207	5,713	+645	-65	-65	+96
703	chr4: 156,022,461-156,022,541	80	+6	n/a	-13	n/a
724	chr4: 172,610,986-172,616,001	5,015	+473	n/a	-1,492	+163
1343	chr8: 112,363,227-112,366,316	3,089	+228	+40	-1,880	n/a
1602	chr10: 56,352,786-56,354,367	1,581	+310	n/a	+8,238	+43
1610	chr10: 58,879,793-58,881,860	2,067	+718	-62	-863	+70
2215	chr16: 86,670,231-86,673,169	2,938	+993	n/a	-277	n/a
2326	chr18: 36,513,891-36,520,749	6,858	+527	+687	-1,362	-456
2529	chr21: 43,794,795-43,797,732	2,937	-30	+330	-492	+64
2575	chr22: 37,624,053-37,628,634	4,581	+1,148	-138	-1,784	-404

1. Only CNVs for which the exact breakpoints were identified by Sanger sequencing are presented here.

2. The breakpoint difference is stated relative to the genome. Thus for the proximal breakpoints, a '+' prefix indicates that the breakpoint estimated by either McCarroll *et al* (2008) or the NGS was further upstream than the true breakpoint.

5.4 Discussion

The aim of the present study was to develop a method for detecting CNVs in target enrichment data generated from NGS experiments. There were two reasons for this study: firstly, rare CNVs have a significant role in schizophrenia genetic aetiology and can inform on the biology of the disorder. Secondly, NGS is a recently introduced ultra high-throughput technology that when applied to DNA samples has the capacity to detect single nucleotide variants and has the potential to detect copy-number variation at greater resolution that existing high-throughput methods such as aCGH or SNP chip-based methods. This capacity exists because the quantity of reads generated during a sequencing reaction will vary depending on the copy number of template DNA. Thus a localised loss or gain in read depth of coverage corresponds to deletion or gain of sequence. This has been demonstrated previously using sequence data from whole genomes (Yoon *et al* 2009, Alkan *et al* 2009). The method described in this study method could potentially be used in target enriched sequence data, e.g. candidate genes or whole exome.

In the present study, the NGS was used to investigate a test set of sixteen previously identified CNVs ranging in size from 70bp to 119,009bp. The targeted CNVs had been previously identified in the CEU samples using a specialised genotyping array (Affymetrix 6.0 SV). The locations and copy-numbers of each CNV were determined by data generated from the genotyping array and the results made publicly-available on the Database of Genomic Variants (McCarroll *et al* 2008). Fourteen of the targeted CNVs were deletions in which individuals possessed two, one or zero copies of the deleted sequence. All of the deletion CNVs segregated in a Mendelian fashion and were in Hardy-Weinberg equilibrium (HWE) in the CEU sample. The two other targeted CNVs had been previously identified as duplications of which it was possible to show that one of them, CNV 468, Mendelised and was in HWE (the other, CNV 2546, having multiple alleles was difficult to assess). All targeted CNVs were identified independently by other research groups using hybridization arrays. The study by McCarroll *et al* (2008) made no attempt to map the exact breakpoints of the reported CNVs and thus the start and stop co-ordinates were only possible at the resolution of the probes on the SNP-array. A further complication to accurately mapping CNVs using SNP array data is that not only does the position of the probes determine whether a CNV is detected but also the binding sites of the primers used for the PCR reaction.

Read depth was determined across each CNV by counting the number of reads in non-overlapping 100bp-long windows. This window size was used because initial investigations showed that 50bp windows were subject to great variation in read depth and that 200bp was too large to

detect the smaller CNVs. In addition, 100bp windows have been used in NGS whole genome sequence CNV detection methods (Yoon *et al* 2009). Initial visualisation of the changes in read depth across CNVs clearly indicated almost total lack of coverage at homozygous deletion sites. In non-deleted sequence, the read depth varied such that loss in coverage caused by heterozygous deletions was more subtle and complicated by the effect of local GC-content of the read depth.

This study has shown how GC-content influences depth of coverage. From the data generated, 0.55 is the optimum GC content. A 50% loss in coverage occurs after sequence <0.3 and >0.7 GC. Because fewer reads will map to sequence of very high or very low GC-content this will not only affect the sensitivity to detect CNVs but also SNPs. Interestingly, the optimum GC-content for coverage using the hybrid selection method is ~0.55 (Gnirke *et al* 2009) compared with non-targeted sequencing (also using the Illumina Genome Analyser II) for which the optimum GC-content is ~0.4 (Alkan *et al* 2009, figure 5.1).

The analysis of NGS read depth for CNVs firstly involved correction for GC-content based upon data generated from two 2Kb flanking regions which were located away from any reported CNVs and repetitive sequence. For the correction method applied to read counts, windows were categorised according to their GC-content and the reads counts per window were adjusted accordingly. The identification of CNVs in adjusted read-count data was performed using the DNAcopy script which was implemented on the R software package. This approach successfully identified 12 out of the 16 CNVs using a total of 158,000bp of sequence in each of nine individuals Of the four CNVs which were not detected, CNV 2215 was not covered at all by baits and CNV 2546, initially reported as a duplication, was probably an inversion. These figures are comparable to Yoon *et al* (2009) who detected 33-89% of CNVs using chromosome 1 sequencing data (249Mb) in five individuals.

In the present study heterozygous deletions were detected as a series of windows with consistently lower read depth than that of the surrounding undeleted sequence. Most importantly CNV could be detected even when there were no individuals who were homozygous for the deletion (CNVs 337, 451, 682, 1610, 2326). The shortest event detected was at CNV 1602 where the deletion was estimated by NGS analysis to be 600bp in length and accurately estimated the position of the distal breakpoint. (The proximal breakpoint was not covered by the bait and therefore no assessment was possible.) Thus 600bp represents the lower limits of the detection threshold for this method. At CNV 2575 the position of the CNV was accurately identified although one 'false positive' event was also generated. The Kruskal-Wallis p-value for this event was 0.002 which is a high p-value compared to many of the others although similar to those which identified CNVs 682 (p=0.0026) and 1610 (p=0.019).

The NGS method used in the present study had a resolution of 100bp because reads were counted and analysed in windows of this size. Where baited regions over-lapped a known breakpoint, the NGS analysis method identified the position of nine breakpoints to within 100bp of the true breakpoints. Seven of the breakpoints estimated using the NGS method were inaccurate by over 100bp. The biggest difference was for CNV 2326 where the estimated proximal breakpoint was over 600bp upstream from the true breakpoint. A possible reason for such inaccuracies may be runs of consecutive 100bp windows with extreme highs or lows of GC-content for which fewer reads can be generated and thus provide less data for the analysis.

Determination of the exact breakpoints of twelve of the targeted CNVs not only allowed for comparison with SNP-chip and NGS results but also surprisingly identified short undeleted sequences in the middle of the deletion. Because the maximum length of an internal sequence in this study was 18bp (CNV 84) it is unlikely that it would be detected in the NGS analysis. Recently Conrad *et al* (2010) have fine-mapped the breakpoints of 400 common CNVs which not only match those identified in the present study but also identify internal sequence in approximately one-third of all common CNVs.

The CNVs which were not identified by NGS analysis in this study were either small or duplications. The smallest CNV tested was 80bp long (after confirmation by Sanger sequencing) and in this case the decrease in read depth in a single window was not detected as a CNV by DNAcopy. This would appear as a potential problem when applying this method to targeted re-sequencing of candidate genes where individual exons <100bp in length are being analyzed. However, this depth of coverage CNV detection method would be applied in parallel with paired-end mapping CNV detection methods. Paired-end methods should detect CNVs that have at least one breakpoint within the boundary of a target exon and are suited to identifying indels and small CNVs. The depth of coverage CNV detection method being developed here is required to detect larger CNVs that knock out an entire target exon where paired sequence reads would not map. A further modification of the current method could be to analyse windows of overlapping rather than adjacent sequence.

The capacity of NGS to detect deletions at the scale of a single exon thus remains to be determined. The present study has provided pilot data on how CNVs affect read depth of coverage generated by NGS using the Illumina Genome Analyser II. In an on-going extension of this work, this method is being applied to further publicly-available data. Firstly, whole exome data is now available on a number of test samples including eight HapMap samples (Ng *et al* 2009b). Secondly, Conrad *et al* (2010) have produced a set of genotypes for validated common CNVs as part of a large WTCCC CNV study. Included in this data-set are CNV genotypes for HapMap samples. I have cross-referenced these CNVs with known exon positions to identify CNVs which should be present in

exome sequencing data. My colleagues, Dr Cormican and Dr Heron, are currently testing the sensitivity and specificity of our method at detecting these CNVs in the HapMap whole exome sequence data.

Appendix to Chapter 5: CNV detection by Next-generation sequencing

Sections A5.1 and A5.2 were undertaken and written by Dr Elizabeth Heron

A5.1 Normalisation for Sample and GC-Content

As we are not interested in the variability of read count due to GC-content, as it may mask the variability due to copy number variation which is our primary interest, it is important to eliminate or account for this GC-content variability. Also, the total read counts can vary from sample to sample and it is important to also take this into account. In order to do this we start by examining the read counts. The most natural distribution for count data is a Poisson distribution, which we will consider here. As in some cases we may have limited data, we do not consider each GC-content percentage, but rather we bin the GC-content into $g = 1, \ldots, G$ bins. For each sample i, we let r_{ijg} denote the j^{th} read count having GC-content in bin g, $j = 1, \ldots, N_{ig}$, where the number of read counts with GC-content g in sample i is given by N_{ig}. We model this read count as being Poisson distributed with mean λ_{ig}:

$r_{ijg} \sim \text{Poisson}(\lambda_{ig})$, $j = 1, \ldots, N_{ig}$.

If we can estimate this mean read count λ_{ig} for person i, bin g, then we can use this estimate to normalise the read count data. A straight forward and computationally simple approach is to use Bayes theorem which allows for the estimation of the posterior distribution of λ_{ig} employing the use of a conjugate Gamma prior distribution as follows:

$$\begin{aligned} P(\lambda_{ig}|\{r_{ijg}\},\alpha,\beta) &\propto \prod_{j=1}^{N_{ig}} L(r_{ijg}|\lambda_{ig})P(\lambda_{ig}) \\ &= \prod_{j=1}^{N_{ig}} \frac{exp(-\lambda_{ig})\lambda_{ig}^{r_{ijg}}}{r_{ijg}!} \frac{exp(-\beta\lambda_{ig})\lambda_{ig}^{(\alpha-1)}}{\Gamma(\alpha)} \\ &\sim \text{Gamma}(\sum_{j=1}^{N_{ig}} r_{ijg} + \alpha, \beta + N_{ig}), \end{aligned}$$

where $P(.|.)$ denotes the posterior distribution, $L(.|.)$ denotes the likelihood function and $P(.)$ denotes the prior distribution. A uniform prior with $\alpha, \beta = 1$ is a reasonable non-informative prior choice for λ_{ig}.

To normalise the data, for each sample, each read count with GC-content in bin g is then divided by the posterior estimate of λ_{ig}. This method of normalisation accounts for both the sample variability and the variability due to GC-content.

A5.2 CNV Detection

After the interval read counts have been normalised for both sample and GC-content variability, the next step is to detect potential copy number variable regions across samples. We utilise 1 the R package: DNAcopy in this process. DNAcopy is a package designed for analysing array Comparative Genomic Hybridization (array CGH) data. The function of DNAcopy that is most useful for our purposes with sequencing read counts is the detection, within samples, across the non-overlapping intervals, of change-points in the normalised read counts. The detection of these change-points results in segments of the genome which are groups of contiguous intervals having approximately the same normalised read count. Segments with increased or decreased normalised read counts are considered as potential copy number variable regions. After the detection of segments with similar normalised read counts within individuals, the mean normalised read count for these segments is then compared across samples, in order to detect segments that have variable mean normalised read counts. This is done as follows, after all segments have been detected within samples, it is highly unlikely the same segments will have been detected in all samples, therefore the union of all segments across all samples is taken. For each of these segments the Kruskal-Wallis non-parametric test is used to detect the segments where at least one sample has a different mean normalised read count. This segment is then considered to be a potential copy-number variable region among these samples. 2

A5.3. Diagrams detection of CNVs by analysis of NGS data (supplementary to those shown in the main text)

CNV 337

The 4,158bp CNV reported by McCarroll *et al* (2008) was baited to include 2,776bp upstream and 434bp downstream although a 1,056bp long section just upstream from the reported CNV was omitted due to repetitive sequence (figure 5A.1a). It was not possible to determine the true breakpoints of CNV 337 by Sanger sequencing, nor was there any report of an indel in the region. The NGS CNV detection method identified a single CNV which was a deletion (Kruskal-Wallis $p=2.32x10^{-41}$, figure 5A.9b). The NGS analysis indicated that the proximal breakpoint of CNV 337 extended beyond that estimated by McCarroll *et al* (2008) at least as a far as 2,776bp upstream where the baited region ended. The distal breakpoint as determined by NGS analysis was 19bp longer than that estimated by McCarroll *et al* (2008). The mean normalized read count of CNV 337 for each of the nine individuals analysed was concordant with the genotype calls determined by McCarroll *et al* (2008) (Figure 5A.9C). Thus, although there was inaccuracy in the initial position of the CNV by McCarroll *et al* (2008) which lead to the true CNV been incompletely baited, the NGS detection method accurately identified the CNV.

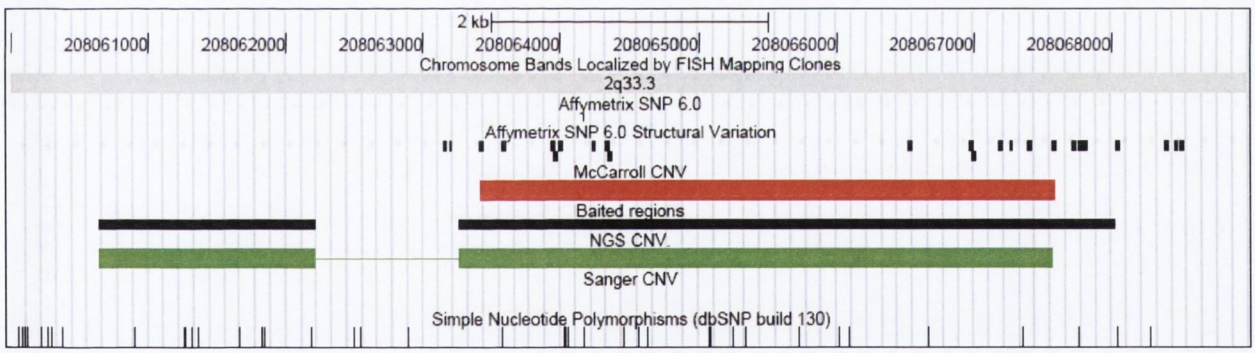


Figure 5A.1a: CNV 337 on the UCSC Genome Browser showing locations determined by McCarroll *et al* (2008) and in-house using NGS. Also show is the region baited and and the position of probes on the genotyping array used by McCarroll *et al* (2008) for CNV detection. It was not possible to determine the exact breakpoints of CNV 337 using Sanger sequencing.

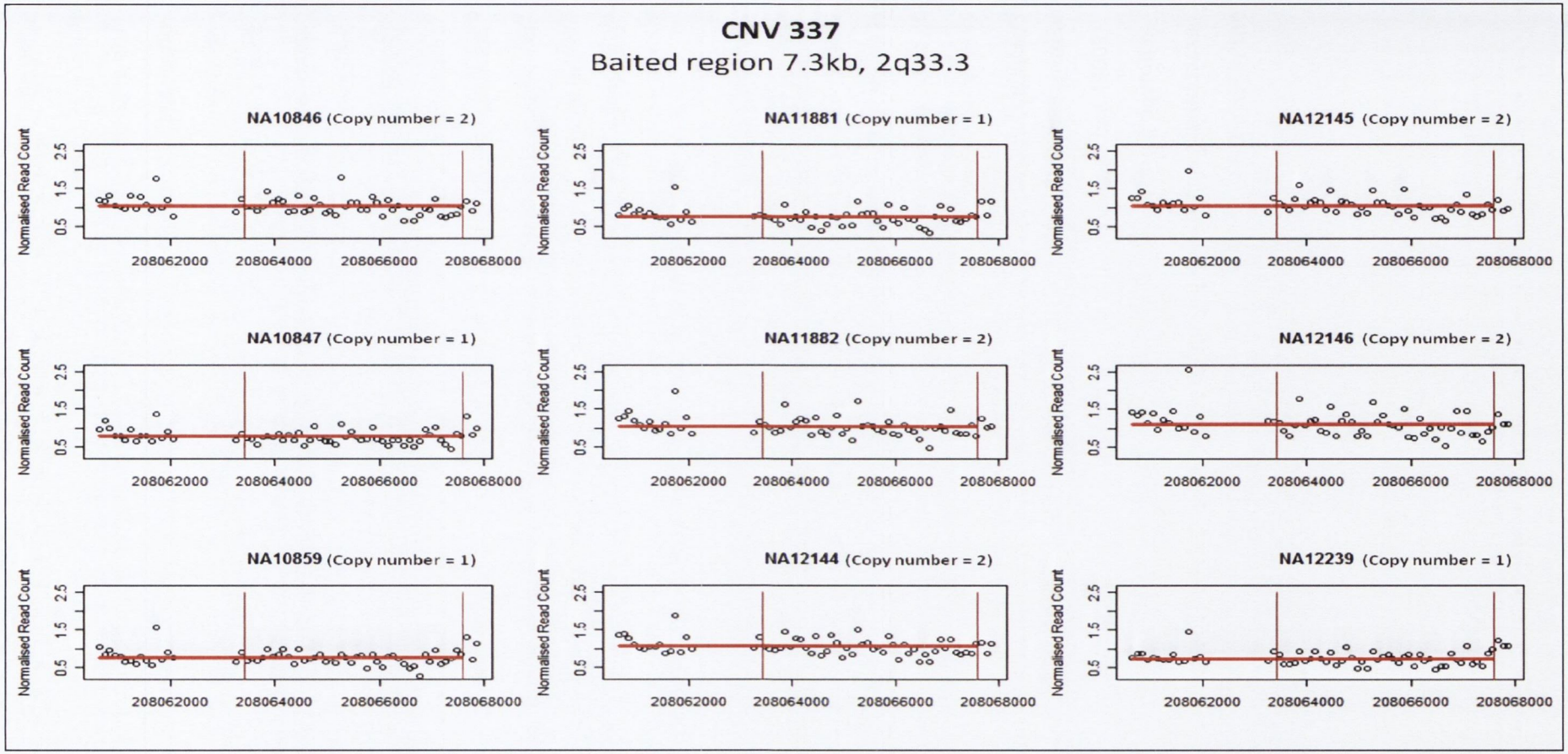


Figure 5A.1b: Detection of CNV 337 from normalised NGS read count data. The x-axis is distance on chromosome 2. The y-axis is the normalised read count. Each data point represents the normalised read count for a 100bp window. Horizontal red bars show the position of the CNV determined by the DNAcopy program using normalised NGS read count data. Vertical pink lines indicate the position of the CNV estimated by McCarroll *et al* (2008) using SNP genotyping. It was not possible to determine the exact breakpoints of CNV 337 using Sanger sequencing.

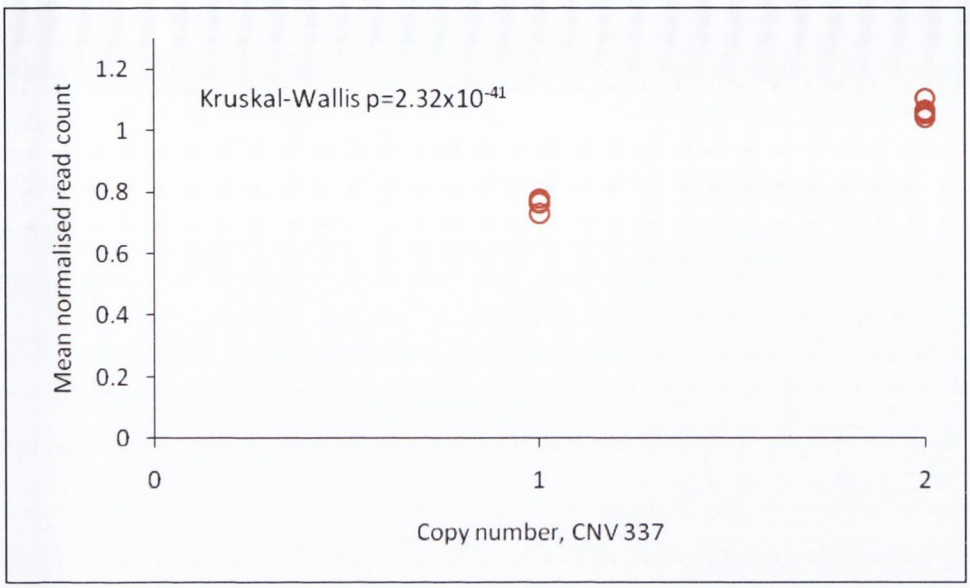


Figure 5A.1c: Scatter-plot showing relationship between copy-number at CNV 337 (determined by McCarroll *et al* (2008)) and mean normalised read count across windows in the CNV determined by NGS analysis.

CNV 451

The 4,936bp CNV reported by McCarroll *et al* (2008) was baited to include 719bp upstream and with the last 282bp omitted due to repetitive sequence (figure 5.9a). By Sanger sequencing of an individual who was homozygous for the deletion (NA10859) the CNV was found to be 645bp longer at the proximal end and 96bp longer at the distal end than reported by McCarroll *et al* (2008). Using read count data generated from NGS, the CNV detection algorithm identified a single CNV which was a deletion (Kruskal-Wallis $p=1.57x10^{-18}$, figure 5.9b). The proximal and distal breakpoints as determined by NGS were 65bp longer and 379bp shorter, respectively than the actual CNV breakpoints (as determined by Sanger sequencing). The mean normalized read count of CNV 451 for each of the nine individuals analysed was concordant with the genotype calls determined by McCarroll *et al* (2008).

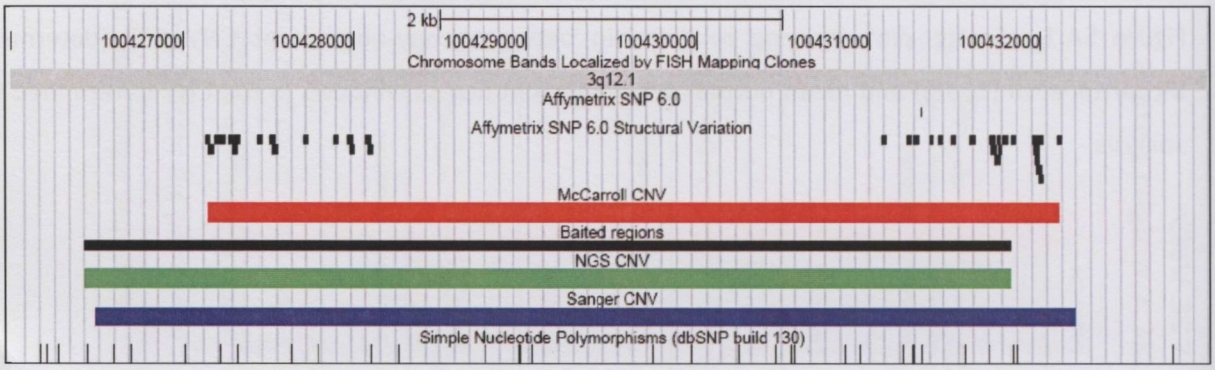


Figure 5A.2a: CNV 451 on the UCSC Genome Browser showing locations determined by McCarroll *et al* (2008) and in-house using NGS and Sanger sequencing. Also show is the region baited and variants in dbSNP (130) including an indel which corresponds to CNV 451 as determined by Sanger sequencing in-house.

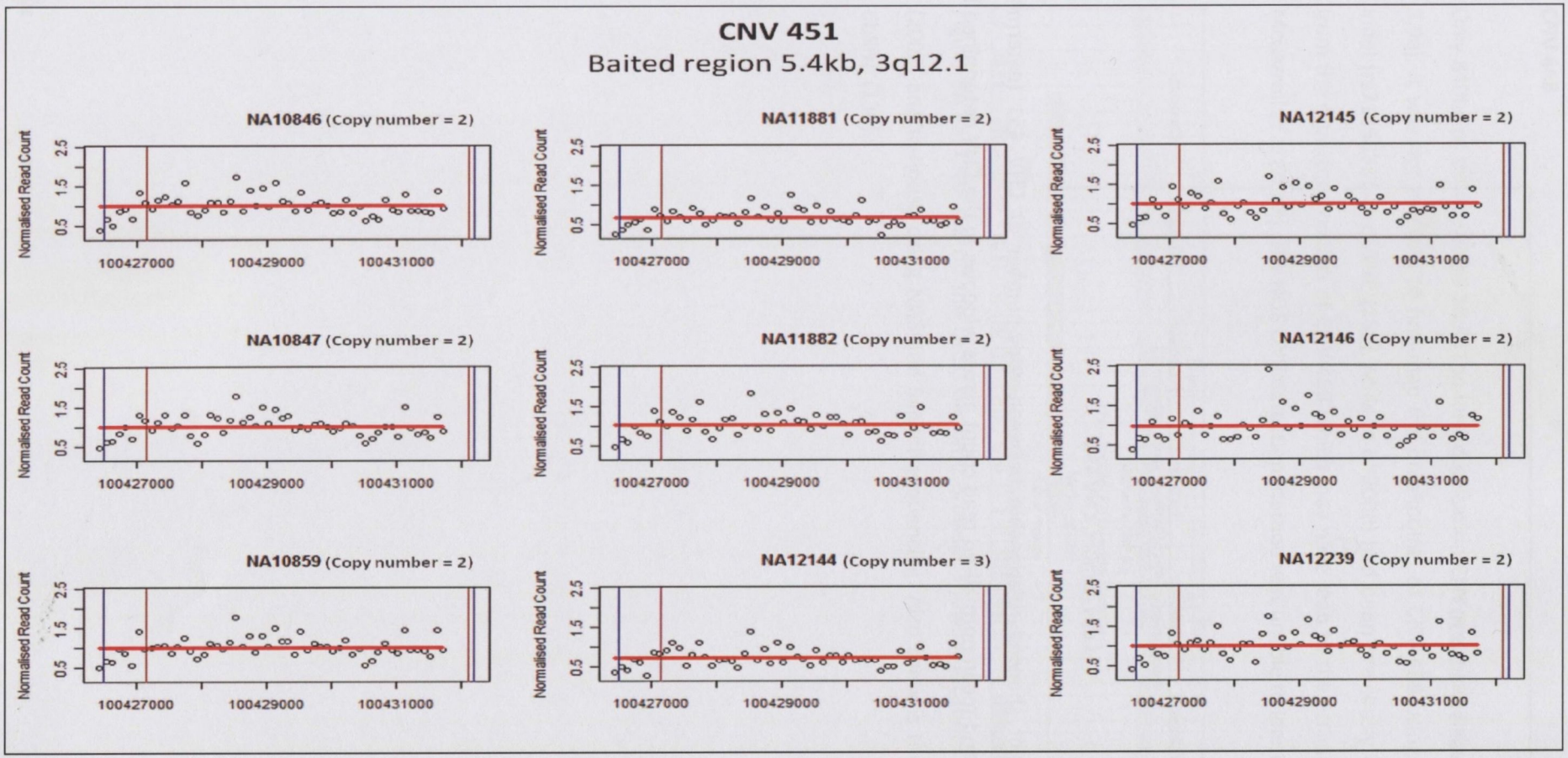


Figure 5A.2b: Detection of CNV 451 from normalised NGS read count data. The x-axis is distance on chromosome 3. The y-axis is the normalised read count. Each data point represents the normalised read count for a 100bp window. Horizontal red bars show the position of the CNV determined by the DNAcopy program using normalised NGS read count data. Vertical pink lines indicate the position of the CNV estimated by McCarroll *et al* (2008) using SNP genotyping. Vertical blue lines indicate the position of the CNV as determined by Sanger sequencing which represents the most accurate of all methods used.

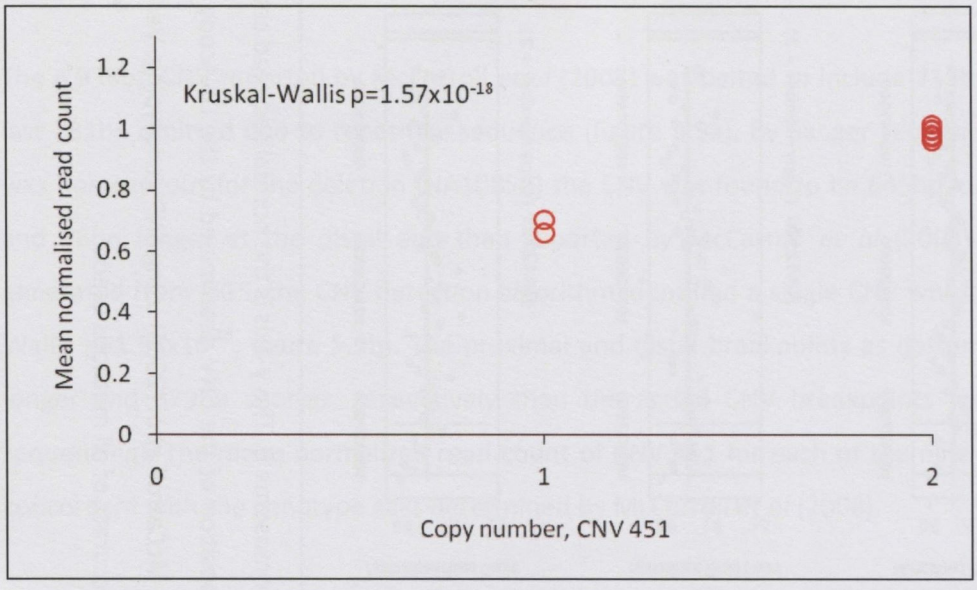


Figure 5A.2c: Scatter-plot showing relationship between copy-number at CNV 451 (determined by McCarroll *et al* (2008)) and mean normalised read count across windows in the CNV determined by NGS analysis.

CNV 468

Only 840bp of the reported 16,841bp long duplication CNV was baited due to repetitive sequence (figure 5.9a). It was not possible to fine-map the breakpoints of CNV 468 using Sanger sequencing. A 9,472bp indel (rs71658331 in dbSNP (130), Mills *et al* 2008) had been previously identified 7,000bp downstream from the reported position of CNV 468 which may have lead to the erroneous positioning of CNV 468 by McCarroll *et al* (2008). The NGS CNV detection method was unable to identify a CNV.

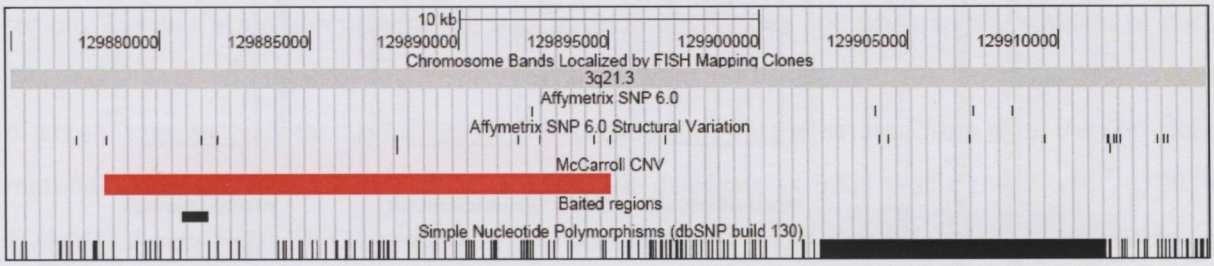


Figure 5A.3a: CNV 468 on the UCSC Genome Browser showing locations determined by McCarroll *et al* (2008) and in-house using NGS and Sanger sequencing. Also show is the region baited and variants in dbSNP (130).

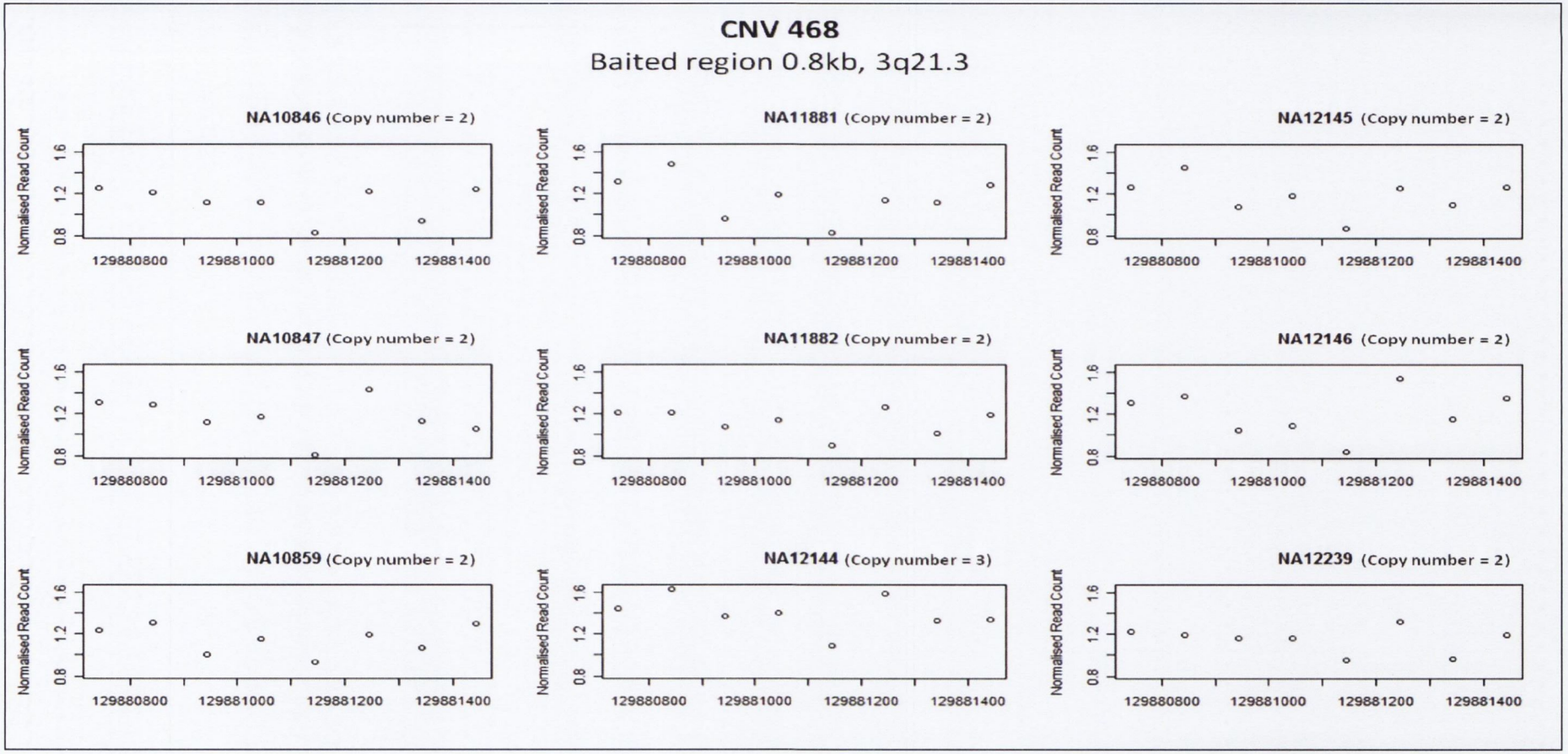


Figure 5A.3b: Screening of CNV 468 using normalised NGS read count data. The x-axis is distance on chromosome 3. The y-axis is the normalised read count. Each data point represents the normalised read count for a 100bp window. The plots show a region within CNV 468 which was baited. The NGS CNV detection method did not find any evidence of a CNV nor was it possible to determine the exact breakpoints using Sanger sequencing.

