Elucidating functional pathways associated with schizophrenia risk through analysis of brain gene expression

Eilís Hannon

A thesis submitted to Cardiff University for the degree of

Doctor of Philosophy

School of Medicine & School of Computer Science & Informatics

Cardiff University

October 2013

Declaration and statements

Declaration

This work has not previously been accepted in substance for any degree and is not concurrently submitted in candidature for any degree.

Signed ……

Date ……..

Statement 1

This thesis is being submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy.

Signed ……

Date ……

Statement 2

This thesis is the result of my own independent investigations, except where otherwise stated. References are given where other sources are acknowledged. A bibliography is appended.

Signed ……

Date ……

Statement 3

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

Signed ……

Date ……

Summary

Schizophrenia is a highly heritable, common psychiatric disorder. Although onset generally occurs during adolescence, multiple lines of evidence point to a neurodevelopmental insult that occurs many years prior to the presentation of symptoms. Many different approaches have been used to elucidate the genetic risk factors and their impact; however, few unequivocal facts have been established. With a considerable amount of data publically available, integrative approaches look to leverage multiple data sources to identify coherent themes. This thesis investigates the neurodevelopmental hypothesis of schizophrenia by incorporating results from genome-wide association studies and copy number variation studies with gene expression datasets, with the overall aim of identifying functional pathways that may be disrupted in the aetiology of the disorder.

This study used foetal and developmental expression datasets of the human brain and statistical approaches to characterise the expression profiles of schizophrenia risk genes. Both spatial profiles in the mid-foetal brain and temporal profiles across development were considered. Data from genome-wide association studies and copy number variation studies were used to test for an enrichment of risk genes; in addition the genetic overlap with bipolar disorder identified through genome-wide association studies was used for validation. Gene sets with a common expression profile enriched for schizophrenia variants were used to identify biological pathways and assessed for their polygenic contribution to schizophrenia risk.

The results of this thesis converged on a common developmental expression profile for schizophrenia risk genes. Genes identified with this profile were shown to harbour multiple, common risk variants for schizophrenia and were implicated in epigenetic processes relating to the regulation of gene transcription. Together this suggests that schizophrenia associated genes are involved in brain development, particularly during foetal stages, and may play a role in the regulation of this process.

Table of Contents

List of Figures

List of Tables

Abbreviations

ADHD - attention deficit hyperactivity disorder

ANOVA - analysis of variance

ARC - activity-regulated cytoskeleton

BPD - bipolar disorder

CNS - central nervous system

CNV - copy number variant

DNA - deoxyribonucleic acid

GABA - gamma-aminobutyric acid

GCTA - genome-wide complex trait analysis

GEO - Gene Expression Omnibus

GO - Gene Ontology

GWAS - genome-wide association study

ISC - International Schizophrenia Consortium

LD - linkage disequilibrium

MAF - minor allele frequency

Mon - months

MGS - Molecular Genetics of Schizophrenia

MHC - major histocompatibility complex

mRNA - messenger ribonucleic acid

NCBI - National Center for Biotechnology Information

NMDAR - N-methyl-D-aspartate receptor

PCW- post conception weeks

PGC - Psychiatric Genomics Consortium

PMI - post-mortem interval

RIN - RNA integrity number

RMA - robust multi-array average

RNA - ribonucleic acid

SCZ - schizophrenia

SE - standard error

SNP - single nucleotide polymorphism

SNV - single nucleotide variant

Acknowledgments

Firstly, I would like to thank my supervisors, Dr Andrew Pocklington, Prof. Peter Holmans, Dr Richard White and Dr Andrew Jones for giving me the opportunity to work towards this PhD. I am proud of the journey I have taken over the last three years and appreciate the advice and guidance given to me throughout. This has enabled me to develop a skillset, which will equip and serve me well as I continue to develop throughout my career in research. I very much enjoyed and benefited from the research environment at Cardiff and am grateful to all who have contributed to that. In particular, I wish to thank all members of the BBU and the schizophrenia and psychosis teams for their day to day contributions, help, and entertainment.

I am eternally grateful to my family, Dad, Holly, Grandma and Granddad you seen me through all of my academic achievements so far. Your confidence and belief in me pushes me ever forward and I continue to aspire to make you all proud. Phoebe and Phil, thank you for your patience, friendship and enthusiasm while we lived together. I am grateful for all the trips/nights out, exercise classes and dinners we have done together over the years, providing the much needed day to day support and down-time that allowed me to reach my potential. Then follows a long list of friends; your interest in my studies and confidence in my ability have inspired me to keep going and ultimately get to this point. I look forward to celebrating with you all.

Chapter 1: Introduction

1.1 Introduction to disorders

1.1.1 Schizophrenia

Schizophrenia (SCZ) is a complex psychiatric disorder with a median lifetime prevalence estimate of 0.4% and a median lifetime morbid risk estimate of 0.7% (McGrath *et al.*, 2008). The primary diagnostic feature is an episode of psychosis characterised by hallucinations, delusions, disorganised thought or speech, with additional negative symptoms that affect normal functioning such as affective flattening, anhedonia (inability to experience pleasure) or avolition (lack of motivation), also required for diagnosis (Andreasen, 1995, Linden, 2011). Deficits in motor and cognitive domains (Heinrichs and Zakzanis, 1998, Dickinson *et al.*, 2007) are also often part of this heterogeneous disorder. This range of clinical features means psychiatrists can see a variety of different presentations, with potentially two patients having few common symptoms.

The onset of SCZ typically occurs during adolescence or early adulthood, generally occurring later in females (Hafner *et al.*, 1994). Incidence rates are higher in males compared to females with a relative risk of approximately 1.4 (Aleman *et al.*, 2003, McGrath *et al.*, 2008). Current diagnosis is based on matching observed or reported symptoms with descriptions provided by the World Health Organisation in the International Classification of Diseases (World Health Organization, 1993) or the American Psychiatric Association's Diagnostic and Statistical Manual of Mental Disorders (American Psychiatric Association, 2000, American Psychiatric Association, 2013), both of which have been revised over the years to reflect research findings.

1.1.2 Bipolar disorder

Bipolar disorder (BPD) is also a common psychiatric disorder with lifetime prevalence estimates ranging from 2-4.4% depending upon the clinical subtypes included

(Kessler *et al.*, 2005, Merikangas *et al.*, 2007, Kessler *et al.*, 2012). It is characterised by contrasting episodes of depression and mania, sometimes including psychotic symptoms (Goodwin and Jamison, 1990), with periods of remission. Similar to SCZ, onset generally occurs during early adulthood (Joyce, 1984) and is diagnosed using the aforementioned diagnostic manuals.

1.1.3 Schizophrenia and bipolar disorder as separate diagnostic entities

At the end of the 19th century German psychiatrist Emil Kraepelin first described the two diagnostic entities *dementia praecox* and manic depression (Kraepelin, 1899), which are considered the forerunners to SCZ and BPD. This dichotomy has since remained, but research findings are challenging such a format (Craddock and Owen, 2005).

An overlap in clinical features is not uncommon with SCZ patients presenting symptoms of depression (Zisook *et al.*, 1999, Majadas *et al.*, 2012) and the first-rank symptoms introduced by Schneider (Schneider, 1959), which encompass hallucinations and delusions, present in some BPD patients (Rosen *et al.*, 2011). Moreover, the idea that these disorders 'breed true' has been challenged through family and molecular genetic studies providing evidence for a shared genetic susceptibility (Lichtenstein *et al.*, 2009, Purcell *et al.*, 2009). The overlap of associated risk factors will be discussed further in the following sections.

1.2 Neurodevelopmental hypothesis

SCZ is classed as a neurodevelopmental disorder (Murray and Lewis, 1987, Weinberger, 1987) in which a disruption during the critical period of brain development plays a major role in the disease aetiology. This is unlikely to be sufficient to cause SCZ and may be combined with genetic predisposition or an additional insult during adolescence (Bayer *et al.*, 1999, Keshavan, 1999). BPD has typically been considered an adult disorder, although in light of evidence supporting a nosological overlap with SCZ a neurodevelopmental insult has been proposed as

part of BPD aetiology (Nasrallah, 1991). Currently, evidence for BPD as a neurodevelopmental disorder is inconsistent, with methodological issues limiting the ability to draw cohesive conclusions (Sanches *et al.*, 2008).

1.2.1 Human brain development

Human brain development is a sensitive sequence of processes, coordinated by genetic, epigenetic and environmental influences (Tau and Peterson, 2010). By the end of the third post conception week (PCW), the earliest neural progenitor cells have migrated to the neural plate signalling the start of brain development. The neural plate becomes the neural tube, which grows and changes from its cylindrical shape to eventually form the cerebral shape of the brain we are familiar with as well as the rest of the central nervous system (Stiles and Jernigan, 2010).

The establishment of the regions of the brain happens progressively by a process called neural patterning. The first subdivision forms the prosencephalon (forebrain), mesencephalon (midbrain) and rhombencephalon (hindbrain). By the end of the embryonic period (8^{th} gestational week) both the prosencephalon and rhombencephalon further subdivide into the telencephalon and diencephalon and the metencephalon and myelencephalon respectively. Neural patterning, which gives rise to the regional organisation of the central nervous system, is coordinated by the expression level of signalling molecules (Hoch *et al.*, 2009). These molecules prompt neural progenitors to differentiate into neurons appropriate for each region (Hebert and Fishell, 2008), a process which continues postnatally. The early stages of brain development are characterised by neurogenesis, migration and differentiation (Stiles and Jernigan, 2010).

While growth in total brain volume slows after early childhood (Dekaban, 1978, Brain Development Cooperative Group, 2012), the human brain still undergoes structural changes with maturational processes active up to and during adolescence. Structural magnetic resonance imaging has shown that white matter volume continues increasing throughout the second and third decade of life (Lenroot *et al.*, 2007,

Koolschijn and Crone, 2013). In contrast, grey matter volume reaches its maximum during childhood, although the exact time point varies between lobes, before decreasing through adolescence (Gogtay *et al.*, 2004, Lenroot *et al.*, 2007, Koolschijn and Crone, 2013).

White matter is composed of myelin, a fatty sheath that insulates axons to improve the transmission of impulses between neurons; therefore structural brain scans that identify changes in the volume of white matter reflect the accumulation and loss of myelination (Giedd, 2004). The increases of white matter are consistent with the knowledge that myelination continues into adolescence (Benes *et al.*, 1994). Moreover, increased white matter volume indicates more efficient communication across the brain (Paus, 2005). Myelination is a progressive event that is active during the same developmental stages as the regressive mechanism of synaptic pruning, both of which may play a role in these structural changes (Sowell *et al.*, 2001). Synaptic pruning, which occurs throughout childhood and adolescence (Huttenlocher, 1979), will also improve the efficiency of communication within the brain.

Given the common onset of SCZ during teenage years and early adulthood there has been much interest in this time period. As brain development continues for at least two decades, insults early on may have immediate effects or some latency before any impacts manifest. Aetiological hypotheses regarding adolescent neurodevelopment have been proposed such as faulty synaptic pruning (Feinberg, 1982). Further, there has been much focus on the prefrontal cortex as it is one of the last brain regions to mature during adolescence (Gogtay *et al.*, 2004) and is related to executive functions that are affected in SCZ (Bozikas *et al.*, 2006).

1.2.2 Evidence of neurodevelopmental antecedents for schizophrenia

Premorbid impairments

Support for a neurodevelopmental contribution to SCZ aetiology comes from many different studies. Birth cohort studies, as well as other study designs, find childhood

impairments in motor (Jones *et al.*, 1994, Cannon *et al.*, 2002a, Clarke *et al.*, 2011, Dickson *et al.*, 2012), social (Jones *et al.*, 1994, Bearden *et al.*, 2000, Schiffman *et al.*, 2004) and cognitive (Jones *et al.*, 1994, Cannon *et al.*, 2002a, Dickson *et al.*, 2012) domains amongst those who go on to develop the disorder. These could be potential indictors of those at higher risk of developing SCZ with aberrant brain development leading to both these premorbid features and SCZ symptoms later in life.

As the name may suggest, Kraepelin's initial descriptions of *dementia praecox* included deterioration of cognitive function (Kraepelin, 1899). In later revisions, Kraepelin himself began to recognise that functional recovery did occur in some patients (Kraepelin, 1919). This deterioration aspect of SCZ has been a contentious issue with the most recent meta-analysis finding no evidence for such a decline during disease onset, further supporting a developmental insult giving rise to SCZ symptoms (Bora and Murray, 2013).

Pre- and perinatal events

Complications during pregnancy and delivery are proposed as risk factors for SCZ that may have a negative impact on brain development. Studies have found increased rates of specific complications in those who go on to develop SCZ including low Apgar scores, a scale used immediately after birth to assess the health of newborn babies (Cannon *et al.*, 2002a), small for gestational age (Cannon *et al.*, 2002a), preeclampsia (Cannon *et al.*, 2002b, Byrne *et al.*, 2007), hypoxia or ischaemia (Zornberg *et al.*, 2000, Cannon *et al.*, 2002a) and low birth weight (Cannon *et al.*, 2002b) amongst others. These findings are primarily based on prospective population-based samples which, given a lifetime prevalence of 0.4% (McGrath *et al.*, 2008) require large cohorts to obtain even a handful of SCZ cases. With a small number of cases power is then limited to detect the small effect sizes (odds ratio < 2) that are often associated with these risk factors (Cannon *et al.*, 2002b). In addition, simultaneously investigating many different individual complications increases the multiple testing burden, and therefore the chances of a spurious association if the significance level is not adjusted to reflect this. A combination of these issues means

that robust replicated findings for specific obstetric complications as yet have not been identified but a general increased rate has been reported (Cannon *et al.*, 2002a).

Exposure in the womb to viruses or infections is one specific insult which may impact on brain development that has received much attention. A recent meta-analysis supported previous reports of an increased rate of SCZ in offspring whose mothers were exposed to infectious agents including Human herpesvirus 2, *Toxoplasma gondii,* Borna disease virus, or Human endogenous retrovirus-W during pregnancy (Arias *et al.*, 2012). The association for Borna disease virus was based on the largest sample with over 2000 cases and controls, followed by the samples for Human herpesvirus 2 and *Toxoplasma gondii,* the latter of which had evidence of a publication bias. However, given that no correction was applied to take into account the multiple exposures that were tested, further replication would be required for each of these exposures. Additionally, larger samples would allow a more accurate estimate of the risk associated with these exposures.

Population-based cohort studies have reported similar increased risk of SCZ in the offspring of mothers hospitalised for infection (Nielsen *et al.*, 2013), or diagnosed with a bacterial (Sorensen *et al.*, 2009) or respiratory infection (Brown *et al.*, 2000a) during pregnancy. As with the obstetric complications these need to have large sample sizes to ensure enough SCZ cases, with all of these studies having at least 75 affected individuals. Studies of the impact of maternal illness have attempted to narrow down the period of increased risk, leading to suggestions that the earliest stages of pregnancy may be the most vulnerable (Khandaker *et al.*, 2013).

Famine studies look at nutritional deficiency during pregnancy and the impact this may have on developing illness later in life. These are uncommon events but one such study showed that those conceived during the lowest point of the Dutch Hunger Winter famine during 1945 had an increased risk of developing SCZ (Susser *et al.*, 1996, Hoek *et al.*, 1998). A similar study of the Chinese Great Leap Forward famine also found that people born during the latter stages, from 1959-1961, and

therefore conceived during the famine, also had a greater risk of SCZ (St Clair *et al.*, 2005). In both studies risk was estimated from birth rates, death rates and SCZ diagnoses from psychiatric hospital records of large geographical regions with minimal migration, and therefore had very considerable sample sizes. Famine is a broad exposure and can encompass a range of deficiencies, including particular nutrients or vitamins, as well as other factors such as maternal stress, which may have an impact, making it challenging to identify how exposure to famine may affect brain development and ultimately SCZ risk.

One potentially more revealing approach is the study of minor physical anomalies, such as cleft palate, low-seated ears or furrowed tongue. As these characteristics are established during gestation, they are proposed to be indicative of disruptions during the period of early brain development (Lobato *et al.*, 2001). SCZ patients have been shown in meta-analyses to have an increased number of minor physical anomalies (Weinberg *et al.*, 2007, Xu *et al.*, 2011). Both of these studies obtained total samples of over 1000 cases from 11 and 14 studies respectively, which suggest that individual studies to date have had quite small sample sizes. Within these, secondary analyses of specific anomalies or anomalies affecting a particular area of the body were sometimes also tested but the number of studies and hence total sample size for these was further reduced. Therefore interpretation of these results requires additional caution.

A similar idea is behind the study of dermatoglyphic anomalies, as they are also established during foetal development and thought to be markers of insults during this timeframe (Lobato *et al.*, 2001). As with minor physical anomalies dermatoglyphic abnormalities are more common in SCZ patients compared to controls (Bramon *et al.*, 2005). This includes evidence for two different dermatoglyphic abnormalities which implicate different developmental time points, suggesting that disruptions potentially causing SCZ could occur at multiple points of foetal development (Golembo-Smith *et al.*, 2012). These findings were also based on meta-analyses of many small studies producing combined samples of sizes similar to

those for the studies of minor physical anomalies and hence further studies should look to confirm these findings.

Although lots of studies have been performed investigating the role of pre- and perinatal risk factors, the majority of these were based on small samples (< 100 cases or controls). Meta-analyses look to synthesize the existing evidence but even when combined, the samples amount to less than 2000 cases or controls. While further evidence is required to strengthen the association for each specific risk factor and to accurately quantify the increased risk, as a group they indicate that disruptions during foetal brain development play a role in the development of SCZ.

These insults are estimated to have small effect sizes (generally ≤ 2), and therefore as the risk to the general population for SCZ is low these will only increase the risk of SCZ by an additional few per cent. In addition, there will be many others who experience the same complications but do not develop SCZ (Clarke *et al.*, 2006) so they have limited predictive power (Lewis and Levitt, 2002). Each factor, therefore, contributes to the risk but is not enough to cause it outright; a combination of environmental and genetic factors play a part in an individual's liability. Genetic risk factors will be discussed further in Section 1.3.

1.2.3 Evidence of neurodevelopmental antecedents for bipolar disorder

For BPD the evidence for a neurodevelopment insult is inconsistent. Early studies did not find evidence for premorbid impairments in those that went on to develop BPD (Zammit *et al.*, 2004, Reichenberg *et al.*, 2005). More recent studies found lower cognitive scores in those who went on to develop BPD but these were not significantly worse than controls (Osler *et al.*, 2007, Seidman *et al.*, 2012). These studies did find a significant difference between SCZ and healthy controls but the number of individuals in the SCZ group was greater than that in the BPD group, meaning there was reduced power to detect a difference between BPD and controls.

Two studies found an association between obstetric complications and BPD although both were based on a combined sample of less than 50 (Kinney *et al.*, 1993, Kinney *et al.*, 1998), hence a subsequent meta-analysis of eight studies with over 500 individuals was not significant (Scott *et al.*, 2006). Very little has been reported for prenatal exposure to infection for those who go on to develop BPD, with mostly negative findings from modest sample sizes that would not have been able to detect small effects (Machon *et al.*, 1997, Stober *et al.*, 1997, Mortensen *et al.*, 2011). Rates of BPD were found to be higher in those exposed prenatally to the Dutch Hunger Winter (Brown *et al.*, 2000b) although this was based on hospitalisation records and therefore could reflect a more severely affected subgroup, which may be more similar to SCZ.

There is also a paucity of studies looking at the occurrence of minor physical anomalies in BPD groups with early studies not finding any association (Green *et al.*, 1994, Trixler *et al.*, 2001). The most recent reported a significantly higher incidence of anomalies affecting the mouth, feet and head, which although the largest to date, was based on less than 200 subjects (Akabaliev *et al.*, 2011). In sum, the evidence reported so far is inconclusive as to whether BPD can be considered a neurodevelopmental disorder, primarily due to a lack of studies with adequate sample sizes. Therefore a comparison with SCZ pre- or perinatal risk factors at this stage would be inappropriate.

1.3 Genetics of schizophrenia and bipolar disorder

1.3.1 Heritability and family studies

Heritability is currently estimated from twin studies at about 80-85% for SCZ (Cardno and Gottesman, 2000, Sullivan *et al.*, 2003) and marginally higher at 85-93% for BPD (McGuffin *et al.*, 2003, Kieseppa *et al.*, 2004). A large scale study taking advantage of the Swedish registry system showed that families where the proband had SCZ had an increased risk for both SCZ and BPD, with a similar finding in families where the proband had BPD (Lichtenstein *et al.*, 2009). Full siblings had a relative risk of 9.0 for SCZ and 3.7 for BPD if their sibling had SCZ, whereas if their sibling had BPD the

relative risk was 7.9 for BPD and 3.9 for SCZ. While this supports a shared genetic component and discredits the idea that these disorders breed true, they also showed that there are unique genetic factors for each disorder.

1.3.2 Association studies

Earliest genetic studies: linkage and candidate gene studies

The search for associated genes and variants for SCZ and BPD has followed a similar course. Linkage studies were one of the first analysis tools used to identify chromosomal regions associated with disease (Teare and Barrett, 2005). Using a family pedigree, linkage analysis looks for regions with the same genetic variation present in affected family members and not present in unaffected family members. The test is based on the probability of loci within these regions segregating together to identify regions where this is highly unlikely to have happened by chance.

A limitation of this approach is that it only indicates broad regions of chromosomes and not specific genes, with one review summarising that the combined results of two meta-analyses for SCZ implicates around 4000 genes (Tandon *et al.*, 2008). Another limitation, perhaps the most important for psychiatric disorders, is that it is not appropriate to identify variants with small effect sizes (Risch and Merikangas, 1996). While regions have been identified through this method for both SCZ and BPD (Badner and Gershon, 2002, Ng *et al.*, 2009), the replication of these has been limited and no specific genes have been identified.

Candidate gene studies were the next step, which looked at variants within genes hypothesised as relevant for either SCZ or BPD. These looked at single nucleotide polymorphisms (SNPs), which are single deoxyribonucleic acid (DNA) base changes that vary between individuals. If these differences occur in coding sequences they can alter the resulting amino acid sequence, which may have functional consequences. Alternatively, if they are found in non-coding regions they may affect regulatory processes such as transcription factor binding and ultimately gene expression. Such variants are associated to disease generally through statistical tests

that compare the allele frequencies between affected and unaffected individuals looking for significant differences.

Most of the candidate genes considered were based on drug mechanisms that were successful at reducing symptoms such as dopamine receptor genes, which are the targets of antipsychotics. Genes involved in other neurotransmitter systems or brain development, based on the neurodevelopmental hypothesis, were also considered (Linden, 2011). Serotoninergic genes were amongst the most studied for BPD (Seifuddin *et al.*, 2012), in addition to genes involved in the dopaminergic, glutamatergic and gamma-aminobutyric acid (GABA) pathways. However, regardless of the biological plausibility of these genes none of these studies were fruitful in terms of associating with high levels of confidence specific mutations with SCZ (Allen *et al.*, 2008) or BPD (Seifuddin *et al.*, 2012). The limitations of these studies included sample size, which meant there was only power to detect variants with large effect sizes, and a lack of statistical rigour (Chanock *et al.*, 2007). As a result a large body of conflicting studies were reported with limited information about either SCZ or BPD.

Genome-wide association studies: single locus analyses

As for many areas of medicine, the sequencing of the human genome (Lander *et al.*, 2001, Venter *et al.*, 2001) was heralded as the start of a new and more productive era in psychiatric genetics (Corvin and Gill, 2003). Candidate gene studies had introduced a bias to the literature as only genes with a prior hypothesis, based on varying types and strength of evidence, were investigated. Genome-wide association studies (GWAS) allow the investigation of in excess of a million SNPs at sites across the genome in a single study (Beaudet and Belmont, 2008). The main benefit of this is that no prior hypotheses are required and potentially novel findings in genes not previously considered may be unearthed.

GWAS are designed to identify common (> 1%) variants that are associated with an increased risk to complex diseases (Bush and Moore, 2012) and are more powerful to detect markers of small effect compared to linkage studies (Risch and Merikangas,

1996). They have been made possible not only through the sequencing of the human genome but also through the characterisation of linkage disequilibrium (LD) between SNPs (International HapMap Consortium, 2005). This knowledge means that only a subset of SNPs, called tag SNPs, need to be directly genotyped or tested in order to look for association across the genome (Hirschhorn and Daly, 2005). A by-product of this is that it is unlikely that any associated SNP identified in a GWAS will be the causal marker, more that they tag the true functional variant.

Like candidate gene studies, statistical tests compare the allele frequencies for each SNP separately between cases and controls to see if they are significantly different (Hirschhorn and Daly, 2005). Power to detect an association of a true disease variant is related to both the frequency and effect size of the SNP, being lower for rarer alleles and smaller effect sizes (Klein, 2007, Spencer *et al.*, 2009). GWAS are not appropriate for looking at very rare SNPs as these cannot be adequately tagged.

While the ability to study hundreds of thousands of variants simultaneously is very efficient, it introduces the caveat of multiple testing. Neighbouring loci are known not to be independent, so a Bonferroni correction to prevent false positives based on the number of SNPs tested would be too severe (Pe'er *et al.*, 2008). Based on a UK Caucasian sample, across the autosomes it has been estimated that there are effectively 693,138 independent tests equating to a genome-wide significant threshold of 7.2×10^{-8} (Dudbridge and Gusnanto, 2008). This is in line with the HapMap Consortium's estimate of 150 effective tests per 500kb of the genome, equivalent to 900,000 tests (human genome assumed to be 3000Mb) and a p value threshold of 5.5×10^{-8} (International HapMap Consortium, 2005). Generally a p value $< 5 \times 10^{-8}$ is accepted as a genome-wide significant result.

Initial GWAS for schizophrenia and bipolar disorder

Despite much early enthusiasm with many studies, small by today's standards, being performed, it is only in the last few years that robust associations have been reported. The first genome-wide significant variants for SCZ were linked to *ZNF804A*

(Williams *et al.*, 2011b), *NRGN* (Stefansson *et al.*, 2009, Steinberg *et al.*, 2011), *TCF4* (Stefansson *et al.*, 2009, Steinberg *et al.*, 2011), *VRK2* (Steinberg *et al.*, 2011) and the extensive major histocompatibility complex (MHC) region on chromosome 6 (Purcell *et al.*, 2009, Shi *et al.*, 2009, Stefansson *et al.*, 2009). Comparable results were found in BPD GWAS for *DGKH* (Baum *et al.*, 2008), *ANK3* (Ferreira *et al.*, 2008), *CACNA1C* (Ferreira *et al.*, 2008) and *NCAN* (Cichon *et al.*, 2011).

Even in combination these genome-wide significant results only explained a small part of the genetic heritability of both disorders, and like many other complex diseases the majority was still unaccounted for (Manolio *et al.*, 2009). A susceptibility model containing a few common variants of large effect (genotypic relative risk > 1.3) was effectively ruled out of SCZ aetiology, as it was demonstrated that there was sufficient power in existing studies to identify these if they existed (Shi *et al.*, 2009). Although alternative genetic models involving rare alleles were also proposed (McClellan *et al.*, 2007), GWAS were now focused on identifying many risk alleles, implicating many genes, with small effect sizes (~1.1) that only make a small contribution to an individual's risk. This polygenic architecture of common variants was successfully demonstrated using the results of the International Schizophrenia Consortium (ISC) GWAS for genetic prediction. In a completely independent dataset, scores based on the number of risk alleles identified in the ISC GWAS, including those below the threshold for genome-wide significance, were found to be significantly higher in cases compared to controls (Purcell *et al.*, 2009). This demonstrated that common variants tagged in current studies do play a part in the genetics of SCZ (Wray and Visscher, 2010). The majority of these variants have not yet reached genome-wide significance due to limited power, but could in principle be unearthed in larger sample sizes (Sullivan *et al.*, 2012).

Consortium GWAS

For both disorders international collaboration through the Psychiatric Genomics Consortium (PGC) has brought together samples from across the world, primarily of European ancestry, which were analysed in the largest studies at that time. In 2011,

results were published from a stage 1 sample of 9394 SCZ cases and 12462 controls and an independent stage 2 of 8442 SCZ cases and 21397 controls. Seven loci were significant at a genome-wide threshold from a meta-analysis of both stages. Five of these were novel associations, located in or nearest to *MIR137*, *PCGEM1, CSMD1*, *MMP16*, and *CNNM2* or *NT5C2*, in addition to further support for the MHC region on chromosome 6 and a region containing *TCF4* and *CCDC68* (Ripke *et al.*, 2011). The BPD sample comprised of 7481 cases and 9250 controls in stage 1 with 4496 independent cases and 42422 controls used for replication and identified a novel locus in *ODZ4* in addition to further support for *CACNA1C* (Sklar *et al.*, 2011).

Since these publications, meta-analyses of the PGC SCZ study with additional samples has identified a further 14 novel loci including *SDCCAG8* (Hamshere *et al.*, 2013), *MAD1L1*, *TSNARE1*, *AKT*, and *FONG* (Ripke *et al.*, 2013), as well support for *CACNA1C* and the *ITIH3/4* region (Hamshere *et al.*, 2013, Ripke *et al.*, 2013). A meta-analysis of the BPD PGC sample with an independent sample has also obtained genome-wide significance for a SNP in *TRPC4AP* and a region on chromosome 12 between *RHEBL1* and *DHH* (Green *et al.*, 2012). Genome-wide significance has also been reported for SNPs in *TRANK1*, *LMAN2L* and *PTGFR* from a meta-analysis of a European BPD sample and a small Asian BPD sample (Chen *et al.*, 2013a).

Estimates of SNP heritability

Secondary analyses of GWAS data include quantifying how much heritability is accounted for by the common SNPs investigated with current microarray technology, referred to as the SNP heritability or chip heritability. The ISC found that their polygenic genetic prediction procedure using their dataset as the discovery sample from which risk alleles were identified, and predicting in the independent Molecular Genetics of Schizophrenia (MGS) dataset could explain approximately 3% of the variance for SCZ risk. By simulating samples approximately similar to the true ISC and MGS data for a range of genetic models they estimated that at least one third of the genetic variance could be explained by common variants found on genotype chips (Purcell *et al.*, 2009).

An alternative methodology which is part of the genome-wide complex trait analysis (GCTA) toolkit (Yang *et al.*, 2011), based on a linear mixed model, estimated that 23% of the liability variance could be attributed to SNPs using the PGC SCZ data (Lee *et al.*, 2012a). This method has also been applied to BPD PGC data where the SNP heritability was calculated to be 25% (Lee *et al.*, 2013). Both of these estimates for SNP heritability further imply that common variants play a sizeable role in SCZ and BPD aetiology and increasing the sample size for GWAS will help to identify them (Lee *et al.*, 2012a).

Approximate Bayesian methods have been developed as an extension to Purcell *et al.*'s polygenic approach to estimate the proportion of variance attributable to common variants as well as estimate the number of SNPs involved and the distribution of their effect sizes and frequencies (Stahl *et al.*, 2012). This approach produced the largest estimate of SNP heritability at 43% (assuming population risk estimate of 0.01) and further, predicted that SCZ would have 8300 independent SNPs that account for 50% of the variance in liability (Ripke *et al.*, 2013).

Evidence of genetic overlap between schizophrenia and bipolar disorder and estimating genetic correlation

Molecular evidence of overlapping genetic associations for SCZ and BPD is growing, particularly for common variants. Prior to the publication of the PGC studies, GWAS datasets had been used to show that an overlap did exist and was present among SNPs tagged in current studies. The ISC tested their polygenic score approach based on SCZ associated alleles in a BPD study and found that BPD cases, like SCZ cases, had higher scores compared to controls. Moreover, this was not the case for type one or type two diabetes, coronary artery disease, rheumatoid arthritis, Crohn's disease or hypertension sufferers, supporting an overlap in the common variants between SCZ and BPD (Purcell *et al.*, 2009). This analysis has been repeated with all of the PGC datasets, including attention deficit hyperactivity disorder (ADHD), autism and major depressive disorder, for all pairs of the five disorders. Overlap of common

variants was found between all disorders, however the most successful pairing used SCZ associated alleles to predict BPD status or vice versa (Smoller *et al.*, 2013).

An alternative approach again using the PGC GWAS datasets looked to quantify the shared genetic relationships between these five disorders. This methodology (Lee *et al.*, 2012b) was similar to the GCTA linear mixed model approach used to calculate the SNP based heritability of complex disorders and like the polygenic approach found the pairing of SCZ and BPD had the largest genetic correlation, estimated at 0.68, of all the pairs of PGC disorders (Lee *et al.*, 2013). This estimate of genetic correlation was consistent with that estimated from the Swedish national family study of 0.6 (Lichtenstein *et al.*, 2009).

Identifying overlapping common variants

Initial attempts to identify the common genetic variants between BPD and SCZ, took risk factors identified in one disorder and investigated them in the second, requiring a lower level of significance (one tailed $p < 0.05$) for evidence of an association. For example, rs1006737 found within *CACNA1C*, identified in a BPD GWAS (Ferreira *et al.*, 2008) also showed an association with SCZ (Green *et al.*, 2010). Further, SNPs from the MHC region and *NRGN* identified for SCZ have been shown to be associated with BPD, although in the same study variants from *TCF4* were not associated with BPD and no association was found between *ANK3* SNPs and SCZ (Williams *et al.*, 2011a).

To date, genome-wide significance in separate GWAS for both SCZ and BPD has been reported for SNPs in *NCAN* (Cichon *et al.*, 2011, Ripke *et al.*, 2013), and *CACNA1C* (Ferreira *et al.*, 2008, Sklar *et al.*, 2011, Hamshere *et al.*, 2013, Ripke *et al.*, 2013). Combined analyses, where both SCZ and BPD individuals are compared to healthy controls, can also be used to identify variants that contribute to the shared genetic component while increasing the power to detect those with smaller effect sizes. With this approach, genome-wide significance has been obtained for SNPs in *ZNF804A* (O'Donovan *et al.*, 2008), *ANK3* (Sklar *et al.*, 2011), and the *ITIH3-ITIH4*

region (Sklar *et al.*, 2011). A GWAS based on a psychosis phenotype including SCZ, BPD and related psychoses also found a novel genome-wide significant SNP in the 16p11.2 region which was associated with gene expression at *MAPK3* (Steinberg *et al.*, 2012).

The genetic overlap of common variants for SCZ and BPD has been used advantageously in conditional false discovery rate analyses which have increased power to detect SNPs associated to both disorders (Andreassen *et al.*, 2013). Using the PGC data this approach, at a false discovery rate of 0.05, identified 58 SCZ loci including 51 novel loci and 35 BPD loci, of which, 30 were novel. Further a conjunction p value, taken as the maximum of the two disorder p values, was used to identify 14 loci with pleiotropic effects for both disorders including genes previously identified such as *CACNA1C* and *ITIH4* as well as novel candidates such as *PPM1F* and *IFI44*.

Biological pathways associated with GWAS results

Given few markers have surpassed genome-wide significance it may be a little premature to look for common functions but one recurring theme of genes identified as top hits in SCZ or BPD GWAS are those relating to calcium channels such as *CACNA1C* or *CACNB2* (Ripke *et al.*, 2013). Both of these genes, along with others relating to calcium channel activity, have also been implicated in autism, ADHD and major depressive disorder through a combined GWAS of psychiatric disorders suggesting that this functional process may be disrupted in many disorders (Smoller *et al.*, 2013).

Gene-based tests

For genes to be associated through a GWAS generally they require a SNP located within or proximal to them to be significant genome-wide. Although the number is steadily increasing, so far few SNPs have been found at the genome-wide significance level compared to the large number of genes expected to be associated

to either SCZ or BPD. Gene-based tests look to combine SNPs located across a gene to see if as a group they confer risk and are likely to be an informative tool (Neale and Sham, 2004).

Looking for significant genes as opposed to significant SNPs is a more powerful study design as it reduces the multiple testing burden (Li *et al.*, 2011) and multiple methods have been tested. One of the simplest approaches is to identify the smallest SNP p value, generally after applying a correction for the number of SNPs within that gene. Other methods look at combining SNP association p values such as the product of P, or the closely associated truncated product of P (Zaykin *et al.*, 2002). The presence of LD means that the individual SNP tests are not independent and permutations have been used to calculate empirical significance for each gene, however this has been found not to be a sufficient correction for the product of P approaches (Moskvina *et al.*, 2012).

An alternative methodology made use of Brown's method (Brown, 1975) for combining test statistics that are not independent. By incorporating the correlation between markers, permutations are not required to ascertain significance and hence this approach is very efficient to run (Moskvina *et al.*, 2011). Thus far the performance of each of these methods has rarely been evaluated, and despite being applied to real datasets these methods have yet to be used to investigate or identify novel genes for SCZ or BPD aetiology.

1.3.3 Copy number variation studies

Copy number variants (CNVs) are a common class of structural variant where large segments of the genome are deleted or duplicated, altering the number of copies of any genes within the affected region. The availability of raw genome-wide SNP intensity data meant that algorithms could be developed to detect CNVs so that they could be investigated in the context of human disease (McCarroll *et al.*, 2006, Redon *et al.*, 2006).

Copy number variants in schizophrenia

SCZ patients have been shown to have an increased number or burden of CNVs across the genome, which are generally larger and more likely to affect loci that are hit rarely in the general population when compared to CNVs found in controls (International Schizophrenia Consortium, 2008, Walsh *et al.*, 2008, Kirov *et al.*, 2009a). The effect sizes of CNVs are typically larger than those found in GWAS for SNPs and hence specific loci at which SCZ patients have a significantly increased rate of CNVs have been found and are displayed in Table 1.1. There are likely additional rarer CNV loci that will be identified as sample sizes increase (Malhotra and Sebat, 2012). Individuals with SCZ also suffer a higher number of *de novo* CNVs (Xu *et al.*, 2008, Malhotra *et al.*, 2011, Kirov *et al.*, 2012), new mutations that have arisen in the individuals' DNA that are not found in their parents, which may explain some non-familial cases of SCZ.

Deletion CNV loci	**Schizophrenia**	**Bipolar disorder**
1q21.1	(International Schizophrenia Consortium, 2008, Stefansson *et al.*, 2008, Kirov *et al.*, 2009a, Levinson *et al.*, 2011)	
2p16.3 (NRXN)	(Kirov *et al.*, 2009b, Levinson *et al.*, 2011)	
3q29	(Mulle *et al.*, 2010, Levinson *et al.*, 2011)	
15q11.2	(Stefansson *et al.*, 2008, Kirov *et al.*, 2009a)	
15q13.3	(International Schizophrenia Consortium, 2008, Stefansson *et al.*, 2008, Kirov *et al.*, 2009a, Levinson *et al.*, 2011, Vacic *et al.*, 2011)	
17q12	(Moreno-De-Luca *et al.*, 2010)	
22q11.2	(International Schizophrenia Consortium, 2008, Mulle *et al.*, 2010, Levinson *et al.*, 2011, Vacic *et al.*, 2011)	
Duplication CNV loci		
7q36.3 (VIPR2)	(Levinson *et al.*, 2011, Vacic *et al.*, 2011)	
16p11.2	(McCarthy *et al.*, 2009, Levinson *et al.*, 2011, Vacic *et al.*, 2011)	(McCarthy *et al.*, 2009)
16p13.1	(Ingason *et al.*, 2011)	

Table 1.1: CNV loci associated with schizophrenia or bipolar disorder.

Copy number variants in bipolar disorder

Findings for specific loci for BPD are much less frequent. Table 1.1 shows one locus, which has a significantly increased rate, a locus that is also associated with SCZ. Further, burden analyses for BPD patients have produced inconsistent results. Some report evidence for an increased frequency of rare CNVs compared to controls (Zhang *et al.*, 2009, Priebe *et al.*, 2012) whereas others find no difference (Grozeva *et al.*, 2010, McQuillin *et al.*, 2011). Similar to SCZ, a higher rate of *de novo* CNVs have been reported in BPD patients (Malhotra *et al.*, 2011). Where increased rates of CNVs have been reported for BPD, they were strongest in the subset of individuals with an earlier age of onset (Zhang *et al.*, 2009, Malhotra *et al.*, 2011, Priebe *et al.*, 2012). Therefore CNVs may play a smaller role in BPD aetiology than in SCZ and be specific to those with an earlier age of onset which is associated with a more severe subtype including more psychotic features, suicide attempts, rapid cycling and worse mania symptoms (Schurhoff *et al.*, 2000, Azorin *et al.*, 2013).

Biological pathways associated with CNVs

One of the early CNV findings was that genes hit by CNVs found in SCZ were overrepresented in neurodevelopmental pathways (Walsh *et al.*, 2008), supporting the hypothesis that an insult during early development plays a role in SCZ aetiology. Although the initial study did not control for the size of the genes hit by each CNV or the size of the CNVs themselves, a subsequent investigation controlling for these factors has shown that neuronal-activity genes were enriched in this dataset (Raychaudhuri *et al.*, 2010). Pathway analyses for genes hit by *de novo* CNVs have also reported enrichment in brain development categories as well synaptic genes (Malhotra *et al.*, 2011) in particular for the N-methyl-D-aspartate receptor (NMDAR) and activity-regulated cytoskeleton (ARC) postsynaptic complexes (Kirov *et al.*, 2012).

CNVs have been shown to play a role in other neurodevelopmental disorders including autism (Sebat *et al.*, 2007), ADHD (Williams *et al.*, 2010) and intellectual disability (Cooper *et al.*, 2011). Moreover, CNV loci have been shown to confer risk of multiple developmental disorders, for example 16p11.2 also confers risk for autism (Marshall *et al.*, 2008, Weiss *et al.*, 2008) and mental retardation (Ballif *et al.*, 2007, Ghebranious *et al.*, 2007). This suggests that CNVs are non-specific risk factors and that they may underlie some of the common phenotypes of these disorders such as cognitive impairments (Van Den Bossche *et al.*, 2012).

1.3.4 Rare variants: exome and whole genome sequencing

As discussed in Section 1.3.2, power to detect rare variants in GWAS is reduced compared to common variants, meaning that sequencing is the best approach to capture variants at the lower end of the frequency spectrum (Eberle *et al.*, 2007). Developments in technology and the shift to next-generation sequencing techniques have enabled the study of rare or moderately rare variants. Whole genome sequencing, which provides the entire sequence of an individual's DNA, is considered the gold standard but current prices make this prohibitive for the large sample sizes required to study rare variants (Cirulli and Goldstein, 2010). Although both the cost and sequencing time are falling rapidly, exome sequencing, which only looks at ~1% of the genome containing protein-coding regions (Teer and Mullikin, 2010), has been used as an interim solution.

Currently, no single nucleotide variants (SNVs) have been robustly associated with an increased risk for SCZ (Need *et al.*, 2012). Exome sequencing studies of trios (proband and both parents) have identified *de novo* SNVs in SCZ patients and reported higher nonsynonymous-to-synonymous (Xu *et al.*, 2012) and nonsense-to-missense ratios (Girard *et al.*, 2011), as well as an increased likelihood of carrying a mutation predicted as damaging (Gulsuner *et al.*, 2013) suggesting these variants are involved in the pathogenesis of SCZ. Although ongoing, no studies of SNVs have been published for BPD.

1.3.5 Gene expression studies

After identifying genes implicated in the aetiology of SCZ or BPD through GWAS or CNV studies, and likely in the future exome or whole genome sequencing, the functions or mechanisms of these genes need to be clarified, particularly in relation to disease pathology. Gene expression studies quantify the abundance of gene transcripts in a tissue or cell of interest. These studies are an intermediate between genotype and phenotype and may provide valuable information in identifying the mechanisms relevant to disease aetiology.

Microarray studies of schizophrenia post-mortem brains

Numerous studies have been undertaken comparing post-mortem SCZ brains to control brains. A variety of approaches have been used including real-time quantitative polymerase chain reaction (RT-qPCR) and in situ hybridisation (ISH), which while more accurate can only be used to look at a handful of genes. Microarray technology can be used to assay the whole transcriptome in a hypothesis free manner, to identify genes with significantly different expression levels in disease brains compared to control brains which may then be taken forward as candidate risk genes. Pathway analysis is generally also performed on these sets of candidate genes to infer potentially disrupted biological mechanisms for SCZ.

From the studies performed so far, there has been limited replication, particularly at the individual gene level, primarily due to methodological differences such as microarray platform, age of sample, cause of death, tissue dissected and statistical protocols (Sequeira *et al.*, 2012). Sample size is another issue as post-mortem brain samples are of limited availability (Mistry *et al.*, 2012) with few studies including more than 35 SCZ or control brains. More recent studies have combined data from existing studies to obtain sample sizes of over 100 brains (Mistry *et al.*, 2012, Perez-Santiago *et al.*, 2012), but this is still much smaller than those used for association studies and will be limited when trying to detect small changes.

Pathways implicated	Studies comparing schizophrenia post-mortem brains to controls	Studies comparing bipolar disorder post-mortem brains to controls
GABA	(Mirnics *et al.*, 2000, Glatt *et al.*, 2005) (Hakak *et al.*, 2001, Hashimoto *et al.*, 2008)	
Glutamate	(Mirnics *et al.*, 2000, Maycox *et al.*, 2009) (Bowden *et al.*, 2008)	
Immune response	(Saetre *et al.*, 2007, Shao and Vawter, 2008, Mistry *et al.*, 2012, Roussos *et al.*, 2012) (Arion *et al.*, 2007, Barnes *et al.*, 2011)	(Ryan *et al.*, 2006, Shao and Vawter, 2008)
Mitochondria and energy metabolism	(Mirnics *et al.*, 2000, Prabakaran *et al.*, 2004, Glatt *et al.*, 2005, Katsel *et al.*, 2005b, Mistry *et al.*, 2012) (Middleton *et al.*, 2002, Altar *et al.*, 2005, Iwamoto *et al.*, 2005, Khaitovich *et al.*, 2008)	(Konradi *et al.*, 2004) (Iwamoto *et al.*, 2005, Sun *et al.*, 2006)
Myelination and oligodendrocytes	(Katsel *et al.*, 2005b) (Hakak *et al.*, 2001, Tkachev *et al.*, 2003, Aston *et al.*, 2004, Sugai *et al.*, 2004)	(Tkachev *et al.*, 2003)
Neurogenesis, neurodevelopment	(Shao and Vawter, 2008, Maycox *et al.*, 2009) (Hakak *et al.*, 2001, Aston *et al.*, 2004, Bowden *et al.*, 2008)	(Nakatani *et al.*, 2006, Shao and Vawter, 2008)
Stress response	(*Choi *et al.*, 2008)	(Iwamoto *et al.*, 2004)
Synapse and signalling	(Mirnics *et al.*, 2000, Prabakaran *et al.*, 2004, Katsel *et al.*, 2005b, Maycox *et al.*, 2009, Perez-Santiago *et al.*, 2012, Roussos *et al.*, 2012) (Hakak *et al.*, 2001, Vawter *et al.*, 2001, Aston *et al.*, 2004, Altar *et al.*, 2005, Barnes *et al.*, 2011)	(Ryan *et al.*, 2006, Chen *et al.*, 2013b) (Iwamoto *et al.*, 2004)
Transcription and translation	(*Choi *et al.*, 2008, Roussos *et al.*, 2012) (Vawter *et al.*, 2001, Aston *et al.*, 2004)	(Iwamoto *et al.*, 2004)

Table 1.2: Functional pathways identified from genes differentially expressed between schizophrenia or bipolar disorder post-mortem brains and controls.
In each row first set of studies showed statistical overrepresentation of differentially expressed genes in a particular pathway, second set are additional studies that describe genes differentially expressed in these pathways. (*Choi *et al.*, 2008) study was based on a psychosis sample.

Another caveat to these studies is the presence and variability of medication exposure, both in terms of type and cumulative dosage, within the SCZ samples. Antipsychotics have already been shown to affect both proteins and metabolites in

SCZ human post-mortem brains (Guest *et al.*, 2010, Chan *et al.*, 2011) and alter gene expression in primate and rat brains (Healy and Meador-Woodruff, 1997, Schmitt *et al.*, 2003, O'Connor *et al.*, 2007). The effect of medication on gene expression studies could introduce both false positive or false negative results either by creating differences not primarily caused by SCZ pathology or by normalising disease-related changes so that they are not detected (Mistry *et al.*, 2012).

Despite all these limitations some broadly consistent themes have emerged when considering genes in functionally related groups shown in Table 1.2. Genes dysregulated between SCZ post-mortem brains and control brains include those involved in immune response pathways, mitochondria and energy metabolism, myelination and oligodendrocytes, neurogenesis and neurodevelopment, stress response, synapse and signalling, and transcription and translation. There is also support for disrupted neurotransmitter systems including glutamate and GABA.

The initial focus of gene expression studies for SCZ was on samples from the prefrontal cortex, in particular the dorsolateral prefrontal cortex. This was based on the rationale that it continues developing through to adolescence (Gogtay *et al.*, 2004, Lenroot and Giedd, 2006), the time point when SCZ typically presents, and impairments in executive functions, which are common in SCZ patients (Bozikas *et al.*, 2006) have been attributed to this region (Orellana and Slachevsky, 2013). Interestingly though, the number of expression differences found in this region is amongst the lowest (Katsel *et al.*, 2005a, Roussos *et al.*, 2012) and in fact the highest number of differences were found in the temporal regions (Katsel *et al.*, 2005a).

Microarray studies of bipolar disorder post-mortem brains

Fewer studies have considered BPD brains including a maximum of 35 post-mortem brains. These have found an overlap with SCZ studies in the pathways implicated including myelination and oligodendrocyte, mitochondria and energy metabolism, synapse and signalling and immune response; see Table 1.2.

Gene expression studies directly comparing expression profiles between the two disorders are infrequent. Despite one study finding only a few genes differentially expressed in SCZ and BPD that overlapped (Iwamoto *et al.*, 2004), a subsequent study found more overlapping than expected by chance (Shao and Vawter, 2008). While these were enriched for nervous system development, cell death and immune categories, only one set of controls was used.

Next-generation sequencing: RNA-Seq

Technological advances mean there is currently a shift to next-generation expression profiling in the form of RNA-Seq. Studies so far have supported the themes identified with microarrays such as neural development, mitochondrial function, synapse vesicle trafficking (Wu *et al.*, 2012) and inflammatory response (Fillman *et al.*, 2013) with genes in these pathways being dysregulated between SCZ and control brains. However, these were based on similar sample sizes to the initial microarray studies with 9 and 20 matched cases and controls respectively, therefore larger samples will be needed in future studies to validate these findings.

Temporal expression profiles of genes associated to schizophrenia

An alternative approach has considered the expression profiles of SCZ candidate genes throughout brain development of healthy individuals (Colantuoni *et al.*, 2008, Choi *et al.*, 2009, Harris *et al.*, 2009). All of these studies were before the publication of the SCZ PGC GWAS when only a handful of genes were robustly associated, which questions any reported findings for SCZ associated genes. There is, however, some overlap in the functional classes enriched for genes whose expression was associated with age and those enriched for genes differentially expressed between SCZ and control brains such as neurodevelopmental processes (Choi *et al.*, 2009, Harris *et al.*, 2009), synaptic activity (Mistry and Pavlidis, 2010), neurotransmitter systems (Harris *et al.*, 2009, Mistry and Pavlidis, 2010) and energy metabolism (Harris *et al.*, 2009). One transcriptomics study has found that the expression profiles of genes whose expression is associated with age can differentiate SCZ cases from controls

(Torkamani *et al.*, 2010). Therefore in light of successful GWAS studies these sorts of analyses should be repeated. Such studies have not been undertaken for BPD genes as far as I am aware.

Alternative splicing

Alternative splicing is a mechanism that produces different gene transcripts by including or excluding exons of a gene, affecting the proteins coded for and ultimately the gene's function. These alternatively spliced variants mean that the approximately 22,000 genes in the human genome can code for many more proteins. This process is estimated to occur in approximately 95% of human genes with more than one exon (Pan *et al.*, 2008). In most cases, it is tissue specific and in human adults occurs most often in the brain (Yeo *et al.*, 2004).

Splicing is known to be important for many processes during brain development including synaptogenesis, as well as affecting ion channel and neurotransmitter proteins in mature neurons (Li *et al.*, 2007). Therefore, disruption to this mechanism could be relevant to SCZ and BPD, as both calcium channel genes and neurotransmitter systems are thought to be involved in the pathogenesis. The vast majority of studies have only considered at most a couple of candidate splicing genes. Genes including *GRIN1* (Le Corre *et al.*, 2000), *GRM3* (Sartorius *et al.*, 2008), and *GABRB2* (Huntsman *et al.*, 1998, Zhao *et al.*, 2009) have been shown to be abnormally spliced between either SCZ or BPD and control post-mortem brains. Additional aberrantly spliced genes for SCZ include those related to brain development (Law *et al.*, 2006, Gibbons *et al.*, 2009).

A genome-wide study looked at two brain regions for SCZ, the prefrontal cortex and caudate head, for 20 SCZ samples and 20 control samples and identified 43 and 31 transcripts as alternatively spliced. Functional analysis found one biological pathway overrepresented for the 31 transcripts identified in the caudate head, 'Agrin in Postsynaptic Differentiation' (Cohen *et al.*, 2012). However due to the paucity of

studies in this area, it is hard to assess the reliability of these findings and the impact splicing has on psychiatric disorders.

1.3.6 Biological pathway analysis

Pathway analysis or functional analysis looks to identify the common processes of a set of genes related to a trait from association or transcriptome studies. They are commonly used as the final step of a genetic study to infer some meaning or interpretation of molecular results for the disease of interest. Many genes, but still not all, have been functionally characterised and categorised into pathways that describe the mechanisms or biological functions they are part of. Resources such as the Gene Ontology (GO) database (Ashburner *et al.*, 2000) or Kyoto Encyclopaedia of Genes and Genomes (KEGG) (Kanehisa, 1997, Kanehisa and Goto, 2000) provide this information so that researchers can integrate it with the results of their genetic studies.

The GO Consortium was established to formulate and organise a controlled, structured hierarchy of species independent annotation terms separated into molecular function, biological process and cellular component ontologies (Ashburner *et al.*, 2000) which over time has become the largest such resource (du Plessis *et al.*, 2011). Terms can be represented by a directed acyclic graph, which demonstrates the hierarchical relationships between parent and child terms (Gene Ontology Consortium, 2001).

Genes are associated to annotations through existing evidence from the literature or other databases. Each association has an evidence code that documents whether it was based on experimental evidence and the type of experiment (Gene Ontology Consortium, 2001) e.g. introducing a mutation to the gene or affecting the expression of the gene, or computational information such as sequence orthology or sequence alignment with genes that have highly confident functional information. Evidence codes are assigned by expert curators of the database who assess the

available evidence and therefore can be used as an indicator of the level of confidence in that association.

One evidence code exists for gene annotation associations that have not been manually curated and only exist as a result of automated annotation (Rhee *et al.*, 2008). Even though these constitute the vast majority (> 95%) of annotation terms in the database (du Plessis *et al.*, 2011, Khatri *et al.*, 2012) some researchers may wish to exclude such associations as the quality of the evidence has not yet be assessed. Although these are generally classed as unreliable, comparing successive versions of the GO database showed that many of these terms were subsequently verified to have experimental evidence, further, this proportion was in line with those manually curated and assessed to have computation evidence which were then promoted to have experimental evidence (Skunca *et al.*, 2012).

Commonly, gene sets of interest are tested to see if they overlap with a biological pathway more than would be expected by chance when compared to a background list. The choice of the background or reference list is important as it should reflect the genes involved in the experiment, not just all genes in the genome as this will introduce errors into the results (Rhee *et al.*, 2008). The limitation of this approach is that after selecting only the most significant genes, all members of this set are considered equal and assumed to be independent (Khatri *et al.*, 2012). More complex methods look to incorporate the effect size associated to each gene, by treating each pathway as a gene set for which a summary statistic reflecting the combined effect size and empirical significance through permutations can then be calculated (Subramanian *et al.*, 2005).

1.3.7 Integrating the 'omics

As discussed above many different molecular techniques have been used to investigate the biological causes of SCZ and BPD. Despite a lot of different datasets being published, there is still a paucity of clear facts for SCZ or BPD. In light of this, attempts have been made to integrate data from many different approaches

including GWAS, transcriptomics (including from multiple tissues and organisms), linkage, CNVs and animal models. So far this has involved two approaches, the first is a gene discovery methodology and the second is a validation exercise (Niculescu, 2013).

The first approach uses algorithms to combine information from different sources generally producing scores to indicate the strength of evidence across each data type which can then be used to rank genes (Patel *et al.*, 2010, Ayalew *et al.*, 2012, Zhao *et al.*, 2013). Validation of this approach is shown by genotyping SNPs within the genes with the strongest evidence across the sources and testing them an independent dataset (Patel *et al.*, 2010, Ayalew *et al.*, 2012, Zhao *et al.*, 2013). An alternative approach produced a network of genes with a strong likelihood of sharing the phenotype of interest (Gilman *et al.*, 2012). These lists of candidate genes can then be used to identify functional pathways and so far have found enrichment in processes relating to neurodevelopment (Gilman *et al.*, 2012) and glutamate receptor signalling (Ayalew *et al.*, 2012) consistent with gene expression studies.

This multifaceted approach is also being incorporated into genetic studies as data are first published. One such example is a CNV study that used annotations derived from proteomics data as the basis of pathway analyses (Kirov *et al.*, 2012) or recent sequencing studies that have incorporated expression and proteomics data to characterise the potential risk genes identified in order to validate and interpret their findings (Xu *et al.*, 2012, Gulsuner *et al.*, 2013).

Obviously the results of these investigations are completely reliant on the quality of the contributing data. Generally they will work best when all genes have been studied equally, although attempts have been made to avoid this 'popularity bias' by implementing a maximum score for each contributing factor (Ayalew *et al.*, 2012). Therefore as more high-quality data resources are published these combinatorial approaches may prove extremely informative in identifying common themes and mechanistic targets for treatment.

1.4 Aims and objectives of thesis

The overall aim of this thesis was to integrate analyses of neurodevelopmental gene expression data with GWAS and CNV results for SCZ and BPD to identify functional pathways. The research question addressed within this was whether genes whose expression showed either spatial or temporal variability were associated with SCZ or BPD.

1.5 Outline of subsequent chapters

Chapter 2 investigates spatial gene expression patterns and alternative splicing across the mid-foetal brain to identify associations for SCZ or BPD risk genes identified through GWAS and CNV studies. Gene expression characteristics associated with either disorder were then used to identify pathways from the GO database.

Chapter 3 looks at temporal expression profiles identified in an expression dataset covering the full scope of human brain development for SCZ and BPD associated genes. Similar to Chapter Two, functional analysis was performed for any expression profiles associated to SCZ or BPD.

Finally, Chapter 4 develops the framework of the polygenic model described in Section 1.3.2 before calculating polygenic scores based on the gene sets associated with SCZ from the expression work. These gene set polygenic scores were then tested to see if they predicted SCZ status and whether they were a better predictor than scores calculated across all SNPs.

Chapter 2: The expression of schizophrenia and bipolar disorder risk genes during human foetal brain development

2.1 Introduction

2.1.1 Background

The human brain relies on a complex series of molecular mechanisms and environmental inputs in order to fully mature. These developmental mechanisms start as early as two weeks post gestation (Stiles and Jernigan, 2010) and can be extremely vulnerable to disruptions. Such an insult during this period is posited to be a factor in the aetiology of both SCZ and BPD (Murray and Lewis, 1987, Weinberger, 1987, Nasrallah, 1991), and the evidence behind this neurodevelopmental hypothesis was discussed in more detail in the Introduction, Section 1.2. One part of this is an increased rate of minor physical anomalies in patients reported in both disorders (Akabaliev *et al.*, 2011, Xu *et al.*, 2011), although more consistently so for SCZ, which are suggestive of disruptions during foetal brain development (Lobato *et al.*, 2001). Insults during this time frame not only lead to the dysmorphogenesis of external features but may also predispose the individual to a psychiatric or mood disorder in later life (Guy *et al.*, 1983). Therefore this chapter was interested in whether expression patterns during foetal brain development are informative to the aetiology of SCZ and BPD.

In their study Johnson *et al.* investigated transcriptional patterns in developing human brains. They found that 76% of genes were expressed in at least one brain region in mid-foetal brains, of which 33% were differentially expressed and 28% showed evidence of alternative splicing across the five major brain structures considered: neocortex, cerebellum, hippocampus, striatum and thalamus. The set of differentially expressed genes were enriched for human-accelerated conserved noncoding sequences, defined as small regions of the genome with more human-specific substitutions than expected by chance (Prabhakar *et al.*, 2006). This

enrichment remained, although to a lesser degree, when excluding genes associated with chimpanzee-accelerated conserved noncoding sequences, suggesting that this set of genes may be responsible for human-specific attributes (Johnson *et al.*, 2009).

Despite SCZ being associated with reduced fecundity (Power *et al.*, 2013) the disorder is still present in today's society and population prevalence rates appear uniform across the globe (Jablensky *et al.*, 1992). As an explanation of these facts, SCZ has been proposed to be a consequence of human evolution (Crow, 1997). In line with this theory the mutations associated with the 'speciation events' that gave rise to the appearance of humans would also play a role in the development of the disorder. Further, given that the disorder is equally prevalent around the world (Jablensky *et al.*, 1992) the associated mutations must have arisen prior to the divergence of today's populations. Therefore, based on this hypothesis and Johnson *et al.*'s findings that genes differentially expressed in the human foetal brain were enriched for human-specific noncoding sequences, this chapter looked at risk genes for both SCZ and BPD, based on the strongest genetic evidence to date, and investigated whether they displayed any characteristic pattern of expression in the human foetal brain.

2.1.2 Outline

Aim

The research question considered in this chapter was whether genes associated with either SCZ or BPD had common expression profiles, either consistent or variable, in the human foetal brain.

Datasets

The primary dataset used in this chapter was that of the Johnson study (Johnson *et al.*, 2009) which contained expression data sampled from four human mid-foetal brains (18-23 weeks gestation). For each brain, samples were taken from thirteen different brain regions, listed in Table 2.1 with their abbreviations, from both

hemispheres. These data, referred to as the Johnson dataset, were obtained with the Affymetrix Human Exon 1.0 ST chip, which provides expression values at both the gene and exon level. Further details on data processing can be found at the end of this chapter in the Methods section. Unless referred to otherwise all analyses used the gene level expression values. In Johnson *et al.*'s original study they controlled for individual and hybridisation date differences, hence these variables were also included in this work.

	Full name	**Abbreviation**
Neocortex: prefrontal cortex	Dorsolateral prefrontal cortex	DLPFC
	Medial prefrontal cortex	MPFC
	Orbital prefrontal cortex	OPFC
	Ventrolateral prefrontal cortex	VLPFC
Neocortex: non-frontal regions	Motor-somatosensory cortex	MS
	Parietal association cortex	PAS
	Temporal auditory cortex	TAU
	Temporal association cortex	TAS
	Occipital visual neocortex	OCC
Hippocampus		HIP
Striatum		STR
Mediodorsal thalamus		THAL
Cerebellum		CBL

Table 2.1: Brain regions included in the Johnson dataset with abbreviations.

A second publically available microarray dataset from a study by Kang *et al.* was used for replication, where gene expression values were obtained for human brains covering the full range of development (Kang *et al.*, 2011). Like the Johnson dataset, multiple regions from both hemispheres were available and the same microarray chip was used to derive the expression values. Twelve regions overlapped with the Johnson dataset and these are identified in Table 2.2. Throughout the text the Johnson abbreviations will be used for common regions but readers can refer to Table 2.2 for the nomenclature used in the Kang manuscript.

Five independent individuals that fell in the same gestational period (18-23 weeks) as the Johnson dataset were extracted to form a replication dataset, referred to as the Kang dataset. The dataset downloaded for this work was already normalised,

Kang *et al.* having used the same software package as the Johnson study. In their study Kang *et al.* included covariates to adjust for sample RNA integrity (RIN) and post-mortem interval (PMI), therefore these were taken into account in this work also. In order to keep the analyses across the two datasets consistent, ideally the same covariates would have been used for both the Johnson and Kang dataset. This was not possible as the additional sample information provided with the datasets did not contain the relevant data. Johnson *et al.* included a covariate to control for individual differences based on the observation that after brain region, this factor contributed the next highest proportion of the variation in the expression data. This covariate could have been included in the analyses of the Kang dataset and will capture a variety of differences between the samples, including lifestyle differences which are hard to measure and factors relating to the post-mortem brain such as PMI. However, including it in the initial analyses may remove some the general variation we were interested in detecting.

	Full name	**Abbreviation**	**Johnson equivalent**
Neocortex: prefrontal cortex	Dorsolateral prefrontal cortex	DLPFC	DLPFC
	Medial prefrontal cortex	MPFC	MPFC
	Orbital prefrontal cortex	OPFC	OPFC
	Ventrolateral prefrontal cortex	VLPFC	VLPFC
Neocortex: non-frontal regions	Primary motor cortex	M1C	
	Primary sensory cortex	S1C	
	Primary auditory cortex	A1C	TAU
	Primary visual cortex	V1C	OCC
	Posterior inferior parietal cortex	IPC	PAR
	Superior temporal cortex	STC	TAS
	Inferior temporal cortex	ITC	
Hippocampus		HIP	HIP
Amygdala		AMY	
Striatum		STR	STR
Mediodorsal thalamus		THAL	THAL
Cerebellum		CBL	CBL

Table 2.2: Brain regions included in the Kang dataset with abbreviations and Johnson equivalents.

Analyses reported in this thesis were based on two different types of genetic variants. SNP association results were taken from the PGC studies, the largest GWAS

available for both SCZ (Ripke *et al.*, 2011) and BPD (Sklar *et al.*, 2011). The SNP p values for each PGC study were combined into two summary statistics for each gene. Firstly, gene-wide p values based on Brown's formula (Brown, 1975) which allows for correlations between SNPs were calculated and provided by V. Escott-Price (Moskvina *et al.*, 2011). Secondly, the p values of all SNPs within a gene were corrected for multiple testing using Simes' procedure (Simes, 1986) and the most significant one taken. Simes' method was developed to be less conservative than the Bonferroni method and was promoted for situations comprising of many highly correlated test statistics.

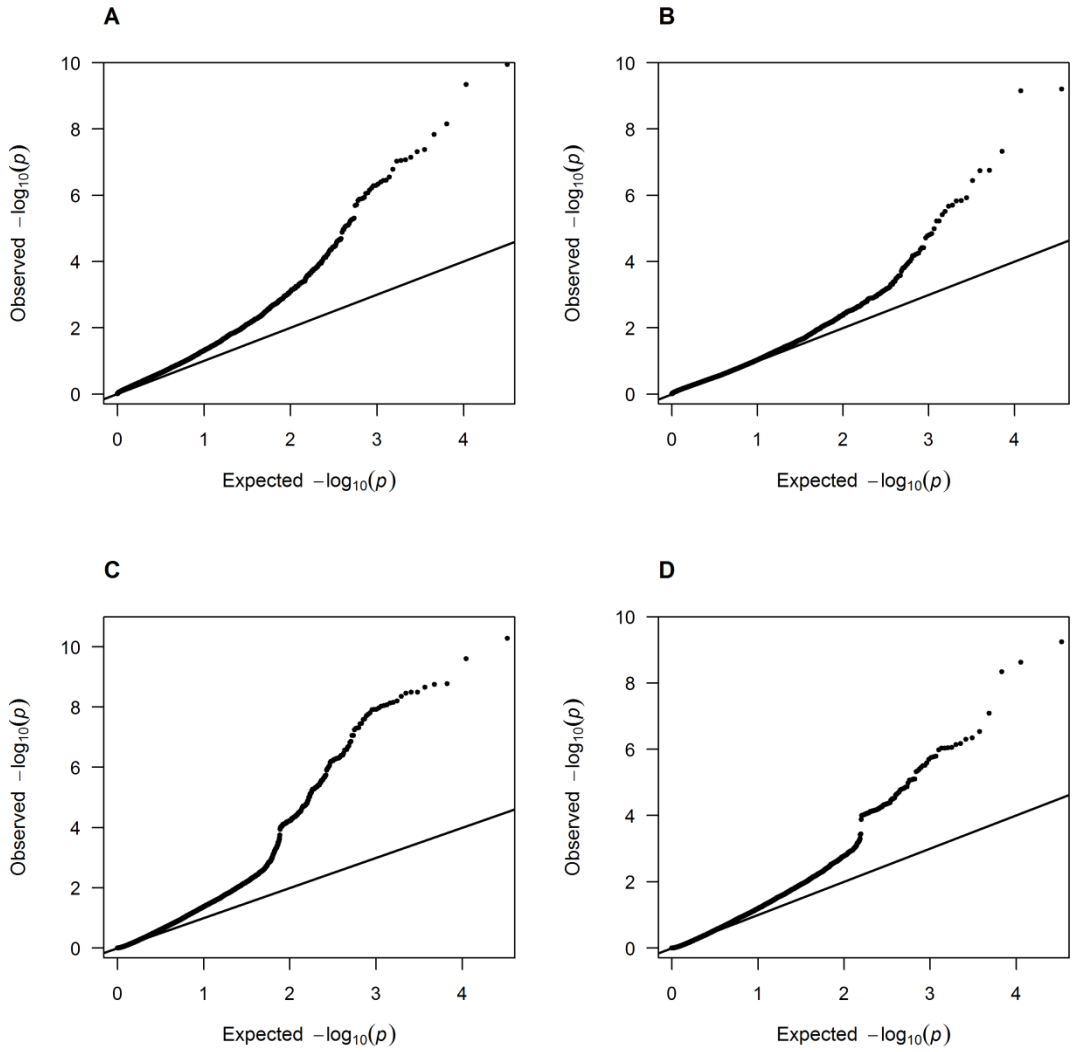

Figure 2.1: QQ plot to demonstrate distribution of gene-wide logP.
Panels A and B are QQ plots plotted with SCZ gene-wide logP; panels C and D are BPD gene-wide logP. Panels A and C are Brown' logP; panels B and D are Simes' logP.

Brown's p values are designed to take into account multiple independent SNP signals within a gene, whereas Simes' p values would be better if a gene has a single highly significant SNP result. While consistent enrichment across both types of p value was desirable, as they should both be capturing true association signal, stronger enrichments with Brown's p values might suggest that many of the genes harbour multiple semi-independent associated variants consistent with a polygenic model. Neither type of p value was biased by gene size, important for this study as brain genes are generally large (Raychaudhuri *et al.*, 2010). Including p values based on these two methods allows an assessment of the performance of each as well as an opportunity to establish the robustness of the results reported. All gene-wide p values were –log10 transformed and will be referred to as Brown's logP or Simes' logP. Figure 2.1 presents QQ plots for all four sets of gene-wide p values. These figures show that all four sets are inflated and pull away from the null line quite early on. It can also be seen that there is greater inflation in the SCZ logP compared to the BPD logP, compare panels A and C with B and D respectively, likely reflecting the larger sample size and greater power to detect small effects. In addition, the Brown's logP (panels A and B) show less inflation than the Simes' logP (panels C and D).

The genetic overlap of SCZ and BPD, particularly for common variants, is well documented and was discussed in the Introduction Section 1.3.2. The inclusion of BPD gene-wide p values in addition to SCZ gene-wide p values provides a level of genetic replication; however the PGC samples contained overlapping controls and hence were not entirely independent.

CNV data, solely for SCZ, from both the ISC (International Schizophrenia Consortium, 2008) and MGS (Levinson *et al.*, 2011) collaborations were also included. CNVs found in SCZ individuals were compared to see if they overlapped with any control CNV, those that did not were classed as singletons. Table 2.3 displays the counts of CNVs found in cases and controls as well as the number of deletions, duplications and singleton CNVs for both datasets. The ISC study was performed on two different chips, Affymetrix 5.0 and Affymetrix 6.0, whereas the MGS was done solely on Affymetrix 6.0, therefore chip differences were controlled for in addition to study

differences in all analyses. The numbers presented in Table 2.3 do not match those in the corresponding manuscripts, as the raw data were available in-house the size filter was relaxed by E. Rees to include all CNVs larger than 15kb. These CNVs were then annotated with the genes hit by A. Pocklington.

	All CNVs		CNVs that hit at least 1 gene	
	Cases	**Controls**	**Cases**	**Controls**
ISC	7288	7397	3499	3351
ISC: deletion	3693	3843	1372	1270
ISC: duplications	3595	3554	2127	2081
ISC: singletons	1081	n/a	601	n/a
ISC: singletons & deletions	524	n/a	234	n/a
ISC: singletons & duplications	557	n/a	367	n/a
MGS	4847	5366	2353	2549
MGS: deletion	2738	2964	1086	1111
MGS: duplications	2109	2402	1267	1438
MGS: singletons	969	n/a	513	n/a
MGS: singletons & deletions	516	n/a	235	n/a
MGS: singletons & duplications	453	n/a	278	n/a

Table 2.3: Counts of CNVs from ISC and MGS studies.

Outline of analysis

Various different expression profiles were identified in the Johnson and Kang datasets. Sets of genes with generally consistent or variable expression profiles across the mid-foetal brain were identified based on summary statistics for each gene and tested for an enrichment of SCZ and BPD associated variants. More specific variable profiles of genes characteristic of each brain region were also identified and tested in a similar manner. Each set of genes was tested for an enrichment of gene-wide p values based on common variants in the largest GWAS studies. In addition, genes with these expression profiles were tested to see if they were hit more frequently by CNVs found in SCZ individuals compared to those found in healthy controls, and within SCZ CNVs comparing genes hit by singleton CNVs to all

remaining case CNVs. This methodology was repeated in a second expression dataset for replication and permutations were performed to ascertain if significant findings across the datasets were unlikely to have occurred by chance.

P values reflecting evidence of alternative splicing were derived for each gene to reflect both global and region specific splicing. Alternatively spliced genes were then tested for enrichment with SCZ and BPD common risk variants and SCZ CNVs. Finally, sets of genes that were enriched for SCZ or BPD associated variants were used to identify GO pathways relevant for disease aetiology.

2.2 Results

2.2.1 Global pattern of gene expression and common risk variants

In order to characterise global expression patterns across the mid-foetal brain, three summary scores were calculated for each gene: the mean, scaled mean and coefficient of variation. In their work, Johnson *et al.* showed that many genes were co-expressed across the nine neocortical regions they included. To prevent the neocortex from being over-represented in the summary statistics the median value across all neocortical samples, for all individual and hemisphere pairings, was taken for each gene. The mean was then computed across the neocortical medians and all non-neocortical samples for each gene, to identify those highly expressed. Alongside this a scaled version of the mean was calculated, where it was divided by the maximum expression value for that gene. This restricts all values for this metric to fall between 0 and 1, with larger values representing more consistent expression relative to the maximum value for a given gene. Ultimately this will identify highly consistent but not necessarily highly expressed genes. In addition the coefficient of variation was also calculated. This measure was included to identify more variable expression profiles and hence, is inversely correlated to the scaled mean. It would be expected, therefore, that their results would be correlated also. The scaled mean is looking for consistency with its maximum expression value whereas the coefficient of variation is just looking for general consistency or variation. For genes where the expression values are either highly consistent or highly variable, the ranks from these

metrics should be negatively correlated. However, this may not be the case for genes in between these two extremes and agreement between these two metrics would present a more robust finding. These three measures will be referred to as global metrics.

A regression approach was used to test for correlations between the gene expression global metrics and SCZ or BPD gene-wide p values. The linear model took each global metric in turn as the independent variable and either Brown's or Simes' logP as the dependent variable for SCZ and BPD separately. These results are displayed in Table 2.4.

Strongly significant positive relationships were found between mean expression and SCZ Brown's logP ($p = 1.07 \times 10^{-10}$) and BPD Brown's logP ($p = 4.89 \times 10^{-6}$) indicating that genes with high means and therefore highly expressed have smaller association p values. An even more significant positive relationship was observed between the scaled mean and both SCZ ($p = 4.58 \times 10^{-14}$) and BPD ($p = 3.06 \times 10^{-8}$). Conversely, highly significant negative relationships were found when testing the coefficient of variation (SCZ $p = 3.03 \times 10^{-13}$; BPD $p = 1.99 \times 10^{-7}$), implying genes not variably expressed were associated to these disorders. By definition if a gene is not variably expressed it must be consistently expressed. Therefore, all three global metrics have identified an enrichment of SCZ and BPD common variants in genes with high consistent expression across the foetal brain. The same pattern of results was seen with the Simes' logP, again with strong levels of significance, although less significant than with Brown's logP.

Despite the high significance of the regression models between the global metrics calculated in the mid-foetal brain and gene-wide p values for SCZ and BPD, the correlation coefficients were small with absolute values between 0.03 and 0.07. Table 2.4 shows that these were marginally higher for SCZ compared to BPD, and generally also slightly higher for Brown's logP compared to Simes' logP.