CNV 682

The 1,219bp deletion CNV reported by McCarroll *et al* (2008) was omitted erroneously from the 7,511bp baited region due to repetitive sequnece. It was not possible to identify the true CNV breakpoints using Sanger sequencing. However, a large indel had been previously identified (rs72258653, Mills *et al* (2003)) which overlapped significantly both with the CNV reported by McCarroll *et al* (2008) and with the baited region. This was likely to represent the true position of CNV 682 and was partially covered by the baits. The breakpoints estimated by McCarroll *et al* (2008) were 194bp and 7,217bp shorter at the proximal and distal ends respectively than rs72258653. Using read count data generated from NGS, the CNV detection method identified three CNV events which were deletions (Kruskal-Wallis $p=7.27x10^{-7}$, 0.0026 and $1.26x10^{-9}$, figure 5.9b). The normalised read count for all three of these deletions corresponded to the genotypes of CNV 682 as determined by McCarroll *et al* (2008). There was an 800bp gap in between the first and second CNVs identified by NGS which appear to have a very low GC-content from visualisation of the UCSC Genome Browser which was probably affecting the accuracy of the detection method. The proximal breakpoint of the CNV could not be estimated by NGS as the baits did not extend far enough although the distal breakpoint was estimated to be 201bp shorter than the previously reported indel.

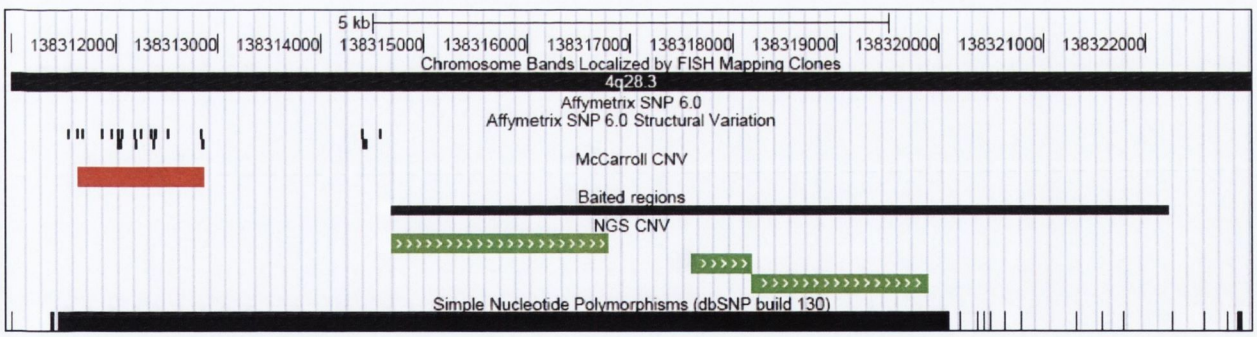


Figure 5A.4a: CNV 682 on the UCSC Genome Browser showing locations determined by McCarroll *et al* (2008) and in-house using NGS and Sanger sequencing. Also show is the region baited and variants in dbSNP (130) including an indel which likely corresponds to the true breakpoints of CNV 682.

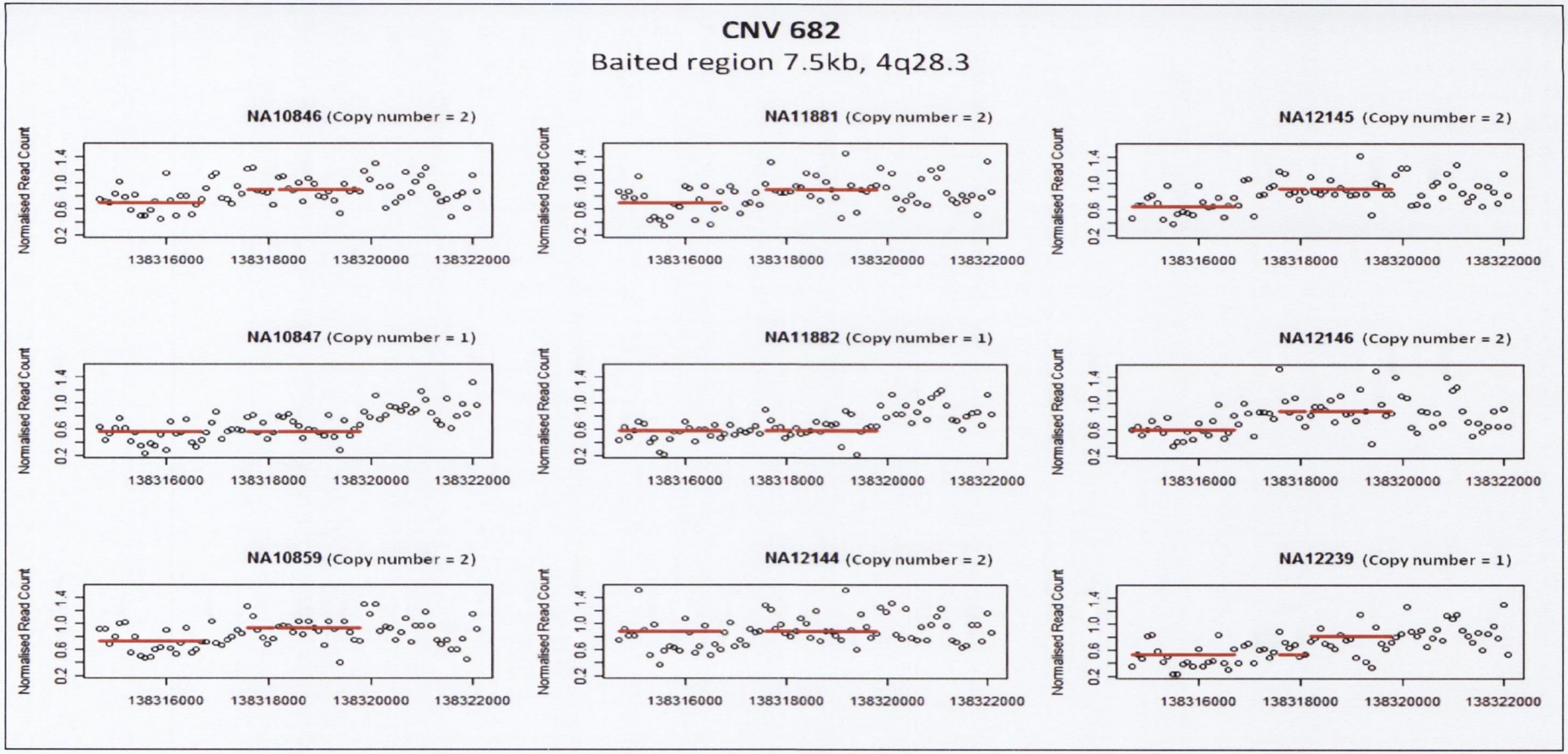


Figure 5A.4b: Detection of CNV 682 from normalised NGS read count data. The x-axis is distance on chromosome 4. The y-axis is the normalised read count. Each data point represents the normalised read count for a 100bp window. Horizontal red bars show the position of the CNV determined by the DNAcopy program using normalised NGS read count data.

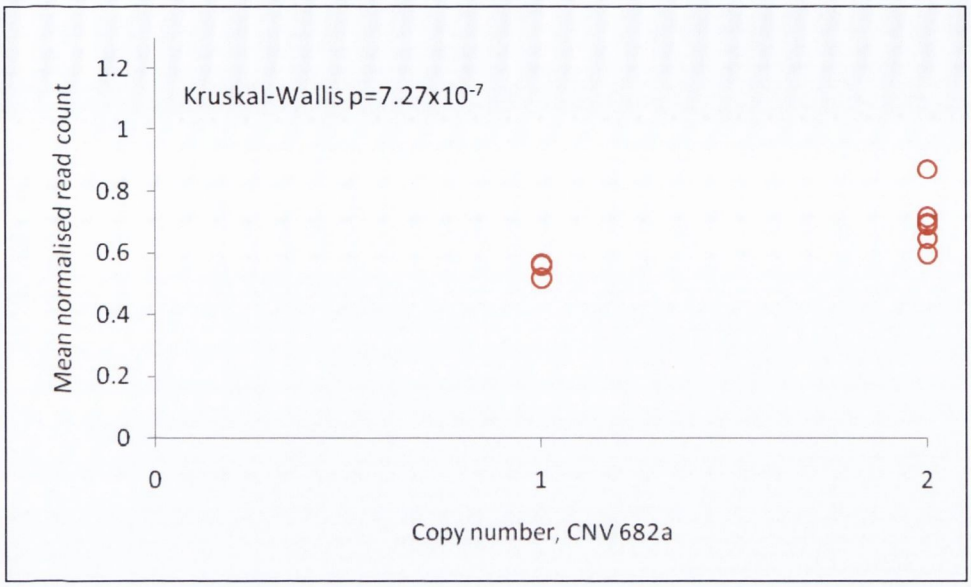


Figure 5A.4c: Scatter-plot showing relationship between copy-number at CNV 682 (determined by McCarroll *et al* (2008)) and mean normalised read count across windows in the CNV determined by NGS analysis (first CNV event detected).

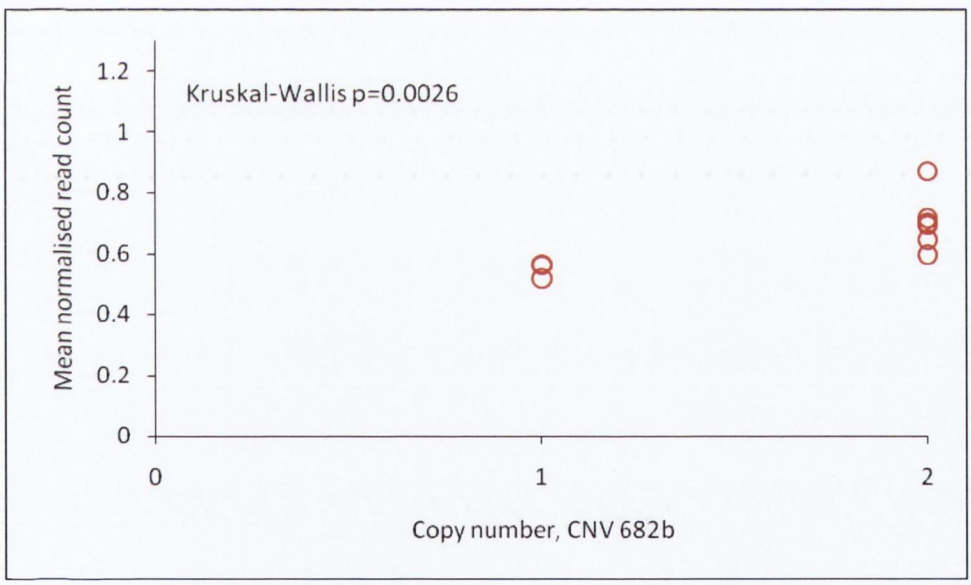


Figure 5A.4d: Scatter-plot showing relationship between copy-number at CNV 682 (determined by McCarroll *et al* (2008)) and mean normalised read count across windows in the CNV determined by NGS analysis (second CNV event detected).

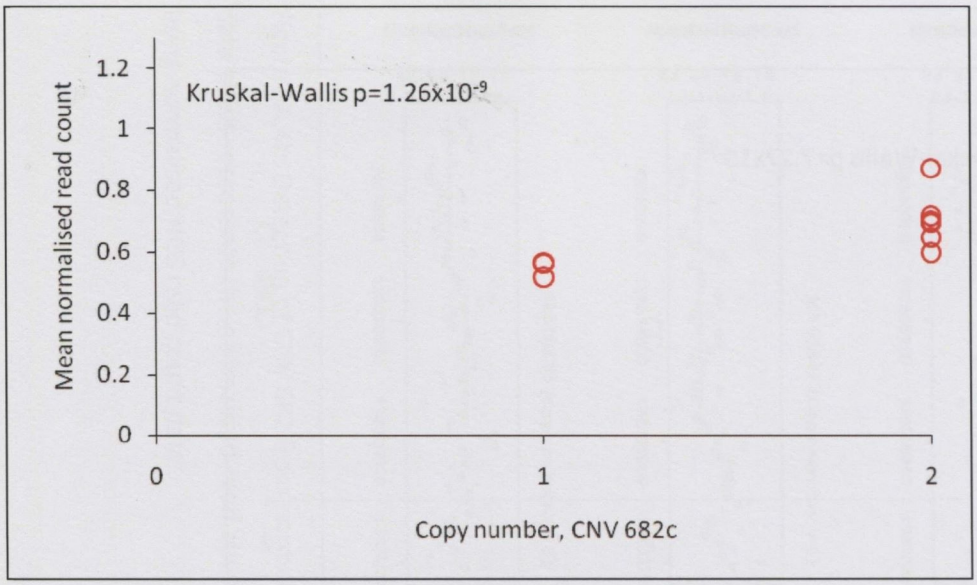


Figure 5A.4e: Scatter-plot showing relationship between copy-number at CNV 682 (determined by McCarroll *et al* (2008)) and mean normalised read count across windows in the CNV determined by NGS analysis (third CNV event detected).

CNV 724

The 3,049bp deletion CNV reported by McCarroll *et al* (2008) was partially baited due to repetitive sequence. The distal 1,499bp was covered with an additional 3,622bp downstream (figure 5.9a). By Sanger sequencing of an individual who was homozygous for the deletion (NA10859) the CNV was found to be 473bp longer at the proximal end and 1,492bp longer at the distal end than reported by McCarroll *et al* (2008). The breakpoints as determined by Sanger sequencing aligned with those of a previously reported indel (rs72281132, Mills *et al* 2003). Using read count data generated from NGS, the CNV detection algorithm identified a single CNV which was a deletion (Kruskal-Wallis $p=1.07x10^{-47}$, figure 5.9b). The proximal breakpoint of CNV 724 could not be determined by NGS analysis because baits did not extend far enough although. NGS analysis did however indicate that the CNV extended at least 2,209bp proximally from the position determined by McCarroll *et al* (2008). The distal breakpoint as determined by NGS analysis was 163bp longer than the actual CNV (as determined by Sanger sequencing) which was probably due to the very low GC-content in the final window (0.24) affecting the accuracy of the detection method. The mean normalized read count of CNV 724 for each of the nine individuals analysed was concordant with the genotype calls determined by McCarroll *et al* (2008).

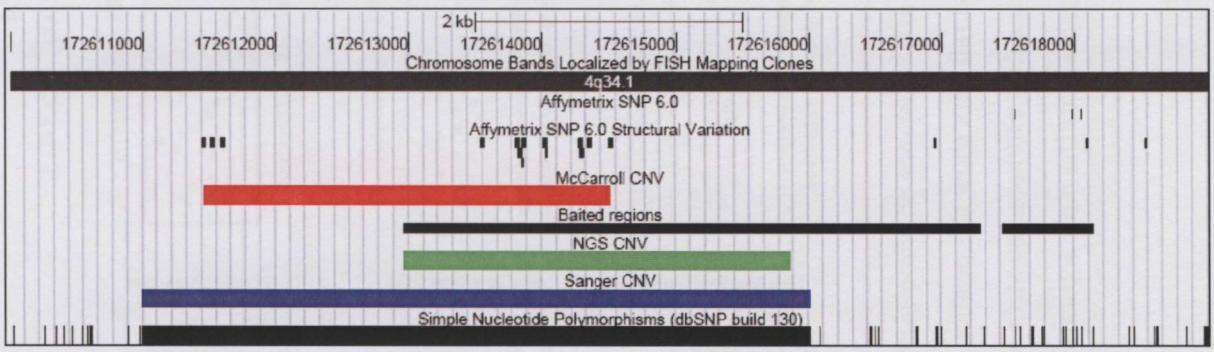


Figure 5A.5a: CNV 724 on the UCSC Genome Browser showing locations determined by McCarroll *et al* (2008) and in-house using NGS and Sanger sequencing. Also show is the region baited and variants in dbSNP (130) including an indel which corresponds to CNV 435 as determined by Sanger sequencing in-house.

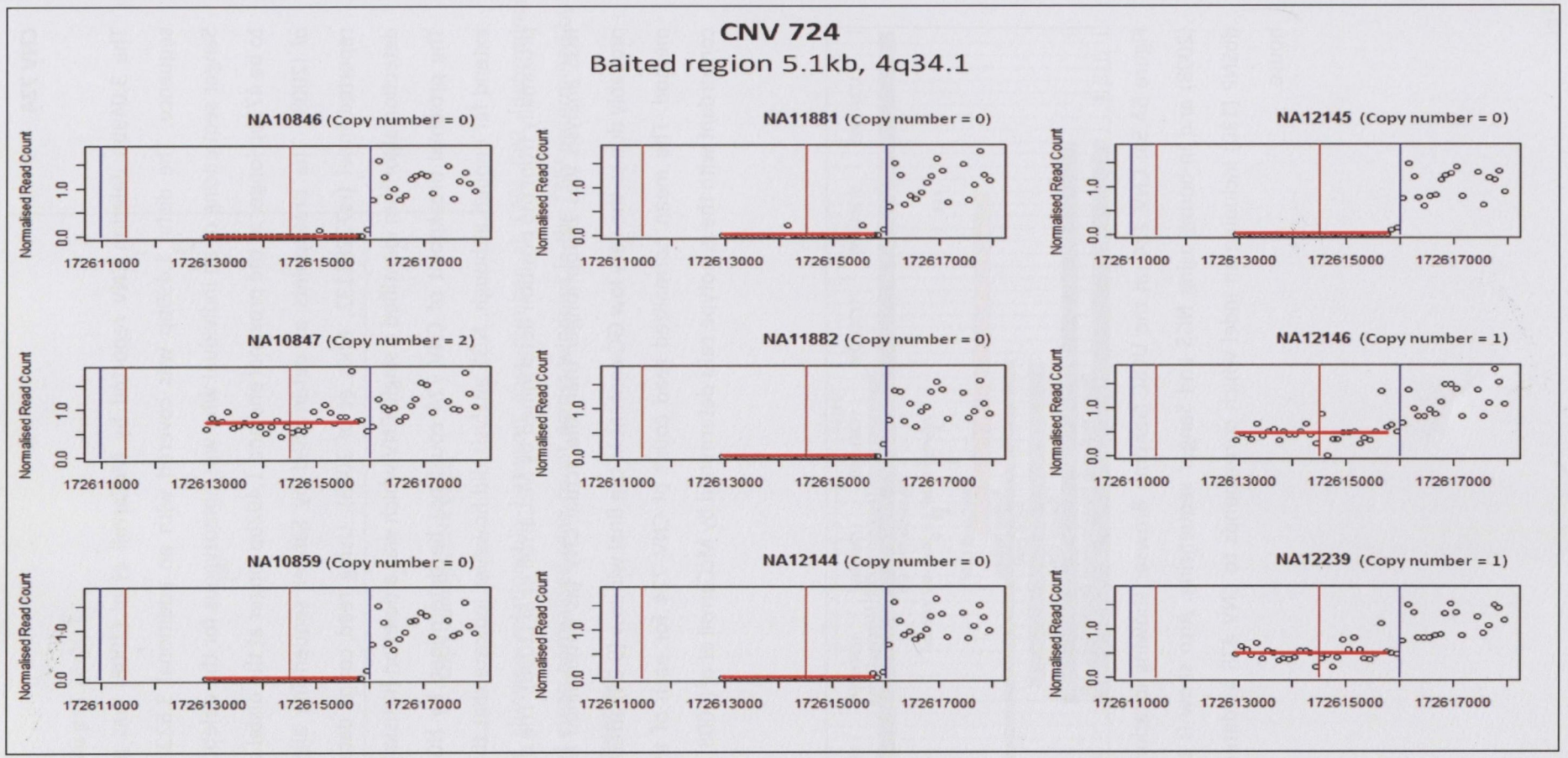


Figure 5A.5b: Detection of CNV 724 from normalised NGS read count data. The x-axis is distance on chromosome 4. The y-axis is the normalised read count. Each data point represents the normalised read count for a 100bp window. Horizontal red bars show the position of the CNV determined by the DNAcopy program using normalised NGS read count data. Vertical pink lines indicate the position of the CNV estimated by McCarroll *et al* (2008) using SNP genotyping. Vertical blue lines indicate the position of the CNV as determined by Sanger sequencing which represents the most accurate of all methods used.

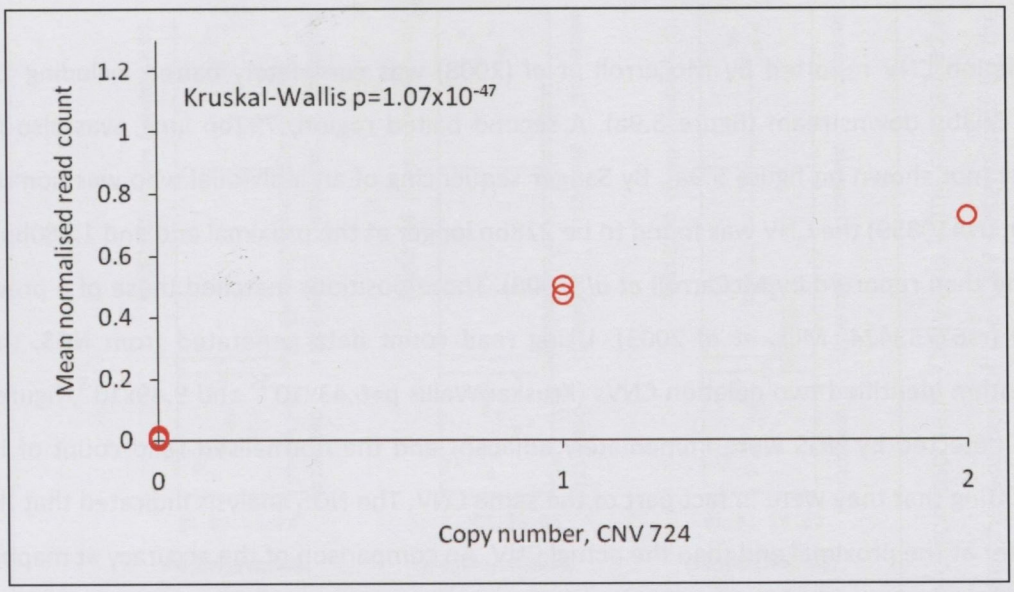


Figure 5A.5c: Scatter-plot showing relationship between copy-number at CNV 724 (determined by McCarroll *et al* (2008)) and mean normalised read count across windows in the CNV determined by NGS analysis.

CNV 1343

The 981bp deletion CNV reported by McCarroll *et al* (2008) was completely baited including 1,788bp upstream and 303bp downstream (figure 5.9a). A second baited region, 792bp long, was also located 3,095bp distally (not shown on figure 5.9a). By Sanger sequencing of an individual who was homozygous for the deletion (NA10859) the CNV was found to be 228bp longer at the proximal end and 1,880bp longer at the distal end than reported by McCarroll *et al* (2008). These positions matched those of a previously-reported indel (rs67753424, Mills *et al* 2003). Using read count data generated from NGS, the CNV detection algorithm identified two deletion CNVs (Kruskal-Wallis $p=6.43x10^{-15}$ and $9.49x10^{-5}$, figure 5.9b). The two CNVs detected by NGS were immediately adjacent and the normalised read count of in each individual indicating that they were in fact part of the same CNV. The NGS analysis indicated that the CNV was 40bp shorter at the proximal end than the actual CNV. An comparison of the accuracy at mapping the distal breakpoint could not be made as the baits did not extend this far. The mean normalized read count of CNV 1343 for each of the nine individuals analysed was concordant with the genotype calls determined by McCarroll *et al* (2008).

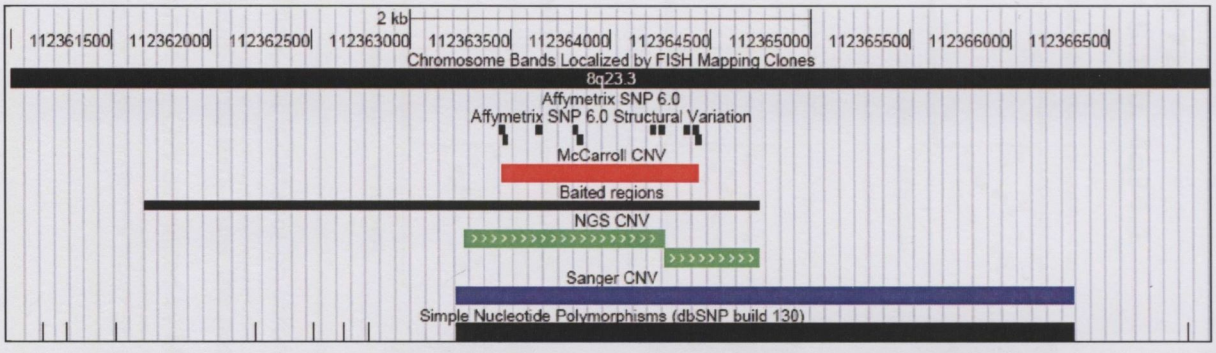


Figure 5A.6a: CNV 1343 on the UCSC Genome Browser showing locations determined by McCarroll *et al* (2008) and in-house using NGS and Sanger sequencing. Also show is the region baited and variants in dbSNP (130) including an indel which corresponds to CNV 1343 as determined by Sanger sequencing in-house.

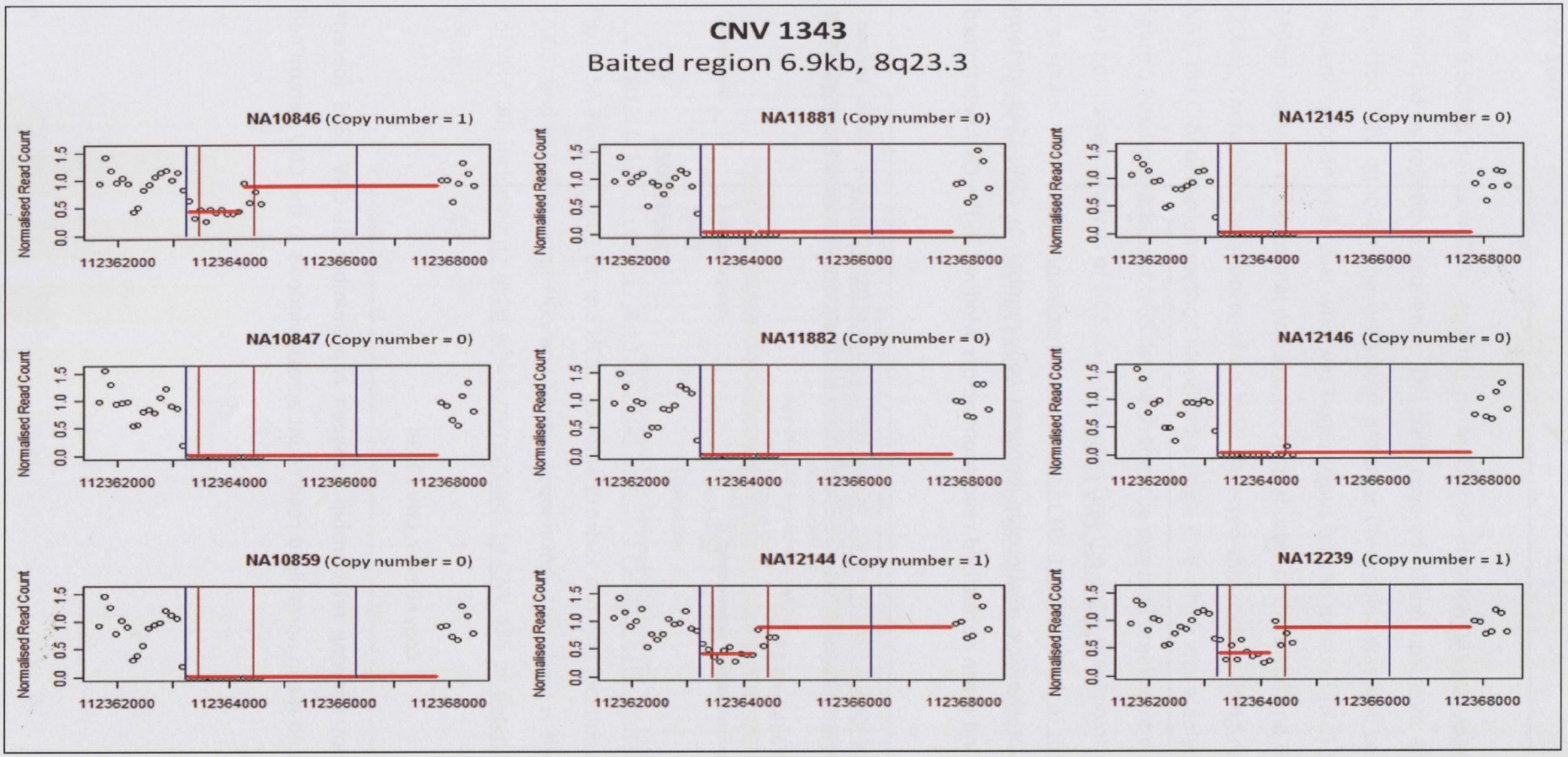


Figure 5A.6b: Detection of CNV 1343 from normalised NGS read count data. The x-axis is distance on chromosome 8. The y-axis is the normalised read count. Each data point represents the normalised read count for a 100bp window. Horizontal red bars show the position of the CNV determined by the DNAcopy program using normalised NGS read count data. Vertical pink lines indicate the position of the CNV estimated by McCarroll *et al* (2008) using SNP genotyping. Vertical blue lines indicate the position of the CNV as determined by Sanger sequencing which represents the most accurate of all methods used.

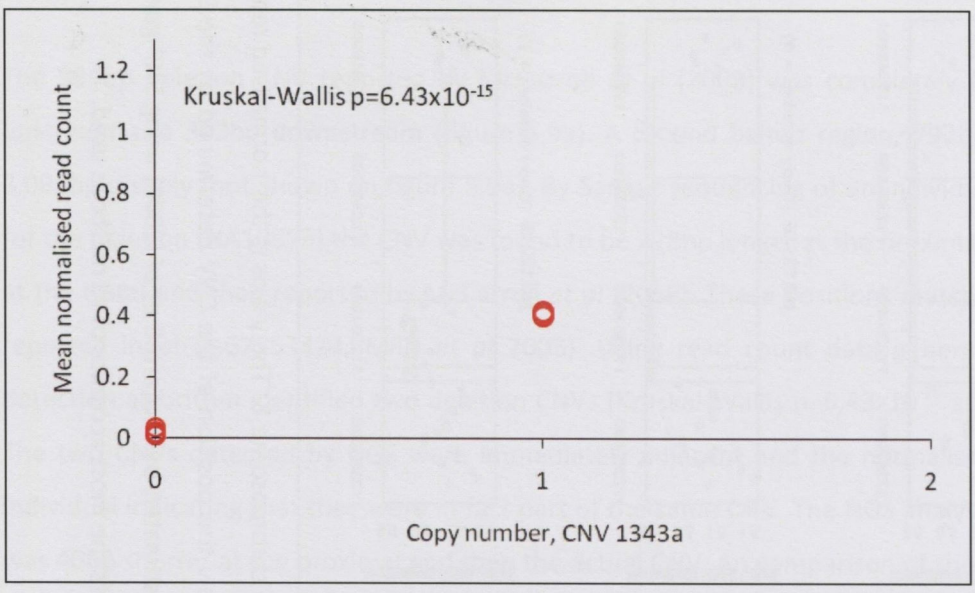


Figure 5A.6c: Scatter-plot showing relationship between copy-number at CNV 1343 (determined by McCarroll *et al* (2008)) and mean normalised read count across windows in the CNV determined by NGS analysis (first section).

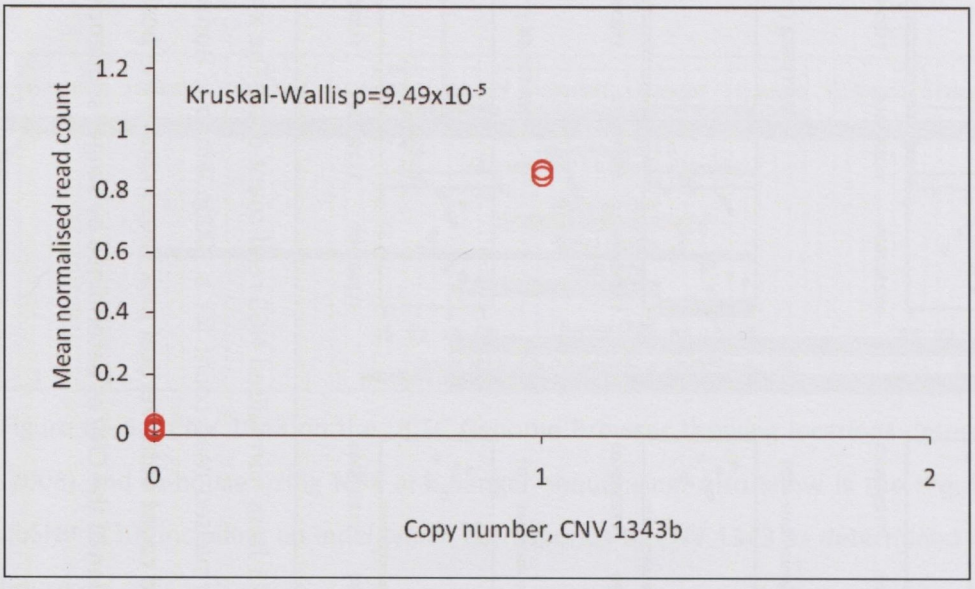


Figure 5A.6d: Scatter-plot showing relationship between copy-number at CNV 1343 (determined by McCarroll *et al* (2008)) and mean normalised read count across windows in the CNV determined by NGS analysis (second section).