		Schizophrenia		Bipolar disorder	
		Brown's	**Simes'**	**Brown's**	**Simes'**
Mean	**P value**	1.07×10^{-10}	6.25×10^{-8}	4.89×10^{-6}	2.29×10^{-6}
	Correlation Coeff.	0.0562	0.0502	0.0390	0.0438
		+	+	+	+
Scaled mean	**P value**	4.58×10^{-14}	1.47×10^{-7}	3.06×10^{-8}	2.10×10^{-5}
	Correlation Coeff.	0.0657	0.0488	0.0473	0.0395
		+	+	+	+
Coeff. of variation	**P value**	3.03×10^{-13}	2.98×10^{-6}	1.99×10^{-7}	6.24×10^{-5}
	Correlation Coeff.	-0.0635	-0.0433	-0.0444	-0.0371
		-	-	-	-
Excluding genes in MHC region					
Mean	**P value**	2.69×10^{-15}	7.47×10^{-13}	3.46×10^{-6}	3.66×10^{-6}
	Correlation Coeff.	0.0692	0.0670	0.0399	0.0433
		+	+	+	+
Scaled mean	**P value**	4.31×10^{-16}	1.05×10^{-9}	1.73×10^{-8}	1.04×10^{-5}
	Correlation Coeff.	0.0711	0.0570	0.0484	0.0412
		+	+	+	+
Coeff. of variation	**P value**	5.27×10^{-15}	1.23×10^{-7}	5.27×10^{-7}	3.49×10^{-5}
	Correlation Coeff.	-0.0684	-0.0494	-0.0456	-0.0387
		-	-	-	-

Table 2.4: Linear regression results and correlation coefficients testing global metrics calculated in the Johnson dataset with gene-wide logP.

Figure 2.2 presents scatterplots of the six relationships between the global metrics and Brown's logP. These associations do not look as convincingly linear as the regression p values may suggest, showing a large degree of noise. This is consistent with the small correlation coefficients reported in Table 2.4, indicating that although these results were highly significant, there are SCZ or BPD risk genes for which this relationship is not apparent. Corresponding scatterplots for Simes' logP can be seen in Appendix Figure 7.1 and show similar relationships.

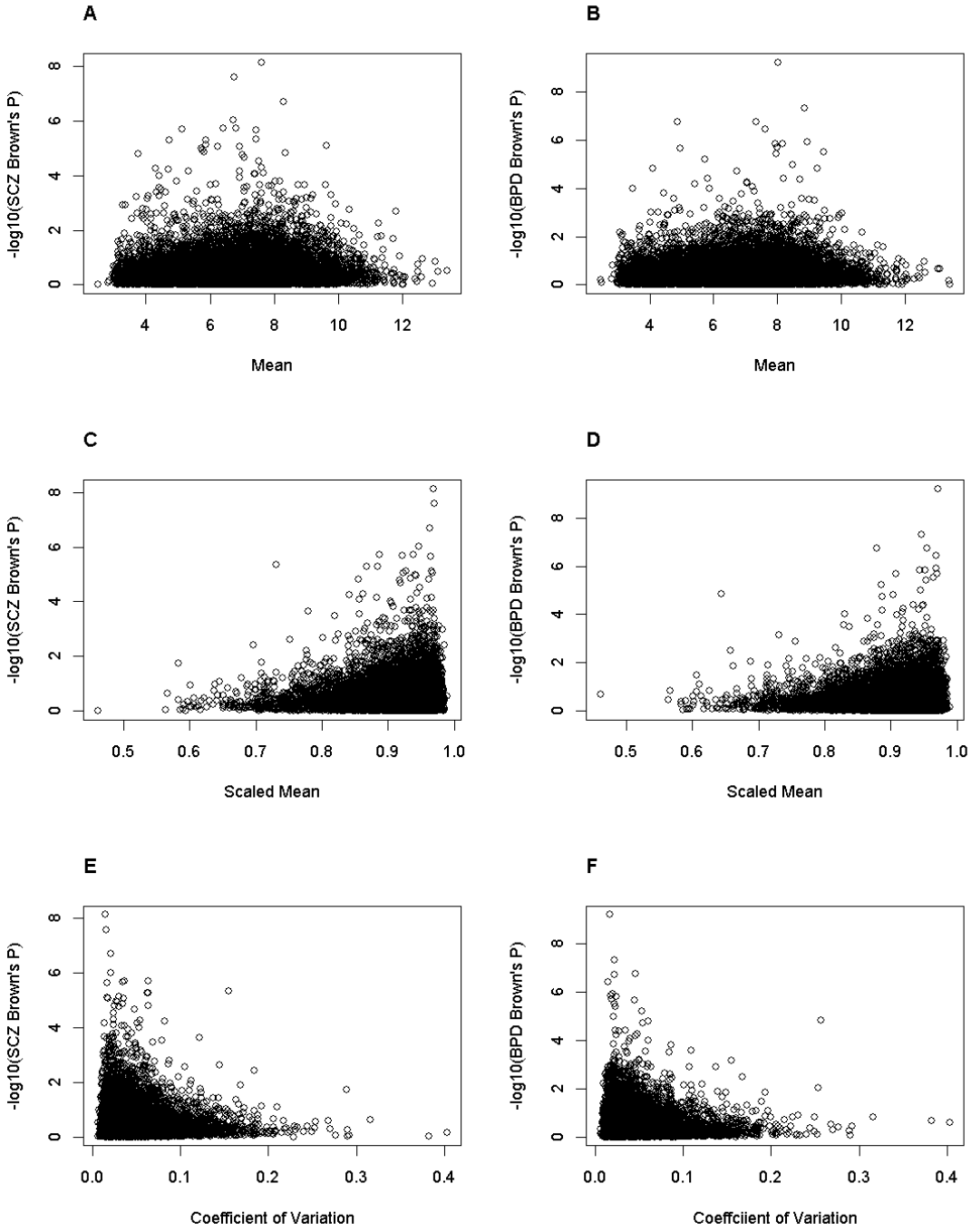

Figure 2.2: Scatterplots of relationships between global metrics calculated in the Johnson dataset and Brown's gene-wide p values.
Panels A, C and E plot SCZ Brown's logP against global metrics; panels B, D and F plot BPD Brown's logP against global metrics. Panels A and B plot mean expression across mid-foetal brain; panels C and D plot scaled mean; panels E and F plot coefficient of variation against gene-wide logP.

These analyses were repeated removing genes found in the MHC region which is strongly associated to SCZ. Due to the strong LD within this region many genes are likely to have highly significant gene-wide p values and may cause spurious results if

these genes are also highly co-expressed. Even without the MHC genes the same pattern of significance remained, shown in Table 2.4. While the BPD results were not greatly altered, the linear relationships with SCZ logP were a couple of orders of magnitude more significant and the correlation coefficients slightly stronger. The association between SNPs in this region is predominantly with SCZ rather than BPD (Bergen *et al.*, 2012) and therefore a greater effect on the SCZ regression models would be expected. Interestingly though these results suggest that the MHC region was not causing false positive results, in fact it was having an effect in the opposite direction and suppressing the signal.

Linear regression assumes that the errors have a normal distribution; despite linear regression being fairly robust, the possibility remains that the results from a parametric approach with this assumption may be biased. This problem is irrelevant in a nonparametric test such as the Mann-Whitney test, as no extra weight is given to highly significant p values, just a higher ranking. Therefore a rank-based approach was used to validate the results found with the parametric methods. Genes were ranked separately by each of the three global metrics before the top n% (5, 10…50%) was selected and tested for more significant p values against the bottom 50%.

In Figure 2.3 panel A, a clear enrichment of more significant SCZ Brown's p values can be seen in the sets of highest ranked genes by either the mean or scaled mean, i.e. those highly and consistently expressed. The most significant enrichments occurred when testing the top 20-50%. For example the strongest enrichment was found in the top 45% of genes ranked by mean expression ($p = 5.27 \times 10^{-13}$) and in the top 25% of genes ranked by their scaled mean expression ($p = 1.70 \times 10^{-15}$) showing that the regression results were not due to outliers. Panel C of Figure 2.3 shows that enrichments for BPD signals were also present in gene sets other than just the top 5 or 10% when ranked by either the mean or scaled mean; therefore these regression results were not due to extreme values either. No significant results were found for genes with high coefficients of variation, validating the regression findings that genes consistently expressed across the foetal brain regions were enriched for SCZ and BPD risk variants.

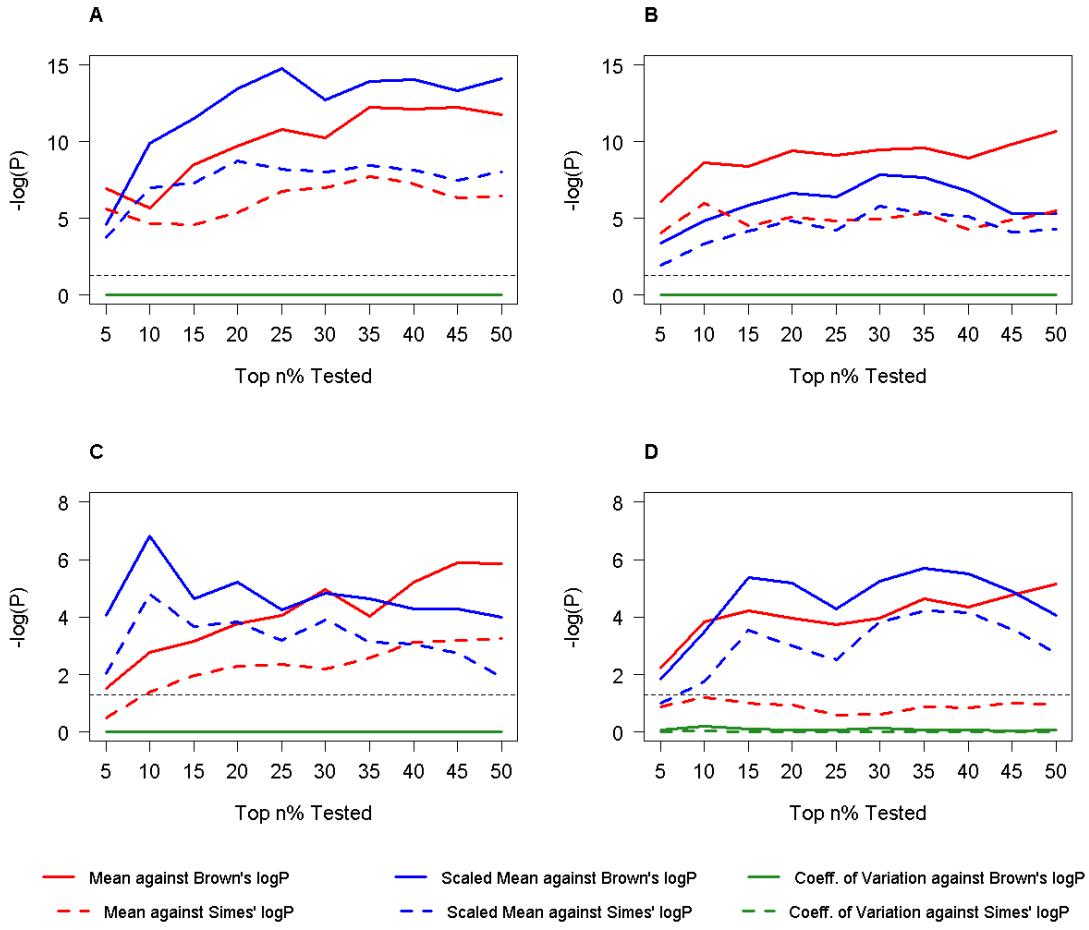

Figure 2.3: Results from Mann-Whitney tests for genes ranked by global metrics calculated in the Johnson and Kang datasets.
Panels A & B are results testing top n% of genes ranked by each global metric in turn against the bottom 50% for smaller SCZ p values; panels C & D are results testing for smaller BPD p values. Panels A & C global metrics were calculated in the Johnson dataset; panels B & D were calculated in Kang dataset. Black dashed line is p = 0.05.

The same pattern of results was seen with Simes' p values, also shown in Figure 2.3, although less significantly compared to results with Brown's p values. Results were generally more significant when testing SCZ p values compared to BPD p values. Removing genes in the MHC region had minimal effects on the results of the Mann-Whitney tests; hence these results are not presented here but can be found in Appendix Figure 7.5.

In the analyses presented so far, expression values for all neocortical regions have been combined into a single measure. To see if the findings reported above held

within the neocortex, the global metrics were calculated across the nine neocortical brain regions. Highly significant positive relationships were again found between the gene-wide logP and the mean or scaled mean expression, and highly significant negative relationships were found with the coefficient of variation. The correlation coefficients associated with these relationships were of a similar strength to those presented in Table 2.4 for global metrics calculated across the mid-foetal brain. This pattern of results, shown in Table 2.5, was found with both Brown's and Simes' p values. As with the global metrics calculated across the five major brain structures, Figure 2.4 and Appendix Figure 7.2 show that there was a fair amount of noise present in the significant relationships presented in Table 2.5.

		Schizophrenia		Bipolar disorder	
		Brown's	**Simes'**	**Brown's**	**Simes'**
Mean	**P value**	2.07×10^{-11}	8.36×10^{-12}	1.38×10^{-6}	1.14×10^{-6}
	Correlation Coeff.	0.0583	0.0617	0.0412	0.0424
		+	+	+	+
Scaled mean	**P value**	6.92×10^{-16}	5.01×10^{-11}	9.62×10^{-7}	2.54×10^{-5}
	Correlation Coeff.	0.0702	0.0594	0.0419	0.0399
		+	+	+	+
Coeff. of variation	**P value**	1.25×10^{-14}	2.22×10^{-11}	1.17×10^{-6}	6.53×10^{-5}
	Correlation Coeff.	-0.0671	-0.0605	-0.0415	-0.0380
		-	-	-	-
Excluding genes in MHC region					
Mean	**P value**	5.86×10^{-16}	7.95×10^{-19}	9.49×10^{-7}	2.80×10^{-6}
	Correlation Coeff.	0.0708	0.0806	0.0421	0.0427
		+	+	+	+
Scaled mean	**P value**	1.29×10^{-20}	5.92×10^{-17}	3.93×10^{-7}	7.50×10^{-6}
	Correlation Coeff.	0.0813	0.0761	0.0436	0.0408
		+	+	+	+
Coeff. of variation	**P value**	6.68×10^{-18}	5.20×10^{-16}	5.75×10^{-7}	1.85×10^{-5}
	Correlation Coeff.	-0.0754	-0.0738	-0.0430	-0.0390
		-	-	-	-

Table 2.5: Linear regression results and correlation coefficients testing global metrics calculated within neocortical regions in the Johnson dataset with gene-wide logP.

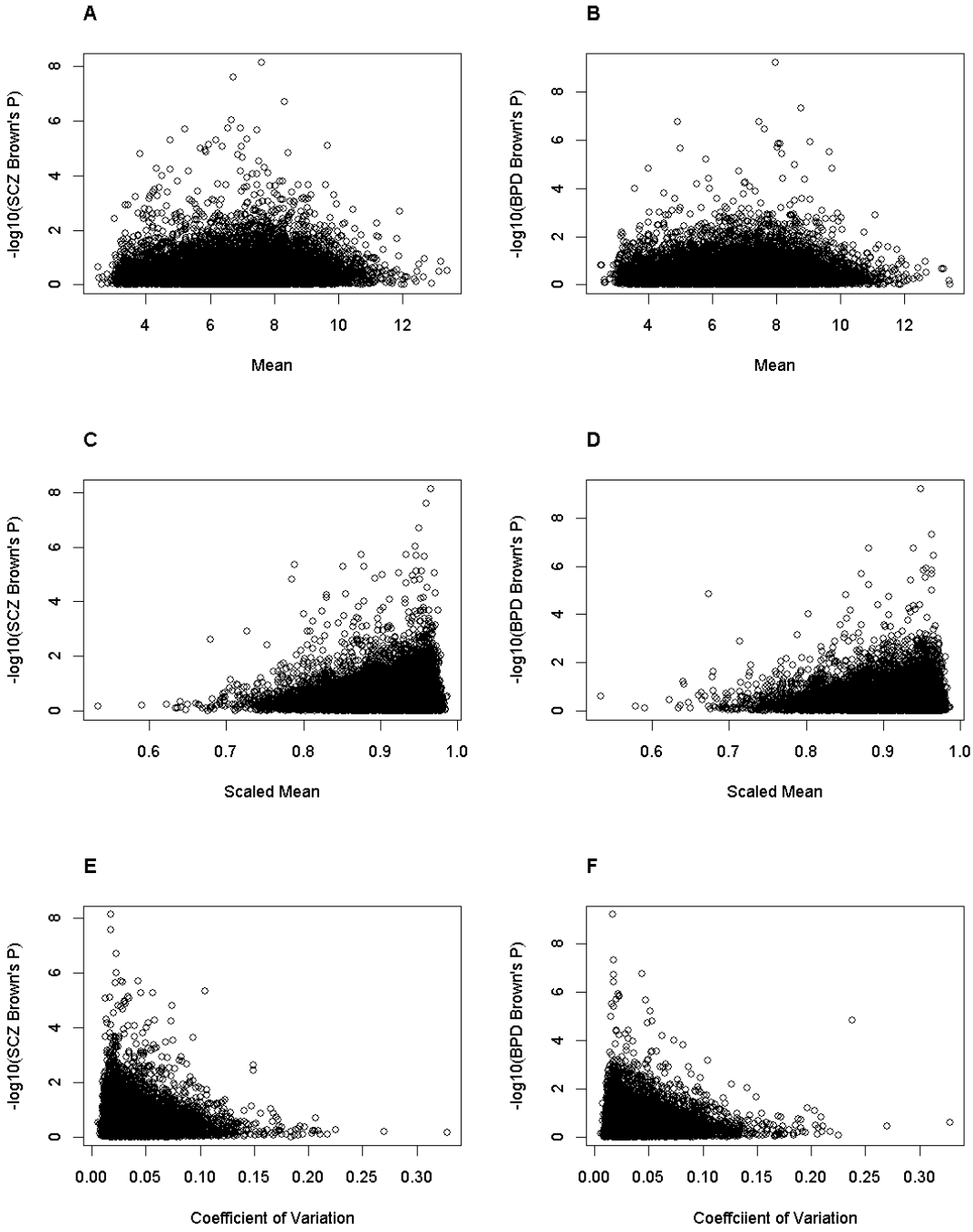

Figure 2.4: Scatterplots of relationships between testing global metrics calculated within neocortical regions in the Johnson dataset and Brown's gene-wide logP.
Panels A, C and E plot SCZ Brown's logP against global metrics; panels B, D and F plot BPD Brown's logP against global metrics. Panels A and B plot mean expression across mid-foetal brain; panels C and D plot scaled mean; panels E and F plot coefficient of variation against gene-wide logP.

Nonparametric tests, presented in Figure 2.5, showed these results were not due to extreme values as enrichments were found in the top 20-50% with both sets of p values, supportive of the general pattern of results. All analyses remained significant

after excluding the MHC genes, shown in Table 2.5 and Appendix Figure 7.5. Therefore the results presented so far show that genes consistently expressed across the neocortex and wider brain structures are enriched for SCZ and BPD common risk variants.

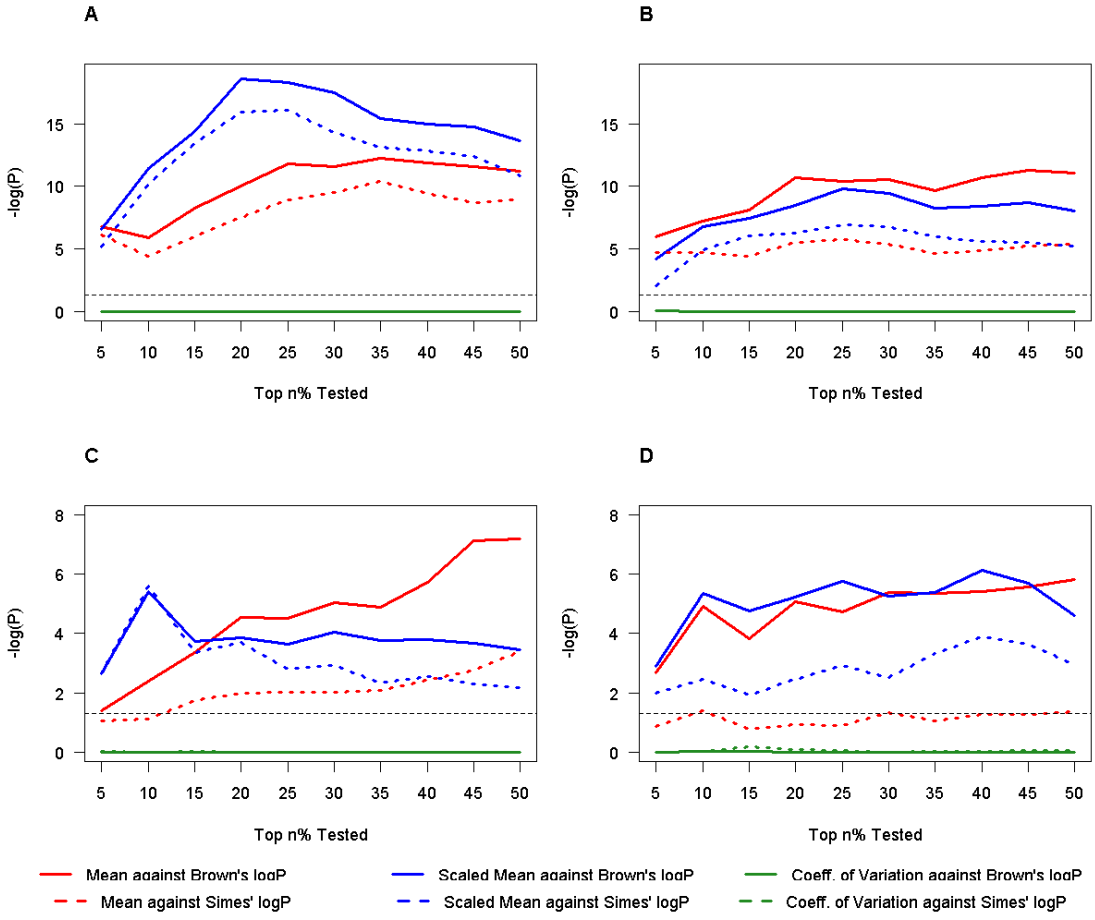

Figure 2.5: Results from Mann-Whitney tests for genes ranked by global metrics within neocortical regions calculated in the Johnson and Kang datasets.
Panels A & B are results testing top n% against the bottom 50% for smaller SCZ p values; panels C & D are results testing for smaller BPD p values. Panels A & C global metrics were calculated in the Johnson dataset; panels B & D were calculated in Kang dataset. Black dashed line is p = 0.05.

Validation in an independent expression dataset

All analyses were repeated in the Kang dataset to look for replications. Table 2.6 presents the results of the linear regression analyses for the global metrics calculated using the median neocortex values. Significant positive relationships were found when testing the mean (SCZ $p = 9.87 \times 10^{-8}$; BPD $p = 3.13 \times 10^{-5}$) and the scaled

mean (SCZ p = 1.57 x 10^{-6}; BPD p = 1.59 x 10^{-5}). As in the Johnson dataset a significant negative relationship was found with the coefficient of variation, although this was only nominally significant with Brown's BPD logP (SCZ p = 0.000149; BPD p = 0.0108). Results with Simes' logP were consistent with those for Brown's logP, although the negative relationship between the coefficient of variation and BPD Simes' logP was not significant. The correlation coefficients reported in Table 2.6 were of a similar magnitude to those reported in Table 2.4 for the Johnson dataset. Figure 2.6 and Appendix Figure 7.3 both show that these linear relationships were subject to a lot of noise and do not hold for all SCZ and BPD risk genes.

		Schizophrenia		Bipolar disorder	
		Brown's	**Simes'**	**Brown's**	**Simes'**
Mean	**P value**	9.87 x 10^{-8}	1.47 x 10^{-8}	3.13 x 10^{-5}	4.97 x 10^{-6}
	Correlation Coeff.	0.0462	0.0510	0.0354	0.042
		+	+	+	+
Scaled mean	**P value**	1.57 x 10^{-6}	2.84 x 10^{-5}	1.59 x 10^{-5}	0.000949
	Correlation Coeff.	0.0416	0.0377	0.0367	0.0297
		+	+	+	+
Coeff. of variation	**P value**	0.000149	0.00822	0.0108	0.102
	Correlation Coeff.	-0.0329	-0.0238	-0.0217	-0.0147
		-	-	-	-
Excluding genes in MHC region					
Mean	**P value**	1.22 x 10^{-11}	8.18 x 10^{-15}	1.87 x 10^{-5}	3.83 x 10^{-6}
	Correlation Coeff.	0.0591	0.0703	0.0366	0.0419
		+	+	+	+
Scaled mean	**P value**	9.12 x 10^{-8}	2.53 x 10^{-7}	9.15 x 10^{-6}	4.49 x 10^{-4}
	Correlation Coeff.	0.0466	0.0467	0.0379	0.0318
		+	+	+	+
Coeff. of variation	**P value**	0.000145	0.00705	0.00885	0.090
	Correlation Coeff.	-0.0331	-0.0244	-0.0224	-0.0154
		-	-	-	-

Table 2.6: Linear regression results and correlation coefficients testing global metrics calculated in the Kang dataset with gene-wide logP.

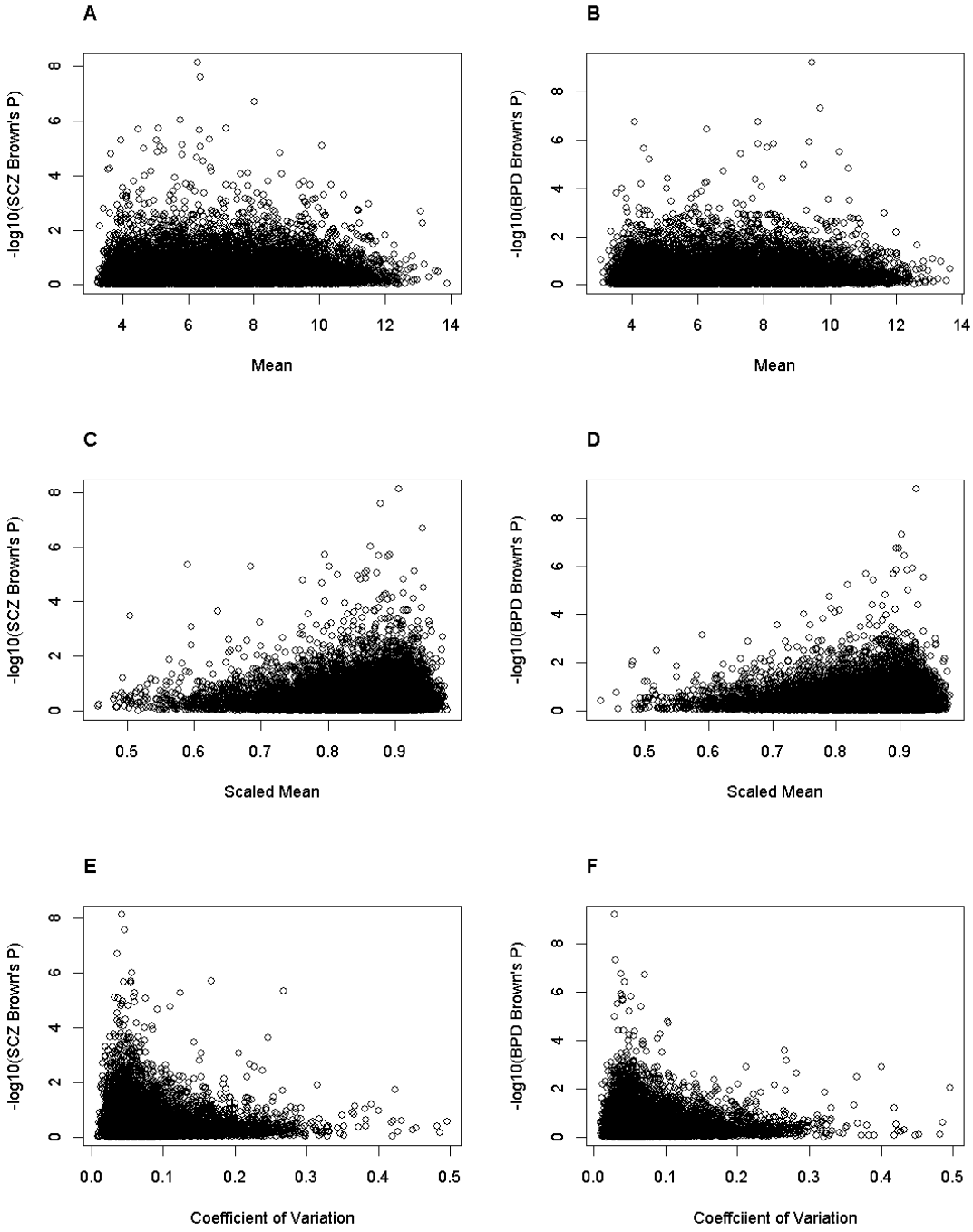

Figure 2.6: Scatterplots of relationships between global metrics calculated in the Kang dataset and Brown's gene-wide logP.
Panels A, C and E plot SCZ Brown's logP against global metrics; panels B, D and F plot BPD Brown's logP against global metrics. Panels A and B plot mean expression across mid-foetal brain; panels C and D plot scaled mean; panels E and F plot coefficient of variation against gene-wide logP.

Removing genes from the MHC region caused the SCZ regression models to become more significant and had a minimal effect on the BPD regression models, see Table 2.6. As in the Johnson dataset this showed that these associations were not driven

by the highly correlated p values of these genes. Mann-Whitney tests verified the regression associations with Brown's logP and SCZ Simes' logP, shown in Figure 2.3 panels B and D. An enrichment of BPD Simes' p values was found in gene sets ranked by their scaled mean, but only a trend for enrichment was found for genes ranked in the top 10% by their mean expression value. Removing the MHC genes from the Mann-Whitney results had minimal impact, shown in Appendix Figure 7.5.

		Schizophrenia		Bipolar disorder	
		Brown's	**Simes'**	**Brown's**	**Simes'**
Mean	**P value**	9.90×10^{-9}	6.58×10^{-10}	1.98×10^{-6}	1.01×10^{-6}
	Correlation Coeff.	0.0497	0.0556	0.0404	0.0440
		+	+	+	+
Scaled mean	**P value**	8.80×10^{-6}	3.66×10^{-5}	1.33×10^{-6}	6.76×10^{-6}
	Correlation Coeff.	0.0385	0.0372	0.411	0.0405
		+	+	+	+
Coeff. of variation	**P value**	6.69×10^{-6}	0.00151	4.79×10^{-6}	0.000187
	Correlation Coeff.	-0.0390	-0.0286	-0.0389	-0.0336
		-	-	-	-
Excluding genes in MHC region					
Mean	**P value**	9.68×10^{-13}	8.63×10^{-17}	8.77×10^{-7}	4.73×10^{-7}
	Correlation Coeff.	0.0622	0.0754	0.0420	0.0456
		+	+	+	+
Scaled mean	**P value**	1.49×10^{-7}	2.42×10^{-8}	4.84×10^{-7}	2.08×10^{-6}
	Correlation Coeff.	0.0458	0.0506	0.0430	0.0430
		+	+	+	+
Coeff. of variation	**P value**	5.13×10^{-8}	3.72×10^{-6}	1.42×10^{-6}	5.17×10^{-5}
	Correlation Coeff.	-0.0475	-0.0419	-0.0412	-0.0367
		-	-	-	-

Table 2.7: Linear regression results and correlation coefficients testing global metrics calculated within neocortical regions in the Kang dataset with gene-wide logP.

Analyses were repeated in the Kang dataset within the neocortical samples and are presented in Table 2.7. Significant linear models were found for both disorders, with both sets of gene-wide p values for all three global metrics, consistent with the results in the Johnson dataset shown in Table 2.5. The results of the rank-based tests with Brown's p values, see Figure 2.5, verified these regression results. Generally results testing Simes' p values showed that the regression associations were not due to extreme values although only nominal enrichments were found for BPD in the top 10, 30 and 50% of genes ranked by their mean expression. Again, scatterplots of

these associations, Figure 2.7 and Appendix Figure 7.4, showed that these signals were particularly noisy. Rerunning these analyses without the MHC genes did not change the pattern of results; see Table 2.7 and Appendix Figure 7.6.

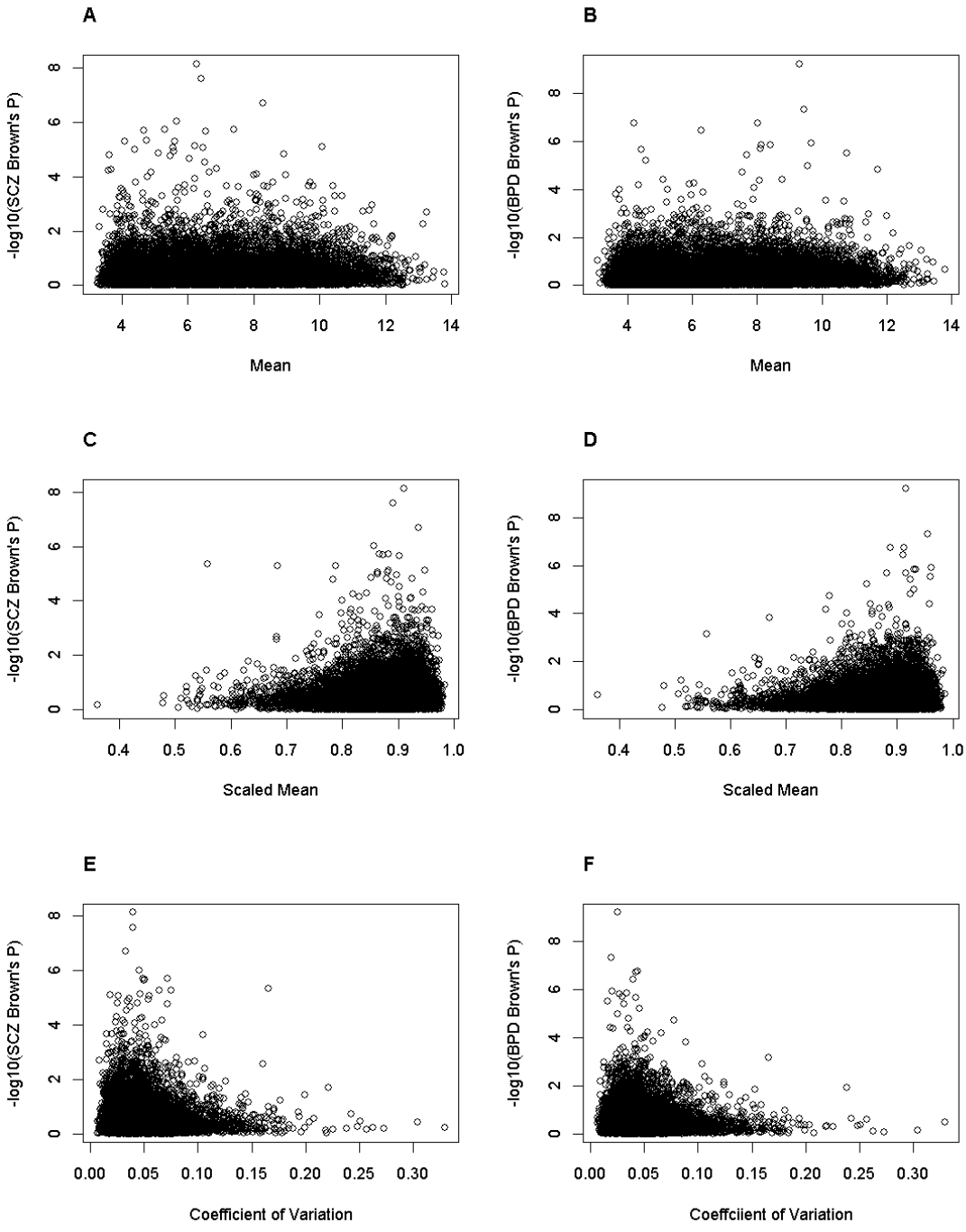

Figure 2.7 Scatterplots of relationships testing global metrics calculated within neocortical regions in the Kang dataset and Brown's gene-wide logP.
Panels A, C and E plot SCZ Brown's logP against global metrics; panels B, D and F plot BPD Brown's logP against global metrics. Panels A and B plot mean expression across mid-foetal brain; panels C and D plot scaled mean; panels E and F plot coefficient of variation against gene-wide logP.

Summary

The analyses presented from two independent expression datasets showed that genes consistently and highly expressed across the human mid-foetal brain were enriched for both SCZ and BPD risk variants. This was found to be the case for gene expression across non-neocortex structures as well as across neocortical regions. Although no formal correction for multiple testing has been applied to these p values they were generally sufficiently small to remain significant even after a conservative Bonferroni correction. Adjusting the significance threshold of $p < 0.05$ for the 12 tests in the top half of Tables 2.4-7 would require a p value of less than 0.00417 to report the tests as significant. Despite these associations being highly significant, they were found to be particularly noisy, meaning that this relationship does not hold for all SCZ and BPD associated genes.

The initial associations found with a parametric regression model were validated with rank-based tests showing that these results were not due to extreme values, neither were the associations due to genes in the MHC region. Testing both Brown's and Simes' gene-wide p values obtained broadly the same results pattern, however more significantly and consistently across SCZ and BPD with Brown's p values.

2.2.2 Regional characteristic gene expression and common risk variants

So far the variation in expression levels across the mid-foetal brain captured through the global metrics has been fairly general. This section investigates genes that exhibit specific patterns of variation, looking for those upregulated or downregulated in an individual brain region. Such genes were referred to as characteristic genes for that region.

Initially a characteristic score was defined based on fold changes between brain region expression values (taking the mean expression across individuals and hemispheres) to calculate a relative enrichment score for each gene in each region (Doyle *et al.*, 2008). Regional characteristic scores were then regressed against the gene-wide logP to identify any region(s) where relative expression levels were

correlated with disease risk. On closer inspection of the results it was found that correlations were typically driven by extreme scores in one or occasionally two individuals.

To properly control for inter-individual variation, along with other sample differences, a linear model was constructed in which these appeared as covariates alongside a binary brain region term that was used to compare the expression values of one region to all others. By fitting this model separately for each brain region, characteristic scores for each gene were derived from the p value and coefficient estimates of this binary term. The magnitude or absolute value of this score indicated how strongly characteristic the gene was of that region and the sign of the score specified whether there was an increase or decrease in expression relative to the average expression across all other brain regions. This formed a scale where large positive values of the characteristic score indicated genes with a highly characteristic increase of expression in that brain region relative to all other brain regions, and large negative values indicated genes with a highly characteristic decrease of expression (further detail in the Methods).

Significant number of associated regions across both microarray datasets

The characteristic scores for each region were tested for a linear relationship with the dependent variable Brown's logP. Significant results with positive coefficients signified that upregulated genes in that brain region were associated with more significant gene-wide p values, whereas significant results with a negative coefficient indicated that downregulated genes in that brain region were associated with more significant gene-wide p values. Before considering the results for each brain region, the true results were compared to permutations to see whether the number of significant models, limited to the 12 brain regions present in both datasets, was greater than that expected by chance.

Briefly, in each dataset the sample labels (containing individual, hemisphere and region information) were permuted before the characteristic scores were

recalculated and tested with SCZ and BPD Brown's logP. Based on 1000 permutations, empirical p values were calculated for the number of significant SCZ or BPD regression models ($p < 0.01$) for both expression datasets, shown in Table 2.8. Empirical p values were also calculated for the number of regions significant across both datasets, and in the same direction, for all pairs of permutations, 10^6 in total, presented in Table 2.9.

Johnson	Number of brain regions	Empirical p value	Brain regions
Schizophrenia p < 0.01	5	0.553	CBL, DLPFC, HIP, MPFC, THAL
Bipolar disorder p < 0.01	3	0.521	HIP, MPFC, THAL
Kang			
Schizophrenia p < 0.01	6	0.533	CBL, HIP, TAS, THAL, PAS, OCC
Bipolar disorder p < 0.01	5	0.458	HIP, THAL, MPFC, PAS, OCC

Table 2.8: Empirical p values for the number of significant regression models between regional characteristic scores and SCZ or BPD Brown's logP.
Based on 12 regions overlapping both datasets and 1000 permutations.

Across both Johnson and Kang	Number of brain regions	Empirical p value	Brain regions
Schizophrenia p < 0.01	3	0.376	CBL, HIP,THAL
Schizophrenia p < 0.01, in same direction	3	0.096	CBL, HIP,THAL
Bipolar disorder p < 0.01	3	0.075	MPFC, HIP, THAL
Bipolar disorder p < 0.01, in same direction	3	0.013	MPFC, HIP, THAL
Schizophrenia & bipolar disorder p < 0.01	2	0.083	HIP, THAL
Schizophrenia & bipolar disorder p < 0.01, in same direction	2	0.025	HIP, THAL

Table 2.9: Empirical p values for the number of significant regression models between regional characteristic scores and SCZ or BPD Brown's logP across both datasets.
Based on 12 regions overlapping both datasets and 10^6 permutations.

The first observation from these results was that events expected to be quite rare, for example Table 2.8 shows 5 of 12 regions having a SCZ regression model p value less than 0.01, had quite a high empirical p value ($p = 0.553$). Expression across neocortical regions has previously been shown to be co-expressed (Johnson *et al.*, 2009), which could reduce the independence of the observations, particularly if the associations indicate differential expression between neocortical regions and non-

neocortical regions. The second observation was that each dataset taken individually did not show more significant associations between brain region characteristic genes and disease p values than expected by chance (all empirical $p > 0.05$).

When looking for consistent results across datasets, two results were greater than expected by chance. BPD Brown's logP were associated with three regional characteristic scores with the same sign of the coefficient: HIP, THAL and MPFC ($p = 0.013$). The second significant result was across both disorders where two regions, HIP and THAL, were significant in the same direction across both datasets ($p = 0.025$).

Table 2.10 shows the results of the regression analyses for the HIP, THAL and MPFC in the Johnson dataset adjusted for testing 13 brain regions by Bonferroni's procedure. This shows that genes with decreased expression or downregulated in the HIP were enriched for SCZ (corrected $p = 1.76 \times 10^{-6}$) and BPD (corrected $p = 0.0186$) Brown's logP. The THAL characteristic scores were also negatively correlated with the SCZ (corrected $p = 0.0178$) and BPD (corrected $p = 0.000534$) Brown's logP and hence it was genes downregulated in this brain region that were associated with more significant gene-wide p values. Conversely, genes with increased expression or upregulated in the MPFC were associated with BPD common variants (corrected $p = 0.0384$), although in this dataset the enrichment was stronger with SCZ Brown's logP (corrected $p = 2.63 \times 10^{-5}$). Table 2.10 also shows that characteristic scores for these three regions were significant when tested with SCZ Simes' logP. Nominal evidence was found with the BPD Simes' logP, which did not remain significant after correction for 13 brain regions, but was importantly in the same direction. The full results for this dataset including all other brain regions can be found in Appendix Table 7.1.

The correlation coefficients for these relationships, also presented in Table 2.10, were small and of a similar magnitude to those reported with the global metrics in Section 2.2.1. The strongest correlation was found between the HIP characteristic

scores and SCZ Brown's logP (r = -0.0459). Generally the correlations were stronger for associations with SCZ logP, as was observed with the global metrics in Section 2.1.1.

		Schizophrenia		Bipolar disorder	
		Brown's	**Simes'**	**Brown's**	**Simes'**
MPFC	**P value**	2.02×10^{-6} (2.63×10^{-5})	3.52×10^{-5} (0.000458)	0.00295 (0.0384)	0.0341 (0.443)
	Correlation Coeff.	0.0414	0.0374	0.0254	0.0192
		+	+	+	+
HIP	**P value**	1.35×10^{-7} (1.76×10^{-6})	1.78×10^{-6} (2.32×10^{-5})	0.00143 (0.0186)	0.0283 (0.368)
	Correlation Coeff.	-0.0459	-0.0432	-0.0272	-0.0198
		-	-	-	-
THAL	**P value**	0.00137 (0.0178)	2.37×10^{-5} (0.000308)	4.10×10^{-5} (0.000534)	0.00668 (0.0868)
	Correlation Coeff.	-0.0279	-0.0382	-0.0350	-0.0245
		-	-	-	-

Table 2.10: Linear regression results and correlation coefficients testing regional characteristic scores calculated in the Johnson dataset with gene-wide logP.
P values in brackets have been corrected for 13 brain regions using Bonferroni's method.

		Schizophrenia		Bipolar disorder	
		Brown's	**Simes'**	**Brown's**	**Simes'**
MPFC	**P value**	0.997	0.324	0.00195 (0.0312)	0.0203 (0.324)
	Correlation Coeff.	-3.17×10^{-5}	0.00888	0.0133	0.0209
		-	+	+	+
HIP	**P value**	8.28×10^{-10} (1.32×10^{-8})	8.78×10^{-8} (1.40×10^{-6})	3.35×10^{-6} (5.36×10^{-5})	9.23×10^{-6} (0.000148)
	Correlation Coeff.	-0.0532	-0.0481	-0.00743	-0.0399
		-	-	-	-
THAL	**P value**	2.66×10^{-5} (0.000425)	6.69×10^{-7} (1.07×10^{-5})	4.55×10^{-6} (7.28×10^{-5})	0.00293 (0.0469)
	Correlation Coeff.	-0.0364	-0.0447	-0.00390	-0.0268
		-	-	-	-

Table 2.11: Linear regression results and correlation coefficients testing regional characteristic scores calculated in the Kang dataset with gene-wide logP.
P values in brackets have been corrected for 16 brain regions using Bonferroni's method, where missing corrected p value was 1.

Table 2.11 presents the results from the same analyses but for the Kang dataset, where the p values were corrected for 16 brain regions. Strong negative correlations

were found between the HIP and THAL characteristic scores and both the BPD and SCZ Brown's logP. It can also be seen that there was a positive correlation between the MPFC characteristic scores and the BPD logP only. Results with the Simes' logP supported these associations. Compared to Table 2.10 the results were more significant in the Kang dataset. The correlation coefficients are again small, suggesting these results do not explain the totality of risk genes for SCZ and BPD. The full results for this dataset can be found in Appendix Table 7.2.

Validation with nonparametric tests

For the same reasons outlined in the global metrics analyses (Section 2.2.1) nonparametric tests were used to verify the regression results. In order to test for an enrichment in either the upregulated or downregulated set of characteristic genes, tests were run as follows. Genes were ranked by the absolute value of their characteristic score i.e. the p value for differential expression of that gene in the region in question, and the top n% (5, 10…50%) selected. These sets were then split into two subgroups, genes with positive characteristic scores indicating increased characteristic expression, and genes with negative characteristic scores indicating decreased characteristic expression. The relevant subset consistent with the coefficient of the significant regression model was then tested against the bottom 50% of genes i.e. those that were not characteristic of that region in either direction, for more significant gene-wide p values.

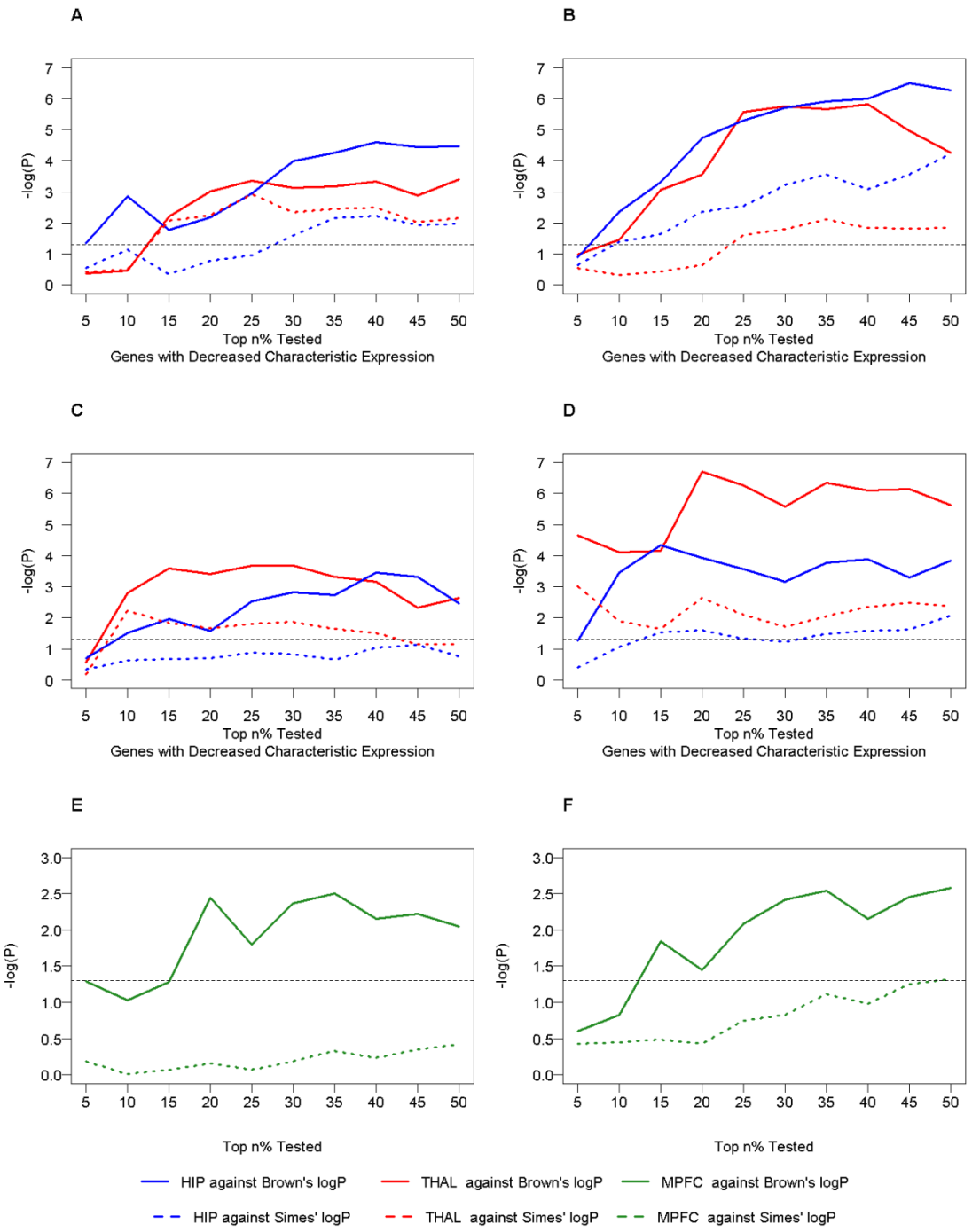

Figure 2.8: Results from Mann-Whitney tests to verify significant regression models with regional characteristic scores calculated in the Johnson and Kang datasets.
Genes were ranked by absolute characteristic scores and the top n% split into positive and negative subsets. The subset consistent with the direction of the significant regression model in Tables 2.10 and 2.11 was tested in a one-sided Mann-Whitney test against the bottom 50%. Panels A - D tested the negative subset i.e. genes with decreased expression; panels E & F tested the positive subset i.e. genes with increased expression. Panels A & B tested for smaller SCZ p values; panels C - F tested for smaller BPD p values. Panels A, C & E characteristic scores were calculated in the Johnson dataset; panels B, D & F were calculated in Kang dataset. Black dashed line is p = 0.05.

Mann-Whitney tests found significantly smaller SCZ or BPD Brown's p values in genes with decreased expression in either the HIP or THAL in the top 15-50% in both the Johnson and Kang datasets; see Figure 2.8 panels A-D. This verifies that the regression results reported for these regions were not due to outliers. Across both datasets enrichments were also found for smaller BPD Brown's p values in genes with increased expression in the MPFC, panels E-F in Figure 2.8. As this was the case for genes in the top 20-50% of characteristic genes for this region, this result was also not caused by outliers.

Simes' p values produced less significant results in the rank-based tests compared to Brown's p values, an observation that was broadly the case in the regression results. Mann-Whitney tests with SCZ Simes' p values were significant for genes with decreased expression in the HIP and THAL in the top 30-50% for both datasets. Results from testing BPD Simes' p values were less significant, with a trend for enrichment in genes with increased expression in the MPFC found only in the Kang dataset. In the top 35-50% of genes ranked by HIP characteristic score in the Kang dataset nominal enrichments were found for BPD Simes' p values, whereas only a trend for significance was found in the top 45% in the Johnson dataset. Enrichments were found in both datasets for genes with decreased expression in the THAL in the top 10-40%.

For the regional characteristic scores showing consistent associations across both datasets, combined scores were calculated using expression data from both the Johnson and Kang studies. This was done to reduce the noise present in either dataset, as genes detected across both studies were more likely to be true signals. All subsequent analyses in this section were performed with these combined characteristic scores. Initially these were used to check that the associations were not driven by genes in the MHC region.

As expected Table 2.12 shows significant linear relationships between the Brown's logP and combined dataset characteristic scores for the HIP, THAL and MPFC. Running the analyses without the MHC genes broadly did not affect the results but

did reduce the significance of several models, such as MPFC in BPD and THAL in SCZ. Generally the same results pattern was found with Simes' logP, however for BPD the associations were only nominally significant, and after removing the MHC genes the MPFC model was no longer significant. While all correlation coefficients were small, the strongest correlation coefficient was for the association between HIP characteristic scores and SCZ Brown's logP.

		Schizophrenia		Bipolar disorder	
		Brown's	**Simes'**	**Brown's**	**Simes'**
MPFC	**P value**			0.000995	0.0362
	Correlation Coeff.			0.0282	0.0190
				+	+
HIP	**P value**	2.39×10^{-7}	5.92×10^{-6}	0.000539	0.00285
	Correlation Coeff.	-0.0450	-0.0410	-0.0296	-0.0270
		-	-	-	-
THAL	**P value**	0.000167	7.00×10^{-6}	2.91×10^{-5}	0.0205
	Correlation Coeff.	-0.0328	-0.0407	-0.0358	-0.0210
		-	-	-	-
Removing MHC genes					
MPFC	**P value**			0.00269	0.0596
	Correlation Coeff.			0.0258	0.0172
				+	+
HIP	**P value**	5.10×10^{-7}	4.40×10^{-6}	0.000505	0.00186
	Correlation Coeff.	-0.0440	-0.0419	-0.0299	-0.0284
		-	-	-	-
THAL	**P value**	0.00148	0.000124	3.29×10^{-5}	0.0166
	Correlation Coeff.	-0.0279	-0.0350	-0.0357	-0.0219
		-	-	-	-

Table 2.12: Linear regression results and correlation coefficients testing regional characteristic scores calculated across both the Johnson and Kang datasets with gene-wide logP.

The results of nonparametric tests, presented in Figure 2.9, supported the regression results with Brown's p values presented in Table 2.12. Genes with decreased expression in the HIP were enriched for more significant SCZ (best p = 1.30×10^{-5} top 35%) and BPD p values (best p = 0.000903 top 20%). Similar results were found for genes with decreased expression in the THAL (SCZ best p = 6.07×10^{-6} top 35%; BPD best p = 2.11×10^{-5} top 30%). The top 25-50% of MPFC characteristic genes with increased expression were enriched for BPD Brown's p values (best p = 0.00181 top

25%).

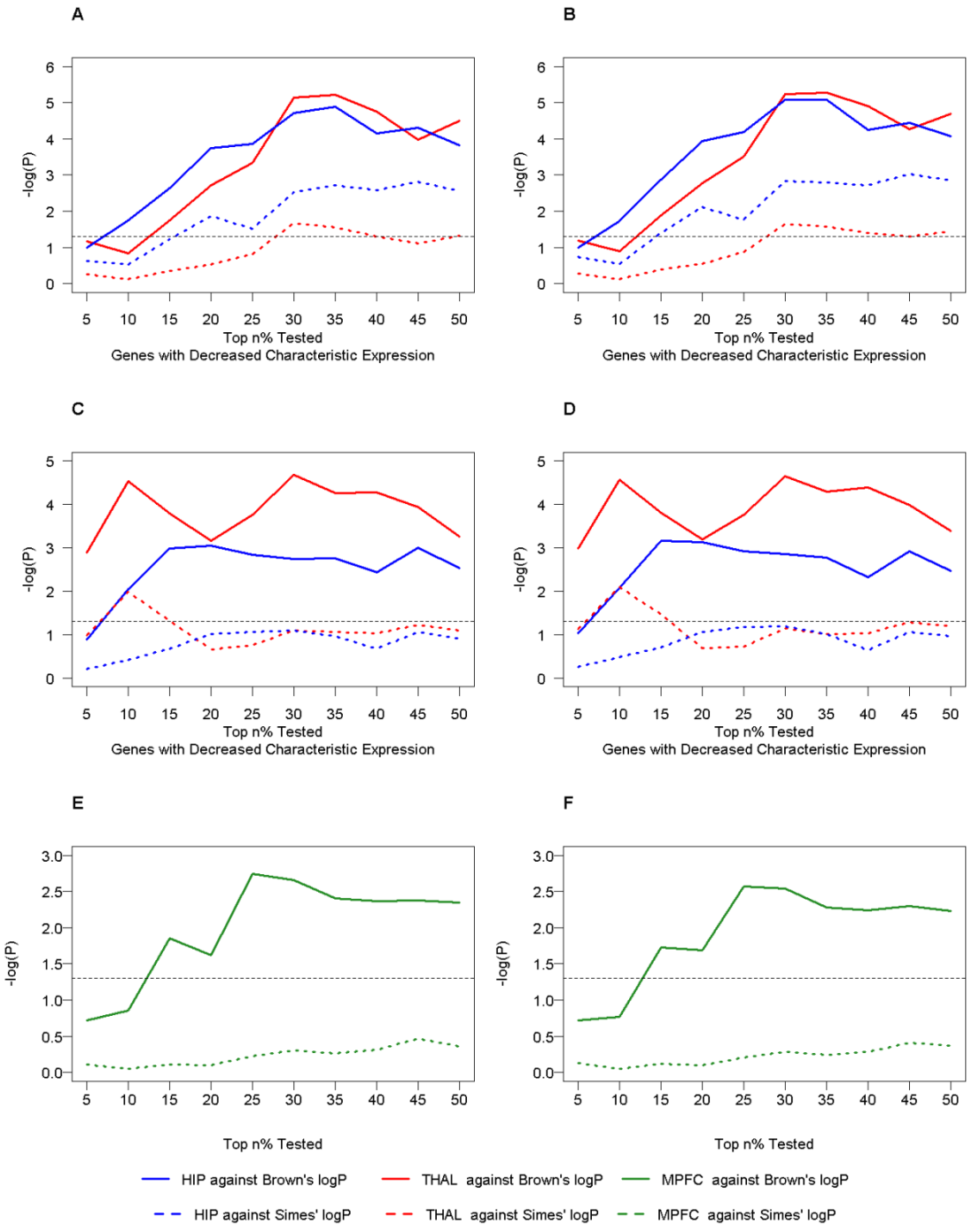

Figure 2.9: Results from Mann-Whitney tests to verify significant regression models with regional characteristic scores calculated across Johnson and Kang datasets.
Genes were ranked by absolute characteristic scores and the top n% split into positive and negative subsets. The subset consistent with the direction of the significant regression model in Table 2.12 was tested in a one-sided Mann-Whitney test against the bottom 50%. Panels A - D tested the negative subset i.e. genes with decreased expression; panels E-F tested the positive subset i.e. genes with increased expression. In all panels results are based on characteristic scores calculated across Johnson and Kang datasets. Panels A & B tested for smaller SCZ p values; panels C - F tested for smaller BPD p values. Panels A, C & E included all genes, panels B, D & F excluded genes in MHC region from analysis. Black dashed line is p = 0.05.

Results from testing SCZ Simes' p values showed only nominal significance when testing genes with decreased expression in either the HIP or THAL. A trend for significance or nominal significance was found with BPD Simes' p values in HIP or THAL characteristic genes and no significance was found in genes with increased expression in the MPFC. Almost identical results were found when removing the MHC genes from the nonparametric analyses; see Figure 2.9 panels B, D & F.

Test for independent associations

Using the combined dataset characteristic scores, the associations for each set were tested simultaneously to see if they were independent. The results of linear regression models for each pair of scores predicting Brown's logP only, as this set of p values had the most consistent enrichments, are presented in Table 2.13.

Dependent variable (Brown's logP)	**Independent variables included in regression model**					
	HIP characteristic score		**THAL characteristic score**		**MPFC characteristic score**	
	P value	**Coeff.**	**P value**	**Coeff.**	**P value**	**Coeff.**
Schizophrenia	1.11×10^{-5}	-	0.00915	-		
Bipolar disorder	0.0112	-	0.000553	-		
Bipolar disorder	0.00307	-			0.00574	+
Bipolar disorder			0.000555	-	0.0215	+
Bipolar disorder	0.0216	-	0.00369	-	0.0420	+

Table 2.13: Linear regression results testing regional characteristic scores calculated across both the Johnson and Kang datasets simultaneously to predict Brown's logP.
Each row represents a separate regression model.

In each pair of characteristic scores, both remained significant although to a lesser degree than when tested individually; compare Table 2.13 to Table 2.12. This indicates that the associations of the characteristic scores were not completely independent. A model was also fitted predicting BPD Brown's logP with all three characteristic scores; this showed that the MPFC score explained little of the signal when including both the THAL and HIP. Therefore as the HIP and THAL were consistent across the two disorders, and largely explain the MPFC signal only these will be considered further.

Expression of enriched gene sets across brain development

In order to compare the results of this Section with that found in Section 2.2.1 for genes consistently expressed across the mid-foetal brain, the expression of the most enriched HIP and THAL characteristic genes was plotted. Based on the results from the Mann-Whitney tests those with negative characteristic scores in the top 35% of genes ranked by their absolute HIP characteristic score were taken as the most enriched set. Figure 2.9 panel A shows that the most significant result with SCZ Brown's gene-wide p values came in the top 35% and panel C shows that the most significant results with BPD Brown's gene-wide p values came in the top 20%, hence the top 35% was chosen as this encompassed the top enrichments for both disorders. Similarly, the genes with negative characteristic scores in the top 35% of absolute THAL characteristic scores were taken as that most enriched set, as the top 35% was most enriched for smaller SCZ Brown's gene-wide p values and the top 30% for smaller BPD Brown's gene-wide p values, also shown in Figure 2.9 panels A and C.

Thus far, only a small part of mid-foetal development has been considered, between 18 and 23 weeks gestation. The Kang replication dataset was taken from a larger transcriptome study covering a much wider range of human brain development and can be used to look at expression profiles over a longer period of development. For each brain region, at each time point, the median expression for the set of enriched genes was plotted from embryonic through to adolescence. For comparison a corresponding value was calculated based on all remaining genes in the dataset that were not in the enriched gene set.

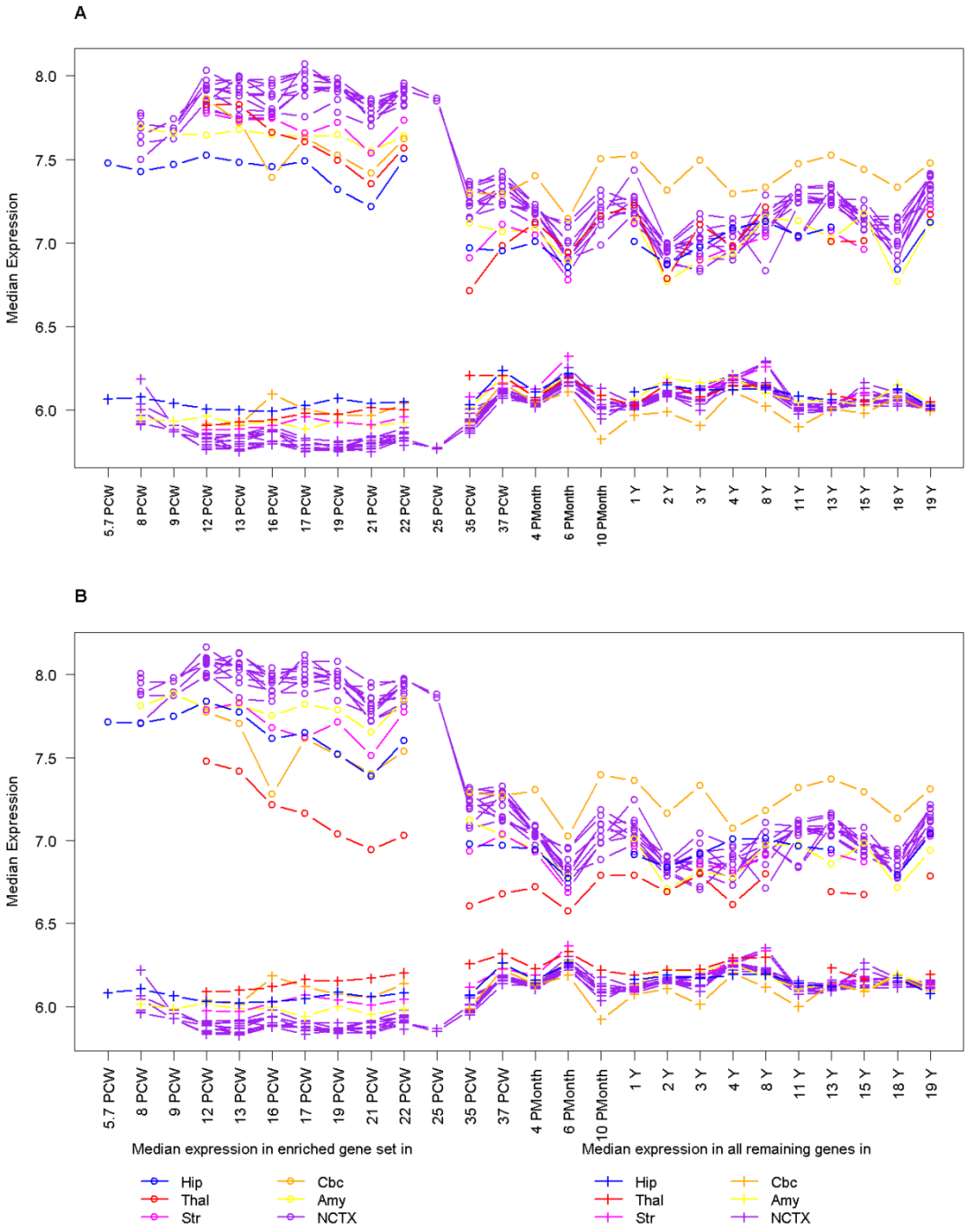

Figure 2.10: Expression across development for most enriched characteristic HIP and THAL gene sets. The sets of genes with decreased expression in either the HIP or THAL most enriched for smaller SCZ Brown's gene-wide p values and BPD Brown's p values were identified from Mann-Whitney tests of characteristic scores calculated across both the Johnson and Kang datasets; panel A top 35% HIP decreased, panel B top 35% THAL decreased. Median expression values for each time point were calculated in more extensive version of Kang dataset, in addition median expression values were also calculated for all remaining genes not part of the enriched gene sets labelled as 'Rest' in figure. PCW – post conception weeks, Mon – months, Y-years, NCTX – neocortex.

Figure 2.10 shows that the set of enriched characteristic HIP genes do have lower expression in the HIP compared to all other regions however, these genes are still

relatively highly expressed in this region, consistent with the results of Section 2.2.1. In fact the decreased expression in the HIP occurs across all foetal samples up to 22 PCW. In this set of genes all brain regions had a peak of expression during foetal development that dropped off before birth, followed by lower expression in postnatal samples. In contrast, genes not part of this enriched set showed a relatively consistent profile across development. A similar pattern was seen when taking the most enriched set of THAL decreased genes. Despite an obvious decrease in expression relative to the neocortical regions in the THAL, these genes were still highly expressed in this region. The decrease in expression for the THAL compared to other samples was greater than seen in the HIP and remained throughout early years.

The expression plots of these two sets of enriched genes corroborates the pairwise regression analyses by suggesting there was an underlying common set of genes with the same development expression profile. Further, it was evident from these graphs that these genes would have higher means than the remaining set of genes and suggests some overlap in these results with those in the Section 2.2.1.

Summary

In summary, two sets of characteristic genes showed enrichment for SCZ common variants: genes downregulated in the HIP and genes downregulated in the THAL. Both of these gene sets, as well as genes upregulated in the MPFC, were enriched for BPD common variants. The association of two sets of characteristic scores (HIP and THAL) with SCZ and BPD, in the same direction across both datasets, was significantly greater than expected when compared to random permutations. The significant enrichment of the MPFC characteristic scores with BPD, in the same direction across both datasets, in addition to the HIP and THAL associations was also more than expected by chance. All three of these relationships were associated with small correlation coefficients, indicating that they do not explain all of SCZ or BPD risk. These results were most consistent with the Brown's p values using both regression

and rank-based tests. Simes' gene-wide p values showed associations in the same direction but not always significantly nor consistently across disorders or tests.

The associations of the HIP, THAL, MPFC characteristic genes with BPD variants were found to be partly overlapping, with the HIP and THAL associations capturing most of the MPFC signal. The SCZ signals in the HIP and THAL characteristic genes were also shown to be overlapping. A common variable profile of expression across development was identified for the enriched sets of genes characteristic of the HIP and THAL, further supporting the idea that they have captured an overlapping set of co-regulated genes. The profile was characterised by a peak of expression during foetal development followed by a relative decrease in postnatal stages and suggests these are developmentally regulated genes, which may be particularly important during foetal brain development.

2.2.3 Global pattern of gene expression and schizophrenia structural variants

In this section CNVs identified in SCZ case control studies were tested to see if they hit genes with common expression profiles. For each CNV the global metrics, introduced in Section 2.2.1, were collated for each gene hit and the minimum, median and maximum calculated. A series of logistic regression models were fitted to compare these metrics between CNVs found in SCZ patients and those found in controls (formulae for these can be found in the Methods). CNV data from the MGS and ISC studies were combined for this analysis, so extra covariates were included to control for study and chip differences. In addition a term for the number of genes hit by that CNV (limited to those found in the expression dataset) was included to control for any CNV size bias, as CNVs that hit more genes are more likely to have an extreme score. Each model was fitted for the full set of CNVs, and the deletions and duplications separately.

No significant differences were found for any of the global metrics (mean, scaled mean or coefficient of variation) when taking the minimum, median or maximum

score of the genes hit by each CNV, see Table 2.14. This was also the case when testing the deletions and duplications separately.

		Compare genes hit by schizophrenia CNVs to control CNVs			Compare singleton schizophrenia CNVs to non-singleton schizophrenia CNVs		
All CNVs		**Min.**	**Med.**	**Max.**	**Min.**	**Med.**	**Max.**
Mean	**P value**	0.594	0.492	0.431	0.320	0.0852	0.0305
	Coeff.	-	-	-	+	+	+
Scaled mean	**P value**	0.319	0.523	0.771	0.236	0.182	0.0401
	Coeff.	-	-	-	+	+	+
Coeff. of variation	**P value**	0.745	0.453	0.131	0.0700	0.179	0.527
	Coeff.	+	+	+	-	-	-
Deletion CNVs							
Mean	**P value**	0.644	0.877	0.762	0.205	0.129	0.150
	Coeff.	-	+	+	+	+	+
Scaled mean	**P value**	0.170	0.573	0.832	0.111	0.0339	0.0198
	Coeff.	-	-	-	+	+	+
Coeff. of variation	**P value**	0.931	0.740	0.240	0.0585	0.0928	0.299
	Coeff.	+	+	+	-	-	-
Duplication CNVs							
Mean	**P value**	0.643	0.377	0.443	0.885	0.364	0.106
	Coeff.	-	-	-	+	+	+
Scaled mean	**P value**	0.973	0.899	0.562	0.902	0.852	0.511
	Coeff.	+	+	+	+	-	+
Coeff. of variation	**P value**	0.745	0.773	0.413	0.480	0.816	0.895
	Coeff.	-	+	+	-	-	+

Table 2.14: Logistic regression results testing CNV status on global metrics calculated in the Johnson dataset.

Studies have shown that not only do SCZ patients have an increased number of CNVs, but also that those found in SCZ patients are rare in the general population (International Schizophrenia Consortium, 2008, Walsh *et al.*, 2008). Hence, the global metrics for each case CNV, calculated as above, were tested to see if singleton CNVs, defined as those that did not overlap with any control CNV, hit genes with common expression characteristics. Similar to above, a logistic regression framework was used for this analysis, to compare singleton case CNVs to the remaining case CNVs.

Nominal p values for the maximum mean values (p = 0.0305) and maximum scaled mean (p = 0.0401) suggest that singleton SCZ CNVs hit genes more highly and consistently expressed compared to more common SCZ CNVs, see Table 2.14. Analyses run separately showed that this was not specific to either deletions or duplications when testing the mean metric. The median (p = 0.0339) and maximum (p = 0.0198) scaled means were significantly higher in singleton deletions compared to non-singleton deletions but not when comparing duplications. However these results were only nominally significant and would not survive multiple testing.

		Compare genes hit by schizophrenia CNVs to control CNVs			**Compare singleton schizophrenia CNVs to non-singleton schizophrenia CNVs**		
All CNVs		**Min.**	**Med.**	**Max.**	**Min.**	**Med.**	**Max.**
Mean	**P value**	0.620	0.594	0.589	0.679	0.197	0.0203
	Coeff.	-	-	-	+	+	+
Scaled mean	**P value**	0.377	0.497	0.473	0.133	0.0430	0.00338
	Coeff.	-	-	-	+	+	+
Coeff. of variation	**P value**	0.503	0.457	0.407	0.155	0.332	0.603
	Coeff.	+	+	+	-	-	-
Deletion CNVs							
Mean	**P value**	0.652	0.333	0.190	0.890	0.617	0.512
	Coeff.	+	+	+	+	+	+
Scaled mean	**P value**	0.572	0.655	0.660	0.334	0.195	0.0566
	Coeff.	-	+	+	+	+	+
Coeff. of variation	**P value**	0.599	0.562	0.699	0.130	0.151	0.269
	Coeff.	-	-	-	-	-	-
Duplication CNVs							
Mean	**P value**	0.116	0.0829	0.127	0.717	0.221	0.0149
	Coeff.	-	-	-	+	+	+
Scaled mean	**P value**	0.456	0.284	0.436	0.308	0.128	0.0289
	Coeff.	-	-	-	+	+	+
Coeff. of variation	**P value**	0.401	0.221	0.196	0.635	0.911	0.708
	Coeff.	+	+	+	-	+	+

Table 2.15: Logistic regression results testing CNV status on global metrics calculated in Kang dataset.

Validation in independent expression dataset

Repeating these analyses in the Kang dataset, presented in Table 2.15, found no significant differences in the global metrics of genes hit when comparing case CNVs to control CNVs. Singleton SCZ CNVs were found to hit genes with higher mean expression, testing the maximum (p = 0.0203), and higher scaled mean expression,

when testing the median (p = 0.0430) and maximum (p = 0.00338). None of these remained significant when limiting the analysis to deletions but when testing the duplications both the maximum mean (p = 0.0149) and maximum scaled mean (p = 0.0289) of the genes hit were significantly higher in the singleton CNVs.

Summary

No significant differences were found in either dataset when comparing SCZ CNVs to those found in healthy controls for the mean, scaled mean or coefficient of variation of the genes hit by each CNV. Both datasets provided nominal evidence that singleton SCZ CNVs hit genes with increased means and scaled means, which after correction for multiple testing would be unlikely to remain significant. Therefore further replication of these results would be needed.

2.2.4 Regional characteristic gene expression and schizophrenia structural variants

Next, SCZ CNVs were compared to control CNVs to see if the genes they hit were more characteristic of particular brain regions. For each brain region, the characteristic scores for each gene hit by each CNV were collated and the minimum, median and maximum were taken. As positive characteristic scores indicated an increase of expression in that region and negative scores indicated a decrease of expression in that region, for any significant model the sign of the coefficient informed which type of characteristic genes were associated. The same regression framework from the previous section was used, controlling for study, chip and the number of genes hit by each CNV. No significant differences ($p < 0.01$) were found between genes hit by SCZ CNVs compared to control CNVs for any brain region's characteristic scores. Further, no significant results were found when separately analysing the deletions and duplications. This was also the case in the Kang dataset results; these results can be found in Appendix Tables 7.3 and 7.5.

Singleton CNVs were compared to all remaining case CNVs for each brain region's characteristic scores. After correcting the analyses in both the Johnson and Kang

datasets for the number of brain regions tested no regression model was significant across both datasets. Hence these results will be discussed no further but can be found in the Appendix Tables 7.4, 7.6.