CNV 1602

The 9,509bp deletion CNV reported by McCarroll *et al* (2008) was targetted with six groups of baits separated by repetitive sequence. The baits covered 638bp upstream and 1,617bp downstream of the reported CNV although the estimated proximal breakpoint was not covered (figure 5.9a). By Sanger sequencing of an individual who was homozygous for the deletion (NA12146) the CNV was found to be 310bp longer at the proximal end and 8,238bp shorter at the distal end than reported by McCarroll *et al* (2008). There were no reports of an indel detected this region. Using read count data generated from NGS, the CNV detection method identified a single CNV which was a deletion (Kruskal-Wallis $p=7.99x10^{-7}$, figure 5.9b). Comparison of NGS data with the true proximal breakpoint was not possible because the site was not baited. Analysis of NGS data indicated that the distal breakpoint was 43bp shorter than the true breakpoint. The mean normalized read count of CNV 1602 for each of the nine individuals analysed was concordant with the genotype calls determined by McCarroll *et al* (2008).

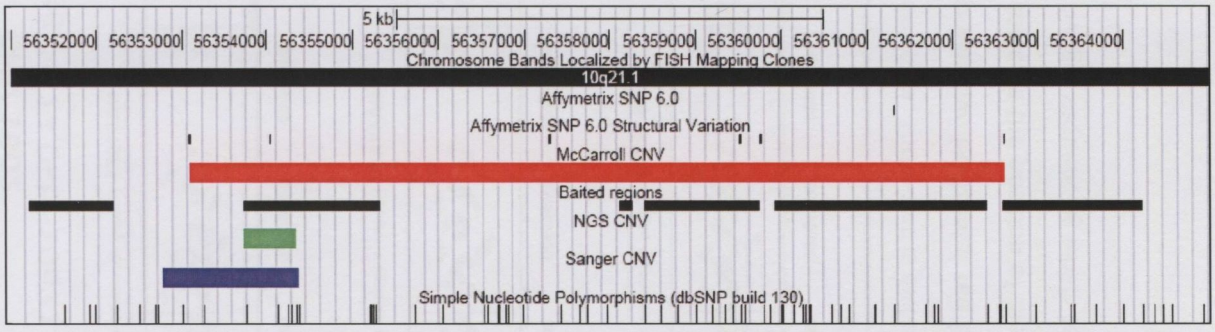


Figure 5A.7a: CNV 1602 on the UCSC Genome Browser showing locations determined by McCarroll *et al* (2008) and in-house using NGS and Sanger sequencing. Also show is the region baited and variants in dbSNP (130) including an indel which corresponds to CNV 435 as determined by Sanger sequencing in-house.

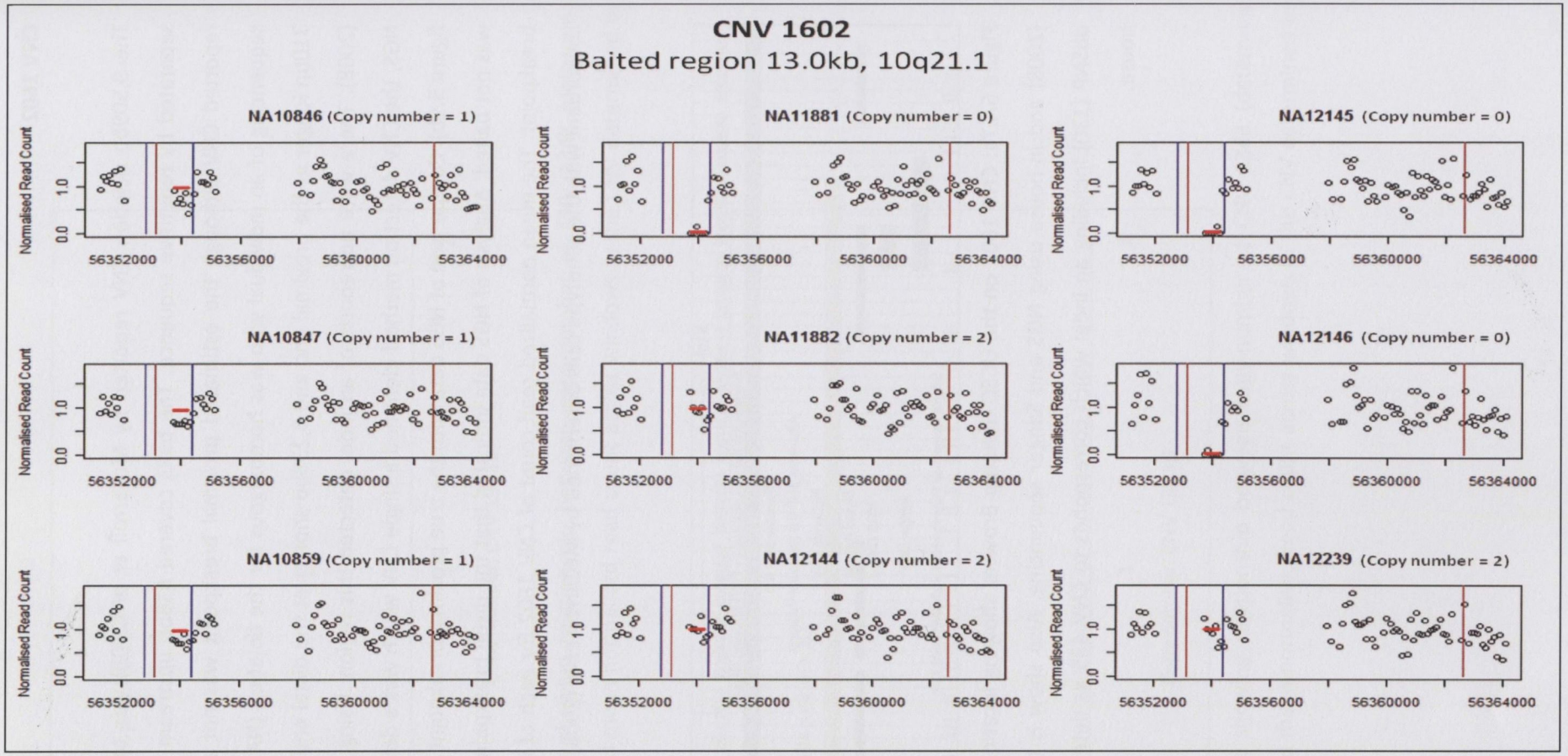


Figure 5A.7b: Detection of CNV 1602 from normalised NGS read count data. The x-axis is distance on chromosome 10. The y-axis is the normalised read count. Each data point represents the normalised read count for a 100bp window. Horizontal red bars show the position of the CNV determined by the DNAcopy program using normalised NGS read count data. Vertical pink lines indicate the position of the CNV estimated by McCarroll *et al* (2008) using SNP genotyping. Vertical blue lines indicate the position of the CNV as determined by Sanger sequencing which represents the most accurate of all methods used.

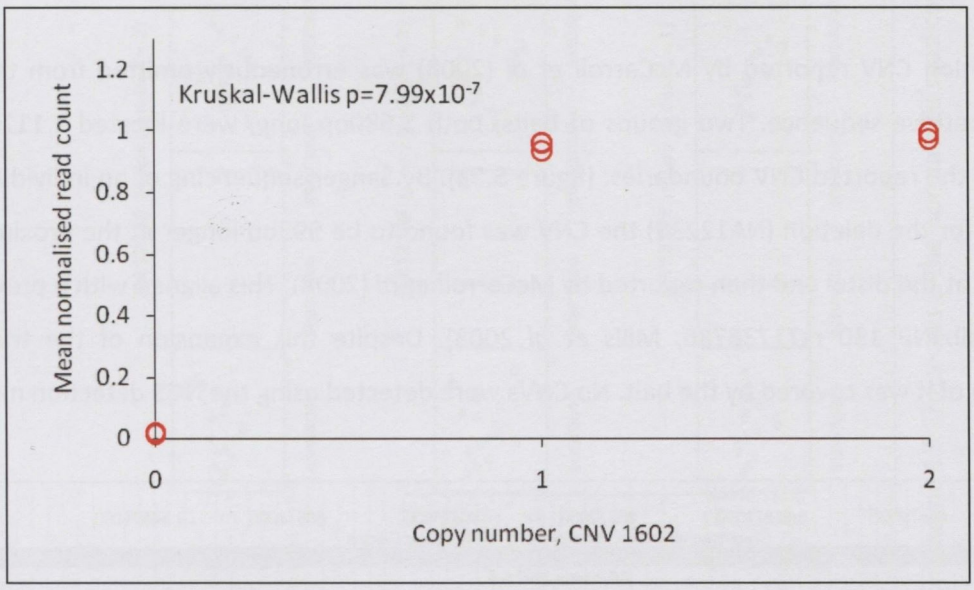


Figure 5A.7c: Scatter-plot showing relationship between copy-number at CNV 1602 (determined by McCarroll *et al* (2008)) and mean normalised read count across windows in the CNV determined by NGS analysis.

CNV 2215

The 1,728bp deletion CNV reported by McCarroll *et al* (2008) was erroneously omitted from targeted baited due to repetitive sequence. Two groups of baits, both 1,680bp long, were located 1,112bp and 673bp away from the reported CNV boundaries. (figure 5.9a). By Sanger sequencing of an individual who was homozygous for the deletion (NA12239) the CNV was found to be 993bp longer at the proximal end and 277bp longer at the distal end than reported by McCarroll *et al* (2008). This aligned with a previously-identified indel (dbSNP 130 rs71736786, Mills *et al* 2003). Despite this expansion of the true CNV breakpoints, none of it was covered by the bait. No CNVs were detected using the NGS detection method.

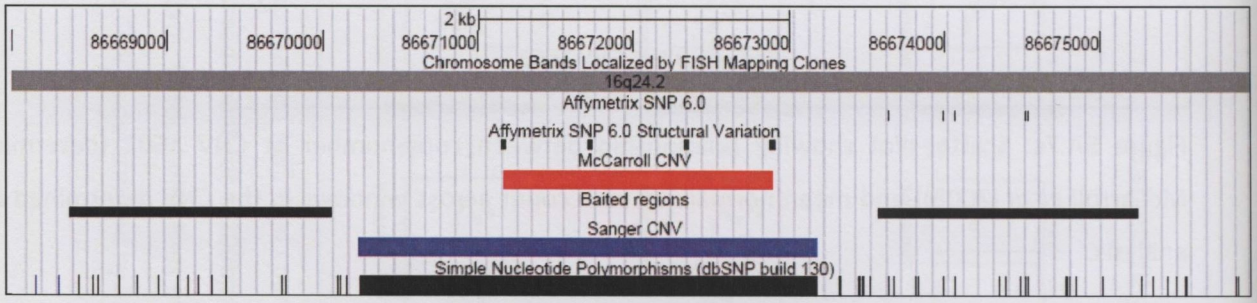


Figure 5A.8a: CNV 2215 on the UCSC Genome Browser showing locations determined by McCarroll *et al* (2008) and in-house using NGS and Sanger sequencing. Also show is the region baited and variants in dbSNP (130) including an indel which corresponds to CNV 2215 as determined by Sanger sequencing in-house.

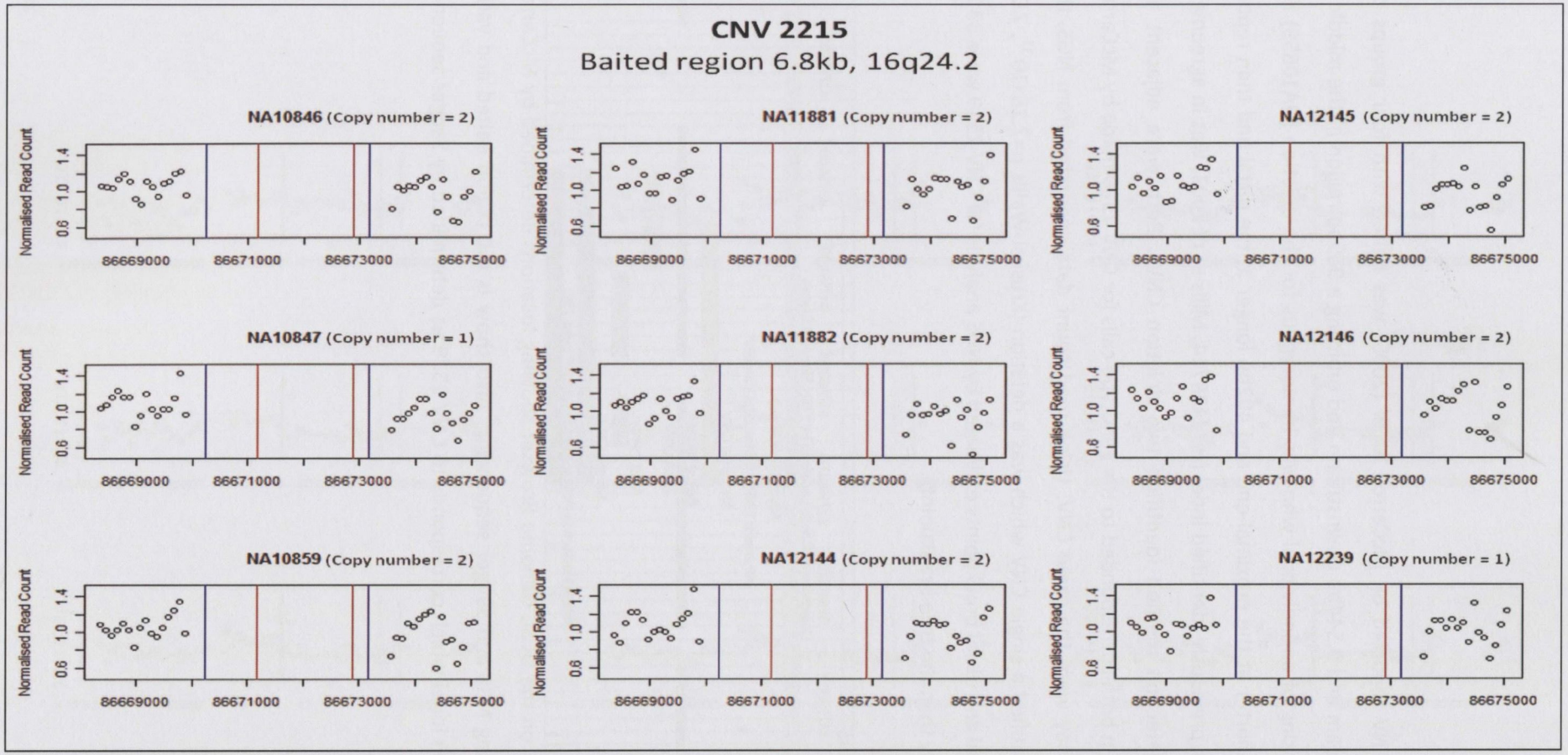


Figure 5A.8b: Screen for CNV 2215 from normalised NGS read count data. The x-axis is distance on chromosome 16. The y-axis is the normalised read count. Each data point represents the normalised read count for a 100bp window. No CNV was detected using the NGS detection method for CNV 2215. Vertical pink lines indicate the position of the CNV estimated by McCarroll *et al* (2008) using SNP genotyping. Vertical blue lines indicate the position of the CNV as determined by Sanger sequencing which represents the most accurate of all methods used.

CNV 2529

The 2,475bp deletion CNV reported by McCarroll *et al* (2008) was baited with four groups of baits including 4,555bp upstream and 2,544bp downstream and omitting a 959bp region in the middle (figure 5.9a). By Sanger sequencing of an individual who was homozygous for the deletion (NA10859) the CNV was found to be 30bp shorter at the proximal end and 492bp longer at the distal end than reported by McCarroll *et al* (2008). A previously-identified indel (rs72438714, Mills *et al* 2003) was in agreement with this result. The NGS detection method identified two deletion CNVs. Both were adjacent and the normalised read counts in both corresponded to the genotype calls for CNV 2529 made by McCarroll *et al* (2008) indicating that they were the same CNV. Using read count data generated from NGS, the CNV detection algorithm identified a single CNV which was a deletion (Kruskal-Wallis $p=7.18x10^{-11}$, $7.05x10^{-9}$, figure 5.9b). The proximal and distal breakpoints estimated by NGS analysis for CNV 2529 were 330bp and 64bp shorter respectively than the true breakpoints.

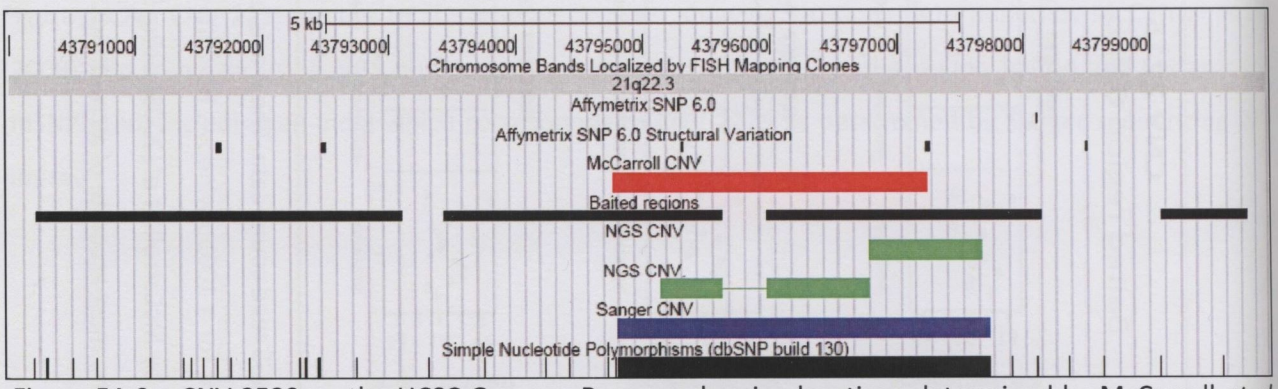


Figure 5A.9a: CNV 2529 on the UCSC Genome Browser showing locations determined by McCarroll *et al* (2008) and in-house using NGS and Sanger sequencing. Also show is the region baited and variants in dbSNP (130) including an indel which corresponds to CNV 2529 as determined by Sanger sequencing in-house.

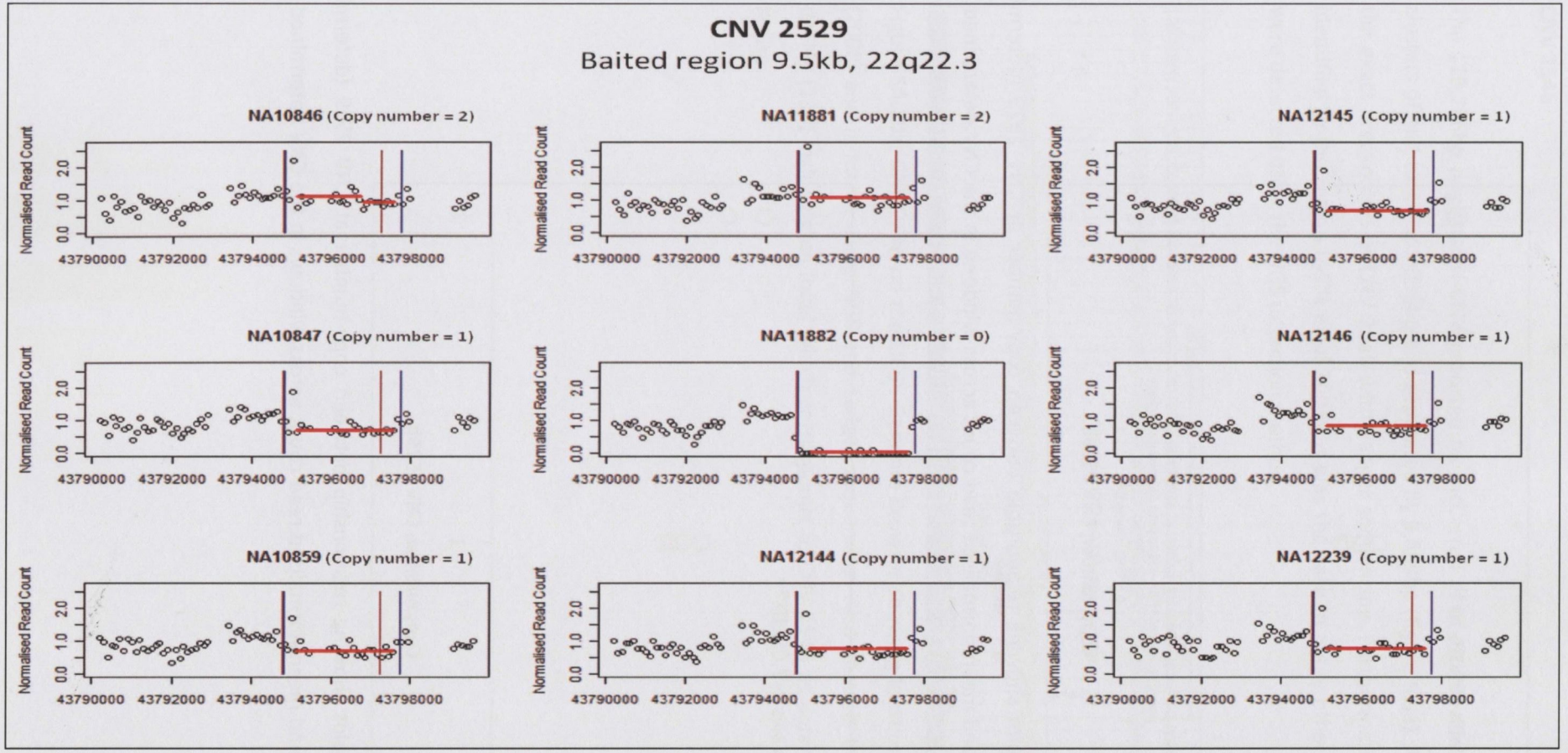


Figure 5A.9b: Detection of CNV 2529 from normalised NGS read count data. The x-axis is distance on chromosome 22. The y-axis is the normalised read count. Each data point represents the normalised read count for a 100bp window. Horizontal red bars show the position of the CNV determined by the DNAcopy program using normalised NGS read count data. Vertical pink lines indicate the position of the CNV estimated by McCarroll *et al* (2008) using SNP genotyping. Vertical blue lines indicate the position of the CNV as determined by Sanger sequencing which represents the most accurate of all methods used.

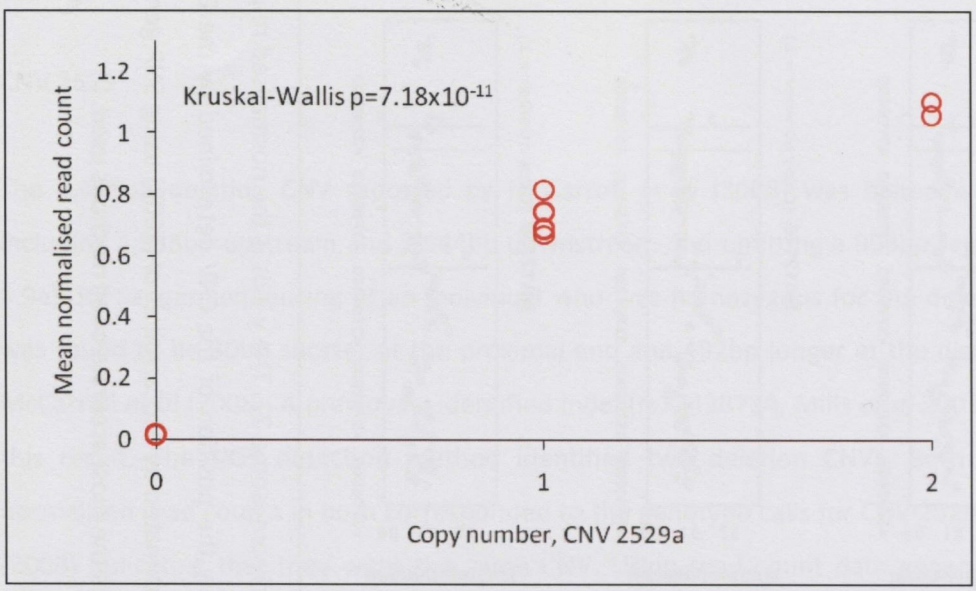


Figure 5A.10c: Scatter-plot showing relationship between copy-number at CNV 2529 (determined by McCarroll *et al* (2008)) and mean normalised read count across windows in the CNV determined by NGS analysis.

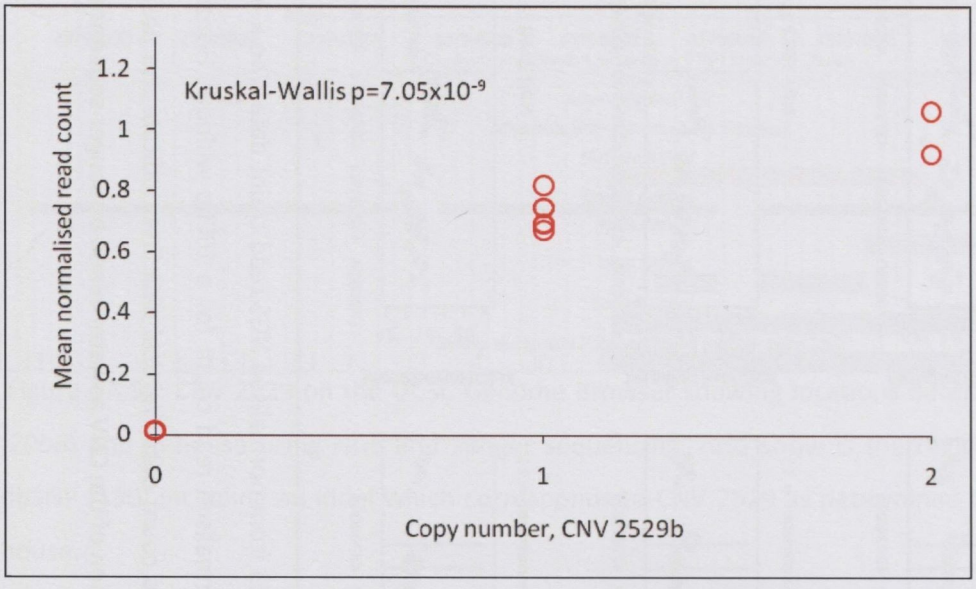
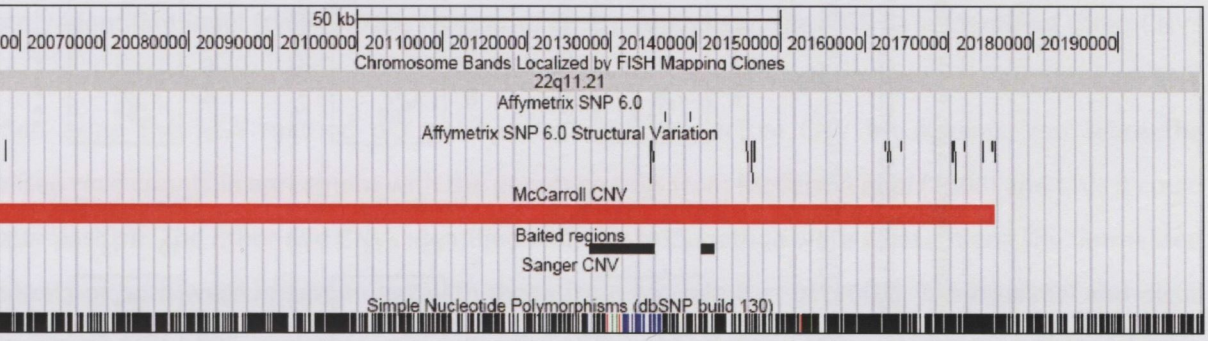


Figure 5A.10d: Scatter-plot showing relationship between copy-number at CNV 2529 (determined by McCarroll *et al* (2008)) and mean normalised read count across windows in the CNV determined by NGS analysis.

19,296bp duplication CNV reported by McCarroll *et al* (2008) was baited in the middle with two rs of bait totalling 9,096bp and separated by 5,689bp (figure 5.9a). It was not possible to determine xact breakpoints of CNV 2546 with Sanger sequencing. Furthermore no indel had been previously fied in the region (Mills *et al* 2003) as was the case in many other of the CNVs studied. No CNVs detected using the NGS detection method.



e 5A.10a: CNV 2546 on the UCSC Genome Browser showing locations determined by McCarroll *et al*) and in-house using NGS and Sanger sequencing. Also show is the region baited and variants in P (130) including an indel which corresponds to CNV 2546 as determined by Sanger sequencing in-e.

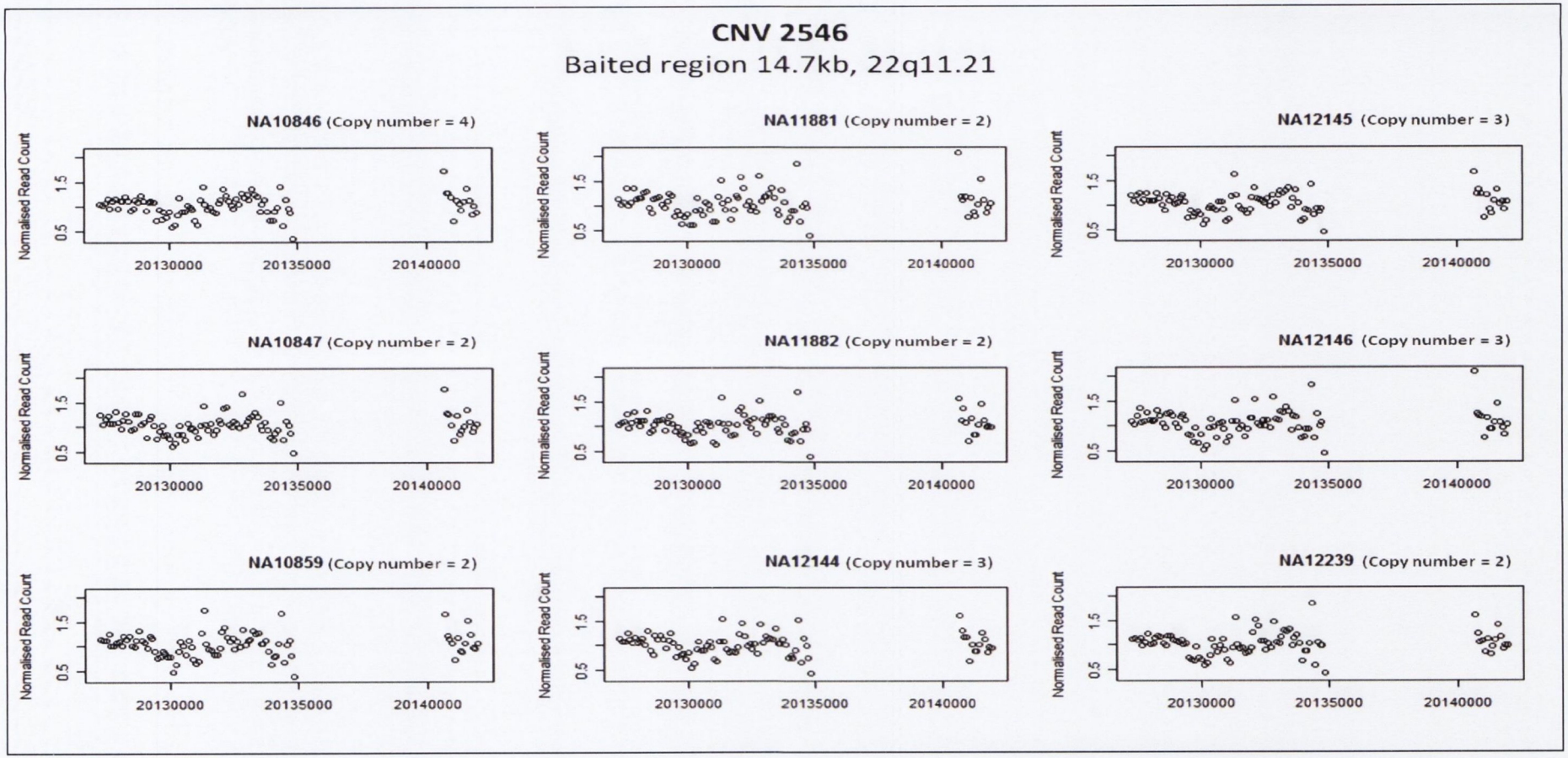


Figure 5A.10b: Screening for CNV 2546 with normalised NGS read count data. The x-axis is distance on chromosome 22. The y-axis is the normalised read count. Each data point represents the normalised read count for a 100bp window. The baited regions were located within the CNV breakpoints as reported by McCarroll *et al* (2008). It was not possible to identify a CNV using the NGS detection method for CNV 2546, nor was it possible to determine the exact breakpoints using Sanger sequencing.

CNV 2575

The 1,649bp deletion CNV reported by McCarroll *et al* (2008) was partially baited by two groups of bait which covered including 2,986bp upstream and 1,473bp downstream and separated by a 1,356bp gap which included the estimated proximal boundary (figure 5.9a). By Sanger sequencing of an individual who was homozygous for the deletion (NA11881) the CNV was found to be 1,148bp longer at the proximal end and 1,784bp longer at the distal end than reported by McCarroll *et al* (2008). This was in agreement with a previously-identified indel (rs66483755, Mills *et al* 2003). Analysis of the NGS data identified three CNVs which were deletions (Kruskal-Wallis $p=2.01x10^{-3}$, $1.37x10^{-15}$ and $4.45x10^{-24}$, figure 5.9b). The first CNV identified using the NGS method did not overlap with the actual CNV breakpoints, nor were the normalised read counts in agreement with the genotype calls from McCarroll *et al* (2008) and thus it could be a false positive. The other two CNVs identified with the NGS method were adjacent and the normalised read counts of both were in agreement with those from McCarroll *et al* (2008). The proximal and distal breakpoints of CNV 2575 as determined using the NGS method were 138bp longer and 404bp shorter than those of the actual CNV.