2.2.5 Alternative splicing and common risk variants

So far the methods used in this chapter for identifying variable expression in the foetal brain would not capture whether multiple isoforms of a gene were being expressed. The previous analyses of variably expressed genes, in Sections 2.2.1 and 2.2.2 were based on a composite value of expression taken from multiple transcripts for each gene, so it is possible that consistently expressed genes were present as alternative isoforms.

The availability of expression data at the exon level from the Affymetrix Human Exon chip enabled methods to be employed to detect for alternative splicing events. The FIRMA (finding isoforms using robust multichip analysis) algorithm (Purdom *et al.*, 2008) was used to identify alternative splicing. FIRMA calculates a score for each exon-sample pairing based on the residual from the estimation step of the robust multichip average (RMA) normalisation procedure to characterise differences between the observed and estimated expression values. The benefit of using this method was that it enabled comparisons between groups, so both global alternative splicing and splicing specific to each brain region could be investigated.

This method was applied separately to each individual in the dataset before being combined. Initially, to look at global splicing across the foetal brain, the FIRMA scores were summarised into a single value for each exon. This was achieved by fitting a linear model for each exon to see if any brain region had a non-zero coefficient. To give each gene an overall splicing p value, all p values from all exons and individuals were collated and the best Simes' corrected one taken.

In order to minimize the number of false positives it has been suggested to remove lowly expressed genes prior to detecting alternative splicing (Affymetrix, 2008). The

FIRMA method was applied to the raw CEL files, which prevented pre-filtering steps from being applied prior to the algorithm itself. Instead genes with only one exon and genes filtered out of the previous analyses as lowly expressed in the foetal brain were removed after calculating FIRMA p values before any further analyses. One caveat of using the FIRMA method was that potential confounders, such as individual and hybridisation date could not be included.

		Schizophrenia		Bipolar disorder	
		Brown's	**Simes'**	**Brown's**	**Simes'**
Johnson	**P value**	0.787	0.419	0.536	0.241
	Correlation Coeff.	0.00236	0.00734	0.00228	0.0107
		+	+	+	+
Kang	**P value**	0.354	0.637	0.847	0.167
	Correlation Coeff.	-0.00806	0.00426	-0.00165	0.0125
		-	+	-	+

Table 2.16: Linear regression results and correlation coefficients testing global splicing logP with gene-wide logP.

The global splicing logP were tested in a regression framework with the Brown's logP to assess if genes with evidence of alterative splicing were enriched for association signal. No significant relationship with either SCZ (p = 0.787) or BPD (p = 0.536) Brown's logP was found, shown in Table 2.16. In the Kang dataset none of the regression models with either the Brown's or Simes' logP were significant.

Brain region splicing p values were also calculated from the FIRMA scores using a linear model to compare the scores from one region to all other regions and deriving a p value for each exon for each brain region. These were combined into gene level brain region splicing p values by selecting the best p value after Simes' correction. Each set of brain region splicing logP were then tested in a linear model with the Brown's logP. Across the Johnson and Kang dataset, no set of region splicing p values were consistently associated with either SCZ or BPD. Therefore these results will not be presented here but can be found in Appendix Tables 7.7 and 7.8.

2.2.6 Alternative splicing and schizophrenia structural variants

The minimum, median and maximum global splicing logP were calculated from the genes hit by each CNV to test if they were significantly different between case and control CNVs. This was not the case, neither were they significantly different between singleton case CNVs and all other case CNVs, Table 2.17.

	Comparing schizophrenia CNVs to control CNVs			Comparing rare schizophrenia CNVs to common schizophrenia CNVs		
All CNVs	**Minimum**	**Median**	**Maximum**	**Minimum**	**Median**	**Maximum**
P value	0.795	0.643	0.460	0.423	0.477	0.737
Coeff.	-	-	-	-	-	-
Deletion CNVs						
P value	0.409	0.520	0.363	0.400	0.419	0.469
Coeff.	-	-	-	+	+	+
Duplication CNVs						
P value	0.962	0.932	0.902	0.105	0.140	0.349
Coeff.	+	-	+	-	-	-

Table 2.17: Logistic regression results testing CNV status on global splicing logP calculated in the Johnson dataset.

None of the twelve overlapping regions between the two expression datasets had significantly different brain region splicing p values between case and control CNVs across both datasets; results presented in Appendix Tables 7.9 and 7.11. Significant models were found when comparing singleton SCZ CNVs to all other SCZ CNVs, although after correction for testing multiple brain regions there were no consistent enrichments across the Johnson and Kang datasets. These results will not be presented here but can be found in Appendix Tables 7.10 and 7.12.

2.2.7 Functional analysis of genes with enriched expression profiles

Significant associations were found between SCZ and BPD common variants and genes with decreased expression in the HIP or THAL, therefore characteristic scores for both of these regions were used to identify functional terms. Annotation terms were taken from the GO database (Ashburner *et al.*, 2000) to identify those that were enriched for genes with decreased expression in either the HIP or THAL, consistent with the direction of the association reported in Section 2.2.2. The GO

terms are the largest resource of pathways (du Plessis *et al.*, 2011) with a wide range of terms and therefore were used here for a hypothesis free analysis.

The GO terms were filtered to those with between 20 and 2000 genes, leaving 3204 unique terms that were tested. A Mann-Whitney test was used for each GO term to compare the HIP and THAL characteristic scores for genes annotated to that term against all remaining genes, to see if they had smaller characteristic scores. In other words these tests looked to identify GO terms with genes that had decreased expression in these regions in line the enrichments reported for SCZ and BPD variants in Section 2.2.2.

A Bonferroni corrected significance threshold of 1.56×10^{-5} was used to identify significant GO terms for the HIP and THAL separately. Many of the GO terms within each set are known to be overlapping and therefore the significance of many of the pathways was not independent. Within each set the pathways were clustered into related groups by identifying the smallest term that captured the signal of any larger terms. Genes in the smallest pathway were removed from all larger pathways and the enrichment analysis was repeated on the remaining genes in each larger pathway. If the larger pathway was no longer significantly enriched for characteristic genes, the smaller pathway was said to explain it. Any pathways explained were combined with the smaller pathway into a merged pathway, and the process repeated until no more pathways could be explained.

The set of terms enriched for lower THAL characteristic scores was the largest with 134 terms compared to 105 with lower HIP characteristic scores. In each set 32 terms were found to explain at least one other term, shown in Figures 2.6 and 2.7, with seventeen terms common to the two figures, highlighted in yellow. Within each panel the terms and clusters have been manually grouped into broad themes, with the same themes appearing in both sets. These themes were ‘Chromosome: structural modification & repair’, ‘Transcription’ and ‘Post-transcriptional RNA processing & transport’.

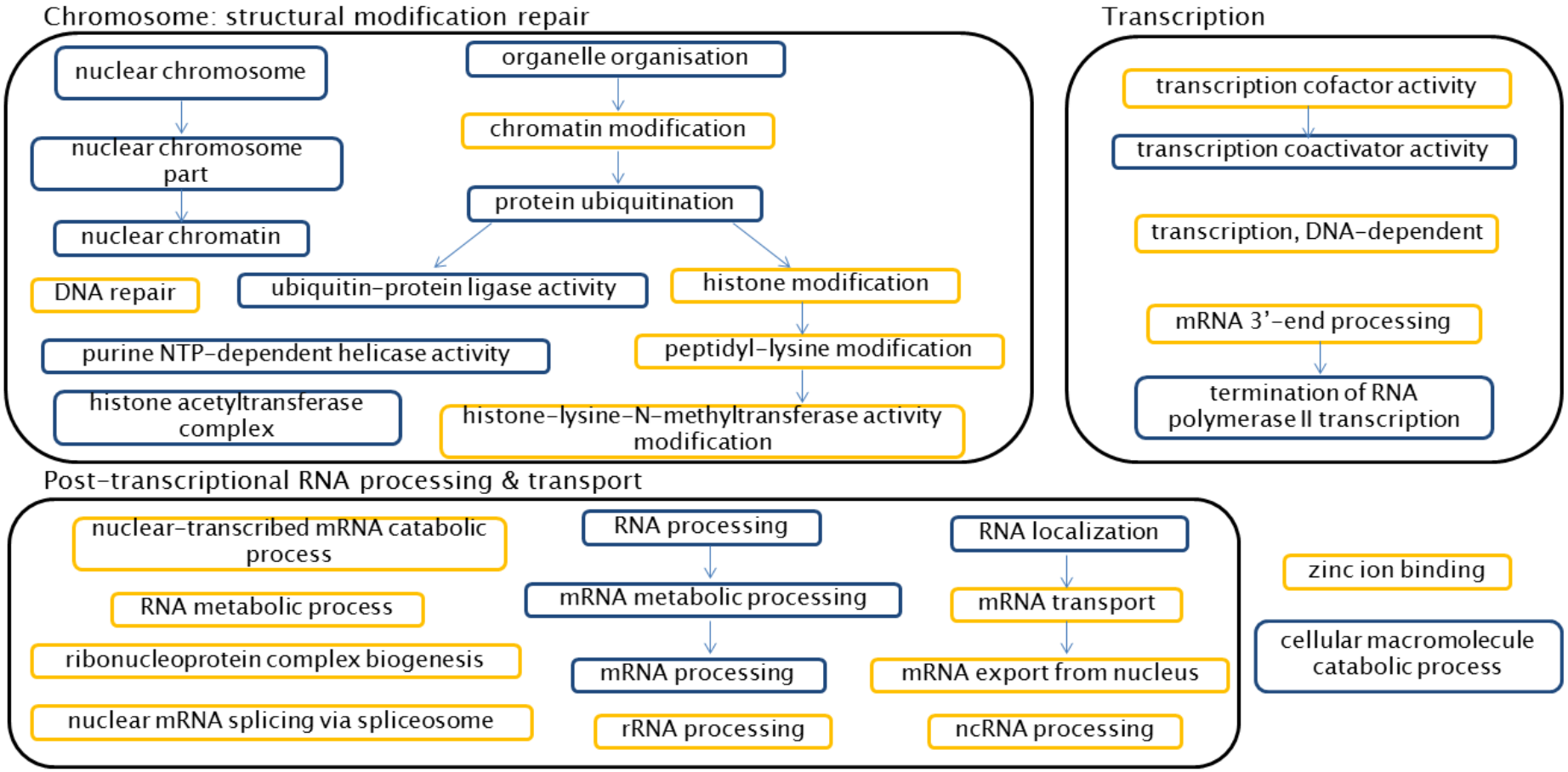

Figure 2.11: Key annotation terms identified from set of significant GO terms with significantly smaller HIP characteristic scores.
Figure shows set of 32 GO terms that explained at least one other term in the set of significant pathways. Terms that did not explain any other term were not included in the Figure. Arrows point from explaining term to the merged pathway it explains i.e. the term pointed to, merged with all other terms it explains. Terms in yellow ovals are also present in Figure 2.12; terms in black boxes were grouped into common themes.

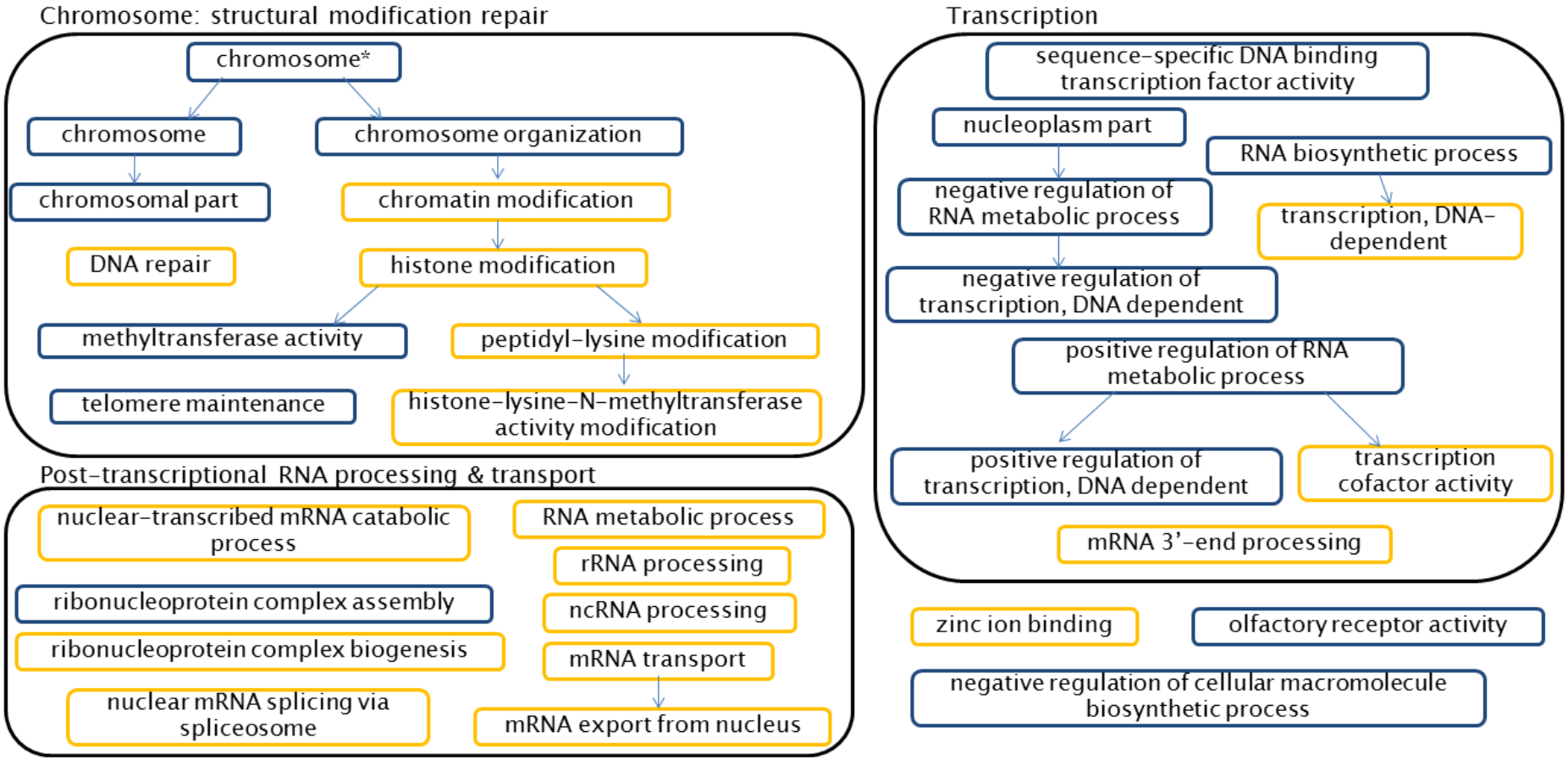

Figure 2.12: Key annotation terms identified from set of significant GO terms with significantly smaller THAL characteristic scores.
Figure shows set of 32 GO terms that explained at least one other term in the set of significant pathways. Terms that did not explain any other term were not included in the Figure. Arrows point from explaining term to the merged pathway it explains i.e. the term pointed to, merged with all other terms it explains. Terms in yellow ovals are also present in Figure 2.11; terms in black boxes were grouped into common themes. *Means merged term explains pathway pointed to.

Having found characteristic gene sets that were associated with SCZ and BPD and identified pathways enriched in these gene sets, the final step was to test whether these pathways captured the association signal of each gene set. The set of genes with the strongest enrichment for SCZ and BPD gene-wide p values identified from the Mann-Whitney tests, the top 35% for the HIP and the top 35 % for THAL, were used as the basis for set-based tests. The HIP decreased genes and the THAL decreased genes were then spilt into two groups, those that were annotated to one of the significant pathways for that region, and those that were not annotated to any of the significant pathways. A set-based test was then performed on each subset to look for an association with SCZ or BPD. This set-based test was equivalent to Brown's method for combining non-independent test statistics and in this case collapses the SNP p values from a GWAS into a single p value to demonstrate the sets' overall association with a phenotype.

The SNP association p values were taken from the PGC GWAS for SCZ and BPD and SNP correlations were calculated from the HapMap data (http://hapmap.ncbi.nlm.nih.gov/). At a gene set level, where SNPs from multiple genes were combined, it was noticed that sets with more SNPs had more significant p values. In order to correct for this, 10000 simulations were performed (see Methods for details) and an adjustment made for the number of SNPs in each set.

Genes with decreased expression in the HIP and in pathways enriched for HIP decreased genes showed evidence for association with both SCZ (adjusted p = 0.00611) and BPD (adjusted p = 0.0290) compared to genes not in significant pathways, results in Table 2.18. This was also the case for genes with decreased expression in the THAL, where genes in significant pathways were associated to both disorders (SCZ adjusted p = 0.000804; BPD adjusted p = 0.000463) whereas genes with decreased expression in the THAL but not in significant pathways were not. This result would not have been expected if either the SCZ or BPD associations or functional gene set enrichments with the HIP or THAL characteristic scores were spurious, and therefore validates these findings.

		Schizophrenia		Bipolar disorder	
		Number of SNPs in set	**Adjusted p value**	**Number of SNPs in set**	**Adjusted p value**
Top 35% HIP decreased genes	**In pathways**	40164	0.00611	76007	0.0290
	Not in pathways	44160	0.145	83138	0.0975
Top 35% THAL decreased genes	**In pathways**	37620	0.000804	69560	0.000463
	Not in pathways	35932	0.134	68408	0.722

Table 2.18: Results of set-based tests for genes found in HIP and THAL enriched sets split into those annotated to a significant pathway and those not.
Set p values were adjusted for number of SNPs in each set.

Further set-based tests were then used to ascertain which of the pathway groups in Figures 2.6 and 2.7 were the most enriched for association signal. For each pathway group a set was formed based on the genes in either the top 35% of HIP decreased, or the top 35% of THAL decreased, and in any of the pathways within that group. After adjusting for the number of SNPs in each set, all pathway groups had a significant association with SCZ, with the most significant p value for the 'Chromosome: structural modification & repair' group (adjusted p = 0.000599). For BPD, two of the three pathway groups were associated with the 'Transcription' group being the exception. The most significant group for BPD was the 'Post-transcriptional RNA processing & transport' group (adjusted p = 3.40×10^{-6}).

	Schizophrenia		Bipolar disorder	
	Number of SNPs in set	**Adjusted p value**	**Number of SNPs in set**	**Adjusted p value**
Chromosome: structural modification & repair	13063	0.000599	24190	0.00475
Transcription	11390	0.0124	20922	0.381
Post-transcriptional RNA processing & transport	18032	0.00983	33032	3.40×10^{-6}

Table 2.19: Results of set-based tests based on pathway groups in Figures 2.11 and 2.12.
Set p values were adjusted for number of SNPs in each set.

2.3 Discussion

2.3.1 Identification of common spatial expression profiles

This chapter presents the first brain-wide and genome-wide study in the human foetal brain for either SCZ or BPD associated genes. In particular, expression profiles across multiple regions of the mid-foetal human brain were investigated to identify sets of genes associated with SCZ or BPD. The first reported association in Section 2.2.1 was between genes consistently highly expressed across the mid-foetal brain, and both SCZ and BPD common risk variants. This finding was then replicated in an independent expression dataset covering the same period of gestation.

While general variation was not enriched for SCZ or BPD common variants, enrichments were found in gene sets identified with specific variable expression profiles in Section 2.2.2. Genes with decreased expression in the HIP or THAL were found to be associated with risk genes for both disorders. In addition, genes with increased expression in the MPFC during this period of foetal development were associated with BPD. These results were also replicated in an independent expression dataset and across the two datasets the probability of these results occurring by chance was nominal, hence these appear to be robust results.

The associations of these three sets of characteristic genes were not independent. The HIP and THAL enriched genes were shown to have the same variable development profile from foetal through to adolescence supporting the idea that there was a common underlying gene set driving the associations. This common profile, shown in Figure 2.10, had high expression during foetal development and lower levels of expression during postnatal stages. The peak of expression of these gene sets was observed to occur during the second trimester suggesting these genes play a role in development processes active during this time frame, before dropping off in the third trimester. The second trimester is an important period of neurogenesis. Neurons are produced at a very high rate and even a minor insult can significantly affect the maturation of these cells influencing the structure and

function of the brain, making this period of brain development particularly vulnerable (Miranda, 2012).

Two recently published studies also made use of the BrainSpan publically available expression datasets to investigate the spatial and temporal patterns of autism risk genes (Parikshak *et al.*, 2013, Willsey *et al.*, 2013). Both used a network-based approach to identify genes that showed common expression profiles in the developing brain. Willsey *et al.* looked at co-expression between nine autism risk genes, selected as having the strongest evidence for association, and all other genes in the dataset. Resulting networks based on co-expression during mid-foetal development (10-24 PCW) and in prefrontal regions were enriched for a second set of autism risk genes, termed probable risk genes as they were based on weaker evidence than the initial high confidence set. This work and the work presented this chapter highlights the importance of gene expression in the human mid-foetal brain for both SCZ and autism.

In contrast, Parikshak *et al.* created a weighted gene co-expression network using correlations between all pairs of genes in the dataset, and identified distinct modules enriched for autism candidate genes that had common developmental expression trajectories. Their module M2, and to some degree M3, showed a temporal expression profile consistent with that in Figure 2.10. Further, both of these modules were enriched for GO pathways relating to the regulation of expression including specific terms identified in Figures 2.11 and 2.12 such as 'zinc ion binding' and 'histone modification'. The similarity of these findings and those presented in this thesis suggest that the developmental trajectory described is not specific to SCZ but may be the case for a number of neurodevelopmental disorders including both SCZ and autism.

This finding is consistent with an early developmental model for SCZ with widespread effects. Prenatal insults associated with SCZ risk are thought to occur during this period of gestation. Minor physical anomalies are presumed to be markers of aberrant brain development during the first or second trimester as they

are formed from the same primordial tissue (Lobato *et al.*, 2001). In particular anomalies of the craniofacial region, for which there is evidence of in SCZ (Ismail *et al.*, 1998, Weinberg *et al.*, 2007), are posited to be the result of an insult during the critical period of 9-16 weeks gestation (Waddington *et al.*, 1998). Differences in fingertip ridge count, observed in SCZ, are also thought to be caused by second trimester insults (Bracha *et al.*, 1992).

The identification of variable profiles across brain regions and genes highly consistently expressed appears contradictory. However, Figure 2.10 shows genes identified with variable expression across brain regions still have higher expression across the mid-foetal brain compared to all other genes in the dataset. These genes would have a higher mean expression value than the remaining genes. The mean was one of the global metrics used to identify consistently expressed genes and was part of the calculation of both the scaled mean and coefficient of variation. Therefore in the mid-foetal brain SCZ and BPD genes in general have higher expression that is relatively lower in the HIP and THAL compared to other brain regions. Moreover, disruptions in the expression of these genes could have fairly global, rather than specific effects on foetal neurodevelopment. Although higher gene expression was observed in neocortical regions when compared to other regions, in particular the HIP and THAL, there was no support for any particular neocortical region being involved in SCZ or BPD pathogenesis.

These associations were found by looking for a correlation between a summary expression metric and gene-level summarised p values calculated from GWAS results. These gene-wide p values which by definition are found in the interval [0, 1] were –log10 transformed to the interval [0, ∞) which emphasises the most significant genes, giving them more weight in the regression model used to test for associations. This transformation creates an unsymmetrical distribution of the gene-wide p values which may cause the assumptions of the regression model to no longer be satisfied, and can introduce outliers that may cause highly significant but noisy associations. In order to minimise the possibility of reporting false positive

results, additional nonparametric tests were used to validate any significant regression results reported in this chapter.

No consistent results were found for any analyses with SCZ CNVs, this may be surprising given CNVs have been shown not only to increase the risk of SCZ but also the risk of other developmental disorders. The results here should not be interpreted to mean that CNVs do not have an impact in the foetal brain; rather they do not appear to have a specific impact. Further, no associations were found between genes with evidence of alternative splicing in the foetal brain and either SCZ or BPD.

2.3.2 Identification of functional pathways from enriched expression gene sets

Pathway analysis in Section 2.2.7 identified three groups of terms associated with genes with decreased expression in the HIP or THAL: 'Chromosome: structural modification & repair', 'Transcription' and 'Post-transcriptional RNA processing & transport'. All three of these groups imply that these genes are involved in the regulation or control of gene expression particularly during the process of transcription. Genes with characteristic expression of the HIP and THAL within these pathways were found to be associated with SCZ and BPD, therefore this suggests that SCZ and BPD risk genes are involved in the regulation of gene expression. The cascade of brain development mechanisms is controlled by gene expression concentrations, therefore given the temporal profile described above for genes enriched for SCZ and BPD common variants, these genes may be involved in the control of expression throughout brain development. Furthermore, a disruption early on could have consequences later in life. The temporal profile of SCZ and BPD genes will be investigated further in the next chapter.

The coherence of the functional results was somewhat noteworthy. Given that GO is incomplete, as not all genes have been fully characterised (Khatri *et al.*, 2012), it would be expected that some true overlaps would be missing and therefore the groupings in Figures 2.6 and 2.7 would not be complete. While the HIP and THAL

scores were shown not to be entirely independent, there will be some variation between them which may lead to the identification of some spurious terms. Considering this, the concordance of the results across both sets increases the confidence of these results. Further the fact that the subset of genes with these characteristic expression profiles also present in the associated pathway groups were enriched for SCZ, and to some extent BPD GWAS signal supports the relevance of these categories for SCZ aetiology.

2.3.3 Comparison of results with schizophrenia and bipolar disorder variants

Given the genetic overlap of common variants discussed in the Introduction Section 1.3.2 it was not surprising that results were consistent for both disorders, and in fact the reason for including both was for genetic replication. Generally there was stronger evidence for enrichment of the association signal with the SCZ gene-wide p values compared to BPD gene-wide p values. In all likelihood this was due to the larger sample size in the SCZ PGC GWAS (approximately 9,400 SCZ cases and 12,500 controls compared to 7,500 BPD cases and 9,300 controls) giving the study more power to detect small effect sizes, which in turn meant there was more power in this study to detect associations.

2.3.4 Comparison of results with Brown's and Simes' gene-wide p values

Two sets of gene-wide p values were used in this chapter, based on Simes' multiple correction procedure (Simes, 1986) and Brown's method for combining correlated test statistics (Brown, 1975). Generally the results were more significant and more consistent with Brown's gene-wide p values compared to Simes' p values. This is supportive of a polygenic model for common variants as this suggests that recognizing and incorporating multiple, semi-independent association signals within genes leads to more significant associations, compared to just using the single best p value.

2.3.5 Summary of chapter findings

The original premise of this study was based on the fact that SCZ has been postulated as a human specific disorder (Crow, 1997) and sequences associated with human evolution were found to be enriched in genes differentially expressed in the foetal brain (Johnson *et al.*, 2009). While enrichments for SCZ risk genes were found in specific variable expression patterns, they were not found for general variation. Regions identified with human-specific evolutionary signatures have been investigated in another study with the PGC data, and no enrichment for SCZ or BPD common variants was found (Bigdeli *et al.*, 2013). The authors conjectured that using common variants limits the power to investigate this theory, and rarer variants identified from sequencing studies may be more appropriate, which may also be the case here.

In sum, the findings reported in this chapter suggest that SCZ risk genes play a role in the regulation of brain development particularly during foetal stages. Therefore in the next chapter, gene expression across the full range of brain development will be investigated to see if genes associated to SCZ or BPD have a common temporal profile.

2.4 Methods

All methods described below were completed with the R statistical language unless otherwise stated.

Preliminary data processing: Johnson

A previously published microarray dataset was downloaded as CEL files from the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO), accession number GSE13344. In total 95 samples from up to thirteen brain regions for four individuals, from both hemispheres, were hybridised to the Affymetrix Human Exon 1.0 chip to obtain genome-wide expression values, these data are referred to as the Johnson dataset. Further details on the quality control,

assessment of the tissue samples and microarray procedure can be found in the original manuscript (Johnson *et al.*, 2009).

Partek Genomics Suite software (Partek GS) was used to process the CEL files into expression values. The RMA algorithm with its default settings was used to transform intensity values into exon transcript values. Principal component analysis reduced the data into 3 dimensions to detect for outliers. By visual inspection no outliers were identified, so all samples were included in the analysis. The Tukey biweight one-step algorithm, with default settings, was then used to calculate gene transcript scores from the exon transcript values. At this stage the data were exported out of Partek GS.

All gene transcripts denoted as core and unique were annotated using messenger ribonucleic acid (mRNA) information of the corresponding exon probesets extracted from the annotation files provided online by Affymetrix (http://www.affymetrix.com/support/technical/annotationfilesmain.affx). These were then mapped to Entrez ids in the latest build of the genome (37.3), checking they matched the accompanying gene symbol, by a series of python scripts. Any Entrez ID annotated to a gene by more than 90% of the possible probesets was kept. Genes with multiple or non-unique Entrez IDs were removed, so that each Entrez ID was unique in the final dataset. From here gene transcripts will be referred to as genes. The gene expression values were filtered on a gene by gene basis to remove any genes that were lowly expressed, specifically anything with a median less than 3 or a maximum less than 5 done by another python script. The final dataset comprised of 16,212 genes.

Preliminary data processing: Kang

The Kang replication microarray dataset, accession number GSE25219, was also downloaded from GEO. This dataset contained post-mortem brains from individuals throughout development, but only those that fell between 18 and 23 weeks gestation were extracted to form the replication dataset. The Kang dataset and

Johnson dataset were known to have overlapping individuals; therefore any individual in the Kang dataset that matched an individual in the Johnson dataset based on age, gender and available brain regions was removed to leave five independent individuals. These data were downloaded already normalised, Kang *et al.* having also used Partek GS with the same default settings. The same annotation and filtering procedures were applied as described for the Johnson microarray dataset.

Calculating Brown's gene-wide p values from GWAS results

Brown's gene-wide p values were provided by V. Escott-Price having been calculated as described in (Moskvina *et al.*, 2011). Briefly, after excluding monomorphic SNPs, GWAS p values for all SNPs located within the start and stop position of each gene were combined into a single p value using Brown's method for combining dependent tests (Brown, 1975).

Calculating Simes' gene-wide p values from GWAS results

Gene locations were downloaded from the NCBI website for build 36.3. SNPs were filtered by minor allele frequency (MAF > 0.01) and INFO score (> 0.8) before all SNPs located within each gene's start and stop position along with their association p values for SCZ and BPD from the PGC GWAS were collated. Simes' procedure (Simes, 1986) for multiple comparisions was applied to all SNP p values and the most significant one after correction was taken as the Simes' p value.

Global metrics across the foetal brain

In their initial work Johnson *et al.* included covariates to control for differences between individuals and date of hybridisation between samples. For this analysis the first step was to regress out the effects of these variables. This was done using a linear regression model with the expression value for each gene taken as the dependent variable with individual and scan date the independent variables. As the residuals are standardised to have a mean of 0 they are unrelated to the original expression values, and therefore ranking genes by these values would be effectively

random. Hence for each sample the residuals were added to the intercept of the model for each gene. A similar procedure was applied to all Kang expression values to regress out the effects of RIN and PMI variables, consistent with the covariates used in their analysis.

Based on Johnson *et al.*'s findings explained in the main text, all samples from neocortical regions were combined by taking the median expression value for each individual, separately for each hemisphere, to replace the individual neocortical sample values. For each gene, across all individuals, the mean and scaled mean (mean divided by the maximum expression value) were calculated across the neocortical medians and non-neocortical samples. The coefficient of variation (standard deviation divided by the mean) was calculated separately for each individual and then averaged into a single value for each gene, to remove any effects of individual variation. In addition, the mean, scaled mean and coefficient of variation were calculated as described here but for the subset of neocortical samples.

Brain region characteristic scores

Brain region characteristic scores were derived from a linear model controlling for individual, hemisphere as well as the additional confounders specific to each dataset, shown in Equation 2.1.

$$\mathrm{exp}_{\mathrm{ijk}} = \mathrm{individual}_{\mathrm{i}} + \mathrm{hemisphere}_{\mathrm{k}} + \mathrm{confounder}_{\mathrm{ijk}} + \mathrm{brainregion}_{\mathrm{jl}}$$

Equation 2.1 Regression model used to calculate characteristic scores for brain region l, where $\mathrm{exp}_{\mathrm{ijk}}$ is the gene expression value for individual i, in brain region j and hemisphere k, and $\mathrm{brainregion}_{\mathrm{jl}}$ is a binary variable denoting whether that sample came from the brain region l or not i.e. are brain regions j and l the same.

In the Johnson dataset an additional covariate was included for hybridisation date, whereas for the Kang microarray data the additional covariates included were RIN and PMI. The model was fitted separately for each brain region, for each gene across all samples. The brain region term was a binary variable indicating whether the

expression value was taken from the brain region in question or not. Three samples from three individuals were excluded from the model fitting as the hemisphere was denoted NA. The characteristic p value was taken as the p value for the binary brain region term in the model. A characteristic score was defined for each gene as the –log10(characteristic p value) for the brain region term multiplied by the sign of the coefficient associated with the same term.

For regional characteristic scores found to be associated with either SCZ or BPD gene-wide p values, combined characteristic scores were calculated across the Johnson and Kang datasets. This used the model in Equation 2.1 and included hemisphere and individual covariates. There was no need to include a study covariate, as the individual term absorbed these differences. In addition the extra confounders were excluded, as these were not the same across the two datasets.

Alternative splicing - FIRMA

The FIRMA algorithm was used to identify alternative splicing (Purdom *et al.*, 2008). This method was implemented in R as part of the aroma.affymetrix package (Bengtsson *et al.*, 2008), following an online vignette provided by the authors (http://www.aroma-project.org/vignettes/FIRMA-HumanExonArrayAnalysis), to each individual separately as there was no option to include potential confounders.

After generating the FIRMA scores, the authors have proposed using the limma package (Smyth, 2005) in R, to fit a linear model and consider F–statistics to see if any tissue type, in this case brain region, has a non-zero coefficient. The p values for each exon and individual were combined into an overall global splicing p value by selecting the best one after Simes' correction. To calculate brain region splicing p values, the FIRMA scores from each brain region were compared to all others with a linear model and combined as described for the global splicing p values. Genes with only one exon were removed, as were genes lowly expressed prior to testing for associations with either SCZ or BPD variants.

Testing for enrichment with gene-wide p values

Mean expression, scaled mean expression, coefficient of variation, brain region characteristic scores and FIRMA gene level logP were tested in separate models with the PGC gene-wide logP. This was done for all four combinations of disease (SCZ and BPD) and summarised p value (Brown's and Simes'). Some of the SCZ Brown's p values had a value of 0 and could not be log transformed so were excluded from this analysis.

Nonparametric Mann-Whitney tests were used for validation of any significant regression results. Genes were ranked by mean expression, scaled mean expression, coefficient of variation and splicing p values separately and the top 5, 10,…, 50% was tested against the bottom 50% in a one-tailed test for smaller gene-wide p values. For regional characteristic scores found to be associated with either SCZ or BPD across both datasets, genes were ranked by the absolute values of the score and the top 5,10,…,50% taken. Each set of genes was then separated based on the sign of the characteristic score into two subgroups. If the characteristic score was positive, the genes detected had increased expression in that brain region relative to the average of all other brain regions. Alternatively if the score was negative the genes had decreased expression relative to the other brain regions. Depending on the coefficient of the significant regression model the appropriate subset was tested with a one sided Mann-Whitney test for smaller gene-wide p values against genes ranked in the bottom 50% by absolute characteristic score.

Overlap of significantly associated regions across datasets

Brain regions not present in both the Johnson and Kang datasets were removed for this analysis leaving 12: DLPFC, MPFC, OPFC, VLPFC, TAU, TAS, PAS, OCC, HIP, THAL, CBL and STR. Separately for each dataset the sample labels, containing region, individual and hemisphere information, were permuted 1000 times and the characteristic regression model in Equation 2.1 fitted for each region. Characteristic scores were calculated as for the true data and tested with a linear regression model predicting SCZ and BPD Brown's logP for each brain region.

Empirical p values were calculated for each dataset by counting how many times in the 1000 permuted datasets, the same number or more, significant models occurred ($p < 0.01$). All pairs of permutations (10^6 in total) were then considered and the number of regions significant in both datasets and in the same direction was counted so that empirical p values could be calculated. In addition, the number of regions significant for both SCZ and BPD, in the same direction across both datasets, was also counted.

CNV logistic regression

CNVs from the ISC (International Schizophrenia Consortium, 2008) and MGS (Levinson *et al.*, 2011) study were provided annotated with the genes hit by each by A. Pocklington based on builds 35.1 and 36.3 respectively. For each CNV the corresponding global metrics, characteristic scores and –log 10 FIRMA gene-level splicing p values for each gene hit were taken, and the minimum, median and maximum identified. Only CNVs that hit genes found in the expression datasets were used for these analyses. In order to test for a significant difference in each of these metrics the following logistic regression model was fitted across the combined set of CNVs from the ISC and MGS. The p value was taken from the M(genes hit) term and the coefficient considered to identify the direction of any associations. Separate models were fitted for all CNVs in addition to the deletions and duplications subsets.

$$\text{case CNV} = \text{M(genes hit)} + \text{N} + \text{study} + \text{chip}$$

Equation 2.2 Logistic regression model to test if genes hit by SCZ CNVs have common characteristic profile, where case CNV is the case status of the individual the CNV was found in, M(genes hit) is either the minimum, median or maximum metric of the genes hit and N is the number of genes in the expression dataset hit by the CNV.

Each SCZ CNV was compared to all control CNVs in the ISC and MGS datasets to see if there was any overlap with any control CNV or was unique to SCZ cases. Those that did not overlap with any control CNV were classed as singletons. These singleton

CNVs were then compared to all remaining SCZ CNVs with a logistic regression model defined as follows.

$$\text{singleton CNV} = \text{M(genes hit)} + \text{N} + \text{study} + \text{chip}$$

Equation 2.3 Logistic regression model to test if genes hit by singleton SCZ CNVs have common characteristic profile, where singleton CNV denotes whether the CNV was singleton or not, M(genes hit) is either the minimum, median or maximum metric of the genes hit and N is the number of genes in the expression dataset hit by the CNV.

As with the previous CNV logistic regression model, the p value was taken from the M(genes hit) term and the coefficient considered to identify the direction of any associations. This model was fitted for all CNVs, as well as deletions and duplications separately.

Functional analysis

Using the HIP and THAL combined characteristic scores (those calculated across both the Kang and Johnson datasets) functional pathways from the GO database were tested to identify those with genes with decreased expression in either the HIP or THAL. A file containing annotation categories and their associated genes was provided by P. Holmans. Each annotation term was tested with a one-sided Mann-Whitney test to see if genes within that category had lower HIP or THAL characteristic scores than genes not in the category. Only genes annotated to at least one GO term and terms with between 20 and 2000 genes were considered in these Mann-Whitney tests.

Significant terms identified from a Bonferroni corrected p value of 1.56×10^{-5} were combined into groups of related terms, to identify the predominant functions represented. In each set the smallest pathway was selected and all other pathways with any overlapping genes identified. The Mann-Whitney test was rerun for each of the larger pathways, removing any genes that overlapped with the smaller pathway to see if it remained significant ($p < 0.01$). If it did not, the smaller pathway was said to explain the larger pathway. All explained larger pathways were combined with the

smaller pathway they were explained by into a merged pathway and the process was repeated until no more pathways could be explained.

Set-based tests

Genes with decreased expression in the top 35% of HIP and THAL characteristic genes were identified as the most significantly enriched sets from the Mann-Whitney tests in Figure 2.9 for both SCZ and BPD Brown's p values. Each set was split into two groups, those that overlapped with a significant GO pathway found with that characteristic score and those that did not. For both subsets, in pathways and not in pathways, set-based p values were calculated. In addition each group of terms in Figures 2.6 and 2.7 had a set-based p value calculated for genes that intersected with either the top HIP or THAL enriched sets.

GWAS results from PGC SCZ and BPD were filtered to remove SNPs with MAF less than 0.01, INFO score less than 0.8 and all SNPs in the MHC region (chr6:25000000-35000000). All remaining genic SNP p values were corrected for genomic inflation. Set-based p values were calculated from these corrected p values using the undocumented –set-screen command in PLINK (Brown, 1975, Purcell *et al.*, 2007, Moskvina *et al.*, 2011). HapMap genotype data release 23 for Europeans only were downloaded (http://hapmap.ncbi.nlm.nih.gov/) and used for the LD statistics involved in the calculation.

Simulations for set-based tests

After observing a correlation between set p value and the number of SNPs in each set, simulations were used to adjust the set p values. Simulated sets with between 20 and 3000 random genes were created and a set-based p value calculated for each. A regression model was then fitted to predict the set log p value from the number of SNPs. Using this relationship, for each true set the predicted log p value was calculated and subtracted from the true log p value. If the residual was positive i.e. the true p value was more significant than the predicted p value then the residual

was unlogged to derive the adjusted p value. A separate set of simulations and adjustment formula was performed for the PGC SCZ and the PGC BPD.

Chapter 3: Expression patterns of schizophrenia and bipolar disorder risk genes throughout human brain development

3.1 Introduction

3.1.1 Background

The human brain continues to develop throughout postnatal life (Stiles and Jernigan, 2010). During early years the brain rapidly increases in volume (Dobbing and Sands, 1973, Matsuzawa *et al.*, 2001, Knickmeyer *et al.*, 2008), while synaptogenesis and pruning modify connectivity throughout the first two decades of life (Glantz *et al.*, 2007). Therefore, as the brain is still developing through the period of onset for SCZ and BPD in adolescence or early adulthood it is reasonable to look at expression across these stages in addition to the earliest stages of brain development.

Kang *et al.* conducted a study covering the full range of brain development including foetal post-mortem brains and found 86% of genes surveyed were expressed in at least one region in one of their defined development stages (Kang *et al.*, 2011). Almost 90% of these expressed genes were differentially expressed across development stages with the major differences occurring between foetal stages and either postnatal or adult stages. An accompanying study, based on more individuals but with just one prefrontal cortex sample per individual, showed that the highest rates of change in expression occurred during foetal development (Colantuoni *et al.*, 2011).

Results from Chapter 2 found that genes enriched for either SCZ or BPD common variants were highly expressed in the mid-foetal brain. Visual inspection of the median expression of these genes across development showed a variable profile with peak expression values during the second trimester and a drop in expression during the third trimester. In this chapter a more formal analysis was conducted to see if genes associated with either SCZ or BPD were enriched in sets of genes with

common developmental expression profiles. The findings in the previous chapter suggest that SNPs contribute to prenatal expression, however in this chapter all stages across the full range of human life were considered to see if they also contributed to the expression patterns in other stages of development such as adolescence.

3.1.2 Outline

Aim

In this chapter, genes with variable profiles across the full range of human brain development were tested for enrichment of SCZ or BPD risk genes, in order to characterise a developmental expression profile. Results from Chapter 2 suggest that expression patterns during prenatal stages will be important for SCZ and BPD. Both SCZ and BPD present symptoms around adolescence coinciding with the maturation of brain structures, such as the prefrontal cortex, associated with the executive functions that are commonly impaired in SCZ (Orellana and Slachevsky, 2013) therefore differences at this time point would also be of interest.

Datasets

BrainSpan: Atlas of Developing Human Brain (http://www.brainspan.org) is a publically available internet resource providing RNA-Seq expression data to enable studies of the developing brain transcriptome. This resource contains gene expression data on 41 individuals from 8 PCW to 41 years with up to 16 different brain regions for each individual. Microarray expression data taken from the Kang *et al.* study introduced in the previous chapter were also used. This dataset provided expression values for multiple brain regions, from both hemispheres, for 57 individuals from embryonic to late adulthood (Kang *et al.*, 2011). Almost all of the BrainSpan individuals were also present in the Kang dataset; therefore this was used for technical not biological replication.

RNA-Seq data are generated through the direct sequencing of transcripts. The resulting data are more reliable and accurate than that of microarrays, which suffer from cross hybridisation and are limited to detecting the transcripts represented on the chip (Wang *et al.*, 2009). In both datasets individuals were classified into 15 development stages from embryonic to late adulthood as defined in the original Kang *et al.* manuscript, shown in Table 3.1. No individuals in the BrainSpan data were part of the embryonic, middle adulthood or late adulthood stages.

Development stage	**Time period**	**Development stage**	**Time period**
Embryonic	[4 pcw, 8 pcw)	Neonatal and early infancy	[birth, 6 mon)
Early foetal A	[8 pcw, 10 pcw)	Late infancy	[6 mon, 12 mon)
Early foetal B	[10 pcw, 13 pcw)	Early childhood	[12 mon, 6 yr)
Early mid-foetal A	[13 pcw, 16 pcw)	Middle and late childhood	[6 yr, 12 yr)
Early mid-foetal B	[16 pcw, 19 pcw)	Adolescence	[12 yr, 20 yr)
Late mid-foetal	[19 pcw, 24 pcw)	Young adulthood	[20 yr, 40 yr)
Late foetal	[24 pcw, 38 pcw)	Middle adulthood	[40 yr, 60 yr)
		Late adulthood	[60 yr,]

Table 3.1: Development stages as defined by Kang *et al.*
pcw – post conception weeks, mon – months, yr- years.

GWAS data from the SCZ and BPD PGC studies (Ripke *et al.*, 2011, Sklar *et al.*, 2011), summarised as described in Chapter 2 into Brown's and Simes' gene-wide p values, were also used in this chapter to test for associations. In addition, GWAS data from similar large studies for Alzheimer's disease (Harold *et al.*, 2009), and Parkinson's disease (UK sample from (Nalls *et al.*, 2011)) were combined into gene-wide p values using Brown's method by V.Escott-Price and permission given for use. All summarised p values were -log10 transformed and shall be referred to as logP. CNV data from the ISC (International Schizophrenia Consortium, 2008) and MGS (Levinson *et al.*, 2011) studies were also included as an alternative type of variant, to test for enrichment in a manner similar to that of the previous chapter.

Outline of analysis

Temporal expression patterns were quantified using regression in a similar fashion to the brain region analysis of Chapter 2. Characteristic scores for each gene in each development stage were derived from linear models, to indicate whether genes were upregulated or downregulated during each period. These were tested for relationships with SCZ and BPD gene-wide logP and genes hit by SCZ CNVs.

One unifying explanation of the polygenic nature of SCZ or BPD whereby many variants that only slightly increase an individual's risk of SCZ or BPD are distributed across both genes and the genome, is that these variants work in combination and affect common pathways or mechanisms (Sullivan, 2012). In order, therefore, to have a noticeable influence it would be expected that risk genes would be co-expressed. This was tested by taking genes highly co-expressed with risk genes identified from the PGC GWAS, the most powerful study and therefore most likely to be true associations, to investigate if they were enriched for association signal. If enrichment was found in the co-expressed gene sets their developmental profile could be looked at and results compared with that of the previous analysis of development stage characteristic genes, ideally looking for a convergent profile.

All these methods were performed initially in the BrainSpan RNA-Seq dataset and followed up in the Kang microarray dataset. Both parametric and nonparametric tests were used to validate results. Functional analysis was then performed on genes enriched for SCZ or BPD variants to elucidate potential mechanisms and interpret these results.

3.2 Results

3.2.1 Development stage characteristic gene expression and common risk variants

In Chapter 2, genes with expression characteristic of each brain region in the foetal brain were identified. In this chapter a very similar approach was used in the BrainSpan RNA-Seq dataset to identify genes with an increase or decrease of

expression during each development stage described in Table 3.1, relative to all other time points. Here characteristic scores, defined in the same way as the previous chapter, were taken from a linear model with a binary development stage term and an additional covariate controlling for brain region differences.

For each development stage, a linear regression model was used to test for a significant association between its characteristic scores and Brown's logP. Figure 3.1 panel A shows these results, where positive relationships indicating upregulation are represented by bars above the origin and negative relationships indicating downregulation are represented by bars below the origin. There was a clear pattern to these results, which broadly suggests that SCZ risk genes are upregulated in early and mid-foetal stages after which their expression levels decrease until early childhood during which they are downregulated. Nine of the twelve stages had significant linear relationships with SCZ Brown's logP after correcting for testing twelve models. In particular the most significant enrichments were found in genes characteristic of early childhood (corrected $p = 1.55 \times 10^{-15}$) and early mid-foetal A (corrected $p = 9.36 \times 10^{-9}$). When testing BPD Brown's logP, the directions of effect were consistent with the results for SCZ for ten stages, overall showing a parallel pattern of results, although less significantly. Only two stages were significant after correcting the number of development stages tested, both of which were also significantly associated with smaller SCZ p values. As in Chapter 2, the correlation coefficients associated with these significant relationships were small, all absolute values < 0.07; see Appendix Table 8.1. Generally these were larger for SCZ logP than BPD logP consistent with the higher number of significant developmental stages.

The same pattern of enrichment was seen across all stages when testing SCZ Simes' logP instead of Brown's logP, see Figure 3.1 panel B. For SCZ seven stages had significant regression models, all of which were significant with the Brown's logP in the same direction. Early childhood (corrected $p = 1.49 \times 10^{-7}$) and early mid-foetal A (corrected $p = 1.36 \times 10^{-7}$) were again the most significant stages. For BPD although no regression models were significant after correcting for twelve development

stages, the directions of effect were consistent with the SCZ results and supported the pattern described.

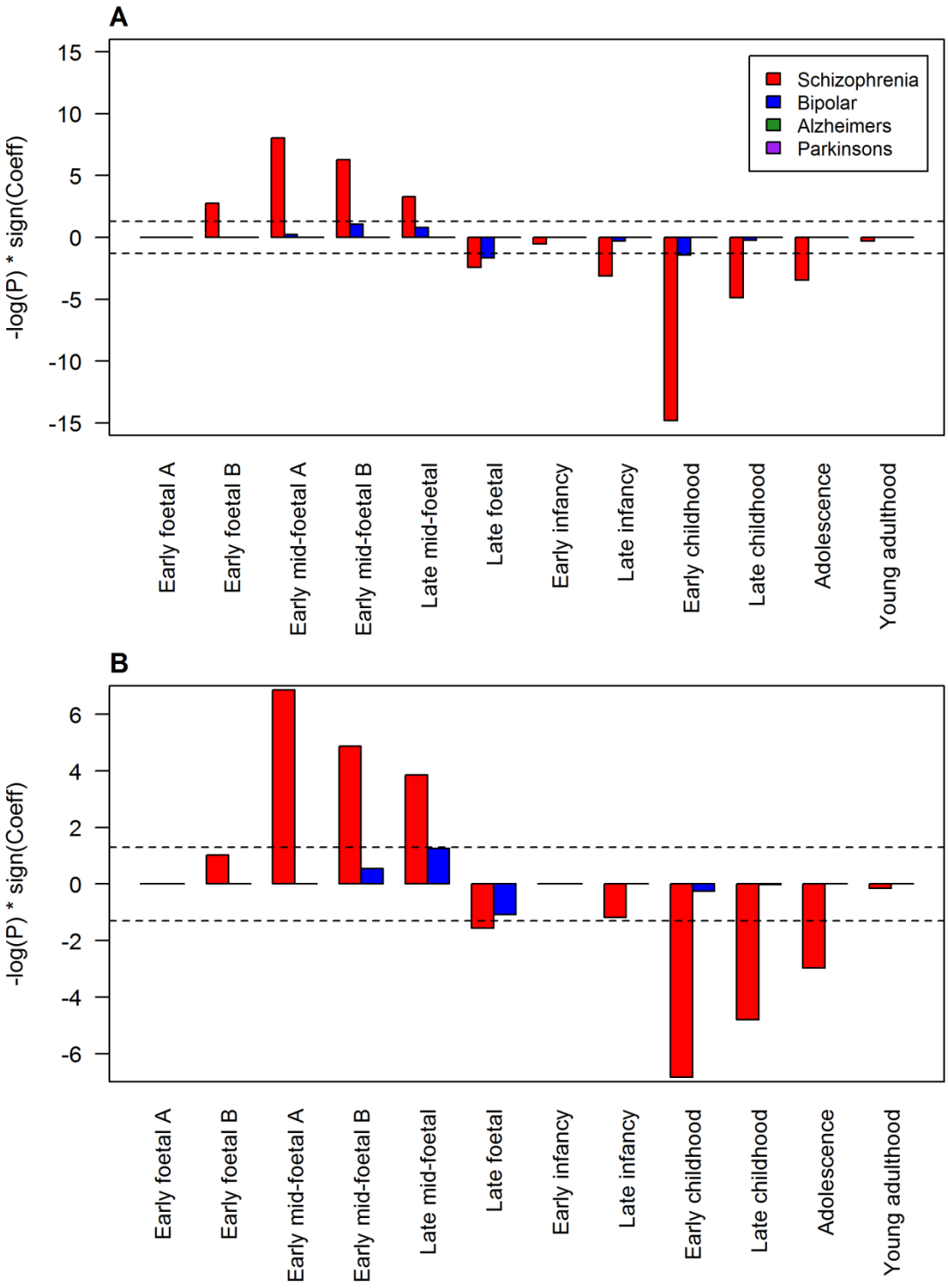

Figure 3.1: Linear regression results testing development stage characteristic scores calculated in the BrainSpan RNA-Seq dataset.
P values were –log10 transformed and multiplied by the sign of the coefficient, therefore bars above the origin indicate positive regression coefficients; bars below the origin indicate negative regression coefficients. Panel A tested Brown's logP; panel B tested Simes' logP. All p values were corrected for 12 development stages using Bonferroni's method. Black dashed line is p = 0.05.

As in Chapter 2 rank-based tests were used to verify the associations reported between characteristic scores and gene-wide p values with a regression approach. Mann-Whitney tests showed that genes with increased expression during foetal development and genes with decreased expression in postnatal stages were enriched for smaller gene-wide p values; see Appendix Figures 8.1 and 8.2. All significant regression models in Figure 3.1 were verified and shown not to be due to extreme values, supporting the results profile already described.

As in the previous chapter these analyses were repeated removing genes located in the MHC. While this reduced the significance of some of the linear regression models, the overall results profile remained. The impact on the Mann-Whitney tests was minimal and hence it can be concluded that this results pattern was not due to the correlated association of genes in the MHC, results presented in Appendix Figures 8.3-8.5.

Comparison with other disorders

The above analyses suggest that SCZ risk genes have a variable profile across brain development with increased expression during mid-foetal stages and decreased expression from birth. Findings with BPD p values were only nominally significant but consistent with this profile. In this section other psychiatric disorders were tested to assess whether the association of a developmental expression profile was specific to SCZ and BPD. Parkinson's disease and Alzheimer's disease are adult neurodegenerative disorders with onset much later than SCZ and BPD, rarely occurring before 50 for Parkinson's (de Lau and Breteler, 2006) or 65 for Alzheimer's disease (Reitz *et al.*, 2011), therefore Brown's p values for these diseases were used as negative controls.

No significant results were observed for Parkinson's Brown's logP, shown in Figure 3.1 panel A. Initially Alzheimer's disease Brown's logP were associated with genes with increased expression during early childhood (corrected $p = 6.93 \times 10^{-7}$),

although this was not validated with the Mann-Whitney tests. Further investigation found that the regression signal was biased by two highly significant Brown's logP for *APOE* and *APOC1*. When these genes were removed from the analysis there were no significant enrichments for Alzheimer's disease logP. This advocates the use of both parametric and nonparametric approaches to verify that associations were not due to extreme values.

Technical validation with microarray data

Characteristic scores were also calculated for each development stage in the Kang microarray dataset from linear models that included extra covariates for hemisphere, PMI and RIN. Figure 3.2 presents the results of testing for enrichment with these characteristic scores. These results showed that genes with smaller SCZ Brown's p values were upregulated in early foetal development and downregulated during early childhood, consistent with the results in the RNA-Seq dataset. Results with BPD Brown's p values were consistent with risk genes being downregulated during middle and late childhood. In contrast to Figure 3.1, genes upregulated during adolescence and young adulthood were enriched for more significant SCZ p values which was opposite to the direction of effect reported in the RNA-Seq dataset. Therefore, although the microarray data during early foetal B and childhood was coherent with the results described in the RNA-Seq dataset, the results profile across the development stages was not technically validated across the two datasets.

Results testing Simes' p values in Figure 3.2 panel B were broadly consistent with downregulation during early childhood for SCZ risk genes. Interestingly, early mid-foetal B characteristic scores were positively associated with Simes' SCZ logP rather than early foetal B which were associated with Brown's SCZ p values. Although there was no association with young adulthood scores, there was an enrichment in genes with decreased expression during late adulthood. BPD Simes' p values were also associated with genes downregulated during late childhood consistent with that found with Brown's p values. Mann-Whitney tests with both Brown's and Simes' p values verified most of these findings; see Appendix Figures 8.6 and 8.7.

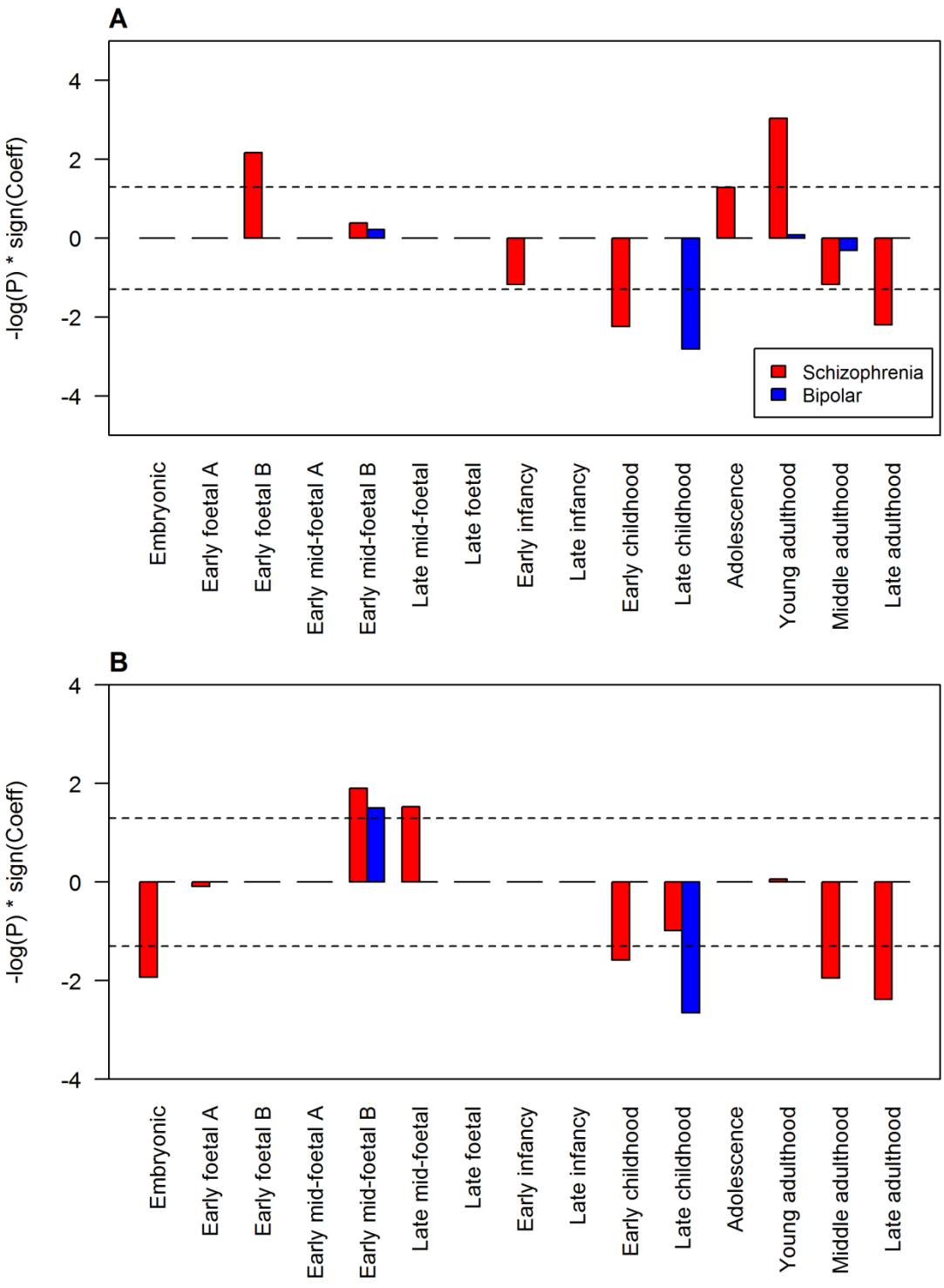

Figure 3.2: Linear regression results testing development stage characteristic scores calculated in Kang microarray dataset.

P values were –log10 transformed and multiplied by the sign of the coefficient, therefore bars above the origin indicate positive regression coefficients; bars below the origin indicate negative regression coefficients. Panel A tested Brown's logP; panel B tested Simes' logP. All p values were corrected for 15 development stages using Bonferroni's method. Black dashed line is p = 0.05.

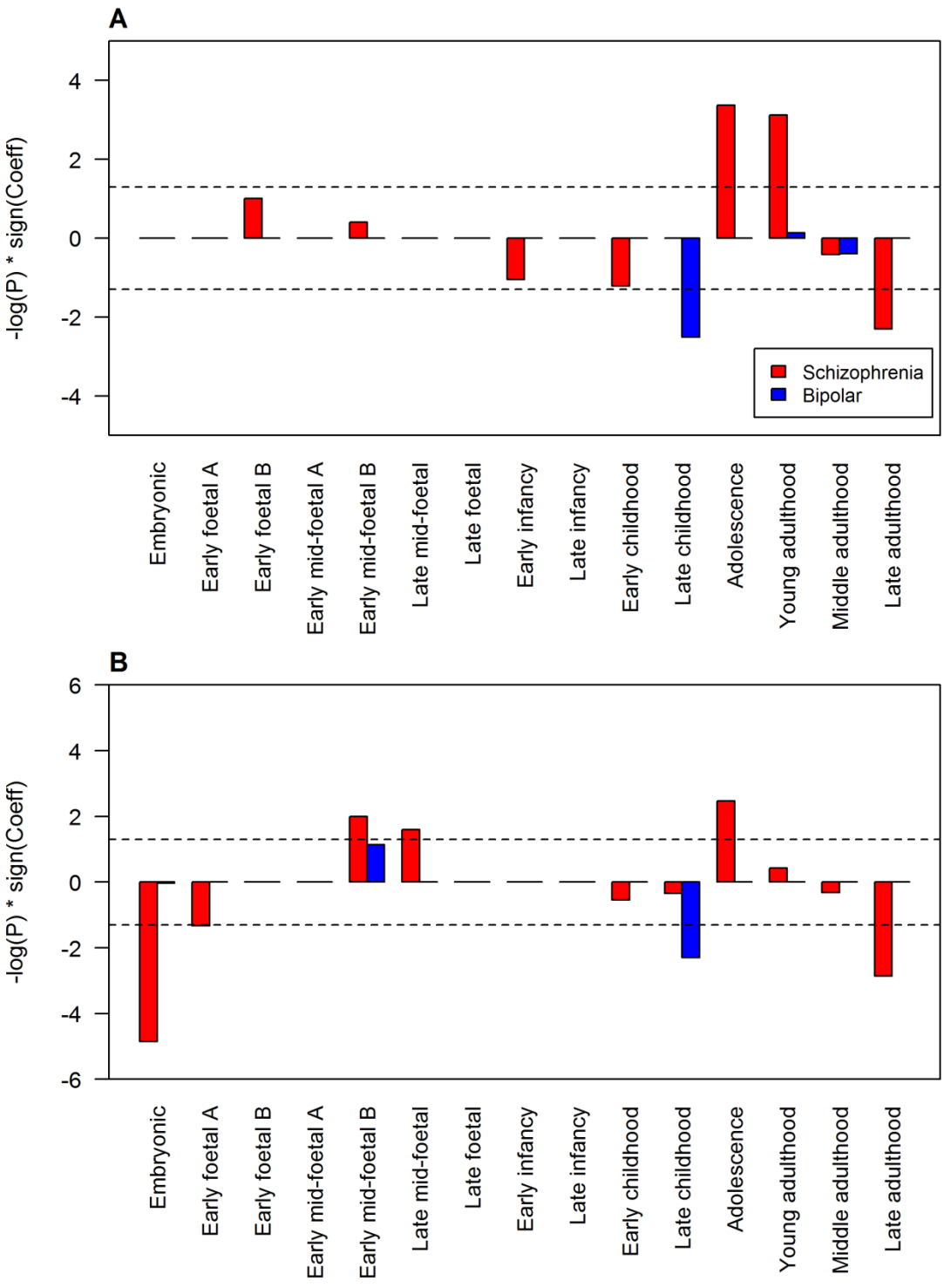

Figure 3.3: Linear regression results testing development stage characteristic scores calculated in Kang microarray dataset, excluding MHC genes.

P values were –log10 transformed and multiplied by the sign of the coefficient, therefore bars above the origin indicate positive regression coefficients; bars below the origin indicate negative regression coefficients. Analysis was run excluding MHC genes. Panel A tested Brown's logP; panel B tested Simes' logP. All p values were corrected for 15 development stages using Bonferroni's method. Black dashed line is p = 0.05.

Removing genes from the MHC region produced a similar profile of results, however genes with increased expression in foetal development or decreased expression during early postnatal years were no longer significantly enriched for SCZ Brown's logP, see panel A of Figure 3.3. There was, however, still enrichment for BPD Brown's logP in genes with decreased expression during late childhood. Further, there was still an enrichment for more significant SCZ Simes' p values in genes with increased expression during mid-foetal development, see panel B of Figure 3.3. Mann-Whitney tests broadly supported the significant associations in Figure 3.3; see Appendix Figures 8.8 and 8.9. Therefore in this dataset, the MHC genes appear to be having a greater effect on the results.

Effect of post-mortem interval covariate

Two potential confounders associated with the quality of the post-mortem brain samples were included when calculating the characteristic scores. One of these, PMI was observed to correlate with the age of the sample, and therefore may have been removing some of the temporal effects the characteristic scores were designed to capture. To examine the impact of this, characteristic scores were recalculated omitting this variable and tested as previously described.

Excluding the PMI variable produced a results profile more in line with that found in the RNA-Seq dataset, see Figure 3.4. Genes upregulated during early and mid-foetal development and genes downregulated in early postnatal years were enriched for smaller SCZ Brown's p values. The results testing BPD p values were consistent with this profile, as were results testing Simes' logP. Again, the correlation coefficients, shown in Appendix Table 8.5 were small, all absolute values < 0.05, suggesting that these expression patterns are not true for all risk genes. Rank-based tests verified all significant regression models; Appendix Figures 8.10 and 8.11. Therefore, the PMI covariate does appear to absorbing some of the temporal expression profile.

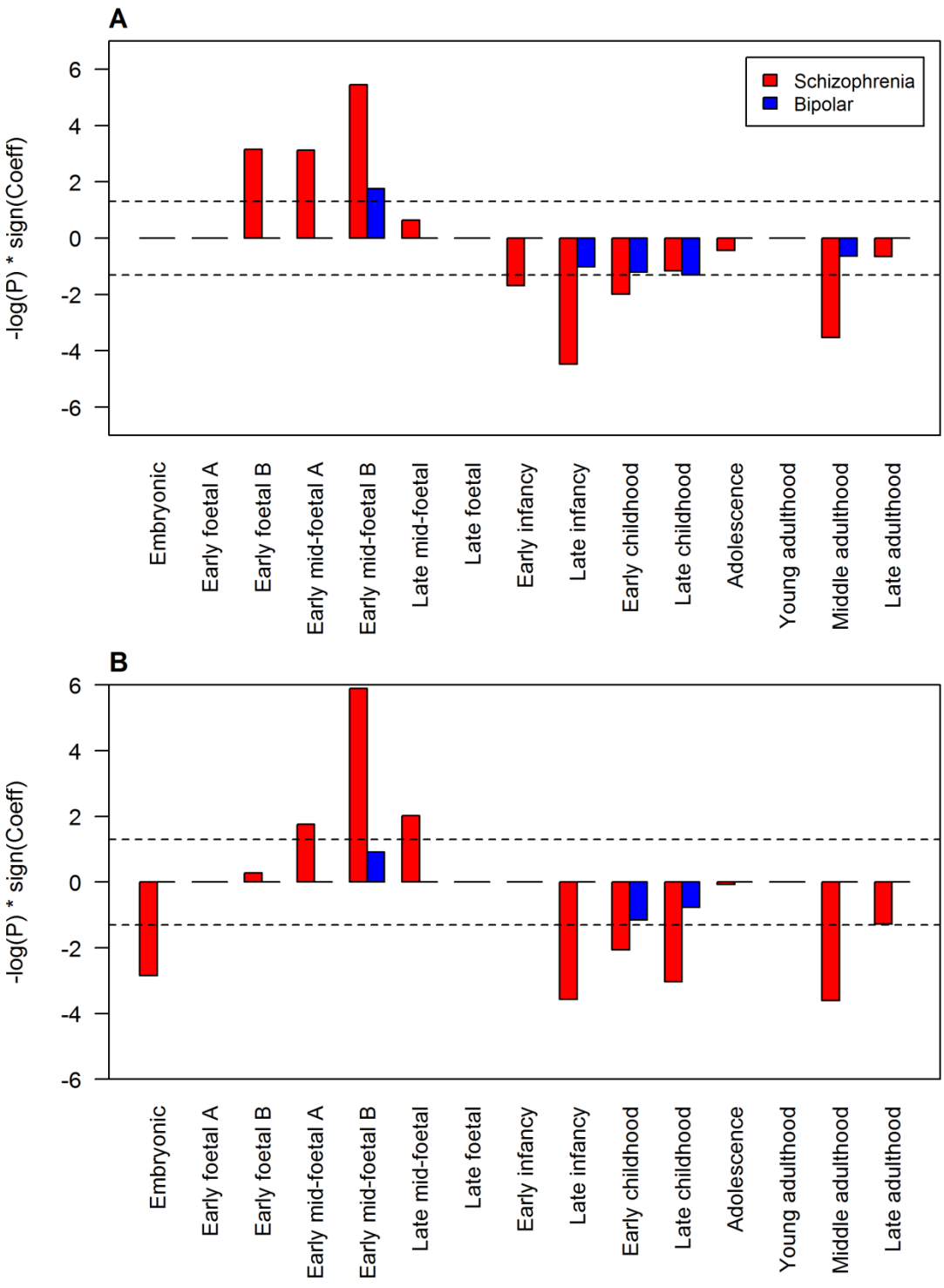

Figure 3.4: Linear regression results testing development stage characteristic scores calculated in Kang microarray dataset without PMI covariate.

P values were –log10 transformed and multiplied by the sign of the coefficient, therefore bars above the origin indicate positive regression coefficients; bars below the origin indicate negative regression coefficients. Panel A tested Brown's logP; panel B tested Simes' logP. All p values were corrected for 15 development stages using Bonferroni's method. Black dashed line is p = 0.05.

Removing genes located in the MHC region did reduce the significance of some of the regression models, but the same broad pattern was present indicating that genes outside of the MHC region associated to SCZ had increased expression in mid-foetal stages and decreased expression after birth, presented in Appendix Figure 8.12. These associations without the MHC genes were also verified with Mann-Whitney tests, see Appendix Figures 8.13 and 8.14.

These results do suggest that the PMI covariate was removing some of the temporal effects, however it was not possible to disentangle whether the temporal expression profiles enriched for SCZ variants were genuine or due to the differences in PMI between the samples.

Test for independent associations

So far each characteristic profile has been considered separately and a general pattern of results has been found, where genes with increased expression during early and mid-foetal development, and genes with decreased expression in postnatal stages were associated with SCZ common variants. As characteristic scores were calculated relative to the expression in all other stages, they were to some degree correlated; therefore the reported associations may be correlated also. In the RNA-Seq dataset the strongest enrichment was found with early childhood scores (uncorrected $p = 1.29 \times 10^{-16}$). The eight other significant stages were tested in a pairwise manner alongside the early childhood scores to see whether both scores remained significant and the associations were independent.

Table 3.2 shows that in almost all models (each one represented by a row) the early childhood scores remained highly significant whereas the paired characteristic scores did not. The exception was late foetal, which also remained significant. This showed that the associations demonstrated in Figure 3.1 generally were not independent, and were detecting enrichment in similar sets of genes which had increased expression through foetal development followed by a decrease through birth and early postnatal years to the lowest expression values in early childhood.

Development stage	Development stage		Early childhood	
	P value	Coeff.	P value	Coeff.
Early foetal B	0.569	-	1.66×10^{-13}	-
Early mid-foetal A	0.179	+	1.22×10^{-8}	-
Early mid-foetal B	0.359	+	3.56×10^{-10}	-
Late mid-foetal	0.329	+	3.79×10^{-13}	-
Late foetal	0.00206	-	7.36×10^{-16}	-
Late infancy	0.189	+	1.90×10^{-13}	-
Late childhood	0.114	-	6.57×10^{-12}	-
Adolescence	0.749	-	8.82×10^{-13}	-

Table 3.2: Linear regression results testing development stage characteristic scores calculated in the BrainSpan RNA-Seq dataset simultaneously predicting SCZ Brown's logP.
Each row represents a separate regression model.

In the Kang microarray dataset, the most significant stage was young adulthood (uncorrected $p = 6.13 \times 10^{-5}$) so all other significant stages were tested simultaneously with this stage to predict SCZ Brown's logP. Interestingly, and in contrast to the RNA-Seq results, all stages remained significant implying that these were independent associations, Table 3.3.

	Development stage		Young adulthood	
	P value	Coeff.	P value	Coeff.
Early foetal B	7.08×10^{-5}	+	9.81×10^{-6}	+
Early childhood	0.000347	-	5.57×10^{-5}	+
Late adulthood	2.05×10^{-6}	-	3.12×10^{-7}	+

Table 3.3: Linear regression results testing development stage characteristic scores calculated in Kang microarray dataset simultaneously predicting SCZ Brown's logP.
Each row represents a separate regression model.

	Development stage		Early mid-foetal B	
	P value	Coeff.	P value	Coeff.
Early foetal B	0.117	+	0.000385	+
Early mid-foetal A	0.183	+	0.000527	+
Early infancy	0.0763	-	9.55×10^{-6}	+
Late infancy	0.0328	-	0.00286	+
Early childhood	0.00205	-	6.77×10^{-7}	+
Middle adulthood	0.368	-	0.00234	+

Table 3.4: Linear regression results testing development stage characteristic scores calculated in Kang microarray dataset without PMI covariate simultaneously predicting SCZ Brown's logP.
Each row represents a separate regression model.

These pairwise analyses were also done with the microarray characteristic scores calculated without the PMI covariate, where the most significantly associated stage was early mid-foetal B (uncorrected $p = 2.39 \times 10^{-7}$). Most stages, see Table 3.4, were not significant after controlling for the early mid-foetal association implying that many of these associations were not entirely independent.

Summary

Genes with increased expression during mid-foetal development and decreased expression around birth and in postnatal stages were shown to be enriched for common SCZ risk variants. This was primarily found in the RNA-Seq dataset and verified with both parametric and nonparametric tests. When these analyses were repeated in a microarray dataset containing overlapping individuals, although the number of significant stages was less, those that were significant were consistent with the results profile for SCZ risk genes described in the RNA-Seq dataset. The lack of complete coherence between the RNA-Seq and microarray datasets was in part due to the inclusion of the PMI covariate, which was observed to correlate with the age of the sample. Removing this covariate produced a results profile more similar to that found with the RNA-Seq data. In addition, microarray data have reduced sensitivity and are not able to detect the small differences that RNA-Seq data can, which may also have contributed to the more significant results in the RNA-Seq dataset.

The association of each set of characteristic scores in the RNA-Seq dataset or microarray dataset without the PMI covariate was not completely independent and suggests that a common set of genes underlie the significant results pattern. Generally only nominally significant results were found when testing BPD p values but these were consistent with the pattern seen with SCZ p values, whereas no significant associations were found for either Alzheimer's disease or Parkinson's disease. Therefore, SCZ risk genes have been shown to have a common variable expression profile across brain development.

3.2.2 Schizophrenia risk genes co-expression models

So far specific expression profiles across brain development have been identified and enrichments found for SCZ associated genes. In this section genes co-expressed with known SCZ risk genes were identified to see if they were enriched for disease risk.

Ten independent genome-wide significant SNPs were reported in the PGC manuscript across the two stages of the study, with the nearest gene for each identified (Ripke *et al.*, 2011). These genes were chosen as having the strongest current evidence as they were identified in the largest published GWAS for SCZ to date. Of these *TRIM26* is located in the MHC region and was excluded from consideration (and eventual analyses) as it is unclear where the true association in this region is. Two genes, *PCGEM1* and *MIR137*, were not present in the BrainSpan dataset leaving seven genes to be considered: *CSMD1*, *CNNM2*, *NT5C2*, *TCF4*, *CCDC68*, *MMP16* and *STT3A*.