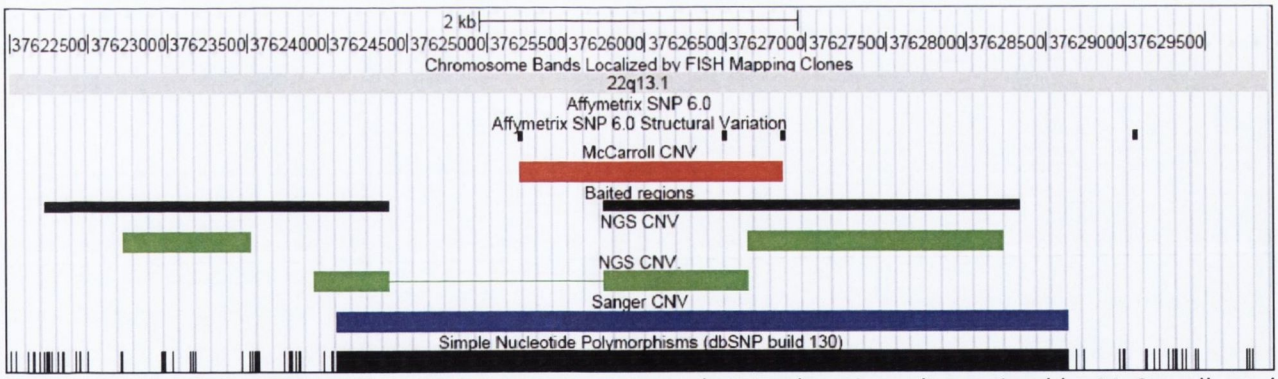


Figure 5A.11a: CNV 2575 on the UCSC Genome Browser showing locations determined by McCarroll *et al* (2008) and in-house using NGS and Sanger sequencing. Also show is the region baited and variants in dbSNP (130) including an indel which corresponds to CNV 435 as determined by Sanger sequencing in-house.

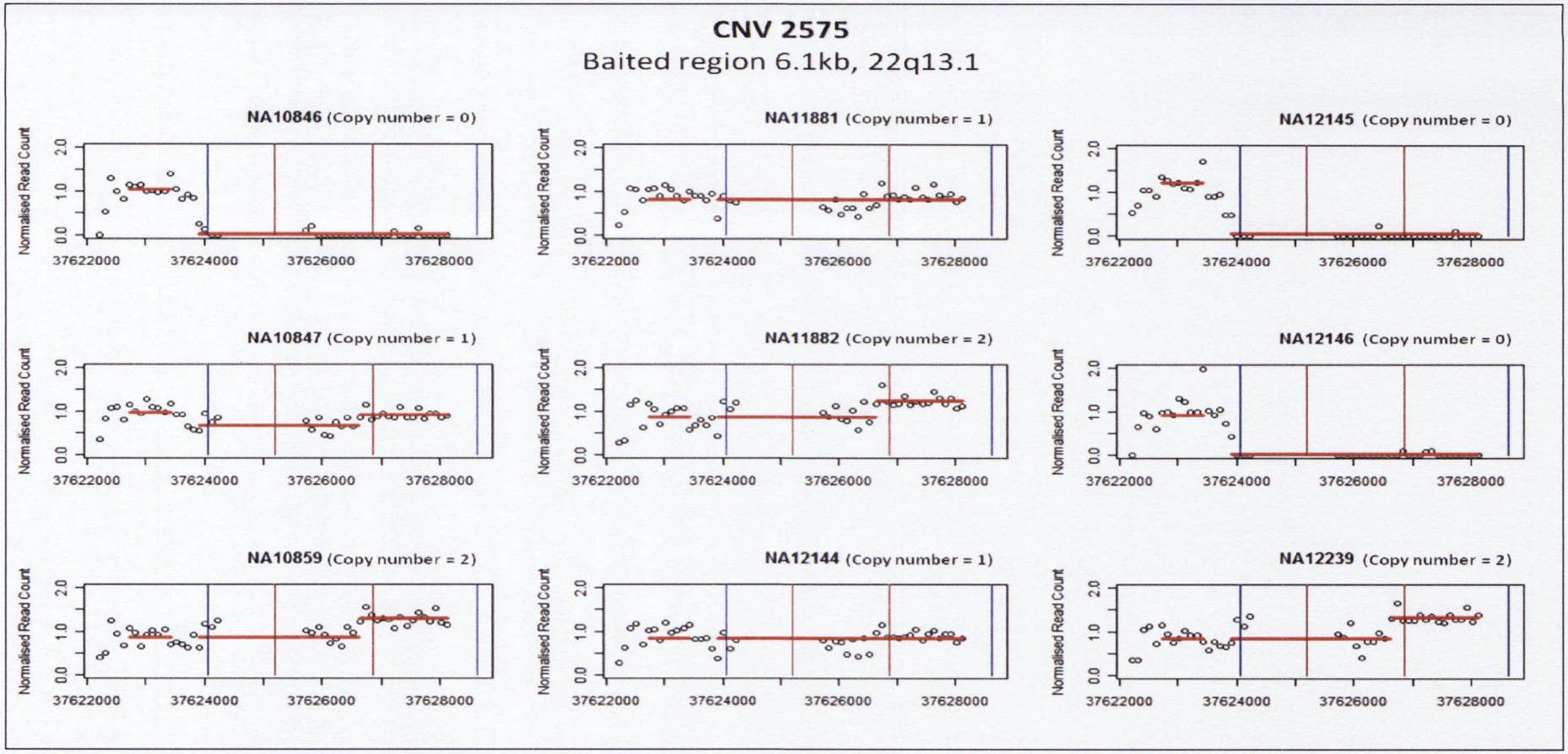


Figure 5A.11b: Detection of CNV 2575 from normalised NGS read count data. The x-axis is distance on chromosome 22. The y-axis is the normalised read count. Each data point represents the normalised read count for a 100bp window. Horizontal red bars show the position of the CNV determined by the DNAcopy program using normalised NGS read count data. Vertical pink lines indicate the position of the CNV estimated by McCarroll *et al* (2008) using SNP genotyping. Vertical blue lines indicate the position of the CNV as determined by Sanger sequencing which represents the most accurate of all methods used.

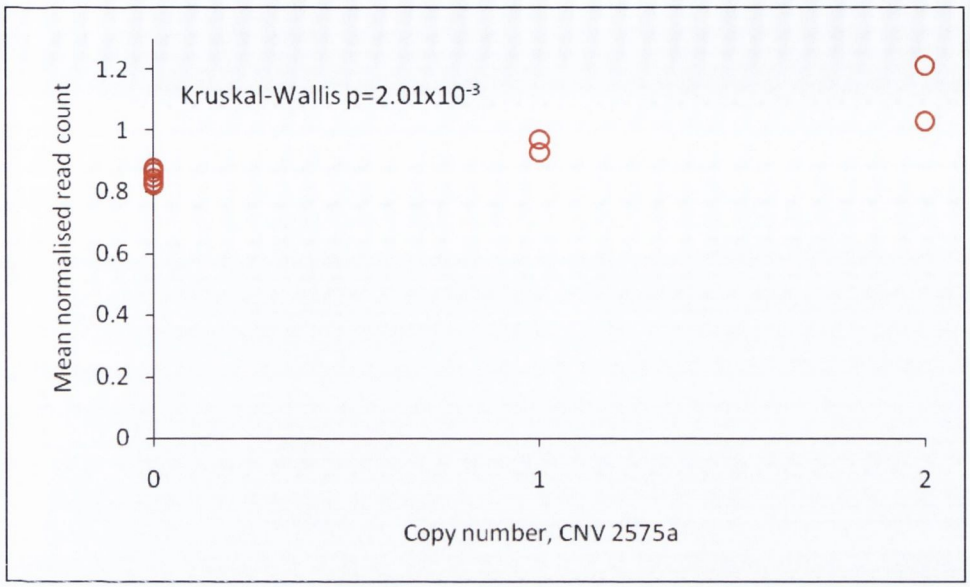


Figure 5A.11c: Scatter-plot showing relationship between copy-number at CNV 2575 (determined by McCarroll *et al* (2008)) and mean normalised read count across windows in the CNV determined by NGS analysis (first section).

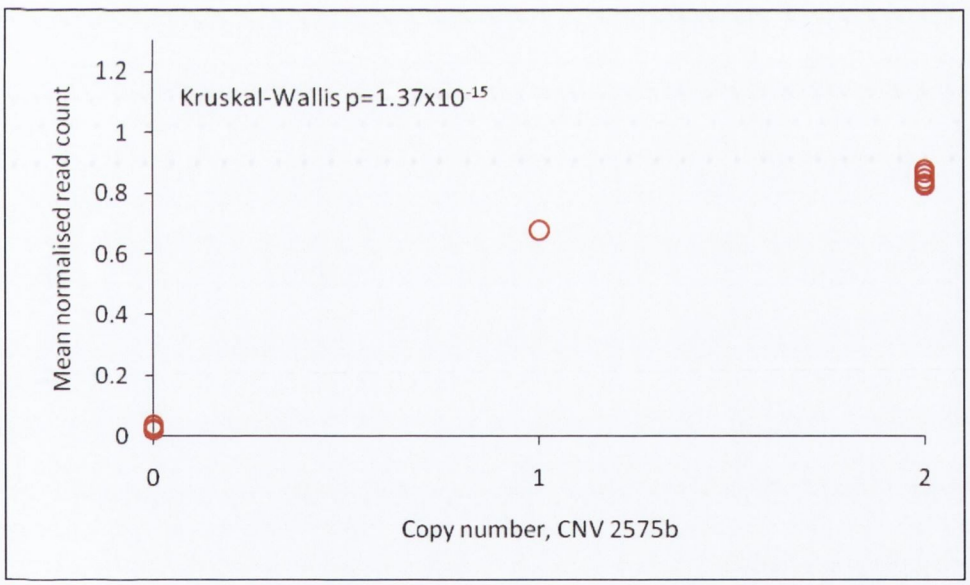


Figure 5A.11d: Scatter-plot showing relationship between copy-number at CNV 2575 (determined by McCarroll *et al* (2008)) and mean normalised read count across windows in the CNV determined by NGS analysis (second section).

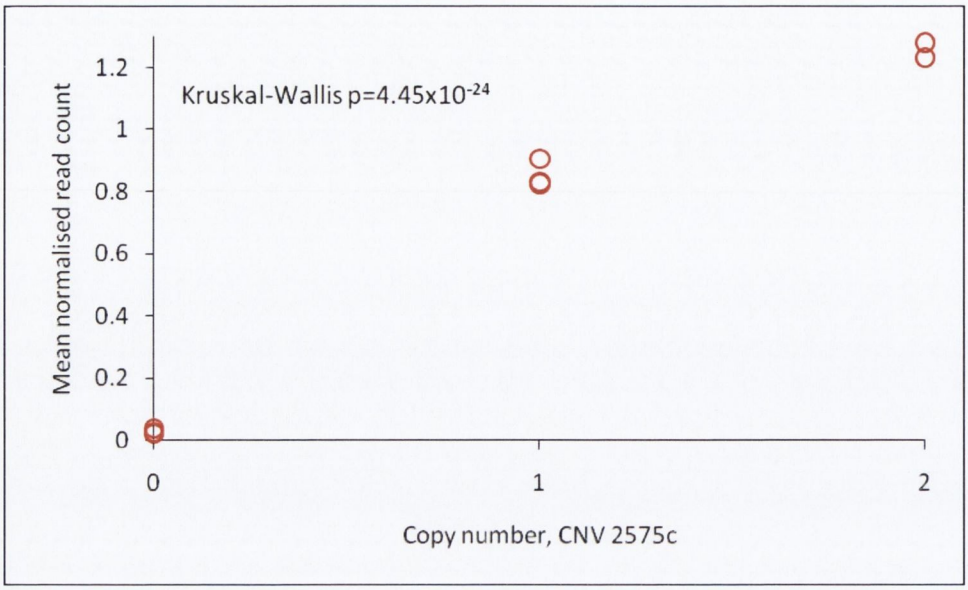


Figure 5A.11d: Scatter-plot showing relationship between copy-number at CNV 2575 (determined by McCarroll *et al* (2008)) and mean normalised read count across windows in the CNV determined by NGS analysis (third section).

Chapter 6: General Discussion

6.1 Recent developments in schizophrenia genetics research

6.1.1 Genome-wide association studies

Since this study was started, there has been a massive change to the methods used for the investigation of schizophrenia genetics. This has included technological advances e.g. dense SNP genotyping arrays, ready availability of user-friendly analytical packages for managing large datasets and greater collaboration within the field generating much larger sample sizes. These changes have brought real progress in the identification of risk variants for the disorder. For example, common risk variation has been confirmed at the *HLA* locus and at the genes *ZNF804A*, *TCF4* and *NRGN* (ISC 2009, O'Donovan *et al* 2008, Stefansson *et al* 2009). As with other common complex disorders the identified effect sizes of the associated variants are small and larger samples will be required to ensure that such studies have the power to generate p-values which withstand correction for multiple testing. For other complex phenotypes such as Type I Diabetes and height, sample sizes of up to 100,000 have allowed identification of effects accounting for 30-40% of genetic variance (Barrett *et al* 2008, 2009; Lango Allen *et al* 2010). Efforts to develop similarly large datasets in schizophrenia are currently underway through the Psychiatric GWAS Consortium (Psychiatric GWAS Consortium 2009, Sullivan 2010).

6.1.2 Copy-number variation

Data from SNP genotyping arrays has been used to identify rare copy-number variants which may not only cause schizophrenia (ISC 2008, Stefansson *et al* 2008) but also other psychiatric and developmental disorders such as autism and attention-deficit/hyper-activity disorder (Guilmatre *et al* 2009, Williams *et al* 2010). Recent estimates suggest that these rare copy-number variants account for ~1% of schizophrenia cases which is higher than has been reported for other non-psychiatric common diseases (Sebat *et al* 2009, Feng & Zhu 2010). CNV studies using SNP arrays have focussed on relatively large structural variants (e.g. >100 Kb, ISC 2008). Smaller variants may make a

significant contribution to susceptibility but have been relatively under-investigated. This is likely to change with the availability of next-generation sequencing methods.

6.1.3 Next-generation sequencing

No studies of schizophrenia have yet been published using this technology which allows for more efficient variant detection and at a massive scale to previous dideoxy terminator sequencing. It is a powerful tool in the search for 'the missing heritability' of schizophrenia and can be applied to the fine mapping of common variant association signals or to the search for rare high penetrant risk mutations. It is currently being used in the Neuropsychiatric Genetics laboratory, TCD to screen all exons of 215 genes in 350 schizophrenia cases and 350 controls for rare causative mutations. The genes were selected from results of genome-wide association studies, CNV studies and functional prioritisation.

6.2 Summary of the aims of the project

Evidence from genetic epidemiology supported the view that schizophrenia is a complex genetic disorder of substantial heritability, which likely involves combined action of multiple common genetic variants of modest effect interacting with each other and with environmental risk (Risch & Baron 1984, Risch & Merikangas 1996). The present study began with the aim of identifying schizophrenia susceptibility genes using candidate gene association studies for which the selection criteria were based upon positional and functional information. Following the GWAS performed by the ISC, the aim shifted to the characterisation of novel susceptibility variants. Finally, the introduction of next-generation sequencing and an increasing emphasis on the role of rare variants in schizophrenia necessitated a study into the detection of copy-number variation using the new ultra-high throughput technology.

6.3 Rationale for the identification of candidate susceptibility loci

6.3.1 Use of linkage meta-analysis data

The positional evidence was provided by genome-wide linkage studies. When this project was started twenty genome-wide linkage scans had been performed in schizophrenia and subjected to a meta-analysis (Levinson *et al* 2003, Lewis *et al* 2003). This meta-analysis identified nine regions with significant evidence of linkage and a further ten with suggested evidence of linkage. Linkage data was not available for the in-house sample. Thus, all nineteen of the regions with evidence of linkage from the meta-analysis were used as a positional guide for the selection of candidate genes for an association study. This approach was supported by the identification of the *DTNBP1* (Straub *et al* 2002) and the replication of association signals at *NRG1* (Corvin *et al* 2004), both in schizophrenia samples from Ireland.

In using the meta-analysis results it was assumed that these data would accurately predict the positions of susceptibility loci in the in-house sample. However, genetic heterogeneity caused by the inclusion of different types of samples (e.g. extended pedigrees, trios) and of samples from disparate populations (e.g. Finland, Palau, Costa Rica) may have introduced noise into the meta-analysis.

A second meta-analysis has more recently been performed, this time incorporating results from thirty-two linkage studies (Ng *et al* 2009a). However, this second study only identified ten bins significant or suggested significant evidence of linkage (compared to nineteen in the previous study). Thus the regions assumed to have evidence of linkage from the first study may not have been completely accurate. An extensive comparison between the results of the two studies is impeded by the fact that they used different 'bin' co-ordinates. Nevertheless significant results were identified in both studies at chromosome 2, 118.7–152 Mb which broadly marks the location of the *ZNF804A* gene and at chromosome 5, 141.8-162.1 Mb. The linkage peak on chromosome 15q, which lead to the selection of the *HOMER2* for association analysis, was not significant in the second linkage meta-analysis study. It will be interesting to compare the results of these linkage studies with those of genome-wide association studies of schizophrenia. However, making a final judgement on the validity of using positional information from the linkage data to guide candidate gene selection will require a much better understanding of the genetic architecture of schizophrenia than is currently available.

6.3.2 Functional prioritisation of candidate genes

Observational studies of schizophrenia have contributed significantly to the understanding of its biology. At the start of this study, the three major schizophrenia genes which had been identified by linkage and showed promising results from replication association studies (*DTNBP1*, *NGR1* and *G72/DAO*) all had important functions at glutamatergic synapses (Harrison & Owen 2003, Harrison & Weinberger 2005). Because of the convergence of the genetic evidence in supporting the 'glutamate hypothesis of schizophrenia', genes involved in glutamate synaptic function, were preferentially selected for investigation (Coyle 1999).

The approach taken for the identification of functionally-relevant genes was one of data-mining in which two on-line databases were queried. A proteomic study of a mouse synapse was also used to supplement the synaptic gene list (Shrimpf *et al* 2005). By primarily considering only genes which were common to both search criteria spurious results could be excluded. An additional step could have been to screen out genes, which were also important in other neuro-transmitter systems because they are not strictly part of the glutamate hypothesis of schizophrenia. An association study on the resulting list of genes would have been a fuller exploration of the glutamate hypothesis.

Although great advancements have been made in the last ten years, one major limitation of candidate gene selection is insufficient knowledge and dissemination of gene/protein function. Interestingly, many established disease genes had no known function prior to their identification as a disease gene. Examples of this include Disrupted-in-Schizophrenia-1 (Millar *et al* 2000), huntingtin (MacDonald *et al* 1993), 'Fat mass and obesity-associated' (Frayling *et al* 2007) and *ORMDL3* (in asthma, Moffatt *et al* 2007). It is to be hoped that as the range of identified risk variants for schizophrenia expands, this will inform more targeted investigation of specific molecular pathways and mechanisms. Essentially, information derived from GWAS and sequencing studies can be 'mined' based on hypotheses with greater prior probability as understanding of molecular aetiology develops.

6.4 Candidate gene studies of schizophrenia

6.4.1 Candidate gene association study in the GASP schizophrenia sample

In the present study a tagging-SNP approach was used for selecting SNPs to genotype and test for association, which was supplemented using data generated in-house on predicted functional SNPs for which genotype data was not available in HapMap. Sixty-six SNPs in nine genes were genotyped in 375 schizophrenia cases and 812 controls with an average tag coverage per gene of 83.5%. Of the nine genes studied, a total of five nominally-significant associations were identified at *SYNGAP1*, *CNDP1* and *HOMER2*. Although the single associated SNP at *SYNGAP1* had the lowest p-value, it was located in a neighbouring gene, which may have complicated interpretation and functional analysis. Thus, the *HOMER2* gene was taken forward for further investigation because multiple independent association signals were identified at the locus.

6.4.2 Association studies of the *HOMER2* gene

In the first stage, twelve SNPs at the *HOMER2* gene were studied, which gave 67% coverage of the common variation at the locus based on HapMap data. Three LD-independent SNPs were significantly associated with the phenotype. Following a second stage of association analysis, coverage increased to 84% and a stronger association was identified in a predicted functional site 8bp 3′ of exon 6 (rs2306428, p=0.006, OR=0.61). Three SNPs on different haplotypes were tested for association in a combined Irish sample (GASP+RPGI) of which only rs2306428 was significantly associated (p=0.008, OR=0.73). Replication of this association was tested in the ISC GWAS sample through its proxy on the Affymetrix 6.0 chip (rs17158194, r^2=1.0 in CEU, not genotyped on the 5.0 chip). Thus replication was tested in 1,287 cases and 1,128 controls (from Bulgaria, Edinburgh and Sweden). A significant association with schizophrenia was identified at rs17158194 and in the same direction as that identified in the combined GASP+RPGI sample (χ^2=5.53, p=0.019, OR=0.77, 95% C.I. 0.60-0.96).

The association signal at *HOMER2* in the ISC GWAS is not the most significant association in that dataset and is potentially an example of the limitations of GWAS analysis. Most GWAS studies have identified only a small fraction of disease susceptibility which may be present because individual effects are too small to pass stringent significance tests, sample sizes are too small, or a

lack of power to assay causal variants with a low minor allele frequency (Yang *et al* 2010). Recent data suggests that a substantial proportion of variance in schizophrenia susceptibility is polygenic and due to many variants of small effect (ISC 2009). Larger sample sizes may address this issue, as has recently been demonstrated for height (Lango Allen *et al* 2010).

6.4.3 Association analysis of *CDC42* gene

Although genome-wide association studies are primarily intended for a 'hypothesis-free' screen for susceptibility loci, they can also be used to test existing hypotheses. Given the evidence for the involvement of *HOMER2* in schizophrenia, the association signals at interactors of the Homer2 protein were investigated in the ISC GWAS results. The strongest association signal of the interactors was at a SNP 15Kb from the *CDC42* gene (rs2473277, $p_{CMH}=2.003\times10^{-6}$, OR=0.845) which was also the 10th strongest signal in the ISC GWAS. *CDC42* was known to have an important role in dendrite development (Threadgill *et al* 1997) and be under-expressed in schizophrenia brain (Hill *et al* 2008). The GAIN and WTCCC2 schizophrenia GWAS studies were used for replication analysis of rs2473277. They had 25% and 23% power respectively to replicate the association identified in the ISC study and thus it was not surprising that rs2473277 was not significantly associated with schizophrenia in either sample. At the current time, for the reasons outlined in the last section it is difficult to assess the extent of the role that this variant has in schizophrenia susceptibility, if any. It remains an interesting candidate and both *HOMER2* and *CDC42* are included in the on-going sequencing project of candidate genes at the Neuropsychiatric Genetics laboratory, TCD (see details below).

6.4.4 Summary of association studies

During the course of this project, association studies moved from candidate-gene to GWAS studies. This has already started to revolutionise schizophrenia genetics with the identification of common susceptibility variants at genes that would not previously have been considered (e.g. *ZNF804A*). One unexpected result from GWAS across multiple disorders is the detection of strong association signals that are distant from any known gene. Such SNPs might be associated with schizophrenia because they alter regulatory sequences of a distant gene (or group of genes). It highlights the fact that the functional annotation of the human genome is still a work in progress.

The number of SNPs now genotyped on a single platform is over one million which, with additional imputation, provides extensive coverage of common genetic variation in the human genome. There are still a few problems with even genome coverage. Firstly, a significant proportion of the genome has not been sequenced or investigated for common SNPs and thus is refractory to investigation for its role in disease. Secondly, the density of genotyped SNPs across different genes may vary (e.g. 0.08-0.56 SNPs per kb for interactants of Homer2 genotyped on the Affymetrix 6.0, table 4.1). This has implications for coverage when investigating candidate genes and for pathway analyses. This is a problem which has partially been overcome with the availability of imputation data for current platforms (Spencer *et al* 2009). One reason for the variation in SNP density is that large-scale SNP discovery projects have only targeted genes of known or predicted function. However, this problem should be gradually alleviated with complete re-sequencing of human genomes for polymorphism discovery using next-generation sequencing e.g. in the 1000 Genomes Project (Via *et al* 2010, http://www.1000genomes.org/page.php).

6.5 Bioinformatic assessment of disease-associated variants

The purpose of identification of disease-susceptibility genes is to inform on disease pathogenesis, which can improve diagnostics and lead to rationalised drug development. Because of the global dissemination of biological knowledge through the internet, especially of large-scale array-based experiments, inference can be made on the biological properties of sequence and also the effects of variation in that sequence. Such *in silico* predictions are particularly useful because analysis can be used to generate specific hypotheses which can be tested in the laboratory. Thus accurate predictions of sequence function can greatly improve the time- and cost-efficiency of laboratory work.

6.5.1 Potential effect of splicing on a schizophrenia-associated SNP at HOMER2 (rs2306428)

SNP rs2306428 is located 8bp 3' of *HOMER2* exon 6 which is an important location in Homer2 protein function. In response to neuronal stimulation, the first 164bp of exon 6 are deleted which causes a frameshift due to the premature introduction of a stop codon (Soloviev *et al* 2000b).

Truncated Homer2 proteins are thus expressed that possess the N-terminal EVH1 domain bind to post-synaptic adaptor and receptor molecules (Shiraishi *et al* 2003). However, they lack the C-terminal coiled-coil domain required for self-binding, thus separating the binding partners and preventing excito-toxicity (Brakeman *et al* 1997).

Because rs2306428 was located close to the boundary of exon 6, it was investigated for an effect on splicing using the 'ESRsearch' tool (http://ast.bioinfo.tau.ac.il/, after Zhang & Chasin 2004). The result of the ESRsearch analysis indicated that the rare allele of rs2306428, which was more frequent in controls, abolished a predicted splice-enhancer binding site.

However, the binding sequence is in intronic sequence, occurs at the opposite end of the exon to the natural truncation and, the direction of association indicates that the rare allele which removes the binding-site is protective. This appears paradoxical because the disruption of normal protein function is generally regarded as a contributory factor to disease rather than being protective. Thus, reduction of Homer2 truncation in response to neuronal stimulation may protect an otherwise susceptible individual against schizophrenia. Alternatively it cannot discounted that rs2306428 is in LD with an as yet unknown causative susceptibility variant.

6.5.2 Bioinformatic assessment of rs2473277 at the *CDC42* gene

The availability of the results of a genome-wide association study of global gene expression in lymphoblast cell lines (mRNA-by-SNP Browser, Dixon *et al* 2008) meant that a relationship between genotype at rs2473277 at *CDC42* expression level was rapidly established in which the schizophrenia susceptibility allele was also associated with reduced *CDC42* expression. *CDC42* expression levels in the CEU (generated and made available through the Genevar project, Stranger *et al* 2005) were also tested against genotype at rs2473277 and supported the earlier finding. One potential caveat of this analysis was that SNPs other than rs2473277 showed a stronger effect on *CDC42* expression. Whilst the existence of multiple *cis*-acting variants at a single gene has been demonstrated (Gruber & Long 2009), it is more difficult to explain why such SNPs at *CDC42* would not also be associated with schizophrenia.

Other studies of genome-wide studies of gene expression could also have been accessed to support these results. One has recently been performed in human cortical tissue and identified the greatest effect on *CDC42* expression from a *trans*-acting variant (Webster *et al* 2009, with secondary analysis undertaken by Dr Lynn Cochran and Dr Richard Anney both at the Neuropsychiatric Genetics laboratory, TCD). Although the signal would not have reached genome-wide significance, it is located

within in the gene encoding protein tyrosine phosphatase receptor, type D which is in involved in dendrite outgrowth in the hippocampus and thus biologically-plausible (Mizuno *et al* 1993, Wang & Bixby 1999, see Chapter 4, Discussion section).

SNP rs2473277 was not in a predicted functional site but was found to be in high LD with a SNP which was (rs2473307, r^2 0.93 in HapMap CEU). The predicted transcription-binding site was identified from visual inspection of the 'Regulation' tracks in the UCSC Genome Browser and then interrogated using 'TFSEARCH' to find that the susceptibility allele at rs2473307 removes binding sites for GATA-X, GATA-2 or c-Ets transcription factors. Although this seemed like reasonable evidence upon which to start functional experiments, there were SNPs in the 5' region of *CDC42* which could have also been functional and in LD with rs2473277. At this point it may have been beneficial to re-sequence functional sites in *CDC42* and integrate the data in the existing HapMap as had been undertaken previously with the *HOMER2* gene.

6.6 Laboratory testing of associated SNPs with predicted functionality

6.6.1 Correlation of HOMER2 expression with genotype at rs2306428

The associated and predicted functional SNP rs2306428 at *HOMER2* was predicted to have an effect on splicing whereby carriers of the rare, protective allele would have a higher expression of long, relative to short Homer2 isoforms. To investigate this, *HOMER2* expression in CEU lymphoblastoid cells was determined by reverse transcription of the mRNA and quantitative PCR. Two independent PCR reactions were performed – one in which the primers bound only to long *HOMER2* isoforms and one in which the primers bound to both long and short *HOMER2* isoforms. No significant effect of genotype at rs2306428 was observed on expression level of *HOMER2* as determined by either primer pair, nor on the difference between values obtained by each primer pair. As in the replication analysis of rs2306428, its low minor allele frequency would have reduced the power to identify an effect. Furthermore, the specialised splicing of *HOMER2* mRNA to form truncated proteins occurs just after neuronal stimulation – in the present study the cells were neither neuronal nor stimulated. One alternative would have been to clone exon 6 harbouring one of each allele at rs2306428 into an 'Exon Trap' vector from which the RNA production from stimulated neuronal cells could have been determined.

6.6.2 Reporter gene and electrophoretic mobility shift assays with *CDC42* SNP rs2473307

A SNP at the *CDC42* gene, in high LD with rs2473277, rs2473307 was tested for effect on reporter gene expression. A significant increase in activity from the base-line occurred in the presence of the protective allele only in a neuronal cell line but not in a non-neuronal cell-line. Binding of transcription factors to rs2473307 was tested using an electrophoretic mobility shift assay. However, the results were not conclusive or supportive of the effect of rs2473277 on protein binding. Thus the results of the reporter gene expression assay require further functional validation. A classical explanation for the difference in the result between two cell types would be that the effect was tissue-specific. However, the protein synthesis rates in the CHO cells are far greater than those of the SHSY-5Y. This means that results might not be comparable between the two and in addition, that the SHSY-5Y cells might be more sensitive to differences in the concentration of test plasmids. One method of resolving this problem would be to replicate the experiment using different clones.

6.7 Endophenotype investigation of the associated SNPs at *HOMER2* and *CDC42*

The concept of endophenotype analysis is based on the assumption that a complex phenotype (e.g. schizophrenia) is composed of heritable components, termed endophenotypes, and that these will have a simpler genetic architecture which may facilitate gene discovery. We tested three specific heritable domains of cognitive performance, which cover general IQ, memory and attention. These tap neurocognitive processes which can in part be tested in animal systems and may be more amenable to investigation than 'schizophrenia' the clinical disorder. For this analysis we had access to a reasonably large sample of clinical cases (349) and controls (230). For this study, *HOMER2* rs2306428 and *CDC42* rs2473277 were tested for effects on fifteen neuropsychological variables covering general cognitive ability (IQ), memory and attention.

No significant effects of genotype at rs2306428 in the *HOMER2* gene were identified on any of the neuropsychological variables tested. Because this SNP has a low minor allele frequency it is difficult to detect significant associations. At *CDC42* rs2473277 significant effects were identified in Working Memory and Episodic Memory ($p<0.05$). In both instances, heterozygosity at rs2473277 conferred enhanced performance. These results came from a combined case and control sample and no significant effect was detected in the two separate samples. Replication in a sample of similar or larger size is required. Given that the strongest association with *CDC42* expression was at rs2268177

($p=6.3\times10^{-37}$, in the Dixon *et al* (2008) study) and the central importance of *CDC42* expression level on senesence (Kerber *et al* 2009, Wang *et al* 2007b), this SNP may be a major influence on human longevity.

It is difficult to estimate the power of the sample used in the present study because the effect size of the contributing variants is unknown. From the present study we can exclude the *HOMER2* SNP from having a large effect on variance in performance on these tasks. It is yet to be determined to what extent different common risk variants impact on different aspects of cognitive performance, or whether these are more highly penetrant effects.

6.9 Detection of copy-number variation by next-generation sequencing

Nucleotide sequencing provides a map of genetic variation in an individual at the resolution of a single base-pair and thus is an extremely powerful tool for the investigation of disease genetics. Traditional dideoxy terminator sequencing has recently been surpassed by 'next-generation' sequencing technologies which have vastly higher throughput and reduced operating costs (Mardis *et al* 2008 for review). Rare-copy number variation has informed significantly on the biology of schizophrenia (Merikangas *et al* 2009 for review) and is detectable by next-generation sequencing because the quantity of data produced is dependent on the copy-number of the target DNA.

In Chapter 5 of the present study, the ability to detect known copy-number variants in the HapMap CEU sample was tested using next-generation sequencing. The role of the author in this group project was firstly to identify known CNVs which were suitable for investigation. This was achieved through screening-out CNVs using population genetic parameters (minor allele frequency, Mendelisation, Hardy-Weinberg equilibrium). The number of sequence reads generated per 100bp window was used as raw data, normalised by GC-content and then screened for consistent reductions in signal strength which would indicate the presence of a CNV. Of fourteen targeted CNVs (which were not affected by experimental error), twelve were detected. Importantly, the normalised read depth values correlated with the known genotypes and CNVs were accurately identified even when only heterozygous deletions were present in the nine tested individuals. Of the two CNVs which were not detected, one was only 80bp long and the second, a duplication present only as a heterozygote in a single tested individual and was 840bp long. Given that the shortest CNV detected was 600bp in length, the lower size limit for CNV detection using the method developed in the present study remains to be determined and there are likely to be other influencing factors such as GC-content and total length of sequence analysed.

A second role of the author was to determine the exact break-points of the CNVs using dideoxy terminator sequencing. This was necessary primarily to provide a comparison as to how accurate the next-generation sequencing was. Here, of the sixteen break-points covered by 'bait', next-generation sequencing identified nine to within the same 100bp analysis window. Of the remaining seven break-points, the worst result from the next-generation sequencing estimated the break-point to be 600bp away from the true breakpoint. Because the CNVs were originally identified through a specialised genotyping array (Affymetrix 6.0 SV, McCarroll *et al* 2008), determination of the exact break-points using dideoxy terminator sequencing provided a comparison for this method as well. This showed that on average the genotyping array approach estimated the start or stop location of the CNV to within 1115bp of its true location.