Genes co-expressed with schizophrenia risk genes

Taking the expression profile for each of these genes across brain development, a linear regression framework was used to compare the expression of all remaining genes in the dataset. P values were calculated for all genes in the dataset, excluding those considered for the co-expression model, indicating how closely their expression profile correlated across development with the expression of each risk gene, further details in the Methods. These co-expression p values, referred to as model p values, were then used to test for association with the Brown's logP.

Genes whose expression correlated with either *CNNM2* ($p = 1.27 \times 10^{-6}$), *NT5C2* ($p = 1.58 \times 10^{-6}$), *MMP16* ($p = 0.0104$) or *TCF4* ($p = 5.71 \times 10^{-5}$) were enriched for SCZ associated common variants, while the *CSMD1* model logP showed a trend for significance ($p = 0.0643$), results in Table 3.5. Only *CCDC68* had a negative coefficient. Similar to all previous analyses with gene-wide logP the correlation coefficients were small. Nonparametric tests were then used to verify these associations, see Figure 3.4. *CNNM2*, *CSMD1*, *TCF4, NT5C2* and *MMP16* all showed

evidence of enrichment with this approach that was not driven by extreme values, while *STT3A* did not. In sum, five of the seven genes considered at this stage were enriched for SCZ risk supportive of the idea that genes harbouring associated variants are co-expressed.

Gene Model	**CSMD1**	**CNNM2**	**NT5C2**	**CCDC68**
P value	0.0643	1.27×10^{-6}	1.58×10^{-6}	0.139
Correlation Coeff.	0.0164	0.0430	0.0426	-0.0132
	+	+	+	-
Gene Model	**MMP16**	**TCF4**	**STT3A**	
P value	0.0104	5.71×10^{-5}	0.276	
Correlation Coeff.	0.0227	0.0357	0.00968	
	+	+	+	

Table 3.5: Linear regression results and correlation coefficients testing single SCZ risk gene co-expression model logP calculated in the BrainSpan RNA-Seq dataset with SCZ Brown's logP.

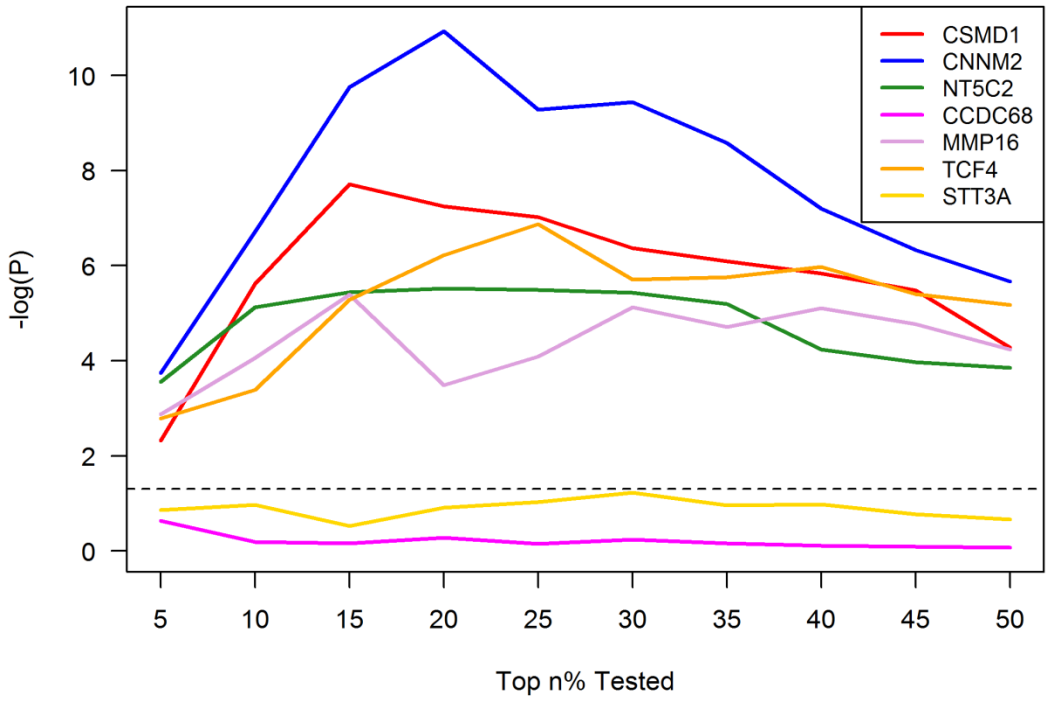

Figure 3.5: Results from Mann-Whitney tests for genes ranked by single SCZ risk gene co-expression model p values calculated in the BrainSpan RNA-Seq dataset.
Genes ranked by SCZ single risk gene co-expression model p values and top n% tested for smaller SCZ Brown's p values against bottom 50%. Black dashed line is p = 0.05.

Identifying alternative genes

Typically, the gene located closest to each associated SNP is assumed to be the most functionally relevant, however this may not always be the case. Taking the two genes for which co-expression did not index disease association (*CCDC68* and *STT3A*), additional proximal genes to the SNP associated in PGC study were tested to see if they showed evidence for a correlation between co-expression and SCZ association. SNPs in strong LD with those reported in the PGC paper were identified using the data from HapMap phases 1, 2 & 3.

For *CCDC68*, a region was taken ~350 kb either side of rs12966547 to include the two neighbouring genes *RAB27B* and *TCF4*. Five SNPs were found to be in perfect LD (r^2 = 1, D' = 1) with the genome-wide significant SNP. Three of these were found closest to *CCDC68* (rs4131791, rs4309482, rs12969453) and the other two approximately halfway (>100 kb) between *CCDC68* and *TCF4* (rs11874716, rs4891131) for which co-expression had already been shown to index SCZ association, see Table 3.4. If risk enrichment of co-expressed genes was used as a measure of SCZ relevance, then it would predict that the association at rs12966547 is most likely through TCF4.

For *STT3A*, a region of 100kb on chromosome 11 containing *EI24* and *CHEK1* around rs548181 was taken. In this region five SNPs were found in perfect LD, three of which were closest to or within *STT3A* (rs503288, rs513209, rs540723), and the remaining two found in *CHEK1* (rs540436, rs569766). Genes co-expressed with *CHEK1* were identified using the method described but these were not associated with SCZ association (p = 0.916; r = -0.000942).

Co-expression with multiple schizophrenia risk genes

In this section the co-expression model was extended to investigate if identifying co-expression with multiple risk genes was a more significant predictor of association signal. A combined model based on the genes whose expression pattern successfully predicted SCZ association should reduce any noise present assuming there is a consistent expression pattern amongst these risk genes.

P values were obtained for all genes as a measure of their co-expression with *CNNM2*, *CSMD1*, *MMP16*, *NT5C2* and *TCF4*. A significant positive relationship ($p = 1.06 \times 10^{-7}$) between SCZ Brown's logP and model logP was found that was more significant and had a larger correlation coefficient than identifying co-expressed genes using any gene individually (compare Table 3.6 with Table 3.5). Similarly the nonparametric test results were all highly significant with the best p-value when testing the top 25% ($p = 9.30 \times 10^{-13}$) see Appendix Figure 8.15 panel A.

	Schizophrenia		**Bipolar disorder**	
	Brown's	**Simes'**	**Brown's**	**Simes'**
P value	1.06×10^{-7}	3.92×10^{-5}	0.00522	0.00104
Correlation Coeff.	0.0472	0.0378	0.0243	0.0302
	+	+	+	+
Excluding MHC genes				
P value	7.41×10^{-9}	2.82×10^{-7}	0.00252	0.000512
Correlation Coeff.	0.0516	0.0476	0.0264	0.0322
	+	+	+	+

Table 3.6: Linear regression results and correlation coefficients testing SCZ risk genes co-expression model logP across development calculated in the BrainSpan RNA-Seq dataset with gene-wide logP. Based on co-expression model with *CNNM2*, *CSDM1*, *MMP16*, *NT5C2* and *TCF4*.

	Schizophrenia		**Bipolar disorder**	
	Brown's	**Simes'**	**Brown's**	**Simes'**
P value	1.29×10^{-6}	0.000126	0.00922	0.00273
Correlation Coeff.	0.0430	0.0353	0.0226	0.0276
	+	+	+	+
Excluding MHC genes				
P value	1.23×10^{-7}	1.89×10^{-6}	0.00458	0.0014
Correlation Coeff.	0.0472	0.0441	0.0248	0.0296
	+	+	+	+

Table 3.7: Linear regression results and correlation coefficients testing SCZ risk genes co-expression model logP across brain regions calculated in the BrainSpan RNA-Seq dataset with gene-wide logP. Based on co-expression model with *CNNM2*, *CSDM1*, *MMP16*, *NT5C2* and *TCF4*.

This association test was also rerun removing genes in the MHC, Table 3.6 and Appendix Figure 8.15. The associations were marginally stronger (regression $p = 7.41 \times 10^{-9}$; Mann-Whitney $p = 9.33 \times 10^{-14}$ top 25%), so genes within this region were not driving this result. An association was also found between BPD Brown's logP and the SCZ model logP. This was both with ($p = 0.00522$) and without the MHC genes (p =

0.00252). The SCZ model p values were also correlated with the Simes' logP for SCZ and BPD, a result that withstood removing the MHC genes.

Variation across brain regions

The models used so far have looked at temporal co-expression but due to the nature of this dataset spatial co-expression could also be investigated, using a similar regression framework to try to identify which is most relevant for capturing SCZ association. Regression analysis found a slightly weaker enrichment ($p = 1.29 \times 10^{-6}$) with a smaller correlation coefficient than when correlating across development, results presented in Table 3.7, that was verified with rank-based tests shown in Appendix Figure 8.15. This model was also enriched for BPD p values ($p = 0.00922$), and was also significant with Simes' logP (SCZ $p = 0.000126$; BPD $p = 0.00273$). Further, all models remained significant after removing genes from the MHC.

When fitting a regression model to predict SCZ Brown's logP with the model logP measuring co-expression across development and co-expression across brain regions simultaneously, both terms were a lot less significant but the model across development remained just significant ($p = 0.0201$), whereas the model across brain regions did not ($p = 0.445$). Both models appear to be capturing very similar sets of genes that were co-regulated across development stages and brain regions. Therefore, only variation across development will be considered further in this section.

Comparison of results with Section 3.2.1

In order to see if these results converged with those in Section 3.2.1 where development stage characteristic scores were tested for association with gene-wide p values, the characteristic scores of genes identified from the SCZ co-expression model were examined. The set of genes most enriched for SCZ association was identified as the top 25% from Appendix Figure 8.15 panel A and the median development stage characteristic scores for this set of genes were plotted in Figure

3.6. An identical pattern to that seen in Figure 3.1 showed that the results from this section converged with the results in Section 3.2.1. Both sections found enrichment for SCZ and BPD common variants in genes with increased expression during early and mid-foetal development that dropped off prior to birth and had decreased values through late infancy and childhood before increasing through adolescence and young adulthood.

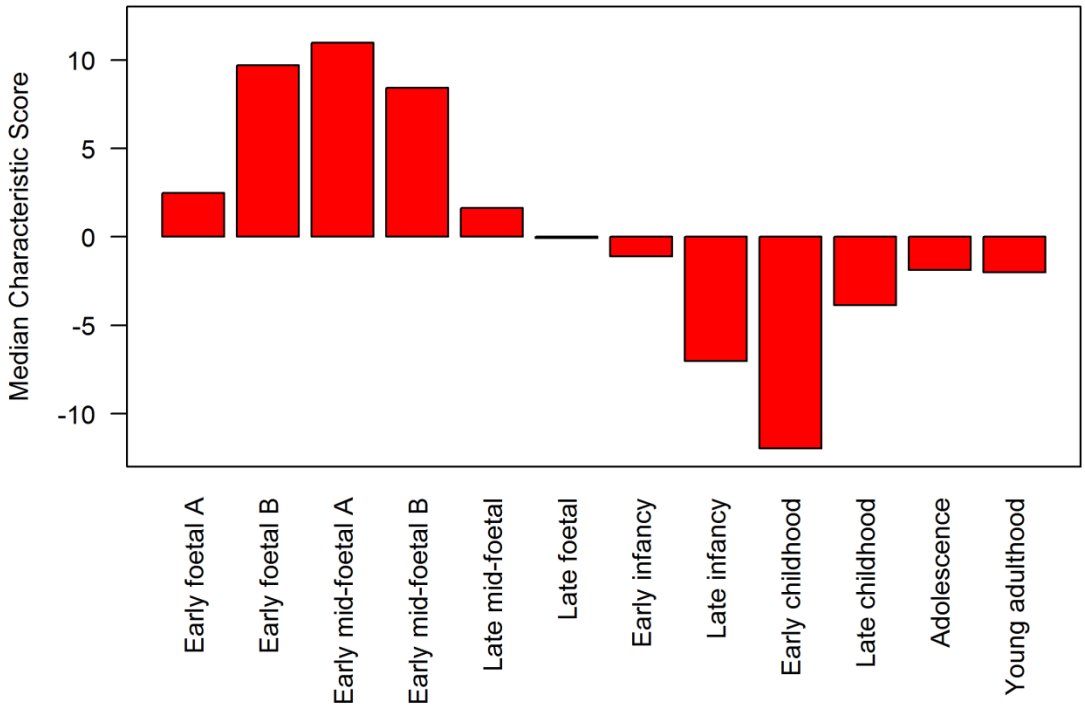

Figure 3.6: Median development stage characteristic scores calculated in the BrainSpan RNA-Seq dataset for most enriched gene set identified from SCZ risk genes co-expression model.
Based on the top 25% of genes ranked by SCZ risk genes co-expression model p values identified as most enriched gene set from the Mann-Whitney tests. For each development stage, the median characteristic score of this set of genes was calculated.

MIR137 targets

One particularly interesting top hit in the SCZ PGC GWAS was *MIR137,* as prior evidence suggests this gene plays a role in brain development, in particular regulating adult neurogenesis (Szulwach *et al.*, 2010) and neuron maturation (Smrt *et al.*, 2010). Therefore, disruptions to this gene may form part of the neurodevelopmental hypothesis of SCZ's aetiology. The PGC authors found that SNPs

in or near predicted targets of *MIR137* were enriched for association (Ripke *et al.*, 2011), a finding that was validated with results from a subsequent meta-analysis of SCZ GWAS (Ripke *et al.*, 2013). *MIR137* could not be included in the SCZ co-expression model as it was not present in the expression dataset.

A list of 301 genes predicted as targets for *MIR137,* obtained from TargetScan (Lewis *et al.*, 2005), were tested with a Mann-Whitney test to see if they were more closely co-expressed across development with the top hits in the PGC GWAS compared to all other genes. This was found to be the case ($p = 5.54 \times 10^{-22}$) demonstrating that these target genes more closely resemble the developmental profile of the SCZ genes than the genes not predicted as targets.

Technical validation in microarray dataset

The same approach was applied using the microarray data, although only six genes (*CCDC68*, *CNNM2*, *CSMD1*, *NT5C2*, *STT3A*, and *TCF4*) were considered, as *MMP16* was not present in the final dataset. Co-expression model logP were not associated with SCZ Brown's logP (all $p > 0.05$) for any of the six genes considered, shown in Table 3.8.

	CCDC68	**CNNM2**	**CSMD1**	**NT5C2**	**STT3A**	**TCF4**
P-value	0.0866	0.598	0.190	0.288	0.434	0.977
Correlation Coeff	-0.0149	0.00460	-0.0114	0.00927	0.00682	0.000257
	-	+	-	+	+	+

Table 3.8: Linear regression results and correlation coefficients testing single SCZ risk gene co-expression model logP calculated in Kang microarray dataset with SCZ Brown's logP.

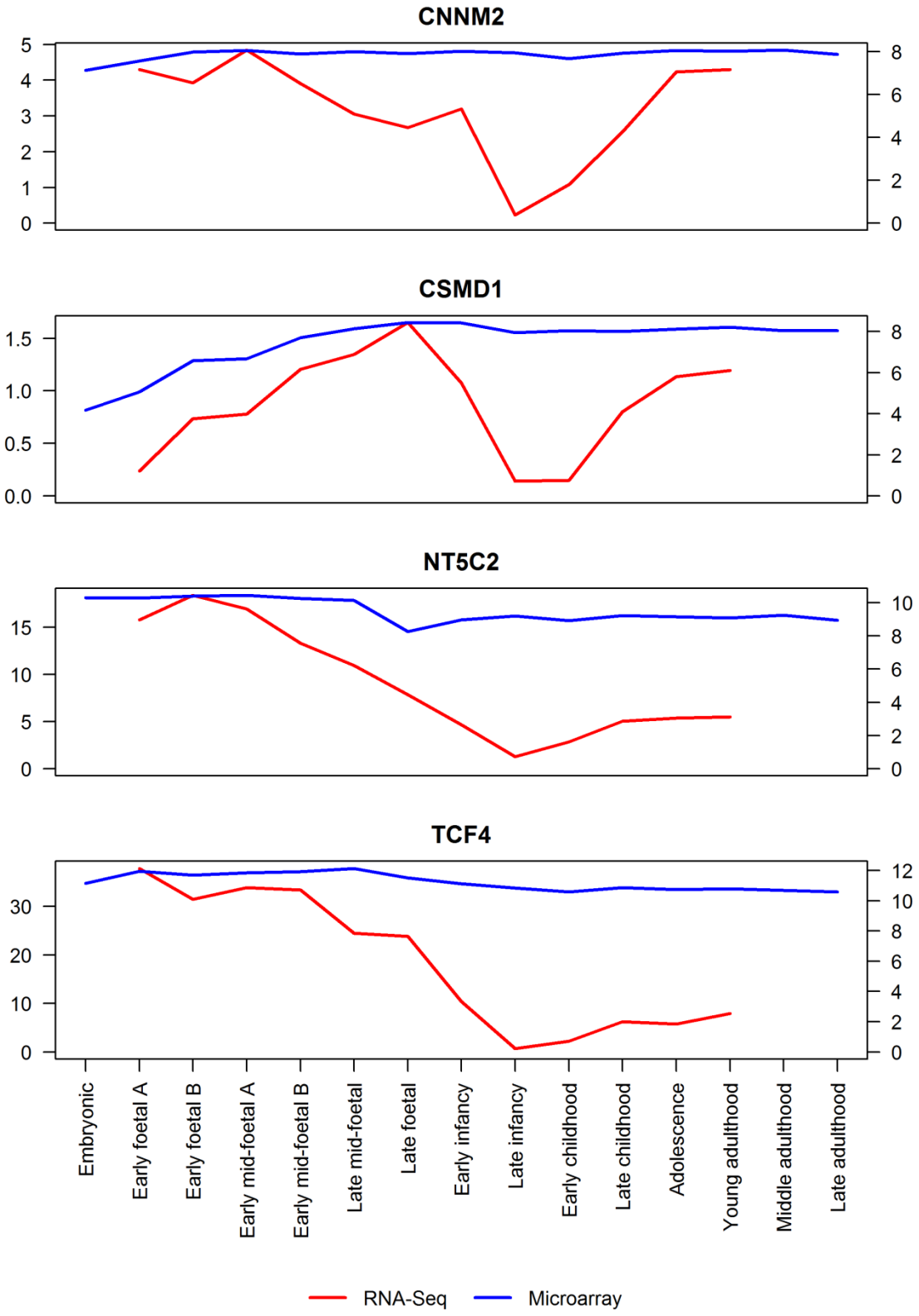

Figure 3.7: SCZ risk genes comparing microarray and RNA-Seq expression values.
Risk genes selected from PGC GWAS whose co-expression indexed association in the RNA-Seq dataset and were also present in the microarray dataset. Median expression values calculated for each development stage in each dataset, scale on left for RNA-Seq data, scale on right for microarray data.

In order to compare these results to those found in the RNA-Seq dataset, the microarray and RNA-Seq expression values for common genes were plotted on the same graph but with scales appropriate to each type, see Figure 3.7. The red lines that represent the RNA-Seq values show much more variability in expression values across development compared to the microarray values in blue. RNA-Seq expression values are known to have a much larger dynamic range than microarray expression values (Wang *et al.*, 2009), which is partly why they are more accurate. For some genes however, such as *NT5C2* or *TCF4*, the microarray expression values appear to be slightly higher in foetal stages compared to postnatal stages. The reduced sensitivity of microarrays meant that the linear model was unable to detect these subtle expression differences that could be picked up when using data from next-generation sequencing technologies. Therefore these results could not be technically validated in the Kang dataset.

Summary

Genes identified as having an expression profile that follows the same trajectory across development as *CNNM2*, *CSMD1*, *MMP16*, *NT5C2* and *TCF4*, through the SCZ co-expression model, were enriched for SCZ gene-wide p values. These five genes were identified as genome-wide significant in the largest GWAS to date and are presumed most likely to be true associations. Initially genes co-expressed with each gene separately were shown to be enriched for SCZ variants, but the association with SCZ gene-wide p values was more significant when considering these five genes together. These results were validated with both parametric and nonparametric approaches. Further, genes predicted as targets for *MIR137*, another genome-wide significant hit for SCZ, were also enriched for genes co-expressed with these SCZ risk genes. The results in this section for co-expressed genes were not validated in the microarray data, although this was likely due to technical differences in the expression values each method obtains.

Genes identified as co-expressed with these genes were shown to have a profile of development stage characteristic scores consistent with the results in Section 3.2.1.

Therefore the two different approaches in Sections 3.2.1 and 3.2.2 have identified the same expression profile for SCZ risk genes, characterised by highest expression values during foetal development, followed by a drop in expression around birth to the lowest values around late infancy and early childhood before increasing again.

3.2.3 Bipolar disorder risk genes co-expression models

An equivalent approach was performed considering BPD risk genes identified from the PGC study. Four genome-wide significant SNPs were associated in either the primary analysis or the combined primary and replication meta-analysis (Sklar *et al.*, 2011). These were located closest to *ANK3*, *SYNE1*, *ODZ4*, and *CACNA1C*, all of which were present and expressed in the RNA-Seq data. To begin with separate models were fitted for each gene and all remaining genes in the expression dataset were given a p value for how highly co-expressed they were with each risk gene.

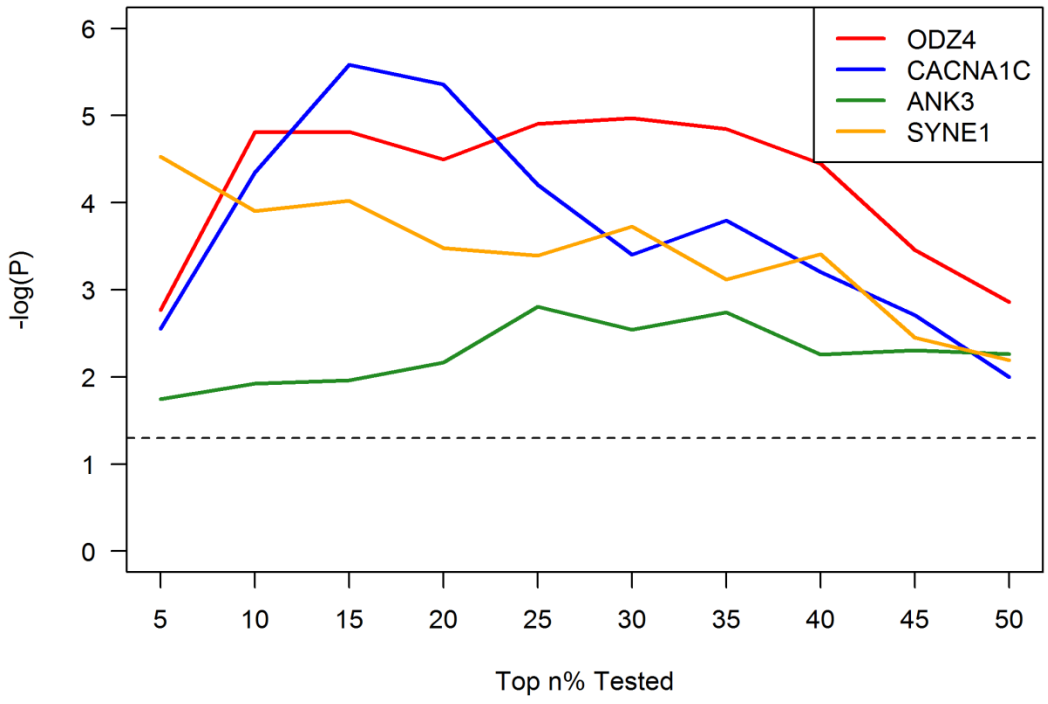

Figure 3.8: Results from Mann-Whitney tests for genes ranked by single BPD risk gene co-expression model p values calculated in the BrainSpan RNA-Seq dataset.
Genes ranked by BPD single risk gene co-expression model p values and top n% tested for smaller BPD Brown's p values against bottom 50%. Black dashed line is p = 0.05.

Significant positive correlations were found between the co-expression model logP and BPD Brown's logP for all four genes (*ANK3* p = 0.0199, r = 0.0202; *CACNA1C* p =

3.63 x 10^{-6}, r = 0.0402; *ODZ4* p = 3.30 x 10^{-5}, r = 0.0361; *SYNE1* p = 2.05 x 10^{-5}, r = 0.0370). Nonparametric tests verified the associations for each gene model logP, for all top proportions of genes tested, see Figure 3.8. Therefore co-expression with any of the four genes identified from the PGC study was associated with more significant BPD associations, adding further confidence that these risk genes are true.

	Schizophrenia		Bipolar disorder	
	Brown's	**Simes'**	**Brown's**	**Simes'**
P value	1.97 x 10^{-5}	1.23 x 10^{-5}	3.68 x 10^{-5}	2.35 x 10^{-6}
Correlation Coeff.	0.0379	0.0402	0.0358	0.0434
	+	+	+	+
Excluding MHC genes				
P value	5.27 x 10^{-7}	7.49 x 10^{-8}	1.89 x 10^{-5}	1.15 x 10^{-6}
Correlation Coeff.	0.0448	0.0498	0.0374	0.0451
	+	+	+	+

Table 3.9: Linear regression results and correlation coefficients testing BPD risk genes co-expression model logP across development calculated in the BrainSpan RNA-Seq dataset with gene-wide logP. Based on co-expression model with *ANK3, CACNA1C, ODZ4,* and *SYNE1*.

	Schizophrenia		Bipolar disorder	
	Brown's	**Simes'**	**Brown's**	**Simes'**
P value	0.00289	0.000154	7.05 x 10^{-5}	1.96 x 10^{-5}
Correlation Coeff.	0.0265	0.0348	0.0304	0.0393
	+	+	+	+
Excluding MHC genes				
P value	0.000111	8.31 x 10^{-6}	4.08 x 10^{-5}	1.16 x 10^{-5}
Correlation Coeff.	0.0345	0.0413	0.0358	0.0406
	+	+	+	+

Table 3.10: Linear regression results and correlation coefficients testing BPD risk genes co-expression model logP across brain regions calculated in the BrainSpan RNA-Seq dataset with gene-wide logP. Based on co-expression model with *ANK3, CACNA1C, ODZ4,* and *SYNE1*.

Two further models were fitted including all four of these genes, one measuring temporal co-expression and the other spatial co-expression, results in Tables 3.9 and 3.10. A marginally stronger association was found with BPD Brown's logP and temporal co-expression logP (p = 3.68 x 10^{-5}) rather than spatial co-expression logP (p = 7.05 x 10^{-5}). The results of the Mann-Whitney tests verified these enrichments, see Appendix Figure 8.16, and showed that they were not due to extreme values.

Genes identified with each model were also associated with SCZ Brown's logP with both parametric and nonparametric tests. A more significant result was found when considering variation across development ($p = 1.97 \times 10^{-5}$) compared to development across brain regions ($p = 0.00289$), see Tables 3.9 and 3.10. Enrichments for both SCZ and BPD were also found with Simes' logP, and all enrichments remained after removing genes located in the MHC. All significant regression models were validated with nonparametric tests and shown not to be due to extreme values, see Appendix Figures 8.16.

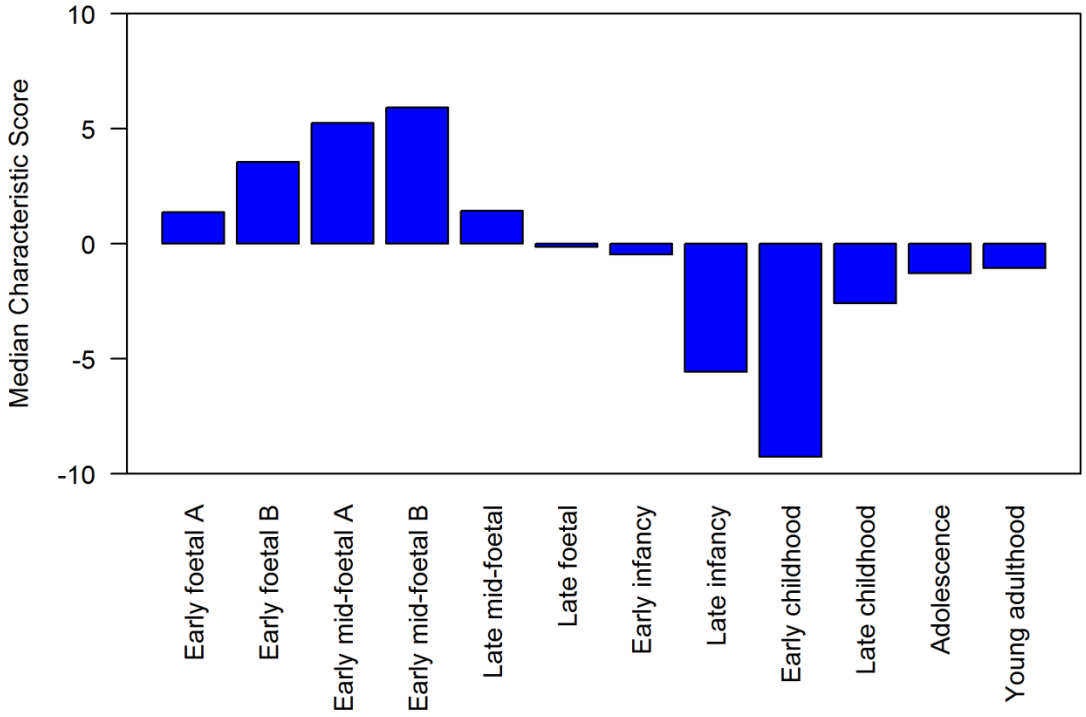

Figure 3.9: Median development stage characteristic scores calculated in the BrainSpan RNA-Seq dataset for most enriched gene set identified from BPD risk genes co-expression model.
Based on the top 40% of genes ranked by BPD risk genes co-expression model p values identified as most enriched gene set from the Mann-Whitney tests. For each development stage, the median characteristic score of this set of genes was calculated.

As with the SCZ co-expression model the most enriched set of genes identified through the BPD co-expression model were taken and their median characteristic scores were plotted, see Figure 3.9. The graph looked very similar to that of Figure 3.1 and Figure 3.6 showing that these genes had very similar expression characteristics to those found to be enriched in Sections 3.2.1 and 3.2.2. Therefore

the results of the BPD co-expression model also converged with those reported in Sections 3.2.1 and 3.2.2. This would imply that both the SCZ and BPD co-expression models were selecting fairly similar sets of genes. The median characteristic scores in Figure 3.9 were not as extreme as in those in Figure 3.6 in Section 3.2.2 perhaps suggesting that BPD risk genes do not have as an extreme temporal expression profile compared to SCZ risk genes.

When the logP from the SCZ co-expression model and BPD co-expression model were tested simultaneously to predict BPD Brown's logP, both terms were considerably less significant than when tested separately with the BPD model logP remaining significant (SCZ model logP p = 0.490; BPD model logP p = 0.00139). A similar result was found when predicting SCZ Brown's logP with the SCZ co-expression model logP remaining significant (SCZ model logP p = 0.000412; BPD model logP p = 0.491).

This follows what was observed in Figures 3.6 and 3.9 that both models were capturing an overlapping set of genes. Interestingly, despite this high degree of overlap, the BPD model was most enriched for BPD signal whereas the SCZ model was most enriched for SCZ signal. This perhaps suggests that the models were picking up slightly different temporal expression characteristics for each disorder although this was not evident from Figures 3.6 and 3.9. The only notable difference was that the median expression characteristic scores for the BPD co-expressed genes were not as extreme as those for the SCZ co-expressed genes.

Validation with microarray data

This approach was repeated in the microarray dataset, although only two genes *ANK3*, *SYNE1* were present. Neither regression result with *SYNE1* model logP (p = 0.597, r = 0.00452) or ANK3 model logP (p = 0.803, r = -0.00214) was significant and hence no further models were fitted. As with the SCZ model, the lack of significance in this dataset compared to the RNA-Seq dataset was attributed to the smaller dynamic range of microarray expression values observed when comparing the

expression of these two genes across the two technologies; see Appendix Figure 8.17.

Summary

Genes whose expression profiles correlated with strongly associated BPD risk genes *ANK3*, *CACNA1C*, *ODZ4* and *SYNE1*, identified in the largest GWAS, were found to be enriched for genes with common variants associated to SCZ and BPD. The enrichment was slightly stronger when identifying variation across development stages, compared to variation across brain regions. Applying these methods to microarray data did not find the same enrichments, and as in Section 3.2.2 this was thought to be due to the smaller dynamic range of these expression values. The profiles of these genes share the same characteristics with the genes in the SCZ co-expression model, shown statistically when the association of these two models was not independent, and were consistent with the results described in Section 3.2.1.

3.2.4 Development stage characteristic gene expression and schizophrenia structural variants

CNVs found in SCZ patients were compared to those found in healthy controls to see if the genes hit were more characteristic of a particular developmental stage. A logistic regression model was used, as described in Chapter 2, comparing the minimum, median and maximum characteristic score of the genes hit by each CNV for each development stage. Genes hit by SCZ CNVs were not more characteristic of any stage compared to those hit by control CNVs, further no significant results were found when testing the deletions or duplications. These results can be found in Appendix Table 8.7.

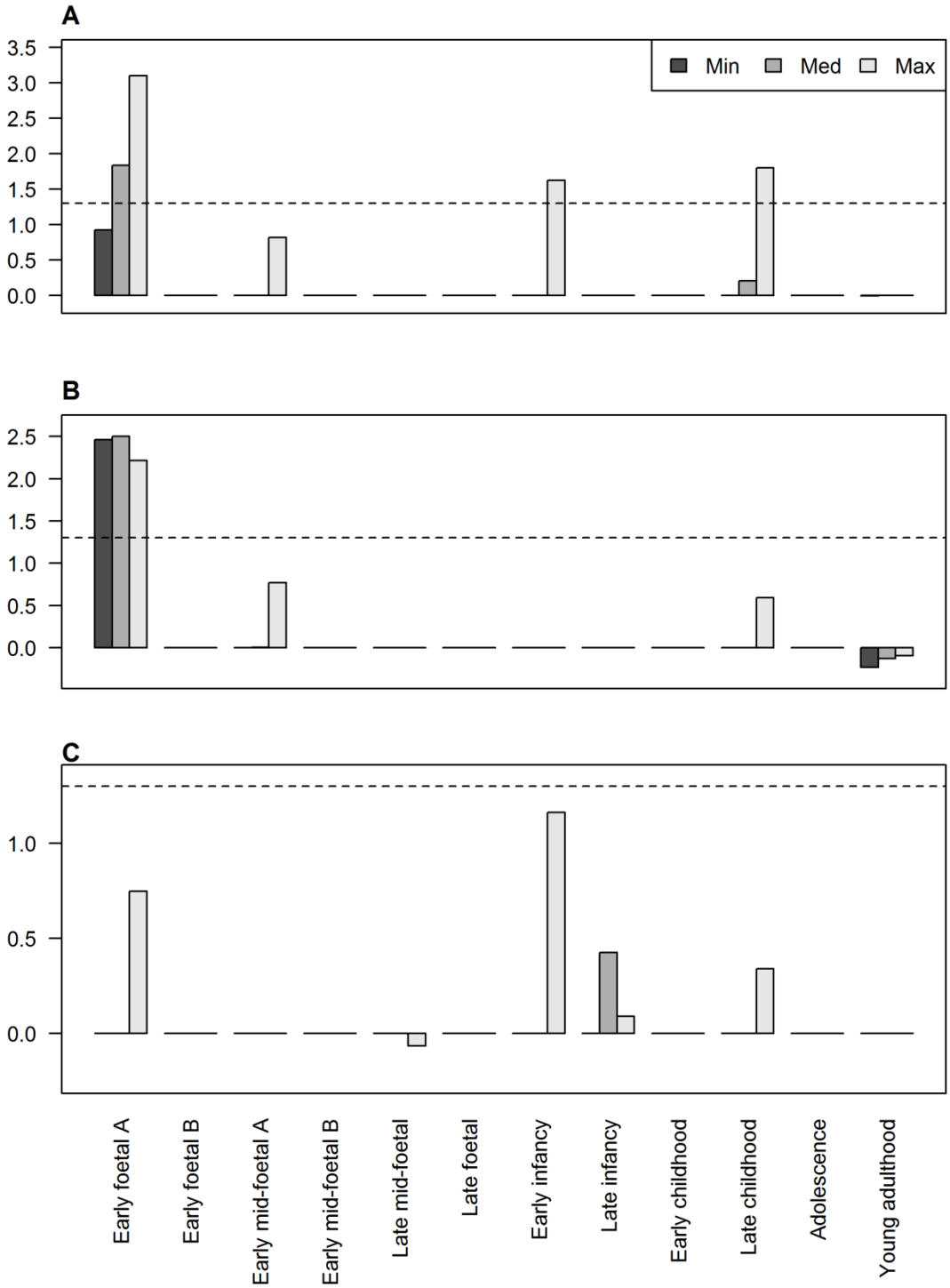

Figure 3.10: Logistic regression results testing CNV singleton status on development stage characteristic scores calculated in the BrainSpan RNA-Seq dataset.
Panel A is all CNVs, panel B deletions, and panel C duplications. P values were corrected for 12 development stages using Bonferroni's method. Black dashed line is 0.05.

Singleton SCZ CNVs were compared to all remaining SCZ CNVs to see if they were enriched for genes characteristic of any development stage. Only a couple of

marginally significant results were found with no obvious pattern to the results. The most significant result suggested singleton SCZ CNVs hit genes with higher early foetal A characteristic scores (median metric corrected p = 0.0145; maximum metric corrected p = 0.000788). This result was also found in the deletions (minimum metric corrected p = 0.00344; median metric corrected p = 0.00315; maximum metric corrected p = 0.00608).

Technical validation with microarray dataset

Repeating these analyses in the Kang microarray dataset showed no significant differences between the genes hit by case CNVs compared to control CNVs for any set of development stage characteristic scores, results in Appendix Table 8.9. When testing genes hit by singleton SCZ CNVs, again there was no pattern to the results however a significant result was found for early foetal A (median metric corrected p = 0.0327; maximum metric corrected p = 0.0109), see Figure 3.11. This was in the same direction as in the RNA-Seq data and was also significant when testing just the deletions. This result remained when excluding the PMI covariate results in Appendix Table 8.12 or Appendix Figure 8.18.

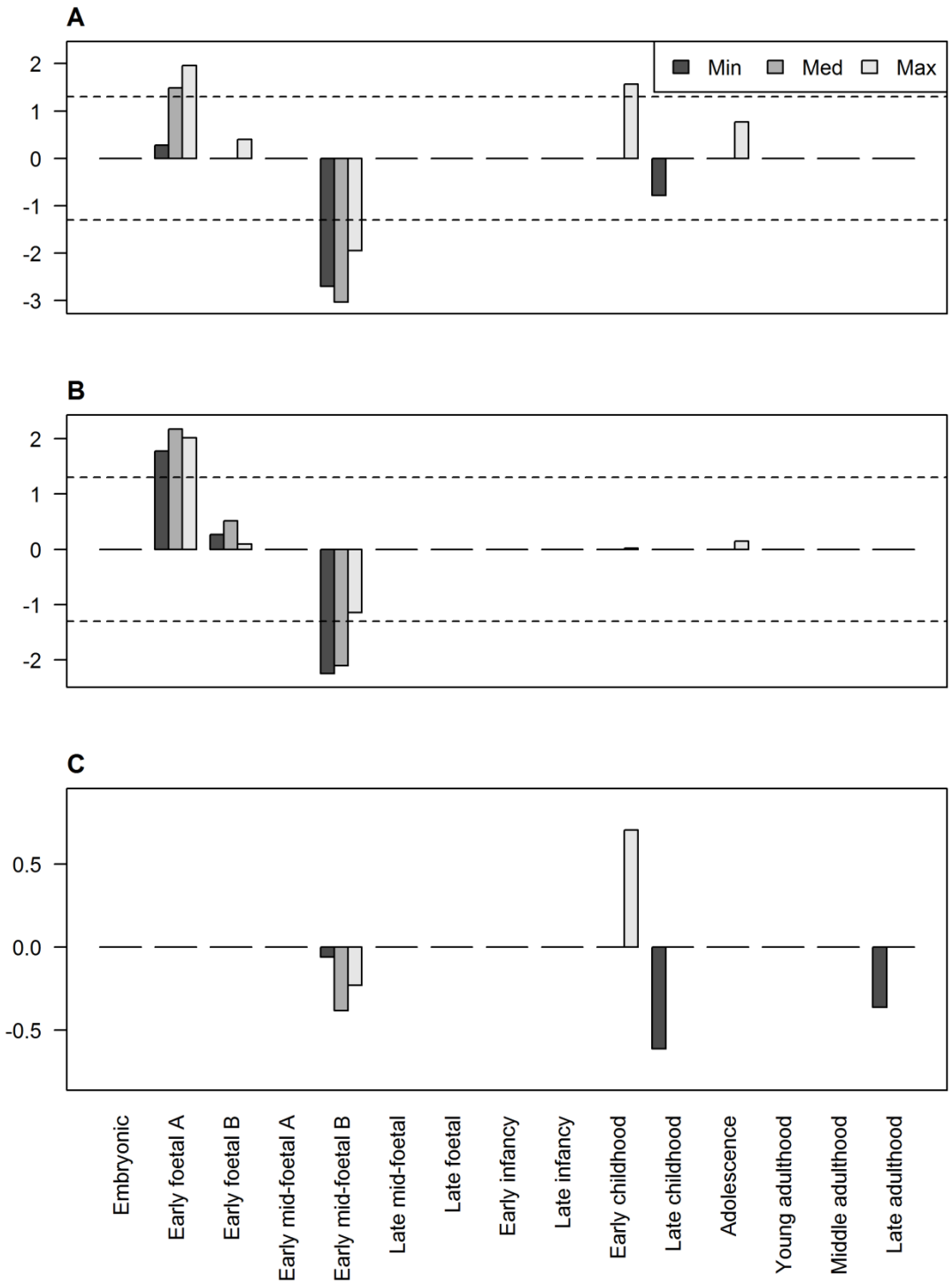

Figure 3.11: Logistic regression results testing CNV singleton status on development stage characteristic scores calculated in Kang microarray dataset.
Panel A is all CNVs, panel B deletions, and panel C duplications. P values were corrected for 15 development stages using Bonferroni's method. Black dashed line is 0.05.

Summary

Singleton SCZ CNVs were found to hit genes with higher early foetal A characteristic scores, compared to all remaining SCZ CNVs with characteristic scores calculated in either the RNA-Seq dataset or the microarray dataset. This was particularly the case for deletions. This suggests genes disrupted by singleton SCZ CNVs have increased expression in the brain during this early development stage. While a similar finding was reported for genes with common variants associated to SCZ, this was in later foetal stages, with no association found for early foetal A characteristic scores.

3.2.5 Functional analysis of genes with enriched expression profiles

As the SCZ co-expression model in the RNA-Seq dataset captured fundamentally the same enriched expression profile found through the separate development stage characteristic scores in Section 3.2.1 through one metric, p values from this model were used for functional analysis. Genes annotated to each functional category were compared to all remaining genes in the expression dataset to see if they were more closely co-expressed with SCZ risk genes. Of 3085 unique terms from the GO database with between 20 and 200 genes, 219 had significantly lower ranked model p values at a Bonferroni corrected threshold of 1.62×10^{-5}. This set of terms was clustered in groups following the procedure described in the previous chapter.

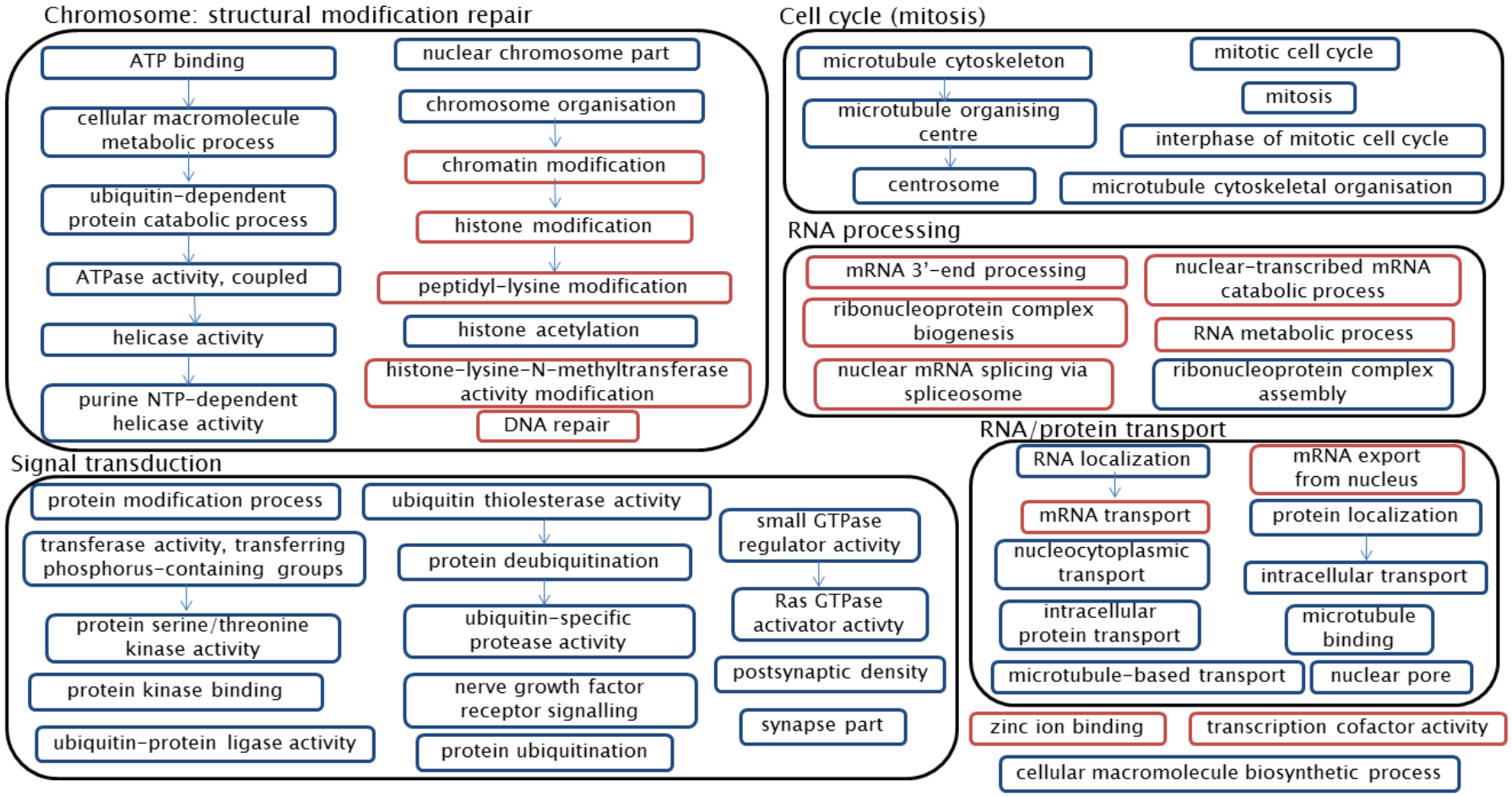

Figure 3.12: Key annotation terms identified from set of significant GO terms with smaller SCZ risk genes co-expression model p values.
Figure shows set of 54 GO terms that explained at least one other term in the set of significant pathways. Terms that did not explain any other term were not included in the Figure. Arrows point from explaining term to the merged pathway it explains i.e. the term pointed to, merged with all other terms it explains. Terms in red ovals are also present in Figures 2.6 and 2.7 in the previous chapter; terms in black boxes were grouped into common themes.

Fifty-four terms explained the enrichment of at least one other larger GO term or merged GO pathway. Figure 3.12 displays these terms where they have been manually grouped into five common themes: 'Chromosome: structural modification & repair', 'Cell cycle (mitosis)', 'RNA processing', 'RNA/protein transport', and 'Signal transduction'. Similar to the set-based tests performed in the previous chapter, the combined association signal of genes with the temporal expression profile was compared between those in these functional categories against genes not in any pathway to confirm that these pathways did contain SCZ association signal. The set of genes with the most significantly smaller SCZ Brown's gene-wide p values identified from the Mann-Whitney tests for the top 5-50% of genes ranked by their SCZ co-expression model p values, shown in Appendix Figure 8.15 panel A to be the top 25%, were taken as the set of genes with temporally variable expression. Separate set-based tests were then performed to identify which terms may be more important for SCZ aetiology.

		Schizophrenia		Bipolar disorder	
		Number of SNPs in pathway	**Adjusted p value**	**Number of SNPs in pathway**	**Adjusted p value**
Top 25% of SZ Model genes	**In pathways**	93728	0.00195	177267	0.00325
	Not in pathways	28507	0.625	54006	0.271
	Cell cycle (mitosis)	9131	0.00129	17400	0.148
	Chromosome: structural modification & repair	23283	0.0262	42535	0.00390
	RNA/protein transport	19394	0.206	38920	0.324
	RNA processing	10753	0.0706	19677	0.567
	Signal transduction	38289	0.0543	73974	0.0206

Table 3.11: Results of set-based tests for genes found in most enriched gene set from SCZ co-expression model and in significantly enriched pathways.
Set p values adjusted for number of SNPs in each set.

Table 3.11 shows that genes with the temporal expression profile and in significant pathways were associated to both SCZ (adjusted p = 0.00195) and BPD (adjusted p = 0.00325). In contrast genes with the temporal expression profile but not in an

enriched pathway, as a group, were not associated to either SCZ or BPD. Of the pathway groups 'Cell cycle (mitosis)' and 'Chromosome: structural modification & repair' had a significant set-based p values when combining SCZ GWAS results. Interestingly 'Chromosome: structural modification & repair' and 'Signal transduction' were enriched for BPD GWAS signal. Two groups, 'RNA/protein transport' and 'RNA processing' were not significant for either SCZ or BPD.

3.3 Discussion

3.3.1 Identification of common developmental expression profile

An expression profile that shows variation across development has been identified for SCZ risk genes. Gene sets identified with increased characteristic expression during early to late mid-foetal development, and gene sets with decreased characteristic expression in postnatal stages were enriched for common SCZ risk variants. These associations were found not to be independent suggesting there is an underlying common set of genes with both aspects of these expression profiles.

Separate analyses showed that genes with expression that correlated closely with five strongly associated SCZ genes (*CNNM2*, *NT5C2*, *TCF4*, *MMP16* and *CSMD1*) identified from the largest published GWAS (Ripke *et al.*, 2011) were enriched for risk variants. Genes identified with this approach showed the same characteristic expression profile as the results in the first section of this chapter. Together, these two approaches suggest that risk genes for SCZ are characterised by a peak of expression during foetal development followed by a gradual decrease starting prior to birth and continuing through to the lowest values around early childhood, before beginning to increase again through adolescence.

Additional analyses found that genes predicted as *MIR137* targets were more similar to this developmental profile than remaining genes in the dataset. As these genes have previously been shown to be enriched for SCZ common variants (Ripke *et al.*, 2011, Ripke *et al.*, 2013) this further supports the described temporal profile for SCZ associated genes.

While previous studies have suggested SCZ genes have age-related expression profiles (Colantuoni *et al.*, 2008, Choi *et al.*, 2009, Harris *et al.*, 2009), these were based on samples covering a smaller period of brain development with entirely postnatal individuals. In addition, these were conducted prior to the GWAS era and the list of SCZ risk genes were primarily based on literature reviews. Therefore this work strengthens these findings, and extends them by describing the temporally variable profile identified.

Gene sets identified with common spatial profiles in mid-foetal brains and enriched for SCZ and BPD common variants in Chapter 2 showed a developmental profile consistent with that identified here, see Figure 2.10. Two exome sequencing studies have looked at the temporal expression profile of genes with *de novo* mutations and observed higher expression in prenatal samples (Xu *et al.*, 2012, Gulsuner *et al.*, 2013), in line with the findings reported here. Another study using GWAS, CNV and SNV data to create networks of genes based on the likelihood of them contributing to SCZ aetiology, also reported higher prenatal expression for genes within the resulting clusters (Gilman *et al.*, 2012). All of these studies used either the BrainSpan or Kang datasets and while it is encouraging to identify the same expression characteristics, replication in an independent expression dataset for this expression profile would still be warranted. These studies

The temporal profile described is consistent with neurodevelopment models of SCZ which posit that a disruption early on in development interacts with normal maturational processes during adolescence (Weinberger, 1987). While upregulation of expression during foetal development has previously been reported, downregulation during early childhood is a novel finding for SCZ associated genes. If disruptions to a risk gene or genes during this time point were causative of SCZ, this would suggest that the window for developmental insults that increase the risk of SCZ extends beyond the period around birth. It would also imply that developmental delays would only occur after this time point, a hypothesis that could be investigated in a prospective birth cohort. An alternative hypothesis would be that altered gene

expression during this developmental period may be a result of an earlier disruption to brain development, and mediates the relationship between developmental insults and SCZ.

Genes with this developmental expression profile were also enriched for BPD common variants. Fewer and less significant results were found when testing genes characteristic of each development stage with BPD gene-wide p values. As in the previous chapter this was probably due to the smaller sample size of the BPD PGC GWAS, however, the directions of effect were consistent with the SCZ results. Genes identified with the SCZ co-expression model also showed enrichment for BPD common variants. Further, a BPD co-expression model based on genome-wide significant BPD genes (*ANK3*, *CACNA1C*, *ODZ4*, *SYNE1*) captured genes with a similar temporal profile to the SCZ risk genes and was associated with more significant SCZ and BPD gene-wide p values.

A few significant results were found when looking at genes hit by singleton SCZ CNVs. While singleton deletions were found to hit genes with increased expression in early foetal A in both the RNA-Seq and microarray dataset, neither of these results was particularly strong. This result was broadly in line with the general finding that SCZ risk genes have increased expression during early foetal development, although this was at a later developmental stage, perhaps suggesting that CNVs affect earlier developmental processes.

3.3.2 Identification of functional pathways from enriched expression gene sets

GO terms enriched for genes with this temporal profile were identified and grouped into broad themes including 'Chromosome: structural modification and repair', 'RNA processing' and 'RNA/protein transport', which suggests that these genes play a role in transcription and the control of related processes. Of these three groups only the 'Chromosome: structural modification and repair' group, when intersected with genes with the temporal profile, showed association to SCZ and BPD. A second pathway group 'Cell cycle (mitosis)' when intersected with genes with the temporal

profile, was also enriched for SCZ common variants. Taken together these two pathway groups may suggest a role for abnormal cellular proliferation in SCZ during the division stage of the cell cycle. Increased cellular proliferation has been reported in olfactory neurosphere-derived cells from SCZ patients compared to controls, which may impact on the early stages of neurodevelopment (Fan *et al.*, 2012).

Thirteen individual GO terms that explained other terms overlapped with those identified in the previous chapter for genes with decreased expression in the HIP and THAL, highlighted in red in Figure 3.12. Due to the inter-dependence of the GO hierarchy it was not possible to test if this was more than would be expected by chance, however it provides some validation that not only did the different approaches in the two chapters converge on the same temporal profile, but that they also identified the same functional pathways. In particular the common pathways relate to epigenetic processes, which along with the temporal profile suggest that these genes may play a role in the regulation of human brain development. Therefore, they may be particularly vulnerable to early insults affecting the course of normal development.

Two groups of pathways, 'Cell cycle (mitosis)' and 'Signal transduction' were only found in this chapter for genes whose expression correlated with the temporal profile and were not picked up with the foetal spatial profiles. Synaptic genes have previously been implicated in SCZ pathology (Kirov *et al.*, 2012, Perez-Santiago *et al.*, 2012, Gulsuner *et al.*, 2013), and alterations to synaptic machinery during synapse formation and pruning, which continues into adulthood, have been proposed as part of the neurodevelopmental model (Mirnics *et al.*, 2000, Mirnics *et al.*, 2001). Although the 'Signal transduction' group contained a few terms relating to the synapse, the genes in these pathways with the temporal expression profile only showed a trend for SCZ GWAS signal.

3.3.3 Comparison of results with Brown's and Simes' gene-wide p values

Two methods to summarise GWAS results into gene-wide p values were used in this chapter to test for association of expression profiles with either SCZ or BPD. Generally the Brown's gene-wide p values were more significant particularly, with the RNA-Seq dataset, in Sections 3.2.1 and 3.2.2. This is consistent with the results in the previous chapter suggesting that using a p value that can take into account multiple signals is more powerful. Interestingly, the Simes' p values were more significant when testing the BPD co-expression models in Section 3.2.3.

3.3.4 Technical replication across RNA-Seq and microarray expression data

Both RNA-Seq and microarray expression datasets, with an overlap of individuals, were used in this chapter. Significant results with the development stage characteristic scores calculated in the microarray dataset were consistent with the results profile found with the RNA-Seq data, although there were not as many significant stages. When previously analysed, two additional covariates were included for the microarray dataset and that same procedure was followed here, whereas no such covariates were provided with the BrainSpan data. The removal of the PMI covariate when calculating the characteristic scores produced a results profile more in line with that found with the RNA-Seq data suggesting that some of the temporal effects were being captured by this variable. As the majority of the samples overlapped the two datasets, it may therefore be that the differences in PMI caused the observed expression profile. This was unlikely to be the sole factor in producing the temporal profile as even after inclusion of this covariate, the results still showed enrichment for SCZ risk genes in those with increased expression during early and mid-foetal stages or decreased expression around early childhood.

Technical replication was not found with the microarray data for any of the co-expression models. The next-generation sequencing technologies used to generate RNA-Seq expression data are more sensitive than microarray approaches and produce a larger range of expression values (Wang *et al.*, 2009), observed for the genes used in the co-expression models in Figures 3.7 and 8.17. Therefore the linear

model approach used here was not able to capture the subtle differences across development with the microarray data that could be picked up with the RNA-Seq data.

One of the main strengths of this study was the use of the most comprehensive expression datasets covering human brain development with multiple samples for each individual. The scarce availability of post-mortem brain samples makes this sort of resource rare and therefore there was no independent expression dataset even close to covering the same developmental window, with the same range of brain regions.

An alternative approach would be to use another GWAS dataset for genetic replication. Results presented here with BPD gene-wide p values support the findings for SCZ and provide a level of replication. The PGC GWAS results were used as the basis for the gene-wide p values as it was the largest available study. Since publication the PGC have recruited an additional 35 SCZ sample collections on top of the initial 17 studies to further increase the sample size. Access to this dataset was provided and preliminary results support the findings reported in this chapter.

3.3.5 Summary of chapter findings

This and the previous chapter have identified common spatial and temporal expression profiles of SCZ risk genes, and through these common biological pathways of potential relevance to SCZ. Moreover, the gene sets with common spatial profiles in the previous chapter had a developmental expression profile that matched the findings in this chapter for SCZ risk genes. Generally the results were more significant when testing Brown's gene-wide p values compared to Simes', particularly for the SCZ co-expression model in Section 3.2.2. This suggests that many of the genes identified with the developmental profile were better represented by a p value that takes into account multiple, semi-independent common variants that increase the risk of SCZ or BPD rather than a p value based on only the most

significant SNP. This is consistent with a polygenic contribution of these genes, which will be investigated in the next chapter.

3.4 Methods

Preliminary data processing: BrainSpan RNA-Seq

The BrainSpan RNA-Seq expression dataset was downloaded from an online resource (http://www.brainspan.org) already normalised to RPKM (reads per kilobase of exon model per million mapped reads) units. Lowly expressed genes were removed as any gene with a maximum value less than 1. Alongside the expression values, gene annotations were provided including Entrez IDs. These were added to the dataset and genes without Entrez IDs or non-unique Entrez IDs were removed.

Preliminary data processing: Kang microarray

This dataset was downloaded from the GEO database accession number GSE25219. Genes were annotated and filtered following the procedure described in the previous chapter applied to all samples.

Development stage characteristic scores

A linear model was fitted to identify genes characteristic of each development stage, shown in Equation 3.1.

$$\exp_{ij} = \text{brain region}_j + \text{development stage}_{ik}$$

Equation 3.1 Regression model to calculate characteristic scores for development stage k, where $\exp_{ij}$ is the expression value for individual i from brain region j, brain region is a categorical term and the development stage term is a binary indicator variable denoting whether individual i was classed in stage k or not.

Separate models were fitted for each development stage, for each gene. Only samples taken from brain regions present in nine out of the twelve development stages were used when fitting the model, removing ten regions and leaving sixteen.

Most of those removed were proxies for regions not yet developed in the foetal brains and hence only present in the foetal individuals. Development stage characteristic scores were derived based on the same formula used in Chapter 2, denoting the magnitude of the differential expression as well as whether it was increased or decreased expression. A similar model was fitted for the Kang microarray dataset, including additional covariates for hemisphere, RIN and PMI.

Testing for enrichment with gene-wide p values

Brown's and Simes' logP, were tested in a linear regression framework predicted in turn by the characteristic scores for each development stage. Significant results were verified with a rank-based Mann-Whitney test. Genes were ranked by the absolute value of the characteristic score for each development stage and the top n% (5, 10…50%) selected. Each set was then separated by the sign of the characteristic score into those with increased characteristic expression (positive coefficient) and decreased characteristic expression (negative coefficient). The relevant subset, depending of the direction of the association in the linear regression was then tested with a one-sided Mann-Whitney test against the bottom 50% for more significant gene-wide p values. Analyses were repeated removing genes located in the MHC region, specifically those found in the chr6:25000000-35000000 region.

Schizophrenia co-expression models

$$\exp_{ij} = \text{SCZ gene}_{ij} + \text{brain region}_j + \text{individual}_i$$

$$\exp_{ij} = \text{brain region}_j + \text{individual}_i$$

Equation 3.2 Formula for regression models compared to obtain a model p value for co-expression with SCZ risk genes, where $\exp_{ij}$ is the expression value from individual i and brain region j. SCZ gene_{ij} is the expression value for that risk gene in individual i and brain region j with categorical covariates to control for brain region and individual differences.

Genes identified from the SCZ PGC GWAS (Ripke *et al.*, 2011) were used as the basis to identify sets of co-expressed genes. For each risk gene, linear models of the form in Equations 3.2 were compared for all remaining genes in the dataset with an

ANOVA to obtain a p value for co-expression, referred to as the co-expression model p value.

In the Kang dataset extra covariates for hemisphere, PMI and RIN were also included. Genes whose co-expression indexed association signal were combined into a single model shown in Equation 3.3, which was also compared with an ANOVA to a model without the five risk gene terms.

$$\mathrm{exp_{ij}} = \mathrm{CNNM2_{ij}} + \mathrm{CSMD1_{ij}} + \mathrm{MMP16_{ij}} + \mathrm{NT5C2_{ij}} + \mathrm{TCF4_{ij}} + \mathrm{brain\ region_j} + \mathrm{individual_i}$$

Equation 3.3 Formula for regression models with multiple SCZ risk genes to identify genes co-expressed across development, where $\mathrm{exp_{ij}}$ is the expression value from individual i and brain region j, $\mathrm{CSMD1_{ij}}$ etc. is the expression value for that risk gene in individual i and brain region j with categorical covariates to control for brain region and individual differences.

Linear models shown in Equation 3.4 were used to investigate co-expression across brain regions, where a brain region term in Equation 3.3 was replaced with a development stage term. A p value was calculated from an ANOVA for all remaining genes in the expression dataset as a measure of co-expression.

$$\mathrm{exp_{ij}} = \mathrm{CNNM2_{ij}} + \mathrm{CSMD1_{ij}} + \mathrm{MMP16_{ij}} + \mathrm{NT5C2_{ij}} + \mathrm{TCF4_{ij}} + \mathrm{development\ stage_i} + \mathrm{individual_i}$$

Equation 3.4 Formula for regression models with multiple SCZ risk genes to identify genes co-expressed across brain regions, where $\mathrm{exp_{ij}}$ is the expression value from individual i and brain region j, $\mathrm{CSMD1_{ij}}$ etc. is the expression value for that risk gene in individual i and brain region j with categorical covariates to control for development stage and individual differences.

MIR137 targets

A list of *MIR137* predicted targets was obtained using TargetScan (Lewis *et al.*, 2005) with a probability of conserved target set to greater than or equal to 0.9. Of the 301 genes, 275 had a SCZ model p values derived from ANOVAs comparing Equations 3.3, with *TCF4* and *CSMD1* automatically excluded as they were used to fit the model. This set was tested with a one-sided Mann-Whitney test for more significant SCZ co-

expression across development model p values compared to all other genes in the expression dataset.

Bipolar disorder co-expression models

Four genes identified in the BPD PGC GWAS (Sklar *et al.*, 2011) were considered for the BPD co-expression models. All four of these genes were present in the BrainSpan dataset and the models in Equations 3.5 were compared to derive a p value for co-expression across development for all remaining genes.

$$\exp_{ij} = \text{BPD gene}_{ij} + \text{brain region}_j + \text{individual}_i$$
$$\exp_{ij} = \text{brain region}_j + \text{individual}_i$$

Equation 3.5 Formula for regression models compared to obtain a model p value for co-expression with BPD risk genes, where $\exp_{ij}$ is the expression value from individual i and brain region j. BPD gene_{ij} is the expression value for that risk gene in individual i and brain region j with categorical covariates to control for brain region and individual differences.

Co-expression with all genes was found to predict BPD association; hence all four were combined into two models. Equation 3.6 was used to derive p values for co-expression across development and Equation 3.7 was used to look at co-expression across brain regions.

$$\exp_{ij} = \text{ANK3}_{ij} + \text{CACNA1C}_{ij} + \text{ODZ4}_{ij} + \text{SYNE1}_{ij} + \text{brain region}_j + \text{individual}_i$$

Equation 3.6 Formula for regression models with multiple BPD risk genes to identify genes co-expressed across development, where $\exp_{ij}$ is the expression value from individual i and brain region j, ANK3_{ij} etc. is the expression value for that risk gene in individual i and brain region j with categorical covariates to control for brain region and individual differences.

$$\exp_{ij} = \text{ANK3}_{ij} + \text{CACNA1C}_{ij} + \text{ODZ4}_{ij} + \text{SYNE1}_{ij} + \text{development stage}_i + \text{individual}_i$$

Equation 3.7 Formula for regression models with multiple BPD risk genes to identify genes co-expressed across brain regions, where $\exp_{ij}$ is the expression value from individual i and brain region j, ANK3_{ij} etc. is the expression value for that risk gene in individual i and brain region j with categorical covariates to control for development stage and individual differences.

CNV logistic regression

For each CNV from the ISC and MGS study, characteristic scores for all genes hit from both the BrainSpan and Kang datasets were collated and the minimum, median and maximum for each development stage identified. Logistic regression models as described in Chapter 2 Equations 2.2 and 2.3 were used to test for an association between the development stage characteristic scores and genes hit by SCZ CNVs.

Functional analysis

Model p values from the RNA-Seq SCZ co-expression model with *CNNM2, CSMD1*, *MMP16*, *NT5C2* and *TCF4*, across development shown in Equation 3.3, were used to identify relevant functional pathways. Categories from the GO database were tested with one-sided Mann-Whitney tests for significantly smaller model p values. Adjusting for 3085 unique terms with between 20 and 2000 genes a Bonferroni corrected p value threshold of 1.62×10^{-5} was used to identify significant terms. This set of significant terms was merged into clusters following the procedure described in the previous chapter. The resulting clusters were manually grouped into common themes.

Set-based tests

Sets were based on the most significant set of genes with the temporal profile, identified as the top 25% from the Mann-Whitney tests shown in Appendix Figure 8.15. This set was split into two subsets; genes that were annotated to one of the 219 significant pathways and those that were not. Further sets for the five pathway groups in Figure 3.12 were created based on the intersect of genes in the top 25% identified through the SCZ co-expression model and genes annotated to any pathway within that group.

For each set a combined set-based p value was calculated from the PGC SCZ and BPD GWAS results, which was subsequently adjusted for the number of SNPs in each set based on the procedure described in the Methods section of the previous chapter.

Chapter 4: Developing the polygenic model for application to expression derived gene sets

4.1 Introduction

4.1.1 Background

Technological advances in the form of genome-wide SNP chips provided new opportunities to those looking for genetic risk factors in psychiatric disorders, as candidate gene studies had proved to be limited in their success. By vastly increasing the number of genetic markers included in a single study, the threshold for statistical significance had to become more stringent to account for the number of tests performed. The effective number of independent tests per genome has been estimated (International HapMap Consortium, 2005, Dudbridge and Gusnanto, 2008) in order to establish an industry standard level of genome-wide significance, to minimise false positive findings that would not be replicated. Initial SCZ GWAS failed to identify markers at this threshold, with sample size the limiting factor.

In 2009 three studies were published simultaneously, one of which, conducted by the ISC, obtained genome-wide significance for an imputed SNP in the MHC region (Purcell *et al.*, 2009). Each of these three studies then incorporated the data from the other two studies to see if any of their top hits could attain genome-wide significance. SNPs that passed this threshold in the meta-analyses were located on chromosomes 6p22.1 part of the MHC region (Shi *et al.*, 2009, Stefansson *et al.*, 2009), 11q24.2 (Stefansson *et al.*, 2009) and 18q21.2 (Stefansson *et al.*, 2009). The MGS study concluded that they had satisfactory power to detect any common variants of large effect (relative risk > 1.3) and the fact that they had not meant there were likely few, if any, to detect (Shi *et al.*, 2009). Instead it was postulated that many common markers were involved, each with a small or moderate contribution to the risk of developing SCZ.

Using their GWAS results the ISC derived a simple mathematical model that demonstrated this polygenic effect of thousands of markers, each of small effect. The principle of this model was to calculate a score, termed the polygenic score, for each individual in an independent dataset as the sum of their risk alleles at each SNP, weighted by the log odds ratio obtained from the initial GWAS. These scores were shown to be significantly different between cases and controls explaining ~3% of the variance (Purcell *et al.*, 2009).

Simulations were used to narrow down the possible genetic models that could have produced these results. These considered the proportion of associated markers that were actually causal compared to those that just tagged the relevant marker, as well as the distributions of allele frequencies and effect sizes of the associated SNPs. The simulations consistent with the true results showed that a minimum of one third of the genetic heritability could be explained by the polygenic contribution of common SNPs (Purcell *et al.*, 2009). This was much greater than the observed 3% which was impacted by the accumulation of sampling errors of the estimated effect sizes (Dudbridge, 2013). More recently this methodology was repeated with the PGC GWAS results derived from a sample with more than double the number of individuals and was found to explain 6% of the variance (Ripke *et al.*, 2011). An increased sample size meant that more true variants would be included in the score and that the odds ratios would be more accurately estimated, therefore the proportion of variance explained was improved.

In the previous chapters, sets of genes with common expression profiles across the mid-foetal brain or brain development have been shown to be enriched for SCZ association signal using gene level summarised p values. In this chapter, genes with a common expression profile were investigated further to see if they harboured a polygenic signal that could better discriminate between cases and controls than the genome-wide polygenic score. As described above, the current procedure for calculating polygenic scores is quite simplistic, so the initial sections in this chapter focus on developing this model. Three adaptions to the model were investigated to incorporate population information, LD relationships between markers or SNP-SNP

interactions to see if the addition of this information improved the discrimination between SCZ cases and controls.

4.1.2 Outline

Aim

The main aim of this chapter was to compare polygenic scores calculated from gene sets identified in the previous chapters to those calculated across all available SNPs to see whether they capture any additional SCZ polygenic signal. In order to make the best possible estimate of the signal captured, a prior aim was to explore whether the inclusion of population information, LD between SNPs or SNP-SNP interactions could improve the fit of the polygenic model described by the ISC.

Datasets

Genotype data from both the ISC and MGS studies were available for use and in all applications the polygenic model was trained in the biggest GWAS, the ISC, to obtain the most accurate estimates of effect size, and the MGS dataset was used as the target or test dataset. To avoid overestimating the predictive ability of the polygenic scores it is important that the training and target dataset are entirely independent (Powell and Zietsch, 2011). The ISC dataset contained 3322 cases and 3587 controls combined from 8 separate studies (Purcell *et al.*, 2009). Data were obtained post quality control to remove problematic SNPs and individuals, with 739995 SNPs left.
The MGS European-American dataset contained 2681 cases and 2653 controls and had 671422 SNPs after quality control (Shi *et al.*, 2009).

Both datasets were filtered to the set of 661356 overlapping SNPs to ease subsequent model developments and make them comparable. This was different to the procedure described in the ISC paper, where SNP filtering and pruning were performed on all SNPs in the ISC data which meant that some of the SNPs used to calculate the polygenic scores would not be found in the MGS data. All subsequent filtering steps were performed using the same thresholds detailed by the ISC; all

SNPs with a MAF < 0.02 or a missing genotype rate > 0.01 in the ISC study were removed. This left 275265 SNPs, with the majority excluded because some studies in the ISC were genotyped on a different chip with only ~380000 SNPs.

SNPs were then pruned in a pairwise manner based on their r^2 statistic using a sliding window of 200 SNPs and 1000kb. The ISC chose a threshold of $r^2 < 0.25$ to create an independent set of SNPs and hence that threshold was used here for comparison. A set of 42113 SNPs with no pairwise $r^2 > 0.25$ and all with a SCZ association p value < 0.5 were used to calculate polygenic scores. These will be referred to as the independently associated SNP set or SNP_{IA} which was used to compare the adaptations of the polygenic model described in this chapter.

Outline of analysis

The first adaptation considered population stratification. In the original application, the two Swedish studies were combined into a single population and all remaining studies taken as six other populations. This population stratification was controlled for in the association test to produce a single odds ratio for each SNP. Here, the polygenic framework was reformulated to allow multiple odds ratios for these different populations and incorporate this extra population information.

The second development described looked at including the LD structure between SNPs to allow the inclusion of more SNPs when calculating the polygenic scores. In the framework proposed by the ISC, stringent pruning was applied to ensure SNPs were independent and prevent overestimating the effects of correlated SNPs. Here an alternative method was implemented that allows for LD when estimating the odds ratios, meaning that more SNPs can be retained and ultimately more information included.

Finally the method was extended to calculate a polygenic score based on interactions between independently associated SNPs. Polygenic scores were calculated for each individual in the MGS dataset for each adaptation to be

compared to those calculated as described in the original formulation. Each set of scores was tested to see how strongly they predicted case control status in a logistic regression test. P values and Nagelkerke's R^2 values (Nagelkerke, 1991) were used to compare the performance of each model to the original ISC framework.

After identifying which adaptations of the polygenic model improved the prediction of case control status and the amount of variance explained, these were used to calculate polygenic scores for all SNPs within a set of genes with a common temporal expression profile identified as enriched for SCZ common variants in Chapter 3. The gene set polygenic scores were tested simultaneously with genome-wide scores in a logistic regression model to see if it was a significant predictor after allowing for the genome-wide score. This would inform whether this gene-set contained any SCZ signal not captured by the genome-wide score.

4.2 Results

4.2.1 Standard polygenic scores

All SNPs in the SNP_{IA} set were used to calculate polygenic scores as described by the ISC, referred to here as the standard polygenic. Odds ratios were estimated from a logistic regression model that predicted case control status by the number of minor alleles at each SNP and included covariates to control for the seven populations. Output from these association analyses was used to calculate the polygenic scores in PLINK (Purcell *et al.*, 2007), which divides the final polygenic score by the number of SNPs to give a mean score per SNP. These scores were found to be significantly different between cases and controls, $p = 5.41 \times 10^{-31}$, $R^2 = 0.0345$. This result was mildly more significant than that reported by the ISC, which was probably due to a larger set of SNPs, obtained from a slightly different filtering procedure, being used. This was the baseline result to which all subsequent modifications were compared.

4.2.2 Population weighted polygenic scores

After implementing a strict quality control procedure to genotype data to remove many sources of possible false positives, population stratification becomes the main concern for spurious results in GWAS (Tian *et al.*, 2008a). In a case control study design if the population background of the cases and controls are not well matched, differences in allele frequencies between these populations may incorrectly associate a SNP to the disorder (Knowler *et al.*, 1988, Campbell *et al.*, 2005). This can be particularly problematic when markers are thought to only confer a small or modest effect on disease risk and studies are therefore only looking for subtle differences in allele frequencies such as the case for SCZ. As a result, GWAS methods have been developed either to take into account ancestral information or to reduce the resulting inflation of test statistics (Devlin and Roeder, 1999, Pritchard *et al.*, 2000, Price *et al.*, 2006). These considerations are now part of the routine of association analyses, as the ISC study was a collaboration of eight other studies a Cochran-Mantel-Haenszel test, which allows for known categorical populations, was used.

In the polygenic model there is the assumption that the alleles identified in the training dataset are also associated with disease risk in the target dataset, with the same direction of effect. While this methodology has been shown to work across different ethnicities, demonstrated by the ISC between European Americans and African Americans, the results were stronger if individuals in both datasets come from the same population (Purcell *et al.*, 2009). Both the ISC and MGS have been filtered to only contain European individuals, but even within European subpopulations allele frequencies are known to differ and can cause false positive results (Seldin *et al.*, 2006, Tian *et al.*, 2008b). Within the ISC each study represents a subpopulation whose individuals were shown to cluster together in a plot of the first two multidimensional scaling components calculated from their genotypes, see Figure 4.1. This plot shows that the samples originating from the British Isles, either Scotland, England or Ireland, group together but also overlap, whereas the Bulgarian, Portuguese and Swedish samples form distinct groups away from each

other and the British Isles cluster. Taking account of this subpopulation structure, and therefore any associated differences, should produce a better estimate of disease susceptibility for each SNP, for each individual. In the first adaption, this subpopulation information was incorporated to calculate personalised odds ratios for each individual in the MGS based on how well their genotypes matched each of the populations in the ISC study.

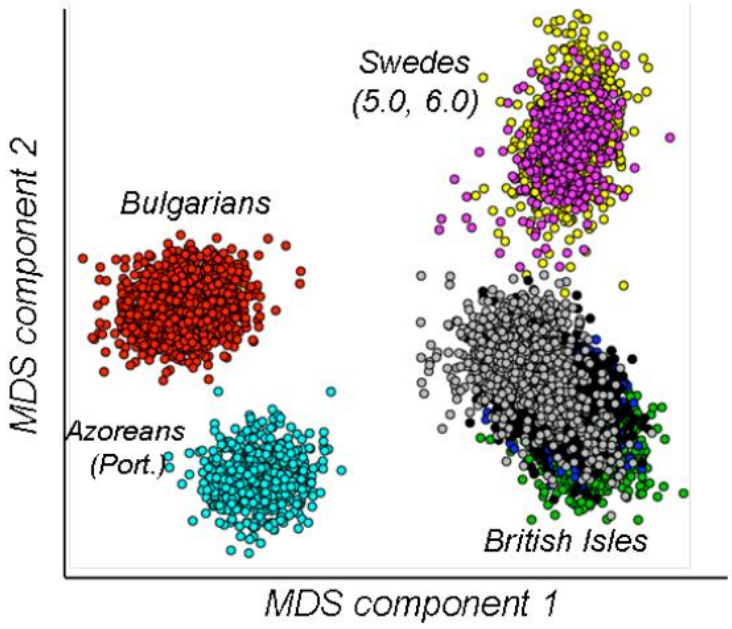

Figure 4.1: Multidimensional scaling plot of individuals in ISC dataset.
This is a two dimensional representation of the ISC genotype data where samples that are most similar will be placed closest together, and samples that show the most differences furthest apart. Each point in this plot represents an individual that is coloured by the sample collection the individual originated from. Taken from Supplementary Figure S1 in (Purcell *et al.*, 2009).

Log odds ratios, θ_{jk}, for each population j and SNP k were estimated from logistic regression models only including individuals from that population. Seven likelihood values w_{ij} were then calculated for each individual i in the MGS data, using the probability of their genotypes occurring in population j of the ISC data, see Equation 4.1. The genotype probabilities were taken from the ISC data and calculated for the pruned SNP set ($r^2 < 0.25$). If a genotype was not observed in a population, the probability for that SNP was set to a value less than if there had be a single occurrence, i.e. set to a value smaller than what could have been observed (Cardiff:

0.0008, Dublin: 0.0008, Edinburgh: 0.001, Portugal: 0.001, Sweden: 0.001, UCL: 0.0009, Aberdeen: 0.0007).

$$w_{ij} = \sum_{k} -\log(\alpha_{ijk})$$

Equation 4.1 Formula for calculation of population weights w_{ij} for individual i in population j, where α_{ijk} is the probability of individual i's genotype at SNP k in population j; α_{ijk} were taken from the genotype frequencies in the ISC dataset which was used as the training dataset.

These w_{ij} were then standardised so that the sum of the weights across all populations for each individual totalled 1. To calculate the weighted log odds ratio θ_{wik} for individual i at SNP k, the weight for each population was multiplied by the relevant population log odds ratio and summed across the populations (Equation 4.2).

$$\theta_{w_{ik}} = \sum_{j} w_{ij}\theta_{jk}$$

Equation 4.2 Formula for the calculation of individual weighted log odds ratios θ_{wik} for individual i at SNP k, where w_{ij} is as in Equation 4.1 and θ_{jk} is the log odds ratio for population j at SNP k.

Polygenic scores were calculated based on SNPs in the SNP_{IA} subset (42118). For each individual a population weighted polygenic score or P_w was calculated using these weighted log odds ratios multiplied by the number of associated alleles at that SNP shown in Equation 4.3. If the genotype for an individual was missing at a particular SNP the expected value was calculated and multiplied by the weighted odds ratio. Consistent with the implementation in PLINK for polygenic scores the P_w were divided by the number of SNPs to create a mean score per SNP.

$$P_{w_i} = \sum_{k} \theta_{w_{ik}}\, \eta_{ik}$$

Equation 4.3 Formula for the calculation of population weighted polygenic scores P_{wi} for individual i, where θ_{wik} is as in Equation 4.2 and η_{ik} is the number of associated alleles for individual i at SNP k.

The P_w were found to be significantly different between cases and controls; p = 7.48 x 10^{-30}; Nagelkerke's R^2 = 0.0331, which was marginally less significant than the standard polygenic baseline result. This population weighted approach assumes populations have different effect sizes at all SNPs. Therefore this method may be more beneficial when only applied to SNPs for which this was the case. Further, for SNPs that do not show any differences or only weak differences between populations, this method will introduce inaccuracies when estimating the effect sizes due to smaller sample sizes for the individual populations.

To identify a relevant subset of SNPs to apply this approach to, two association regression models, shown in Equations 4.4, with and without interactions between the number of minor alleles and population covariates were compared. P values were taken from a chi squared 6 degree of freedom test used to compare the fit of these two models. Multiple p value thresholds (0.1, 0.05, 0.01, 0.005 and 0.001) were used to select SNPs to apply the population weighted approach to, with all remaining SNPs taking their odds ratios from Equation 4.4.A.

$$\text{status} = \text{nMA} + \omega_1 + \omega_2 + \omega_3 + \omega_4 + \omega_5 + \omega_6 \qquad \textbf{A}$$

$$\text{status} = \text{nMA} + \omega_1 + \omega_2 + \omega_3 + \omega_4 + \omega_5 + \omega_6 + \text{nMA} * \omega_1 + \text{nMA} * \omega_2 + \text{nMA} * \omega_3 + \text{nMA} * \omega_4 + \text{nMA} * \omega_5 + \text{nMA} * \omega_6 \qquad \textbf{B}$$

Equation 4.4 Formula for two regression models compared to identify SNPs with differences in effect sizes across populations and that population weighted log odds ratios should be calculated for, where status is SCZ case control status, nMA is the number of minor alleles at test SNP and ω_j are binary covariates for population j.

Table 4.1 shows that applying this population weighted approach to an informed subset of SNPs does marginally improve the significance compared to applying it to all SNPs and explains slightly more of the variance. The best result was obtained when using a threshold of 0.05 to select SNPs to apply the weighted approach to; however across the thresholds the results were broadly similar. In addition this result was also slightly more significant than the baseline polygenic result.

Threshold to select SNPs for weighted approach	All SNPs	P < 0.1	P < 0.05	P < 0.01	P < 0.005	P < 0.001
Number of SNPs with weighted approach	42113	4646	2374	480	258	51
P value	6.43×10^{-30}	3.28×10^{-31}	2.98×10^{-31}	4.46×10^{-31}	3.52×10^{-31}	4.99×10^{-31}
Coeff.	+	+	+	+	+	+
R^2	0.0332	0.0348	0.0348	0.0346	0.0347	0.0346

Table 4.1: Logistic regression results testing population weighted polygenic scores.
Population weighted approach applied to subsets of SNPs with significant difference in effect sizes between populations for different significance thresholds. Compare to baseline result of $p = 5.41 \times 10^{-31}$, $R^2 = 0.0345$.

The lack of a noticeable improvement of the population weighted model introduced here compared to the standard polygenic model could have been due to the inability to detect true differences between populations, either because there were none or there was not adequate power to detect them. Alternatively, it may be that taking these differences into account did not improve the model, perhaps because the differences are small. The number of SNPs detected with significant differences across the populations in the ISC data was compared to 100 random permutations where the population structure had been removed. Population was permuted within cases and within controls separately to retain the same number of cases and controls per population but remove population differences across the sample. Equations 4.4 were fitted for each SNP in the permuted dataset and the number of SNPs with significant differences was counted.

Table 4.2 shows that at each threshold, the number of SNPs with population differences in the true ISC data was greater than any permutation, except $p < 0.01$ where one permutation had at least as many significant SNPs. This result is consistent with polymorphisms affecting the different populations within the ISC study to different extents. Alternatively, these results may reflect differing LD relationships between the causal SNP and the tag SNP across the populations. Where this correlation is weaker, the evidence for association and effect size between the tag SNP and SCZ disease status would smaller. In these instances, the model used to detect population differences would be unable to distinguish if it was true effect size

difference or merely a result of differing LD relationships. In either scenario, the additional noise from estimating the effect size for each population from a smaller sample size, meant allowing for these differences in the polygenic framework did not greatly improve discrimination between SCZ cases and controls.

Threshold for significant differences across populations	**0.1**	**0.05**	**0.01**	**0.005**	**0.001**
Number of SNPs in ISC with differences across populations	8805	4491	923	499	106
Empirical p value based on 100 permutations	< 0.01	< 0.01	0.01	< 0.01	< 0.01

Table 4.2: Empirical p values for number of SNPs with significant differences in effect size across populations in the ISC.
Tested by comparing models in Equation 4.4 and comparing results in true ISC data to 100 permutations where population structure was randomised.

4.2.3 LD adjusted polygenic scores

In the current framework when summing across SNPs, any pair of SNPs in even moderate LD will cause an overestimation of the number of independent effects. To prevent this, the original application was performed on a set of stringently pruned SNPs, where no pair of SNPs had an $r^2 > 0.25$. This reduced the set of SNPs that passed quality control by approximately 70%, losing information from SNPs that was not captured by those they were partially correlated with. This could be particularly relevant when calculating polygenic scores for gene sets later in this chapter, which were based on a smaller set of SNPs. Results reported in Chapter 3 were more significant when testing for enrichment with summarised p values based on Brown's method, reflecting semi-independent effects from across the gene. Stringent pruning of SNPs within this set may therefore lose some of this information.

When using logistic regression as the test for association analysis, case control status is regressed on the number of minor alleles for the test SNP, denoted here as SNP_{test}. Within this framework covariates can easily be included to control for population structure with scope to include any other possible confounders. In order to account for the correlation between SNPs, an extra covariate for a SNP in high LD with

SNP_{test}, was included in the regression framework as follows. SNPs were considered in the order they were located along each chromosome. The first SNP_{test} on each chromosome was treated as usual with its odds ratio estimated from a logistic regression model. For each subsequent SNP_{test}, the SNP within the previous 200 SNPs and 1000 kb that was in strongest LD with SNP_{test} (identified by r^2 calculated in controls only) was taken, denoted from here as SNP_{LD1}.

$$\text{status} \sim \text{SNP}_{\text{test}} + \text{SNP}_{\text{LD1}} + \omega$$

Equation 4.5 Logistic regression model used to adjust SNP_{test}'s odds ratio for $SNP_{LD,}$ where status is SCZ case control status, SNP_{test} and SNP_{LD1} are the number of minor alleles at these SNPs and ω is the population covariates.

The number of minor alleles of SNP_{LD1} was then included as a covariate in the association test for SNP_{test} so that the odds ratio would be adjusted for LD between these two SNPs, see Equation 4.5. Therefore, if both of these SNPs were included when calculating the polygenic scores, the effects should no longer be overestimated. These polygenic scores will be referred to as LD adjusted polygenic scores or P_{LD}.

As regression is not effective at estimating parameters for highly correlated variables, some LD based pruning was required. Pruning was applied at six r^2 thresholds (0.8, 0.82, 0.84, 0.86, 0.88 and 0.9) to see if this produced any variation in the results. Table 4.3 displays the number of SNPs in each set.

Pruning threshold	**0.8**	**0.82**	**0.84**	**0.86**	**0.88**	**0.9**
Number of SNPs after LD pruning	161793	164722	167939	171251	174923	178899
Number of LD pruned SNPs with p < 0.5	84397	85954	87651	89405	91359	93475

Table 4.3: Number of SNPs after LD based pruning in sets used to calculate LD adjusted polygenic scores.

For each of these six SNP sets the first polygenic model fitted was based on odds ratios from logistic regression models not adjusted for any LD SNPs, results in Table 4.4. At all thresholds the polygenic scores were significantly higher in SCZ cases compared to controls, with the most significant difference at $r^2 < 0.84$ ($p = 3.35 \times 10^{-42}$). All of these results were more significant than the baseline result, and explained ~4.7-4.8% of the variance compared to ~3.5% in the baseline model. The inclusion of extra SNPs to calculate polygenic scores will increase the significant difference between cases and controls, however these may be an overestimation of the true effects due to the inclusion of correlated SNPs.

Pruning threshold	**0.8**	**0.82**	**0.84**	**0.86**	**0.88**	**0.9**
Number of SNPs	84397	85954	87651	89405	91359	93475
P value	1.10×10^{-41}	1.72×10^{-41}	3.35×10^{-42}	1.39×10^{-41}	1.56×10^{-41}	3.00×10^{-41}
Coeff.	+	+	+	+	+	+
R^2	0.0475	0.0473	0.0482	0.0474	0.0473	0.0470

Table 4.4: Logistic regression results testing polygenic scores not adjusted for any LD SNPs.
Compare to baseline result of $p = 5.41 \times 10^{-31}$, $R^2 = 0.0345$.

Pruning threshold	**0.8**	**0.82**	**0.84**	**0.86**	**0.88**	**0.9**
Total number of SNPs	84397	85954	87651	89405	91359	93475
P value	2.26×10^{-39}	2.04×10^{-38}	3.41×10^{-39}	2.89×10^{-37}	1.97×10^{-37}	1.66×10^{-36}
Coeff.	+	+	+	+	+	+
R^2	0.0447	0.0435	0.0445	0.0421	0.0423	0.0412

Table 4.5: Logistic regression results testing LD adjusted polygenic scores, where all SNPs were adjusted for an LD SNP.
Compare to baseline result of $p = 5.41 \times 10^{-31}$, $R^2 = 0.0345$.

P_{LD} produced a more significant difference between cases and controls than the standard polygenic scores from an LD pruned subset (the baseline comparison model). Table 4.5 shows this was true for all SNP sets, which were LD pruned at different thresholds. Generally the results across these thresholds were fairly stable, but more significant for SNP sets with $r^2 < 0.8$, 0.82 or 0.84. The variance explained with this method increased by up to 1% to 4.5% in the 0.8 and 0.84 pruned SNP sets compared to the baseline result.

Compared to the results in Table 4.4 for the same sets of SNPs, but not adjusted for any LD SNPs, the results in Table 4.5 were a couple of orders of magnitude less significant. While in theory the LD adjusted model should be more accurate, including extra covariates, particularly correlated covariates, reduces the accuracy of the effect size estimates. Therefore this model may be less significant compared to the unadjusted model with the same number of SNPs, as it has introduced additional errors.

Within the SNP sets, some SNPs will effectively have no correlation with any other SNP. Theoretically the adjustment for LD SNPs with weak correlation should be minimal. To confirm that controlling unnecessarily for SNPs not in LD did not stifle the association, in a second iteration LD adjusted odds ratios were only calculated if there was a SNP_{LD1} with evidence of correlation with SNP_{test}, identified by $r^2 > 0.25$. With this criterion, adjusted odds ratios were calculated for between 54% and 59% of SNPs with the percentage increasing as the pruning threshold became less conservative.

Pruning threshold	**0.8**	**0.82**	**0.84**	**0.86**	**0.88**	**0.9**
Number of SNPs adjusted for an LD SNP	46070 (54.6%)	47714 (55.5%)	49493 (56.5%)	51308 (57.4%)	53303 (58.3%)	55438 (59.3%)
P value	1.74×10^{-39}	1.40×10^{-38}	1.65×10^{-39}	1.67×10^{-37}	1.11×10^{-37}	8.45×10^{-37}
Coeff.	+	+	+	+	+	+
R^2	0.0448	0.0437	0.0449	0.0424	0.0426	0.0415

Table 4.6: Logistic regression results testing LD adjusted polygenic scores, where SNPs where only adjusted for an LD SNP if $r^2 > 0.25$.
Compare to baseline result of $p = 5.41 \times 10^{-31}$, $R^2 = 0.0345$.

Only adjusting SNPs with a neighbouring SNP in sufficiently high LD produced nominally more significant differences between cases and controls compared to adjusting all SNPs, shown in Table 4.6. In principle and practice there was no benefit to accounting for weak correlation between SNPs. As it was possible that smaller r^2 values could have occurred by chance, in all subsequent applications a lower bound

was required when calculating adjusted odds ratios. This lower bound was set to $r^2 > 0.25$ as this was the pruning threshold used to create a set of independent SNPs. An additional benefit to this decision was a reduction in computation time.

These results may still be an overestimate of the SCZ signal due to the aggregation of non-independent effects. Therefore this method was further extended to adjust the odds ratios for a second or even third LD SNP, denoted SNP_{LD2} and SNP_{LD3} respectively. Partial correlations were used to identify the best SNP_{LD2} such that the correlation with SNP_{test} was independent to that already captured by SNP_{LD1}, and SNP_{LD3} such that the correlation was independent to SNP_{LD1} and SNP_{LD2}. SNP_{LD1}, SNP_{LD2} and SNP_{LD3} were only included if there was evidence of LD with the SNP_{test}, identified as r^2 or partial correlation greater than 0.25. Therefore SNPs were adjusted for zero, one, two or three other SNPs, depending on LD structure. In each pruning set pairs of SNPs with a partial correlation greater than the original pruning threshold were ignored.

Pruning threshold	**0.8**	**0.82**	**0.84**	**0.86**	**0.88**	**0.9**
Total number of SNPs used in model	84397	85954	87651	89405	91359	93475
Number of SNPs adjusted for 2 LD SNPs	12542 (14.9%)	12951 (15.1%)	13371 (15.3%)	13757 (15.4%)	14108 (15.4%)	14434 (15.4%)
P value	1.63×10^{-37}	1.91×10^{-36}	3.38×10^{-37}	1.49×10^{-34}	3.69×10^{-35}	6.90×10^{-34}
Coeff.	+	+	+	+	+	+
R^2	0.0424	0.0411	0.0420	0.0388	0.0396	0.0380

Table 4.7: Logistic regression results testing LD adjusted polygenic scores, where SNPs were adjusted for up to two LD SNPs.
Compare to baseline result of $p = 5.41 \times 10^{-31}$, $R^2 = 0.0345$.

For each pruning threshold ~15% of all SNPs included were adjusted for two LD SNPs, results presented in Table 4.7. Controlling for a second LD SNP continued to significantly discriminate cases from controls but a couple of orders of magnitude less significantly than including one LD SNP, with a lower estimate for the variance explained. As with controlling for one LD SNP, the results were fairly consistent

across the pruning thresholds and again the more conservative thresholds (0.8, 0.82 and 0.84) had the most significant results.

Pruning threshold	**0.8**	**0.82**	**0.84**	**0.86**	**0.88**	**0.9**
Number of SNPs adjusted for 3 LD SNPs	2818 (3.3%)	2887 (3.4%)	2959 (3.4%)	3015 (3.4%)	3092 (3.4%)	3093 (3.3%)
P value	1.47×10^{-36}	1.00×10^{-34}	2.54×10^{-36}	4.46×10^{-34}	1.06×10^{-33}	1.30×10^{-32}
Coeff.	+	+	+	+	+	+
R^2	0.0412	0.0390	0.0410	0.0382	0.0378	0.0365

Table 4.8: Logistic regression results testing LD adjusted polygenic scores, where SNPS were adjusted for up to three LD SNPs.
Compare to baseline result of $p = 5.41 \times 10^{-31}$, $R^2 = 0.0345$.

When controlling for a third LD SNP, SNP_{LD3}, the proportion of SNPs for which three independent LD SNPs were identified was ~3%. This appears to have captured most of the LD structure for most SNPs in this dataset and hence no more LD SNPs were sought. The P_{LD} were still significantly different between cases and controls, shown in Table 4.8, across all pruning thresholds although the variability of the results was greater than seen with zero, one or two LD SNPs.

The decrease in significance of p values as a second and then a third LD SNP was included compared to the models adjusted for zero or one LD SNPs was again, likely due to increasingly poorly estimated effect sizes. However, even with poorly estimated effect sizes the increased number of SNPs in each of these sets meant all LD adjusted models were more significant than the baseline model on a more stringently pruned subset. The most significant result controlling for up to three LD SNPs in the $r^2 < 0.8$ SNP set explained more than 4% of the variance, which was an improvement of half a per cent on the baseline comparison.

Comparing pruning thresholds

Regression models are known not to be good at handling highly correlated variables, which can cause unreliable estimates of the coefficients or effect size. The principle

here was to include SNPs in LD, i.e. those with a correlation between them, in the regression framework and take the coefficients from these as weights for the polygenic model. Therefore it was prudent to investigate the potential impact of this on the coefficient estimates. As an attempt to prevent extreme violations of this assumption, mild pruning based on the r^2 statistic between SNPs was performed. Here the output of the logistic regression, in particular the standard errors (SE) of the coefficients for SNP_{test}, was examined to check the thresholds chosen were appropriate.

Large SE can be caused by multicolinearity and signify that unreliable estimates of the odds ratio were used to calculate the polygenic scores. For regression models with one, two or three LD SNPs the distribution of SE were plotted for each pruning threshold and in particular the tails of the distribution were looked at for extreme SE. As the output of the logistic regression with no LD SNPs was the same across all pruning thresholds, the distributions were virtually identical, varying only by the number of SNPs (and which SNPs) included so just one plot with the most SNPs was produced for comparison.

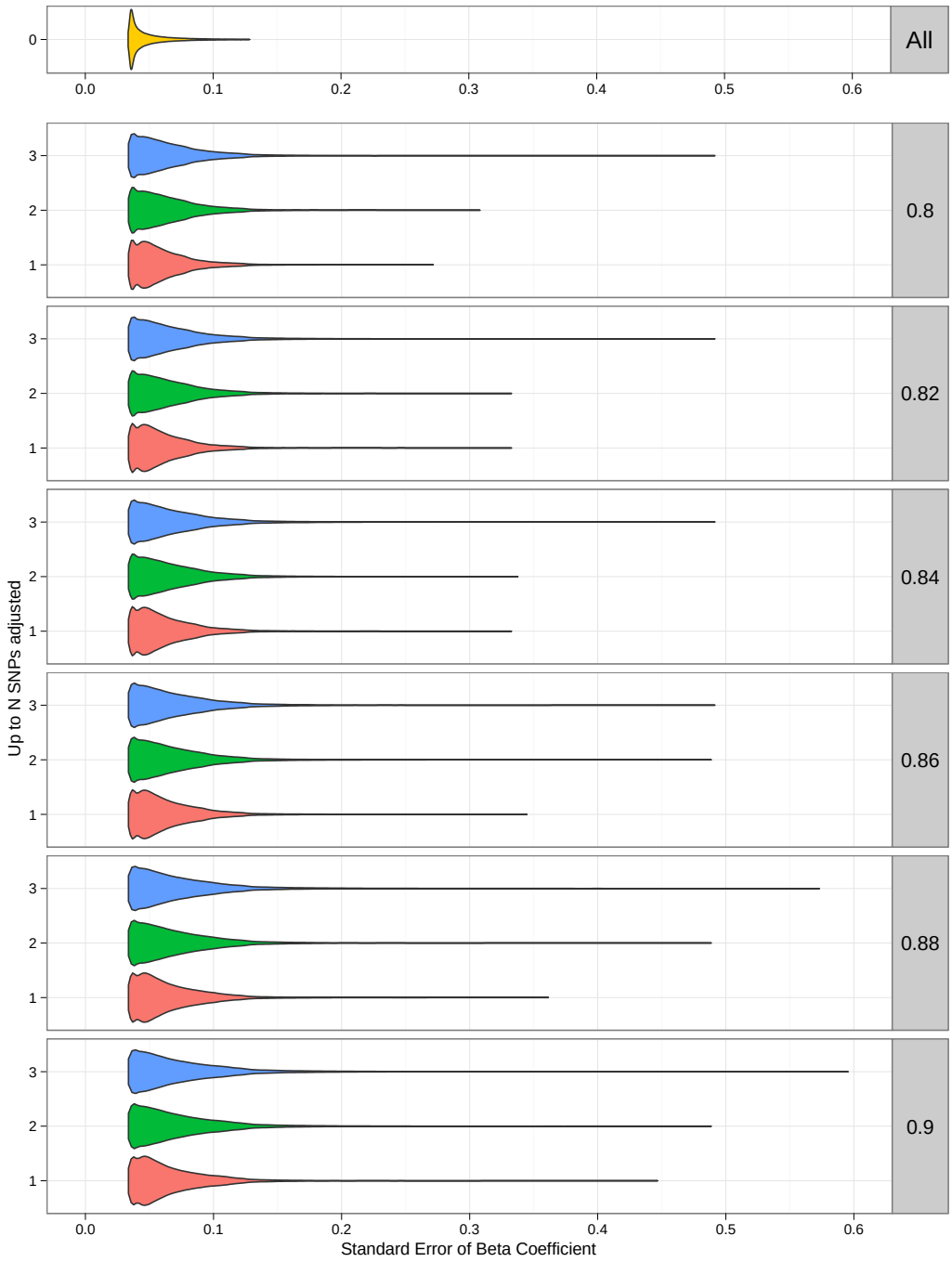

Figure 4.2: Violin plots of SE for beta coefficients for SNPs in LD adjusted polygenic models.
Each plot shows the range and frequency distribution of SE and represents a different polygenic model from Tables 4.6-4.8.Plots include SE from all SNPs used to calculate polygenic scores which will be adjusted for up to N SNPs on y axis. The violin plot at the top (yellow) is of SE from SNP set with r^2 < 0.9 not adjusted for any LD SNPs and is included for comparison.

When no LD SNPs were included, no multicolinearity was introduced and the SE for the coefficient estimates were all less than 0.13, although the distribution shows a

long right tail with a 95th percentile of 0.089, see Figure 4.2. Generally the inclusion of LD SNPs shifts the distribution to the right, increases the inter-quartile range or spread of the data and the length of the right tail. This effect was intensified as the pruning threshold became more relaxed and the possibility of higher correlations between SNPs increases. Across the more conservative thresholds (0.8, 0.82 and 0.84) the SE distributions were fairly similar when one, two or three SNPs were included in the regression. For the less conservative thresholds (0.88 and 0.9) the inclusion of the second and third LD SNPs produced the longest tails, with the 0.86 results intermediate to these two groups.

Based on these observations and trying to retain as many SNPs as possible, a threshold of 0.84 appears the correct balance for these data as it did not have many SE larger than those observed at the more conservative thresholds. This SNP set produced only marginally less significant results when including any number of LD SNPs than the best results with the 0.8 threshold. In fact the three most conservative thresholds were always more significant than the three least conservative. Despite more relaxed pruning generating larger SNP sets, the higher SE associated with including more strongly correlated markers meant that the higher proportion of SNPs with unstable estimates of effect size introduced noise.

The effect of correlated markers on estimates of the odds ratios will vary from dataset to dataset as larger samples will be able to handle multicolinearity, as well as multiple LD SNPs, better than smaller sample sizes. The size of the ISC sample appears to be big enough to avoid huge SE (> 1) even when including up to three extra SNPs correlated with the test SNP. However the increased error sizes appear to accumulate when summing over multiple SNPs and cause a decrease in significance as additional LD SNPs were included.

4.2.4 SNP-SNP interaction polygenic scores

The final adaptation to the polygenic model was to extend it to include interactions between pairs of SNPs. Significant interactions for all SNPs in the SNP_{IA} set with

another independently associated SNP within this set were identified in the following manner. SNPs were ordered by their chromosomal position and for each SNP, taken as SNP_{test}, all SNPs with $r^2 < 0.25$ from the preceding 200 and within 1000kb were considered. Of these SNP_{int} was selected as the one with the strongest individual evidence of association in the ISC GWAS, defined as the smallest association p value. As this SNP_{int} is part of the SNP_{IA} set its p value < 0.5 and therefore has some evidence of an association to SCZ. This approach ensures that for each SNP a unique interaction was considered. Equation 4.6 shows the logistic regression model fitted to estimate the odds ratios and significance of each interaction.

$$\text{status} \sim \text{SNP}_{\text{test}} + \text{SNP}_{\text{int}} + \text{SNP}_{\text{test}} * \text{SNP}_{\text{int}} + \omega$$

Equation 4.6 Logistic regression model used to identify significant SNP-SNP interactions between SNP_{test} and SNP_{int}, which are independently associated to SCZ and estimate the associated odds ratio, where status is SCZ case control status, SNP_{test} and SNP_{int} are the number of minor alleles at these SNPs and ω is the population covariates.

A SNP based polygenic score and interaction based polygenic score were calculated separately. The SNP score, S_A, was calculated as described previously, shown in Equation 4.7, where for all SNPs without a significant interaction the odds ratios were taken from the standard logistic regression model used for testing association shown in Equation 4.4A. For SNPs with a significant interaction the odds ratios were taken from Equation 4.6 so that they were adjusted for the contribution of the interaction.

$$S_{A_i} = \sum_k \beta_k \, \eta_{ik}$$

Equation 4.7 Formula to calculate SNP based polygenic score S_{Ai} for individual i, where β_k is the coefficient for SNP_{test} taken from Equation 4.6 for significant interactions or Equation 4.4A otherwise and η_{ik} is the number of associated alleles for individual i at SNP k.

The interaction score, S_I, was calculated for each significant interaction as the product of the number of minor alleles at each SNP multiplied by the natural logarithm of the associated odds ratio from Equation 4.6, shown in Equation 4.8.

$$S_{I_i} = \sum_{int} \beta_{int} \eta_{int_i} \eta_{test_i}$$

Equation 4.8: Formula to calculate interaction based polygenic score S_{Ii} for individual i, where η_{test} is the number of associated alleles at SNP_{test}, η_{int} is the number of associated alleles at SNP_{int} and β_{int} is the coefficient for the interaction term from Equation 4.6.

In any instance where the genotype was missing the expected value was taken based on the allele frequencies in the target dataset. The polygenic scores, as implemented in PLINK, are divided by the number of SNPs to give a mean score per SNP. Hence, S_I was divided by the number of significant interactions to give the mean score per interaction.

Interaction threshold		**0.5**	**0.1**	**0.05**	**0.01**	**0.005**	**0.001**
Model testing SNP score: S_A	**P value**	1.84×10^{-19}	1.18×10^{-19}	6.31×10^{-21}	1.33×10^{-30}	1.86×10^{-30}	4.19×10^{-31}
	Coeff.	+	+	+	+	+	+
	R^2	0.0203	0.0205	0.0219	0.0328	0.0326	0.0333
Model testing interaction score: S_I	**P value**	0.284	0.286	0.289	0.293	0.293	0.285
	Coeff.	-	-	-	-	-	-
	R^2	0.000288	0.000286	0.000283	0.000278	0.000277	0.000286

Table 4.9: Logistic regression results testing polygenic SNP scores and polygenic interaction scores separately.
Compare to baseline result of $p = 5.41 \times 10^{-31}$, $R^2 = 0.0345$.

Six thresholds (0.5, 0.1, 0.05, 0.01, 0.005 and 0.001) were used to identify significant interactions. The distribution of interaction p values was fairly uniform and at each threshold, the proportions that were significant were just less than would be expected by chance. Firstly the S_A and S_I were tested separately in logistic regression models predicting case control status. Table 4.9 shows the SNP scores, S_A, all significantly predicted case control status, although less significantly than in the original framework. This discrepancy was due to a proportion of the SNPs using odds ratios adjusted for a significant interaction. Fewer interactions will be found at more

stringent thresholds and therefore less of the SNP odds ratios will have been adjusted, meaning these SNP scores become closer to the baseline polygenic scores as the threshold became more significant. None of the interaction scores, S_I, significantly predicted case control status, regardless of the threshold used to select interactions for inclusion.

Both of these scores were then tested simultaneously in a joint model to see if the inclusion of the interaction term in addition to the SNP score improved the fit of the model, results presented in Table 4.10. Only at a threshold of 0.5 for significant interactions there was a trend for the interaction score to predict case control status. At all other thresholds the SNP score was significant but the interaction score was not. In sum, the inclusion of a polygenic score based on SNP-SNP interactions did not improve on just a SNP based score.

Interaction threshold		**0.5**	**0.1**	**0.05**	**0.01**	**0.005**	**0.001**
Joint model	**P value**	3.44×10^{-19}	8.30×10^{-19}	7.02×10^{-20}	1.56×10^{-29}	1.53×10^{-29}	2.44×10^{-30}
	R^2	0.0212	0.0207	0.0220	0.0329	0.0329	0.0338
SNP score	**P value**	5.27×10^{-20}	1.28×10^{-19}	1.05×10^{-20}	1.89×10^{-30}	1.84×10^{-30}	2.93×10^{-31}
	Coeff.	+	+	+	+	+	+
Interaction score	**P value**	0.0571	0.319	0.714	0.515	0.287	0.175
	Coeff.	+	+	+	-	-	-

Table 4.10: Logistic regression results testing polygenic SNP scores and polygenic interaction scores jointly.
Compare to baseline result of $p = 5.41 \times 10^{-31}$, $R^2 = 0.0345$.

4.2.5 Summary of polygenic model adaptations

Of the adaptations investigated here, the population weighted approach performed similarly to the baseline result and the inclusion of SNP-SNP interactions did not improve the model fit. Alternatively, including the LD relationships between SNPs meant that more than double the number of SNPs could be retained, which improved the significant difference in polygenic scores between cases and controls. Allowing for up to three LD SNPs captured the majority of the LD relationships and improved the variance explained by over half a per cent to more than 4%.

All of these approaches require large sample sizes to accurately estimate the odds ratios and may suffer from introducing additional inaccuracies. Despite this, in the LD adjusted model the relaxed LD pruning threshold meant more SNPs could be included and this outweighed the increased errors when estimating the effect sizes. In the next section the LD adjusted approach will be used to test gene sets identified from the expression work in the previous chapters. The pruning threshold of $r^2 < 0.84$ will be used as it was the best balance of maximising the number of SNPs and obtaining reasonable SE in the results presented so far.

4.2.6 Application to gene set identified in Chapter 3

In the previous chapter a temporal expression profile was identified for SCZ risk genes with increased expression during early and mid-foetal stages followed by a decrease of expression to the lowest values in early postnatal years before increasing though adolescence and adulthood. This was captured by identifying genes co-expressed with robustly associated SCZ genes over brain development through a linear model. Here the most enriched set, identified from the Mann-Whitney tests shown in Appendix Figure 8.15 panel A as the top 25% of genes ranked by their SCZ co-expression model p values calculated from Equation 3.3, were taken and polygenic scores calculated for the set.

These scores, calculated using the method as described by the ISC in a pruned SNP set ($r^2 < 0.25$), were found to be significantly higher in SCZ cases compared to controls ($p = 6.34 \times 10^{-12}$; $R^2 = 0.0118$). This confirms the results in the previous chapter that genes identified with this common developmental expression profile harbour multiple common risk variants associated with SCZ.

Scores calculated in a less stringently pruned SNP set ($r^2 < 0.84$) were a more significant predictor of case control status ($p = 1.02 \times 10^{-13}$; $R^2 = 0.0118$) which was as to be expected, as more SNPs were included to calculate the score. Polygenic scores were also calculated using the LD adjusted method introduced in Section 4.2.3 for up

to three LD SNPs. Of these, presented in Table 4.11, the most significant result was when up to two LD SNPs were adjusted for ($p = 3.97 \times 10^{-9}$; $R^2 = 0.00866$). Unlike with the genome-wide scores there was not a monotonic increase in p values as the number of LD SNPs included increased, further none of the LD adjusted results were more significant than the $r^2 < 0.25$ pruned unadjusted result.

Pruning threshold	**0.25**	**0.84**			
Odds ratios adjusted for up to N SNPs	**0**	**0**	**1**	**2**	**3**
Number of SNPs used to calculate scores	5362	10162	10162 (5019[a])	10162 (3746[a], 1273[b])	10162 (3746[a], 994[b], 279[c])
P value	6.34×10^{-12}	1.02×10^{-13}	1.38×10^{-7}	3.97×10^{-9}	6.59×10^{-8}
Coeff.	+	+	+	+	+
R^2	0.0118	0.0138	0.00695	0.00866	0.00730

Table 4.11: Logistic regression results testing expression gene set polygenic scores using unadjusted and LD adjusted methods.
[a] number of SNPs adjusted for 1 LD SNP, [b] number of SNPs adjusted for 2 LD SNPs, [c] number of SNPs adjusted for 3 LD SNPs.

Pruning threshold		**0.25**	**0.84**			
Odds ratios adjusted for up to N SNPs		**0**	**0**	**1**	**2**	**3**
Joint model	**P value**	1.42×10^{-32}	2.99×10^{-44}	1.77×10^{-41}	4.52×10^{-39}	1.01×10^{-37}
	R^2	0.0363	0.0494	0.0463	0.0436	0.0421
Genome-wide polygenic score	**P value**	2.01×10^{-23}	1.98×10^{-33}	1.19×10^{-36}	9.84×10^{-33}	1.44×10^{-32}
	Coeff.	+	+	+	+	+
	R^2	0.0248	0.0361	0.0396	0.0352	0.350
Gene set polygenic score	**P value**	0.00673	0.0252	0.0152	0.0111	0.0315
	Coeff.	+	+	+	+	+
	R^2	0.000186	0.00127	0.00150	0.00164	0.000117

Table 4.12: Logistic regression results jointly testing genome-wide and gene set polygenic scores.

The final step was to see if these gene set polygenic scores captured anything additional to the whole-genome polygenic score. As it was calculated with more SNPs the whole-genome polygenic score was, as expected, the more significant term, see Table 4.12. However in all five models the gene set score also remained significant implying it captured some SCZ signal not present in the whole-genome score. The gene set score was most significant in the more stringently pruned

unadjusted method. However, overall, for all models presented in Table 4.11, using both scores to predict case control status only explained slightly more of the variance compared to just using the whole genome score.

The gene set scores were also compared to genome-wide measures calculated only from genic SNPs, results shown in Table 4.13. In this scenario the gene set score was not a significant predictor when the whole genome score was included in any of the five models tested. This implies that the gene set score was a proxy for the genome-wide genic score and that there was additional genic signal outside this set. It also suggests that there was a polygenic signal for SCZ in both genic and non-genic SNPs, although this was weaker in the latter set. Therefore when the gene set score was tested against the whole genome score, which included both genic and non-genic SNPs, it remained significant as the genome-wide score had introduced a weaker signal from the non-genic SNPs. Whereas when the gene set score was tested against a genome-wide score calculated from just genic SNPs it did not contain any signal not already captured and hence was not significant.

Pruning threshold		**0.25**	**0.84**			
Odds ratios adjusted for up to N SNPs		**0**	**0**	**1**	**2**	**3**
Joint model	**P value**	1.61×10^{-27}	2.63×10^{-35}	8.32×10^{-36}	1.61×10^{-32}	3.80×10^{-31}
	R^2	0.0306	0.0394	0.0399	0.0362	0.0347
Genome-wide polygenic score	**P value**	2.61×10^{-18}	2.08×10^{-24}	6.12×10^{-31}	4.01×10^{-26}	6.25×10^{-26}
	Coeff.	+	+	+	+	+
	R^2	0.0190	0.0259	0l.0332	0.0278	0.0276
Gene set polygenic score	**P value**	0.400	0.862	0.148	0.233	0.462
	Coeff.	+	+	+	+	+
	R^2	0.000179	7.62×10^{-6}	0.000531	0.000361	0.000137

Table 4.13: Logistic regression results jointly testing genic genome-wide and gene set polygenic scores.

4.3 Discussion

4.3.1 Extensions of polygenic model

Of the adaptations described here the most successful was the inclusion of LD relationships between markers. The main benefit of this method was that the initial pruning threshold to obtain a set of independent SNPs could be relaxed, and therefore more SNPs were included when calculating the scores. From these extra SNPs additional information was retained that previously may have been missed.

Theoretically the LD adjusted model should be more accurate than a model based on the same set of SNPs that incorrectly assumes the association of each SNP is independent. However, the more complex LD adjusted model will be limited by poorer estimates of effect size. As additional LD SNPs were controlled for, from zero to three, the p values for predicting case control status became less significant. The inclusion of each additional SNP as a covariate in the logistic regression model introduced larger SE for the odds ratios with a higher proportion of SE at the extreme end of the distribution, see Figure 4.2. This effect will be further enhanced when the covariates are correlated, such as the case here. When summing across SNPs for the polygenic score, the combination of even small errors becomes significant (Dudbridge, 2013), while the SE only increased for the subset of SNPs adjusted for LD SNPs the accumulative impact of this will affect the predictive ability of the polygenic scores. This would be improved by larger sample sizes, which would improve the accuracy of the effect size estimates and be able to handle correlated covariates better.

This application, however, has shown that this approach is plausible with real data, and that the inclusion of additional SNPs still improves the discrimination of the scores even if the effect sizes have been estimated with reduced accuracy when compared to the stringently pruned, unadjusted baseline result. Ideally, a simulations procedure would be performed to directly compare the approach introduced here to the simple pruned version introduced by the ISC in scenarios

where the true variants and LD structure is known. This was beyond the time-frame of this project but would be an interesting extension.

Although SNP chips only contain a few hundred thousand SNPs, the markers are carefully selected to capture genetic variation across the majority of the genome (Hirschhorn and Daly, 2005). Based on the genotypes identified from a SNP chip experiment, many other SNPs can be imputed using the relationships between the markers and probability models. Imputation will further increase the information available to calculate polygenic scores, but the LD structure will need to be taken into account as the number of correlated associations from a GWAS based on an imputed dataset will be greater. The LD adjusted method proposed in this thesis may, therefore, enable the polygenic score approach to be applied to studies where imputation has been used.

For SCZ the polygenic model had already been shown to be highly significant, even for a stringently pruned SNP set. The LD adjusted method with a more relaxed pruning threshold improved the significance of this result further, by a couple of orders of magnitude. The benefit of this approach may be more relevant for scenarios where polygenic scores were only nominally associated to an outcome measure and the inclusion of extra SNPs through less conservative SNP pruning may help identify true signals. Such scenarios could include taking GWAS results from one phenotype to predict a different trait or disorder, or testing the polygenic nature of a subset of genes for example from a functional pathway. Caution is advised however, and users would have to consider how far the pruning threshold could be relaxed and how many LD SNPs could be included given their sample size, particularly, if this was already a factor of the initial inconclusive results.

Sample size was also a factor in the calculation of the population weighted and interaction polygenic scores. Subdividing the ISC into seven smaller subpopulations will have introduced inaccuracies when estimating the effect sizes for each population, which would have been further exacerbated when summed across populations to calculate the population weighted odds ratios, and then across SNPs

to calculate the final polygenic scores. To accurately estimate interaction odds ratios a much larger sample size is required relative to that needed when estimating SNP odds ratios, in particular for small effects which is what the polygenic model is designed to capture. In both of these approaches with the current sample size the inaccuracies of the effect sizes will have introduced noise deflating the significant difference between cases and controls and may explain why neither adaptation performed better than the baseline comparison model.

SNP-SNP interactions have been suggested to explain some of the missing heritability of complex disorders (Manolio *et al.*, 2009). It is likely in this dataset that there was not enough power to identify truly associated interactions. In the implementation described here the SNP-SNP interactions were selected as the strongest independent association from a set of neighbouring SNPs. This obviously ignored any possible interactions between SNPs on different chromosomes. Alternative ways of choosing pairs of SNPs could be further investigated and for example may look at interactions within the same gene, or between functionally related genes which may be more likely to be associated to disease. A score based on these may be more effective at discriminating between cases and controls.

An additional consideration for all of these adaptations is the additional computational time required. The polygenic method as described in the ISC paper is part of the PLINK analysis package (Purcell *et al.*, 2007) and post GWAS is very quick to implement. Each adaptation here requires additional association analyses on top of the initial GWAS to obtain an alternative effect size estimate, although in each case this was only applied to a subset of SNPs. The population weighted approach was the most computationally intensive as it also requires the calculation of the population likelihood values for each individual in the target dataset. The LD adjusted method was the easiest to implement because after adjusting the relevant odds ratios the PLINK –-score function can be used to calculate the target individuals LD adjusted polygenic scores.

4.3.2 Assessing polygenic contribution of gene set identified from Chapter 3

The rationale behind these extensions was to identify possible improvements to the polygenic framework for use with a set of genes informed as relevant for SCZ aetiology from the gene expression work in Chapters 2 and 3. Polygenic scores calculated based on such a gene set significantly predicted SCZ case control status. This was in line with the enrichment of this set for gene-wide p values combined using Brown's approach, which takes in account multiple signals within a gene in Chapter 3. Comparing the gene set polygenic scores to scores calculated from SNPs across the genome showed that they were representative of a score based on all genic SNPs suggesting SCZ signal is concentrated in genes, consistent with another study showing enrichment of association signal in genic elements for many complex diseases including SCZ and BPD (Schork *et al.*, 2013). While it has previously been shown that common variants associated through GWAS with SCZ are over-represented in brain-expressed genes (Lee *et al.*, 2012a), it would be of interest to see if the subset of genes with a temporal profile across brain development used here harbour more polygenic signal compared to random gene sets with similar numbers of SNPs. Set-based tests in Section 3.2.5 in the previous chapter, found that this set of genes contained SCZ association signal after correcting for the number of SNPs within the set, suggesting that this would be the case.

4.3.3 Summary of chapter findings

In sum, this chapter has shown that genetic prediction based on a subset of genes identified through gene expression profiles as enriched for SCZ signal can be used to discriminate between cases and controls. This suggests that within this set of genes there are likely many variants that are currently sub-threshold but as sample sizes increase will be identified as associated to SCZ. This is consistent with the findings in the previous chapter that genes within this set, identified by their co-expression with SCZ risk genes, are good candidates for SCZ aetiology. Further, functional analysis on this set of genes, as performed in Chapter 3, may help understand the biological causes of SCZ.

Chapter 5: Discussion

5.1 Identification of temporal expression profile for schizophrenia risk genes

SCZ is generally regarded as a neurodevelopmental disorder where aberrant brain development may cause symptoms to present later in life (Murray and Lewis, 1987, Weinberger, 1987). Increased rates of minor physical anomalies (Xu *et al.*, 2011) and dermatoglyphic anomalies (Golembo-Smith *et al.*, 2012) suggest that at least in some individuals a disruption occurs during gestation, prior to the formal onset of symptoms during adolescence. Further, association studies have identified genetic risk factors in genes related to brain development such as *MIR137* (Ripke *et al.*, 2011) and genes hit by CNVs found in SCZ patients have been shown to be overrepresented in pathways relating to brain development (Walsh *et al.*, 2008, Raychaudhuri *et al.*, 2010).

The purpose of this thesis was to investigate the neurodevelopmental hypothesis of SCZ, integrating transcriptomic, GWAS and CNV data to identify functional pathways. The main finding was a developmental expression profile for genes associated to SCZ. This was characterised by increased expression during foetal development, in particular during the second trimester, before a decrease prior to birth that continued to the lowest expression values around late infancy and early childhood before increasing through adolescence. This is supportive of the neurodevelopmental model for SCZ, where an insult during gestation, perhaps during the second trimester when gene expression values were greatest, may affect development to the extent that psychiatric symptoms present later in life. As discussed in Chapter 2, the second trimester has previously been suggested as a vulnerable time point for insults related to an increased risk of SCZ, with minor physical anomalies and dermatoglyphic abnormalities considered as markers for disruptions during this time frame (Lobato *et al.*, 2001).

Previous studies have investigated the temporal profile of risk genes for SCZ (Colantuoni *et al.*, 2008, Choi *et al.*, 2009, Harris *et al.*, 2009) however, this is the first study to include prenatal samples, covering the full range of human brain development up until late adulthood. Moreover, this study included multiple brain regions for each individual where previous studies have generally focused on samples from the prefrontal cortex. Chapter 2 focused on a period of mid-foetal development and identified spatial expression profiles, notably those with decreased expression in the HIP or THAL, enriched for both SCZ and BPD common variants. Gene sets with these expression patterns were also shown to have a variable expression profile across brain development with high expression during foetal development, which starts to decreases prior to birth to lower values during postnatal years until adolescence, shown in Figure 2.10.

Chapter 3 identified the same developmental expression profile described in Chapter 2 and extended it by considering samples up to late adulthood through two complementary approaches. Firstly, specific characteristic expression profiles across development were identified. Genes with increased expression during foetal development and decreased expression in early postnatal years were associated with more significant SCZ gene-wide p values. Secondly, sets of genes co-expressed with SCZ risk genes were shown to be enriched for SCZ common variants and their characteristic expression profiles matched the results of the first set of analyses. These results suggest that SCZ risk genes play a role in the development of the human brain, particularly during foetal stages when expression values were highest, but also during adolescence.

Generally, previous studies have only considered linear relationships between expression values and age for SCZ risk genes in healthy post-mortem brains (Colantuoni *et al.*, 2008, Choi *et al.*, 2009). The strategy used here to identify genes characteristic of each development stage was more in line with that of Harris *et al.* who looked for genes with their maximum or minimum expression values during the period of onset for SCZ (Harris *et al.*, 2009). The study presented in this thesis, however, considered fifteen separate development stages from early foetal through

to late adulthood. By using a regression framework for the analyses, sample differences or potential confounders could be taken into account. The enriched profile was non-linear, suggesting that to truly capture the variability of these genes simply correlating expression values with age is not sufficient.

After identifying genes associated with age, only one other study did a formal analysis to demonstrate that age dependent genes were enriched for genes associated to SCZ (Choi *et al.*, 2009). All of these studies had identified SCZ risk genes from reviews of the literature before many, if any, robust associations had been reported, questioning the validity of these lists. In the analyses presented here, results from the largest published and therefore the most reliable GWAS to date were used as a measure of SCZ or BPD association. Therefore this study has extended these works by showing that genes associated to SCZ do vary across the full range of human life and has described the trajectory of these genes.

Recently other studies have used the BrainSpan RNA-Seq or Kang microarray data to interpret their studies into the genetic causes of SCZ (Gilman *et al.*, 2012, Xu *et al.*, 2012, Gulsuner *et al.*, 2013). Their descriptions are consistent with the results reported here, finding relatively higher prenatal expression for SCZ risk genes (Gilman *et al.*, 2012, Xu *et al.*, 2012) as well as an increase during early adulthood (Gulsuner *et al.*, 2013). These findings were based on simple comparisons of prenatal and postnatal expression (Gilman *et al.*, 2012, Xu *et al.*, 2012) or just descriptions of expression plots (Gulsuner *et al.*, 2013). In contrast, the techniques used in this thesis considered each development stage separately and therefore can be more specific about the developmental trajectory of these genes.

Results in Chapters 3 and 4 showed that genes with this temporal profile contained multiple, independent common variants associated with an increased risk of SCZ. This is consistent with a previous finding that central nervous system (CNS) expressed genes are enriched for common variants associated to SCZ relative to their genomic length (Lee *et al.*, 2012a). This study estimated that their set of CNS expressed genes contained around 7% of variation of liability to SCZ. Although the

gene set investigated in Chapter 4 was approximately the same size it only explained ~1% of the variance. This value will be affected by the accuracy of the effect size estimates (Dudbridge, 2013) and is therefore an underestimate which will increase as the sample size of the discovery sample increases (Lee *et al.*, 2012a).

5.2 Identification of temporal expression profile for bipolar disorder risk genes

While typically BPD has been considered an adult disorder, increasing evidence of an overlap with SCZ has meant that it has also been investigated for neurodevelopmental antecedents with currently inconclusive findings (Sanches *et al.*, 2008). The developmental expression profile was primarily associated with more significant SCZ GWAS gene-wide p values. In Chapter 2, gene sets enriched for both SCZ and BPD common variants were observed to have this temporal expression profile. All results in Chapter 3 for BPD were consistent with the SCZ results but were generally less significant. The reduced significance in the BPD results can likely be explained by the GWAS results coming from a smaller study compared to SCZ, meaning that there was less power to detect associations for BPD in this study. However, this may also be evidence for a milder neurodevelopmental disruption compared to SCZ. The results presented here suggest that genes associated with increased risk for BPD also exhibit this temporal expression pattern, consistent with the shared genetic aetiology between SCZ and BPD particularly for common variants. Interestingly when testing the SCZ and BPD co-expression models simultaneously, a stronger enrichment for SCZ variants was found with the SCZ co-expression model whereas a stronger enrichment for BPD variants was found with the BPD co-expression model. Although these tests showed that both the SCZ and BPD co-expression models were detecting similar sets of genes, these models may be capturing some genuine differences in risk genes and further investigation would be warranted to clarify this.

5.3 Identification of functional pathways from genes with common expression profiles

Gene expression is seen as an intermediate between genotype and phenotype and therefore commonly used to infer biological mechanisms affected in SCZ aetiology (Hakak *et al.*, 2001, Katsel *et al.*, 2005b, Maycox *et al.*, 2009). Part of the aim of this thesis was to identify functional pathways from the integration of GWAS, CNV and transcriptomic data; hence expression profiles associated with SCZ and BPD were subject to pathway analysis.

Functional analysis of genes with the described temporal profile identified five groups of functional terms: 'Chromosome: structural modification & repair', 'RNA processing', 'RNA/protein transport', 'Cell cycle (mitosis)' and 'Signal transduction'. Three of these were also identified in Chapter 2, however only genes within the 'Chromosome: structural modification & repair' with the temporal profile or enriched spatial profiles in the mid-foetal brain had a significant SCZ association in both chapters. Therefore, these pathways were the most consistent across the two chapters, suggesting that SCZ risk genes play a role in epigenetic regulation through processes such as histone modification, methylation or acetylation. These mechanisms can either enhance or repress gene expression which, if disrupted, may impact on brain development and have functional consequences as the brain matures during adolescence.

Two other pathway groups 'RNA processing', and 'RNA/protein transport', also relating to the control of gene expression, were identified in both Chapters 2 and 3 but were only found to be enriched for SCZ association in Chapter 2. The lack of association in Chapter 3 may be explained by the fact that the spatial profile scores were calculated across two independent datasets to reduce the amount of noise and the number of spuriously associated terms. The temporal characteristic scores in Chapter 3 were only calculated within one dataset and therefore may contain some random variation, diffusing the SCZ association.

While predominantly interested in consistent themes across Chapters 2 and 3, the identification of the 'Cell cycle (mitosis)' functional group in Chapter 3 may also be relevant, as it was enriched for SCZ association. These terms are relevant to the production of neurons, which continues through the second trimester. Although these terms were not identified in the functional analysis in Chapter 2, Figure 2.10 showed that the HIP and THAL characteristic genes had a peak of expression during the second trimester coinciding with the period of neurogenesis, suggesting that SCZ risk genes are involved in this process. During this period the rate of production of neurons is high, meaning disruptions could potentially impact on the structure and function of the mature brain (Miranda, 2012).

5.4 Future work

Although Chapter 3 used both microarray and RNA-Seq data to provide technical replication, these datasets were not independent. Moreover, not all results were verified across both technologies for example, testing the association of genes whose expression correlated with strongly associated SCZ genes in Section 3.1.2. As an independent expression dataset covering the same age range was not available, the primary objective of any future work would be replication of the described temporal profile for SCZ and BPD risk genes.

The availability of an additional temporal expression dataset would also allow further investigation of two unanswered questions in this work. Firstly, it would ascertain if the temporal profile identified was an artefact of the different PMI between the prenatal and postnatal samples. Although including this variable as a covariate still supported the expression pattern described, it would be beneficial to confirm this in an independent dataset. Secondly, the current expression dataset had few samples between 22 and 35 PCW. It was during this time frame that expression levels started to decrease and a dataset with more complete coverage of this developmental period would help specify further when the drop in expression occurs.

An alternative replication approach would be to use an independent GWAS dataset for replication, but again no such published independent dataset exists, at least one with the same level of confidence. The PGC have since expanded their SCZ mega-analysis from 17 to 52 studies in version two, although this is yet to be published. This is obviously not independent to the GWAS already used here, but certainly is more powerful and is a useful resource to check the findings reported here hold up. Preliminary work shows that they do.

The current work could easily be extended to incorporate sequencing data alongside the GWAS and CNV data. Exome sequencing and eventually whole genome sequencing data will become increasingly prevalent and will be able to identify a range of different mutations in those affected by SCZ or BPD (Gershon *et al.*, 2011). The challenge will be distinguishing those that are disease causing from those that are part of natural variation; bioinformatic tools that predict the functional consequences of any mutation or variants that are found in multiple affected individuals may point researchers towards the right candidates (Ku *et al.*, 2013). Current exome sequencing studies for SCZ have identified genes with *de novo* functional variants and have investigated the temporal expression of these genes (Xu *et al.*, 2012, Gulsuner *et al.*, 2013). An alternative approach would be to calculate gene-wide measures of rare variants through burden tests, or even a combined measure of rare and common variation for example using the Combined Multivariate and Collapsing test (Li and Leal, 2008) and test in a similar fashion to the gene-wide p values calculated from GWAS data used here.

While this study included high-quality data from a variety of different studies, it remains a bioinformatics study. Functional mechanisms identified would need to be experimentally validated and tested to truly understand how they relate to SCZ aetiology. If the profile was confirmed in an independent dataset the next step would be to investigate how these findings relate to SCZ brains. One simple extension would be to see if genes with this temporal profile are differentially expressed between SCZ and control brains. Expression changes in SCZ brains can be influenced by many external factors, such as medication effects or lifestyle

differences such as alcohol or drug exposure, which makes it challenging to identify which changes are related to the primary disease processes from secondary responses. Ideally these issues would need to be addressed first in order to compile a list of genes that are truly dysregulated in SCZ before integrating with the analyses in this thesis.

As a similar developmental study of SCZ brains would be impossible, an alternative would be to look in peripheral tissues such as blood, which could then be compared to healthy controls. One benefit to this approach would be that the same individual could be followed up over time, removing the issue of individual variation. As SCZ is not diagnosed until adolescence, this would need to be a prospective study based on a large population cohort such as the Avon Longitudinal Study of Parents and Children (ALSPAC) (Golding *et al.*, 2001). Though even then, based on the prevalence of SCZ the number of cases is likely to be small. Such a study could lead to the identification of a biomarker which would have clinical utility (Gladkevich *et al.*, 2004, Chana *et al.*, 2013). However, any findings would always need to be linked back to the tissue of interest, the brain, in order to develop our understanding of the causes of SCZ.

A second alternative would be to look in animal models, which could be used to directly assess the impact of prenatal insults on gene expression. Such studies have already been undertaken, with rodents subjected to in utero exposure to infection assessed for expression changes compared to control offspring. Despite many genes showing differences, these are rarely consistent across animal models (Schijndel and Martens, 2010). Further, any findings will also need to be related back to the human brain.

The results presented in this thesis would recommend a follow-up investigation of expression quantitative trait loci or eQTLs in the developing human brain. This study has emphasised the importance of and the regulation of gene expression in brain development during mid-foetal and early childhood stages. If genotypes were available for the same samples, it could be tested whether expression at these time

points is genetically regulated. Further, it may be tested whether SNPs regulating gene expression at these time points are associated to SCZ in GWAS studies. These analyses would have the potential to explain some of the functional implications of genetic variants shown to increase risk of developing SCZ and would tie in with the transcriptional regulation themes arising from the GO analysis.

Given the enrichment of these genes in terms relating to epigenetic processes, one direction would be to investigate epigenetic marks, such as histone methylation levels. Based ideally on the same sample, changes in methylation levels across development could be measured to see if these correlated with the expression findings. For example, histone methylation marks associated with repression of gene expression such as dimethylation of histone H3 at lysine 9 (H3K9me2) may be predicted to be associated with SCZ risk genes during postnatal years when expression values were lowest. Epigenetic mechanisms are dynamic processes throughout development that modify gene expression, generally through changes to chromatin structure or DNA methylation and play an important role in brain development (Fagiolini *et al.*, 2009, Ma *et al.*, 2010). Epigenetic changes have been reported as a result of prenatal stresses such as exposure to famine (Tobi *et al.*, 2009) and therefore may explain the link between these and SCZ risk. Differences in DNA methylation have been found between SCZ post-mortem brains and control post-mortem brains (Mill *et al.*, 2008) and it may be of interest to investigate if these epigenetic changes are related to the gene expression changes documented here.

5.5 Concluding statement

In sum, a developmental expression profile has been identified for genes containing common variants for SCZ and BPD. This profile and pathway analyses suggest that genes associated to SCZ and BPD play a role in human brain development and the regulation of related processes. This is consistent with the neurodevelopmental hypothesis where a disruption to these processes may impact on brain function in later life.

Chapter 6: References

AFFYMETRIX 2008. Identifying and validating alternative splicing events: an introduction to managing data provided by GeneChip Exon arrays. Available at: http://www.affymetrix.com/support/technical/technotesmain.affx.

AKABALIEV, V., SIVKOV, S., MANTARKOV, M. & AHMED-POPOVA, F. 2011. Minor physical anomalies in patients with bipolar I disorder and normal controls. *J Affect Disord,* 135**,** 193-200.

ALEMAN, A., KAHN, R. S. & SELTEN, J. P. 2003. Sex differences in the risk of schizophrenia: evidence from meta-analysis. *Arch Gen Psychiatry,* 60**,** 565-71.

ALLEN, N. C., BAGADE, S., MCQUEEN, M. B., IOANNIDIS, J. P., KAVVOURA, F. K., KHOURY, M. J., TANZI, R. E. & BERTRAM, L. 2008. Systematic meta-analyses and field synopsis of genetic association studies in schizophrenia: the SzGene database. *Nat Genet,* 40**,** 827-34.

ALTAR, C. A., JURATA, L. W., CHARLES, V., LEMIRE, A., LIU, P., BUKHMAN, Y., YOUNG, T. A., BULLARD, J., YOKOE, H., WEBSTER, M. J., KNABLE, M. B. & BROCKMAN, J. A. 2005. Deficient hippocampal neuron expression of proteasome, ubiquitin, and mitochondrial genes in multiple schizophrenia cohorts. *Biol Psychiatry,* 58**,** 85-96.

AMERICAN PSYCHIATRIC ASSOCIATION 2000. *Diagnostic and statistical manual of mental disorders: DSM-IV-TR®*. Washington: American Psychiatric Publishing.

AMERICAN PSYCHIATRIC ASSOCIATION 2013. *Diagnostic and statistical manual of mental disorders, Fifth Edition (DSM-5)*. Washington: American Psychiatric Publishing.

ANDREASEN, N. C. 1995. Symptoms, signs, and diagnosis of schizophrenia. *Lancet,* 346**,** 477-81.

ANDREASSEN, O. A., THOMPSON, W. K., SCHORK, A. J., RIPKE, S., MATTINGSDAL, M., KELSOE, J. R., KENDLER, K. S., O'DONOVAN, M. C., RUJESCU, D., WERGE, T., SKLAR, P., RODDEY, J. C., CHEN, C. H., MCEVOY, L., DESIKAN, R. S., DJUROVIC, S. & DALE, A. M. 2013. Improved detection of common variants associated with schizophrenia and bipolar disorder using pleiotropy-informed conditional false discovery rate. *PLoS Genet,* 9**,** e1003455.

ARIAS, I., SORLOZANO, A., VILLEGAS, E., DE DIOS LUNA, J., MCKENNEY, K., CERVILLA, J., GUTIERREZ, B. & GUTIERREZ, J. 2012. Infectious agents associated with schizophrenia: a meta-analysis. *Schizophrenia Research,* 136**,** 128-36.

ARION, D., UNGER, T., LEWIS, D. A., LEVITT, P. & MIRNICS, K. 2007. Molecular evidence for increased expression of genes related to immune and chaperone function in the prefrontal cortex in schizophrenia. *Biol Psychiatry,* 62**,** 711-21.

ASHBURNER, M., BALL, C. A., BLAKE, J. A., BOTSTEIN, D., BUTLER, H., CHERRY, J. M., DAVIS, A. P., DOLINSKI, K., DWIGHT, S. S., EPPIG, J. T., HARRIS, M. A., HILL, D. P., ISSEL-TARVER, L., KASARSKIS, A., LEWIS, S., MATESE, J. C., RICHARDSON, J. E., RINGWALD, M., RUBIN, G. M. & SHERLOCK, G. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet,* 25**,** 25-9.

ASTON, C., JIANG, L. & SOKOLOV, B. P. 2004. Microarray analysis of postmortem temporal cortex from patients with schizophrenia. *J Neurosci Res,* 77**,** 858-66.

AYALEW, M., LE-NICULESCU, H., LEVEY, D. F., JAIN, N., CHANGALA, B., PATEL, S. D., WINIGER, E., BREIER, A., SHEKHAR, A., AMDUR, R., KOLLER, D., NURNBERGER, J. I., CORVIN, A., GEYER, M., TSUANG, M. T., SALOMON, D., SCHORK, N. J., FANOUS, A. H., O'DONOVAN, M. C. & NICULESCU, A. B. 2012. Convergent functional genomics of schizophrenia: from comprehensive understanding to genetic risk prediction. *Mol Psychiatry,* 17**,** 887-905.

AZORIN, J. M., BELLIVIER, F., KALADJIAN, A., ADIDA, M., BELZEAUX, R., FAKRA, E., HANTOUCHE, E., LANCRENON, S. & GOLMARD, J. L. 2013. Characteristics and profiles of bipolar I patients according to age-at-onset: Findings from an admixture analysis. *J Affect Disord*, 150, 993-1000.

BADNER, J. A. & GERSHON, E. S. 2002. Meta-analysis of whole-genome linkage scans of bipolar disorder and schizophrenia. *Mol Psychiatry,* 7**,** 405-11.

BALLIF, B. C., HORNOR, S. A., JENKINS, E., MADAN-KHETARPAL, S., SURTI, U., JACKSON, K. E., ASAMOAH, A., BROCK, P. L., GOWANS, G. C., CONWAY, R. L., GRAHAM, J. M., JR., MEDNE, L., ZACKAI, E. H., SHAIKH, T. H., GEOGHEGAN, J., SELZER, R. R., EIS, P. S., BEJJANI, B. A. & SHAFFER, L. G. 2007. Discovery of a previously unrecognized microdeletion syndrome of 16p11.2-p12.2. *Nat Genet,* 39**,** 1071-3.

BARNES, M. R., HUXLEY-JONES, J., MAYCOX, P. R., LENNON, M., THORNBER, A., KELLY, F., BATES, S., TAYLOR, A., REID, J., JONES, N., SCHROEDER, J., SCORER, C. A., DAVIES, C., HAGAN, J. J., KEW, J. N., ANGELINETTA, C., AKBAR, T., HIRSCH, S., MORTIMER, A. M., BARNES, T. R. & DE BELLEROCHE, J. 2011. Transcription and pathway analysis of the superior temporal cortex and anterior prefrontal cortex in schizophrenia. *J Neurosci Res,* 89**,** 1218-27.

BAUM, A. E., AKULA, N., CABANERO, M., CARDONA, I., CORONA, W., KLEMENS, B., SCHULZE, T. G., CICHON, S., RIETSCHEL, M., NOTHEN, M. M., GEORGI, A., SCHUMACHER, J., SCHWARZ, M., ABOU JAMRA, R., HOFELS, S., PROPPING, P., SATAGOPAN, J., DETERA-WADLEIGH, S. D., HARDY, J. & MCMAHON, F. J. 2008. A genome-wide association study implicates diacylglycerol kinase eta (DGKH) and several other genes in the etiology of bipolar disorder. *Mol Psychiatry,* 13**,** 197-207.

BAYER, T. A., FALKAI, P. & MAIER, W. 1999. Genetic and non-genetic vulnerability factors in schizophrenia: the basis of the "two hit hypothesis". *J Psychiatr Res,* 33**,** 543-8.

BEARDEN, C. E., ROSSO, I. M., HOLLISTER, J. M., SANCHEZ, L. E., HADLEY, T. & CANNON, T. D. 2000. A prospective cohort study of childhood behavioral deviance and language abnormalities as predictors of adult schizophrenia. *Schizophr Bull,* 26**,** 395-410.

BEAUDET, A. L. & BELMONT, J. W. 2008. Array-based DNA diagnostics: let the revolution begin. *Annu Rev Med,* 59**,** 113-29.

BENES, F. M., TURTLE, M., KHAN, Y. & FAROL, P. 1994. Myelination of a key relay zone in the hippocampal formation occurs in the human brain during childhood, adolescence, and adulthood. *Arch Gen Psychiatry,* 51**,** 477-84.

BENGTSSON, H., SIMPSON, K., BULLARD, J. & HANSEN, K. 2008. *aroma. affymetrix: A generic framework in R for analyzing small to very large Affymetrix data sets*

in bounded memory. Tech Report #745, Department of Statistics, University of California, Berkeley.

BERGEN, S. E., O'DUSHLAINE, C. T., RIPKE, S., LEE, P. H., RUDERFER, D. M., AKTERIN, S., MORAN, J. L., CHAMBERT, K. D., HANDSAKER, R. E., BACKLUND, L., OSBY, U., MCCARROLL, S., LANDEN, M., SCOLNICK, E. M., MAGNUSSON, P. K., LICHTENSTEIN, P., HULTMAN, C. M., PURCELL, S. M., SKLAR, P. & SULLIVAN, P. F. 2012. Genome-wide association study in a Swedish population yields support for greater CNV and MHC involvement in schizophrenia compared with bipolar disorder. *Mol Psychiatry,* 17**,** 880-6.

BIGDELI, T. B., FANOUS, A. H., RILEY, B. P., REIMERS, M., CHEN, X., KENDLER, K. S. & BACANU, S. A. 2013. On schizophrenia as a "disease of humanity". *Schizophrenia Research,* 143**,** 223-4.

BORA, E. & MURRAY, R. M. 2013. Meta-analysis of Cognitive Deficits in Ultra-high Risk to Psychosis and First-Episode Psychosis: Do the Cognitive Deficits Progress Over, or After, the Onset of Psychosis? *Schizophr Bull*. Advance online publication, doi: 10.1093/schbul/sbt085.

BOWDEN, N. A., SCOTT, R. J. & TOONEY, P. A. 2008. Altered gene expression in the superior temporal gyrus in schizophrenia. *BMC Genomics,* 9**,** 199.

BOZIKAS, V. P., KOSMIDIS, M. H., KIOSSEOGLOU, G. & KARAVATOS, A. 2006. Neuropsychological profile of cognitively impaired patients with schizophrenia. *Compr Psychiatry,* 47**,** 136-43.

BRACHA, H. S., TORREY, E. F., GOTTESMAN, II, BIGELOW, L. B. & CUNNIFF, C. 1992. Second-trimester markers of fetal size in schizophrenia: a study of monozygotic twins. *Am J Psychiatry,* 149**,** 1355-61.

BRAIN DEVELOPMENT COOPERATIVE GROUP 2012. Total and regional brain volumes in a population-based normative sample from 4 to 18 years: the NIH MRI Study of Normal Brain Development. *Cereb Cortex,* 22**,** 1-12.

BRAMON, E., WALSHE, M., MCDONALD, C., MARTIN, B., TOULOPOULOU, T., WICKHAM, H., VAN OS, J., FEARON, P., SHAM, P. C., FANANAS, L. & MURRAY, R. M. 2005. Dermatoglyphics and Schizophrenia: a meta-analysis and investigation of the impact of obstetric complications upon a-b ridge count. *Schizophrenia Research,* 75**,** 399-404.

BROWN, A. S., SCHAEFER, C. A., WYATT, R. J., GOETZ, R., BEGG, M. D., GORMAN, J. M. & SUSSER, E. S. 2000a. Maternal exposure to respiratory infections and adult schizophrenia spectrum disorders: a prospective birth cohort study. *Schizophr Bull,* 26**,** 287-95.

BROWN, A. S., VAN OS, J., DRIESSENS, C., HOEK, H. W. & SUSSER, E. S. 2000b. Further evidence of relation between prenatal famine and major affective disorder. *Am J Psychiatry,* 157**,** 190-5.

BROWN, M. B. 1975. A method for combining non-independent, one-sided tests of significance. *Biometrics,* 31**,** 987-992.

BUSH, W. S. & MOORE, J. H. 2012. Chapter 11: Genome-Wide Association Studies. *PLoS Comput Biol,* 8**,** e1002822.

BYRNE, M., AGERBO, E., BENNEDSEN, B., EATON, W. W. & MORTENSEN, P. B. 2007. Obstetric conditions and risk of first admission with schizophrenia: a Danish national register based study. *Schizophrenia Research,* 97**,** 51-9.

CAMPBELL, C. D., OGBURN, E. L., LUNETTA, K. L., LYON, H. N., FREEDMAN, M. L., GROOP, L. C., ALTSHULER, D., ARDLIE, K. G. & HIRSCHHORN, J. N. 2005. Demonstrating stratification in a European American population. *Nat Genet,* 37**,** 868-72.

CANNON, M., CASPI, A., MOFFITT, T. E., HARRINGTON, H., TAYLOR, A., MURRAY, R. M. & POULTON, R. 2002a. Evidence for early-childhood, pan-developmental impairment specific to schizophreniform disorder: results from a longitudinal birth cohort. *Arch Gen Psychiatry,* 59**,** 449-56.

CANNON, M., JONES, P. B. & MURRAY, R. M. 2002b. Obstetric complications and schizophrenia: historical and meta-analytic review. *Am J Psychiatry,* 159**,** 1080-92.

CARDNO, A. G. & GOTTESMAN, II 2000. Twin studies of schizophrenia: from bow-and-arrow concordances to star wars Mx and functional genomics. *Am J Med Genet,* 97**,** 12-7.

CHAN, M. K., TSANG, T. M., HARRIS, L. W., GUEST, P. C., HOLMES, E. & BAHN, S. 2011. Evidence for disease and antipsychotic medication effects in post-mortem brain from schizophrenia patients. *Mol Psychiatry,* 16**,** 1189-202.

CHANA, G., BOUSMAN, C. A., MONEY, T. T., GIBBONS, A., GILLETT, P., DEAN, B. & EVERALL, I. P. 2013. Biomarker investigations related to pathophysiological pathways in schizophrenia and psychosis. *Front Cell Neurosci,* 7**,** 95.

CHANOCK, S. J., MANOLIO, T., BOEHNKE, M., BOERWINKLE, E., HUNTER, D. J., THOMAS, G., HIRSCHHORN, J. N., ABECASIS, G., ALTSHULER, D., BAILEY-WILSON, J. E., BROOKS, L. D., CARDON, L. R., DALY, M., DONNELLY, P., FRAUMENI, J. F., JR., FREIMER, N. B., GERHARD, D. S., GUNTER, C., GUTTMACHER, A. E., GUYER, M. S., HARRIS, E. L., HOH, J., HOOVER, R., KONG, C. A., MERIKANGAS, K. R., MORTON, C. C., PALMER, L. J., PHIMISTER, E. G., RICE, J. P., ROBERTS, J., ROTIMI, C., TUCKER, M. A., VOGAN, K. J., WACHOLDER, S., WIJSMAN, E. M., WINN, D. M. & COLLINS, F. S. 2007. Replicating genotype-phenotype associations. *Nature,* 447**,** 655-60.