In summary, the method was competent at detecting large homozygous and heterozygous deletions. It was developed with the intention of screening the exons of 215 candidate genes sequenced using the Illumina Genome Analyser II for CNVs in schizophrenia patients. This ability will be tested bioinformatically using publicly-available data from whole-exome resequencing (Ng *et al* 2009a). Once proven, the method can be applied to data generated from sequencing of candidate schizophrenia genes and used by researchers in other projects.

6.10 Future directions for schizophrenia genetics research

6.10.1 Mapping of causative loci from genome-wide association signals

Association signals often do not come directly from the causative variant. Thus, the next stage is to identify SNPs which are in LD with the genotyped one. This is commonly done by 'expanding' the LD using patterns inferred from a reference sample such as the CEU. The genotype data in the CEU may be supplemented with data generated by resequencing in the region associated interval. This approach would be most beneficial if the region has not previously undergone an extensive screen of common variation, otherwise most common variants should already have been identified and genotyped in the CEU. The associated interval may be very large e.g. 5Mb in the HLA region, in which case it could be more efficient to sequence functional sites such as exons, promoters and other untranslated regions. The problem with this approach is that discovery of a functional variant in high LD with the associated variant does not exclude the possibility that there may be a third functional and linked SNP which is located outside of the sequenced sites. Complete genome sequence in

reference samples (e.g. 1,000 Genomes data) should give complete maps of LD at associated loci. However, LD can also confound efforts to identify the causative SNP where multiple putative functional SNPs all lie on one associated haplotype. Fine-mapping in additional samples with different background patterns of LD may help identify the causative SNP, assuming there is not genetic heterogeneity between populations of different ethnic background.

The *ZNF804A* gene contains a common variant that is highly associated with schizophrenia (Williams *et al* 2010). Investigation of the role of rare variants in the *ZNF804A* gene in schizophrenia susceptibility has not yielded any positive findings (Dwyer *et al* 2010). This approach was not unreasonable as genes which harbour common susceptibility variants (identified through GWAS) were previously known to harbour rare, highly-penetrant variants for the same disorder e.g. *SNCA*, *MAPT* and *LRRK2* in Parkinson's Disease (Satake *et al* 2009, Simón-Sánchez *et al* 2009, Edwards *et al* 2010). The complete re-sequencing of schizophrenia genes identified through GWAS should provide a very interesting insight into the spectrum of both common and rare susceptibility variants at these loci.

Although the three GWAS of Parkinson's disease cited above did not screen out affected individuals with a family history, it is unlikely that rare, highly penetrant mutations in the associated genes would have been the cause of the association signals. In fact it is more likely that the inclusion of individuals with a family history of disease reduces the power of an association study because common variants are less likely to be the cause of disease in these individuals. Such an influence may be important in GWAS of schizophrenia because approximately one-third of patients have an affected first- or second-degree relative.

6.10.2 Interpretation of results from exome sequencing

The sequencing of exons of candidate functional genes will no doubt identify many novel variants. It is of great importance to demonstrate which ones cause disease and this task will be influenced by the number of variants identified. Recently, exon sequencing of 18,654 genes in 200 control individuals identified 53,081 coding SNPs with a minor allele frequency greater than 0.02 (Li *et al* 2010). In the context of sequencing association studies, this represents a considerable quantity of data to test for causation. The issue becomes prioritization: first, where variants are identified as having a population frequency greater than e.g. 0.01, they are unlikely to represent highly penetrant variants. This will require substantial public databasing of genome sequence information, as has recently been proposed for genomic structural variation (Church *et al* 2010). Second, identification

of rare causative variants has traditionally been achieved through linkage mapping and positional cloning in an extended pedigree. Having evidence that a mutation segregates with disease in a family is strong evidence to implicate the variant, without necessarily proving causation. In schizophrenia, collection of large pedigrees is not common and so this approach may not be fully possible currently, but suggests a requirement for further follow-up of families of mutation carriers. This information could be combined with information about *de novo* events where parental DNA is available. The exploitation of large scale genome sequence information in these scenarios will require substantial international infrastructure to provide information on mutations in large case and control collections.

An alternative approach could be a gene-based prioritisation method in which the total number of rare coding variants detected in each gene could be compared between cases and controls. Here, a gene which harbours multiple unique variants in the case sample but none or few in the control sample could be considered *prima facie* as a causative candidate. Weighting could also be placed on individual variants themselves using prior knowledge of the effects of coding sequence variation.

Inferring function of coding variants is generally simpler than it is for intronic and extra-genic ones. Synonymous variants tend to be non-functional although in some instances they can lead to disease by altering RNA secondary structure (O'Driscoll *et al* 2003) or by introducing a splice-site (Faa et al, 2010). Non-synonymous variants (which can be sub-divided into mis-sense and non-sense mutations) are more likely to have a functional affect, especially if they show cross-species conservation. A number of on-line tools are available and commonly used to score the effect of mutations of protein structure e.g. 'PMut' (Ferrer-Costa *et al* 2005, http://mmb.pcb.ub.es/PMut/) and 'PolyPhen-2' (http://genetics.bwh.harvard.edu/pph2/).

The exon resequencing study by Li *et al* (2010) identified an enrichment of rare non-synonymous variants compared to synonymous variants which is indicative of strong purifying selection. The authors interpreted this as evidence for a role of recessive non-synonymous variants in disease. The implication of rare, homozygous coding polymorphisms in schizophrenia pathogenesis is an exciting possibility because it represents an as yet unknown intermediate genetic contribution to heritability between common variants of small effect e.g. from *ZNF804A,* and very rare, private mutations e.g. the *DISC1* translocation.

Next-generation sequencing technologies also have the ability to study epigenetic modifications such as methylation which profoundly affect gene function and are responsive to the environment (Hurd & Nelson 2009). In addition to such approaches to investigate sequence variation

at the level of the gene, the next decade will likely see the emergence of many more methods of analysis of pathways and epistasis.

6.10.3 Functional studies of candidate susceptibility variants

In the present study, mRNA quantification, gene reporter and electrophoretic mobility shift assays were used to test a functional hypotheses on candidate SNPs. Such methods are the tip of the iceberg in the exploration of disease mechanisms with a genetic origin. For example, the use of *post mortem* brain has recently shown that *cis*-acting effects of common genetic variation on *ZNF804A*, *NOS1*, *RGS4*, *TCF4* and *AKT1* expression differ by region (Buonocore *et al* 2010). This highlights the importance of cell type in *in vitro* studies as well as providing a rare insight into the apparent regional specificity of schizophrenia psychopathology. Thus increased availability of *post mortem* brain from individuals with schizophrenia would greatly assist in the study of disease mechanisms.
Mouse models of candidate gene disruption have been an invaluable source of information because they are amenable to behavioural and neurological investigation at a level that is not possible in human subjects (O'Tuathaigh & Waddington 2010). This means that the function of the expressed proteins can be studied *in vivo* and the biochemical pathway which ultimately leads to disease investigated for genetic aberrations in genes which encode the constituent proteins.

Current techniques to investigate disease pathways are suited for single genes. However, the data generated by genome-wide association and resequencing experiments is likely to be reflective of the many common susceptibility variants, which each have a subtle effect on gene function and varies temporally and by cell type. New technical developments have enabled high-throughput, high-content (HT-HC) cellular screens which can analyse and combine data from millions of individual experiments. The type of experiments may range from studies of epigenetic modifications to cell-cell interactions. Of relevance to schizophrenia might be the ability to quantify axonal growth (Sepp *et al* 2008), protein interactions of candidate genes or the effects of gene suppression using RNA-interference (RNAi). Significantly, HT-HC can be multi-dimensional in which the simultaneous quantification of several different processes provides multiple avenues for identifying processes that may lead to disease (Jain & Heutink 2010 for review).

HT-HC can be used to study many different cell types. The present study was limited in that the only neuronal cell type used was SHSY-5Y, which was originally isolated from a bone marrow neuroblastoma and thus may not be an accurate representation of a cortical or hippocampal neuron. Embryonic stem (ES) cells and induced pluripotent stem (iPS) cells can potentially be grown

into any cell type and in unlimited supply (Salero & Hatten 2007). The use of ES and iPS cell lines in future experiments on candidate genes may provide a much more accurate view of gene function and dysfunction within the brain of schizophrenia patients.

Finally, the identification of rare, highly-penetrant mutations in a small but significant subset of schizophrenia patients creates an *in vivo* example of single-gene disruption which, subject to appropriate ethical guidelines, provides an opportunity for exciting clinical research.

6.11 Conclusion

Whilst the heritability of schizophrenia is still elusive, the technology now exists for its discovery in the form of genome-wide association studies and next-generation sequencing. Furthermore, with technological advancements in functional genomics, the combined functional action of multiple susceptibility variants can be studied. Eventually, it is hoped that through the identification of genetic susceptibility variants, perhaps clustering in specific pathways, improved diagnostics and therapeutics will become a reality for schizophrenia, as is in the case for other common diseases (Aartsma-Rus & van Ommen 2009, Cuny 2009 for reviews). The convergence of informatics, high-throughput biochemical analysis and increasing sample size is generating vast amounts of data. Such experiments must be designed and analysed both imaginatively and with strict rigour if current and future investments are to deliver a better outcome for patients.

On-line resources

1000 Genomes Project

http://www.1000genomes.org/page.php

Affymetrix (microarray probe sequences)

http://www.affymetrix.com/Auth/analysis/downloads/data/

Cis-ter (**Cis**-elementer clus**ter** finder)

http://zlab.bu.edu/~mfrith/cister.shtml

Database of Genomic Variants (DGV)

http://projects.tcag.ca/variation/

DAVID (The **D**atabase for **A**nnotation, **V**isualization and **I**ntegrated **D**iscovery)

http://david.abcc.ncifcrf.gov/home.jsp

ESE (exonic spice enhancer) Search tool

http://ast.bioinfo.tau.ac.il/

Genetic Power Calculator

http://pngu.mgh.harvard.edu/~purcell/gpc/

Genevar project

http://www.sanger.ac.uk/humgen/genevar/

Haploview

http://www.broadinstitute.org/haploview

Michigan University QTL software

http://www.sph.umich.edu/csg/liang/

PLINK

http://pngu.mgh.harvard.edu/~purcell/plink/

TFsearch

http://www.cbrc.jp/research/db/TFSEARCH

The International Schizophrenia Consortium (ISC)

http://pngu.mgh.harvard.edu/isc/

The UCSC Genome Browser

http://genome.ucsc.edu/

References

1. Aartsma-Rus A & van Ommen GJ (2009) Less is more: therapeutic exon skipping for Duchenne muscular dystrophy. Lancet Neurol 8(10):873-5
2. Abi-Dargham A & Moore H (2003) Prefrontal DA transmission at D1 receptors and the pathology of schizophrenia. Neuroscientist 9:404-416
3. Abou-Sleiman PM, Hanna MG, Wood NW (2006) Genetic association studies of complex neurological diseases. J Neurol Neurosurg Psychiatry 77(12):1302-4
4. Agarwal N, Port JD, Bazzocchi M, Renshaw PF (2010) Update on the use of MR for assessment and diagnosis of psychiatric diseases. Radiology 255(1):23-41
5. Ajima R, Kajiya K, Inoue T, Tani M, Shiraishi-Yamaguchi Y, Maeda M, Segawa T, Furuichi T, Sutoh K, Yokota J (2007) HOMER2 binds MYO18B and enhances its activity to suppress anchorage independent growth. Biochem Biophys Res Commun 356(4):851-6
6. Aleman A, Kahn RS, Selten JP (2003) Sex differences in the risk of schizophrenia: evidence from meta-analysis. Arch Gen Psychiatry 60(6):565-7
7. Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, Hormozdiari F, Kitzman JO, Baker C, Malig M, Mutlu O, Sahinalp SC, Gibbs RA, Eichler EE (2009) Personalized copy number and segmental duplication maps using next-generation sequencing. Nat Genet 41(10):1061-7
8. Allen NC, Bagade S, McQueen MB, Ioannidis JP, Kavvoura FK, Khoury MJ, Tanzi RE, Bertram L (2008) Systematic meta-analyses and field synopsis of genetic association studies in schizophrenia: the SzGene database. Nat Genet 40(7):827-34
9. Altshuler D, Daly MJ, Lander ES (2008) Genetic mapping in human disease. Science 322(5903):881-8
10. Amann LC, Gandal MJ, Halene TB, Ehrlichman RS, White SL, McCarren HS, Siegel SJ (2010) Mouse behavioral endophenotypes for schizophrenia. Brain Res Bull 83(3-4):147-61
11. American Psychiatric Association (2000) Diagnostic and statistical manual of mental disorders, 4th edition. Washington, DC
12. Andreasen NC (1982) Negative symptoms in schizophrenia. Definition and reliability. Arch Gen Psychiatry 39(7):784-8
13. Anis NA, Berry SC, Burton NR, Lodge D (1983) The dissociative anaesthetics, ketamine and phencyclidine, selectively reduce excitation of central mammalian neurones by N-methyl-aspartate. Br J Pharmacol 79:565-575

14. Balion CM, Benson C, Raina PS, Papaioannou A, Patterson C, Ismaila AS (2007) Brain-type carnosinase in dementia: a pilot study. BMC Neurol 7:38
15. Barrett JC, Clayton DG, Concannon P, Akolkar B, Cooper JD, Erlich HA, Julier C, Morahan G, Nerup J, Nierras C, Plagnol V, Pociot F, Schuilenburg H, Smyth DJ, Stevens H, Todd JA, Walker NM, Rich SS; Type 1 Diabetes Genetics Consortium (2009) Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. Nat Genet 41(6):703-7
16. Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21(2):263-5
17. Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, Rioux JD, Brant SR, Silverberg MS, Taylor KD, Barmada MM, Bitton A, Dassopoulos T, Datta LW, Green T, Griffiths AM, Kistner EO, Murtha MT, Regueiro MD, Rotter JI, Schumm LP, Steinhart AH, Targan SR, Xavier RJ; NIDDK IBD Genetics Consortium, Libioulle C, Sandor C, Lathrop M, Belaiche J, Dewit O, Gut I, Heath S, Laukens D, Mni M, Rutgeerts P, Van Gossum A, Zelenika D, Franchimont D, Hugot JP, de Vos M, Vermeire S, Louis E; Belgian-French IBD Consortium; Wellcome Trust Case Control Consortium, Cardon LR, Anderson CA, Drummond H, Nimmo E, Ahmad T, Prescott NJ, Onnie CM, Fisher SA, Marchini J, Ghori J, Bumpstead S, Gwilliam R, Tremelling M, Deloukas P, Mansfield J, Jewell D, Satsangi J, Mathew CG, Parkes M, Georges M, Daly MJ (2008) Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat Genet 40(8):955-62
18. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Keira Cheetham R, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu X, Maisinger KS, Murray LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR, Rasolonjatovo IM, Reed MT, Rigatti R, Rodighiero C, Ross MT, Sabot A, Sankar SV, Scally A, Schroth GP, Smith ME, Smith VP, Spiridou A, Torrance PE, Tzonev SS, Vermaas EH, Walter K, Wu X, Zhang L, Alam MD, Anastasi C, Aniebo IC, Bailey DM, Bancarz IR, Banerjee S, Barbour SG, Baybayan PA, Benoit VA, Benson KF, Bevis C, Black PJ, Boodhun A, Brennan JS, Bridgham JA, Brown RC, Brown AA, Buermann DH, Bundu AA, Burrows JC, Carter NP, Castillo N, Chiara ECM, Chang S, Neil Cooley R, Crake NR, Dada OO, Diakoumakos KD, Dominguez-Fernandez B, Earnshaw DJ, Egbujor UC, Elmore DW, Etchin SS, Ewan MR, Fedurco M, Fraser LJ, Fuentes Fajardo KV, Scott Furey W, George D, Gietzen KJ, Goddard CP, Golda GS, Granieri PA, Green DE, Gustafson DL, Hansen NF, Harnish K, Haudenschild CD, Heyer NI, Hims MM, Ho JT, Horgan AM, Hoschler K, Hurwitz S, Ivanov DV, Johnson MQ, James T, Huw Jones TA, Kang GD, Kerelska TH, Kersey AD, Khrebtukova I, Kindwall AP, Kingsbury Z, Kokko-Gonzales PI, Kumar A, Laurent MA, Lawley CT, Lee SE, Lee X, Liao AK, Loch JA, Lok M, Luo S, Mammen RM, Martin JW, McCauley PG, McNitt P, Mehta P, Moon KW, Mullens JW,

Newington T, Ning Z, Ling Ng B, Novo SM, O'Neill MJ, Osborne MA, Osnowski A, Ostadan O, Paraschos LL, Pickering L, Pike AC, Chris Pinkard D, Pliskin DP, Podhasky J, Quijano VJ, Raczy C, Rae VH, Rawlings SR, Chiva Rodriguez A, Roe PM, Rogers J, Rogert Bacigalupo MC, Romanov N, Romieu A, Roth RK, Rourke NJ, Ruediger ST, Rusman E, Sanches-Kuiper RM, Schenker MR, Seoane JM, Shaw RJ, Shiver MK, Short SW, Sizto NL, Sluis JP, Smith MA, Ernest Sohna Sohna J, Spence EJ, Stevens K, Sutton N, Szajkowski L, Tregidgo CL, Turcatti G, Vandevondele S, Verhovsky Y, Virk SM, Wakelin S, Walcott GC, Wang J, Worsley GJ, Yan J, Yau L, Zuerlein M, Mullikin JC, Hurles ME, McCooke NJ, West JS, Oaks FL, Lundberg PL, Klenerman D, Durbin R, Smith AJ (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456:53-59

19. Bhattacharjee A,Hodges E, Gordon B, Rooks M, Xuan Z, Brizuela L, McCombie WR, Hannon G (2009) Complementing Next Generation Sequencing Technologies with Agilent's SureSelect DNA Capture Array. Application note, Agilent Technologies
20. Bleuler PE (1911) Dementia Praecox or the group of Schizophrenias. Translated by J Zinkin. New York. International Universities Press, 1911
21. Bliss TV & Collingridge GL (1993) A synaptic model of memory: long-term potentiation in the hippocampus. Nature 361:31-39
22. Böök JA (1953) Schizophrenia as a gene mutation. Acta Genet Stat Med 4(2-3):133-9
23. Bowtell DD (199) Options available - from start to finish - for obtaining expression data by microarray. Nat Genet 21(1 Suppl):25-32
24. Brakeman PR, Lanahan AA, O'Brien R, Roche K, Barnes CA, Huganir RL, Worley PF (1997) Homer: a protein that selectively binds metabotropic glutamate receptors. Nature 386(6622):284-8
25. Bramon E, Croft RJ, McDonald C, Virdi GK, Gruzelier JG, Baldeweg T, Sham PC, Frangou S, Murray RM (2004) Mismatch negativity in schizophrenia: a family study. Schizophr Res 67(1):1-10
26. Bray NJ, Buckland PR, Williams NM, Williams HJ, Norton N, Owen MJ, O'Donovan MC (2003) A haplotype implicated in schizophrenia susceptibility is associated with reduced *COMT* expression in human brain. Am J Hum Genet 73(1):152-61
27. Brown S (1997) Excess mortality of schizophrenia. A meta-analysis. Br J Psychiatry 171:502-8
28. Bruner J & Tauc L (1965) [Synaptic Plasticity Involved in the Habituation Process in Aplysia.]. J Physiol (Paris) 57:230-231
29. Buonanno A & Fischbach GD (2001) Neuregulin and ErbB receptor signaling pathways in the nervous system. Curr Opin Neurobiol 11(3):287-96

30. Buonocore F, Hill MJ, Campbell CD, Oladimeji PB, Jeffries AR, Troakes C, Hortobagyi T, Williams BP, Cooper JD, Bray NJ (2010) Effects of *cis*-regulatory variation differ across regions of the adult human brain. Hum Mol Genet 19(22):4490-6
31. Burdick KE, Gunawardane N, Woodberry K, Malhotra AK (2009) The role of general intelligence as an intermediate phenotype for neuropsychiatric disorders. Cogn Neuropsychiatry 14(4-5):299-311
32. Burdick KE, Lencz T, Funke B, Finn CT, Szeszko PR, Kane JM, Kucherlapati R, Malhotra AK (2006) Genetic variation in DTNBP1 influences general cognitive ability. Hum Mol Genet 15(10):1563-8
33. Bustillo JR, Rowland LM, Mullins P, Jung R, Chen H, Qualls C, Hammond R, Brooks WM, Lauriello J (2010) 1H-MRS at 4 tesla in minimally treated early schizophrenia. Mol Psychiatry 15:629-636
34. Camargo LM, Collura V, Rain JC, Mizuguchi K, Hermjakob H, Kerrien S, Bonnert TP, Whiting PJ, Brandon NJ (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):74-8
35. Campbell CD, Ogburn EL, Lunetta KL, Lyon HN, Freedman ML, Groop LC, Altshuler D, Ardlie KG, Hirschhorn JN (2005) Demonstrating stratification in a European American population. Nat Genet 37(8):868-72
36. Cannon M, Kendell R, Susser E, Jones P (2003) Prenatal and perinatal risk factors for schizophrenia. In: The Epidemiology of Schizophrenia. Murray RM, Jones PB, Susser E, van Os J, Cannon M, eds. University Press, Cambridge, UK
37. Cardno AG & Gottesman II (2000) Twin studies of schizophrenia: from bow-and-arrow concordances to Star Wars Mx and functional genomics. Am J Med Gen 97:12-17
38. Cardno AG, Marshall EJ, Coid B, Macdonald AM, Ribchester TR, Davies NJ, Venturi P, Jones LA, Lewis SW, Sham PC, Gottesman II, Farmer AE, McGuffin P, Reveley AM, Murray RM (1999) Heritability estimates for psychotic disorders: the Maudsley twin psychosis series. Arch Gen Psychiatry. 56(2):162-8
39. Cardon LR & Palmer LJ (2003) Population stratification and spurious allelic association. Lancet 361(9357):598-604
40. Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA (2003) Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. Am J Hum Genet 74: 106-120
41. Carlsson A & Lindqvist M (1962) In-vivo decarboxylation of alpha-methyl DOPA and alpha-methyl metatyrosine. Acta Physiol Scand 54:87-94
42. Carpten J, Nupponen N, Isaacs S, Sood R, Robbins C, Xu J, Faruque M, Moses T, Ewing C, Gillanders E, Hu P, Bujnovszky P, Makalowska I, Baffoe-Bonnie A, Faith D, Smith J, Stephan D,

Wiley K, Brownstein M, Gildea D, Kelly B, Jenkins R, Hostetter G, Matikainen M, Schleutker J, Klinger K, Connors T, Xiang Y, Wang Z, De Marzo A, Papadopoulos N, Kallioniemi OP, Burk R, Meyers D, Grönberg H, Meltzer P, Silverman R, Bailey-Wilson J, Walsh P, Isaacs W, Trent J (2002) Germline mutations in the ribonuclease L gene in families showing linkage with HPC1. Nat Genet 30(2):181-4

43. Carter NP (1994) Cytogenetic analysis by chromosome painting. Cytometry 18:2-10
44. Chan RC, Wang Y, Zhao Q, Yan C, Xu T, Gong QY, Manschreck TC (2010) Neurological soft signs in individuals with schizotypal personality features. Aust NZ J Psychiatry 44:800-804
45. Chen L, Liao G, Yang L, Campbell K, Nakafuku M, Kuan CY, Zheng Y (2006) Cdc42 deficiency causes Sonic hedgehog-independent holoprosencephaly. Proc Natl Acad Sci USA 103:16520-16525
46. Chen X & Sullivan PF (2003) Single nucleotide polymorphism genotyping: biochemistry, protocol, cost and throughput. Pharmacogenomics J 3(2):77-96
47. Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M, Burdick JT (2005) Mapping determinants of human gene expression by regional and genome-wide association. Nature 437(7063):1365-9
48. Chisa JL & Burke DT (2007) Mammalian mRNA splice-isoform selection is tightly controlled. Genetics 175(3):1079-87
49. Choi KH, Zepp ME, Higgs BW, Weickert CS, Webster MJ (2009) Expression profiles of schizophrenia susceptibility genes during human prefrontal cortical development. J Psychiatry Neurosci 34:450-458
50. Chow EW, Bassett AS, Weksberg R (1994) Velo-cardio-facial syndrome and psychotic disorders: implications for psychiatric genetics. Am J Med Genet 54(2):107-12
51. Chumakov I, Blumenfeld M, Guerassimenko O, Cavarec L, Palicio M, Abderrahim H, Bougueleret L, Barry C, Tanaka H, La Rosa P, Puech A, Tahri N, Cohen-Akenine A, Delabrosse S, Lissarrague S, Picard FP, Maurice K, Essioux L, Millasseau P, Grel P, Debailleul V, Simon AM, Caterina D, Dufaure I, Malekzadeh K, Belova M, Luan JJ, Bouillot M, Sambucy JL, Primas G, Saumier M, Boubkiri N, Martin-Saumier S, Nasroune M, Peixoto H, Delaye A, Pinchot V, Bastucci M, Guillou S, Chevillon M, Sainz-Fuertes R, Meguenni S, Aurich-Costa J, Cherif D, Gimalac A, Van Duijn C, Gauvreau D, Ouellette G, Fortier I, Raelson J, Sherbatich T, Riazanskaia N, Rogaev E, Raeymaekers P, Aerssens J, Konings F, Luyten W, Macciardi F, Sham PC, Straub RE, Weinberger DR, Cohen N, Cohen D (2002) Genetic and physiological data implicating the new human gene G72 and the gene for D-amino acid oxidase in schizophrenia. Proc Natl Acad Sci USA 99(21):13675-80

52. Church DM, Lappalainen I, Sneddon TP, Hinton J, Maguire M, Lopez J, Garner J, Paschall J, DiCuccio M, Yaschenko E, Scherer SW, Feuk L, Flicek P (2010) Public data archives for genomic structural variation. Nat Genet 42(10):813-4
53. Clark AG (2003) Finding genes underlying risk of complex disease by linkage disequilibrium mapping. Curr Opin Genet Dev 13(3):296-302
54. Cohen BD, Rosenbaum G, Luby ED, Gottlieb JS (1962) Comparison of phencyclidine hydrochloride (Sernyl) with other drugs. Simulation of schizophrenic performance with phencyclidine hydrochloride (Sernyl), lysergic acid diethylamide (LSD-25), and amobarbital (Amytal) sodium; II. Symbolic and sequential thinking. Arch Gen Psychiatry 6:395-401
55. Cohen M, Hartlage PL, Krawiecki N, Roesel RA, Carter AL, Hommes FA (1985) Serum carnosinase deficiency: a non-disabling phenotype? J Ment Defic Res 29 (Pt 4):383-9
56. Collingridge GL, Isaac JT, Wang YT (2004) Receptor trafficking and synaptic plasticity. Nat Rev Neurosci 5:952-962
57. Collins A (2009) Allelic association: linkage disequilibrium structure and gene mapping. Mol Biotechnol 41(1):83-9
58. Collins FS, Guyer MS, Charkravarti A (1997) Variations on a theme: cataloging human DNA sequence variation. Science 278(5343):1580-1
59. Conrad DF, Andrews TD, Carter NP, Hurles ME, Pritchard JK (2006) A high-resolution survey of deletion polymorphism in the human genome. Nat Genet 38:75-81
60. Conrad DF, Bird C, Blackburne B, Lindsay S, Mamanova L, Lee C, Turner DJ, Hurles ME (2010) Mutation spectrum revealed by breakpoint sequencing of human germline CNVs. Nat Genet 42:385-391
61. Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P, Fitzgerald T, Hu M, Ihm CH, Kristiansson K, Macarthur DG, Macdonald JR, Onyiah I, Pang AW, Robson S, Stirrups K, Valsesia A, Walter K, Wei J, Tyler-Smith C, Carter NP, Lee C, Scherer SW, Hurles ME (2010) Origins and functional impact of copy number variation in the human genome. Nature 464:704-712
62. Cook EH Jr & Scherer SW (2008) Copy-number variations associated with neuropsychiatric conditions. Nature 455(7215):919-23
63. Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M (2009) Mapping complex disease traits with global gene expression. Nat Rev Genet 10(3):184-94
64. Cooper GM, Zerr T, Kidd JM, Eichler EE, Nickerson DA (2008) Systematic assessment of copy number variant detection via genome-wide SNP genotyping. Nat Genet 40:1199-1203

65. Corder EH, Saunders AM, Strittmatter WJ, Schmechel DE, Gaskell PC, Small GW, Roses AD, Haines JL, Pericak-Vance MA (1993) Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. Science 261(5123):921-3
66. Corvin A, Donohoe G, Nangle JM, Schwaiger S, Morris D, Gill M (2008) A dysbindin risk haplotype associated with less severe manic-type symptoms in psychosis. Neurosci Lett 431(2):146-9
67. Coyle JT & Tsai G (2004) NMDA receptor function, neuroplasticity, and the pathophysiology of schizophrenia. Int Rev Neurobiol 59:491-515
68. Coyle JT (1996) The glutamatergic dysfunction hypothesis for schizophrenia. Harv Rev Psychiatry 3(5):241-53
69. Coyle JT (2006) Glutamate and schizophrenia: beyond the dopamine hypothesis. Cell Mol Neurobiol 26(4-6):365-84
70. Craddock N & Owen MJ (2005) The beginning of the end for the Kraepelinian dichotomy. Br J Psychiatry 186:364-6
71. Craddock N & Owen MJ (2010) The Kraepelinian dichotomy - going, going... but still not gone. Br J Psychiatry 196(2):92-5
72. Craddock N, Hurles ME, Cardin N, Pearson RD, Plagnol V, Robson S, Vukcevic D, Barnes C, Conrad DF, Giannoulatou E, Holmes C, Marchini JL, Stirrups K, Tobin MD, Wain LV, Yau C, Aerts J, Ahmad T, Andrews TD, Arbury H, Attwood A, Auton A, Ball SG, Balmforth AJ, Barrett JC, Barroso I, Barton A, Bennett AJ, Bhaskar S, Blaszczyk K, Bowes J, Brand OJ, Braund PS, Bredin F, Breen G, Brown MJ, Bruce IN, Bull J, Burren OS, Burton J, Byrnes J, Caesar S, Clee CM, Coffey AJ, Connell JM, Cooper JD, Dominiczak AF, Downes K, Drummond HE, Dudakia D, Dunham A, Ebbs B, Eccles D, Edkins S, Edwards C, Elliot A, Emery P, Evans DM, Evans G, Eyre S, Farmer A, Ferrier IN, Feuk L, Fitzgerald T, Flynn E, Forbes A, Forty L, Franklyn JA, Freathy RM, Gibbs P, Gilbert P, Gokumen O, Gordon-Smith K, Gray E, Green E, Groves CJ, Grozeva D, Gwilliam R, Hall A, Hammond N, Hardy M, Harrison P, Hassanali N, Hebaishi H, Hines S, Hinks A, Hitman GA, Hocking L, Howard E, Howard P, Howson JM, Hughes D, Hunt S, Isaacs JD, Jain M, Jewell DP, Johnson T, Jolley JD, Jones IR, Jones LA, Kirov G, Langford CF, Lango-Allen H, Lathrop GM, Lee J, Lee KL, Lees C, Lewis K, Lindgren CM, Maisuria-Armer M, Maller J, Mansfield J, Martin P, Massey DC, McArdle WL, McGuffin P, McLay KE, Mentzer A, Mimmack ML, Morgan AE, Morris AP, Mowat C, Myers S, Newman W, Nimmo ER, O'Donovan MC, Onipinla A, Onyiah I, Ovington NR, Owen MJ, Palin K, Parnell K, Pernet D, Perry JR, Phillips A, Pinto D, Prescott NJ, Prokopenko I, Quail MA, Rafelt S, Rayner NW, Redon R, Reid DM, Renwick, Ring SM, Robertson N, Russell E, St Clair D, Sambrook JG, Sanderson JD, Schuilenburg H, Scott CE, Scott R, Seal S, Shaw-Hawkins S, Shields BM,

Simmonds MJ, Smyth DJ, Somaskantharajah E, Spanova K, Steer S, Stephens J, Stevens HE, Stone MA, Su Z, Symmons DP, Thompson JR, Thomson W, Travers ME, Turnbull C, Valsesia A, Walker M, Walker NM, Wallace C, Warren-Perry M, Watkins NA, Webster J, Weedon MN, Wilson AG, Woodburn M, Wordsworth BP, Young AH, Zeggini E, Carter NP, Frayling TM, Lee C, McVean G, Munroe PB, Palotie A, Sawcer SJ, Scherer SW, Strachan DP, Tyler-Smith C, Brown MA, Burton PR, Caulfield MJ, Compston A, Farrall M, Gough SC, Hall AS, Hattersley AT, Hill AV, Mathew CG, Pembrey M, Satsangi J, Stratton MR, Worthington J, Deloukas P, Duncanson A, Kwiatkowski DP, McCarthy MI, Ouwehand W, Parkes M, Rahman N, Todd JA, Samani NJ, Donnelly P (2010) Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature 464:713-720

73. Craig DW, Pearson JV, Szelinger S, Sekar A, Redman M, Corneveaux JJ, Pawlowski TL, Laub T, Nunn G, Stephan DA, Homer N, Huentelman MJ (2008) Identification of genetic variants using bar-coded multiplexed sequencing. Nat Methods 5:887-893
74. Crepel A, Breckpot J, Fryns JP, De la Marche W, Steyaert J, Devriendt K, Peeters H (2010) DISC1 duplication in two brothers with autism and mild mental retardation. Clin Genet 77(4):389-94
75. Cui H, Nishiguchi N, Yanagi M, Fukutake M, Mouri K, Kitamura N, Hashimoto T, Shirakawa O, Hishimoto A (2010) A putative *cis*-acting polymorphism in the *NOS1* gene is associated with schizophrenia and NOS1 immunoreactivity in the post mortem brain. Schizophr Res 121(1-3):172-8
76. Cummings E, Donohoe G, McDonald C, Dinan TG, O'Neill FA, O'Callaghan E, Waddington JL, Murphy KC, Gill M, Morris DW, Corvin A (2010) Clinical symptomatology and the psychosis risk gene *ZNF804A*. Schizophr Res 122(1-3):273-5
77. Cuny GD (2009) Kinase inhibitors as potential therapeutics for acute and chronic neurodegenerative conditions. Curr Pharm Des 15(34):3919-39
78. Davidson LL & Heinrichs RW (2003) Quantification of frontal and temporal lobe brain-imaging findings in schizophrenia: a meta-analysis. Psychiatry Res. 2003 Feb 15;122(2):69-87
79. Davies BM & Beech HR (1960) The effect of 1-arylcylohexylamine (sernyl) on twelve normal volunteers. J Ment Sci 106:912-924
80. Davis KL, Kahn RS, Ko G, Davidson M (1991) Dopamine in schizophrenia: a review and reconceptualization. Am J Psychiatry 148:1474-1486
81. de Smith AJ, Tsalenko A, Sampas N, Scheffer A, Yamada NA, Tsang P, Ben-Dor A, Yakhini Z, Ellis RJ, Bruhn L, Laderman S, Froguel P, Blakemore AI (2007) Array CGH analysis of copy number variation identifies 1284 new genes variant in healthy white males: implications for association studies of complex diseases. Hum Mol Genet 16:2783-2794

82. Dellinger AE, Saw SM, Goh LK, Seielstad M, Young TL, Li YJ (2010) Comparative analyses of seven algorithms for copy number variant identification from single nucleotide polymorphism arrays. Nucleic Acids Res 38(9):e105
83. Deloukas P, Schuler GD, Gyapay G, Beasley EM, Soderlund C, Rodriguez-Tomé P, Hui L, Matise TC, McKusick KB, Beckmann JS, Bentolila S, Bihoreau M, Birren BB, Browne J, Butler A, Castle AB, Chiannilkulchai N, Clee C, Day PJ, Dehejia A, Dibling T, Drouot N, Duprat S, Fizames C, Fox S, Gelling S, Green L, Harrison P, Hocking R, Holloway E, Hunt S, Keil S, Lijnzaad P, Louis-Dit-Sully C, Ma J, Mendis A, Miller J, Morissette J, Muselet D, Nusbaum HC, Peck A, Rozen S, Simon D, Slonim DK, Staples R, Stein LD, Stewart EA, Suchard MA, Thangarajah T, Vega-Czarny N, Webber C, Wu X, Hudson J, Auffray C, Nomura N, Sikela JM, Polymeropoulos MH, James MR, Lander ES, Hudson TJ, Myers RM, Cox DR, Weissenbach J, Boguski MS, Bentley DR (1998) A physical map of 30,000 human genes. Science 282(5389):744-6
84. Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol 4(5):P3
85. Dibbens LM, Mullen S, Helbig I, Mefford HC, Bayly MA, Bellows S, Leu C, Trucks H, Obermeier T, Wittig M, Franke A, Caglayan H, Yapici Z, Sander T, Eichler EE, Scheffer IE, Mulley JC, Berkovic SF (2009) Familial and sporadic 15q13.3 microdeletions in idiopathic generalized epilepsy: precedent for disorders with complex inheritance. Hum Mol Genet 18:3626-3631
86. Dixon AL, Liang L, Moffatt MF, Chen W, Heath S, Wong KC, Taylor J, Burnett E, Gut I, Farrall M, Lathrop GM, Abecasis GR, Cookson WO (2008) A genome-wide association study of global gene expression. Nat Genet 39(10):1202-7
87. Donis-Keller H, Green P, Helms C, Cartinhour S, Weiffenbach B, Stephens K, Keith TP, Bowden DW, Smith DR, Lander ES, Botstein D, Akots G, Rediker KS, Gravius T, Brown VA, Rising MB, Parker C, Powers JA, Watt DE, Kauffman ER, Bricker A, Phipps P, Muller-Kahle H, Fulton TR, Ng S, Schumm JW, Braman JC, Knowlton RG, Barker DF, Crooks SM, Lincoln SE, Daly MJ, Abrahamson J (1987). A genetic linkage map of the human genome. Cell 51(2):319-37
88. Donohoe G, Morris DW, De Sanctis P, Magno E, Montesi JL, Garavan HP, Robertson IH, Javitt DC, Gill M, Corvin AP, Foxe JJ (2008) Early visual processing deficits in dysbindin-associated schizophrenia. Biol Psychiatry 63(5):484-9
89. du Bois TM, Huang XF (2007) Early brain development disruption from NMDA receptor hypofunction: relevance to schizophrenia. Brain Res Rev 53:260-270
90. du Manoir S, Speicher MR, Joos S, Schröck E, Popp S, Döhner H, Kovacs G, Robert-Nicoud M, Lichter P, Cremer T (1993) Detection of complete and partial chromosome gains and losses by comparative genomic in situ hybridization. Hum Genet 90(6):590-610

91. Dwyer S, Williams H, Holmans P, Moskvina V, Craddock N, Owen MJ, O'Donovan MC (2010) No evidence that rare coding variants in *ZNF804A* confer risk of schizophrenia. Am J Med Genet B Neuropsychiatr Genet *in press*
92. Eastwood SL & Harrison PJ (1995) Decreased synaptophysin in the medial temporal lobe in schizophrenia demonstrated using immunoautoradiography. Neuroscience 69:339-343
93. Edwards TL, Scott WK, Almonte C, Burt A, Powell EH, Beecham GW, Wang L, Züchner S, Konidari I, Wang G, Singer C, Nahab F, Scott B, Stajich JM, Pericak-Vance M, Haines J, Vance JM, Martin ER (2010) Genome-wide association study confirms SNPs in *SNCA* and the *MAPT* region as common risk factors for Parkinson disease. Ann Hum Genet 74(2):97-109
94. Eeles RA, Kote-Jarai Z, Giles GG, Olama AA, Guy M, Jugurnauth SK, Mulholland S, Leongamornlert DA, Edwards SM, Morrison J, Field HI, Southey MC, Severi G, Donovan JL, Hamdy FC, Dearnaley DP, Muir KR, Smith C, Bagnato M, Ardern-Jones AT, Hall AL, O'Brien LT, Gehr-Swain BN, Wilkinson RA, Cox A, Lewis S, Brown PM, Jhavar SG, Tymrakiewicz M, Lophatananon A, Bryant SL; UK Genetic Prostate Cancer Study Collaborators; British Association of Urological Surgeons, Section of Oncology; UK ProtecT Study Collaborators, Horwich A, Huddart RA, Khoo VS, Parker CC, Woodhouse CJ, Thompson A, Christmas T, Ogden C, Fisher C, Jamieson C, Cooper CS, English DR, Hopper JL, Neal DE, Easton DF (2008) Multiple newly identified loci associated with prostate cancer susceptibility. Nat Genet 40(3):316-21
95. Egan CM, Sridhar S, Wigler M, Hall IM (2007) Recurrent DNA copy number variation in the laboratory mouse. Nat Genet 39(11):1384-9
96. Epstein DJ (2009) *Cis*-regulatory mutations in human disease. Brief Funct Genomic Proteomic 8(4):310-6
97. Esslinger C, Walter H, Kirsch P, Erk S, Schnell K, Arnold C, Haddad L, Mier D, Opitz von Boberfeld C, Raab K, Witt SH, Rietschel M, Cichon S, Meyer-Lindenberg A (2009) Neural mechanisms of a genome-wide supported psychosis variant. Science 324(5927):605
98. Evans T, Brown ML, Fraser ED, Northup JK (1986) Purification of the major GTP-binding proteins from human placental membranes. J Biol Chem 261(15):7052-9
99. Faa' V, Coiana A, Incani F, Costantino L, Cao A, Rosatelli MC (2010) A synonymous mutation in the CFTR gene causes aberrant splicing in an italian patient affected by a mild form of cystic fibrosis. J Mol Diagn 12(3):380-3
100. Fallgatter AJ, Herrmann MJ, Hohoff C, Ehlis AC, Jarczok TA, Freitag CM, Deckert J (2006) *DTNBP1* (dysbindin) gene variants modulate prefrontal brain function in healthy individuals. Neuropsychopharmacology 31(9):2002-10

101. Fan JB, Oliphant A, Shen R, Kermani BG, Garcia F, Gunderson KL, Hansen M, Steemers F, Butler SL, Deloukas P, Galver L, Hunt S, McBride C, Bibikova M, Rubano T, Chen J, Wickham E, Doucet D, Chang W, Campbell D, Zhang B, Kruglyak S, Bentley D, Haas J, Rigault P, Zhou L, Stuelpnagel J, Chee MS (2003) Highly parallel SNP genotyping. Cold Spring Harb Symp Quant Biol 68:69-78
102. Fanous AH & Kendler KS (2005) Genetic heterogeneity, modifier genes, and quantitative phenotypes in psychiatric illness: searching for a framework. Mol Psychiatry 10(1):6-1
103. Fanous AH, van den Oord EJ, Riley BP, Aggen SH, Neale MC, O'Neill FA, Walsh D, Kendler KS (2005) Relationship between a high-risk haplotype in the DTNBP1 (dysbindin) gene and clinical features of schizophrenia. Am J Psychiatry 162(10):1824-32
104. Faris RE and Dunham HW (1939) *Mental disorders in urban areas: An ecological study of schizophrenia and other psychoses*. The University of Chicago Press, Chicago/London
105. Farmer AE, McGuffin P, Gottesman II (1987) Twin concordance for DSM-III schizophrenia. Scrutinizing the validity of the definition. Arch Gen Psychiatry 44(7):634-41
106. Fedurco M, Romieu A, Williams S, Lawrence I, Turcatti G (2006) BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. Nucleic Acids Res 34:e22
107. Feinberg, I (1982) Schizophrenia: caused by a fault in programmed synaptic elimination during adolescence? J Psychiatr Res 17:319-334
108. Feinberg, I (1990) Cortical pruning and the development of schizophrenia. Schizophr Bull 16:567-570
109. Feng T & Zhu X (2010) Genome-wide searching of rare genetic variants in WTCCC data. Hum Genet 128(3):269-280
110. Ferrer-Costa C, Gelpí JL, Zamakola L, Parraga I, de la Cruz X, Orozco M (2005) PMUT: a web-based tool for the annotation of pathological mutations on proteins. Bioinformatics. 2005 Jul 15;21(14):3176-8
111. Fischer, M (1973) Genetic and environmental factors in schizophrenia. Acta Psychiatrica Scandinavia 238: 1-158
112. Flint J & Munafò MR (2007) The endophenotype concept in psychiatric genetics. Psychol Med 37(2):163-180
113. Fouser LA, Friesen JD (1986) Mutations in a yeast intron demonstrate the importance of specific conserved nucleotides for the two stages of nuclear mRNA splicing. Cell 45(1):81-93
114. Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, Lindgren CM, Perry JR, Elliott KS, Lango H, Rayner NW, Shields B, Harries LW, Barrett JC, Ellard S, Groves CJ, Knight B, Patch AM, Ness AR, Ebrahim S, Lawlor DA, Ring SM, Ben-Shlomo Y, Jarvelin MR, Sovio U, Bennett AJ, Melzer

D, Ferrucci L, Loos RJ, Barroso I, Wareham NJ, Karpe F, Owen KR, Cardon LR, Walker M, Hitman GA, Palmer CN, Doney AS, Morris AD, Smith GD, Hattersley AT, McCarthy MI (2007) A common variant in the *FTO* gene is associated with body mass index and predisposes to childhood and adult obesity. Science 316(5826):889-894

115. Freimer N, Sabatti C (2004) The use of pedigree, sib-pair and association studies of common diseases for genetic mapping and epidemiology. Nat Genet. 2004 Oct;36(10):1045-1051
116. Friedman LS, Ostermeyer EA, Szabo CI, Dowd P, Lynch ED, Rowell SE, King MC (1994) Confirmation of BRCA1 by analysis of germline mutations linked to breast and ovarian cancer in ten families. Nat Genet 8(4):399-404
117. Friston KJ, Frith CD, Fletcher P, Liddle PF, Frackowiak RS (1996) Functional topography: multidimensional scaling and functional connectivity in the brain. Cereb Cortex 6(2):156-64
118. Frith MC, Hansen U, Weng Z (2001) Detection of cis-element clusters in higher eukaryotic DNA. Bioinformatics 17(10):878-889
119. Fusar-Poli P, Perez J, Broome M, Borgwardt S, Placentino A, Caverzasi E, Cortesi M, Veggiotti P, Politi P, Barale F, McGuire P (2007) Neurofunctional correlates of vulnerability to psychosis: a systematic review and meta-analysis. Neurosci Biobehav Rev 31(4):465-84
120. Gefen A, Cohen R, Birk OS (2010) Syndrome to Gene (S2G): *in silico* identification of candidate genes for human diseases. Hum Mutat 31(3):229-236
121. Gill M, Donohoe G, Corvin A (2010) What have the genomics ever done for the psychoses? Psychol Med 40(4):529-40
122. Glantz LA & Lewis DA (2000) Decreased dendritic spine density on prefrontal cortical pyramidal neurons in schizophrenia. Arch Gen Psychiatry 57(1):65-73
123. Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, Fennell T, Giannoukos G, Fisher S, Russ C, Gabriel S, Jaffe DB, Lander ES, Nusbaum C (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol 27(2):182-189
124. Gold B, Kirchhoff T, Stefanov S, Lautenberger J, Viale A, Garber J, Friedman E, Narod S, Olshen AB, Gregersen P, Kosarin K, Olsh A, Bergeron J, Ellis NA, Klein RJ, Clark AG, Norton L, Dean M, Boyd J, Offit K (2008) Genome-wide association study provides evidence for a breast cancer risk locus at 6q22.33. Proc Natl Acad Sci USA 105(11):4340-5
125. Gottesman II & Bertelsen A (1989) Confirming unexpressed genotypes for schizophrenia. Risks in the offspring of Fischer's Danish identical and fraternal discordant twins. Arch Gen Psychiatry 46:867-872

126. Gottesman II & Gould TD (2003) The endophenotype concept in psychiatry: etymology and strategic intentions. Am J Psychiatry 160:636-645

127. Gottesman II & Shields J (1967) A polygenic theory of schizophrenia. Proc Natl Acad Sci USA 58(1):199-205

128. Gottesman II (1991) *Schizophrenia Genesis: The Origin of Madness*. New York: Freeman

129. Grant SF, Reid DM, Blake G, Herd R, Fogelman I, Ralston SH (1996) Reduced bone density and osteoporosis associated with a polymorphic Sp1 binding site in the collagen type I alpha 1 gene. Nat Genet 14(2):203-5

130. Gray NW, Fourgeaud L, Huang B, Chen J, Cao H, Oswald BJ, Hémar A, McNiven MA (2003) Dynamin 3 is a component of the postsynapse, where it interacts with mGluR5 and Homer. Curr Biol 13(6):510-5

131. Gruber JD & Long AD (2009) *Cis*-regulatory variation is typically polyallelic in *Drosophila*. Genetics 181(2):661-70

132. Guilmatre A, Dubourg C, Mosca AL, Legallic S, Goldenberg A, Drouin-Garraud V, Layet V, Rosier A, Briault S, Bonnet-Brilhault F, Laumonnier F, Odent S, Le Vacon G, Joly-Helas G, David V, Bendavid C, Pinoit JM, Henry C, Impallomeni C, Germano E, Tortorella G, Di Rosa G, Barthelemy C, Andres C, Faivre L, Frébourg T, Saugier Veber P, Campion D (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

133. Gusella JF, Tanzi RE, Bader PI, Phelan MC, Stevenson R, Hayden MR, Hofman KJ, Faryniarz AG, Gibbons K (1985) Deletion of Huntington's disease-linked G8 (D4S10) locus in Wolf-Hirschhorn syndrome. Nature 318:75-78

134. Hahn CG, Wang HY, Cho DS, Talbot K, Gur RE, Berrettini WH, Bakshi K, Kamins J, Borgmann-Winter KE, Siegel SJ, Gallop RJ, Arnold SE (2006) Altered neuregulin 1-erbB4 signaling contributes to NMDA receptor hypofunction in schizophrenia. Nat Med 12(7):824-8

135. Harrison PJ (1999) The neuropathology of schizophrenia. A critical review of the data and their interpretation. Brain 122 (Pt 4):593-624

136. Harrison PJ (2004) The hippocampus in schizophrenia: a review of the neuropathological evidence and its pathophysiological implications. Psychopharmacology (Berlin) 174(1):151-62

137. Harrison PJ, Law AJ, Eastwood SL (2004) Glutamate receptors and transporters in the hippocampus in schizophrenia. Ann N Y Acad Sci 1003:94-101

138. Harrison PJ, Weinberger DR (2005) Schizophrenia genes, gene expression, and neuropathology: on the matter of their convergence. Mol Psychiatry 10:40-68; image 45

139. Hawi Z, Gibson S, Straub RE, Walsh D, Kendler KS, Gill M (1999) Schizophrenia and HLA: No association with PCR-SSOP typed classical loci in a large Irish familial sample. Am J Med Genet 88(4):422-9

140. Hayashi-Takagi A, Takaki M, Graziane N, Seshadri S, Murdoch H, Dunlop AJ, Makino Y, Seshadri AJ, Ishizuka K, Srivastava DP, Xie Z, Baraban JM, Houslay MD, Tomoda T, Brandon NJ, Kamiya A, Yan Z, Penzes P, Sawa A (2010) Disrupted-in-Schizophrenia 1 (DISC1) regulates spines of the glutamate synapse via Rac1. Nat Neurosci 13(3):327-32

141. Heasman SJ & Ridley AJ (2008) Mammalian Rho GTPases: new insights into their functions from *in vivo* studies. Nat Rev Mol Cell Biol 9(9):690-701

142. Heinemeyer T, Wingender E, Reuter I, Hermjakob H, Kel AE, Kel OV, Ignatieva EV, Ananko EA, Podkolodnaya OA, Kolpakov FA, Podkolodny NL, Kolchanov NA (1998) Databases on transcriptional regulation: TRANSFAC, TRRD and COMPEL. Nucleic Acids Res 26:362-367

143. Helbig I, Mefford HC, Sharp AJ, Guipponi M, Fichera M, Franke A, Muhle H, de Kovel C, Baker C, von Spiczak S, Kron KL, Steinich I, Kleefuss-Lie AA, Leu C, Gaus V, Schmitz B, Klein KM, Reif PS, Rosenow F, Weber Y, Lerche H, Zimprich F, Urak L, Fuchs K, Feucht M, Genton P, Thomas P, Visscher F, de Haan GJ, Moller RS, Hjalgrim H, Luciano D, Wittig M, Nothnagel M, Elger CE, Nurnberg P, Romano C, Malafosse A, Koeleman BP, Lindhout D, Stephani U, Schreiber S, Eichler EE, Sander T (2009) 15q13.3 microdeletions increase risk of idiopathic generalized epilepsy. Nat Genet 41:160-162

144. Hellman LM, Fried MG (2007) Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions. Nat Protoc 2:1849-1861

145. Heston LL (1966) Psychiatric disorders in foster home reared children of schizophrenic mothers. Br J Psych 112: 819-825

146. Hill JJ, Hashimoto T, Lewis DA (2006) Molecular mechanisms contributing to dendritic spine alterations in the prefrontal cortex of subjects with schizophrenia. Mol Psychiatry 11:557-566

147. Hill SK, Beers SR, Kmiec JA, Keshavan MS, Sweeney JA (2004) Impairment of verbal memory and learning in antipsychotic-naïve patients with first-episode schizophrenia. Schizophr Res 68(2-3):127-36

148. Hinds DA, Kloek AP, Jen M, Chen X, Frazer KA (2006) Common deletions and SNPs are in linkage disequilibrium in the human genome. Nat Genet 38:82-85

149. Hodgkinson CA, Enoch MA, Srivastava V, Cummins-Oman JS, Ferrier C, Iarikova P, Sankararaman S, Yamini G, Yuan Q, Zhou Z, Albaugh B, White KV, Shen PH, Goldman D. Genome-wide association identifies candidate genes that influence the human electroencephalogram. Proc Natl Acad Sci U S A. 2010 May 11;107(19):8695-700

150. Honea R, Crow TJ, Passingham D, Mackay CE (2005) Regional deficits in brain volume in schizophrenia: a meta-analysis of voxel-based morphometry studies. Am J Psych 162(12):2233-45

151. Huang da W, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4(1):44-57

152. Hurd PJ & Nelson CJ (2009) Advantages of next-generation sequencing versus the microarray in epigenetic research. Brief Funct Genomic Proteomic 8(3):174-83

153. Ide M, Lewis DA (2010) Altered cortical CDC42 signaling pathways in schizophrenia: implications for dendritic spine deficits. Biol Psychiatry 68:25-32

154. Inaba Y, Tian QB, Okano A, Zhang JP, Sakagami H, Miyazawa S, Li W, Komiyama A, Inokuchi K, Kondo H, Suzuki T (2004) Brain-specific potential guanine nucleotide exchange factor for Arf, synArfGEF (Po), is localized to postsynaptic density. J Neurochem 89(6):1347-57

155. International HapMap Consortium (2003) The International HapMap Project. Nature 426(6968):789-96

156. International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437(7063):1299-320

157. International HapMap Consortium, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, Hu W, Li C, Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q, Zhao H, Zhao H, Zhou J, Gabriel SB, Barry R, Blumenstiel B, Camargo A, Defelice M, Faggart M, Goyette M, Gupta S, Moore J, Nguyen H, Onofrio RC, Parkin M, Roy J, Stahl E, Winchester E, Ziaugra L, Altshuler D, Shen Y, Yao Z, Huang W, Chu X, He Y, Jin L, Liu Y, Shen Y, Sun W, Wang H, Wang Y, Wang Y, Xiong X, Xu L, Waye MM, Tsui SK, Xue H, Wong JT, Galver LM, Fan JB, Gunderson K, Murray SS, Oliphant AR, Chee MS, Montpetit A, Chagnon F, Ferretti V, Leboeuf M, Olivier JF, Phillips MS, Roumy S, Sallée C, Verner A, Hudson TJ, Kwok PY, Cai D, Koboldt DC, Miller RD, Pawlikowska L, Taillon-Miller P, Xiao M, Tsui LC, Mak W, Song YQ, Tam PK, Nakamura Y, Kawaguchi T, Kitamoto T, Morizono T, Nagashima A, Ohnishi Y, Sekine A, Tanaka T, Tsunoda T, Deloukas P, Bird CP, Delgado M, Dermitzakis ET, Gwilliam R, Hunt S, Morrison J, Powell D, Stranger BE, Whittaker P, Bentley DR, Daly MJ, de Bakker PI, Barrett J, Chretien YR, Maller J, McCarroll S, Patterson N, Pe'er I, Price A, Purcell S, Richter DJ, Sabeti P, Saxena R, Schaffner SF, Sham PC, Varilly P, Altshuler D, Stein LD, Krishnan L, Smith AV, Tello-Ruiz MK, Thorisson GA, Chakravarti A, Chen PE, Cutler DJ, Kashuk CS, Lin S, Abecasis GR, Guan W, Li Y, Munro HM, Qin ZS, Thomas DJ, McVean G, Auton A, Bottolo L, Cardin N, Eyheramendy S, Freeman C, Marchini J, Myers S, Spencer C, Stephens M, Donnelly P, Cardon LR, Clarke G, Evans DM, Morris AP, Weir BS,

Tsunoda T, Mullikin JC, Sherry ST, Feolo M, Skol A, Zhang H, Zeng C, Zhao H, Matsuda I, Fukushima Y, Macer DR, Suda E, Rotimi CN, Adebamowo CA, Ajayi I, Aniagwu T, Marshall PA, Nkwodimmah C, Royal CD, Leppert MF, Dixon M, Peiffer A, Qiu R, Kent A, Kato K, Niikawa N, Adewole IF, Knoppers BM, Foster MW, Clayton EW, Watkin J, Gibbs RA, Belmont JW, Muzny D, Nazareth L, Sodergren E, Weinstock GM, Wheeler DA, Yakub I, Gabriel SB, Onofrio RC, Richter DJ, Ziaugra L, Birren BW, Daly MJ, Altshuler D, Wilson RK, Fulton LL, Rogers J, Burton J, Carter NP, Clee CM, Griffiths M, Jones MC, McLay K, Plumb RW, Ross MT, Sims SK, Willey DL, Chen Z, Han H, Kang L, Godbout M, Wallenburg JC, L'Archevêque P, Bellemare G, Saeki K, Wang H, An D, Fu H, Li Q, Wang Z, Wang R, Holden AL, Brooks LD, McEwen JE, Guyer MS, Wang VO, Peterson JL, Shi M, Spiegel J, Sung LM, Zacharia LF, Collins FS, Kennedy K, Jamieson R, Stewart J (2008) A second generation human haplotype map of over 3.1 million SNPs. Nature 449(7164):851-61

158. International Schizophrenia Consortium (2008) Rare chromosomal deletions and duplications increase risk of schizophrenia. Nature 455(7210):237-41
159. International Schizophrenia Consortium, Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, Sklar P (2009) Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 460(7256):748-52
160. Ishibashi T, Ogawa S, Hashiguchi Y, Inoue Y, Udo H, Ohzono H, Kato A, Minakami R, Sugiyama H (2005) A novel protein specifically interacting with Homer2 regulates ubiquitin-proteasome systems. J Biochem 137(5):617-23
161. Jablensky A (2000) Epidemiology of schizophrenia: the global burden of disease and disability. Eur Arch Psychiatry Clin Neurosci 250(6):274-85
162. Jablensky A, Hugler H, Von Cranach M, Kalinov K (1993) Kraepelin revisited: a reassessment and statistical analysis of dementia praecox and manic-depressive insanity in 1908. Psychol Med 23(4):843-58
163. Jain S & Heutink (2010) From single genes to gene networks: High-throughput-high-content screening for neurological disease. Neuron 68: 207-217
164. Javitt DC & Zukin SR (1991) Recent advances in the phencyclidine model of schizophrenia. Am J Psychiatry. 148:1301–308
165. Jenuth JP (2000) The NCBI: publicly available tools and resources on the Web. Methods Mol Biol 132:301-12
166. Johnson DI & Pringle JR (1990) Molecular characterization of CDC42, a Saccharomyces cerevisiae gene involved in the development of cell polarity. J Cell Biol 111(1):143-52
167. Jönsson E, Lannfelt L, Sokoloff P, Schwartz JC, Sedvall G (1993) Lack of association between schizophrenia and alleles in the dopamine D3 receptor gene. Acta Psychiatr Scand 87:345-349

168. Jönsson EG, Zhang F, Nimgaonkar VL, Rudert WA, Sedvall GC (1998) Lack of association between schizophrenia and HLA DQB1 alleles in a Swedish sample. Schizophr Res 29:293-296

169. Johnstone EC, Crow TJ, Frith CD, Husband J, Kreel L (1976) Cerebral ventricular size and cognitive impairment in chronic schizophrenia. Lancet 2(7992):924-6.

170. Kallioniemi A, Kallioniemi OP, Piper J, Tanner M, Stokke T, Chen L, Smith HS, Pinkel D, Gray JW, Waldman FM (1994) Detection and mapping of amplified DNA sequences in breast cancer by comparative genomic hybridization. Proc Natl Acad Sci USA 91(6):2156-60

171. Kang R, Wan J, Arstikaitis P, Takahashi H, Huang K, Bailey AO, Thompson JX, Roth AF, Drisdel RC, Mastro R, Green WN, Yates JR, 3rd, Davis NG, El-Husseini A (2008) Neural palmitoyl-proteomics reveals dynamic synaptic palmitoylation. Nature 456:904-909

172. Karayiorgou M, Morris MA, Morrow B, Shprintzen RJ, Goldberg R, Borrow J, Gos A, Nestadt G, Wolyniec PS, Lasseter VK, et al. (1995) Schizophrenia susceptibility associated with interstitial deletions of chromosome 22q11. Proc Natl Acad Sci USA 92(17):7612-6

173. Kato A, Ozawa F, Saitoh Y, Fukazawa Y, Sugiyama H, Inokuchi K (1998) Novel members of the Vesl/Homer family of PDZ proteins that bind metabotropic glutamate receptors. J Biol Chem 273(37):23969-75

174. Kaymaz N & van Os J (2009) Heritability of structural brain traits an endophenotype approach to deconstruct schizophrenia. Int Rev Neurobiol. 2009;89:85-130

175. Kendler KS & Diehl SR (1993) The genetics of schizophrenia: a current, genetic-epidemiologic perspective. Schizophr Bull 19(2):261-85

176. Kendler KS (1983) Overview: a current perspective on twin studies of schizophrenia. Am J Psychiatry 140:1413-1425

177. Kendler KS, Gruenberg AM, Kinney DK (1994) Independent diagnoses of adoptees and relatives as defined by DSM-III in the provincial and national samples of the Danish Adoption Study of Schizophrenia. Arch General Psych 51(6): 456-68

178. Kendler KS, McGuire M, Gruenberg AM, O'Hare A, Spellman M, Walsh D (1993) The Roscommon Family Study. I. Methods, diagnosis of probands, and risk of schizophrenia in relatives. Arch Gen Psychiatry 50:527-540

179. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D (2002) The human genome browser at UCSC. Genome Res 12(6):996-1006

180. Kerber RA, O'Brien E, Cawthon RM (2009) Gene expression profiles associated with aging and mortality in humans. Aging Cell 8:239-250

181. Keshavan MS, Anderson S, Pettegrew JW (1994) Is schizophrenia due to excessive synaptic pruning in the prefrontal cortex? The Feinberg hypothesis revisited. J Psychiatr Res 28(3):239-65

182. Ketty SS (1988) Schizophrenic illness in the families of schizophrenic adoptees: findings from the Danish national sample. Schizophr Bull 14(2): 217-22

183. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N, Teague B, Alkan C, Antonacci F, Haugen E, Zerr T, Yamada NA, Tsang P, Newman TL, Tüzün E, Cheng Z, Ebling HM, Tusneem N, David R, Gillett W, Phelps KA, Weaver M, Saranga D, Brand A, Tao W, Gustafson E, McKernan K, Chen L, Malig M, Smith JD, Korn JM, McCarroll SA, Altshuler DA, Peiffer DA, Dorschner M, Stamatoyannopoulos J, Schwartz D, Nickerson DA, Mullikin JC, Wilson RK, Bruhn L, Olson MV, Kaul R, Smith DR, Eichler EE (2008) Mapping and sequencing of structural variation from eight human genomes. Nature 453(7191):56-64

184. Kim N, Lim D, Lee S, Kim H (2005) ASePCR: alternative splicing electronic RT-PCR in multiple tissues and organs. Nucleic Acids Res 33(Web Server issue):W681-5

185. Kim S, Webster MJ (2010) The Stanley neuropathology consortium integrative database: a novel, web-based tool for exploring neuropathological markers in psychiatric disorders and the biological processes associated with abnormalities of those markers. Neuropsychopharmacology 35(2):473-82

186. Kimble GA (1993) Evolution of the nature-nurture issue in the history of psychology. In: Plomin R & McClearn GE, Nature Nurture and Psychology. American Psychological Association Press, Washington DC

187. Kirov G, Grozeva D, Norton N, Ivanov D, Mantripragada KK, Holmans P, The International Schizophrenia Consortium, The Wellcome Trust Case Control Consortium, Craddock N, Owen MJ, O'Donovan MC (2009) Support for the involvement of large copy number variants in the pathogenesis of schizophrenia. Hum Mol Genet 18(8):1497-503

188. Kitada T, Asakawa S, Hattori N, Matsumine H, Yamamura Y, Minoshima S, Yokochi M, Mizuno Y, Shimizu N (1998) Mutations in the parkin gene cause autosomal recessive juvenile parkinsonism. Nature 392(6676):605-8

189. Kleinjan DJ & Coutinho P (2009) *Cis*-ruption mechanisms: disruption of cis-regulatory control as a cause of human genetic disease. Brief Funct Genomic Proteomic 8(4):317-32

190. Knight HM, Pickard BS, Maclean A, Malloy MP, Soares DC, McRae AF, Condie A, White A, Hawkins W, McGhee K, van Beck M, MacIntyre DJ, Starr JM, Deary IJ, Visscher PM, Porteous DJ, Cannon RE, St Clair D, Muir WJ, Blackwood DH (2009) A cytogenetic abnormality and rare coding variants identify ABCA13 as a candidate gene in schizophrenia, bipolar disorder, and depression. Am J Hum Genet 85(6):833-46

191. Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L, Taillon BE, Chen Z, Tanzer A, Saunders AC, Chi J, Yang F, Carter NP, Hurles ME, Weissman

SM, Harkins TT, Gerstein MB, Egholm M, Snyder M (2007) Paired-end mapping reveals extensive structural variation in the human genome. Science 318:420-426

192. Kræpelin E (1919). Dementia praecox and paraphrenia (translated by R M Barclay). Facsimile edition, Kreiger, New York, 1971
193. Kruglyak L & Lander ES (1995) High-resolution genetic mapping of complex traits. Am J Hum Genet 56(5):1212-23
194. Kullmann DM, Lamsa KP (2007) Long-term synaptic plasticity in hippocampal interneurons. Nat Rev Neurosci 8:687-699
195. Kwan T, Benovoy D, Dias C, Gurd S, Provencher C, Beaulieu P, Hudson TJ, Sladek R, Majewski J (2008) Genome-wide analysis of transcript isoform variation in humans. Nat Genet 40(2):225-31
196. Kwan T, Benovoy D, Dias C, Gurd S, Serre D, Zuzan H, Clark TA, Schweitzer A, Staples MK, Wang H, Blume JE, Hudson TJ, Sladek R, Majewski J (2007) Heritability of alternative splicing in the human genome. Genome Res 17(8):1210-8
197. Lander ES & Kruglyak L (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat Genet 11(3):241-7
198. Lander ES (1996) The new genomics: global views of biology. Science 274(5287):536-9
199. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA,

Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blöcker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ; International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. Nature 409(6822):860-921

200. Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, Willer CJ, Jackson AU, Vedantam S, Raychaudhuri S, Ferreira T, Wood AR, Weyant RJ, Segrè AV, Speliotes EK, Wheeler E, Soranzo N, Park JH, Yang J, Gudbjartsson D, Heard-Costa NL, Randall JC, Qi L, Vernon Smith A, Mägi R, Pastinen T, Liang L, Heid IM, Luan J, Thorleifsson G, Winkler TW, Goddard ME, Sin Lo K, Palmer C, Workalemahu T, Aulchenko YS, Johansson A, Carola Zillikens M, Feitosa MF, Esko T, Johnson T, Ketkar S, Kraft P, Mangino M, Prokopenko I, Absher D, Albrecht E, Ernst F, Glazer NL, Hayward C, Hottenga JJ, Jacobs KB, Knowles JW, Kutalik Z, Monda KL, Polasek O, Preuss M, Rayner NW, Robertson NR, Steinthorsdottir V, Tyrer JP, Voight BF, Wiklund F, Xu J, Hua Zhao J, Nyholt DR, Pellikka N, Perola M, Perry JR, Surakka I, Tammesoo ML, Altmaier EL, Amin N, Aspelund T, Bhangale T, Boucher G, Chasman DI, Chen C, Coin L, Cooper MN, Dixon AL, Gibson Q, Grundberg E, Hao K, Juhani Junttila M, Kaplan LM, Kettunen J, König IR, Kwan T, Lawrence RW, Levinson DF, Lorentzon M, McKnight B, Morris AP, Müller M, Suh Ngwa J, Purcell S, Rafelt S, Salem RM, Salvi E, Sanna S, Shi J, Sovio U, Thompson JR, Turchin MC, Vandenput L, Verlaan DJ, Vitart V, White CC, Ziegler A, Almgren P, Balmforth AJ, Campbell H, Citterio L, De Grandi A, Dominiczak A, Duan J, Elliott P, Elosua R, Eriksson JG, Freimer NB, Geus EJ, Glorioso N, Haiqing S, Hartikainen AL, Havulinna AS, Hicks AA, Hui J, Igl W, Illig T, Jula A, Kajantie E, Kilpeläinen TO, Koiranen M, Kolcic I, Koskinen S, Kovacs P, Laitinen J, Liu J, Lokki ML, Marusic A, Maschio A, Meitinger T, Mulas A, Paré G, Parker AN, Peden JF, Petersmann A, Pichler I, Pietiläinen KH, Pouta A, Ridderstråle M, Rotter JI, Sambrook JG, Sanders AR, Oliver Schmidt C, Sinisalo J, Smit JH, Stringham HM, Bragi Walters G, Widen E, Wild SH, Willemsen G, Zagato L, Zgaga L, Zitting P,

Alavere H, Farrall M, McArdle WL, Nelis M, Peters MJ, Ripatti S, van Meurs JB, Aben KK, Ardlie KG, Beckmann JS, Beilby JP, Bergman RN, Bergmann S, Collins FS, Cusi D, den Heijer M, Eiriksdottir G, Gejman PV, Hall AS, Hamsten A, Huikuri HV, Iribarren C, Kähönen M, Kaprio J, Kathiresan S, Kiemeney L, Kocher T, Launer LJ, Lehtimäki T, Melander O, Mosley TH Jr, Musk AW, Nieminen MS, O'Donnell CJ, Ohlsson C, Oostra B, Palmer LJ, Raitakari O, Ridker PM, Rioux JD, Rissanen A, Rivolta C, Schunkert H, Shuldiner AR, Siscovick DS, Stumvoll M, Tönjes A, Tuomilehto J, van Ommen GJ, Viikari J, Heath AC, Martin NG, Montgomery GW, Province MA, Kayser M, Arnold AM, Atwood LD, Boerwinkle E, Chanock SJ, Deloukas P, Gieger C, Grönberg H, Hall P, Hattersley AT, Hengstenberg C, Hoffman W, Mark Lathrop G, Salomaa V, Schreiber S, Uda M, Waterworth D, Wright AF, Assimes TL, Barroso I, Hofman A, Mohlke KL, Boomsma DI, Caulfield MJ, Adrienne Cupples L, Erdmann J, Fox CS, Gudnason V, Gyllensten U, Harris TB, Hayes RB, Jarvelin MR, Mooser V, Munroe PB, Ouwehand WH, Penninx BW, Pramstaller PP, Quertermous T, Rudan I, Samani NJ, Spector TD, Völzke H, Watkins H, Wilson JF, Groop LC, Haritunians T, Hu FB, Kaplan RC, Metspalu A, North KE, Schlessinger D, Wareham NJ, Hunter DJ, O'Connell JR, Strachan DP, Wichmann HE, Borecki IB, van Duijn CM, Schadt EE, Thorsteinsdottir U, Peltonen L, Uitterlinden AG, Visscher PM, Chatterjee N, Loos RJ, Boehnke M, McCarthy MI, Ingelsson E, Lindgren CM, Abecasis GR, Stefansson K, Frayling TM, Hirschhorn JN. (2010) Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature 467(7317):832-8

201. Laruelle M, Abi-Dargham A, van Dyck CH, Gil R, D'Souza CD, Erdos J, McCance E, Rosenblatt W, Fingado C, Zoghbi SS, Baldwin RM, Seibyl JP, Krystal JH, Charney DS, Innis RB (1996) Single photon emission computerized tomography imaging of amphetamine-induced dopamine release in drug-free schizophrenic subjects. Proc Natl Acad Sci USA 93(17):9235-40
202. Laruelle M, Kegeles LS, Abi-Dargham A (2003) Glutamate, dopamine, and schizophrenia: from pathophysiology to treatment. Ann N Y Acad Sci 1003:138-158
203. Law AJ, Kleinman JE, Weinberger DR, Weickert CS (2007) Disease-associated intronic variants in the ErbB4 gene are related to altered ErbB4 splice-variant expression in the brain in schizophrenia. Hum Mol Genet 16(2):129-41
204. Lee KH, Williams LM, Breakspear M, Gordon E (2003) Synchronous gamma activity: a review and contribution to an integrative neuroscience model of schizophrenia. Brain Res Brain Res Rev 41(1):57-7
205. Levinson DF, Levinson MD, Segurado R, Lewis CM (2003) Genome scan meta-analysis of schizophrenia and bipolar disorder, part I: Methods and power analysis. Am J Hum Genet 73(1):17-33

206. Lewis CM, Levinson DF, Wise LH, DeLisi LE, Straub RE, Hovatta I, Williams NM, Schwab SG, Pulver AE, Faraone SV, Brzustowicz LM, Kaufmann CA, Garver DL, Gurling HM, Lindholm E, Coon H, Moises HW, Byerley W, Shaw SH, Mesen A, Sherrington R, O'Neill FA, Walsh D, Kendler KS, Ekelund J, Paunio T, Lönnqvist J, Peltonen L, O'Donovan MC, Owen MJ, Wildenauer DB, Maier W, Nestadt G, Blouin JL, Antonarakis SE, Mowry BJ, Silverman JM, Crowe RR, Cloninger CR, Tsuang MT, Malaspina D, Harkavy-Friedman JM, Svrakic DM, Bassett AS, Holcomb J, Kalsi G, McQuillin A, Brynjolfson J, Sigmundsson T, Petursson H, Jazin E, Zoëga T, Helgason T. Genome scan meta-analysis of schizophrenia and bipolar disorder, part II: Schizophrenia. Am J Hum Genet 73(1):34-48

207. Lewis DA & Hashimoto T (2007) Deciphering the disease process of schizophrenia: the contribution of cortical GABA neurons. Int Rev Neurobiol 78:109-31

208. Lewis DA & Levitt P (2002) Schizophrenia as a disorder of neurodevelopment. Ann Rev Neurosci 25: 409-432

209. Lewis DA & Moghaddam B (2006) Cognitive dysfunction in schizophrenia: convergence of gamma-aminobutyric acid and glutamate alterations. Arch Neurol 63(10):1372-6

210. Lewis DA, & González-Burgos G (2008) Neuroplasticity of neocortical circuits in schizophrenia. Neuropsychopharmacology 33(1):141-65

211. Lewis DA, Glantz LA, Pierri JN, Sweet RA (2003) Altered cortical glutamate neurotransmission in schizophrenia: evidence from morphological studies of pyramidal neurons. Ann N Y Acad Sci 1003:102–112

212. Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 18:1851-1858

213. Li R, Li Y, Fang X, Yang H, Wang J, Kristiansen K (2009) SNP detection for massively parallel whole-genome resequencing. Genome Res 19:1124-1132

214. Li Y, Vinckenbosch N, Tian G, Huerta-Sanchez E, Jiang T, Jiang H, Albrechtsen A, Andersen G, Cao H, Korneliussen T, Grarup N, Guo Y, Hellman I, Jin X, Li Q, Liu J, Liu X, Sparsø T, Tang M, Wu H, Wu R, Yu C, Zheng H, Astrup A, Bolund L, Holmkvist J, Jørgensen T, Kristiansen K, Schmitz O, Schwartz TW, Zhang X, Li R, Yang H, Wang J, Hansen T, Pedersen O, Nielsen R, Wang J (2010) Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. Nat Genet *in press*

215. Locke DP, Sharp AJ, McCarroll SA, McGrath SD, Newman TL, Cheng Z, Schwartz S, Albertson DG, Pinkel D, Altshuler DM, Eichler EE (2006) Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. Am J Hum Genet 79:275-290

216. Loewi, O (1921) Über humorale Übertragbarkeit der Herznervenwirkung. I. Pflügers Archiv 189, pp. 239-242

217. Loranger AW (1984) Sex difference in age at onset of schizophrenia. Arch Gen Psychiatry 41(2):157-61

218. Luby ED, Cohen BD, Rosenbaum G, Gottlieb JS, Kelley R (1959) Study of a new schizophrenomimetic drug; sernyl. AMA Arch Neurol Psychiatry 81:363-369

219. Lucito R, Healy J, Alexander J, Reiner A, Esposito D, Chi M, Rodgers L, Brady A, Sebat J, Troge J, West JA, Rostan S, Nguyen KC, Powers S, Ye KQ, Olshen A, Venkatraman E, Norton L, Wigler M (2003) Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation. Genome Res 13:2291-2305

220. Luo L, Hensch TK, Ackerman L, Barbel S, Jan LY, Nung Jan Y (1996) Differential effects of the Rac GTPase on Purkinje cell axons and dendritic trunks and spines. Nature 379:837-840

221. Luo Z & Geschwind DH (2001) Microarray applications in neuroscience. Neurobiol Dis 8(2):183-93

222. Lupski JR (1998) Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. Trends Genet 14:417-422

223. Lupski JR, de Oca-Luna RM, Slaugenhaupt S, Pentao L, Guzzetta V, Trask BJ, Saucedo-Cardenas O, Barker DF, Killian JM, Garcia CA, Chakravarti A, Patel PI (1991) DNA duplication associated with Charcot-Marie-Tooth disease type 1A. Cell 66(2):219-32

224. MacDonald ME, Ambrose CM, Duyao MP, Myers RH, Lin C, Srinidhi L, Barnes G, Taylor SA, James M, Groot N, MacFarlane H, Jenkins B, Anderson MA, Wexler NS, Gusella JF, Bates GP, Baxendale S, Hummerich H, Kirby S, North M, Youngman S, Mott R, Zehetner G, Sedlacek Z, Poustka A, Frischauf A-M, Lehrach H, Buckler AJ, Church D, Doucette-Stamm L, O'Donovan MC, Riba-Ramirez L, Shah M, Stanton VP, Strobel SA, Draths KM, Wales JL, Dervan P, Housman DE, Altherr M, Shiang R, Thompson L, Fielder T, Wasmuth JJ, Tagle D, Valdes J, Elmer L, Allard M, Castilla L, Swaroop M, Blanchard K, Collins FS, Snell R, Holloway T, Gillespie K, Datson N, Shaw D, Harper PS (1992) A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. Cell 72(6):971-83

225. Maglott D, Ostell J, Pruitt KD, Tatusova T (2005) Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res 33(Database issue):D54-8

226. Mah S, Nelson MR, Delisi LE, Reneland RH, Markward N, James MR, Nyholt DR, Hayward N, Handoko H, Mowry B, Kammerer S, Braun A (2006) Identification of the semaphorin receptor PLXNA2 as a candidate for susceptibility to schizophrenia. Mol Psychiatry 11(5):471-8

227. Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, Howard E, Shendure J, Turner DJ (2010) Target-enrichment strategies for next-generation sequencing. Nat Methods 7:111-118
228. Maniatis, N (2007) Linkage disequilibrium maps and disease-association mapping. Methods Mol Biol 376:109-21
229. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM. (2009) Finding the missing heritability of complex diseases. Nature 461(7265):747-53
230. Mardis ER (2008) Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet 9:387-402
231. Marks PW & Kwiatkowski DJ (1996) Genomic Organization and Chromosomal Location of MurineCdc42. Genomics 38:13-18
232. Maston GA, Evans SK, Green MR (2006) Transcriptional regulatory elements in the human genome. Annu Rev Genomics Hum Genet 7:29-59
233. McCarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J, Wysoker A, Shapero MH, de Bakker PI, Maller JB, Kirby A, Elliott AL, Parkin M, Hubbell E, Webster T, Mei R, Veitch J, Collins PJ, Handsaker R, Lincoln S, Nizzari M, Blume J, Jones KW, Rava R, Daly MJ, Gabriel SB, Altshuler D (2008) Integrated detection and population-genetic analysis of SNPs and copy number variation. Nat Genet 40(10):1166-74
234. McCarthy SE, Makarov V, Kirov G, Addington AM, McClellan J, Yoon S, Perkins DO, Dickel DE, Kusenda M, Krastoshevsky O, Krause V, Kumar RA, Grozeva D, Malhotra D, Walsh T, Zackai EH, Kaplan P, Ganesh J, Krantz ID, Spinner NB, Roccanova P, Bhandari A, Pavon K, Lakshmi B, Leotta A, Kendall J, Lee YH, Vacic V, Gary S, Iakoucheva LM, Crow TJ, Christian SL, Lieberman JA, Stroup TS, Lehtimaki T, Puura K, Haldeman-Englert C, Pearl J, Goodell M, Willour VL, Derosse P, Steele J, Kassem L, Wolff J, Chitkara N, McMahon FJ, Malhotra AK, Potash JB, Schulze TG, Nothen MM, Cichon S, Rietschel M, Leibenluft E, Kustanovich V, Lajonchere CM, Sutcliffe JS, Skuse D, Gill M, Gallagher L, Mendell NR, Craddock N, Owen MJ, O'Donovan MC, Shaikh TH, Susser E, Delisi LE, Sullivan PF, Deutsch CK, Rapoport J, Levy DL, King MC, Sebat J (2009) Microduplications of 16p11.2 are associated with schizophrenia. Nat Genet 41:1223-1227
235. McClellan JM, Susser E, King MC (2007) Schizophrenia: a common disease caused by multiple rare alleles. Br J Psychiatry 190:194-9
236. McEntee WJ, Crook TH (1993) Glutamate: its role in learning, memory, and the aging brain. Psychopharmacology (Berl) 111:391-401

237. McGue M, Gottesman II, Rao DC (1983) The transmission of schizophrenia under a multifactorial threshold model. Am J Hum Genet 35(6):1161-78

238. Mefford HC, Sharp AJ, Baker C, Itsara A, Jiang Z, Buysse K, Huang S, Maloney VK, Crolla JA, Baralle D, Collins A, Mercer C, Norga K, de Ravel T, Devriendt K, Bongers EM, de Leeuw N, Reardon W, Gimelli S, Bena F, Hennekam RC, Male A, Gaunt L, Clayton-Smith J, Simonic I, Park SM, Mehta SG, Nik-Zainal S, Woods CG, Firth HV, Parkin G, Fichera M, Reitano S, Lo Giudice M, Li KE, Casuga I, Broomer A, Conrad B, Schwerzmann M, Raber L, Gallati S, Striano P, Coppola A, Tolmie JL, Tobias ES, Lilley C, Armengol L, Spysschaert Y, Verloo P, De Coene A, Goossens L, Mortier G, Speleman F, van Binsbergen E, Nelen MR, Hochstenbach R, Poot M, Gallagher L, Gill M, McClellan J, King MC, Regan R, Skinner C, Stevenson RE, Antonarakis SE, Chen C, Estivill X, Menten B, Gimelli G, Gribble S, Schwartz S, Sutcliffe JS, Walsh T, Knight SJ, Sebat J, Romano C, Schwartz CE, Veltman JA, de Vries BB, Vermeesch JR, Barber JC, Willatt L, Tassabehji M, Eichler EE (2008) Recurrent rearrangements of chromosome 1q21.1 and variable pediatric phenotypes. N Engl J Med 359:1685-1699

239. Merikangas AK, Corvin AP, Gallagher L (2009) Copy-number variants in neurodevelopmental disorders: promises and challenges. Trends Genet 25(12):536-44

240. Metzker ML (2010) Sequencing technologies - the next generation. Nat Rev Genet 11:31-46

241. Meyer-Lindenberg A, Poline JB, Kohn PD, Holt JL, Egan MF, Weinberger DR, Berman KF (2001) Evidence for abnormal cortical functional connectivity during working memory in schizophrenia. Am J Psychiatry 158:1809-1817

242. Millar JK, Wilson-Annan JC, Anderson S, Christie S, Taylor MS, Semple CA, Devon RS, Clair DM, Muir WJ, Blackwood DH, Porteous DJ (2000) Disruption of two novel genes by a translocation co-segregating with schizophrenia. Hum Mol Genet 9(9):1415-23

243. Millar JK, Wilson-Annan JC, Anderson S, Christie S, Taylor MS, Semple CA, Devon RS, Clair DM, Muir WJ, Blackwood DH, Porteous DJ (2000) Disruption of two novel genes by a translocation co-segregating with schizophrenia. Hum Mol Genet 9(9):1415-23

244. Mills RE, Luttig CT, Larkins CE, Beauchamp A, Tsui C, Pittard WS, Devine SE (2006) An initial map of insertion and deletion (INDEL) variation in the human genome. Genome Res 16:1182-1190

245. Mitchell KJ & Porteous DJ (2009) GWAS for psychiatric disease: is the framework built on a solid foundation? Mol Psychiatry 14(8):740-1

246. Mitchell KJ, Porteous DJ (2010) Rethinking the genetic architecture of schizophrenia. Psychol Med:1-14

247. Mizuno K, Hasegawa K, Katagiri T, Ogimoto M, Ichikawa T, Yakura H (1993) MPTP delta, a putative murine homolog of HPTP delta, is expressed in specialized regions of the brain and in the B-cell lineage. Mol Cell Biol 13:5513-5523

248. Moffatt MF, Kabesch M, Liang L, Dixon AL, Strachan D, Heath S, Depner M, von Berg A, Bufe A, Rietschel E, Heinzmann A, Simma B, Frischer T, Willis-Owen SA, Wong KC, Illig T, Vogelberg C, Weiland SK, von Mutius E, Abecasis GR, Farrall M, Gut IG, Lathrop GM, Cookson WO (2007) Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. Nature448(7152):470-3

249. Morel B (1860) *Traité des maladies mentales*. Paris, 1852-1853; 2nd edition, 1860

250. Morgan TH (1911) The Origin of Nine Wing Mutations in Drosophila. Science 33:496-499

251. Mortensen PB, Juel K (1993) Mortality and causes of death in first admitted schizophrenic patients. Br J Psychiatry 163:183-189

252. Mueller JC (2004) Linkage disequilibrium for different scales and applications. Brief Bioinform 5:355-364

253. Murphy KC (2002) Schizophrenia and velo-cardio-facial syndrome. Lancet 359:426-430

254. Murphy KC, Jones LA, Owen MJ (1999) High rates of schizophrenia in adults with velo-cardio-facial syndrome. Arch Gen Psychiatry 56(10):940-5

255. Murray RM, Jones P, O'Callaghan E, Takei N, Sham P (1992) Genes, viruses and neurodevelopmental schizophrenia. J Psychiatr Res26(4):225-35

256. Myers AJ, Gibbs JR, Webster JA, Rohrer K, Zhao A, Marlowe L, Kaleem M, Leung D, Bryden L, Nath P, Zismann VL, Joshipura K, Huentelman MJ, Hu-Lince D, Coon KD, Craig DW, Pearson JV, Holmans P, Heward CB, Reiman EM, Stephan D, Hardy J (2007) A survey of genetic human cortical gene expression. Nat Genet 39(12):1494-9

257. Negishi M & Katoh H (2005) Rho family GTPases and dendrite plasticity. Neuroscientist 11(3):187-91

258. Ng MY, Levinson DF, Faraone SV, Suarez BK, DeLisi LE, Arinami T, Riley B, Paunio T, Pulver AE, Irmansyah, Holmans PA, Escamilla M, Wildenauer DB, Williams NM, Laurent C, Mowry BJ, Brzustowicz LM, Maziade M, Sklar P, Garver DL, Abecasis GR, Lerer B, Fallin MD, Gurling HM, Gejman PV, Lindholm E, Moises HW, Byerley W, Wijsman EM, Forabosco P, Tsuang MT, Hwu HG, Okazaki Y, Kendler KS, Wormley B, Fanous A, Walsh D, O'Neill FA, Peltonen L, Nestadt G, Lasseter VK, Liang KY, Papadimitriou GM, Dikeos DG, Schwab SG, Owen MJ, O'Donovan MC, Norton N, Hare E, Raventos H, Nicolini H, Albus M, Maier W, Nimgaonkar VL, Terenius L, Mallet J, Jay M, Godard S, Nertney D, Alexander M, Crowe RR, Silverman JM, Bassett AS, Roy MA, Mérette C, Pato CN, Pato MT, Roos JL, Kohn Y, Amann-Zalcenstein D, Kalsi G, McQuillin A, Curtis D,

Brynjolfson J, Sigmundsson T, Petursson H, Sanders AR, Duan J, Jazin E, Myles-Worsley M, Karayiorgou M, Lewis CM (2009a) Meta-analysis of 32 genome-wide linkage studies of schizophrenia. Mol Psychiatry 14(8):774-85

259. Ng SB, Bigham AW, Buckingham KJ, Hannibal MC, McMillin MJ, Gildersleeve HI, Beck AE, Tabor HK, Cooper GM, Mefford HC, Lee C, Turner EH, Smith JD, Rieder MJ, Yoshiura K, Matsumoto N, Ohta T, Niikawa N, Nickerson DA, Bamshad MJ, Shendure J (2010a) Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. Nat Genet 42:790-793
260. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA, Shendure J, Bamshad MJ (2010b) Exome sequencing identifies the cause of a mendelian disorder. Nat Genet 42:30-35
261. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE, Bamshad M, Nickerson DA, Shendure J (2009b) Targeted capture and massively parallel sequencing of 12 human exomes. Nat Genet 461:272-276
262. Nicole S, White PS, Topaloglu H, Beigthon P, Salih M, Hentati F, Fontaine B (1999) The human CDC42 gene: genomic organization, evidence for the existence of a putative pseudogene and exclusion as a SJS1 candidate gene. Human Genetics 105:98-103
263. Niwa M, Kamiya A, Murai R, Kubo K, Gruber AJ, Tomita K, Lu L, Tomisato S, Jaaro-Peled H, Seshadri S, Hiyama H, Huang B, Kohda K, Noda Y, O'Donnell P, Nakajima K, Sawa A, Nabeshima T (2010) Knockdown of DISC1 by in utero gene transfer disturbs postnatal dopaminergic maturation in the frontal cortex and leads to adult behavioral deficits. Neuron 65:480-489
264. Nobes CD & Hall A (1995) Rho, rac and cdc42 GTPases: regulators of actin structures, cell adhesion and motility. Biochem Soc Trans 23(3):456-9
265. Norton N, Williams HJ, Williams NM, Spurlock G, Zammit S, Jones G, Jones S, Owen R, O'Donovan MC, Owen MJ (2003) Mutation screening of the Homer gene family and association analysis in schizophrenia. Am J Med Genet B Neuropsychiatr Genet 120B (1):18-21
266. O'Driscoll M, Ruiz-Perez VL, Woods CG, Jeggo PA, Goodship JA (2003) A splicing mutation affecting expression of ataxia-telangiectasia and Rad3-related protein (ATR) results in Seckel syndrome. Nat Genet. 2003 Apr;33(4):497-501
267. O'Dushlaine CT, Dolan C, Weale ME, Stanton A, Croke DT, Kalviainen R, Eriksson K, Kantanen AM, Gibson RA, Hosford D, Sisodiya SM, Gill M, Corvin AP, Morris DW, Delanty N, Cavalleri GL (2008) An assessment of the Irish population for large-scale genetic mapping studies involving epilepsy and other complex diseases. Eur J Hum Genet 16(2):176-83
268. Ohta T & Kimura M (1969) Linkage disequilibrium due to random genetic drift. Genet Res 13: 47-55

269. Olenik C, Barth H, Just I, Aktories K, Meyer DK (1997) Gene expression of the small GTP-binding proteins RhoA, RhoB, Rac1, and Cdc42 in adult rat brain. Brain Res Mol Brain Res 52(2):263-9

270. Olney JW, Farber NB (1995) NMDA antagonists as neurotherapeutic drugs, psychotogens, neurotoxins, and research tools for studying schizophrenia. Neuropsychopharmacology 13:335-345

271. Orozco LD, Cokus SJ, Ghazalpour A, Ingram-Drake L, Wang S, van Nas A, Che N, Araujo JA, Pellegrini M, Lusis AJ (2009) Copy number variation influences gene expression and metabolic traits in mice. Hum Mol Genet 18(21):4118-29

272. Otani S, Daniel H, Roisin MP, Crepel F (2003) Dopaminergic modulation of long-term synaptic plasticity in rat prefrontal neurons. Cereb Cortex 13:1251-1256

273. O'Tuathaigh CM, Waddington JL (2010) Mutant mouse models: phenotypic relationships to domains of psychopathology and pathobiology in schizophrenia. Schizophr Bull 36(2):243-5

274. Owen MJ, Craddock N, Jablensky A (2007) The genetic deconstruction of psychosis. Schizophr Bull 33:905-911

275. Owen MJ, Craddock N. & O'Donovan MC (2005) Schizophrenia: genes at last? Trends in Genetics 21: 518 -525

276. Palotie A, Vaisanen P, Ott J, Ryhanen L, Elima K, Vikkula M, Cheah K, Vuorio E, Peltonen L (1989) Predisposition to familial osteoarthrosis linked to type II collagen gene. Lancet 1:924-927

277. Parisiadou L, Bethani I, Michaki V, Krousti K, Rapti G, *et al* (2008) Homer2 and Homer3 interact with amyloid precursor protein and inhibit Abeta production. Neurobiol Dis 30: 353–364

278. Patel NH, Vyas NS, Puri BK, Nijran KS, Al-Nahhas A (2010) Positron emission tomography in schizophrenia: a new perspective. J Nucl Med 51(4):511-20

279. Paunio T, Ekelund J, Varilo T, Parker A, Hovatta I, Turunen JA, Rinard K, Foti A, Terwilliger JD, Juvonen H, Suvisaari J, Arajarvi R, Suokas J, Partonen T, Lonnqvist J, Meyer J, Peltonen L (2001) Genome-wide scan in a nationwide study sample of schizophrenia families in Finland reveals susceptibility loci on chromosomes 2q and 5q. Hum Mol Genet 10:3037–3048

280. Pedersen CB, Mortensen PB (2001) Evidence of a dose-response relationship between urbanicity during upbringing and schizophrenia risk. Arch Gen Psychiatry 58(11):1039-46

281. Perez-Iratxeta C, Wjst M, Bork P, Andrade MA (2005) G2D: a tool for mining genes associated with disease. BMC Genet 6:45

282. Perry GH, Ben-Dor A, Tsalenko A, Sampas N, Rodriguez-Revenga L, Tran CW, Scheffer A, Steinfeld I, Tsang P, Yamada NA, Park HS, Kim JI, Seo JS, Yakhini Z, Laderman S, Bruhn L, Lee C (2008) The fine-scale and complex architecture of human copy-number variation. Am J Hum Genet 82(3):685-95

283. Perry GH, Ben-Dor A, Tsalenko A, Sampas N, Rodriguez-Revenga L, Tran CW, Scheffer A, Steinfeld I, Tsang P, Yamada NA, Park HS, Kim JI, Seo JS, Yakhini Z, Laderman S, Bruhn L, Lee C (2008) The fine-scale and complex architecture of human copy-number variation. Am J Hum Genet 82:685-695

284. Pertz OC, Wang Y, Yang F, Wang W, Gay LJ, Gristenko MA, Clauss TR, Anderson DJ, Liu T, Auberry KJ, Camp DG 2nd, Smith RD, Klemke RL (2008) Spatial mapping of the neurite and soma proteomes reveals a functional Cdc42/Rac regulatory network. Proc Natl Acad Sci USA. 105(6):1931-6

285. Philbert RA & Gershelfeld HK (2007) Theraputic potential of targetting gene variants in schizophrenia. Expert Rev Neurotheraputics 7(7):757-760

286. Pilowsky LS, Bressan RA, Stone JM, Erlandsson K, Mulligan RS, Krystal JH, Ell PJ (2006) First in vivo evidence of an NMDA receptor deficit in medication-free schizophrenic patients. Mol Psychiatry 11(2):118-9

287. Pinkel D, Segraves R, Sudar D, Clark S, Poole I, Kowbel D, Collins C, Kuo WL, Chen C, Zhai Y, Dairkee SH, Ljung BM, Gray JW, Albertson DG. (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nat Genet 20(2):207-11

288. Pinto D, Marshall C, Feuk L, Scherer SW (2007) Copy-number variation in control population cohorts. Hum Mol Genet 16 Spec No. 2:R168-173

289. Polymeropoulos MH, Lavedan C, Leroy E, Ide SE, Dehejia A, Dutra A, Pike B, Root H, Rubenstein J, Boyer R, Stenroos ES, Chandrasekharappa S, Athanassiadou A, Papapetropoulos T, Johnson WG, Lazzarini AM, Duvoisin RC, Di Iorio G, Golbe LI, Nussbaum RL (1997) Mutation in the alpha-synuclein gene identified in families with Parkinson's disease. Science 276:2045-2047

290. Porteous DJ, Thomson P, Brandon NJ, Millar JK (2006) The Genetics and Biology of Disc1 - An Emerging Role in Psychosis and Cognition. Biological Psychiatry 60:123-131

291. Price AL, Zaitlen NA, Reich D, Patterson N (2010) New approaches to population stratification in genome-wide association studies. Nat Rev Genet 11:459-463

292. Psychiatric GWAS Consortium Steering Committee (2009) A framework for interpreting genome-wide association studies of psychiatric disorders. Mol Psychiatry 14(1):10-7

293. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81:559-575

294. Purves D & Lichtman JW (1980) Elimination of synapses in the developing nervous system. Science 210:153-157

295. Quitsch A, Berhorster K, Liew CW, Richter D, Kreienkamp HJ (2005) Postsynaptic shank antagonizes dendrite branching induced by the leucine-rich repeat protein Densin-180. J Neurosci 25:479-487

296. Rahimov F, Marazita ML, Visel A, Cooper ME, Hitchler MJ, Rubini M, Domann FE, Govil M, Christensen K, Bille C, Melbye M, Jugessur A, Lie RT, Wilcox AJ, Fitzpatrick DR, Green ED, Mossey PA, Little J, Steegers-Theunissen RP, Pennacchio LA, Schutte BC, Murray JC (2008) Disruption of an AP-2alpha binding site in an IRF6 enhancer is associated with cleft lip. Nat Genet 40:1341-1347

297. Ramón y Cajal, S (1899) Comparative Study of the Sensory Areas of the Human Cortex. BiblioLife, 2010

298. Renwick JH (1971) Assignment and map-positioning of human loci using chromosomal variation. Ann Hum Genet 35:79-97

299. Rieff HI, Corfas G (2006) ErbB receptor signalling regulates dendrite formation in mouse cerebellar granule cells in vivo. Eur J Neurosci 23:2225-2229

300. Riordan JR, Rommens JM, Kerem B, Alon N, Rozmahel R, Grzelczak Z, Zielenski J, Lok S, Plavsic N, Chou JL, et al. (1989) Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. Science 245:1066-1073

301. Rioux JD, Xavier RJ, Taylor KD, Silverberg MS, Goyette P, Huett A, Green T, Kuballa P, Barmada MM, Datta LW, Shugart YY, Griffiths AM, Targan SR, Ippoliti AF, Bernard EJ, Mei L, Nicolae DL, Regueiro M, Schumm LP, Steinhart AH, Rotter JI, Duerr RH, Cho JH, Daly MJ, Brant SR (2007) Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. Nat Genet 39(5):596-604

302. Risch N & Baron M (1984) Segregation analysis of schizophrenia and related disorders. Am J Hum Genet 36(5):1039-59

303. Risch N & Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273(5281):1516-7

304. Roberts, E (1972) Prospects for research on schizophrenia. A hypotheses suggesting that there is a defect in the GABA system in schizophrenia. Neurosci Res Program Bull 10(4):468-82

305. Rohatgi R, Ma L, Miki H, Lopez M, Kirchhausen T, Takenawa T, Kirschner MW (1999) The interaction between N-WASP and the Arp2/3 complex links Cdc42-dependent signals to actin assembly. Cell 97(2):221-31

306. Rosenbaum G, Cohen BD, Luby ED, Gottlieb JS, Yelen D (1959) Comparison of sernyl with other drugs: simulation of schizophrenic performance with sernyl, LSD-25, and amobarbital (amytal) sodium; I. Attention, motor function, and proprioception. Arch Gen Psychiatry 1:651-656

307. Rossi S, Masotti D, Nardini C, Bonora E, Romeo G, Macii E, Benini L, Volinia S (2006) TOM: a web-based integrated approach for identification of candidate disease genes. Nucleic Acids Res 34(Web Server issue):W285-92

308. Rudin (1916) *Zur Vererbung und Neuentslehung der dementia praecox*. Berlin, Springer.

309. Saha S, Chant D, Welham J, McGrath J (2005) A systematic review of the prevalence of schizophrenia. PLoS Med 2(5): e141

310. Salero E & Hatten ME (2007). Differentiation of ES cells into cerebellar neurons. Proc Natl Acad Sci USA 104: 2997–3002

311. Satake W, Nakabayashi Y, Mizuta I, Hirota Y, Ito C, Kubo M, Kawaguchi T, Tsunoda T, Watanabe M, Takeda A, Tomiyama H, Nakashima K, Hasegawa K, Obata F, Yoshikawa T, Kawakami H, Sakoda S, Yamamoto M, Hattori N, Murata M, Nakamura Y, Toda T (2009) Genome-wide association study identifies common variants at four loci as genetic risk factors for Parkinson's disease. Nat Genet 41(12):1303-7

312. Schneider K. Clinical psychopathology (trans M W Hamilton). Grune and Stratton, New York, 1959

313. Schrimpf SP, Meskenaite V, Brunner E, Rutishauser D, Walther P, Eng J, Aebersold R, Sonderegger P (2005) Proteomic analysis of synaptosomes using isotope-coded affinity tags and mass spectrometry. Proteomics 5(10):2531-41