CHEN, D. T., JIANG, X., AKULA, N., SHUGART, Y. Y., WENDLAND, J. R., STEELE, C. J., KASSEM, L., PARK, J. H., CHATTERJEE, N., JAMAIN, S., CHENG, A., LEBOYER, M., MUGLIA, P., SCHULZE, T. G., CICHON, S., NOTHEN, M. M., RIETSCHEL, M., MCMAHON, F. J., KELSOE, J. R., GREENWOOD, T. A., NIEVERGELT, C. M., MCKINNEY, R., SHILLING, P. D., SCHORK, N. J., SMITH, E. N., BLOSS, C. S., NURNBERGER, J. I., JR., EDENBERG, H. J., FOROUD, T., KOLLER, D. L., GERSHON, E. S., LIU, C., BADNER, J. A., SCHEFTNER, W. A., LAWSON, W. B., NWULIA, E. A., HIPOLITO, M., CORYELL, W., RICE, J., BYERLEY, W., BERRETTINI, W. H., POTASH, J. B., ZANDI, P. P., MAHON, P. B., MCINNIS, M. G., ZOLLNER, S., ZHANG, P., CRAIG, D. W., SZELINGER, S. & BARRETT, T. B. 2013a. Genome-wide association study meta-analysis of European and Asian-ancestry samples identifies three novel loci associated with bipolar disorder. *Mol Psychiatry,* 18**,** 195-205.

CHEN, H., WANG, N., ZHAO, X., ROSS, C. A., O'SHEA, K. S. & MCINNIS, M. G. 2013b. Gene expression alterations in bipolar disorder postmortem brains. *Bipolar Disord,* 15**,** 177-87.

CHOI, K. H., ELASHOFF, M., HIGGS, B. W., SONG, J., KIM, S., SABUNCIYAN, S., DIGLISIC, S., YOLKEN, R. H., KNABLE, M. B., TORREY, E. F. & WEBSTER, M. J.

2008. Putative psychosis genes in the prefrontal cortex: combined analysis of gene expression microarrays. *BMC Psychiatry,* 8**,** 87.

CHOI, K. H., ZEPP, M. E., HIGGS, B. W., WEICKERT, C. S. & WEBSTER, M. J. 2009. Expression profiles of schizophrenia susceptibility genes during human prefrontal cortical development. *J Psychiatry Neurosci,* 34**,** 450-8.

CICHON, S., MUHLEISEN, T. W., DEGENHARDT, F. A., MATTHEISEN, M., MIRO, X., STROHMAIER, J., STEFFENS, M., MEESTERS, C., HERMS, S., WEINGARTEN, M., PRIEBE, L., HAENISCH, B., ALEXANDER, M., VOLLMER, J., BREUER, R., SCHMAL, C., TESSMANN, P., MOEBUS, S., WICHMANN, H. E., SCHREIBER, S., MULLER-MYHSOK, B., LUCAE, S., JAMAIN, S., LEBOYER, M., BELLIVIER, F., ETAIN, B., HENRY, C., KAHN, J. P., HEATH, S., HAMSHERE, M., O'DONOVAN, M. C., OWEN, M. J., CRADDOCK, N., SCHWARZ, M., VEDDER, H., KAMMERER-CIERNIOCH, J., REIF, A., SASSE, J., BAUER, M., HAUTZINGER, M., WRIGHT, A., MITCHELL, P. B., SCHOFIELD, P. R., MONTGOMERY, G. W., MEDLAND, S. E., GORDON, S. D., MARTIN, N. G., GUSTAFSSON, O., ANDREASSEN, O., DJUROVIC, S., SIGURDSSON, E., STEINBERG, S., STEFANSSON, H., STEFANSSON, K., KAPUR-POJSKIC, L., ORUC, L., RIVAS, F., MAYORAL, F., CHUCHALIN, A., BABADJANOVA, G., TIGANOV, A. S., PANTELEJEVA, G., ABRAMOVA, L. I., GRIGOROIU-SERBANESCU, M., DIACONU, C. C., CZERSKI, P. M., HAUSER, J., ZIMMER, A., LATHROP, M., SCHULZE, T. G., WIENKER, T. F., SCHUMACHER, J., MAIER, W., PROPPING, P., RIETSCHEL, M. & NOTHEN, M. M. 2011. Genome-wide association study identifies genetic variation in neurocan as a susceptibility factor for bipolar disorder. *Am J Hum Genet,* 88**,** 372-81.

CIRULLI, E. T. & GOLDSTEIN, D. B. 2010. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet,* 11**,** 415-25.

CLARKE, M. C., HARLEY, M. & CANNON, M. 2006. The role of obstetric events in schizophrenia. *Schizophr Bull,* 32**,** 3-8.

CLARKE, M. C., TANSKANEN, A., HUTTUNEN, M., LEON, D. A., MURRAY, R. M., JONES, P. B. & CANNON, M. 2011. Increased risk of schizophrenia from additive interaction between infant motor developmental delay and obstetric complications: evidence from a population-based longitudinal study. *Am J Psychiatry,* 168**,** 1295-302.

COHEN, O. S., MCCOY, S. Y., MIDDLETON, F. A., BIALOSUKNIA, S., ZHANG-JAMES, Y., LIU, L., TSUANG, M. T., FARAONE, S. V. & GLATT, S. J. 2012. Transcriptomic analysis of postmortem brain identifies dysregulated splicing events in novel candidate genes for schizophrenia. *Schizophrenia Research,* 142**,** 188-99.

COLANTUONI, C., HYDE, T. M., MITKUS, S., JOSEPH, A., SARTORIUS, L., AGUIRRE, C., CRESWELL, J., JOHNSON, E., DEEP-SOBOSLAY, A., HERMAN, M. M., LIPSKA, B. K., WEINBERGER, D. R. & KLEINMAN, J. E. 2008. Age-related changes in the expression of schizophrenia susceptibility genes in the human prefrontal cortex. *Brain Structure & Function,* 213**,** 255-271.

COLANTUONI, C., LIPSKA, B. K., YE, T., HYDE, T. M., TAO, R., LEEK, J. T., COLANTUONI, E. A., ELKAHLOUN, A. G., HERMAN, M. M., WEINBERGER, D. R. & KLEINMAN, J. E. 2011. Temporal dynamics and genetic control of transcription in the human prefrontal cortex. *Nature,* 478**,** 519-23.

COOPER, G. M., COE, B. P., GIRIRAJAN, S., ROSENFELD, J. A., VU, T. H., BAKER, C., WILLIAMS, C., STALKER, H., HAMID, R., HANNIG, V., ABDEL-HAMID, H., BADER, P., MCCRACKEN, E., NIYAZOV, D., LEPPIG, K., THIESE, H., HUMMEL, M., ALEXANDER, N., GORSKI, J., KUSSMANN, J., SHASHI, V., JOHNSON, K., REHDER, C., BALLIF, B. C., SHAFFER, L. G. & EICHLER, E. E. 2011. A copy number variation morbidity map of developmental delay. *Nat Genet,* 43**,** 838-46.

CORVIN, A. & GILL, M. 2003. Psychiatric genetics in the post-genome age. *Br J Psychiatry,* 182**,** 95-6.

CRADDOCK, N. & OWEN, M. J. 2005. The beginning of the end for the Kraepelinian dichotomy. *The British Journal of Psychiatry,* 186**,** 364-366.

CROW, T. J. 1997. Is schizophrenia the price that Homo sapiens pays for language? *Schizophrenia Research,* 28**,** 127-41.

DE LAU, L. M. & BRETELER, M. M. 2006. Epidemiology of Parkinson's disease. *Lancet Neurol,* 5**,** 525-35.

DEKABAN, A. S. 1978. Changes in brain weights during the span of human life: relation of brain weights to body heights and body weights. *Ann Neurol,* 4**,** 345-56.

DENNISON, M., WHITTLE, S., YUCEL, M., VIJAYAKUMAR, N., KLINE, A., SIMMONS, J. & ALLEN, N. B. 2013. Mapping subcortical brain maturation during adolescence: evidence of hemisphere- and sex-specific longitudinal changes. *Dev Sci,* 16**,** 772-91.

DEVLIN, B. & ROEDER, K. 1999. Genomic control for association studies. *Biometrics,* 55**,** 997-1004.

DICKINSON, D., RAMSEY, M. E. & GOLD, J. M. 2007. Overlooking the obvious: a meta-analytic comparison of digit symbol coding tasks and other cognitive measures in schizophrenia. *Arch Gen Psychiatry,* 64**,** 532-42.

DICKSON, H., LAURENS, K. R., CULLEN, A. E. & HODGINS, S. 2012. Meta-analyses of cognitive and motor function in youth aged 16 years and younger who subsequently develop schizophrenia. *Psychol Med,* 42**,** 743-55.

DOBBING, J. & SANDS, J. 1973. Quantitative growth and development of human brain. *Arch Dis Child,* 48**,** 757-67.

DOYLE, J. P., DOUGHERTY, J. D., HEIMAN, M., SCHMIDT, E. F., STEVENS, T. R., MA, G., BUPP, S., SHRESTHA, P., SHAH, R. D., DOUGHTY, M. L., GONG, S., GREENGARD, P. & HEINTZ, N. 2008. Application of a translational profiling approach for the comparative analysis of CNS cell types. *Cell,* 135**,** 749-62.

DU PLESSIS, L., SKUNCA, N. & DESSIMOZ, C. 2011. The what, where, how and why of gene ontology--a primer for bioinformaticians. *Brief Bioinform,* 12**,** 723-35.

DUDBRIDGE, F. 2013. Power and predictive accuracy of polygenic risk scores. *PLoS Genet,* 9**,** e1003348.

DUDBRIDGE, F. & GUSNANTO, A. 2008. Estimation of significance thresholds for genomewide association scans. *Genet Epidemiol,* 32**,** 227-34.

EBERLE, M. A., NG, P. C., KUHN, K., ZHOU, L., PEIFFER, D. A., GALVER, L., VIAUD-MARTINEZ, K. A., LAWLEY, C. T., GUNDERSON, K. L., SHEN, R. & MURRAY, S. S. 2007. Power to detect risk alleles using genome-wide tag SNP panels. *PLoS Genet,* 3**,** 1827-37.

FAGIOLINI, M., JENSEN, C. L. & CHAMPAGNE, F. A. 2009. Epigenetic influences on brain development and plasticity. *Curr Opin Neurobiol,* 19**,** 207-12.

FAN, Y., ABRAHAMSEN, G., MCGRATH, J. J. & MACKAY-SIM, A. 2012. Altered cell cycle dynamics in schizophrenia. *Biol Psychiatry,* 71**,** 129-35.

FEINBERG, I. 1982. Schizophrenia: caused by a fault in programmed synaptic elimination during adolescence? *J Psychiatr Res,* 17**,** 319-34.

FERREIRA, M. A., O'DONOVAN, M. C., MENG, Y. A., JONES, I. R., RUDERFER, D. M., JONES, L., FAN, J., KIROV, G., PERLIS, R. H., GREEN, E. K., SMOLLER, J. W., GROZEVA, D., STONE, J., NIKOLOV, I., CHAMBERT, K., HAMSHERE, M. L., NIMGAONKAR, V. L., MOSKVINA, V., THASE, M. E., CAESAR, S., SACHS, G. S., FRANKLIN, J., GORDON-SMITH, K., ARDLIE, K. G., GABRIEL, S. B., FRASER, C., BLUMENSTIEL, B., DEFELICE, M., BREEN, G., GILL, M., MORRIS, D. W., ELKIN, A., MUIR, W. J., MCGHEE, K. A., WILLIAMSON, R., MACINTYRE, D. J., MACLEAN, A. W., ST, C. D., ROBINSON, M., VAN BECK, M., PEREIRA, A. C., KANDASWAMY, R., MCQUILLIN, A., COLLIER, D. A., BASS, N. J., YOUNG, A. H., LAWRENCE, J., FERRIER, I. N., ANJORIN, A., FARMER, A., CURTIS, D., SCOLNICK, E. M., MCGUFFIN, P., DALY, M. J., CORVIN, A. P., HOLMANS, P. A., BLACKWOOD, D. H., GURLING, H. M., OWEN, M. J., PURCELL, S. M., SKLAR, P. & CRADDOCK, N. 2008. Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. *Nat Genet,* 40**,** 1056-8.

FILLMAN, S. G., CLOONAN, N., CATTS, V. S., MILLER, L. C., WONG, J., MCCROSSIN, T., CAIRNS, M. & WEICKERT, C. S. 2013. Increased inflammatory markers identified in the dorsolateral prefrontal cortex of individuals with schizophrenia. *Mol Psychiatry,* 18**,** 206-14.

GENE ONTOLOGY CONSORTIUM 2001. Creating the gene ontology resource: design and implementation. *Genome Res,* 11**,** 1425-33.

GERSHON, E. S., ALLIEY-RODRIGUEZ, N. & LIU, C. 2011. After GWAS: searching for genetic risk for schizophrenia and bipolar disorder. *Am J Psychiatry,* 168**,** 253-6.

GHEBRANIOUS, N., GIAMPIETRO, P. F., WESBROOK, F. P. & REZKALLA, S. H. 2007. A novel microdeletion at 16p11.2 harbors candidate genes for aortic valve development, seizure disorder, and mild mental retardation. *Am J Med Genet A,* 143A**,** 1462-71.

GIBBONS, A. S., THOMAS, E. A. & DEAN, B. 2009. Regional and duration of illness differences in the alteration of NCAM-180 mRNA expression within the cortex of subjects with schizophrenia. *Schizophr Res,* 112**,** 65-71.

GIEDD, J. N. 2004. Structural magnetic resonance imaging of the adolescent brain. *Ann N Y Acad Sci,* 1021**,** 77-85.

GILMAN, S. R., CHANG, J., XU, B., BAWA, T. S., GOGOS, J. A., KARAYIORGOU, M. & VITKUP, D. 2012. Diverse types of genetic variation converge on functional gene networks involved in schizophrenia. *Nat Neurosci,* 15**,** 1723-8.

GIRARD, S. L., GAUTHIER, J., NOREAU, A., XIONG, L., ZHOU, S., JOUAN, L., DIONNE-LAPORTE, A., SPIEGELMAN, D., HENRION, E., DIALLO, O., THIBODEAU, P., BACHAND, I., BAO, J. Y., TONG, A. H., LIN, C. H., MILLET, B., JAAFARI, N., JOOBER, R., DION, P. A., LOK, S., KREBS, M. O. & ROULEAU, G. A. 2011.

Increased exonic de novo mutation rate in individuals with schizophrenia. *Nat Genet,* 43**,** 860-3.

GLADKEVICH, A., KAUFFMAN, H. F. & KORF, J. 2004. Lymphocytes as a neural probe: potential for studying psychiatric disorders. *Prog Neuropsychopharmacol Biol Psychiatry,* 28**,** 559-76.

GLANTZ, L. A., GILMORE, J. H., HAMER, R. M., LIEBERMAN, J. A. & JARSKOG, L. F. 2007. Synaptophysin and postsynaptic density protein 95 in the human prefrontal cortex from mid-gestation into early adulthood. *Neuroscience,* 149**,** 582-91.

GLATT, S. J., EVERALL, I. P., KREMEN, W. S., CORBEIL, J., SASIK, R., KHANLOU, N., HAN, M., LIEW, C. C. & TSUANG, M. T. 2005. Comparative gene expression analysis of blood and brain provides concurrent validation of SELENBP1 upregulation in schizophrenia. *Proc Natl Acad Sci U S A,* 102**,** 15533-8.

GOGTAY, N., GIEDD, J. N., LUSK, L., HAYASHI, K. M., GREENSTEIN, D., VAITUZIS, A. C., NUGENT, T. F., 3RD, HERMAN, D. H., CLASEN, L. S., TOGA, A. W., RAPOPORT, J. L. & THOMPSON, P. M. 2004. Dynamic mapping of human cortical development during childhood through early adulthood. *Proc Natl Acad Sci U S A,* 101**,** 8174-9.

GOLDING, J., PEMBREY, M. & JONES, R. 2001. ALSPAC--the Avon Longitudinal Study of Parents and Children. I. Study methodology. *Paediatr Perinat Epidemiol,* 15**,** 74-87.

GOLEMBO-SMITH, S., WALDER, D. J., DALY, M. P., MITTAL, V. A., KLINE, E., REEVES, G. & SCHIFFMAN, J. 2012. The presentation of dermatoglyphic abnormalities in schizophrenia: A meta-analytic review. *Schizophrenia Research,* 142**,** 1-11.

GOODWIN, F. K. & JAMISON, K. R. 1990. *Manic depressive illness*. Oxford Univeristy Press, Incorporated.

GREEN, E. K., GROZEVA, D., JONES, I., JONES, L., KIROV, G., CAESAR, S., GORDON-SMITH, K., FRASER, C., FORTY, L., RUSSELL, E., HAMSHERE, M. L., MOSKVINA, V., NIKOLOV, I., FARMER, A., MCGUFFIN, P., HOLMANS, P. A., OWEN, M. J., O'DONOVAN, M. C. & CRADDOCK, N. 2010. The bipolar disorder risk allele at CACNA1C also confers risk of recurrent major depression and of schizophrenia. *Mol Psychiatry,* 15**,** 1016-22.

GREEN, E. K., HAMSHERE, M., FORTY, L., GORDON-SMITH, K., FRASER, C., RUSSELL, E., GROZEVA, D., KIROV, G., HOLMANS, P., MORAN, J. L., PURCELL, S., SKLAR, P., OWEN, M. J., O'DONOVAN, M. C., JONES, L., JONES, I. R. & CRADDOCK, N. 2012. Replication of bipolar disorder susceptibility alleles and identification of two novel genome-wide significant associations in a new bipolar disorder case-control sample. *Mol Psychiatry*. Advance online publication, doi: 10.1038/mp.2012.142.

GREEN, M. F., SATZ, P. & CHRISTENSON, C. 1994. Minor physical anomalies in schizophrenia patients, bipolar patients, and their siblings. *Schizophr Bull,* 20**,** 433-40.

GROZEVA, D., KIROV, G., IVANOV, D., JONES, I. R., JONES, L., GREEN, E. K., ST CLAIR, D. M., YOUNG, A. H., FERRIER, N., FARMER, A. E., MCGUFFIN, P., HOLMANS, P. A., OWEN, M. J., O'DONOVAN, M. C. & CRADDOCK, N. 2010. Rare copy number variants: a point of rarity in genetic risk for bipolar disorder and schizophrenia. *Arch Gen Psychiatry,* 67**,** 318-327.

GUEST, K. A., DYCK, B. A., SHETHWALA, S. & MISHRA, R. K. 2010. Atypical antipsychotic drugs upregulate synapsin II in the prefrontal cortex of post-mortem samples obtained from patients with schizophrenia. *Schizophrenia Research,* 120**,** 229-31.

GULSUNER, S., WALSH, T., WATTS, A. C., LEE, M. K., THORNTON, A. M., CASADEI, S., RIPPEY, C., SHAHIN, H., NIMGAONKAR, V. L., GO, R. C., SAVAGE, R. M., SWERDLOW, N. R., GUR, R. E., BRAFF, D. L., KING, M. C. & MCCLELLAN, J. M. 2013. Spatial and temporal mapping of de novo mutations in schizophrenia to a fetal prefrontal cortical network. *Cell,* 154**,** 518-29.

GUY, J. D., MAJORSKI, L. V., WALLACE, C. J. & GUY, M. P. 1983. The incidence of minor physical anomalies in adult male schizophrenics. *Schizophr Bull,* 9**,** 571-82.

HAFNER, H., MAURER, K., LOFFLER, W., FATKENHEUER, B., AN DER HEIDEN, W., RIECHER-ROSSLER, A., BEHRENS, S. & GATTAZ, W. F. 1994. The epidemiology of early schizophrenia. Influence of age and gender on onset and early course. *Br J Psychiatry Suppl***,** 29-38.

HAKAK, Y., WALKER, J. R., LI, C., WONG, W. H., DAVIS, K. L., BUXBAUM, J. D., HAROUTUNIAN, V. & FIENBERG, A. A. 2001. Genome-wide expression analysis reveals dysregulation of myelination-related genes in chronic schizophrenia. *Proc Natl Acad Sci U S A,* 98**,** 4746-51.

HAMSHERE, M. L., WALTERS, J. T., SMITH, R., RICHARDS, A. L., GREEN, E., GROZEVA, D., JONES, I., FORTY, L., JONES, L., GORDON-SMITH, K., RILEY, B., O'NEILL, F. A., KENDLER, K. S., SKLAR, P., PURCELL, S., KRANZ, J., MORRIS, D., GILL, M., HOLMANS, P., CRADDOCK, N., CORVIN, A., OWEN, M. J. & O'DONOVAN, M. C. 2013. Genome-wide significant associations in schizophrenia to ITIH3/4, CACNA1C and SDCCAG8, and extensive replication of associations reported by the Schizophrenia PGC. *Mol Psychiatry,* 18**,** 708-12.

HAROLD, D., ABRAHAM, R., HOLLINGWORTH, P., SIMS, R., GERRISH, A., HAMSHERE, M. L., PAHWA, J. S., MOSKVINA, V., DOWZELL, K., WILLIAMS, A., JONES, N., THOMAS, C., STRETTON, A., MORGAN, A. R., LOVESTONE, S., POWELL, J., PROITSI, P., LUPTON, M. K., BRAYNE, C., RUBINSZTEIN, D. C., GILL, M., LAWLOR, B., LYNCH, A., MORGAN, K., BROWN, K. S., PASSMORE, P. A., CRAIG, D., MCGUINNESS, B., TODD, S., HOLMES, C., MANN, D., SMITH, A. D., LOVE, S., KEHOE, P. G., HARDY, J., MEAD, S., FOX, N., ROSSOR, M., COLLINGE, J., MAIER, W., JESSEN, F., SCHURMANN, B., VAN DEN BUSSCHE, H., HEUSER, I., KORNHUBER, J., WILTFANG, J., DICHGANS, M., FROLICH, L., HAMPEL, H., HULL, M., RUJESCU, D., GOATE, A. M., KAUWE, J. S., CRUCHAGA, C., NOWOTNY, P., MORRIS, J. C., MAYO, K., SLEEGERS, K., BETTENS, K., ENGELBORGHS, S., DE DEYN, P. P., VAN BROECKHOVEN, C., LIVINGSTON, G., BASS, N. J., GURLING, H., MCQUILLIN, A., GWILLIAM, R., DELOUKAS, P., AL-CHALABI, A., SHAW, C. E., TSOLAKI, M., SINGLETON, A. B., GUERREIRO, R., MUHLEISEN, T. W., NOTHEN, M. M., MOEBUS, S., JOCKEL, K. H., KLOPP, N., WICHMANN, H. E., CARRASQUILLO, M. M., PANKRATZ, V. S., YOUNKIN, S. G., HOLMANS, P. A., O'DONOVAN, M., OWEN, M. J. & WILLIAMS, J. 2009. Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. *Nat Genet,* 41**,** 1088-93.

HARRIS, L. W., LOCKSTONE, H. E., KHAITOVICH, P., WEICKERT, C. S., WEBSTER, M. J. & BAHN, S. 2009. Gene expression in the prefrontal cortex during adolescence: implications for the onset of schizophrenia. *Bmc Medical Genomics,* 2**,** 14.

HASHIMOTO, T., ARION, D., UNGER, T., MALDONADO-AVILES, J. G., MORRIS, H. M., VOLK, D. W., MIRNICS, K. & LEWIS, D. A. 2008. Alterations in GABA-related transcriptome in the dorsolateral prefrontal cortex of subjects with schizophrenia. *Mol Psychiatry,* 13**,** 147-61.

HEALY, D. J. & MEADOR-WOODRUFF, J. H. 1997. Clozapine and haloperidol differentially affect AMPA and kainate receptor subunit mRNA levels in rat cortex and striatum. *Brain Res Mol Brain Res,* 47**,** 331-8.

HEBERT, J. M. & FISHELL, G. 2008. The genetics of early telencephalon patterning: some assembly required. *Nat Rev Neurosci,* 9**,** 678-85.

HEINRICHS, R. W. & ZAKZANIS, K. K. 1998. Neurocognitive deficit in schizophrenia: a quantitative review of the evidence. *Neuropsychology,* 12**,** 426-45.

HIRSCHHORN, J. N. & DALY, M. J. 2005. Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet,* 6**,** 95-108.

HOCH, R. V., RUBENSTEIN, J. L. & PLEASURE, S. 2009. Genes and signaling events that establish regional patterning of the mammalian forebrain. *Semin Cell Dev Biol,* 20**,** 378-86.

HOEK, H. W., BROWN, A. S. & SUSSER, E. 1998. The Dutch famine and schizophrenia spectrum disorders. *Soc Psychiatry Psychiatr Epidemiol,* 33**,** 373-9.

HUNTSMAN, M. M., TRAN, B. V., POTKIN, S. G., BUNNEY, W. E., JR. & JONES, E. G. 1998. Altered ratios of alternatively spliced long and short gamma2 subunit mRNAs of the gamma-amino butyrate type A receptor in prefrontal cortex of schizophrenics. *Proc Natl Acad Sci U S A,* 95**,** 15066-71.

HUTTENLOCHER, P. R. 1979. Synaptic density in human frontal cortex - developmental changes and effects of aging. *Brain Res,* 163**,** 195-205.

INGASON, A., RUJESCU, D., CICHON, S., SIGURDSSON, E., SIGMUNDSSON, T., PIETILAINEN, O. P., BUIZER-VOSKAMP, J. E., STRENGMAN, E., FRANCKS, C., MUGLIA, P., GYLFASON, A., GUSTAFSSON, O., OLASON, P. I., STEINBERG, S., HANSEN, T., JAKOBSEN, K. D., RASMUSSEN, H. B., GIEGLING, I., MOLLER, H. J., HARTMANN, A., CROMBIE, C., FRASER, G., WALKER, N., LONNQVIST, J., SUVISAARI, J., TUULIO-HENRIKSSON, A., BRAMON, E., KIEMENEY, L. A., FRANKE, B., MURRAY, R., VASSOS, E., TOULOPOULOU, T., MUHLEISEN, T. W., TOSATO, S., RUGGERI, M., DJUROVIC, S., ANDREASSEN, O. A., ZHANG, Z., WERGE, T., OPHOFF, R. A., RIETSCHEL, M., NOTHEN, M. M., PETURSSON, H., STEFANSSON, H., PELTONEN, L., COLLIER, D., STEFANSSON, K. & ST CLAIR, D. M. 2011. Copy number variations of chromosome 16p13.1 region associated with schizophrenia. *Mol Psychiatry,* 16**,** 17-25.

INTERNATIONAL HAPMAP CONSORTIUM 2005. A haplotype map of the human genome. *Nature,* 437**,** 1299-320.

INTERNATIONAL SCHIZOPHRENIA CONSORTIUM 2008. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature,* 455**,** 237-41.

ISMAIL, B., CANTOR-GRAAE, E. & MCNEIL, T. F. 1998. Minor physical anomalies in schizophrenic patients and their siblings. *Am J Psychiatry,* 155**,** 1695-702.

IWAMOTO, K., BUNDO, M. & KATO, T. 2005. Altered expression of mitochondria-related genes in postmortem brains of patients with bipolar disorder or

schizophrenia, as revealed by large-scale DNA microarray analysis. *Hum Mol Genet,* 14**,** 241-53.

IWAMOTO, K., KAKIUCHI, C., BUNDO, M., IKEDA, K. & KATO, T. 2004. Molecular characterization of bipolar disorder by comparing gene expression profiles of postmortem brains of major mental disorders. *Mol Psychiatry,* 9**,** 406-16.

JABLENSKY, A., SARTORIUS, N., ERNBERG, G., ANKER, M., KORTEN, A., COOPER, J. E., DAY, R. & BERTELSEN, A. 1992. Schizophrenia: manifestations, incidence and course in different cultures. A World Health Organization ten-country study. *Psychol Med Monogr Suppl,* 20**,** 1-97.

JOHNSON, M. B., KAWASAWA, Y. I., MASON, C. E., KRSNIK, Z., COPPOLA, G., BOGDANOVIC, D., GESCHWIND, D. H., MANE, S. M., STATE, M. W. & SESTAN, N. 2009. Functional and evolutionary insights into human brain development through global transcriptome analysis. *Neuron,* 62**,** 494-509.

JONES, P., RODGERS, B., MURRAY, R. & MARMOT, M. 1994. Child development risk factors for adult schizophrenia in the British 1946 birth cohort. *Lancet,* 344**,** 1398-402.

JOYCE, P. R. 1984. Age of onset in bipolar affective disorder and misdiagnosis as schizophrenia. *Psychol Med,* 14**,** 145-9.

KANEHISA, M. 1997. A database for post-genome analysis. *Trends Genet,* 13**,** 375-6.

KANEHISA, M. & GOTO, S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res,* 28**,** 27-30.

KANG, H. J., KAWASAWA, Y. I., CHENG, F., ZHU, Y., XU, X., LI, M., SOUSA, A. M., PLETIKOS, M., MEYER, K. A., SEDMAK, G., GUENNEL, T., SHIN, Y., JOHNSON, M. B., KRSNIK, Z., MAYER, S., FERTUZINHOS, S., UMLAUF, S., LISGO, S. N., VORTMEYER, A., WEINBERGER, D. R., MANE, S., HYDE, T. M., HUTTNER, A., REIMERS, M., KLEINMAN, J. E. & SESTAN, N. 2011. Spatio-temporal transcriptome of the human brain. *Nature,* 478**,** 483-9.

KATSEL, P., DAVIS, K. L., GORMAN, J. M. & HAROUTUNIAN, V. 2005a. Variations in differential gene expression patterns across multiple brain regions in schizophrenia. *Schizophrenia Research,* 77**,** 241-52.

KATSEL, P., DAVIS, K. L. & HAROUTUNIAN, V. 2005b. Variations in myelin and oligodendrocyte-related gene expression across multiple brain regions in schizophrenia: a gene ontology study. *Schizophrenia Research,* 79**,** 157-73.

KESHAVAN, M. S. 1999. Development, disease and degeneration in schizophrenia: a unitary pathophysiological model. *J Psychiatr Res,* 33**,** 513-21.

KESSLER, R. C., BERGLUND, P., DEMLER, O., JIN, R., MERIKANGAS, K. R. & WALTERS, E. E. 2005. Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. *Arch Gen Psychiatry,* 62**,** 593-602.

KESSLER, R. C., PETUKHOVA, M., SAMPSON, N. A., ZASLAVSKY, A. M. & WITTCHEN, H. U. 2012. Twelve-month and lifetime prevalence and lifetime morbid risk of anxiety and mood disorders in the United States. *Int J Methods Psychiatr Res,* 21**,** 169-84.

KHAITOVICH, P., LOCKSTONE, H. E., WAYLAND, M. T., TSANG, T. M., JAYATILAKA, S. D., GUO, A. J., ZHOU, J., SOMEL, M., HARRIS, L. W., HOLMES, E., PAABO, S. & BAHN, S. 2008. Metabolic changes in schizophrenia and human brain evolution. *Genome Biol,* 9**,** R124.

KHANDAKER, G. M., ZIMBRON, J., LEWIS, G. & JONES, P. B. 2013. Prenatal maternal infection, neurodevelopment and adult schizophrenia: a systematic review of population-based studies. *Psychol Med,* 43**,** 239-57.

KHATRI, P., SIROTA, M. & BUTTE, A. J. 2012. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput Biol,* 8**,** e1002375.

KIESEPPA, T., PARTONEN, T., HAUKKA, J., KAPRIO, J. & LONNQVIST, J. 2004. High concordance of bipolar I disorder in a nationwide sample of twins. *Am J Psychiatry,* 161**,** 1814-21.

KINNEY, D. K., YURGELUN-TODD, D. A., LEVY, D. L., MEDOFF, D., LAJONCHERE, C. M. & RADFORD-PAREGOL, M. 1993. Obstetrical complications in patients with bipolar disorder and their siblings. *Psychiatry Res,* 48**,** 47-56.

KINNEY, D. K., YURGELUN-TODD, D. A., TOHEN, M. & TRAMER, S. 1998. Pre- and perinatal complications and risk for bipolar disorder: a retrospective study. *J Affect Disord,* 50**,** 117-24.

KIROV, G., GROZEVA, D., NORTON, N., IVANOV, D., MANTRIPRAGADA, K. K., HOLMANS, P., CRADDOCK, N., OWEN, M. J. & O'DONOVAN, M. C. 2009a. Support for the involvement of large copy number variants in the pathogenesis of schizophrenia. *Hum Mol Genet,* 18**,** 1497-503.

KIROV, G., POCKLINGTON, A. J., HOLMANS, P., IVANOV, D., IKEDA, M., RUDERFER, D., MORAN, J., CHAMBERT, K., TONCHEVA, D., GEORGIEVA, L., GROZEVA, D., FJODOROVA, M., WOLLERTON, R., REES, E., NIKOLOV, I., VAN DE LAGEMAAT, L. N., BAYES, A., FERNANDEZ, E., OLASON, P. I., BOTTCHER, Y., KOMIYAMA, N. H., COLLINS, M. O., CHOUDHARY, J., STEFANSSON, K., STEFANSSON, H., GRANT, S. G., PURCELL, S., SKLAR, P., O'DONOVAN, M. C. & OWEN, M. J. 2012. De novo CNV analysis implicates specific abnormalities of postsynaptic signalling complexes in the pathogenesis of schizophrenia. *Mol Psychiatry,* 17**,** 142-53.

KIROV, G., RUJESCU, D., INGASON, A., COLLIER, D. A., O'DONOVAN, M. C. & OWEN, M. J. 2009b. Neurexin 1 (NRXN1) deletions in schizophrenia. *Schizophr Bull,* 35**,** 851-4.

KLEIN, R. J. 2007. Power analysis for genome-wide association studies. *BMC Genet,* 8**,** 58.

KNICKMEYER, R. C., GOUTTARD, S., KANG, C., EVANS, D., WILBER, K., SMITH, J. K., HAMER, R. M., LIN, W., GERIG, G. & GILMORE, J. H. 2008. A structural MRI study of human brain development from birth to 2 years. *J Neurosci,* 28**,** 12176-82.

KNOWLER, W. C., WILLIAMS, R. C., PETTITT, D. J. & STEINBERG, A. G. 1988. Gm3;5,13,14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture. *Am J Hum Genet,* 43**,** 520-6.

KONRADI, C., EATON, M., MACDONALD, M. L., WALSH, J., BENES, F. M. & HECKERS, S. 2004. Molecular evidence for mitochondrial dysfunction in bipolar disorder. *Arch Gen Psychiatry,* 61**,** 300-8.

KOOLSCHIJN, P. C. & CRONE, E. A. 2013. Sex differences and structural brain maturation from childhood to early adulthood. *Dev Cogn Neurosci,* 5**,** 106-18.

KRAEPELIN, E. 1899. *Psychiatrie. Ein Lehrbuch für Studirende und Aerzte, Volume 6.* Leipzig: J. A. Barth.

KRAEPELIN, E. 1919. *Dementia praecox and paraphrenia*. Livingstone.

KU, C.-S., TAN, E. K. & COOPER, D. N. 2013. From the periphery to centre stage: de novo single nucleotide variants play a key role in human genetic disease. *Journal of Medical Genetics,* 50**,** 203-211.

LANDER, E. S., LINTON, L. M., BIRREN, B., NUSBAUM, C., ZODY, M. C., BALDWIN, J., DEVON, K., DEWAR, K., DOYLE, M., FITZHUGH, W., FUNKE, R., GAGE, D., HARRIS, K., HEAFORD, A., HOWLAND, J., KANN, L., LEHOCZKY, J., LEVINE, R., MCEWAN, P., MCKERNAN, K., MELDRIM, J., MESIROV, J. P., MIRANDA, C., MORRIS, W., NAYLOR, J., RAYMOND, C., ROSETTI, M., SANTOS, R., SHERIDAN, A., SOUGNEZ, C., STANGE-THOMANN, N., STOJANOVIC, N., SUBRAMANIAN, A., WYMAN, D., ROGERS, J., SULSTON, J., AINSCOUGH, R., BECK, S., BENTLEY, D., BURTON, J., CLEE, C., CARTER, N., COULSON, A., DEADMAN, R., DELOUKAS, P., DUNHAM, A., DUNHAM, I., DURBIN, R., FRENCH, L., GRAFHAM, D., GREGORY, S., HUBBARD, T., HUMPHRAY, S., HUNT, A., JONES, M., LLOYD, C., MCMURRAY, A., MATTHEWS, L., MERCER, S., MILNE, S., MULLIKIN, J. C., MUNGALL, A., PLUMB, R., ROSS, M., SHOWNKEEN, R., SIMS, S., WATERSTON, R. H., WILSON, R. K., HILLIER, L. W., MCPHERSON, J. D., MARRA, M. A., MARDIS, E. R., FULTON, L. A., CHINWALLA, A. T., PEPIN, K. H., GISH, W. R., CHISSOE, S. L., WENDL, M. C., DELEHAUNTY, K. D., MINER, T. L., DELEHAUNTY, A., KRAMER, J. B., COOK, L. L., FULTON, R. S., JOHNSON, D. L., MINX, P. J., CLIFTON, S. W., HAWKINS, T., BRANSCOMB, E., PREDKI, P., RICHARDSON, P., WENNING, S., SLEZAK, T., DOGGETT, N., CHENG, J. F., OLSEN, A., LUCAS, S., ELKIN, C., UBERBACHER, E., FRAZIER, M., *et al.* 2001. Initial sequencing and analysis of the human genome. *Nature,* 409**,** 860-921.

LAW, A. J., LIPSKA, B. K., WEICKERT, C. S., HYDE, T. M., STRAUB, R. E., HASHIMOTO, R., HARRISON, P. J., KLEINMAN, J. E. & WEINBERGER, D. R. 2006. Neuregulin 1 transcripts are differentially expressed in schizophrenia and regulated by 5' SNPs associated with the disease. *Proc Natl Acad Sci U S A,* 103**,** 6747-52.

LE CORRE, S., HARPER, C. G., LOPEZ, P., WARD, P. & CATTS, S. 2000. Increased levels of expression of an NMDARI splice variant in the superior temporal gyrus in schizophrenia. *Neuroreport,* 11**,** 983-6.

LEE, S. H., DECANDIA, T. R., RIPKE, S., YANG, J., SULLIVAN, P. F., GODDARD, M. E., KELLER, M. C., VISSCHER, P. M. & WRAY, N. R. 2012a. Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nat Genet,* 44**,** 247-50.

LEE, S. H., RIPKE, S., NEALE, B. M., FARAONE, S. V., PURCELL, S. M., PERLIS, R. H., MOWRY, B. J., THAPAR, A., GODDARD, M. E., WITTE, J. S., ABSHER, D., AGARTZ, I., AKIL, H., AMIN, F., ANDREASSEN, O. A., ANJORIN, A., ANNEY, R., ANTTILA, V., ARKING, D. E., ASHERSON, P., AZEVEDO, M. H., BACKLUND, L., BADNER, J. A., BAILEY, A. J., BANASCHEWSKI, T., BARCHAS, J. D., BARNES, M. R., BARRETT, T. B., BASS, N., BATTAGLIA, A., BAUER, M., BAYES, M., BELLIVIER, F., BERGEN, S. E., BERRETTINI, W., BETANCUR, C., BETTECKEN, T., BIEDERMAN, J., BINDER, E. B., BLACK, D. W., BLACKWOOD, D. H., BLOSS, C. S., BOEHNKE, M., BOOMSMA, D. I., BREEN, G., BREUER, R., BRUGGEMAN, R., CORMICAN, P., BUCCOLA, N. G., BUITELAAR, J. K., BUNNEY, W. E., BUXBAUM, J. D., BYERLEY, W. F., BYRNE, E. M., CAESAR, S., CAHN, W., CANTOR, R. M., CASAS, M., CHAKRAVARTI, A., CHAMBERT, K., CHOUDHURY, K., CICHON, S., CLONINGER, C. R., COLLIER, D. A., COOK, E. H., COON, H., CORMAND, B.,

CORVIN, A., CORYELL, W. H., CRAIG, D. W., CRAIG, I. W., CROSBIE, J., CUCCARO, M. L., CURTIS, D., CZAMARA, D., DATTA, S., DAWSON, G., DAY, R., DE GEUS, E. J., DEGENHARDT, F., DJUROVIC, S., DONOHOE, G. J., DOYLE, A. E., DUAN, J., DUDBRIDGE, F., DUKETIS, E., EBSTEIN, R. P., EDENBERG, H. J., ELIA, J., ENNIS, S., ETAIN, B., FANOUS, A., FARMER, A. E., FERRIER, I. N., FLICKINGER, M., FOMBONNE, E., FOROUD, T., FRANK, J., FRANKE, B., FRASER, C., *et al.* 2013. Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nat Genet*, 45, 984-94.

LEE, S. H., YANG, J., GODDARD, M. E., VISSCHER, P. M. & WRAY, N. R. 2012b. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics,* 28**,** 2540-2.

LENROOT, R. K. & GIEDD, J. N. 2006. Brain development in children and adolescents: insights from anatomical magnetic resonance imaging. *Neurosci Biobehav Rev,* 30**,** 718-29.

LENROOT, R. K., GOGTAY, N., GREENSTEIN, D. K., WELLS, E. M., WALLACE, G. L., CLASEN, L. S., BLUMENTHAL, J. D., LERCH, J., ZIJDENBOS, A. P., EVANS, A. C., THOMPSON, P. M. & GIEDD, J. N. 2007. Sexual dimorphism of brain developmental trajectories during childhood and adolescence. *Neuroimage,* 36**,** 1065-73.

LEVINSON, D. F., DUAN, J., OH, S., WANG, K., SANDERS, A. R., SHI, J., ZHANG, N., MOWRY, B. J., OLINCY, A., AMIN, F., CLONINGER, C. R., SILVERMAN, J. M., BUCCOLA, N. G., BYERLEY, W. F., BLACK, D. W., KENDLER, K. S., FREEDMAN, R., DUDBRIDGE, F., PE'ER, I., HAKONARSON, H., BERGEN, S. E., FANOUS, A. H., HOLMANS, P. A. & GEJMAN, P. V. 2011. Copy number variants in schizophrenia: confirmation of five previous findings and new evidence for 3q29 microdeletions and VIPR2 duplications. *Am J Psychiatry,* 168**,** 302-16.

LEWIS, B. P., BURGE, C. B. & BARTEL, D. P. 2005. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell,* 120**,** 15-20.

LEWIS, D. A. & LEVITT, P. 2002. Schizophrenia as a disorder of neurodevelopment. *Annu Rev Neurosci,* 25**,** 409-32.

LI, B. & LEAL, S. M. 2008. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet,* 83**,** 311-21.

LI, M. X., GUI, H. S., KWAN, J. S. & SHAM, P. C. 2011. GATES: a rapid and powerful gene-based association test using extended Simes' procedure. *Am J Hum Genet,* 88**,** 283-93.

LI, Q., LEE, J. A. & BLACK, D. L. 2007. Neuronal regulation of alternative pre-mRNA splicing. *Nat Rev Neurosci,* 8**,** 819-31.

LICHTENSTEIN, P., YIP, B. H., BJORK, C., PAWITAN, Y., CANNON, T. D., SULLIVAN, P. F. & HULTMAN, C. M. 2009. Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study. *Lancet,* 373**,** 234-9.

LINDEN, D. 2012. *The biology of psychological disorders*. Palgrave Macmillan.

LOBATO, M. I., BELMONTE-DE-ABREU, P., KNIJNIK, D., TERUCHKIN, B., GHISOLFI, E. & HENRIQUES, A. 2001. Neurodevelopmental risk factors in schizophrenia. *Braz J Med Biol Res,* 34**,** 155-63.

MA, D. K., MARCHETTO, M. C., GUO, J. U., MING, G. L., GAGE, F. H. & SONG, H. 2010. Epigenetic choreographers of neurogenesis in the adult mammalian brain. *Nat Neurosci,* 13**,** 1338-44.

MACHON, R. A., MEDNICK, S. A. & HUTTUNEN, M. O. 1997. Adult major affective disorder after prenatal exposure to an influenza epidemic. *Arch Gen Psychiatry,* 54**,** 322-8.

MAJADAS, S., OLIVARES, J., GALAN, J. & DIEZ, T. 2012. Prevalence of depression and its relationship with other clinical characteristics in a sample of patients with stable schizophrenia. *Compr Psychiatry,* 53**,** 145-51.

MALHOTRA, D., MCCARTHY, S., MICHAELSON, J. J., VACIC, V., BURDICK, K. E., YOON, S., CICHON, S., CORVIN, A., GARY, S., GERSHON, E. S., GILL, M., KARAYIORGOU, M., KELSOE, J. R., KRASTOSHEVSKY, O., KRAUSE, V., LEIBENLUFT, E., LEVY, D. L., MAKAROV, V., BHANDARI, A., MALHOTRA, A. K., MCMAHON, F. J., NOTHEN, M. M., POTASH, J. B., RIETSCHEL, M., SCHULZE, T. G. & SEBAT, J. 2011. High frequencies of de novo CNVs in bipolar disorder and schizophrenia. *Neuron,* 72**,** 951-63.

MALHOTRA, D. & SEBAT, J. 2012. CNVs: harbingers of a rare variant revolution in psychiatric genetics. *Cell,* 148**,** 1223-41.

MANOLIO, T. A., COLLINS, F. S., COX, N. J., GOLDSTEIN, D. B., HINDORFF, L. A., HUNTER, D. J., MCCARTHY, M. I., RAMOS, E. M., CARDON, L. R., CHAKRAVARTI, A., CHO, J. H., GUTTMACHER, A. E., KONG, A., KRUGLYAK, L., MARDIS, E., ROTIMI, C. N., SLATKIN, M., VALLE, D., WHITTEMORE, A. S., BOEHNKE, M., CLARK, A. G., EICHLER, E. E., GIBSON, G., HAINES, J. L., MACKAY, T. F., MCCARROLL, S. A. & VISSCHER, P. M. 2009. Finding the missing heritability of complex diseases. *Nature,* 461**,** 747-53.

MARSHALL, C. R., NOOR, A., VINCENT, J. B., LIONEL, A. C., FEUK, L., SKAUG, J., SHAGO, M., MOESSNER, R., PINTO, D., REN, Y., THIRUVAHINDRAPDURAM, B., FIEBIG, A., SCHREIBER, S., FRIEDMAN, J., KETELAARS, C. E., VOS, Y. J., FICICIOGLU, C., KIRKPATRICK, S., NICOLSON, R., SLOMAN, L., SUMMERS, A., GIBBONS, C. A., TEEBI, A., CHITAYAT, D., WEKSBERG, R., THOMPSON, A., VARDY, C., CROSBIE, V., LUSCOMBE, S., BAATJES, R., ZWAIGENBAUM, L., ROBERTS, W., FERNANDEZ, B., SZATMARI, P. & SCHERER, S. W. 2008. Structural variation of chromosomes in autism spectrum disorder. *Am J Hum Genet,* 82**,** 477-88.

MATSUZAWA, J., MATSUI, M., KONISHI, T., NOGUCHI, K., GUR, R. C., BILKER, W. & MIYAWAKI, T. 2001. Age-related volumetric changes of brain gray and white matter in healthy infants and children. *Cereb Cortex,* 11**,** 335-42.

MAYCOX, P. R., KELLY, F., TAYLOR, A., BATES, S., REID, J., LOGENDRA, R., BARNES, M. R., LARMINIE, C., JONES, N., LENNON, M., DAVIES, C., HAGAN, J. J., SCORER, C. A., ANGELINETTA, C., AKBAR, M. T., HIRSCH, S., MORTIMER, A. M., BARNES, T. R. & DE BELLEROCHE, J. 2009. Analysis of gene expression in two large schizophrenia cohorts identifies multiple changes associated with nerve terminal function. *Mol Psychiatry,* 14**,** 1083-94.

MCCARROLL, S. A., HADNOTT, T. N., PERRY, G. H., SABETI, P. C., ZODY, M. C., BARRETT, J. C., DALLAIRE, S., GABRIEL, S. B., LEE, C., DALY, M. J. & ALTSHULER, D. M. 2006. Common deletion polymorphisms in the human genome. *Nat Genet,* 38**,** 86-92.

MCCARTHY, S. E., MAKAROV, V., KIROV, G., ADDINGTON, A. M., MCCLELLAN, J., YOON, S., PERKINS, D. O., DICKEL, D. E., KUSENDA, M., KRASTOSHEVSKY, O., KRAUSE, V., KUMAR, R. A., GROZEVA, D., MALHOTRA, D., WALSH, T., ZACKAI, E. H., KAPLAN, P., GANESH, J., KRANTZ, I. D., SPINNER, N. B., ROCCANOVA, P., BHANDARI, A., PAVON, K., LAKSHMI, B., LEOTTA, A., KENDALL, J., LEE, Y. H., VACIC, V., GARY, S., IAKOUCHEVA, L. M., CROW, T. J., CHRISTIAN, S. L., LIEBERMAN, J. A., STROUP, T. S., LEHTIMAKI, T., PUURA, K., HALDEMAN-ENGLERT, C., PEARL, J., GOODELL, M., WILLOUR, V. L., DEROSSE, P., STEELE, J., KASSEM, L., WOLFF, J., CHITKARA, N., MCMAHON, F. J., MALHOTRA, A. K., POTASH, J. B., SCHULZE, T. G., NOTHEN, M. M., CICHON, S., RIETSCHEL, M., LEIBENLUFT, E., KUSTANOVICH, V., LAJONCHERE, C. M., SUTCLIFFE, J. S., SKUSE, D., GILL, M., GALLAGHER, L., MENDELL, N. R., CRADDOCK, N., OWEN, M. J., O'DONOVAN, M. C., SHAIKH, T. H., SUSSER, E., DELISI, L. E., SULLIVAN, P. F., DEUTSCH, C. K., RAPOPORT, J., LEVY, D. L., KING, M. C. & SEBAT, J. 2009. Microduplications of 16p11.2 are associated with schizophrenia. *Nat Genet,* 41**,** 1223-7.

MCCLELLAN, J. M., SUSSER, E. & KING, M. C. 2007. Schizophrenia: a common disease caused by multiple rare alleles. *Br J Psychiatry,* 190**,** 194-9.

MCGRATH, J., SAHA, S., CHANT, D. & WELHAM, J. 2008. Schizophrenia: a concise overview of incidence, prevalence, and mortality. *Epidemiol Rev,* 30**,** 67-76.

MCGUFFIN, P., RIJSDIJK, F., ANDREW, M., SHAM, P., KATZ, R. & CARDNO, A. 2003. The heritability of bipolar affective disorder and the genetic relationship to unipolar depression. *Arch Gen Psychiatry,* 60**,** 497-502.

MCQUILLIN, A., BASS, N., ANJORIN, A., LAWRENCE, J., KANDASWAMY, R., LYDALL, G., MORAN, J., SKLAR, P., PURCELL, S. & GURLING, H. 2011. Analysis of genetic deletions and duplications in the University College London bipolar disorder case control sample. *Eur J Hum Genet,* 19**,** 588-92.

MERIKANGAS, K. R., AKISKAL, H. S., ANGST, J., GREENBERG, P. E., HIRSCHFELD, R. M., PETUKHOVA, M. & KESSLER, R. C. 2007. Lifetime and 12-month prevalence of bipolar spectrum disorder in the National Comorbidity Survey replication. *Arch Gen Psychiatry,* 64**,** 543-52.

MIDDLETON, F. A., MIRNICS, K., PIERRI, J. N., LEWIS, D. A. & LEVITT, P. 2002. Gene expression profiling reveals alterations of specific metabolic pathways in schizophrenia. *J Neurosci,* 22**,** 2718-29.

MILL, J., TANG, T., KAMINSKY, Z., KHARE, T., YAZDANPANAH, S., BOUCHARD, L., JIA, P., ASSADZADEH, A., FLANAGAN, J., SCHUMACHER, A., WANG, S. C. & PETRONIS, A. 2008. Epigenomic profiling reveals DNA-methylation changes associated with major psychosis. *Am J Hum Genet,* 82**,** 696-711.

MIRANDA, R. C. 2012. MicroRNAs and Fetal Brain Development: Implications for Ethanol Teratology during the Second Trimester Period of Neurogenesis. *Front Genet,* 3**,** 77.

MIRNICS, K., MIDDLETON, F. A., LEWIS, D. A. & LEVITT, P. 2001. Analysis of complex brain disorders with gene expression microarrays: schizophrenia as a disease of the synapse. *Trends Neurosci,* 24**,** 479-86.

MIRNICS, K., MIDDLETON, F. A., MARQUEZ, A., LEWIS, D. A. & LEVITT, P. 2000. Molecular characterization of schizophrenia viewed by microarray analysis of gene expression in prefrontal cortex. *Neuron,* 28**,** 53-67.

MISTRY, M., GILLIS, J. & PAVLIDIS, P. 2012. Genome-wide expression profiling of schizophrenia using a large combined cohort. *Mol Psychiatry*.

MISTRY, M. & PAVLIDIS, P. 2010. A cross-laboratory comparison of expression profiling data from normal human postmortem brain. *Neuroscience,* 167**,** 384-95.

MORENO-DE-LUCA, D., MULLE, J. G., KAMINSKY, E. B., SANDERS, S. J., MYERS, S. M., ADAM, M. P., PAKULA, A. T., EISENHAUER, N. J., UHAS, K., WEIK, L., GUY, L., CARE, M. E., MOREL, C. F., BONI, C., SALBERT, B. A., CHANDRAREDDY, A., DEMMER, L. A., CHOW, E. W., SURTI, U., ARADHYA, S., PICKERING, D. L., GOLDEN, D. M., SANGER, W. G., ASTON, E., BROTHMAN, A. R., GLIEM, T. J., THORLAND, E. C., ACKLEY, T., IYER, R., HUANG, S., BARBER, J. C., CROLLA, J. A., WARREN, S. T., MARTIN, C. L. & LEDBETTER, D. H. 2010. Deletion 17q12 is a recurrent copy number variant that confers high risk of autism and schizophrenia. *Am J Hum Genet,* 87**,** 618-30.

MORTENSEN, P. B., PEDERSEN, C. B., MCGRATH, J. J., HOUGAARD, D. M., NORGAARD-PETERSEN, B., MORS, O., BORGLUM, A. D. & YOLKEN, R. H. 2011. Neonatal antibodies to infectious agents and risk of bipolar disorder: a population-based case-control study. *Bipolar Disord,* 13**,** 624-9.

MOSKVINA, V., O'DUSHLAINE, C., PURCELL, S., CRADDOCK, N., HOLMANS, P. & O'DONOVAN, M. C. 2011. Evaluation of an approximation method for assessment of overall significance of multiple-dependent tests in a genomewide association study. *Genet Epidemiol,* 35**,** 861-6.

MOSKVINA, V., SCHMIDT, K. M., VEDERNIKOV, A., OWEN, M. J., CRADDOCK, N., HOLMANS, P. & O'DONOVAN, M. C. 2012. Permutation-based approaches do not adequately allow for linkage disequilibrium in gene-wide multi-locus association analysis. *Eur J Hum Genet,* 20**,** 890-6.

MULLE, J. G., DODD, A. F., MCGRATH, J. A., WOLYNIEC, P. S., MITCHELL, A. A., SHETTY, A. C., SOBREIRA, N. L., VALLE, D., RUDD, M. K., SATTEN, G., CUTLER, D. J., PULVER, A. E. & WARREN, S. T. 2010. Microdeletions of 3q29 confer high risk for schizophrenia. *Am J Hum Genet,* 87**,** 229-36.

MURRAY, R. M. & LEWIS, S. W. 1987. Is schizophrenia a neurodevelopmental disorder? *Br Med J (Clin Res Ed),* 295**,** 681-2.

NAGELKERKE, N. J. D. 1991. A note on a general definition of the coefficient of determination. *Biometrika,* 78**,** 691-692.

NAKATANI, N., HATTORI, E., OHNISHI, T., DEAN, B., IWAYAMA, Y., MATSUMOTO, I., KATO, T., OSUMI, N., HIGUCHI, T., NIWA, S. & YOSHIKAWA, T. 2006. Genome-wide expression analysis detects eight genes with robust alterations specific to bipolar I disorder: relevance to neuronal network perturbation. *Hum Mol Genet,* 15**,** 1949-62.

NALLS, M. A., PLAGNOL, V., HERNANDEZ, D. G., SHARMA, M., SHEERIN, U. M., SAAD, M., SIMON-SANCHEZ, J., SCHULTE, C., LESAGE, S., SVEINBJORNSDOTTIR, S.,

STEFANSSON, K., MARTINEZ, M., HARDY, J., HEUTINK, P., BRICE, A., GASSER, T., SINGLETON, A. B. & WOOD, N. W. 2011. Imputation of sequence variants for identification of genetic risks for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet,* 377**,** 641-9.

NASRALLAH, H. A. 1991. Neurodevelopmental aspects of bipolar affective disorder. *Biol Psychiatry,* 29**,** 1-2.

NEALE, B. M. & SHAM, P. C. 2004. The future of association studies: gene-based analysis and replication. *Am J Hum Genet,* 75**,** 353-62.

NEED, A. C., MCEVOY, J. P., GENNARELLI, M., HEINZEN, E. L., GE, D., MAIA, J. M., SHIANNA, K. V., HE, M., CIRULLI, E. T., GUMBS, C. E., ZHAO, Q., CAMPBELL, C. R., HONG, L., ROSENQUIST, P., PUTKONEN, A., HALLIKAINEN, T., REPO-TIIHONEN, E., TIIHONEN, J., LEVY, D. L., MELTZER, H. Y. & GOLDSTEIN, D. B. 2012. Exome sequencing followed by large-scale genotyping suggests a limited role for moderately rare risk factors of strong effect in schizophrenia. *Am J Hum Genet,* 91**,** 303-12.

NG, M. Y., LEVINSON, D. F., FARAONE, S. V., SUAREZ, B. K., DELISI, L. E., ARINAMI, T., RILEY, B., PAUNIO, T., PULVER, A. E., IRMANSYAH, HOLMANS, P. A., ESCAMILLA, M., WILDENAUER, D. B., WILLIAMS, N. M., LAURENT, C., MOWRY, B. J., BRZUSTOWICZ, L. M., MAZIADE, M., SKLAR, P., GARVER, D. L., ABECASIS, G. R., LERER, B., FALLIN, M. D., GURLING, H. M., GEJMAN, P. V., LINDHOLM, E., MOISES, H. W., BYERLEY, W., WIJSMAN, E. M., FORABOSCO, P., TSUANG, M. T., HWU, H. G., OKAZAKI, Y., KENDLER, K. S., WORMLEY, B., FANOUS, A., WALSH, D., O'NEILL, F. A., PELTONEN, L., NESTADT, G., LASSETER, V. K., LIANG, K. Y., PAPADIMITRIOU, G. M., DIKEOS, D. G., SCHWAB, S. G., OWEN, M. J., O'DONOVAN, M. C., NORTON, N., HARE, E., RAVENTOS, H., NICOLINI, H., ALBUS, M., MAIER, W., NIMGAONKAR, V. L., TERENIUS, L., MALLET, J., JAY, M., GODARD, S., NERTNEY, D., ALEXANDER, M., CROWE, R. R., SILVERMAN, J. M., BASSETT, A. S., ROY, M. A., MERETTE, C., PATO, C. N., PATO, M. T., ROOS, J. L., KOHN, Y., AMANN-ZALCENSTEIN, D., KALSI, G., MCQUILLIN, A., CURTIS, D., BRYNJOLFSON, J., SIGMUNDSSON, T., PETURSSON, H., SANDERS, A. R., DUAN, J., JAZIN, E., MYLES-WORSLEY, M., KARAYIORGOU, M. & LEWIS, C. M. 2009. Meta-analysis of 32 genome-wide linkage studies of schizophrenia. *Mol Psychiatry,* 14**,** 774-85.

NICULESCU, A. B. 2013. Convergent functional genomics of psychiatric disorders. *Am J Med Genet B Neuropsychiatr Genet*, 162, 587-94.

NIELSEN, P. R., LAURSEN, T. M. & MORTENSEN, P. B. 2013. Association between parental hospital-treated infection and the risk of schizophrenia in adolescence and early adulthood. *Schizophr Bull,* 39**,** 230-7.

O'CONNOR, J. A., MULY, E. C., ARNOLD, S. E. & HEMBY, S. E. 2007. AMPA receptor subunit and splice variant expression in the DLPFC of schizophrenic subjects and rhesus monkeys chronically administered antipsychotic drugs. *Schizophrenia Research,* 90**,** 28-40.

O'DONOVAN, M. C., CRADDOCK, N., NORTON, N., WILLIAMS, H., PEIRCE, T., MOSKVINA, V., NIKOLOV, I., HAMSHERE, M., CARROLL, L., GEORGIEVA, L., DWYER, S., HOLMANS, P., MARCHINI, J. L., SPENCER, C. C., HOWIE, B., LEUNG, H. T., HARTMANN, A. M., MOLLER, H. J., MORRIS, D. W., SHI, Y., FENG, G., HOFFMANN, P., PROPPING, P., VASILESCU, C., MAIER, W., RIETSCHEL, M.,

ZAMMIT, S., SCHUMACHER, J., QUINN, E. M., SCHULZE, T. G., WILLIAMS, N. M., GIEGLING, I., IWATA, N., IKEDA, M., DARVASI, A., SHIFMAN, S., HE, L., DUAN, J., SANDERS, A. R., LEVINSON, D. F., GEJMAN, P. V., CICHON, S., NOTHEN, M. M., GILL, M., CORVIN, A., RUJESCU, D., KIROV, G., OWEN, M. J., BUCCOLA, N. G., MOWRY, B. J., FREEDMAN, R., AMIN, F., BLACK, D. W., SILVERMAN, J. M., BYERLEY, W. F. & CLONINGER, C. R. 2008. Identification of loci associated with schizophrenia by genome-wide association and follow-up. *Nat Genet,* 40**,** 1053-5.

ORELLANA, G. & SLACHEVSKY, A. 2013. Executive functioning in schizophrenia. *Front Psychiatry,* 4**,** 35.

OSLER, M., LAWLOR, D. A. & NORDENTOFT, M. 2007. Cognitive function in childhood and early adulthood and hospital admission for schizophrenia and bipolar disorders in Danish men born in 1953. *Schizophrenia Research,* 92**,** 132-41.

PAN, Q., SHAI, O., LEE, L. J., FREY, B. J. & BLENCOWE, B. J. 2008. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet,* 40**,** 1413-5.

PARIKSHAK, N. N., LUO, R., ZHANG, A., WON, H., LOWE, J. K., CHANDRAN, V., HORVATH, S. & GESCHWIND, D. H. 2013. Integrative functional genomic analyses implicate specific molecular pathways and circuits in autism. *Cell,* 155**,** 1008-21.

PATEL, S. D., LE-NICULESCU, H., KOLLER, D. L., GREEN, S. D., LAHIRI, D. K., MCMAHON, F. J., NURNBERGER, J. I., JR. & NICULESCU, A. B., 3RD 2010. Coming to grips with complex disorders: genetic risk prediction in bipolar disorder using panels of genes identified through convergent functional genomics. *Am J Med Genet B Neuropsychiatr Genet,* 153B**,** 850-77.

PAUS, T. 2005. Mapping brain maturation and cognitive development during adolescence. *Trends Cogn Sci,* 9**,** 60-8.

PE'ER, I., YELENSKY, R., ALTSHULER, D. & DALY, M. J. 2008. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet Epidemiol,* 32**,** 381-5.

PEREZ-SANTIAGO, J., DIEZ-ALARCIA, R., CALLADO, L. F., ZHANG, J. X., CHANA, G., WHITE, C. H., GLATT, S. J., TSUANG, M. T., EVERALL, I. P., MEANA, J. J. & WOELK, C. H. 2012. A combined analysis of microarray gene expression studies of the human prefrontal cortex identifies genes implicated in schizophrenia. *J Psychiatr Res,* 46**,** 1464-74.

POWELL, J. E. & ZIETSCH, B. P. 2011. Predicting sensation seeking from dopamine genes: use and misuse of genetic prediction. *Psychol Sci,* 22**,** 413-5.

POWER, R. A., KYAGA, S., UHER, R., MACCABE, J. H., LANGSTROM, N., LANDEN, M., MCGUFFIN, P., LEWIS, C. M., LICHTENSTEIN, P. & SVENSSON, A. C. 2013. Fecundity of patients with schizophrenia, autism, bipolar disorder, depression, anorexia nervosa, or substance abuse vs their unaffected siblings. *JAMA Psychiatry,* 70**,** 22-30.

PRABAKARAN, S., SWATTON, J. E., RYAN, M. M., HUFFAKER, S. J., HUANG, J. T., GRIFFIN, J. L., WAYLAND, M., FREEMAN, T., DUDBRIDGE, F., LILLEY, K. S., KARP, N. A., HESTER, S., TKACHEV, D., MIMMACK, M. L., YOLKEN, R. H., WEBSTER, M. J., TORREY, E. F. & BAHN, S. 2004. Mitochondrial dysfunction in

schizophrenia: evidence for compromised brain metabolism and oxidative stress. *Mol Psychiatry,* 9**,** 684-97, 643.

PRABHAKAR, S., NOONAN, J. P., PAABO, S. & RUBIN, E. M. 2006. Accelerated evolution of conserved noncoding sequences in humans. *Science,* 314**,** 786.

PRICE, A. L., PATTERSON, N. J., PLENGE, R. M., WEINBLATT, M. E., SHADICK, N. A. & REICH, D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet,* 38**,** 904-9.

PRIEBE, L., DEGENHARDT, F. A., HERMS, S., HAENISCH, B., MATTHEISEN, M., NIERATSCHKER, V., WEINGARTEN, M., WITT, S., BREUER, R., PAUL, T., ALBLAS, M., MOEBUS, S., LATHROP, M., LEBOYER, M., SCHREIBER, S., GRIGOROIU-SERBANESCU, M., MAIER, W., PROPPING, P., RIETSCHEL, M., NOTHEN, M. M., CICHON, S. & MUHLEISEN, T. W. 2012. Genome-wide survey implicates the influence of copy number variants (CNVs) in the development of early-onset bipolar disorder. *Mol Psychiatry,* 17**,** 421-32.

PRITCHARD, J. K., STEPHENS, M., ROSENBERG, N. A. & DONNELLY, P. 2000. Association mapping in structured populations. *Am J Hum Genet,* 67**,** 170-81.

PURCELL, S., NEALE, B., TODD-BROWN, K., THOMAS, L., FERREIRA, M. A., BENDER, D., MALLER, J., SKLAR, P., DE BAKKER, P. I., DALY, M. J. & SHAM, P. C. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet,* 81**,** 559-75.

PURCELL, S. M., WRAY, N. R., STONE, J. L., VISSCHER, P. M., O'DONOVAN, M. C., SULLIVAN, P. F. & SKLAR, P. 2009. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature,* 460**,** 748-52.

PURDOM, E., SIMPSON, K. M., ROBINSON, M. D., CONBOY, J. G., LAPUK, A. V. & SPEED, T. P. 2008. FIRMA: a method for detection of alternative splicing from exon array data. *Bioinformatics,* 24**,** 1707-1714.

RAYCHAUDHURI, S., KORN, J. M., MCCARROLL, S. A., ALTSHULER, D., SKLAR, P., PURCELL, S. & DALY, M. J. 2010. Accurately assessing the risk of schizophrenia conferred by rare copy-number variation affecting genes with brain function. *PLoS Genet,* 6**,** e1001097.

REDON, R., ISHIKAWA, S., FITCH, K. R., FEUK, L., PERRY, G. H., ANDREWS, T. D., FIEGLER, H., SHAPERO, M. H., CARSON, A. R., CHEN, W., CHO, E. K., DALLAIRE, S., FREEMAN, J. L., GONZALEZ, J. R., GRATACOS, M., HUANG, J., KALAITZOPOULOS, D., KOMURA, D., MACDONALD, J. R., MARSHALL, C. R., MEI, R., MONTGOMERY, L., NISHIMURA, K., OKAMURA, K., SHEN, F., SOMERVILLE, M. J., TCHINDA, J., VALSESIA, A., WOODWARK, C., YANG, F., ZHANG, J., ZERJAL, T., ZHANG, J., ARMENGOL, L., CONRAD, D. F., ESTIVILL, X., TYLER-SMITH, C., CARTER, N. P., ABURATANI, H., LEE, C., JONES, K. W., SCHERER, S. W. & HURLES, M. E. 2006. Global variation in copy number in the human genome. *Nature,* 444**,** 444-54.

REICHENBERG, A., WEISER, M., RAPP, M. A., RABINOWITZ, J., CASPI, A., SCHMEIDLER, J., KNOBLER, H. Y., LUBIN, G., NAHON, D., HARVEY, P. D. & DAVIDSON, M. 2005. Elaboration on premorbid intellectual performance in schizophrenia: premorbid intellectual decline and risk for schizophrenia. *Arch Gen Psychiatry,* 62**,** 1297-304.

REITZ, C., BRAYNE, C. & MAYEUX, R. 2011. Epidemiology of Alzheimer disease. *Nat Rev Neurol,* 7**,** 137-52.

RHEE, S. Y., WOOD, V., DOLINSKI, K. & DRAGHICI, S. 2008. Use and misuse of the gene ontology annotations. *Nat Rev Genet,* 9**,** 509-15.

RIPKE, S., O'DUSHLAINE, C., CHAMBERT, K., MORAN, J. L., KAHLER, A. K., AKTERIN, S., BERGEN, S. E., COLLINS, A. L., CROWLEY, J. J., FROMER, M., KIM, Y., LEE, S. H., MAGNUSSON, P. K., SANCHEZ, N., STAHL, E. A., WILLIAMS, S., WRAY, N. R., XIA, K., BETTELLA, F., BORGLUM, A. D., BULIK-SULLIVAN, B. K., CORMICAN, P., CRADDOCK, N., DE LEEUW, C., DURMISHI, N., GILL, M., GOLIMBET, V., HAMSHERE, M. L., HOLMANS, P., HOUGAARD, D. M., KENDLER, K. S., LIN, K., MORRIS, D. W., MORS, O., MORTENSEN, P. B., NEALE, B. M., O'NEILL, F. A., OWEN, M. J., MILOVANCEVIC, M. P., POSTHUMA, D., POWELL, J., RICHARDS, A. L., RILEY, B. P., RUDERFER, D., RUJESCU, D., SIGURDSSON, E., SILAGADZE, T., SMIT, A. B., STEFANSSON, H., STEINBERG, S., SUVISAARI, J., TOSATO, S., VERHAGE, M., WALTERS, J. T., LEVINSON, D. F., GEJMAN, P. V., KENDLER, K. S., LAURENT, C., MOWRY, B. J., O'DONOVAN, M. C., OWEN, M. J., PULVER, A. E., RILEY, B. P., SCHWAB, S. G., WILDENAUER, D. B., DUDBRIDGE, F., HOLMANS, P., SHI, J., ALBUS, M., ALEXANDER, M., CAMPION, D., COHEN, D., DIKEOS, D., DUAN, J., EICHHAMMER, P., GODARD, S., HANSEN, M., LERER, F. B., LIANG, K. Y., MAIER, W., MALLET, J., NERTNEY, D. A., NESTADT, G., NORTON, N., O'NEILL, F. A., PAPADIMITRIOU, G. N., RIBBLE, R., SANDERS, A. R., SILVERMAN, J. M., WALSH, D., WILLIAMS, N. M., WORMLEY, B., ARRANZ, M. J., BAKKER, S., BENDER, S., BRAMON, E., COLLIER, D., CRESPO-FACORRO, B., HALL, J., IYEGBE, C., *et al.* 2013. Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat Genet*, 45, 1150-9.

RIPKE, S., SANDERS, A. R., KENDLER, K. S., LEVINSON, D. F., SKLAR, P., HOLMANS, P. A., LIN, D. Y., DUAN, J., OPHOFF, R. A., ANDREASSEN, O. A., SCOLNICK, E., CICHON, S., ST CLAIR, D., CORVIN, A., GURLING, H., WERGE, T., RUJESCU, D., BLACKWOOD, D. H., PATO, C. N., MALHOTRA, A. K., PURCELL, S., DUDBRIDGE, F., NEALE, B. M., ROSSIN, L., VISSCHER, P. M., POSTHUMA, D., RUDERFER, D. M., FANOUS, A., STEFANSSON, H., STEINBERG, S., MOWRY, B. J., GOLIMBET, V., DE HERT, M., JONSSON, E. G., BITTER, I., PIETILAINEN, O. P., COLLIER, D. A., TOSATO, S., AGARTZ, I., ALBUS, M., ALEXANDER, M., AMDUR, R. L., AMIN, F., BASS, N., BERGEN, S. E., BLACK, D. W., BORGLUM, A. D., BROWN, M. A., BRUGGEMAN, R., BUCCOLA, N. G., BYERLEY, W. F., CAHN, W., CANTOR, R. M., CARR, V. J., CATTS, S. V., CHOUDHURY, K., CLONINGER, C. R., CORMICAN, P., CRADDOCK, N., DANOY, P. A., DATTA, S., DE HAAN, L., DEMONTIS, D., DIKEOS, D., DJUROVIC, S., DONNELLY, P., DONOHOE, G., DUONG, L., DWYER, S., FINK-JENSEN, A., FREEDMAN, R., FREIMER, N. B., FRIEDL, M., GEORGIEVA, L., GIEGLING, I., GILL, M., GLENTHOJ, B., GODARD, S., HAMSHERE, M., HANSEN, M., HANSEN, T., HARTMANN, A. M., HENSKENS, F. A., HOUGAARD, D. M., HULTMAN, C. M., INGASON, A., JABLENSKY, A. V., JAKOBSEN, K. D., JAY, M., JURGENS, G., KAHN, R. S., KELLER, M. C., KENIS, G., KENNY, E., KIM, Y., KIROV, G. K., KONNERTH, H., KONTE, B., KRABBENDAM, L., KRASUCKI, R., *et al.* 2011. Genome-wide association study identifies five new schizophrenia loci. *Nat Genet,* 43**,** 969-76.

RISCH, N. & MERIKANGAS, K. 1996. The future of genetic studies of complex human diseases. *Science,* 273**,** 1516-7.

ROSEN, C., GROSSMAN, L. S., HARROW, M., BONNER-JACKSON, A. & FAULL, R. 2011. Diagnostic and prognostic significance of Schneiderian first-rank symptoms: a 20-year longitudinal study of schizophrenia and bipolar disorder. *Compr Psychiatry,* 52**,** 126-31.

ROUSSOS, P., KATSEL, P., DAVIS, K. L., SIEVER, L. J. & HAROUTUNIAN, V. 2012. A system-level transcriptomic analysis of schizophrenia using postmortem brain tissue samples. *Arch Gen Psychiatry***,** 69, 1205-13.

RYAN, M. M., LOCKSTONE, H. E., HUFFAKER, S. J., WAYLAND, M. T., WEBSTER, M. J. & BAHN, S. 2006. Gene expression analysis of bipolar disorder reveals downregulation of the ubiquitin cycle and alterations in synaptic genes. *Mol Psychiatry,* 11**,** 965-78.

SAETRE, P., EMILSSON, L., AXELSSON, E., KREUGER, J., LINDHOLM, E. & JAZIN, E. 2007. Inflammation-related genes upregulated in schizophrenia brains. *BMC Psychiatry,* 7**,** 46.

SANCHES, M., KESHAVAN, M. S., BRAMBILLA, P. & SOARES, J. C. 2008. Neurodevelopmental basis of bipolar disorder: a critical appraisal. *Prog Neuropsychopharmacol Biol Psychiatry,* 32**,** 1617-27.

SARTORIUS, L. J., WEINBERGER, D. R., HYDE, T. M., HARRISON, P. J., KLEINMAN, J. E. & LIPSKA, B. K. 2008. Expression of a GRM3 splice variant is increased in the dorsolateral prefrontal cortex of individuals carrying a schizophrenia risk SNP. *Neuropsychopharmacology,* 33**,** 2626-34.

SCHIFFMAN, J., WALKER, E., EKSTROM, M., SCHULSINGER, F., SORENSEN, H. & MEDNICK, S. 2004. Childhood videotaped social and neuromotor precursors of schizophrenia: a prospective investigation. *Am J Psychiatry,* 161**,** 2021-7.

SCHIJNDEL, J. E. & MARTENS, G. J. 2010. Gene expression profiling in rodent models for schizophrenia. *Curr Neuropharmacol,* 8**,** 382-93.

SCHMITT, A., ZINK, M., PETROIANU, G., MAY, B., BRAUS, D. F. & HENN, F. A. 2003. Decreased gene expression of glial and neuronal glutamate transporters after chronic antipsychotic treatment in rat brain. *Neurosci Lett,* 347**,** 81-4.

SCHNEIDER, K. 1959. *Clinical psychopathology*. New York: Grune & Stratton.

SCHORK, A. J., THOMPSON, W. K., PHAM, P., TORKAMANI, A., RODDEY, J. C., SULLIVAN, P. F., KELSOE, J. R., O'DONOVAN, M. C., FURBERG, H., SCHORK, N. J., ANDREASSEN, O. A. & DALE, A. M. 2013. All SNPs are not created equal: genome-wide association studies reveal a consistent pattern of enrichment among functionally annotated SNPs. *PLoS Genet,* 9**,** e1003449.

SCHURHOFF, F., BELLIVIER, F., JOUVENT, R., MOUREN-SIMEONI, M. C., BOUVARD, M., ALLILAIRE, J. F. & LEBOYER, M. 2000. Early and late onset bipolar disorders: two different forms of manic-depressive illness? *J Affect Disord,* 58**,** 215-21.

SCOTT, J., MCNEILL, Y., CAVANAGH, J., CANNON, M. & MURRAY, R. 2006. Exposure to obstetric complications and subsequent development of bipolar disorder: Systematic review. *Br J Psychiatry,* 189**,** 3-11.

SEBAT, J., LAKSHMI, B., MALHOTRA, D., TROGE, J., LESE-MARTIN, C., WALSH, T., YAMROM, B., YOON, S., KRASNITZ, A., KENDALL, J., LEOTTA, A., PAI, D., ZHANG, R., LEE, Y. H., HICKS, J., SPENCE, S. J., LEE, A. T., PUURA, K., LEHTIMAKI, T., LEDBETTER, D., GREGERSEN, P. K., BREGMAN, J., SUTCLIFFE, J. S., JOBANPUTRA, V., CHUNG, W., WARBURTON, D., KING, M. C., SKUSE, D.,

GESCHWIND, D. H., GILLIAM, T. C., YE, K. & WIGLER, M. 2007. Strong association of de novo copy number mutations with autism. *Science,* 316**,** 445-9.

SEIDMAN, L. J., CHERKERZIAN, S., GOLDSTEIN, J. M., AGNEW-BLAIS, J., TSUANG, M. T. & BUKA, S. L. 2012. Neuropsychological performance and family history in children at age 7 who develop adult schizophrenia or bipolar psychosis in the New England Family Studies. *Psychol Med***,** 43**,** 119-31.

SEIFUDDIN, F., MAHON, P. B., JUDY, J., PIROOZNIA, M., JANCIC, D., TAYLOR, J., GOES, F. S., POTASH, J. B. & ZANDI, P. P. 2012. Meta-analysis of genetic association studies on bipolar disorder. *Am J Med Genet B Neuropsychiatr Genet,* 159B**,** 508-18.

SELDIN, M. F., SHIGETA, R., VILLOSLADA, P., SELMI, C., TUOMILEHTO, J., SILVA, G., BELMONT, J. W., KLARESKOG, L. & GREGERSEN, P. K. 2006. European population substructure: clustering of northern and southern populations. *PLoS Genet,* 2**,** e143.

SEQUEIRA, P. A., MARTIN, M. V. & VAWTER, M. P. 2012. The first decade and beyond of transcriptional profiling in schizophrenia. *Neurobiol Dis,* 45**,** 23-36.

SHAO, L. & VAWTER, M. P. 2008. Shared gene expression alterations in schizophrenia and bipolar disorder. *Biol Psychiatry,* 64**,** 89-97.

SHI, J., LEVINSON, D. F., DUAN, J., SANDERS, A. R., ZHENG, Y., PE'ER, I., DUDBRIDGE, F., HOLMANS, P. A., WHITTEMORE, A. S., MOWRY, B. J., OLINCY, A., AMIN, F., CLONINGER, C. R., SILVERMAN, J. M., BUCCOLA, N. G., BYERLEY, W. F., BLACK, D. W., CROWE, R. R., OKSENBERG, J. R., MIREL, D. B., KENDLER, K. S., FREEDMAN, R. & GEJMAN, P. V. 2009. Common variants on chromosome 6p22.1 are associated with schizophrenia. *Nature,* 460**,** 753-7.

SIMES, R. J. 1986. An improved Bonferroni procedure for multiple tests of significance. *Biometrika,* 73**,** 751-4.

SKLAR, P., RIPKE, S., SCOTT, L. J., ANDREASSEN, O. A., CICHON, S., CRADDOCK, N., EDENBERG, H. J., NURNBERGER, J. I., JR., RIETSCHEL, M., BLACKWOOD, D., CORVIN, A., FLICKINGER, M., GUAN, W., MATTINGSDAL, M., MCQUILLIN, A., KWAN, P., WIENKER, T. F., DALY, M., DUDBRIDGE, F., HOLMANS, P. A., LIN, D., BURMEISTER, M., GREENWOOD, T. A., HAMSHERE, M. L., MUGLIA, P., SMITH, E. N., ZANDI, P. P., NIEVERGELT, C. M., MCKINNEY, R., SHILLING, P. D., SCHORK, N. J., BLOSS, C. S., FOROUD, T., KOLLER, D. L., GERSHON, E. S., LIU, C., BADNER, J. A., SCHEFTNER, W. A., LAWSON, W. B., NWULIA, E. A., HIPOLITO, M., CORYELL, W., RICE, J., BYERLEY, W., MCMAHON, F. J., SCHULZE, T. G., BERRETTINI, W., LOHOFF, F. W., POTASH, J. B., MAHON, P. B., MCINNIS, M. G., ZOLLNER, S., ZHANG, P., CRAIG, D. W., SZELINGER, S., BARRETT, T. B., BREUER, R., MEIER, S., STROHMAIER, J., WITT, S. H., TOZZI, F., FARMER, A., MCGUFFIN, P., STRAUSS, J., XU, W., KENNEDY, J. L., VINCENT, J. B., MATTHEWS, K., DAY, R., FERREIRA, M. A., O'DUSHLAINE, C., PERLIS, R., RAYCHAUDHURI, S., RUDERFER, D., HYOUN, P. L., SMOLLER, J. W., LI, J., ABSHER, D., THOMPSON, R. C., MENG, F. G., SCHATZBERG, A. F., BUNNEY, W. E., BARCHAS, J. D., JONES, E. G., WATSON, S. J., MYERS, R. M., AKIL, H., BOEHNKE, M., CHAMBERT, K., MORAN, J., SCOLNICK, E., DJUROVIC, S., MELLE, I., MORKEN, G., GILL, M., MORRIS, D., QUINN, E., MUHLEISEN, T. W., DEGENHARDT, F. A., MATTHEISEN, M., *et al.* 2011. Large-scale genome-wide

association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nat Genet,* 43**,** 977-83.

SKUNCA, N., ALTENHOFF, A. & DESSIMOZ, C. 2012. Quality of computationally inferred gene ontology annotations. *PLoS Comput Biol,* 8**,** e1002533.

SMOLLER, J. W., CRADDOCK, N., KENDLER, K., LEE, P. H., NEALE, B. M., NURNBERGER, J. I., RIPKE, S., SANTANGELO, S. & SULLIVAN, P. F. 2013. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet,* 381**,** 1371-9.

SMRT, R. D., SZULWACH, K. E., PFEIFFER, R. L., LI, X., GUO, W., PATHANIA, M., TENG, Z. Q., LUO, Y., PENG, J., BORDEY, A., JIN, P. & ZHAO, X. 2010. MicroRNA miR-137 regulates neuronal maturation by targeting ubiquitin ligase mind bomb-1. *Stem Cells,* 28**,** 1060-70.

SMYTH, G. 2005. Limma: linear models for microarray data. *Bioinformatics and computational biology solutions using R and Bioconductor***,** 397-420.

SORENSEN, H. J., MORTENSEN, E. L., REINISCH, J. M. & MEDNICK, S. A. 2009. Association between prenatal exposure to bacterial infection and risk of schizophrenia. *Schizophr Bull,* 35**,** 631-7.

SOWELL, E. R., THOMPSON, P. M., TESSNER, K. D. & TOGA, A. W. 2001. Mapping continued brain growth and gray matter density reduction in dorsal frontal cortex: Inverse relationships during postadolescent brain maturation. *J Neurosci,* 21**,** 8819-29.

SPENCER, C. C. A., 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, e1000477.

ST CLAIR, D., XU, M., WANG, P., YU, Y., FANG, Y., ZHANG, F., ZHENG, X., GU, N., FENG, G., SHAM, P. & HE, L. 2005. Rates of adult schizophrenia following prenatal exposure to the Chinese famine of 1959-1961. *JAMA,* 294**,** 557-62.

STAHL, E. A., WEGMANN, D., TRYNKA, G., GUTIERREZ-ACHURY, J., DO, R., VOIGHT, B. F., KRAFT, P., CHEN, R., KALLBERG, H. J., KURREEMAN, F. A., KATHIRESAN, S., WIJMENGA, C., GREGERSEN, P. K., ALFREDSSON, L., SIMINOVITCH, K. A., WORTHINGTON, J., DE BAKKER, P. I., RAYCHAUDHURI, S. & PLENGE, R. M. 2012. Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. *Nat Genet,* 44**,** 483-9.

STEFANSSON, H., OPHOFF, R. A., STEINBERG, S., ANDREASSEN, O. A., CICHON, S., RUJESCU, D., WERGE, T., PIETILAINEN, O. P., MORS, O., MORTENSEN, P. B., SIGURDSSON, E., GUSTAFSSON, O., NYEGAARD, M., TUULIO-HENRIKSSON, A., INGASON, A., HANSEN, T., SUVISAARI, J., LONNQVIST, J., PAUNIO, T., BORGLUM, A. D., HARTMANN, A., FINK-JENSEN, A., NORDENTOFT, M., HOUGAARD, D., NORGAARD-PEDERSEN, B., BOTTCHER, Y., OLESEN, J., BREUER, R., MOLLER, H. J., GIEGLING, I., RASMUSSEN, H. B., TIMM, S., MATTHEISEN, M., BITTER, I., RETHELYI, J. M., MAGNUSDOTTIR, B. B., SIGMUNDSSON, T., OLASON, P., MASSON, G., GULCHER, J. R., HARALDSSON, M., FOSSDAL, R., THORGEIRSSON, T. E., THORSTEINSDOTTIR, U., RUGGERI, M., TOSATO, S., FRANKE, B., STRENGMAN, E., KIEMENEY, L. A., 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, A. C., GE, D., YOON, J. L., SHIANNA, K. V.,

FREIMER, N. B., CANTOR, R. M., MURRAY, R., KONG, A., GOLIMBET, V., CARRACEDO, A., ARANGO, C., COSTAS, J., JONSSON, E. G., TERENIUS, L., AGARTZ, I., PETURSSON, H., NOTHEN, M. M., RIETSCHEL, M., MATTHEWS, P. M., MUGLIA, P., PELTONEN, L., ST CLAIR, D., GOLDSTEIN, D. B., STEFANSSON, K. & COLLIER, D. A. 2009. Common variants conferring risk of schizophrenia. *Nature,* 460**,** 744-7.

STEFANSSON, H., RUJESCU, D., CICHON, S., PIETILAINEN, O. P., INGASON, A., STEINBERG, S., FOSSDAL, R., SIGURDSSON, E., SIGMUNDSSON, T., BUIZER-VOSKAMP, J. E., HANSEN, T., JAKOBSEN, K. D., MUGLIA, P., FRANCKS, C., MATTHEWS, P. M., GYLFASON, A., HALLDORSSON, B. V., GUDBJARTSSON, D., THORGEIRSSON, T. E., SIGURDSSON, A., JONASDOTTIR, A., JONASDOTTIR, A., BJORNSSON, A., MATTIASDOTTIR, S., BLONDAL, T., HARALDSSON, M., MAGNUSDOTTIR, B. B., GIEGLING, I., MOLLER, H. J., HARTMANN, A., SHIANNA, K. V., GE, D., NEED, A. C., 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., MUHLEISEN, T. W., WANG, A. G., ULLUM, H., DJUROVIC, S., MELLE, I., OLESEN, J., KIEMENEY, L. A., FRANKE, B., SABATTI, C., FREIMER, N. B., GULCHER, J. R., THORSTEINSDOTTIR, U., KONG, A., ANDREASSEN, O. A., OPHOFF, R. A., GEORGI, A., RIETSCHEL, M., WERGE, T., PETURSSON, H., GOLDSTEIN, D. B., NOTHEN, M. M., PELTONEN, L., COLLIER, D. A., ST CLAIR, D. & STEFANSSON, K. 2008. Large recurrent microdeletions associated with schizophrenia. *Nature,* 455**,** 232-6.

STEINBERG, S., DE JONG, S., ANDREASSEN, O. A., WERGE, T., BORGLUM, A. D., MORS, O., MORTENSEN, P. B., GUSTAFSSON, O., COSTAS, J., PIETILAINEN, O. P., DEMONTIS, D., PAPIOL, S., HUTTENLOCHER, J., MATTHEISEN, M., BREUER, R., VASSOS, E., GIEGLING, I., FRASER, G., WALKER, N., TUULIO-HENRIKSSON, A., SUVISAARI, J., LONNQVIST, J., PAUNIO, T., AGARTZ, I., MELLE, I., DJUROVIC, S., STRENGMAN, E., JURGENS, G., GLENTHOJ, B., TERENIUS, L., HOUGAARD, D. M., ORNTOFT, T., WIUF, C., DIDRIKSEN, M., HOLLEGAARD, M. V., NORDENTOFT, M., VAN WINKEL, R., KENIS, G., ABRAMOVA, L., KALEDA, V., ARROJO, M., SANJUAN, J., ARANGO, C., SPERLING, S., ROSSNER, M., RIBOLSI, M., MAGNI, V., SIRACUSANO, A., CHRISTIANSEN, C., KIEMENEY, L. A., VELDINK, J., VAN DEN BERG, L., INGASON, A., MUGLIA, P., MURRAY, R., NOTHEN, M. M., SIGURDSSON, E., PETURSSON, H., THORSTEINSDOTTIR, U., KONG, A., RUBINO, I. A., DE HERT, M., RETHELYI, J. M., BITTER, I., JONSSON, E. G., GOLIMBET, V., CARRACEDO, A., EHRENREICH, H., CRADDOCK, N., OWEN, M. J., O'DONOVAN, M. C., RUGGERI, M., TOSATO, S., PELTONEN, L., OPHOFF, R. A., COLLIER, D. A., ST CLAIR, D., RIETSCHEL, M., CICHON, S., STEFANSSON, H., RUJESCU, D. & STEFANSSON, K. 2011. Common variants at VRK2 and TCF4 conferring risk of schizophrenia. *Hum Mol Genet,* 20**,** 4076-81.

STEINBERG, S., DE JONG, S., MATTHEISEN, M., COSTAS, J., DEMONTIS, D., JAMAIN, S., PIETILAINEN, O. P., LIN, K., PAPIOL, S., HUTTENLOCHER, J., SIGURDSSON, E., VASSOS, E., GIEGLING, I., BREUER, R., FRASER, G., WALKER, N., MELLE, I., DJUROVIC, S., AGARTZ, I., TUULIO-HENRIKSSON, A., SUVISAARI, J., LONNQVIST, J., PAUNIO, T., OLSEN, L., HANSEN, T., INGASON, A., PIRINEN, M.,

STRENGMAN, E., HOUGAARD, D. M., ORNTOFT, T., DIDRIKSEN, M., HOLLEGAARD, M. V., NORDENTOFT, M., ABRAMOVA, L., KALEDA, V., ARROJO, M., SANJUAN, J., ARANGO, C., ETAIN, B., BELLIVIER, F., MEARY, A., SCHURHOFF, F., SZOKE, A., RIBOLSI, M., MAGNI, V., SIRACUSANO, A., SPERLING, S., ROSSNER, M., CHRISTIANSEN, C., KIEMENEY, L. A., FRANKE, B., VAN DEN BERG, L. H., VELDINK, J., CURRAN, S., BOLTON, P., POOT, M., STAAL, W., REHNSTROM, K., KILPINEN, H., FREITAG, C. M., MEYER, J., MAGNUSSON, P., SAEMUNDSEN, E., MARTSENKOVSKY, I., BIKSHAIEVA, I., MARTSENKOVSKA, I., VASHCHENKO, O., RALEVA, M., PAKETCHIEVA, K., STEFANOVSKI, B., DURMISHI, N., PEJOVIC MILOVANCEVIC, M., LECIC TOSEVSKI, D., SILAGADZE, T., NANEISHVILI, N., MIKELADZE, N., SURGULADZE, S., VINCENT, J. B., FARMER, A., MITCHELL, P. B., WRIGHT, A., SCHOFIELD, P. R., FULLERTON, J. M., MONTGOMERY, G. W., MARTIN, N. G., RUBINO, I. A., VAN WINKEL, R., KENIS, G., DE HERT, M., RETHELYI, J. M., BITTER, I., TERENIUS, L., JONSSON, E. G., BAKKER, S., VAN OS, J., JABLENSKY, A., LEBOYER, M., BRAMON, E., POWELL, J., MURRAY, R., *et al.* 2012. Common variant at 16p11.2 conferring risk of psychosis. *Mol Psychiatry*. Advance online publication, doi:10.1038/mp.2012.157.

STILES, J. & JERNIGAN, T. L. 2010. The basics of brain development. *Neuropsychol Rev,* 20**,** 327-348.

STOBER, G., KOCHER, I., FRANZEK, E. & BECKMANN, H. 1997. First-trimester maternal gestational infection and cycloid psychosis. *Acta Psychiatr Scand,* 96**,** 319-24.

SUBRAMANIAN, A., TAMAYO, P., MOOTHA, V. K., MUKHERJEE, S., EBERT, B. L., GILLETTE, M. A., PAULOVICH, A., POMEROY, S. L., GOLUB, T. R., LANDER, E. S. & MESIROV, J. P. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A,* 102**,** 15545-50.

SUGAI, T., KAWAMURA, M., IRITANI, S., ARAKI, K., MAKIFUCHI, T., IMAI, C., NAKAMURA, R., KAKITA, A., TAKAHASHI, H. & NAWA, H. 2004. Prefrontal abnormality of schizophrenia revealed by DNA microarray: impact on glial and neurotrophic gene expression. *Ann N Y Acad Sci,* 1025**,** 84-91.

SULLIVAN, P. F. 2012. Puzzling over schizophrenia: schizophrenia as a pathway disease. *Nat Med,* 18**,** 210-1.

SULLIVAN, P. F., DALY, M. J. & O'DONOVAN, M. 2012. Genetic architectures of psychiatric disorders: the emerging picture and its implications. *Nat Rev Genet,* 13**,** 537-51.

SULLIVAN, P. F., KENDLER, K. S. & NEALE, M. C. 2003. Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. *Arch Gen Psychiatry,* 60**,** 1187-92.

SUN, X., WANG, J. F., TSENG, M. & YOUNG, L. T. 2006. Downregulation in components of the mitochondrial electron transport chain in the postmortem frontal cortex of subjects with bipolar disorder. *J Psychiatry Neurosci,* 31**,** 189-96.

SUSSER, E., NEUGEBAUER, R., HOEK, H. W., BROWN, A. S., LIN, S., LABOVITZ, D. & GORMAN, J. M. 1996. Schizophrenia after prenatal famine. Further evidence. *Arch Gen Psychiatry,* 53**,** 25-31.

SZULWACH, K. E., LI, X., SMRT, R. D., LI, Y., LUO, Y., LIN, L., SANTISTEVAN, N. J., LI, W., ZHAO, X. & JIN, P. 2010. Cross talk between microRNA and epigenetic regulation in adult neurogenesis. *J Cell Biol,* 189**,** 127-41.

TANDON, R., KESHAVAN, M. S. & NASRALLAH, H. A. 2008. Schizophrenia, "just the facts" what we know in 2008. 2. Epidemiology and etiology. *Schizophrenia Research,* 102**,** 1-18.

TAU, G. Z. & PETERSON, B. S. 2010. Normal development of brain circuits. *Neuropsychopharmacology,* 35**,** 147-68.

TEARE, D. M. & BARRETT, J. H. 2005. Genetic linkage studies. *Lancet,* 366**,** 1036-44.

TEER, J. K. & MULLIKIN, J. C. 2010. Exome sequencing: the sweet spot before whole genomes. *Hum Mol Genet,* 19**,** R145-51.

TIAN, C., GREGERSEN, P. K. & SELDIN, M. F. 2008a. Accounting for ancestry: population substructure and genome-wide association studies. *Hum Mol Genet,* 17**,** R143-50.

TIAN, C., PLENGE, R. M., RANSOM, M., LEE, A., VILLOSLADA, P., SELMI, C., KLARESKOG, L., PULVER, A. E., QI, L., GREGERSEN, P. K. & SELDIN, M. F. 2008b. Analysis and application of European genetic substructure using 300 K SNP information. *PLoS Genet,* 4**,** e4.

TKACHEV, D., MIMMACK, M. L., RYAN, M. M., WAYLAND, M., FREEMAN, T., JONES, P. B., STARKEY, M., WEBSTER, M. J., YOLKEN, R. H. & BAHN, S. 2003. Oligodendrocyte dysfunction in schizophrenia and bipolar disorder. *Lancet,* 362**,** 798-805.

TOBI, E. W., LUMEY, L. H., TALENS, R. P., KREMER, D., PUTTER, H., STEIN, A. D., SLAGBOOM, P. E. & HEIJMANS, B. T. 2009. DNA methylation differences after exposure to prenatal famine are common and timing- and sex-specific. *Hum Mol Genet,* 18**,** 4046-53.

TORKAMANI, A., DEAN, B., SCHORK, N. J. & THOMAS, E. A. 2010. Coexpression network analysis of neural tissue reveals perturbations in developmental processes in schizophrenia. *Genome Res,* 20**,** 403-12.

TRIXLER, M., TENYI, T., CSABI, G. & SZABO, R. 2001. Minor physical anomalies in schizophrenia and bipolar affective disorder. *Schizophrenia Research,* 52**,** 195-201.

VACIC, V., MCCARTHY, S., MALHOTRA, D., MURRAY, F., CHOU, H. H., PEOPLES, A., MAKAROV, V., YOON, S., BHANDARI, A., COROMINAS, R., IAKOUCHEVA, L. M., KRASTOSHEVSKY, O., KRAUSE, V., LARACH-WALTERS, V., WELSH, D. K., CRAIG, D., KELSOE, J. R., GERSHON, E. S., LEAL, S. M., DELL AQUILA, M., MORRIS, D. W., GILL, M., CORVIN, A., INSEL, P. A., MCCLELLAN, J., KING, M. C., KARAYIORGOU, M., LEVY, D. L., DELISI, L. E. & SEBAT, J. 2011. Duplications of the neuropeptide receptor gene VIPR2 confer significant risk for schizophrenia. *Nature,* 471**,** 499-503.

VAN DEN BOSSCHE, M. J., JOHNSTONE, M., STRAZISAR, M., PICKARD, B. S., GOOSSENS, D., LENAERTS, A. S., DE ZUTTER, S., NORDIN, A., NORRBACK, K. F., MENDLEWICZ, J., SOUERY, D., DE RIJK, P., SABBE, B. G., ADOLFSSON, R., BLACKWOOD, D. & DEL-FAVERO, J. 2012. Rare copy number variants in neuropsychiatric disorders: Specific phenotype or not? *Am J Med Genet B Neuropsychiatr Genet,* 159b**,** 812-22.

VAWTER, M. P., BARRETT, T., CHEADLE, C., SOKOLOV, B. P., WOOD, W. H., 3RD, DONOVAN, D. M., WEBSTER, M., FREED, W. J. & BECKER, K. G. 2001. Application of cDNA microarrays to examine gene expression differences in schizophrenia. *Brain Res Bull,* 55**,** 641-50.

VENTER, J. C., ADAMS, M. D., MYERS, E. W., LI, P. W., MURAL, R. J., SUTTON, G. G., SMITH, H. O., YANDELL, M., EVANS, C. A., HOLT, R. A., GOCAYNE, J. D., AMANATIDES, P., BALLEW, R. M., HUSON, D. H., WORTMAN, J. R., ZHANG, Q., KODIRA, C. D., ZHENG, X. H., CHEN, L., SKUPSKI, M., SUBRAMANIAN, G., THOMAS, P. D., ZHANG, J., GABOR MIKLOS, G. L., NELSON, C., BRODER, S., CLARK, A. G., NADEAU, J., MCKUSICK, V. A., ZINDER, N., LEVINE, A. J., ROBERTS, R. J., SIMON, M., SLAYMAN, C., HUNKAPILLER, M., BOLANOS, R., DELCHER, A., DEW, I., FASULO, D., FLANIGAN, M., FLOREA, L., HALPERN, A., HANNENHALLI, S., KRAVITZ, S., LEVY, S., MOBARRY, C., REINERT, K., REMINGTON, K., ABU-THREIDEH, J., BEASLEY, E., BIDDICK, K., BONAZZI, V., BRANDON, R., CARGILL, M., CHANDRAMOULISWARAN, I., CHARLAB, R., CHATURVEDI, K., DENG, Z., DI FRANCESCO, V., DUNN, P., EILBECK, K., EVANGELISTA, C., GABRIELIAN, A. E., GAN, W., GE, W., GONG, F., GU, Z., GUAN, P., HEIMAN, T. J., HIGGINS, M. E., JI, R. R., KE, Z., KETCHUM, K. A., LAI, Z., LEI, Y., LI, Z., LI, J., LIANG, Y., LIN, X., LU, F., MERKULOV, G. V., MILSHINA, N., MOORE, H. M., NAIK, A. K., NARAYAN, V. A., NEELAM, B., NUSSKERN, D., RUSCH, D. B., SALZBERG, S., SHAO, W., SHUE, B., SUN, J., WANG, Z., WANG, A., WANG, X., WANG, J., WEI, M., WIDES, R., XIAO, C., YAN, C., *et al.* 2001. The sequence of the human genome. *Science,* 291**,** 1304-51.

WADDINGTON, J. L., LANE, A., SCULLY, P. J., LARKIN, C. & O'CALLAGHAN, E. 1998. Neurodevelopmental and neuroprogressive processes in schizophrenia. Antithetical or complementary, over a lifetime trajectory of disease? *Psychiatr Clin North Am,* 21**,** 123-49.

WALSH, T., MCCLELLAN, J. M., MCCARTHY, S. E., ADDINGTON, A. M., PIERCE, S. B., COOPER, G. M., NORD, A. S., KUSENDA, M., MALHOTRA, D., BHANDARI, A., STRAY, S. M., RIPPEY, C. F., ROCCANOVA, P., MAKAROV, V., LAKSHMI, B., FINDLING, R. L., 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, E. E., MELTZER, P. S., NELSON, S. F., SINGLETON, A. B., LEE, M. K., RAPOPORT, J. L., KING, M.-C. & SEBAT, J. 2008. Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science,* 320**,** 539-543.

WANG, Z., GERSTEIN, M. & SNYDER, M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet,* 10**,** 57-63.

WEINBERG, S. M., JENKINS, E. A., MARAZITA, M. L. & MAHER, B. S. 2007. Minor physical anomalies in schizophrenia: a meta-analysis. *Schizophrenia Research,* 89**,** 72-85.

WEINBERGER, D. R. 1987. Implications of normal brain development for the pathogenesis of schizophrenia. *Arch Gen Psychiatry,* 44**,** 660-9.

WEISS, L. A., SHEN, Y., KORN, J. M., ARKING, D. E., MILLER, D. T., FOSSDAL, R., SAEMUNDSEN, E., STEFANSSON, H., FERREIRA, M. A., GREEN, T., PLATT, O. S., RUDERFER, D. M., WALSH, C. A., ALTSHULER, D., CHAKRAVARTI, A., TANZI, R. E., STEFANSSON, K., SANTANGELO, S. L., GUSELLA, J. F., SKLAR, P., WU, B. L. &

DALY, M. J. 2008. Association between microdeletion and microduplication at 16p11.2 and autism. *N Engl J Med,* 358**,** 667-75.

WILLIAMS, H. J., CRADDOCK, N., RUSSO, G., HAMSHERE, M. L., MOSKVINA, V., DWYER, S., SMITH, R. L., GREEN, E., GROZEVA, D., HOLMANS, P., OWEN, M. J. & O'DONOVAN, M. C. 2011a. Most genome-wide significant susceptibility loci for schizophrenia and bipolar disorder reported to date cross-traditional diagnostic boundaries. *Hum Mol Genet,* 20**,** 387-91.

WILLIAMS, H. J., NORTON, N., DWYER, S., MOSKVINA, V., NIKOLOV, I., CARROLL, L., GEORGIEVA, L., WILLIAMS, N. M., MORRIS, D. W., QUINN, E. M., GIEGLING, I., IKEDA, M., WOOD, J., LENCZ, T., HULTMAN, C., LICHTENSTEIN, P., THISELTON, D., MAHER, B. S., MALHOTRA, A. K., RILEY, B., KENDLER, K. S., GILL, M., SULLIVAN, P., SKLAR, P., PURCELL, S., NIMGAONKAR, V. L., KIROV, G., HOLMANS, P., CORVIN, A., RUJESCU, D., CRADDOCK, N., OWEN, M. J. & O'DONOVAN, M. C. 2011b. Fine mapping of ZNF804A and genome-wide significant evidence for its involvement in schizophrenia and bipolar disorder. *Mol Psychiatry,* 16**,** 429-41.

WILLIAMS, N. M., ZAHARIEVA, I., MARTIN, A., LANGLEY, K., MANTRIPRAGADA, K., FOSSDAL, R., STEFANSSON, H., STEFANSSON, K., MAGNUSSON, P., GUDMUNDSSON, O. O., GUSTAFSSON, O., HOLMANS, P., OWEN, M. J., O'DONOVAN, M. & THAPAR, A. 2010. Rare chromosomal deletions and duplications in attention-deficit hyperactivity disorder: a genome-wide analysis. *Lancet,* 376**,** 1401-8.

WILLSEY, A. J., SANDERS, S. J., LI, M., DONG, S., TEBBENKAMP, A. T., MUHLE, R. A., REILLY, S. K., LIN, L., FERTUZINHOS, S., MILLER, J. A., MURTHA, M. T., BICHSEL, C., NIU, W., COTNEY, J., ERCAN-SENCICEK, A. G., GOCKLEY, J., GUPTA, A. R., HAN, W., HE, X., HOFFMAN, E. J., KLEI, L., LEI, J., LIU, W., LIU, L., LU, C., XU, X., ZHU, Y., MANE, S. M., LEIN, E. S., WEI, L., NOONAN, J. P., ROEDER, K., DEVLIN, B., SESTAN, N. & STATE, M. W. 2013. Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. *Cell,* 155**,** 997-1007.

WORLD HEALTH ORGANIZATION 1993. *The ICD-10 classification of mental and behavioural disorders: diagnostic criteria for research*. World Health Organization.

WRAY, N. R. & VISSCHER, P. M. 2010. Narrowing the boundaries of the genetic architecture of schizophrenia. *Schizophr Bull,* 36**,** 14-23.

WU, J. Q., WANG, X., BEVERIDGE, N. J., TOONEY, P. A., SCOTT, R. J., CARR, V. J. & CAIRNS, M. J. 2012. Transcriptome sequencing revealed significant alteration of cortical promoter usage and splicing in schizophrenia. *PLoS One,* 7**,** e36351.

XU, B., IONITA-LAZA, I., ROOS, J. L., BOONE, B., WOODRICK, S., SUN, Y., LEVY, S., GOGOS, J. A. & KARAYIORGOU, M. 2012. De novo gene mutations highlight patterns of genetic and neural complexity in schizophrenia. *Nat Genet,* 44**,** 1365-9.

XU, B., ROOS, J. L., LEVY, S., VAN RENSBURG, E. J., GOGOS, J. A. & KARAYIORGOU, M. 2008. Strong association of de novo copy number mutations with sporadic schizophrenia. *Nat Genet,* 40**,** 880-5.

XU, T., CHAN, R. C. & COMPTON, M. T. 2011. Minor physical anomalies in patients with schizophrenia, unaffected first-degree relatives, and healthy controls: a meta-analysis. *PLoS One,* 6**,** e24129.

YANG, J., LEE, S. H., GODDARD, M. E. & VISSCHER, P. M. 2011. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet,* 88**,** 76-82.

YEO, G., HOLSTE, D., KREIMAN, G. & BURGE, C. B. 2004. Variation in alternative splicing across human tissues. *Genome Biol,* 5, R74.

ZAMMIT, S., ALLEBECK, P., DAVID, A. S., DALMAN, C., HEMMINGSSON, T., LUNDBERG, I. & LEWIS, G. 2004. A longitudinal study of premorbid IQ Score and risk of developing schizophrenia, bipolar disorder, severe depression, and other nonaffective psychoses. *Arch Gen Psychiatry,* 61**,** 354-60.

ZAYKIN, D. V., ZHIVOTOVSKY, L. A., WESTFALL, P. H. & WEIR, B. S. 2002. Truncated product method for combining P-values. *Genet Epidemiol,* 22**,** 170-85.

ZHANG, D., CHENG, L., QIAN, Y., ALLIEY-RODRIGUEZ, N., KELSOE, J. R., GREENWOOD, T., NIEVERGELT, C., BARRETT, T. B., MCKINNEY, R., SCHORK, N., SMITH, E. N., BLOSS, C., NURNBERGER, J., EDENBERG, H. J., FOROUD, T., SHEFTNER, W., LAWSON, W. B., NWULIA, E. A., HIPOLITO, M., CORYELL, W., RICE, J., BYERLEY, W., MCMAHON, F., SCHULZE, T. G., BERRETTINI, W., POTASH, J. B., BELMONTE, P. L., ZANDI, P. P., MCINNIS, M. G., ZOLLNER, S., CRAIG, D., SZELINGER, S., KOLLER, D., CHRISTIAN, S. L., LIU, C. & GERSHON, E. S. 2009. Singleton deletions throughout the genome increase risk of bipolar disorder. *Mol Psychiatry,* 14**,** 376-80.

ZHAO, C., XU, Z., WANG, F., CHEN, J., NG, S. K., WONG, P. W., YU, Z., PUN, F. W., REN, L., LO, W. S., TSANG, S. Y. & XUE, H. 2009. Alternative-splicing in the exon-10 region of GABA(A) receptor beta(2) subunit gene: relationships between novel isoforms and psychotic disorders. *PLoS One,* 4**,** e6977.

ZHAO, Z., WEBB, B. T., JIA, P., BIGDELI, T. B., MAHER, B. S., VAN DEN OORD, E., BERGEN, S. E., AMDUR, R. L., O'NEILL, F. A., WALSH, D., THISELTON, D. L., CHEN, X., PATO, C. N., RILEY, B. P., KENDLER, K. S. & FANOUS, A. H. 2013. Association study of 167 candidate genes for schizophrenia selected by a multi-domain evidence-based prioritization algorithm and neurodevelopmental hypothesis. *PLoS One,* 8**,** e67776.

ZISOOK, S., MCADAMS, L. A., KUCK, J., HARRIS, M. J., BAILEY, A., PATTERSON, T. L., JUDD, L. L. & JESTE, D. V. 1999. Depressive symptoms in schizophrenia. *Am J Psychiatry,* 156**,** 1736-43.

ZORNBERG, G. L., BUKA, S. L. & TSUANG, M. T. 2000. Hypoxic-ischemia-related fetal/neonatal complications and risk of schizophrenia and other nonaffective psychoses: a 19-year longitudinal study. *Am J Psychiatry,* 157**,** 196-202.

Chapter 7: Appendix A

7.1 Additional tables for Chapter 2

		Schizophrenia		Bipolar disorder	
		Brown's	Simes'	Brown's	Simes'
DLPFC	P value	0.00161 (0.0209)	0.0194 (0.252)	0.439	0.304
	Correlation Coeff.	-0.0275	-0.0211	-0.00662	-0.0093
		-	-	-	-
MPFC	P value	2.02×10^{-6} (2.63×10^{-5})	3.52×10^{-5} (0.0005)	0.00295 (0.0384)	0.0341 (0.443)
	Correlation Coeff.	0.0414	0.0374	0.0254	0.0192
		+	+	+	+
OPFC	P value	0.155	0.918	0.626	0.851
	Correlation Coeff.	0.0124	0.000928	-0.00417	-0.00170
		+	+	-	-
VLPFC	P value	0.0110 (0.144)	0.108	0.725	0.303
	Correlation Coeff.	0.0221	0.0145	-0.00300	0.00931
		+	+	-	+
MS	P value	0.00411 (0.0535)	3.71×10^{-6} (4.83×10^{-5})	2.84×10^{-8} (3.69×10^{-7})	6.06×10^{-5} (0.000788)
	Correlation Coeff.	0.0250	0.0418	0.0474	0.0363
		+	+	+	+
PAS	P value	0.0719 (0.935)	0.0103 (0.133)	0.219	0.376
	Correlation Coeff.	0.0157	0.0232	0.0105	0.00802
		+	+	+	+
TAU	P value	0.305	0.500	0.879	0.676
	Correlation Coeff.	0.00894	0.00610	0.00113	0.00378
		+	+	+	+
TAS	P value	0.161	0.381	0.895	0.381
	Correlation Coeff.	0.0122	0.00792	0.00113	-0.00792
		+	+	+	-
OCC	P value	0.120	0.153	0.0165 (0.214)	0.997
	Correlation Coeff.	0.0135	0.00129	0.0205	-3.65×10^{-5}
		+	+	+	-
HIP	P value	1.35×10^{-7} (1.76×10^{-6})	1.78×10^{-6} (2.32×10^{-5})	0.00143 (0.0186)	0.0283 (0.368)
	Correlation Coeff.	-0.0459	-0.0432	-0.0272	-0.0198
		-	-	-	-
STR	P value	0.679	0.689	0.958	0.823
	Correlation Coeff.	0.00360	0.00362	0.0187	-0.00202
		+	+	+	-

THAL	**P value**	0.00137 (0.0178)	2.37×10^{-5} (0.000308)	4.10×10^{-5} (0.000534)	0.00668 (0.0868)
	Correlation Coeff.	-0.0279	-0.0382	-0.0350	-0.0245
		-	-	-	-
CBL	**P value**	0.00865 (0.112)	0.0179 (0.233)	0.204	0.557
	Correlation Coeff.	-0.0229	-0.0214	-0.0109	-0.00532
		-	-	-	-

Table 7.1: Linear regression results and correlation coefficients testing regional characteristic scores calculated in the Johnson dataset with gene-wide logP.
P values in brackets have been corrected for 13 brain regions using Bonferroni's method.

		Schizophrenia		Bipolar disorder	
		Brown's	Simes'	Brown's	Simes'
DLPFC	P value	0.449	0.206	0.353	0.602
	Correlation Coeff.	0.00657	0.0114	0.00412	0.00469
		+	+	+	+
MPFC	P value	0.997	0.324	0.00195 (0.0312)	0.0203 (0.324)
	Correlation Coeff.	-3.17×10^{-5}	0.00888	0.0133	0.0209
		-	+	+	+
OPFC	P value	0.0147 (0.235)	0.0553 (0.884)	0.0155 (0.248)	0.0740
	Correlation Coeff.	0.0212	0.0173	0.0100	0.0161
		+	+	+	+
VLPFC	P value	0.390	0.624	0.632	0.351
	Correlation Coeff.	0.00746	0.00442	0.00251	0.00839
		+	+	+	+
M1C	P value	0.714	0.601	0.0327 (0.523)	0.175
	Correlation Coeff.	-0.00318	-0.00471	0.00983	0.0122
		-	-	+	+
S1C	P value	0.332	0.727	0.111	0.374
	Correlation Coeff.	0.00842	0.00315	0.00880	0.00799
		+	+	+	+
TAU	P value	0.119	0.857	0.208	0.687
	Correlation Coeff.	-0.0135	0.00162	-0.00596	-0.00363
		-	+	-	-
OCC	P value	3.57×10^{-8} (5.70×10^{-7})	5.04×10^{-8} (8.06×10^{-7})	7.34×10^{-5} (0.00118)	0.0105 (0.168)
	Correlation Coeff.	0.0478	0.0490	0.0102	0.0230
		+	+	+	+
IPC	P value	0.000144 (0.00230)	4.98×10^{-5} (0.000797)	8.65×10^{-6} (0.000138)	0.00252 (0.0403)
	Correlation Coeff.	0.0330	0.0365	0.0186	0.0272
		+	+	+	+
TAS	P value	0.00207 (0.0331)	0.00128 (0.0205)	0.0388 (0.620)	0.171
	Correlation Coeff.	0.0267	0.0290	0.008110	0.0123
		+	+	+	+
ITC	P value	0.0436 (0.698)	0.0220 (0.353)	0.0903	0.0494 (0.791)
	Correlation Coeff.	0.0175	0.0206	-0.00913	-0.0177
		+	+	-	-
HIP	P value	8.28×10^{-10} (1.32×10^{-8})	8.78×10^{-8} (1.40×10^{-6})	3.35×10^{-6} (5.36×10^{-5})	9.23×10^{-6} (0.000148)
	Correlation Coeff.	-0.0532	-0.0481	-0.00743	-0.0399
		-	-	-	-
AMY	P value	0.319	0.249	0.208	0.182
	Correlation Coeff.	0.00864	0.0104	0.00256	0.0120
		+	+	+	+

STR	**P value**	0.141	0.756	0.858	0.852
	Correlation Coeff.	0.0128	0.00280	-0.000188	0.00168
		+	+	-	+
THAL	**P value**	2.66×10^{-5} (0.000425)	6.69×10^{-7} (1.07×10^{-5})	4.55×10^{-6} (7.28×10^{-5})	0.00293 (0.0469)
	Correlation Coeff.	-0.0364	-0.0447	-0.00390	-0.0268
		-	-	-	-
CBL	**P value**	0.00345 (0.0552)	0.0383 (0.612)	0.0906	0.462
	Correlation Coeff.	-0.0254	-0.0187	-0.00119	-0.00663
		-	-	-	+

Table 7.2: Linear regression results and correlation coefficients testing regional characteristic scores calculated in the Kang dataset with gene-wide logP.
P values in brackets have been corrected for 16 brain regions using Bonferroni's method.

		All CNVs			Deletions			Duplications		
		Min.	Med.	Max.	Min.	Med.	Max.	Min.	Med.	Max.
DLPFC	P value	0.99	0.780	0.519	0.166	0.285	0.476	0.419	0.543	0.883
	Coeff.	+	-	-	-	-	-	+	+	+
MPFC	P value	0.569	1	0.802	0.274	0.547	0.808	0.742	0.342	0.484
	Coeff.	+	+	+	-	-	+	+	+	+
OPFC	P value	0.477	0.851	0.244	0.206	0.171	0.212	0.635	0.170	0.0501 (0.651)
	Coeff.	+	-	-	+	+	+	-	-	-
VLPFC	P value	0.547	0.264	0.268	0.144	0.163	0.280	0.914	0.669	0.808
	Coeff.	-	-	-	-	-	-	-	-	-
MS	P value	0.917	0.680	0.462	0.0980	0.124	0.136	0.0650 (0.845)	0.163	0.314
	Coeff.	+	-	-	-	-	-	+	+	+
PAS	P value	0.174	0.313	0.704	0.539	0.543	0.595	0.197	0.241	0.533
	Coeff.	+	+	+	+	+	+	+	+	+
TAU	P value	0.547	0.757	1	0.758	0.493	0.422	0.672	0.974	0.859
	Coeff.	+	+	+	+	+	+	+	+	-
TAS	P value	0.100	0.162	0.241	0.168	0.115	0.067 (0.877)	0.471	0.656	0.776
	Coeff.	+	+	+	+	+	+	+	+	+
OCC	P value	0.750	0.365	0.246	0.792	0.804	0.933	0.866	0.448	0.294
	Coeff.	-	-	-	-	-	-	-	-	-
HIP	P value	0.932	0.960	0.728	0.232	0.475	0.718	0.541	0.668	0.405
	Coeff.	-	-	+	-	-	-	+	+	+

STR	**P value**	0.425	0.420	0.428	0.945	0.556	0.288	0.208	0.100	0.078
	Coeff.	-	-	-	+	+	+	-	-	-
THAL	**P value**	0.468	0.555	0.267	0.349	0.319	0.167	0.767	0.699	0.889
	Coeff.	+	+	+	+	+	+	-	-	+
CBL	**P value**	0.988	0.472	0.292	0.851	0.708	0.365	0.381	0.875	0.686
	Coeff.	+	+	+	+	+	+	-	+	+

Table 7.3: Logistic regression results testing CNV case control status on regional characteristic scores calculated in the Johnson dataset.
All p values in brackets were adjusted for 13 brain regions using Bonferroni's method, where missing corrected p value was 1.

		All CNVs			Deletions			Duplications		
		Min.	**Med.**	**Max.**	**Min.**	**Med.**	**Max.**	**Min.**	**Med.**	**Max.**
DLPFC	**P value**	0.306	0.842	0.827	0.654	0.847	0.800	0.319	0.639	0.607
	Coeff.	-	-	-	-	+	+	-	-	-
MPFC	**P value**	0.0285 (0.371)	0.00983 (0.128)	0.0174 (0.226)	0.595	0.797	0.733	0.000411 (0.00535)	0.000117 (0.00152)	0.000835 (0.0109)
	Coeff.	-	-	-	+	+	+	-	-	-
OPFC	**P value**	0.932	0.565	0.263	0.568	0.335	0.199	0.710	0.943	0.724
	Coeff.	+	+	+	+	+	+	-	-	+
VLPFC	**P value**	0.00278 (0.0362)	0.00434 (0.0564)	0.0117 (0.152)	0.180	0.265	0.463	0.00653	0.00518	0.00683
	Coeff.	-	-	-	-	-	-	-	-	-
MS	**P value**	0.648	0.270	0.355	0.00702 (0.0912)	0.00512 (0.0665)	0.00770 (0.100)	0.0860	0.326	0.346
	Coeff.	+	+	+	+	+	+	-	-	-
PAS	**P value**	0.0275 (0.358)	0.0307 (0.399)	0.0210 (0.273)	0.0109 (0.142)	0.0189 (0.246)	0.0253 (0.329)	0.463	0.412	0.237
	Coeff.	+	+	+	+	+	+	+	+	+
TAU	**P value**	0.711	0.205	0.0322	0.0627 (0.815)	0.0217 (0.282)	0.00703 (0.0914)	0.261	0.667	0.611
	Coeff.	+	+	+	+	+	+	-	-	+
TAS	**P value**	0.361	0.992	0.540	0.319	0.506	0.342	0.771	0.468	0.931
	Coeff.	-	+	-	-	-	-	-	+	+
OCC	**P value**	0.00519 (0.0675)	0.0172 (0.224)	0.0547 (0.711)	0.133	0.244	0.329	0.013 (0.133)	0.0159 (0.207)	0.0656 (0.853)
	Coeff.	+	+	+	+	+	+	+	+	+
HIP	**P value**	0.531	0.302	0.270	0.701	0.883	0.719	0.609	0.168	0.0583 (0.758)
	Coeff.	+	+	+	+	+	-	+	+	+

STR	**P value**	0.0748 (0.973)	0.0676 (0.878)	0.233	0.637	0.734	0.946	0.0514 (0.668)	0.0261 (0.339)	0.112
	Coeff.	+	+	+	+	+	+	+	+	+
THAL	**P value**	0.000284 (0.00369)	0.00878 (0.114)	0.0186 (0.242)	0.0385 (0.501)	0.0520 (0.676)	0.0174 (0.226)	0.00215 (0.0280)	0.0633 (0.823)	0.263
	Coeff.	-	-	-	-	-	-	-	-	-
CBL	**P value**	0.0222	0.108	0.246	0.00355 (0.0461)	0.00554 (0.0720)	0.0191 (0.249)	0.760	0.535	0.503
	Coeff.	-	-	-	-	-	-	-	+	+

Table 7.4: Logistic regression results testing CNV singleton status on regional characteristic scores calculated in the Johnson dataset.
All p values in brackets were adjusted for 13 brain regions using Bonferroni's method, where missing corrected p value was 1.

		All CNVs			Deletions			Duplications		
		Min.	Med.	Max.	Min.	Med.	Max.	Min.	Med.	Max.
DLPFC	P value	0.298	0.178	0.0882	0.0453 (0.725)	0.0338 (0.541)	0.0827	0.990	0.964	0.619
	Coeff.	-	-	-	-	-	-	+	+	-
MPFC	P value	0.770	0.484	0.356	0.0308 (0.493)	0.0388 (0.620)	0.0547 (0.875)	0.111	0.243	0.384
	Coeff.	-	-	-	-	-	-	+	+	+
OPFC	P value	0.942	0.730	0.599	0.434	0.306	0.508	0.732	0.596	0.854
	Coeff.	-	-	-	-	-	-	+	+	+
VLPFC	P value	0.730	0.392	0.208	0.0739	0.0477 (0.763)	0.0737	0.395	0.475	0.842
	Coeff.	-	-	-	-	-	-	+	+	+
M1C	P value	0.635	0.521	0.761	0.0259 (0.415)	0.0255 (0.409)	0.187	0.190	0.146	0.228
	Coeff.	-	-	-	-	-	-	+	+	+
S1C	P value	0.636	0.843	0.798	0.475	0.530	0.655	0.236	0.283	0.522
	Coeff.	+	+	-	-	-	-	+	+	+
TAU	P value	0.316	0.207	0.425	0.649	0.621	0.419	0.188	0.197	0.409
	Coeff.	+	+	+	-	+	+	+	+	+
OCC	P value	0.171	0.112	0.0819	0.413	0.710	0.838	0.265	0.127	0.0532 (0.851)
	Coeff.	-	-	-	-	-	+	-	-	-
IPC	P value	0.0848	0.215	0.149	0.920	0.756	0.731	0.0372 (0.595)	0.0480 (0.768)	0.0576 (0.921)
	Coeff.	+	+	+	-	-	+	+	+	+
TAS	P value	0.290	0.187	0.149	0.676	0.397	0.174	0.417	0.233	0.230
	Coeff.	+	+	+	+	+	+	+	+	+

ITC	**P value**	0.326	0.332	0.381	0.0456 (0.729)	0.00722 (0.116)	0.00614 (0.0983)	0.430	0.200	0.338
	Coeff.	+	+	+	+	+	+	-	-	-
HIP	**P value**	0.455	0.646	0.890	0.324	0.357	0.486	0.143	0.178	0.357
	Coeff.	+	+	+	-	-	-	+	+	+
AMY	**P value**	0.433	0.352	0.724	0.381	0.140	0.162	0.872	0.893	0.876
	Coeff.	+	+	+	+	+	+	+	+	+
STR	**P value**	0.873	0.381	0.312	0.665	0.829	0.450	0.411	0.162	0.0956
	Coeff.	-	-	-	+	+	+	-	-	-
THAL	**P value**	0.557	0.909	0.926	0.900	0.536	0.317	0.312	0.402	0.535
	Coeff.	-	-	+	-	+	+	-	-	-
CBL	**P value**	0.916	0.858	0.939	0.673	0.795	0.697	0.286	0.586	0.638
	Coeff.	+	+	+	+	+	+	-	-	+

Table 7.5: Logistic regression results testing CNV case control status on regional characteristic scores calculated in the Kang dataset.
All p values in brackets were adjusted for 16 brain regions using Bonferroni's method, where missing corrected p value was 1.

		All CNVs			Deletions			Duplications		
		Min.	Med.	Max.	Min.	Med.	Max.	Min.	Med.	Max.
DLPFC	P value	0.249	0.203	0.148	0.910	0.869	0.977	0.141	0.0844	0.0767
	Coeff.	-	-	-	+	+	+	-	-	-
MPFC	P value	0.190	0.842	0.890	0.0123	0.0654	0.0872	0.518	0.132	0.178
	Coeff.	+	+	+	+	+	+	-	-	-
OPFC	P value	0.151	0.0836	0.0560 (0.895)	0.883	0.846	0.816	0.0877	0.0353 (0.565)	0.0289 (0.463)
	Coeff.	-	-	-	-	-	-	-	-	-
VLPFC	P value	0.131	0.102	0.0139 (0.222)	0.858	0.986	0.446	0.0433 (0.692)	0.0395 (0.632)	0.0139 (0.223)
	Coeff.	-	-	-	+	-	-	-	-	-
M1C	P value	0.575	0.760	0.495	0.391	0.373	0.172	0.995	0.707	0.859
	Coeff.	+	+	+	+	+	+	+	-	-
S1C	P value	0.127	0.264	0.549	0.569	0.663	0.815	0.120	0.257	0.513
	Coeff.	-	-	-	-	-	-	-	-	-
TAU	P value	0.974	0.986	0.647	0.444	0.570	0.750	0.498	0.640	0.442
	Coeff.	+	-	+	-	-	-	+	+	+
OCC	P value	0.0322	0.159	0.443	0.0297	0.0275	0.0659	0.366	0.799	0.478
	Coeff.	+	+	+	+	+	+	+	-	-
IPC	P value	0.743	0.201	0.0646	0.802	0.762	0.984	0.807	0.146	0.0132 (0.211)
	Coeff.	+	+	+	+	+	+	+	+	+
TAS	P value	0.202	0.237	0.0302	0.596	0.651	0.629	0.200	0.227	0.0126
	Coeff.	+	+	+	+	+	+	+	+	+
ITC	P value	0.441	0.377	0.776	0.832	0.763	0.793	0.198	0.0987	0.523
	Coeff.	+	+	+	-	-	-	+	+	+

HIP	**P value**	0.237	0.220	0.187	0.148	0.309	0.524	0.689	0.414	0.204
	Coeff.	+	+	+	+	+	+	+	+	+
AMY	**P value**	0.660	0.708	0.281	0.830	0.951	0.809	0.442	0.620	0.259
	Coeff.	-	-	-	+	+	+	-	-	-
STR	**P value**	0.213	0.143	0.318	0.698	0.777	0.917	0.192	0.0914	0.174
	Coeff.	+	+	+	+	+	-	+	+	+
THAL	**P value**	0.147	0.663	0.717	0.526	0.604	0.883	0.0183	0.321	0.578
	Coeff.	-	-	-	+	+	+	-	-	-
CBL	**P value**	0.0311 (0.497)	0.0706	0.0675	0.00401	0.00414	0.0104	0.956	0.622	0.945
	Coeff.	-	-	-	-	-	-	-	+	-

Table 7.6: Logistic regression results testing CNV singleton status on regional characteristic scores calculated in the Kang dataset.
All p values in brackets were adjusted for 16 brain regions using Bonferroni's method, where missing corrected p value was 1.

		Schizophrenia		Bipolar disorder	
		Brown's	**Simes'**	**Brown's**	**Simes'**
DLPFC	**P value**	0.00988 (0.128)	0.712	0.987	0.270
	Correlation Coeff.	0.0226	0.00336	0.000145	0.0100
		+	+	+	+
MPFC	**P value**	0.326	0.399	0.233	0.0325 (0.422)
	Correlation Coeff.	0.00861	0.00767	0.0102	0.0195
		+	+	+	+
OPFC	**P value**	0.184	0.0141 (0.183)	0.578	0.902
	Correlation Coeff.	0.0116	0.0223	-0.00477	0.00112
		+	+	-	+
VLPFC	**P value**	0.0444 (0.5777)	0.0213 (0.277)	0.354	0.499
	Correlation Coeff.	0.0176	0.0209	0.00797	0.00615
		+	+	-	+
MS	**P value**	0.0270 (0.350)	0.0886	0.160	0.109
	Correlation Coeff.	0.0194	0.0155	0.0121	0.0146
		+	+	+	+
PAS	**P value**	0.810	0.532	0.919	0.484
	Correlation Coeff.	0.00210	0.00568	0.000879	0.00637
		+	+	+	+
TAU	**P value**	0.370	0.375	0.854	0.563
	Correlation Coeff.	-0.00786	-0.00806	0.00158	0.00527
		-	-	+	+
TAS	**P value**	0.00417 (0.0542)	0.0199 (0.259)	0.613	0.267
	Correlation Coeff.	0.0251	0.0212	0.00435	0.0101
		+	+	+	+
OCC	**P value**	0.360	0.232	0.216	0.507
	Correlation Coeff.	0.00801	0.0109	0.0116	0.00604
		+	+	+	+
HIP	**P value**	0.0443 (0.575)	0.516	0.738	0.792
	Correlation Coeff.	0.0176	0.00591	0.00287	-0.00240
		+	+	+	-
STR	**P value**	0.415	0.0269 (0.349)	0.560	0.237
	Correlation Coeff.	0.00715	0.0201	0.00501	0.0108
		+	+	+	+

THAL	**P value**	0.839	0.116	0.658	0.344
	Correlation Coeff.	0.00178	0.0143	-0.00380	0.00862
		+	+	-	+
CBL	**P value**	0.00651 (0.0846)	0.0601 (0.781)	0.124	0.00476 (0.0619)
	Correlation Coeff.	-0.0238	-0.0171	0.0132	0.0257
		-	-	+	+

Table 7.7: Linear regression results and correlation coefficients testing regional splicing logP calculated in the Johnson dataset with gene-wide logP.
P values in brackets have been corrected for 13 brain regions using Bonferroni's method.

		Schizophrenia		Bipolar disorder	
		Brown's	Simes'	Brown's	Simes'
DLPFC	P value	0.258	0.116	0.612	0.277
	Correlation Coeff.	-0.00985	-0.0142	0.00433	0.00981
		-	-	+	+
MPFC	P value	0.0139 (0.223)	0.186	0.759	0.816
	Correlation Coeff.	-0.0214	-0.0119	-0.00262	0.00210
		-	-	-	+
OPFC	P value	0.0718	0.643	0.117	0.507
	Correlation Coeff.	-0.0157	-0.00419	-0.0134	-0.00599
		-	-	-	-
VLPFC	P value	0.762	0.434	0.563	0.284
	Correlation Coeff.	-0.00264	-0.00707	0.00494	0.00967
		-	-	-	+
M1C	P value	0.498	0.333	0.527	0.608
	Correlation Coeff.	0.00590	-0.00874	0.00540	0.00463
		+	-	+	+
S1C	P value	0.338	0.552	0.780	0.718
	Correlation Coeff.	-0.00834	-0.00537	-0.00238	-0.00326
		-	-	+	-
TAU	P value	0.703	0.823	0.661	0.109
	Correlation Coeff.	-0.00332	0.00202	-0.00375	-0.0145
		-	+	-	-
OCC	P value	0.00318 (0.0508)	0.0487 (0.780)	0.891	0.687
	Correlation Coeff.	0.0257	0.01789	-0.00117	-0.00364
		+	+	-	-
IPC	P value	0.00150 (0.0240)	0.00132 (0.0211)	0.556	0.504
	Correlation Coeff.	-0.0276	-0.0290	-0.00502	-0.00603
		-	-	-	-
TAS	P value	0.0652	0.702	0.778	0.526
	Correlation Coeff.	0.0160	0.00346	-0.00241	0.00572
		+	+	-	+
ITC	P value	0.152	0.222	0.884	0.951
	Correlation Coeff.	-0.0124	-0.0110	0.00125	0.000550
		-	-	+	+
HIP	P value	0.00324 (0.0518)	0.0248 (0.396)	0.482	0.319
	Correlation Coeff.	0.0256	0.0203	0.00600	0.00899
		+	+	+	+
AMY	P value	0.291	0.861	0.198	0.866
	Correlation Coeff.	-0.00919	0.00158	-0.0110	-0.00153
		-	+	-	-
STR	P value	0.756	0.665	0.583	0.546
	Correlation Coeff.	-0.00270	0.00392	-0.00469	0.00546
		-	+	-	+

THAL	**P value**	0.0275 (0.441)	0.0893	0.016 (0.254)	0.0347 (0.556)
	Correlation Coeff.	0.0192	0.0153	0.0206	0.0191
		+	+	+	+
CBL	**P value**	0.0502 (0.803)	0.493	0.878	0.114
	Correlation Coeff.	-0.0170	-0.00619	0.00131	0.0143
		-	-	+	+

Table 7.8: Linear regression results and correlation coefficients testing regional splicing logP calculated in the Kang dataset with gene-wide logP.
P values in brackets have been corrected for 16 brain regions using Bonferroni's method.

		All CNVs			Deletions			Duplications		
		Min.	Med.	Max.	Min.	Med.	Max.	Min.	Med.	Max.
DLPFC	P value	0.465	0.609	0.886	0.130	0.163	0.607	0.936	0.681	0.647
	Coeff.	-	-	-	-	-	-	+	+	+
MPFC	P value	0.177	0.261	0.995	0.983	0.655	0.305	0.0440 (0.572)	0.0613 (0.797)	0.718
	Coeff.	-	-	+	-	+	+	-	-	-
OPFC	P value	0.241	0.220	0.0789	0.177	0.115	0.0122 (0.159)	0.643	0.757	0.891
	Coeff.	-	-	-	-	-	-	-	-	-
VLPFC	P value	0.0434	0.113	0.698	0.156	0.209	0.352	0.135	0.331	0.261
	Coeff.	-	-	-	-	-	+	-	-	-
MS	P value	0.894	0.959	0.973	0.704	0.664	0.787	0.465	0.393	0.961
	Coeff.	+	-	+	+	+	+	-	-	+
PAS	P value	0.436	0.424	0.651	0.272	0.508	0.987	0.678	0.574	0.817
	Coeff.	-	-	-	-	-	-	-	-	-
TAU	P value	0.628	0.640	0.459	0.701	0.878	0.815	0.535	0.555	0.134
	Coeff.	+	+	+	-	-	-	+	+	+
TAS	P value	0.424	0.283	0.495	0.428	0.421	0.876	0.700	0.680	0.829
	Coeff.	-	-	-	-	-	+	-	-	-
OCC	P value	0.798	0.474	0.448	0.671	0.591	0.930	0.789	0.507	0.441
	Coeff.	-	-	-	-	-	-	-	-	-
HIP	P value	0.937	0.926	0.389	0.274	0.684	0.620	0.457	0.858	0.312
	Coeff.	+	-	+	-	-	+	+	+	+
STR	P value	0.991	0.821	0.812	0.761	0.843	0.917	0.956	0.834	0.987
	Coeff.	+	-	-	+	+	-	-	-	+
THAL	P value	0.318	0.388	0.0790	0.288	0.294	0.0670 (0.871)	0.759	0.745	0.234
	Coeff.	+	+	+	+	+	+	+	+	+

CBL	**P value**	0.657	0.625	0.582	0.758	0.802	0.910	0.226	0.283	0.526
	Coeff.	-	-	-	+	+	+	-	-	-

Table 7.9: Logistic regression results testing CNV case control status on regional splicing logP calculated in the Johnson dataset.
All p values in brackets were adjusted for 13 brain regions using Bonferroni's method, where missing corrected p value was 1.

		All CNVs			Deletions			Duplications		
		Min.	**Med.**	**Max.**	**Min.**	**Med.**	**Max.**	**Min.**	**Med.**	**Max.**
DLPFC	**P value**	0.213	0.186	0.260	0.0556 (0.723)	0.431	0.970	0.973	0.254	0.0118
	Coeff.	+	+	+	+	+	-	+	+	+
MPFC	**P value**	0.265	0.0828	0.00212 (0.0275)	0.302	0.232	0.0826	0.589	0.202	0.0132
	Coeff.	+	+	+	+	+	+	+	+	+
OPFC	**P value**	0.0493 (0.0641)	0.112	0.890	0.539	0.528	0.761	0.0515 (0.670)	0.150	0.980
	Coeff.	-	-	-	-	-	-	-	-	-
VLPFC	**P value**	0.0230 (0.300)	0.000523 (0.00680)	0.0469 (0.609)	0.0455 (0.591)	0.0335 (0.436)	0.149	0.186	0.00522 (0.0679)	0.0940
	Coeff.	+	+	+	+	+	+	+	+	+
MS	**P value**	0.0105	0.0986	0.804	0.000975 (0.0127)	0.00162 (0.0211)	0.0324 (0.421)	0.843	0.166	0.0365 (0.475)
	Coeff.	-	-	+	-	-	-	+	+	+
PAS	**P value**	0.00700 (0.0910)	0.0175 (0.227)	0.0288 (0.0374)	0.0184 (0.239)	0.0162 (0.210)	0.00162 (0.0211)	0.119	0.283	0.681
	Coeff.	-	-	-	-	-	-	-	-	-
TAU	**P value**	0.764	0.347	0.0300 (0.390)	0.304	0.475	0.510	0.184	0.0548 (0.713)	0.00381 (0.0495)
	Coeff.	+	+	+	-	-	-	+	+	+
TAS	**P value**	0.0356 (0.462)	0.0117 (0.152)	0.0769	5.51×10^{-7} (7.16×10^{-6})	6.29×10^{-7} (8.18×10^{-6})	7.38×10^{-5} (0.000959)	0.197	0.535	0.598
	Coeff.	+	+	+	+	+	+	-	-	-
OCC	**P value**	0.000222 (0.00289)	0.000193 (0.00251)	0.0153 (0.199)	0.0599 (0.779)	0.0940	0.115	0.000471 (0.00612)	0.000169 (0.00220)	0.0619 (0.804)
	Coeff.	-	-	-	-	-	-	-	-	-

HIP	**P value**	0.138	0.138	0.0868	0.348	0.297	0.0896	0.246	0.310	0.468
	Coeff.	-	-	-	-	-	-	-	-	-
STR	**P value**	0.0358 (0.466)	0.0784	0.0733 (0.953)	0.00644 (0.0837)	0.00256 (0.0332)	0.00373 (0.0485)	0.00180 (0.235)	0.00338 (0.0439)	0.00254 (0.0330)
	Coeff.	-	-	-	+	+	+	-	-	-
THAL	**P value**	0.694	0.339	0.589	0.0359 (0.467)	0.0223 (0.289)	0.0590 (0.767)	0.212	0.516	0.429
	Coeff.	+	+	+	+	+	+	-	-	-
CBL	**P value**	0.885	0.991	0.367	0.936	0.626	0.650	0.821	0.668	0.136
	Coeff.	-	-	+	-	-	-	-	+	+

Table 7.10: Logistic regression results testing CNV singleton status on regional splicing logP calculated in the Johnson dataset.
All p values in brackets were adjusted for 13 brain regions using Bonferroni's method, where missing corrected p value was 1.

		All CNVs			Deletions			Duplications		
		Min.	Med.	Max.	Min.	Med.	Max.	Min.	Med.	Max.
DLPFC	P value	0.166	0.305	0.752	0.364	0.676	0.583	0.380	0.384	0.324
	Coeff.	+	+	+	+	+	-	+	+	+
MPFC	P value	0.118	0.141	0.628	0.207	0.265	0.180	0.458	0.373	0.911
	Coeff.	+	+	+	+	+	+	+	+	-
OPFC	P value	0.503	0.284	0.140	0.170	0.169	0.239	0.846	0.828	0.533
	Coeff.	-	-	-	+	-	-	+	-	-
VLPFC	P value	0.537	0.371	0.235	0.215	0.209	0.115	0.926	0.832	0.640
	Coeff.	-	-	-	-	-	-	+	+	+
M1C	P value	0.00306 (0.0490)	0.000386 (0.00617)	0.000638 (0.0102)	0.00112 (0.0180)	0.00139 (0.0222)	0.00599 (0.0958)	0.212	0.0502 (0.804)	0.0457 (0.731)
	Coeff.	-	-	-	-	-	-	-	-	-
S1C	P value	0.0743	0.0250 (0.400)	0.920	0.632	0.495	0.370	0.0290 (0.465)	0.0131 (0.210)	0.214
	Coeff.	-	-	-	-	-	+	-	-	-
TAU	P value	0.0645	0.0540 (0.864)	0.156	0.287	0.353	0.237	0.226	0.109	0.416
	Coeff.	+	+	+	+	+	+	+	+	+
OCC	P value	0.157	0.205	0.262	0.632	0.424	0.319	0.0110	0.0197	0.0401
	Coeff.	-	-	-	+	+	+	-	-	-
IPC	P value	0.405	0.208	0.757	0.846	0.940	0.840	0.107	0.0658	0.597
	Coeff.	-	-	-	+	-	+	-	-	-

TAS	**P value**	0.506	0.732	0.878	0.549	0.545	0.968	0.974	0.846	0.571
	Coeff.	+	+	+	+	+	-	+	-	+
ITC	**P value**	0.249	0.0704	0.314	0.0414	0.0332	0.581	0.854	0.535	0.452
	Coeff.	-	-	-	-	-	-	+	-	-
HIP	**P value**	0.896	0.815	0.589	0.341	0.422	0.854	0.496	0.709	0.235
	Coeff.	+	-	+	-	-	-	+	+	+
AMY	**P value**	0.778	0.982	0.704	0.583	0.527	0.847	0.481	0.555	0.387
	Coeff.	+	+	+	-	-	-	+	+	+
STR	**P value**	0.805	0.641	0.889	0.554	0.436	0.486	0.986	0.967	0.736
	Coeff.	-	-	-	-	-	-	+	-	+
THAL	**P value**	0.0503	0.480	0.0988	0.697	0.499	0.180	0.656	0.578	0.107
	Coeff.	+	+	+	+	+	+	+	+	+
CBL	**P value**	0.502	0.496	0.499	0.933	0.939	0.918	0.135	0.192	0.346
	Coeff.	-	-	-	-	-	+	-	-	-

Table 7.11: Logistic regression results testing CNV case control status on regional splicing logP calculated in the Kang dataset.
All p values in brackets were adjusted for 16 brain regions using Bonferroni's method, where missing corrected p value was 1.

		All CNVs			Deletions			Duplications		
		Min.	**Med.**	**Max.**	**Min.**	**Med.**	**Max.**	**Min.**	**Med.**	**Max.**
DLPFC	**P value**	0.000566 (0.00906)	0.00677 (0.108)	0.0995	0.0297 (0.475)	0.0911	0.377	0.00862 (0.138)	0.0347 (0.555)	0.130
	Coeff.	-	-	-	-	-	-	-	-	-
MPFC	**P value**	0.382	0.171	0.113	0.385	0.323	0.234	0.750	0.340	0.295
	Coeff.	+	+	+	+	+	+	+	+	+
OPFC	**P value**	0.000367 (0.00587)	0.000111 (0.00177)	6.25×10^{-5} (0.00100)	0.000229 (0.00367)	1.77×10^{-5} (0.000282)	1.27×10^{-7} (2.03×10^{-6})	0.109	0.123	0.150
	Coeff.	+	+	+	+	+	+	+	+	+
VLPFC	**P value**	0.0770	0.784	0.584	0.988	0.683	0.299	0.0216	0.410	0.712
	Coeff.	-	-	+	-	+	+	-	-	-
M1C	**P value**	0.975	0.856	0.344	0.580	0.615	0.956	0.611	0.884	0.303
	Coeff.	-	+	+	+	+	+	-	-	+
S1C	**P value**	0.177	0.348	0.0445	0.152	0.493	0.0459 (0.734)	0.652	0.513	0.817
	Coeff.	-	-	+	-	-	+	-	-	+
TAU	**P value**	6.77×10^{-7} (1.08×10^{-5})	2.27×10^{-6} (3.64×10^{-5})	0.00268 (0.0428)	3.46×10^{-5} (5.53×10^{-4})	2.69×10^{-5} (4.30×10^{-4})	0.00125 (0.0200)	0.0142 (0.227)	0.0634	0.352
	Coeff.	-	-	-	-	-	-	-	-	-
OCC	**P value**	0.750	0.704	0.239	0.478	0.444	0.143	0.889	0.874	0.773
	Coeff.	+	+	+	+	+	+	-	-	+
IPC	**P value**	0.198	0.185	0.877	0.0247 (0.396)	0.0449 (0.719)	0.0781	0.556	0.809	0.0794
	Coeff.	-	-	+	-	-	-	+	+	+
TAS	**P value**	0.00322 (0.0516)	0.0116 (0.185)	0.0203 (0.325)	0.0106 (0.170)	0.0638	0.130	0.129	0.0816	0.0936
	Coeff.	+	+	+	+	+	+	+	+	+

ITC	**P value**	0.245	0.184	0.400	0.818	0.909	0.565	0.0621 (0.993)	0.0935	0.566
	Coeff.	-	-	-	+	-	-	-	-	-
HIP	**P value**	0.0212 (0.339)	0.222	0.884	0.271	0.403	0.782	0.0406 (0.650)	0.397	0.685
	Coeff.	-	-	+	-	-	-	-	-	+
AMY	**P value**	0.588	0.623	0.667	0.676	0.605	0.613	0.730	0.853	0.355
	Coeff.	-	-	+	-	-	-	-	-	+
STR	**P value**	0.901	0.845	0.554	0.0248 (0.396)	0.0322 (0.515)	0.0482 (0.772)	0.195	0.384	0.772
	Coeff.	-	+	+	+	+	+	-	-	-
THAL	**P value**	0.358	0.615	0.795	0.0738	0.0731	0.163	0.00509 (0.0814)	0.0241 (0.385)	0.150
	Coeff.	-	-	-	+	+	+	-	-	-
CBL	**P value**	0.184	0.200	0.215	0.205	0.154	0.167	0.510	0.691	0.616
	Coeff.	-	-	-	-	-	-	-	-	-

Table 7.12: Logistic regression results testing CNV singleton status on regional splicing logP calculated in the Kang dataset.
All p values in brackets were adjusted for 16 brain regions using Bonferroni's method, where missing corrected p value was 1.

7.2 Additional figures for Chapter 2

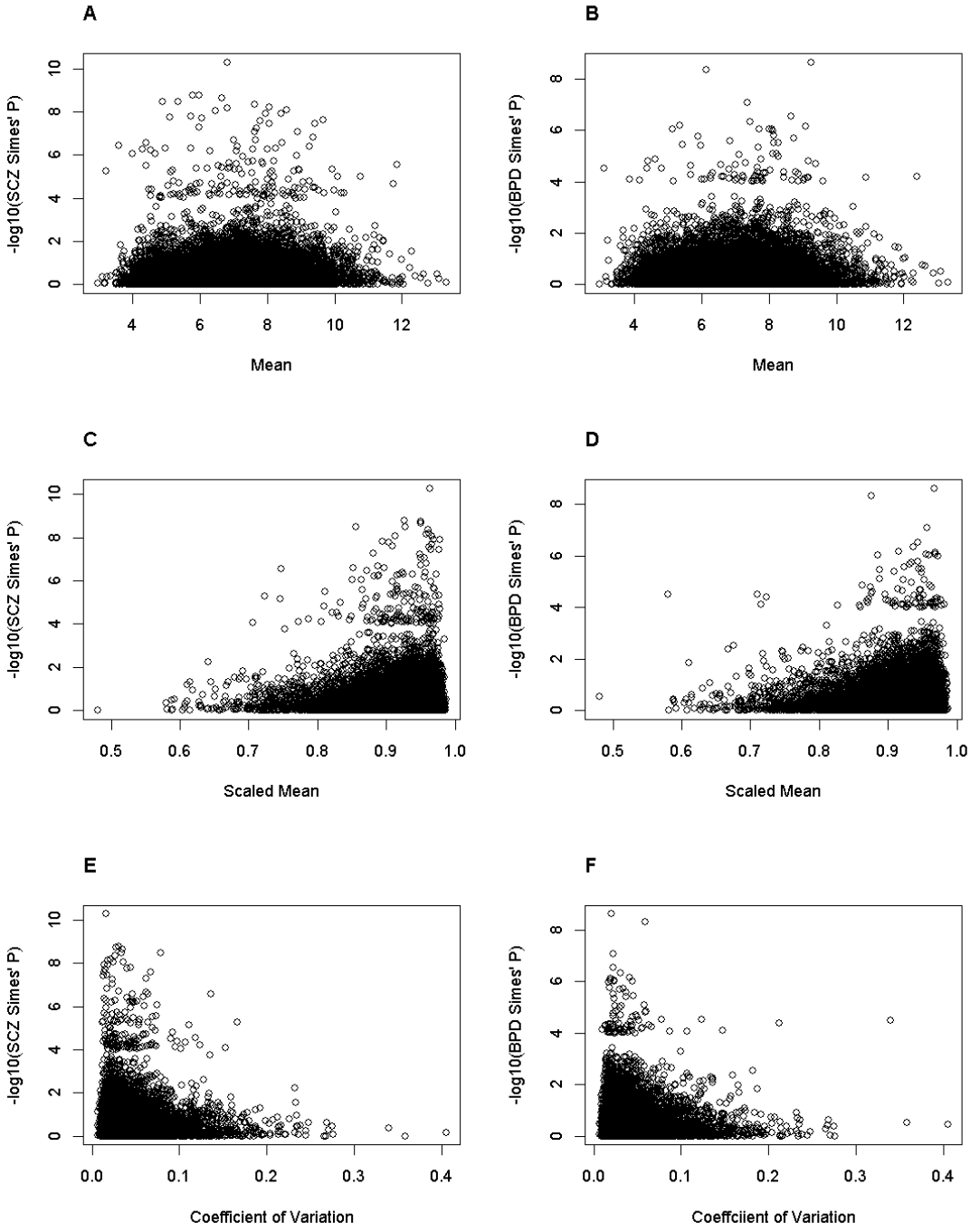

Figure 7.1: Scatterplots of relationships between global metrics calculated in the Johnson dataset and Simes' gene-wide p values.
Panels A, C and E plot SCZ Brown's logP against global metrics; panels B, D and F plot BPD Simes' logP against global metrics. Panels A and B plot mean expression across mid-foetal brain; panels C and D plot scaled mean; panels E and F plot coefficient of variation against gene-wide logP.
.