314. Schuler GD, Boguski MS, Stewart EA, Stein LD, Gyapay G, Rice K, White RE, Rodriguez-Tome P, Aggarwal A, Bajorek E, Bentolila S, Birren BB, Butler A, Castle AB, Chiannilkulchai N, Chu A, Clee C, Cowles S, Day PJ, Dibling T, Drouot N, Dunham I, Duprat S, East C, Edwards C, Fan JB, Fang N, Fizames C, Garrett C, Green L, Hadley D, Harris M, Harrison P, Brady S, Hicks A, Holloway E, Hui L, Hussain S, Louis-Dit-Sully C, Ma J, MacGilvery A, Mader C, Maratukulam A, Matise TC, McKusick KB, Morissette J, Mungall A, Muselet D, Nusbaum HC, Page DC, Peck A, Perkins S, Piercy M, Qin F, Quackenbush J, Ranby S, Reif T, Rozen S, Sanders C, She X, Silva J, Slonim DK, Soderlund C, Sun WL, Tabar P, Thangarajah T, Vega-Czarny N, Vollrath D, Voyticky S, Wilmer T, Wu X, Adams MD, Auffray C, Walter NA, Brandon R, Dehejia A, Goodfellow PN, Houlgatte R, Hudson JR, Jr., Ide SE, Iorio KR, Lee WY, Seki N, Nagase T, Ishikawa K, Nomura N, Phillips C, Polymeropoulos MH, Sandusky M, Schmitt K, Berry R, Swanson K, Torres R, Venter JC, Sikela JM, Beckmann JS, Weissenbach J, Myers RM, Cox DR, James MR, Bentley D, Deloukas P, Lander ES, Hudson TJ (1996) A gene map of the human genome. Science 274:540-546

315. Schwartz RC (1998) Symptomatology and insight in schizophrenia. Psychol Rep 82:227-233

316. Scott EK, Reuter JE, Luo L (2003) Small GTPase Cdc42 is required for multiple aspects of dendritic morphogenesis. J Neurosci 23(8):3118-23

317. Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Maner S, Massa H, Walker M, Chi M, Navin N, Lucito R, Healy J, Hicks J, Ye K, Reiner A, Gilliam TC, Trask B, Patterson N, Zetterberg A, Wigler M (2004) Large-scale copy number polymorphism in the human genome. Science 305:525-528

318. Sebat J, Levy DL, McCarthy SE (2009) Rare structural variants in schizophrenia: one disorder, multiple mutations; one mutation, multiple disorders. Trends Genet 25:528-535

319. Seeman P, Chau-Wong M, Tedesco J, Wong K (1975) Brain receptors for antipsychotic drugs and dopamine: direct binding assays. Proc Natl Acad Sci U S A 72:4376-4380

320. Semple D, Smyth R, Burns J, Darjee R, McIntosh (2006) Oxford Handbook of Pyschiatry. Oxford University Press

321. Sepp KJ, Hong P, Lizarraga SB, Liu JS, Mejia LA, Walsh CA, Perrimon N (2008) Identification of neural outgrowth genes using genome-wide RNAi. PLoS Genet 4(7):e1000111

322. Shen W, Wu B, Zhang Z, Dou Y, Rao ZR, Chen YR, Duan S (2006) Activity-induced rapid synaptic maturation mediated by presynaptic cdc42 signaling. Neuron 50(3):401-14

323. Shen Y, Li N, Wu S, Zhou Y, Shan Y, Zhang Q, Ding C, Yuan Q, Zhao F, Zeng R, Zhu X (2008) Nudel Binds Cdc42GAP to Modulate Cdc42 Activity at the Leading Edge of Migrating Cells. Developmental Cell 14:342-353

324. Shi J, Levinson DF, Duan J, Sanders AR, Zheng Y, Pe'er I, Dudbridge F, Holmans PA, Whittemore AS, Mowry BJ, Olincy A, Amin F, Cloninger CR, Silverman JM, Buccola NG, Byerley WF, Black DW, Crowe RR, Oksenberg JR, Mirel DB, Kendler KS, Freedman R, Gejman PV (2009) Common variants on chromosome 6p22.1 are associated with schizophrenia. Nature 460(7256):753-7

325. Shih RA, Belmonte PL, Zandi PP (2004) A review of the evidence from family, twin and adoption studies for a genetic contribution to adult psychiatric disorders. Int Rev Psychiatry 16:260-283

326. Shinjo K, Koland JG, Hart MJ, Narasimhan V, Johnson DI, Evans T, Cerione RA (1990) Molecular cloning of the gene for the human placental GTP-binding protein Gp (G25K): identification of this GTP-binding protein as the human homolog of the yeast cell-division-cycle protein CDC42. Proc Natl Acad Sci USA 87(24):9853-7

327. Shiraishi Y, Mizutani A, Bito H, Fujisawa K, Narumiya S, Mikoshiba K, Furuichi T (1999) Cupidin, an Isoform of Homer/Vesl, Interacts with the Actin Cytoskeleton and Activated Rho Family Small GTPases and Is Expressed in Developing Mouse Cerebellar Granule Cells. J Neurosci 19:8389-8400

328. Shiraishi Y, Mizutani A, Mikoshiba K, Furuichi T (2003) Coincidence in dendritic clustering and synaptic targeting of homer proteins and NMDA receptor complex proteins NR2B and PSD95 during development of cultured hippocampal neurons. Mol Cell Neurosci 22(2):188-201

329. Shiraishi-Yamaguchi Y, Furuichi T (2007) The Homer family proteins. Genome Biol 8(2):206

330. Shiraishi-Yamaguchi Y, Sato Y, Sakai R, Mizutani A, Knopfel T, Mori N, Mikoshiba K, Furuichi T (2009) Interaction of Cupidin/Homer2 with two actin cytoskeletal regulators, Cdc42 small GTPase and Drebrin, in dendritic spines. BMC Neurosci 10:25

331. Simón-Sánchez J, Schulte C, Bras JM, Sharma M, Gibbs JR, Berg D, Paisan-Ruiz C, Lichtner P, Scholz SW, Hernandez DG, Krüger R, Federoff M, Klein C, Goate A, Perlmutter J, Bonin M, Nalls MA, Illig T, Gieger C, Houlden H, Steffens M, Okun MS, Racette BA, Cookson MR, Foote KD, Fernandez HH, Traynor BJ, Schreiber S, Arepalli S, Zonozi R, Gwinn K, van der Brug M, Lopez G, Chanock SJ, Schatzkin A, Park Y, Hollenbeck A, Gao J, Huang X, Wood NW, Lorenz D, Deuschl G, Chen H, Riess O, Hardy JA, Singleton AB, Gasser T (2009) Genome-wide association study reveals genetic risk underlying Parkinson's disease. Nat Genet 41(12):1308-12

332. Sisk CL & Foster DL (2004) The neural basis of puberty and adolescence. Nat Neurosci 7:1040-1047

333. Snyder SH (1976) The dopamine hypothesis of schizophrenia: focus on the dopamine receptor. Am J Psych 133: 197-202

334. Snyder SH (1981) Dopamine receptors, neuroleptics, and schizophrenia. Am J Psychiatry 138:460-464

335. Spencer CC, Su Z, Donnelly P, Marchini J (2009) Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. PLoS Genet 5(5):e1000477

336. Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 52(3):506-16

337. St Clair D, Blackwood D, Muir W, Carothers A, Walker M, Spowart G, Gosden C, Evans HJ (1990) Association within a family of a balanced autosomal translocation with major mental illness. Lancet 336(8706):13-6

338. Stefansson H, Ophoff RA, Steinberg S, Andreassen OA, Cichon S, Rujescu D, Werge T, Pietiläinen OP, Mors O, Mortensen PB, Sigurdsson E, Gustafsson O, Nyegaard M, Tuulio-Henriksson A, Ingason A, Hansen T, Suvisaari J, Lonnqvist J, Paunio T, Børglum AD, Hartmann A, Fink-Jensen A, Nordentoft M, Hougaard D, Norgaard-Pedersen B, Böttcher Y, Olesen J, Breuer R, Möller HJ, Giegling I, Rasmussen HB, Timm S, Mattheisen M, Bitter I, Réthelyi JM, Magnusdottir BB, Sigmundsson T, Olason P, Masson G, Gulcher JR, Haraldsson M, Fossdal R, Thorgeirsson TE, Thorsteinsdottir U, Ruggeri M, Tosato S, Franke B, Strengman E, Kiemeney LA; Genetic Risk and Outcome in Psychosis (GROUP), Melle I, Djurovic S, Abramova L, Kaleda V, Sanjuan J, de Frutos R, Bramon E, Vassos E, Fraser G, Ettinger U, Picchioni M, Walker N, Toulopoulou T, Need AC, Ge D,

Yoon JL, Shianna KV, Freimer NB, Cantor RM, Murray R, Kong A, Golimbet V, Carracedo A, Arango C, Costas J, Jönsson EG, Terenius L, Agartz I, Petursson H, Nöthen MM, Rietschel M, Matthews PM, Muglia P, Peltonen L, St Clair D, Goldstein DB, Stefansson K, Collier DA (2009) Common variants conferring risk of schizophrenia. Nature 460(7256):744-7

339. Stefansson H, Rujescu D, Cichon S, Pietiläinen OP, Ingason A, Steinberg S, Fossdal R, Sigurdsson E, Sigmundsson T, Buizer-Voskamp JE, Hansen T, Jakobsen KD, Muglia P, Francks C, Matthews PM, Gylfason A, Halldorsson BV, Gudbjartsson D, Thorgeirsson TE, Sigurdsson A, Jonasdottir A, Jonasdottir A, Bjornsson A, Mattiasdottir S, Blondal T, Haraldsson M, Magnusdottir BB, Giegling I, Möller HJ, Hartmann A, Shianna KV, Ge D, Need AC, Crombie C, Fraser G, Walker N, Lonnqvist J, Suvisaari J, Tuulio-Henriksson A, Paunio T, Toulopoulou T, Bramon E, Di Forti M, Murray R, Ruggeri M, Vassos E, Tosato S, Walshe M, Li T, Vasilescu C, Mühleisen TW, Wang AG, Ullum H, Djurovic S, Melle I, Olesen J, Kiemeney LA, Franke B; GROUP, Sabatti C, Freimer NB, Gulcher JR, Thorsteinsdottir U, Kong A, Andreassen OA, Ophoff RA, Georgi A, Rietschel M, Werge T, Petursson H, Goldstein DB, Nöthen MM, Peltonen L, Collier DA, St Clair D, Stefansson K (2008) Large recurrent microdeletions associated with schizophrenia. Nature 455(7210):232-6
340. Stefansson H, Sigurdsson E, Steinthorsdottir V, Bjornsdottir S, Sigmundsson T et al. (2002) Neuregulin 1 and susceptibility to schizophrenia. Am J Hum Genet 71(4): 877-92
341. Stefansson H, Sigurdsson E, Steinthorsdottir V, Bjornsdottir S, Sigmundsson T, Ghosh S, Brynjolfsson J, Gunnarsdottir S, Ivarsson O, Chou TT, Hjaltason O, Birgisdottir B, Jonsson H, Gudnadottir VG, Gudmundsdottir E, Bjornsson A, Ingvarsson B, Ingason A, Sigfusson S, Hardardottir H, Harvey RP, Lai D, Zhou M, Brunner D, Mutel V, Gonzalo A, Lemke G, Sainz J, Johannesson G, Andresson T, Gudbjartsson D, Manolescu A, Frigge ML, Gurney ME, Kong A, Gulcher JR, Petursson H, Stefansson K (2002) Neuregulin 1 and susceptibility to schizophrenia. Am J Hum Genet 71(4):877-92
342. Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, Beazley C, Ingle CE, Dunning M, Flicek P, Koller D, Montgomery S, Tavaré S, Deloukas P, Dermitzakis ET (2007) Population genomics of human gene expression. Nat Genet 39(10):1217-24
343. Straub RE, Jiang Y, MacLean CJ, Ma Y, Webb BT, Myakishev MV, Harris-Kerr C, Wormley B, Sadek H, Kadambi B, Cesare AJ, Gibberman A, Wang X, O'Neill FA, Walsh D, Kendler KS (2002a) Genetic variation in the 6p22.3 gene DTNBP1, the human ortholog of the mouse dysbindin gene, is associated with schizophrenia. Am J Hum Genet 71(2):337-48
344. Straub RE, MacLean CJ, Ma Y, Webb BT, Myakishev MV, Harris-Kerr C, Wormley B, Sadek H, Kadambi B, O'Neill FA, Walsh D, Kendler KS (2002b) Genome-wide scans of three independent

sets of 90 Irish multiplex schizophrenia families and follow-up of selected regions in all families provides evidence for multiple susceptibility genes. Mol Psychiatry 7:542–559

345. Straub RE, Weinberger DR (2006) Schizophrenia genes - famine to feast. Biol Psychiatry 60(2):81-3
346. Sullivan PF & Gejman PV (2010) Response to Mitchell and Porteus. Mol Psychiatry 15:450-452
347. Sullivan PF (2005) The genetics of schizophrenia. PLoS Med 2(7): e212
348. Sullivan PF (2010) The psychiatric GWAS consortium: big science comes to psychiatry. Neuron 68(2):182-6
349. Szumlinski KK, Kalivas PW, Worley PF (2006) Homer proteins: implications for neuropsychiatric disorders. Curr Opin Neurobiol 16(3):251-7
350. Szumlinski KK, Lominac KD, Kleschen MJ, Oleson EB, Dehoff MH, Schwarz MK, Seeburg PH, Worley PF, Kalivas PW (2005) Behavioral and neurochemical phenotyping of Homer1 mutant mice: possible relevance to schizophrenia. Genes Brain Behav 4(5):273-88
351. Tabangin ME, Woo JG, Martin LJ (2009) The effect of minor allele frequency on the likelihood of obtaining false positives. BMC Proc 3 Suppl 7:S41
352. Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123(3):585-95
353. Talbot K, Eidem WL, Tinsley CL, Benson MA, Thompson EW, Smith RJ, Hahn CG, Siegel SJ, Trojanowski JQ, Gur RE, Blake DJ, Arnold SE (2004) Dysbindin-1 is reduced in intrinsic, glutamatergic terminals of the hippocampal formation in schizophrenia. J Clin Invest 113:1353-1363
354. Tam GW, Redon R, Carter NP, Grant SG (2009) The role of DNA copy number variation in schizophrenia. Biol Psychiatry 66:1005-1012
355. Tam GW, van de Lagemaat LN, Redon R, Strathdee KE, Croning MD, Malloy MP, Muir WJ, Pickard BS, Deary IJ, Blackwood DH, Carter NP, Grant SG (2010) Confirmed rare copy number variants implicate novel genes in schizophrenia. Biochem Soc Trans 38(2):445-51
356. Tandon R, Nasrallah HA, Keshavan MS (2010) Schizophrenia, "just the facts" 5. Treatment and prevention. Past, present, and future. Schizophr Res 122(1-3):1-23
357. Tandon R, Keshavan MS, Nasrallah HA (2008) Schizophrenia, "Just the Facts": what we know in 2008 part 1: overview. Schizophr Res 100:4-19
358. Tang TT, Yang F, Chen BS, Lu Y, Ji Y, Roche KW, Lu B (2009) Dysbindin regulates hippocampal LTP by controlling NMDA receptor surface expression. Proc Natl Acad Sci USA 106:21395-21400
359. Tarrier N (2010) Cognitive behavior therapy for schizophrenia and psychosis: current status and future directions. Clin Schizophr Relat Psychoses 4(3):176-84

360. Teufel M, Saudek V, Ledig JP, Bernhardt A, Boularand S, Carreau A, Cairns NJ, Carter C, Cowley DJ, Duverger D, Ganzhorn AJ, Guenet C, Heintzelmann B, Laucher V, Sauvage C, Smirnova T (2003) Sequence identification and characterization of human carnosinase and a closely related non-specific dipeptidase. J Biol Chem 278(8):6521-31

361. Thomas G, Jacobs KB, Yeager M, Kraft P, Wacholder S, Orr N, Yu K, Chatterjee N, Welch R, Hutchinson A, Crenshaw A, Cancel-Tassin G, Staats BJ, Wang Z, Gonzalez-Bosquet J, Fang J, Deng X, Berndt SI, Calle EE, Feigelson HS, Thun MJ, Rodriguez C, Albanes D, Virtamo J, Weinstein S, Schumacher FR, Giovannucci E, Willett WC, Cussenot O, Valeri A, Andriole GL, Crawford ED, Tucker M, Gerhard DS, Fraumeni JF Jr, Hoover R, Hayes RB, Hunter DJ, Chanock SJ (2008) Multiple loci identified in a genome-wide association study of prostate cancer. Nat Genet 40(3):310-5

362. Thompson PM, Sower AC, Perrone-Bizzozero NI (1998) Altered levels of the synaptosomal associated protein SNAP-25 in schizophrenia. Biol Psychiatry 43:239-243

363. Threadgill R, Bobb K, Ghosh A (1997) Regulation of dendritic growth and remodeling by Rho, Rac, and Cdc42. Neuron 19(3):625-34

364. Tian C, Gregersen PK, Seldin MF (2008) Accounting for ancestry: population substructure and genome-wide association studies. Hum Mol Genet 17:R143-150

365. Tkachev D, Mimmack ML, Huffaker SJ, Ryan M, Bahn S (2007) Further evidence for altered myelin biosynthesis and glutamatergic dysfunction in schizophrenia. Int J Neuropsychopharmacol 10:557-563

366. Tsai G, Passani LA, Slusher BS, Carter R, Baer L, Kleinman JE, Coyle JT (1995) Abnormal excitatory neurotransmitter metabolism in schizophrenic brains. Arch Gen Psychiatry 52:829-836

367. Tseng KY, Lewis BL, Lipska BK, O'Donnell P (2007) Post-pubertal disruption of medial prefrontal cortical dopamine-glutamate interactions in a developmental animal model of schizophrenia. Biol Psychiatry 62(7):730-8

368. Tu JC, Xiao B, Naisbitt S, Yuan JP, Petralia RS, Brakeman P, Doan A, Aakalu VK, Lanahan AA, Sheng M, *et al* (1999) Coupling of mGluR/Homer and PSD-95 complexes by the Shank family of postsynaptic density proteins. Neuron 23:583-592

369. Tu JC, Xiao B, Yuan JP, Lanahan AA, Leoffert K, Li M, Linden DJ, Worley PF (1998) Homer binds a novel proline-rich motif and links group 1 metabotropic glutamate receptors with IP3 receptors. Neuron 21:717-726

370. Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM, Haugen E, Hayden H, Albertson D, Pinkel D, Olson MV, Eichler EE (2005) Fine-scale structural variation of the human genome. Nat Genet 37:727-732

371. Uetani N, Kato K, Ogura H, Mizuno K, Kawano K, Mikoshiba K, Yakura H, Asano M, Iwakura Y (2000) Impaired learning with enhanced hippocampal long-term potentiation in PTPdelta-deficient mice. Embo J 19:2775-2785

372. Uhlhaas PJ & Singer W (2010) Abnormal neural oscillations and synchrony in schizophrenia. Nat Rev Neurosci 11(2):100-13

373. Umbricht D & Krljes S (2005) Mismatch negativity in schizophrenia: a meta-analysis. Schizophr Res 76(1):1-23

374. Valente EM, Abou-Sleiman PM, Caputo V, Muqit MM, Harvey K, Gispert S, Ali Z, Del Turco D, Bentivoglio AR, Healy DG, Albanese A, Nussbaum R, Gonzalez-Maldonado R, Deller T, Salvi S, Cortelli P, Gilks WP, Latchman DS, Harvey RJ, Dallapiccola B, Auburger G, Wood NW (2004) Hereditary early-onset Parkinson's disease caused by mutations in PINK1. Science 304:1158-1160

375. Veling W, Susser E, van Os J, Mackenbach JP, Selten JP, Hoek HW (2008) Ethnic density of neighborhoods and incidence of psychotic disorders among immigrants. Am J Psychiatry 165:66-73

376. Verrall L, Burnet PW, Betts JF, Harrison PJ (2010) The neurobiology of D-amino acid oxidase and its involvement in schizophrenia. Mol Psychiatry 15:122-137

377. Via M, Gignoux C, Burchard EG (2010) The 1000 Genomes Project: new opportunities for research and social challenges. Genome Med 2(1):3

378. Walsh T, McClellan JM, McCarthy SE, Addington AM, Pierce SB, Cooper GM, Nord AS, Kusenda M, Malhotra D, Bhandari A, Stray SM, Rippey CF, Roccanova P, Makarov V, Lakshmi B, Findling RL, Sikich L, Stromberg T, Merriman B, Gogtay N, Butler P, Eckstrand K, Noory L, Gochman P, Long R, Chen Z, Davis S, Baker C, Eichler EE, Meltzer PS, Nelson SF, Singleton AB, Lee MK, Rapoport JL, King MC, Sebat J (2008) Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. Science 320(5875):539-43

379. Walters JT, Corvin A, Owen MJ, Williams H, Dragovic M, Quinn EM, Judge R, Smith DJ, Norton N, Giegling I, Hartmann AM, Möller HJ, Muglia P, Moskvina V, Dwyer S, O'Donoghue T, Morar B, Cooper M, Chandler D, Jablensky A, Gill M, Kaladjieva L, Morris DW, O'Donovan MC, Rujescu D, Donohoe G (2010) Psychosis susceptibility gene ZNF804A and cognitive performance in schizophrenia. Arch Gen Psychiatry. 2010 Jul;67(7):692-700

380. Walters RG, Jacquemont S, Valsesia A, de Smith AJ, Martinet D, Andersson J, Falchi M, Chen F, Andrieux J, Lobbens S, Delobel B, Stutzmann F, El-Sayed Moustafa JS, Chevre JC, Lecoeur C, Vatin V, Bouquillon S, Buxton JL, Boute O, Holder-Espinasse M, Cuisset JM, Lemaitre MP, Ambresin AE, Brioschi A, Gaillard M, Giusti V, Fellmann F, Ferrarini A, Hadjikhani N, Campion D, Guilmatre A,

Goldenberg A, Calmels N, Mandel JL, Le Caignec C, David A, Isidor B, Cordier MP, Dupuis-Girod S, Labalme A, Sanlaville D, Beri-Dexheimer M, Jonveaux P, Leheup B, Ounap K, Bochukova EG, Henning E, Keogh J, Ellis RJ, Macdermot KD, van Haelst MM, Vincent-Delorme C, Plessis G, Touraine R, Philippe A, Malan V, Mathieu-Dramard M, Chiesa J, Blaumeiser B, Kooy RF, Caiazzo R, Pigeyre M, Balkau B, Sladek R, Bergmann S, Mooser V, Waterworth D, Reymond A, Vollenweider P, Waeber G, Kurg A, Palta P, Esko T, Metspalu A, Nelis M, Elliott P, Hartikainen AL, McCarthy MI, Peltonen L, Carlsson L, Jacobson P, Sjostrom L, Huang N, Hurles ME, O'Rahilly S, Farooqi IS, Mannik K, Jarvelin MR, Pattou F, Meyre D, Walley AJ, Coin LJ, Blakemore AI, Froguel P, Beckmann JS A new highly penetrant form of obesity due to deletions on chromosome 16p11.2. Nature 463:671-675

381. Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, Ghandour G, Perkins N, Winchester E, Spencer J, Kruglyak L, Stein L, Hsie L, Topaloglou T, Hubbell E, Robinson E, Mittmann M, Morris MS, Shen N, Kilburn D, Rioux J, Nusbaum C, Rozen S, Hudson TJ, Lipshutz R, Chee M, Lander ES (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science 280(5366):1077-82
382. Wang J & Bixby JL (1999) Receptor tyrosine phosphatase-delta is a homophilic, neurite-promoting cell adhesion molecular for CNS neurons. Mol Cell Neurosci 14:370-384
383. Wang L, Yang L, Debidda M, Witte D, Zheng Y (2007) Cdc42 GTPase-activating protein deficiency promotes genomic instability and premature aging-like phenotypes. Proc Natl Acad Sci USA 104:1248-1253
384. Webster JA, Gibbs JR, Clarke J, Ray M, Zhang W, Holmans P, Rohrer K, Zhao A, Marlowe L, Kaleem M, McCorquodale DS, 3rd, Cuello C, Leung D, Bryden L, Nath P, Zismann VL, Joshipura K, Huentelman MJ, Hu-Lince D, Coon KD, Craig DW, Pearson JV, Heward CB, Reiman EM, Stephan D, Hardy J, Myers AJ (2009) Genetic control of human brain transcript expression in Alzheimer disease. Am J Hum Genet 84:445-458
385. Weinberger DR, Berman KF, Torrey EF (1992) Correlations between abnormal hippocampal morphology and prefrontal physiology in schizophrenia. Clin Neuropharmacol 15, Suppl 1 Pt A:393A-394A
386. Wernike, C (1906) *Grundriss der Psychiatrie in klinischen Vorlesungen*. Leipzig, G. Thieme, 1894-1900; 2nd edition, 1906; translated into English
387. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT, Gomes X, Tartaro K, Niazi F, Turcotte CL, Irzyk GP, Lupski JR, Chinault C, Song XZ, Liu Y, Yuan Y, Nazareth L, Qin X, Muzny DM, Margulies M, Weinstock GM, Gibbs RA, Rothberg JM

(2008) The complete genome of an individual by massively parallel DNA sequencing. Nature 452:872-876

388. Whitcombe D, Newton CR, Little S (1998) Advances in approaches to DNA-based diagnostics. Curr Opin Biotechnol 9(6):602-8

389. WHO (2007) International Statistical Classification of Diseases and Related Health Problems 10th Revision. World Health Organization

390. Williams HJ, Norton N, Dwyer S, Moskvina V, Nikolov I, Carroll L, Georgieva L, Williams NM, Morris DW, Quinn EM, Giegling I, Ikeda M, Wood J, Lencz T, Hultman C, Lichtenstein P, Thiselton D, Maher BS; Molecular Genetics of Schizophrenia Collaboration (MGS) International Schizophrenia Consortium (ISC), SGENE-plus, GROUP, Malhotra AK, Riley B, Kendler KS, Gill M, Sullivan P, Sklar P, Purcell S, Nimgaonkar VL, Kirov G, Holmans P, Corvin A, Rujescu D, Craddock N, Owen MJ, O'Donovan MC (2010) Fine mapping of ZNF804A and genome-wide significant evidence for its involvement in schizophrenia and bipolar disorder. Mol Psychiatry. *in press*

391. Williams NM, Rees MI, Holmans P, Norton N, Cardno AG, Jones LA, Murphy KC, Sanders RD, McCarthy G, Gray MY, Fenton I, McGuffin P, Owen MJ (1999) A two-stage genome scan for schizophrenia susceptibility genes in 196 affected sibling pairs. Hum Mol Genet 8:1729–1739

392. Wilson GM, Flibotte S, Chopra V, Melnyk BL, Honer WG, Holt RA (2006) DNA copy-number analysis in bipolar disorder and schizophrenia reveals aberrations in genes involved in glutamate signaling. Hum Mol Genet. 2006 Mar 1;15(5):743-9

393. Winchester L, Yau C, Ragoussis J (2009) Comparing CNV detection methods for SNP arrays. Brief Funct Genomic Proteomic 8:353-366

394. Woo TU, Kim AM, Viscidi E (2008) Disease-specific alterations in glutamatergic neurotransmission on inhibitory interneurons in the prefrontal cortex in schizophrenia. Brain Res 1218:267-77

395. Wratten NS, Memoli H, Huang Y, Dulencin AM, Matteson PG, Cornacchia MA, Azaro MA, Messenger J, Hayter JE, Bassett AS, Buyske S, Millonig JH, Vieland VJ, Brzustowicz LM (2009) Identification of a schizophrenia-associated functional noncoding variant in *NOS1AP*. Am J Psychiatry 166(4):434-41

396. Wurtz RH, Castellucci VF, Nusrala JM (1967) Synaptic plasticity: the effect of the action potential in the postsynaptic neuron. Exp Neurol 18:350-368

397. Xiao B, Tu JC, Petralia RS, Yuan JP, Doan A, Breder CD, Ruggiero A, Lanahan AA, Wenthold RJ, Worley PF (1998) Homer regulates the association of group 1 metabotropic glutamate receptors with multivalent complexes of homer-related, synaptic proteins. Neuron 21(4):707-16

398. Xu B, Roos JL, Levy S, van Rensburg EJ, Gogos JA, Karayiorgou M (2008) Strong association of *de novo* copy number mutations with sporadic schizophrenia. Nat Genet 40:880-885

399. Xu Q, Jia YB, Zhang BY, Zou K, Tao YB, Wang YP, Qiang BQ, Wu GY, Shen Y, Ji HK, Huang Y, Sun XQ, Ji L, Li YD, Yuan YB, Shu L, Yu X, Shen YC, Yu YQ, Ju GZ (2004) Association study of an SNP combination pattern in the dopaminergic pathway in paranoid schizophrenia: a novel strategy for complex disorders. Mol Psychiatry 9:510-521

400. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, Goddard ME, Visscher PM (2010) Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42(7):565-9

401. Yuan JP, Kiselyov K, Shin DM, Chen J, Shcheynikov N, Kang SH, Dehoff MH, Schwarz MK, Seeburg PH, Muallem S, Worley PF (2003) Homer binds TRPC family channels and is required for gating of TRPC1 by IP3 receptors. Cell 114(6):777-89

402. Yoon S, Xuan Z, Makarov V, Ye K, Sebat J (2009) Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Res 19(9):1586-92

403. Zeggini E, Rayner W, Morris AP, Hattersley AT, Walker M, Hitman GA, Deloukas P, Cardon LR, McCarthy MI (2005) An evaluation of HapMap sample size and tagging SNP performance in large-scale empirical and simulated data sets. Nat Genet 37(12):1320-2

404. Zeggini E, Scott LJ, Saxena R, Voight BF, Marchini JL, Hu T, de Bakker PI, Abecasis GR, Almgren P, Andersen G, Ardlie K, Boström KB, Bergman RN, Bonnycastle LL, Borch-Johnsen K, Burtt NP, Chen H, Chines PS, Daly MJ, Deodhar P, Ding CJ, Doney AS, Duren WL, Elliott KS, Erdos MR, Frayling TM, Freathy RM, Gianniny L, Grallert H, Grarup N, Groves CJ, Guiducci C, Hansen T, Herder C, Hitman GA, Hughes TE, Isomaa B, Jackson AU, Jørgensen T, Kong A, Kubalanza K, Kuruvilla FG, Kuusisto J, Langenberg C, Lango H, Lauritzen T, Li Y, Lindgren CM, Lyssenko V, Marvelle AF, Meisinger C, Midthjell K, Mohlke KL, Morken MA, Morris AD, Narisu N, Nilsson P, Owen KR, Palmer CN, Payne F, Perry JR, Pettersen E, Platou C, Prokopenko I, Qi L, Qin L, Rayner NW, Rees M, Roix JJ, Sandbaek A, Shields B, Sjögren M, Steinthorsdottir V, Stringham HM, Swift AJ, Thorleifsson G, Thorsteinsdottir U, Timpson NJ, Tuomi T, Tuomilehto J, Walker M, Watanabe RM, Weedon MN, Willer CJ; Wellcome Trust Case Control Consortium, Illig T, Hveem K, Hu FB, Laakso M, Stefansson K, Pedersen O, Wareham NJ, Barroso I, Hattersley AT, Collins FS, Groop L, McCarthy MI, Boehnke M, Altshuler D (2008) Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. Nat Genet 40(5):638-45

405. Zhang J, Feuk L, Duggan GE, Khaja R, Scherer SW (2006) Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome. Cytogenet Genome Res 115(3-4):205-14

406. Zhang XH & Chasin LA (2004) Computational definition of sequence motifs governing constitutive exon splicing. Genes Dev 18(11):1241-50

Summary

Investigation of glutamatergic synapse genes in schizophrenia susceptibility: from linkage to genome-wide association studies

William Gilks
Department of Psychiatry, Trinity College Dublin
A thesis submitted to the University of Dublin for the degree of Doctor of Philosophy
29th October 2010

Schizophrenia is a common and severe mental illness which is highly heritable. Susceptibility to schizophrenia is likely to be caused by multiple genetic variants in combination with environmental risk factors. The identification of such genetic variants would greatly improve diagnosis and treatment.

The first aim of this study was to identify common susceptibility variants for schizophrenia using a candidate gene case-control association study design. Nine genes were selected based upon positional (linkage data) and functional criteria (*ARHGEF11*, *CNDP1*, *GGCX*, *GLUL*, *HOMER2*, *HTRA2*, *MAPT*, *PEA15* and *SYNGAP1*). Sixty-six tagging SNPs were tested in 375 schizophrenia cases and 812 controls from Ireland. Significant associations were identified at five SNPs in three of the genes (*CNDP1*, *HOMER2*, *SYNGAP1*).

The second aim was to follow-up significant association findings at *HOMER2* with fine-mapping, functional studies and independent replication. The strongest associated SNP at *HOMER2* (rs2306428; p=0.006, OR=0.61) which was predicted to affect splicing. Replication analysis in the International Schizophrenia Consortium (ISC) sample (1,287 cases and 1,129 controls) was supportive of a role for *rs2306428* in schizophrenia susceptibility (p=0.019, OR=0.77). No relationship was identified between genotype of rs2306428 and transcript levels in lymphoblast cells, nor on indices of neurocognitive function which were tested under the endophenotype hypothesis.

The third aim was to investigate HOMER2 interactants in schizophrenia susceptibility using results of the ISC genome-wide association study (GWAS). A SNP at the *CDC42* gene, rs2473227 was highly-associated with schizophrenia ($p=2.00\times10^{-6}$, OR=0.85), in high linkage disequilibrium with a predicted regulatory SNP rs2473307, and correlated with CDC42 mRNA level in two independent studies. Gene reporter assay of rs2473307 indicated that the susceptibility allele reduced expression in a neuronal cell lines. However, this could not be confirmed using an electrophoretic mobility shift assay because binding of transcription factors to either allele was not observed. The heterozygous genotype at rs2473277 was associated with indices of working and episodic memory ($p<0.05$). Further available GWAS data did not replicate the association with schizophrenia at this SNP.

The fourth aim was to investigate the ability of next-generation sequencing to detect copy-number variants (CNV). Sixteen test CNVs were selected from publicly-available data which were verified by population genetic parameters. The true break-points of the selected CNVs were determined by Sanger sequencing. The CNV detection algorithm developed in-house accurately predicted the existence and break-points of many of the CNVs in targeted resequencing data although its resolution appeared to be affected by the size of the CNV and GC-content of the target region.

Since this study was started there have been massive changes in the way in which the genetics of common disease such as schizophrenia. Techniques such as genome-wide association studies, next-generation sequencing and high-throughput high-content analysis are generating vast amounts of data. Such experiments must be designed and analysed both imaginatively and with strict rigour if current and future investments are to deliver a better outcome for patients.