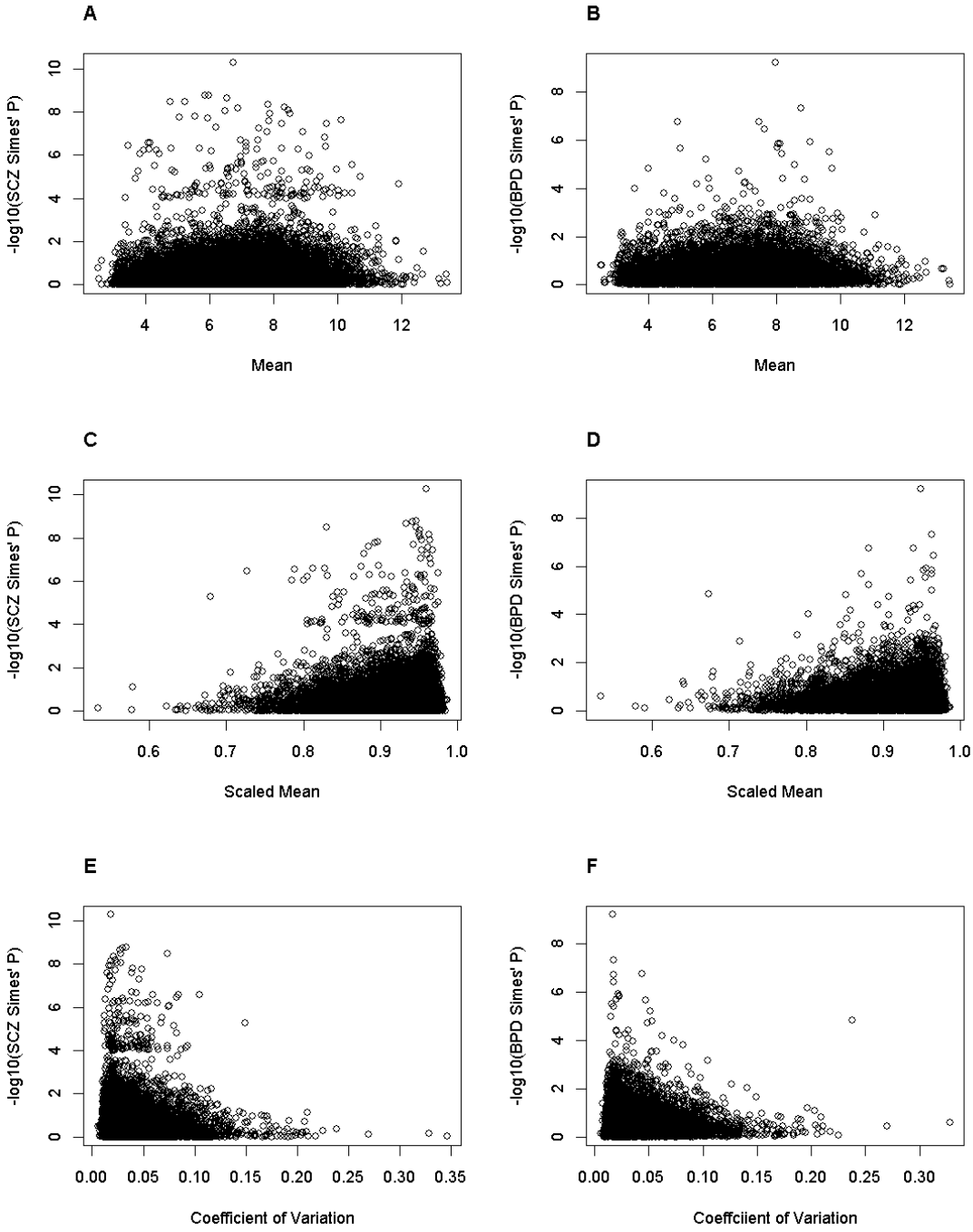

Figure 7.2: Scatterplots of relationships between testing global metrics calculated within neocortical regions in the Johnson dataset and Simes' gene-wide p values.
Panels A, C and E plot SCZ Brown's logP against global metrics; panels B, D and F plot BPD Simes' logP against global metrics. Panels A and B plot mean expression across mid-foetal brain; panels C and D plot scaled mean; panels E and F plot coefficient of variation against gene-wide logP.

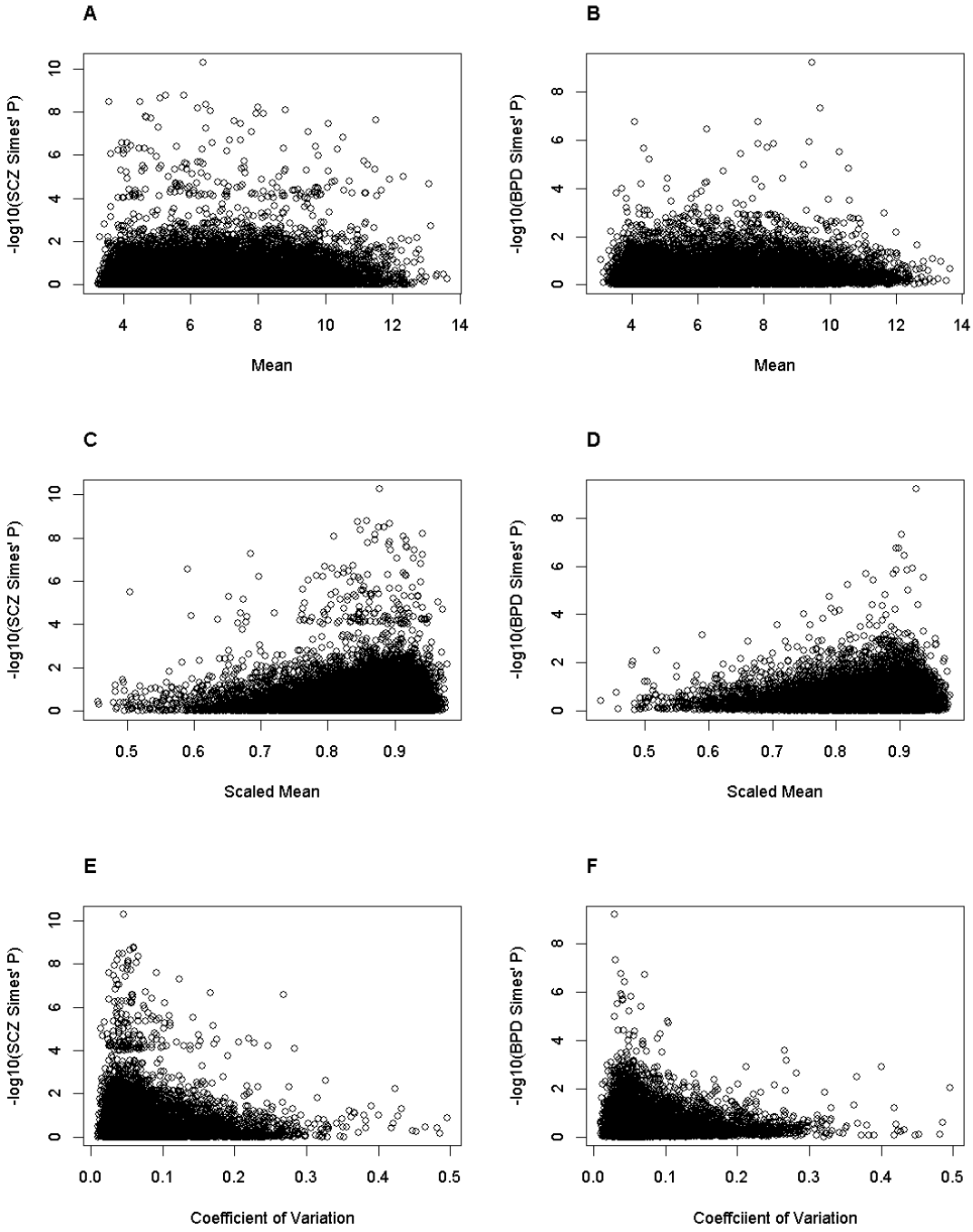

Figure 7.3: Scatterplots of relationships between global metrics calculated in the Kang dataset and Simes' gene-wide p values.
Panels A, C and E plot SCZ Brown's logP against global metrics; panels B, D and F plot BPD Simes' logP against global metrics. Panels A and B plot mean expression across mid-foetal brain; panels C and D plot scaled mean; panels E and F plot coefficient of variation against gene-wide logP.

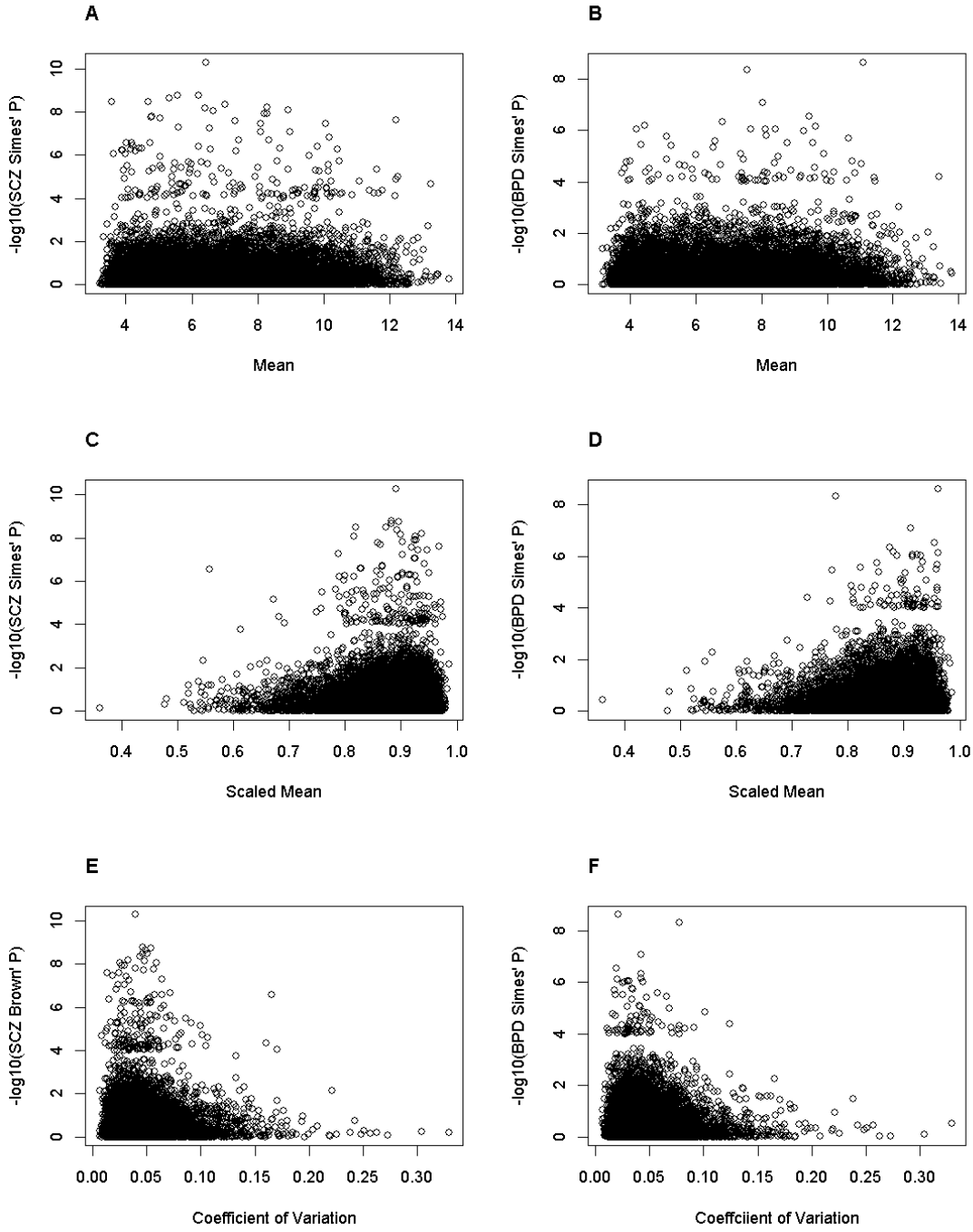

Figure 7.4: Scatterplots of relationships between testing global metrics calculated within neocortical regions in the Kang dataset and Simes' gene-wide p values.
Panels A, C and E plot SCZ Brown's logP against global metrics; panels B, D and F plot BPD Simes' logP against global metrics. Panels A and B plot mean expression across mid-foetal brain; panels C and D plot scaled mean; panels E and F plot coefficient of variation against gene-wide logP.

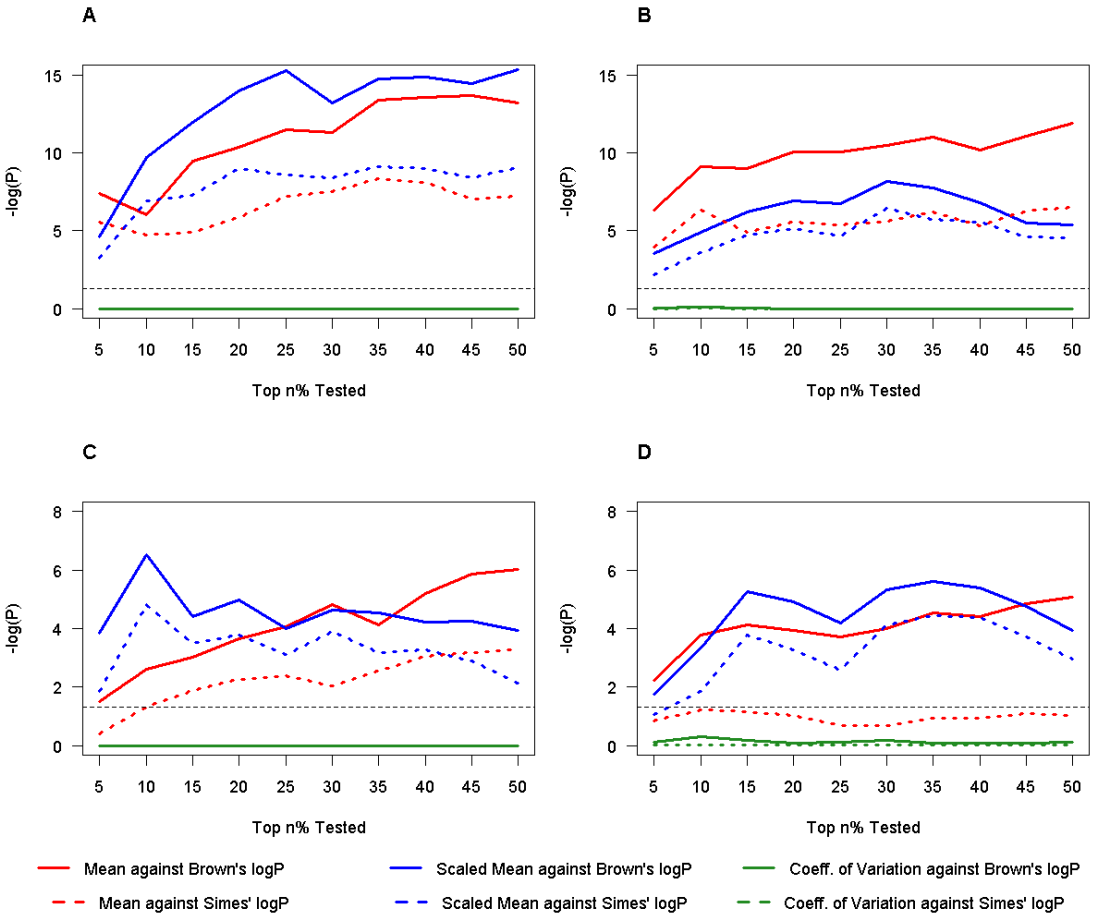

Figure 7.5: Results from Mann-Whitney tests for genes ranked by global metrics calculated in the Johnson and Kang datasets, excluding MHC genes.
Panels A & B are results testing top n% of genes ranked by each global metric in turn against the bottom 50% for smaller SCZ p values; panels C & D are results testing for smaller BPD p values. Panels A & C global metrics were calculated in the Johnson dataset; panels B & D were calculated in Kang dataset. Analyses were run excluding MHC genes. Black dashed line is 0.05.

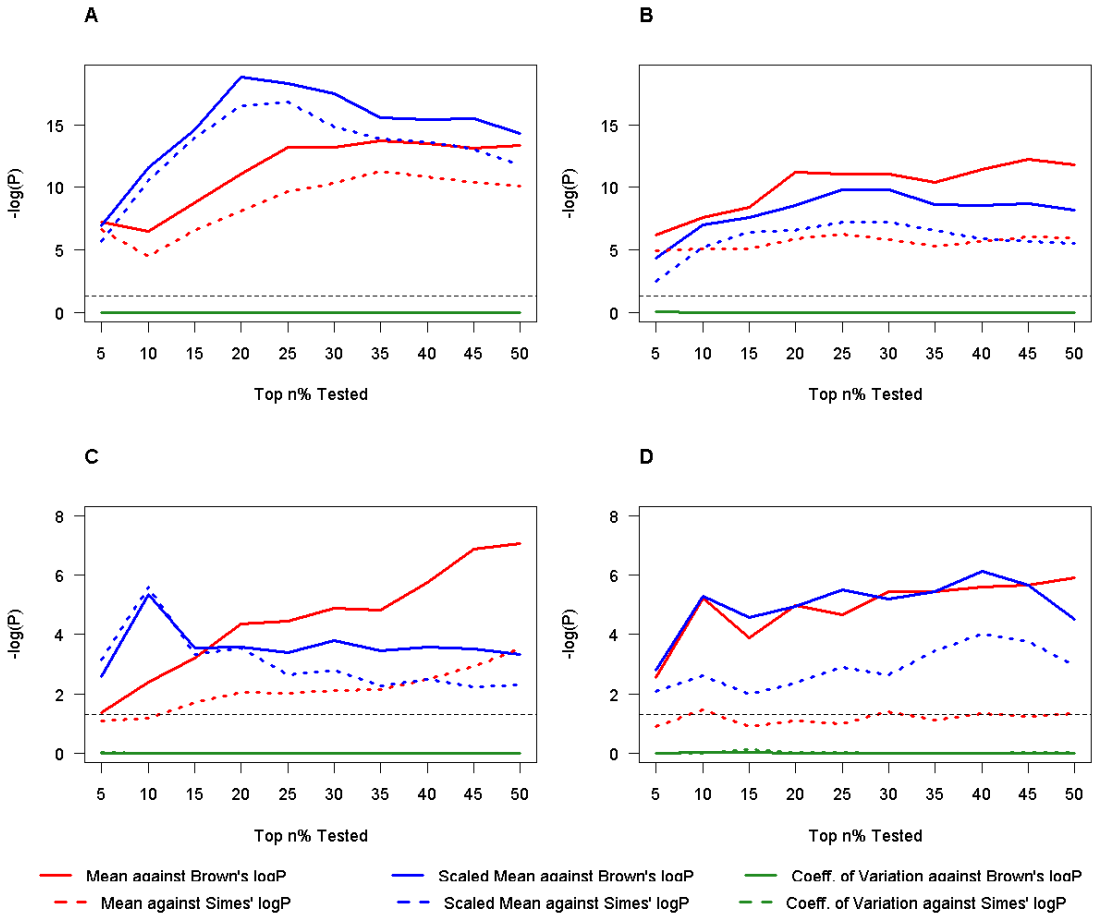

Figure 7.6: Results from Mann-Whitney tests for genes ranked by global metrics within neocortical regions calculated in the Johnson and Kang datasets, excluding MHC genes.
Panels A & B are results testing top n% of genes ranked by each global metric in turn against the bottom 50% for smaller SCZ p values; panels C & D are results testing for smaller BPD p values. Panels A & C global metrics were calculated in the Johnson dataset; panels B & D were calculated in the Kang dataset. Analyses were run excluding MHC genes. Black dashed line is 0.05.

Chapter 8: Appendix B

8.1 Additional tables for Chapter 3

		Schizophrenia		Bipolar disorder		Parkinson's	Alzheimer's
		Brown's	Simes'	Brown's	Simes'	Brown's	Brown's
Early foetal A	P value	0.844	0.401	0.475	0.0977	0.368	0.408
	Correlation Coeff.	0.0130	-0.00772	-0.00621	-0.0153	-0.00791	-0.00683
		+	-	-	-	-	-
Early foetal B	P value	0.000147 (0.00176)	0.00791 (0.0950)	0.575	0.314	0.237	0.569
	Correlation Coeff.	0.0293	0.0244	-0.00488	-0.00928	-0.0104	-0.00470
		+	+	-	-	-	-
Early mid-foetal A	P value	7.80×10^{-10} (9.36×10^{-9})	1.14×10^{-8} (1.36×10^{-7})	0.0480 (0.576)	0.167	0.895	0.182
	Correlation Coeff.	0.0562	0.0525	0.0172	0.0127	-0.00116	-0.0110
		+	+	+	+	-	-
Early mid-foetal B	P value	4.40×10^{-8} (5.28×10^{-7})	1.13×10^{-6} (1.36×10^{-5})	0.00737 (0.0884)	0.0236 (0.283)	0.309	0.321
	Correlation Coeff.	0.0538	0.0448	0.0233	0.0209	0.00893	-0.00819
		+	+	+	+	+	-
Late mid-foetal	P value	4.45×10^{-5} (0.000534)	1.14×10^{-5} (0.000137)	0.0133 (0.160)	0.00458 (0.0550)	0.117	0.923
	Correlation Coeff.	0.0263	0.0404	0.0215	0.0261	0.0138	0.000800
		+	+	+	+	+	+
Late foetal	P value	0.000324 (0.00389)	0.00228 (0.0274)	0.00188 (0.0225)	0.00688 (0.0826)	0.800	0.109
	Correlation Coeff.	-0.0277	-0.281	-0.0270	-0.0249	-0.00223	-0.0132
		-	-	-	-	-	-

Neonatal and early infancy	**P value**	0.0246 (0.295)	0.142	0.862	0.457	0.127	0.575
	Correlation Coeff.	-0.0305	-0.0135	-0.00151	0.00685	0.0134	0.00462
		-	-	-	-	+	+
Late infancy	**P value**	6.09×10^{-5} (0.000731)	0.00547 (0.0656)	0.0407 (0.489)	0.221	0.258	0.846
	Correlation Coeff.	-0.0365	-0.0256	-0.0178	-0.0113	-0.00993	0.00160
		-	-	-	-	-	+
Early childhood	**P value**	1.29×10^{-16} (1.55×10^{-15})	1.24×10^{-8} (1.49×10^{-7})	0.00301 (0.0361)	0.0467 (0.561)	0.127	0.258
	Correlation Coeff.	-0.0616	-0.0524	-0.0258	-0.0183	-0.0134	0.00932
		-	-	-	-	-	+
Middle and late childhood	**P value**	1.06×10^{-6} (1.27×10^{-5})	1.32×10^{-6} (1.58×10^{-5})	0.0491 (0.589)	0.0791 (0.949)	0.751	0.774
	Correlation Coeff.	-0.0426	-0.0445	-0.0171	-0.0162	0.00279	0.00237
		-	-	-	-	+	+
Adolescence	**P value**	2.91×10^{-5} (0.000350)	8.87×10^{-5} (0.00106)	0.264	0.376	0.912	0.453
	Correlation Coeff.	-0.0375	-0.0361	-0.00970	-0.00816	-0.000966	0.00619
		-	-	-	-	-	+
Young adulthood	**P value**	0.0405 (0.486)	0.0583 (0.700)	0.497	0.394	0.770	0.300
	Correlation Coeff.	-0.0231	-0.0174	0.00590	0.00785	-0.00257	0.00854
		-	-	+	-	-	+

Table 8.1: Linear regression results and correlation coefficients testing development stage characteristic scores calculated in the BrainSpan RNA-Seq dataset with gene-wide logP.
P values in brackets corrected for 12 development stages, where missing corrected p value was 1.

		Schizophrenia		Bipolar disorder	
		Brown's	**Simes'**	**Brown's**	**Simes'**
Early foetal A	**P value**	0.784	0.199	0.339	0.0535 (0.643)
	Correlation Coeff.	-0.00244	-0.0119	-0.00835	-0.0179
		-	-	-	-
Early foetal B	**P value**	0.00143 (0.0228)	0.00999 (0.120)	0.575	0.318
	Correlation Coeff.	0.0285	0.0239	-0.00491	-0.00927
		+	+	-	-
Early mid-foetal A	**P value**	3.54×10^{-8} (5.67×10^{-7})	6.97×10^{-7} (8.36×10^{-6})	0.0618 (0.741)	0.233
	Correlation Coeff.	0.0492	0.0460	0.0163	0.0111
		+	+	+	+
Early mid-foetal B	**P value**	1.04×10^{-6} (1.66×10^{-5})	0.000167 (0.00201)	0.0112 (0.135)	0.0380 (0.456)
	Correlation Coeff.	0.0436	0.0349	0.0222	0.0192
		+	+	+	+
Late mid-foetal	**P value**	0.000377 (0.00603)	0.000237 (0.00284)	0.0175 (0.210)	0.00504 (0.0604)
	Correlation Coeff.	0.0317	0.0341	0.0208	0.0260
		+	+	+	+
Late foetal	**P value**	0.000203 (0.00324)	0.00451 (0.0542)	0.00202 (0.0243)	0.00990 (0.119)
	Correlation Coeff.	-0.0332	-0.0263	-0.0270	-0.0239
		-	-	-	-
Neonatal and early infancy	**P value**	0.258	0.779	0.904	0.298
	Correlation Coeff.	-0.101	0.00260	0.00105	0.00965
		-	+	+	+
Late Infancy	**P value**	9.14×10^{-7} (1.46×10^{-5})	0.000106 (0.00127)	0.0294 (0.353)	0.173
	Correlation Coeff.	-0.0438	-0.0359	-0.0190	-0.012
		-	-	-	-
Early childhood	**P value**	1.22×10^{-17} (1.95×10^{-16})	4.22×10^{-10} (5.07×10^{-9})	0.00292 (0.0351)	0.0419 (0.503)
	Correlation Coeff.	-0.0763	-0.0579	-0.0260	-0.0189
		-	-	-	-
Middle and late childhood	**P value**	2.78×10^{-5} (0.000445)	5.66×10^{-5} (0.000679)	0.0524 (0.629)	0.0727 (0.872)
	Correlation Coeff.	-0.0374	-0.0373	-0.0170	-0.0166
		-	-	-	-
Adolescence	**P value**	0.000417 (0.00667)	0.00111 (0.0133)	0.383	0.573
	Correlation Coeff.	-0.0315	-0.0302	-0.00763	-0.00522
		-	-	-	-

Young adulthood	**P value**	0.126	0.116	0.428	0.328
	Correlation Coeff.	-0.0136	-0.0145	0.00693	0.00907
		-	-	+	+

Table 8.2: Linear regression results and correlation coefficients testing development stage characteristic scores calculated in the BrainSpan RNA-Seq dataset with gene-wide logP, excluding MHC genes.
P values in brackets corrected for 12 development stages, where missing corrected p value was 1.

		Schizophrenia		Bipolar disorder		Parkinson's	Alzheimer's
		Brown's	Simes'	Brown's	Simes'	Brown's	Brown's
Embryonic	P value	0.881	0.000784 (0.0118)	0.454	0.123	0.298	0.222
	Correlation Coeff.	0.00130	-0.0293	0.00641	-0.0135	-0.00896	-0.00986
		+	-	+	-	-	-
Early foetal A	P value	0.655	0.0544 (0.816)	0.800	0.220	0.174	0.261
	Correlation Coeff.	-0.00390	-0.0168	0.00216	-0.0108	-0.0117	-0.00906
		-	-	+	-	-	-
Early foetal B	P value	0.000452 (0.00678)	0.291	0.581	0.135	0.227	0.567
	Correlation Coeff.	0.0306	0.00920	0.00472	-0.0131	-0.0104	-0.00462
		+	+	+	-	-	-
Early mid-foetal A	P value	0.493	0.602	0.446	0.938	0.953	0.505
	Correlation Coeff.	0.00597	0.00454	0.00651	-0.000677	0.000506	0.00538
		+	+	+	-	+	+
Early mid-foetal B	P value	0.0277 (0.415)	0.000825 (0.0124)	0.0405 (0.607)	0.00207 (0.0311)	0.0938	0.506
	Correlation Coeff.	0.0192	0.0291	0.0175	0.0270	0.0144	-0.00536
		+	+	+	+	+	-
Late mid-foetal	P value	0.142	0.00196 (0.0293)	0.926	0.518	0.264	0.721
	Correlation Coeff.	0.0128	0.0270	-0.000798	0.00566	-0.00961	-0.00288
		+	+	-	+	-	-
Late foetal	P value	0.382	0.751	0.461	0.859	0.960	0.0654 (0.980)
	Correlation Coeff.	-0.00762	0.00277	-0.00631	-0.00155	0.000434	0.0149
		-	+	-	-	+	+
Neonatal and early infancy	P value	0.00447 (0.0670)	0.513	0.905	0.538	0.583	0.356
	Correlation	-0.0248	-0.00570	-0.00102	0.00540	-0.00473	0.00745

	Coeff.	-	-	-	+	-	+
Late infancy	**P value**	0.171	0.869	0.381	0.818	0.372	0.742
	Correlation Coeff.	-0.0119	-0.00144	-0.00749	0.00202	-0.00769	0.00265
		-	-	-	+	-	+
Early childhood	**P value**	0.000382 (0.00573)	0.00174 (0.0261)	0.219	0.244	0.658	0.548
	Correlation Coeff.	-0.0310	-0.0273	-0.0105	-0.0102	0.00381	0.00484
		-	-	-	-	+	+
Middle and late childhood	**P value**	0.194	0.00691 (0.104)	0.000102 (0.00153)	0.000148 (0.00222)	0.521	0.0917
	Correlation Coeff.	-0.0113	-0.0235	-0.0332	-0.0332	-0.00553	-0.0136
		-	-	-	-	-	-
Adolescence	**P value**	0.00341 (0.0511)	0.0850	0.592	0.983	0.262	0.335
	Correlation Coeff.	0.0255	0.0150	0.00459	0.000189	0.00967	-0.00778
		+	+	+	+	+	-
Young adulthood	**P value**	6.13×10^{-5} (0.000919)	0.0572 (0.859)	0.0553 (0.830)	0.207	0.154	0.182
	Correlation Coeff.	0.0349	0.0166	0.0164	0.011	0.0123	-0.0108
		+	+	+	+	+	-
Middle adulthood	**P value**	0.00447 (0.0670)	0.000761 (0.0114)	0.0327 (0.490)	0.154	0.359	0.460
	Correlation Coeff.	-0.0248	-0.0293	-0.0183	-0.0125	0.00790	0.00595
		-	-	-	-	+	+
Late adulthood	**P value**	0.000423 (0.00634)	0.000282 (0.00423)	0.161	0.432	0.845	0.746
	Correlation Coeff.	-0.0307	-0.0317	-0.0120	-0.00689	-0.00169	0.00261
		-	-	-	-	-	+

Table 8.3: Linear regression results and correlation coefficients testing development stage characteristic scores calculated in the Kang microarray dataset with gene-wide logP.

P values in brackets corrected for 15 development stages, where missing corrected p value was 1.

		Schizophrenia		Bipolar disorder	
		Brown's	**Simes'**	**Brown's**	**Simes'**
Embryonic	**P value**	0.680	9.37×10^{-7} (1.41×10^{-5})	0.515	0.0619 (0.928)
	Correlation Coeff.	-0.00362	-0.0430	0.00560	-0.0165
		-	-	+	-
Early foetal A	**P value**	0.414	0.00318 (0.0478)	0.841	0.130
	Correlation Coeff.	-0.00716	-0.0259	0.00173	-0.0134
		-	-	+	-
Early foetal B	**P value**	0.00663 (0.0995)	0.907	0.614	0.126
	Correlation Coeff.	0.0238	0.00102	0.00434	-0.0135
		+	+	+	-
Early mid-foetal A	**P value**	0.725	0.947	0.460	0.910
	Correlation Coeff.	0.00308	0.000586	0.00636	-0.00100
		+	+	+	-
Early mid-foetal B	**P value**	0.0263 (0.394)	0.000674 (0.0101)	0.0700	0.00487
	Correlation Coeff.	0.0195	0.0298	0.0156	0.0248
		+	+	+	+
Late mid-foetal	**P value**	0.281	0.00160 (0.0249)	0.991	0.384
	Coeff.	0.00946	0.0276	9.63×10^{-5}	0.00768
		+	+	+	+
Late foetal	**P value**	0.118	0.329	0.351	0.686
	Correlation Coeff.	-0.0137	-0.00857	-0.00802	-0.00356
		-	-	-	-
Neonatal and early infancy	**P value**	0.00601 (0.0902)	0.532	0.920	0.524
	Correlation Coeff.	-0.0241	-0.00549	-0.000863	0.00562
		-	-	-	+
Late infancy	**P value**	0.158	0.991	0.344	0.827
	Correlation Coeff.	-0.0124	0.000101	-0.00814	0.00193
		-	+	-	+
Early childhood	**P value**	0.00412 (0.0619)	0.0194 (0.291)	0.257	0.313
	Correlation Coeff.	-0.0251	-0.0205	-0.00975	-0.00890
		-	-	-	-
Middle and late childhood	**P value**	0.420	0.0306 (0.459)	0.000208 (0.00312)	0.000340 (0.00510)
	Correlation Coeff.	-0.00707	-0.0190	-0.0319	-0.0316
		-	-	-	-
Adolescence	**P value**	2.87×10^{-5} (0.000430)	0.000229 (0.00343)	0.359	0.645
	Correlation Coeff.	0.00367	0.0323	0.00790	0.00406
		+	+	+	+

Young adulthood	**P value**	5.08×10^{-5} (0.000763)	0.0246 (0.369)	0.0494 (0.740)	0.220
	Correlation Coeff.	0.0355	0.0197	0.0169	0.0108
		+	+	+	+
Middle adulthood	**P value**	0.0255 (0.383)	0.0317 (0.476)	0.0267 (0.401)	0.134
	Correlation Coeff.	-0.0196	-0.0189	-0.0191	-0.0132
		-	-	-	-
Late adulthood	**P value**	0.000331 (0.00496)	9.21×10^{-5}	0.158	0.396
	Correlation Coeff.	-0.0315	-0.0343	-0.0121	-0.00749
		-	-	-	-

Table 8.4: Linear regression results and correlation coefficients testing development stage characteristic scores calculated in the Kang microarray dataset with gene-wide logP, excluding MHC genes.
P values in brackets corrected for 15 development stages, where missing corrected p value was 1.

		Schizophrenia		Bipolar disorder	
		Brown's	**Simes'**	**Brown's**	**Simes'**
Embryonic	**P value**	0.776	9.50×10^{-5} (0.00143)	0.678	0.0886
	Correlation Coeff.	-0.00248	-0.0340	0.00355	-0.0149
		-	-	+	-
Early foetal A	**P value**	0.907	0.152	0.735	0.165
	Correlation Coeff.	0.00102	-0.0125	0.00290	-0.0122
		+	-	+	-
Early foetal B	**P value**	4.77×10^{-5} (0.000715)	0.0354 (0.532)	0.200	0.667
	Correlation Coeff.	0.0355	0.0183	0.0110	-0.00378
		+	+	+	-
Early mid-foetal A	**P value**	5.00×10^{-5} (0.000750)	0.00116 (0.0174)	0.123	0.917
	Correlation Coeff.	0.0354	0.0283	0.0132	-0.000913
		+	+	+	-
Early mid-foetal B	**P value**	2.39×10^{-7} (8.59×10^{-8})	8.59×10^{-8} (1.29×10^{-6})	0.00115 (0.0173)	0.00812 (0.122)
	Correlation Coeff.	0.0450	0.0467	0.0278	0.0232
		+	+	+	+
Late mid-foetal	**P value**	0.0152 (0.228)	0.000628 (0.00942)	0.535	0.510
	Correlation Coeff.	0.0212	0.0298	0.00531	0.00577
		+	+	+	+
Late foetal	**P value**	0.264	0.597	0.467	0.770
	Correlation Coeff.	-0.00974	0.00460	-0.00622	-0.00256
		-	+	-	-
Neonatal and early infancy	**P value**	0.00138 (0.0207)	0.356	0.704	0.166
	Correlation Coeff.	-0.0279	-0.0099	0.00325	0.0121
		-	-	+	+
Late infancy	**P value**	2.27×10^{-6} (3.41×10^{-5})	1.83×10^{-5} (2.74×10^{-4})	0.00650 (0.0976)	0.268
	Correlation Coeff.	-0.0412	-0.0373	-0.0233	-0.00971
		-	-	-	-
Early childhood	**P value**	0.000691 (0.0104)	0.000575 (0.00862)	0.00418 (0.0627)	0.00467 (0.0700)
	Correlation Coeff.	-0.0296	-0.0300	-0.0245	-0.0248
		-	-	-	-
Middle and late childhood	**P value**	0.00454 (0.0680)	6.29×10^{-5} (0.000943)	0.00326 (0.0490)	0.0115 (0.172)
	Correlation Coeff.	-0.0247	-0.0349	-0.0252	-0.0221
		-	-	-	-

Adolescence	**P value**	0.0241 (0.361)	0.0565 (0.848)	0.769	0.701
	Correlation Coeff.	-0.0197	-0.0166	-0.00252	0.00337
		-	-	-	+
Young adulthood	**P value**	0.831	0.298	0.666	0.593
	Correlation Coeff.	-0.00186	-0.00907	0.00370	0.00469
		-	-	+	+
Middle adulthood	**P value**	1.95×10^{-5} (0.000293)	1.65×10^{-5} (0.000248)	0.0152	0.143
	Correlation Coeff.	-0.0372	-0.0375	-0.0208	-0.0128
		-	-	-	-
Late adulthood	**P value**	0.0146 (0.219)	0.143	0.186	0.452
	Correlation Coeff.	-0.0213	-0.0255	-0.0113	-0.00659
		-	-	-	-

Table 8.5: Linear regression results and correlation coefficients testing development stage characteristic scores calculated in the Kang microarray dataset without PMI covariate with gene-wide logP.
P values in brackets corrected for 15 development stages, where missing corrected p value was 1

		Schizophrenia		Bipolar disorder	
		Brown's	**Simes'**	**Brown's**	**Simes'**
Embryonic	**P value**	0.442	1.29×10^{-7} (1.94×10^{-6})	0.732	0.0554 (0.831)
	Correlation Coeff.	-0.00674	-0.0463	0.00295	-0.0169
		-	-	+	-
Early foetal A	**P value**	0.722	0.00843 (0.126)	0.829	0.0815
	Correlation Coeff.	-0.00312	-0.0231	0.00186	-0.0154
		-	-	+	-
Early foetal B	**P value**	0.00218 (0.0326)	0.394	0.241	0.583
	Correlation Coeff.	0.0269	0.00748	0.0101	-0.00484
		+	+	+	-
Early mid-foetal A	**P value**	0.00218 (0.0326)	0.0817	0.188	0.724
	Correlation Coeff.	0.0266	0.0153	0.0113	-0.00312
		+	+	+	-
Early mid-foetal B	**P value**	4.87×10^{-5} (0.000730)	0.000209 (0.00313)	0.00433 (0.0649)	0.0267 (0.401)
	Correlation Coeff.	0.0356	0.0325	0.0245	0.0195
		+	+	+	+
Late mid-foetal	**P value**	0.0818	0.00461 (0.0691)	0.566	0.467
	Correlation Coeff.	0.0153	0.0249	0.00493	0.00642
		+	+	+	+
Late foetal	**P value**	0.0617	0.502	0.356	0.582
	Correlation Coeff.	-0.0164	-0.00590	-0.00794	-0.00486
		-	-	-	-
Neonatal and early infancy	**P value**	0.0273 (0.409)	0.925	0.651	0.133
	Correlation Coeff.	-0.0194	0.000826	0.00389	0.0132
		-	+	+	+
Late infancy	**P value**	8.07×10^{-5} (0.00121)	0.000861 (0.0129)	0.00902 (0.135)	0.365
	Correlation Coeff.	-0.0345	-0.0292	-0.0225	-0.00799
		-	-	-	-
Early childhood	**P value**	0.00157 (0.0236)	0.000637 (0.00956)	0.00414 (0.0621)	0.00493 (0.0740)
	Correlation Coeff.	-0.0277	-0.0300	-0.0247	-0.0248
		-	-	-	-
Middle and late childhood	**P value**	0.132	0.0225 (0.338)	0.00681 (0.102)	0.0244 (0.366)
	Correlation Coeff.	-0.0132	-0.0200	-0.0233	-0.0199
		-	-	-	-

Adolescence	**P value**	0.221	0.904	0.916	0.497
	Correlation Coeff.	-0.0107	-0.00106	-0.000911	0.00599
		-	-	-	+
Young adulthood	**P value**	0.651	0.916	0.541	0.524
	Correlation Coeff.	0.00396	-0.000930	0.00526	0.00563
		+	-	+	+
Middle adulthood	**P value**	0.000151 (0.00227)	0.000540 (0.00810)	0.0210 (0.314)	0.176
	Correlation Coeff.	-0.0332	-0.0304	-0.0199	-0.0119
		-	-	-	-
Late adulthood	**P value**	0.0133 (0.200)	0.000867 (0.0130)	0.215	0.491
	Correlation Coeff.	-0.0217	-0.0292	-0.0107	-0.00607
		-	-	-	-

Table 8.6: Linear regression results and correlation coefficients testing development stage characteristic scores calculated in the Kang microarray dataset without PMI covariate with gene-wide logP, excluding MHC genes.
P values in brackets corrected for 15 development stages, where missing corrected p value was 1.

		All CNVs			Deletions			Duplications		
		Min.	Med.	Max.	Min.	Med.	Max.	Min.	Med.	Max.
Early foetal A	P value	0.248	0.511	0.705	0.0644 (0.772)	0.0773	0.118	0.981	0.692	0.166
	Coeff.	+	+	-	+	+	+	+	-	-
Early foetal B	P value	0.329	0.462	0.510	0.434	0.571	0.783	0.455	0.729	0.882
	Coeff.	-	-	-	-	-	-	-	-	-
Early mid-foetal A	P value	0.945	0.864	0.381	0.651	0.815	0.669	0.895	0.581	0.201
	Coeff.	-	+	+	-	-	+	+	+	+
Early mid-foetal B	P value	0.513	0.932	0.828	0.089	0.204	0.501	0.617	0.237	0.176
	Coeff.	-	-	+	-	-	-	+	+	+
Late mid-foetal	P value	0.269	0.357	0.445	0.052 (0.628)	0.072 (0.868)	0.133	0.910	0.627	0.512
	Coeff.	-	-	-	-	-	-	+	+	+
Late foetal	P value	0.437	0.169	0.170	0.282	0.197	0.278	0.725	0.650	0.349
	Coeff.	+	+	+	+	+	+	-	+	+
Neonatal and early infancy	P value	0.475	0.306	0.216	0.433	0.178	0.036	0.918	0.795	0.900
	Coeff.	+	+	+	+	+	+	+	+	+
Late infancy	P value	0.529	0.938	0.712	0.308	0.626	0.879	0.727	0.618	0.473
	Coeff.	-	+	+	-	-	+	-	+	+
Early childhood	P value	0.780	0.771	0.554	0.662	0.865	0.235	0.794	0.852	0.997
	Coeff.	+	+	+	-	+	+	+	+	+

Middle and late childhood	**P value**	0.691	0.857	0.539	0.569	0.184	0.099	0.517	0.213	0.832
	Coeff.	+	+	+	+	+	+	-	-	-
Adolescence	**P value**	0.530	0.695	0.850	0.647	0.826	0.993	0.382	0.612	0.958
	Coeff.	-	-	-	-	-	-	-	-	+
Young adulthood	**P value**	0.606	0.578	0.912	0.736	0.660	0.781	0.329	0.513	0.895
	Coeff.	-	-	-	-	-	-	-	-	+

Table 8.7: Logistic regression results testing CNV case control status on development stage characteristic score calculated in the BrainSpan RNA-Seq dataset.
P values in brackets corrected for 12 development stages, where missing corrected p value was 1.

		All CNVs			Deletions			Duplications		
		Min.	**Med.**	**Max.**	**Min.**	**Med.**	**Max.**	**Min.**	**Med.**	**Max.**
Early foetal A	**P value**	0.00986 (0.118)	0.00121 (0.0145)	6.56×10^{-5} (0.000788)	0.000287 (0.00344)	0.000263 (0.00315)	0.000506 (0.00608)	0.748	0.231	0.0149 (0.179)
	Coeff.	+	+	+	+	+	+	+	+	+
Early foetal B	**P value**	0.916	0.808	0.763	0.589	0.444	0.296	0.793	0.754	0.623
	Coeff.	+	+	+	+	+	+	-	-	-
Early mid-foetal A	**P value**	0.265	0.148	0.0127 (0.152)	0.100	0.0818	0.0141	0.883	0.655	0.218
	Coeff.	+	+	+	+	+	+	+	+	+
Early mid-foetal B	**P value**	0.163	0.314	0.329	0.866	0.864	0.925	0.117	0.163	0.233
	Coeff.	-	-	-	-	+	-	-	-	-
Late mid-foetal	**P value**	0.200	0.145	0.149	0.701	0.799	0.853	0.190	0.0895	0.0714 (0.857)
	Coeff.	-	-	-	-	-	-	-	-	-
Late foetal	**P value**	0.885	0.541	0.355	0.600	0.402	0.206	0.783	0.943	0.757
	Coeff.	+	+	+	+	+	+	-	+	+
Neonatal and early infancy	**P value**	0.989	0.310	0.00198 (0.0238)	0.644	0.979	0.121	0.688	0.198	0.00573 (0.0687)
	Coeff.	+	+	+	-	+	+	+	+	+
Late infancy	**P value**	0.968	0.151	0.0851	0.373	0.779	0.796	0.427	0.0314 (0.376)	0.0679
	Coeff.	+	+	+	-	-	+	+	+	+
Early childhood	**P value**	0.987	0.494	0.561	0.937	0.777	0.364	0.970	0.561	0.958
	Coeff.	-	+	+	-	+	+	+	+	-
Middle and late childhood	**P value**	0.369	0.0512	0.00132 (0.0158)	0.609	0.209	0.0214	0.444	0.143	0.0381
	Coeff.	+	+	+	+	+	+	+	+	+

Adolescence	**P value**	0.216	0.580	0.907	0.553	0.717	0.996	0.281	0.680	0.978
	Coeff.	-	-	+	-	-	+	-	-	-
Young adulthood	**P value**	0.083	0.182	0.346	0.0489 (0.587)	0.0620 (0.744)	0.0675 (0.809)	0.685	0.969	0.775
	Coeff.	-	-	-	-	-	-	-	+	+

Table 8.8: Logistic regression results testing CNV singleton status on development stage characteristic scores calculated in the BrainSpan RNA-Seq dataset.
P values in brackets corrected for 12 development stages, where missing corrected p value was 1.

		All CNVs			Deletion CNVs			Duplication CNVs		
		Min.	Med.	Max.	Min.	Med.	Max.	Min.	Med.	Max.
Embryonic	P value	0.982	0.874	0.409	0.638	0.923	0.979	0.357	0.417	0.523
	Coeff.	+	+	-	-	-	-	+	+	-
Early foetal A	P value	0.532	0.495	0.721	0.704	0.536	0.519	0.605	0.576	0.897
	Coeff.	+	+	+	+	+	+	+	+	+
Early foetal B	P value	0.465	0.746	0.855	0.947	0.719	0.708	0.321	0.673	0.688
	Coeff.	+	+	+	-	+	+	+	+	+
Early mid-foetal A	P value	0.514	0.257	0.487	0.426	0.613	0.887	0.934	0.444	0.748
	Coeff.	-	-	-	-	-	-	-	-	-
Early mid-foetal B	P value	0.676	0.457	0.834	0.304	0.382	0.806	0.864	0.646	0.993
	Coeff.	-	-	-	-	-	-	-	-	+
Late mid-foetal	P value	0.835	0.914	0.650	0.424	0.513	0.838	0.601	0.533	0.605
	Coeff.	+	+	+	-	-	+	+	+	+
Late foetal	P value	0.806	0.495	0.746	0.341	0.801	0.673	0.738	0.185	0.224
	Coeff.	-	+	+	-	-	-	+	+	+
Neonatal and early infancy	P value	0.131	0.0869	0.0221 (0.332)	0.989	0.507	0.138	0.111	0.100	0.0815
	Coeff.	+	+	+	+	+	+	+	+	+
Late infancy	P value	0.342	0.393	0.778	0.764	0.576	0.178	0.181	0.487	0.618
	Coeff.	+	+	+	-	+	+	+	+	-
Early childhood	P value	0.816	0.682	0.748	0.966	0.593	0.263	0.984	0.920	0.937
	Coeff.	+	+	+	-	+	+	+	+	-
Middle and late childhood	P value	0.584	0.456	0.475	0.849	0.587	0.366	0.384	0.132	0.191
	Coeff.	-	-	-	-	+	+	-	-	-

Adolescence	**P value**	0.254	0.646	0.780	0.683	0.457	0.202	0.0457 (0.686)	0.210	0.338
	Coeff.	-	-	-	+	+	+	-	-	-
Young adulthood	**P value**	0.294	0.134	0.123	0.425	0.362	0.889	0.267	0.179	0.0578 (0.867)
	Coeff.	-	-	-	-	-	+	-	-	-
Middle adulthood	**P value**	0.484	0.434	0.745	0.160	0.489	0.830	0.755	0.671	0.889
	Coeff.	-	-	-	-	-	-	+	-	+
Late adulthood	**P value**	0.712	0.614	0.623	0.780	0.472	0.125	0.522	0.121	0.544
	Coeff.	-	-	+	-	+	+	-	-	-

Table 8.9: Logistic regression results testing CNV case control status on development stage characteristic scores calculated in the Kang microarray dataset. P values in brackets corrected for 15 development stages.

		All CNVs			Deletion CNVs			Duplication CNVs		
		Min.	**Med.**	**Max.**	**Min.**	**Med.**	**Max.**	**Min.**	**Med.**	**Max.**
Embryonic	**P value**	0.371	0.160	0.353	0.303	0.310	0.329	0.880	0.257	0.631
	Coeff.	+	+	+	+	+	+	+	+	+
Early foetal A	**P value**	0.0351 (0.527)	0.00218 (0.0327)	0.000724 (0.0109)	0.00113 (0.0170)	0.000446 (0.00669)	0.000648 (0.00972)	0.965	0.287	0.0896
	Coeff.	+	+	+	+	+	+	+	+	+
Early foetal B	**P value**	0.121	0.0722	0.0262 (0.393)	0.0361 (0.541)	0.0205 (0.308)	0.0540 (0.810)	0.726	0.634	0.155
	Coeff.	+	+	+	+	+	+	+	+	+
Early mid-foetal A	**P value**	0.253	0.264	0.487	0.186	0.290	0.478	0.738	0.572	0.713
	Coeff.	+	+	+	+	+	+	+	+	+
Early mid-foetal B	**P value**	0.000133 (0.00200)	6.24×10^{-5} (0.000937)	0.000750 (0.0113)	0.000382 (0.00573)	0.000524 (0.00785)	0.00483 (0.0725)	0.0581 (0.871)	0.0278 (0.416)	0.0394 (0.591)
	Coeff.	-	-	-	-	-	-	-	-	-
Late mid-foetal	**P value**	0.112	0.454	0.684	0.459	0.404	0.300	0.141	0.799	0.676
	Coeff.	-	-	-	-	-	-	-	-	+
Late foetal	**P value**	0.397	0.242	0.207	0.719	0.371	0.547	0.183	0.410	0.262
	Coeff.	-	+	+	+	+	+	-	+	+
Neonatal and early infancy	**P value**	0.0735	0.159	0.825	0.367	0.636	0.803	0.122	0.146	0.562
	Coeff.	+	+	+	+	+	-	+	+	+
Late infancy	**P value**	0.991	0.740	0.616	0.517	0.473	0.753	0.517	0.266	0.394
	Coeff.	-	-	-	+	+	+	-	-	-

Early childhood	**P value**	0.408	0.135	0.00181 (0.0272)	0.351	0.189	0.0636 (0.953)	0.739	0.395	0.0132 (0.198)
	Coeff.	+	+	+	+	+	+	+	+	+
Middle and late childhood	**P value**	0.0111 (0.166)	0.114	0.0703	0.302	0.753	0.514	0.0163 (0.244)	0.0777	0.0913
	Coeff.	-	-	-	-	-	-	-	-	-
Adolescence	**P value**	0.947	0.353	0.0114 (0.171)	0.481	0.349	0.0477 (0.715)	0.532	0.681	0.100
	Coeff.	-	+	+	+	+	+	-	+	+
Young adulthood	**P value**	0.196	0.122	0.419	0.361	0.118	0.116	0.340	0.480	0.828
	Coeff.	-	-	-	-	-	-	-	-	+
Middle adulthood	**P value**	0.447	0.515	0.745	0.464	0.563	0.664	0.846	0.750	0.908
	Coeff.	+	-	-	+	-	-	+	-	-
Late adulthood	**P value**	0.136	0.184	0.151	0.999	0.891	0.550	0.0289 (0.434)	0.0726	0.177
	Coeff.	-	-	-	+	-	-	-	-	-

Table 8.10: Logistic regression results testing CNV singleton status on development stage characteristic scores calculated in the Kang microarray dataset.
P values in brackets corrected for 15 development stages, where missing corrected p value was 1.

		All CNVs			Deletion CNVs			Duplication CNVs		
		Min.	Med.	Max.	Min.	Med.	Max.	Min.	Med.	Max.
Embryonic	**P value**	0.945	0.869	0.400	0.506	0.768	0.816	0.271	0.249	0.583
	Coeff.	-	+	-	-	-	-	+	+	-
Early foetal A	**P value**	0.486	0.392	0.910	0.652	0.470	0.473	0.516	0.435	0.811
	Coeff.	+	+	+	+	+	+	+	+	-
Early foetal B	**P value**	0.894	0.876	0.784	0.356	0.538	0.601	0.352	0.510	0.545
	Coeff.	+	-	+	-	-	-	+	+	+
Early mid-foetal A	**P value**	0.839	0.686	0.678	0.295	0.505	0.861	0.583	0.719	0.879
	Coeff.	-	-	-	-	-	-	+	+	+
Early mid-foetal B	**P value**	0.863	0.789	0.651	0.350	0.583	0.886	0.330	0.223	0.239
	Coeff.	+	+	+	-	-	-	+	+	+
Late mid-foetal	**P value**	0.936	0.903	0.850	0.279	0.444	0.821	0.465	0.597	0.841
	Coeff.	+	-	-	-	-	-	+	+	+
Late foetal	**P value**	0.927	0.813	0.987	0.436	0.672	0.786	0.771	0.388	0.473
	Coeff.	-	+	-	-	-	-	+	+	+
Neonatal and early infancy	**P value**	0.219	0.128	0.178	0.979	0.410	0.118	0.309	0.273	0.678
	Coeff.	+	+	+	-	+	+	+	+	+
Late infancy	**P value**	0.999	0.941	0.944	0.755	0.655	0.323	0.845	0.698	0.565
	Coeff.	+	+	-	-	+	+	-	-	-
Early childhood	**P value**	0.822	0.245	0.436	0.811	0.426	0.401	0.987	0.258	0.369
	Coeff.	+	+	+	+	+	+	+	+	+

Middle and late childhood	**P value**	0.282	0.368	0.301	0.176	0.0581	0.0146	0.758	0.402	0.562
	Coeff.	+	+	+	+	+	+	-	-	-
Adolescence	**P value**	0.421	0.646	0.908	0.590	0.525	0.462	0.998	0.782	0.553
	Coeff.	+	+	-	+	+	+	+	-	-
Young adulthood	**P value**	0.983	0.392	0.455	0.680	0.996	0.394	0.374	0.150	0.116
	Coeff.	+	-	-	+	-	+	-	-	-
Middle adulthood	**P value**	0.767	0.482	0.320	0.335	0.798	0.845	0.923	0.327	0.164
	Coeff.	-	-	-	-	-	+	+	-	-
Late adulthood	**P value**	0.310	0.310	0.958	0.447	0.965	0.304	0.303	0.097	0.397
	Coeff.	-	-	+	-	+	+	-	-	-

Table 8.11: Logistic regression results testing CNV case control status on development stage characteristic scores calculated in the Kang microarray dataset without PMI covariate.

P values in brackets corrected for 15 development stages, where missing corrected p value was 1.

		All CNVs			Deletion CNVs			Duplication CNVs		
		Min.	**Med.**	**Max.**	**Min.**	**Med.**	**Max.**	**Min.**	**Med.**	**Max.**
Embryonic	**P value**	0.374	0.156	0.443	0.354	0.352	0.374	0.792	0.184	0.704
	Coeff.	+	+	+	+	+	+	+	+	+
Early foetal A	**P value**	0.0129 (0.194)	0.000358 (0.00537)	0.000153 (0.00229)	0.000292 (0.00438)	0.000426 (0.00640)	0.00152 (0.0228)	0.867	0.089	0.0136 (0.204)
	Coeff.	+	+	+	+	+	+	+	+	+
Early foetal B	**P value**	0.0713	0.0358 (0.537)	0.0953	0.0137 (0.0228)	0.0161 (0.242)	0.0850	0.686	0.410	0.367
	Coeff.	+	+	+	+	+	+	+	+	+
Early mid-foetal A	**P value**	0.562	0.174	0.442	0.0505 (0.758)	0.0592 (0.887)	0.173	0.423	0.823	0.998
	Coeff.	+	+	+	+	+	+	-	-	-
Early mid-foetal B	**P value**	0.344	0.518	0.932	0.904	0.947	0.865	0.265	0.416	0.971
	Coeff.	-	-	+	-	-	+	-	-	-
Late mid-foetal	**P value**	0.107	0.294	0.377	0.416	0.277	0.223	0.143	0.681	0.939
	Coeff.	-	-	-	-	-	-	-	-	-
Late foetal	**P value**	0.0750	0.737	0.867	0.580	0.989	0.650	0.0706	0.622	0.587
	Coeff.	-	+	+	-	-	-	-	+	+
Neonatal and early infancy	**P value**	0.165	0.283	0.0808	0.111	0.302	0.0417 (0.626)	0.531	0.562	0.578
	Coeff.	-	-	-	-	-	-	-	-	-
Late infancy	**P value**	0.0177 (0.266)	0.0841	0.326	0.207	0.303	0.475	0.0396 (0.593)	0.172	0.513
	Coeff.	-	-	-	-	-	-	-	-	-
Early childhood	**P value**	6.89×10^{-6} (0.000103)	1.17×10^{-5} (0.000175)	5.93×10^{-5} (0.000890)	1.24×10^{-5} (0.000186)	6.05×10^{-6} (9.08×10^{-5})	4.80×10^{-7} (7.20×10^{-6})	0.0193 (0.289)	0.0652 (0.978)	0.337
	Coeff.	+	+	+	+	+	+	+	+	+

Middle and late childhood	**P value**	0.0874	0.698	0.495	0.586	0.938	0.517	0.0590 (0.884)	0.592	0.162
	Coeff.	-	-	+	-	-	-	-	-	+
Adolescence	**P value**	0.857	0.814	0.534	0.588	0.727	0.734	0.534	0.557	0.329
	Coeff.	-	-	-	-	-	-	+	+	+
Young adulthood	**P value**	0.163	0.183	0.450	0.208	0.197	0.0720	0.467	0.515	0.594
	Coeff.	-	-	-	-	-	-	-	-	+
Middle adulthood	**P value**	0.368	0.551	0.812	0.136	0.531	0.808	0.701	0.878	0.573
	Coeff.	+	+	-	+	+	+	-	+	-
Late adulthood	**P value**	0.116	0.214	0.0343 (0.514)	0.694	0.540	0.129	0.0451 (0.677)	0.214	0.154
	Coeff.	-	-	-	-	-	-	-	-	-

Table 8.12: Logistic regression results testing CNV singleton status on development stage characteristic scores calculated in the Kang microarray dataset without PMI covariate.
P values in brackets corrected for 15 development stages, where missing corrected p value was 1.

8.2 Additional figures for Chapter 3

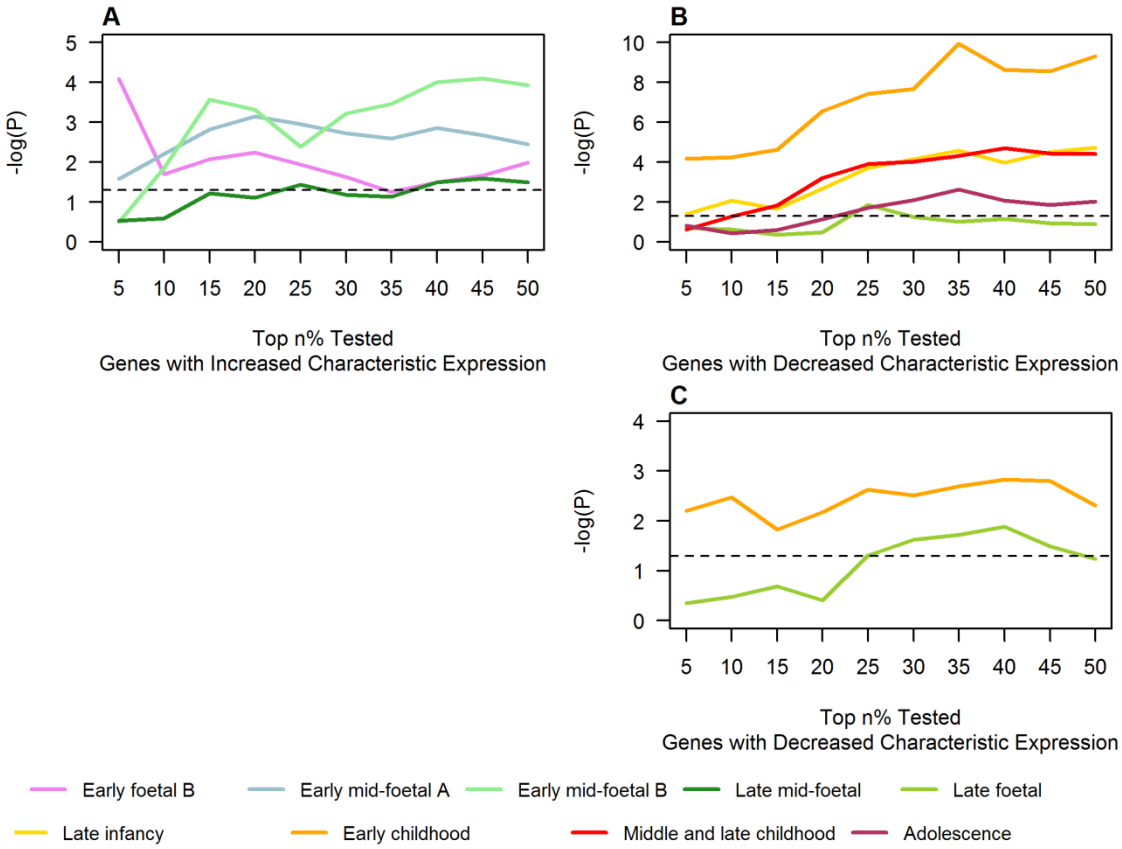

Figure 8.1: Results from Mann-Whitney tests to verify significant regression models between development stage characteristic scores calculated in the BrainSpan RNA-Seq dataset and Brown's p values.
Genes were ranked by absolute characteristic scores and the top n% split into positive and negative subsets. The subset consistent with the direction of the significant regression model in Figure 3.1 was tested in a one-sided Mann-Whitney test against the bottom 50%, panel A tested positive subset; panels B & C tested negative subset. Panels A & B tested for smaller SCZ Brown's p values; panel C tested for smaller BPD Brown's p values. Black dashed line is p = 0.05.

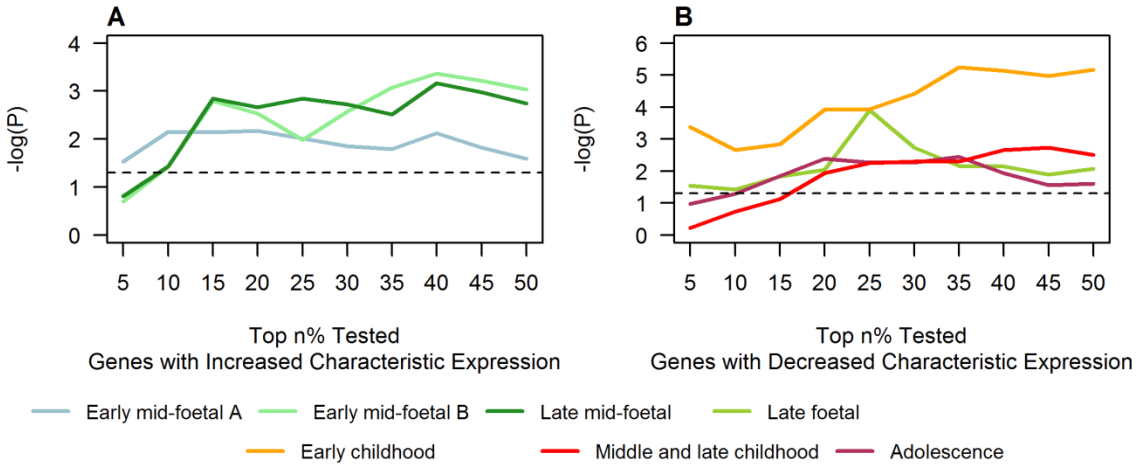

Figure 8.2: Results from Mann-Whitney tests to verify significant regression models between development stage characteristic scores calculated in the BrainSpan RNA-Seq dataset and SCZ Simes' p values.
Genes were ranked by absolute characteristic scores and the top n% split into positive and negative subsets. The subset consistent with the direction of the significant regression model in Figure 3.1 was tested in a one-sided Mann-Whitney test against the bottom 50% for smaller SCZ Simes' p values, panel A tested positive subset; panel B tested negative subset. Black dashed line is p = 0.05.

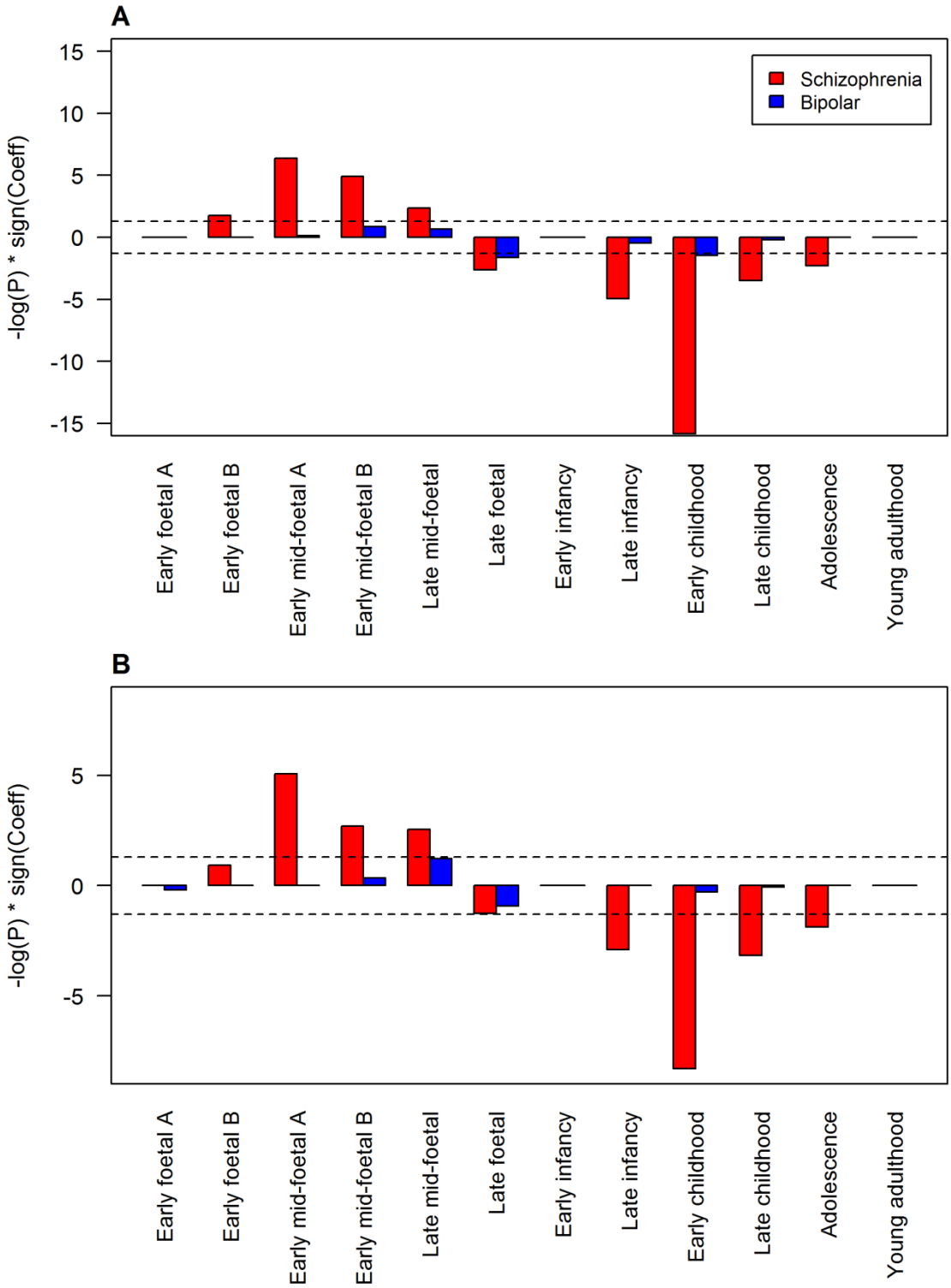

Figure 8.3: Results from linear regression of development stage characteristic scores calculated in the BrainSpan RNA-Seq dataset, excluding MHC genes.
P values were –log10 transformed and multiplied by the sign of the coefficient, therefore bars above the origin indicate positive regression coefficients; bars below the origin indicate negative regression coefficients. Panel A tested Brown's logP; panel B tested Simes' logP. All p values were corrected for 12 development stages using Bonferroni's method. Black dashed line is p = 0.05.

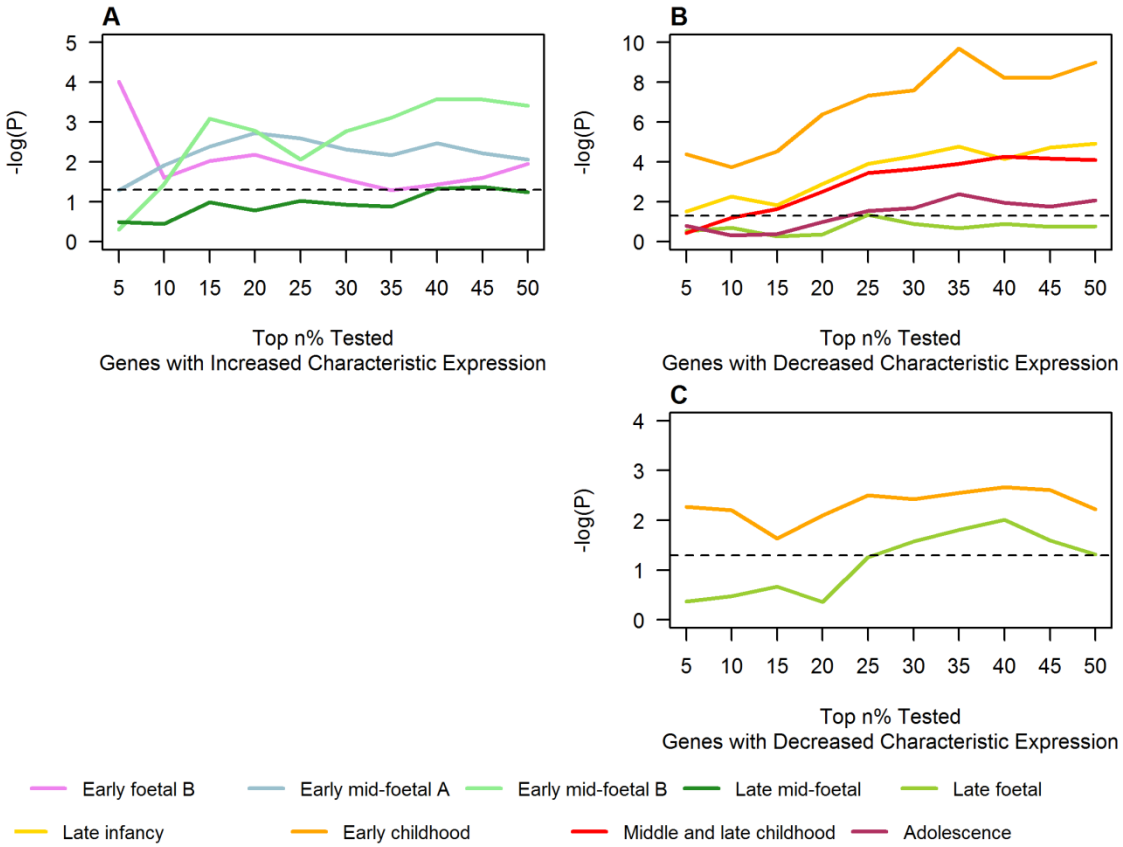

Figure 8.4: Results from Mann-Whitney tests to verify significant regression models between development stage characteristic scores calculated in the BrainSpan RNA-Seq dataset and Brown's p values, excluding MHC genes.
Genes were ranked by absolute characteristic scores and the top n% split into positive and negative subsets. The subset consistent with the direction of the significant regression model in Figure 8.3 was tested in a one-sided Mann-Whitney test against the bottom 50%, panel A tested positive subset; panels B & C tested negative subset. Panels A & B tested SCZ Brown's p values; panel C tested BPD Brown's p values. Black dashed line is p = 0.05.

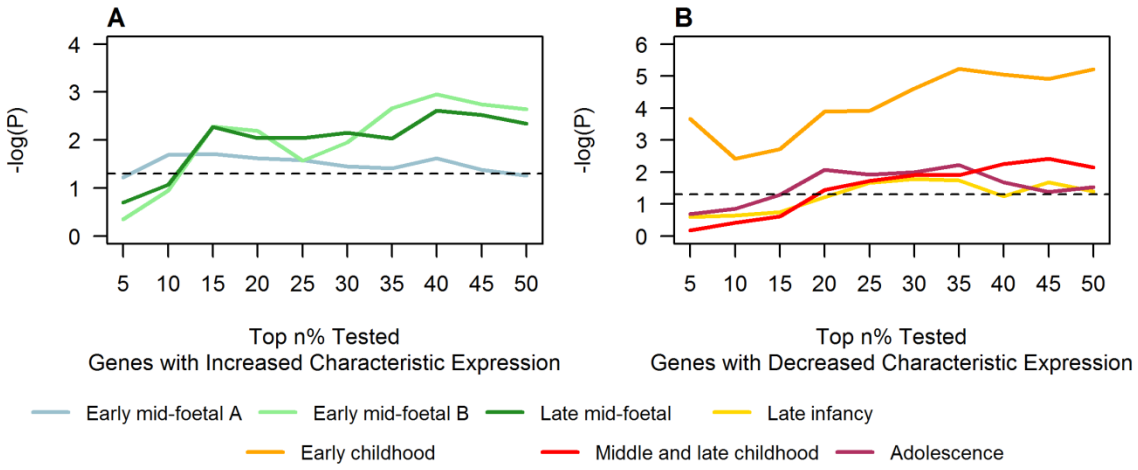

Figure 8.5: Results from Mann-Whitney tests to verify significant regression models between development stage characteristic scores calculated in the BrainSpan RNA-Seq dataset and SCZ Simes' p values, excluding MHC genes.
Genes were ranked by absolute characteristic scores and the top n% split into positive and negative subsets. The subset consistent with the direction of the significant regression model in Figure 8.3 was tested in a one-sided Mann-Whitney test against the bottom 50% for smaller SCZ Simes' p values, panel A tested positive subset; panel B tested negative subset. Black dashed line is p = 0.05.

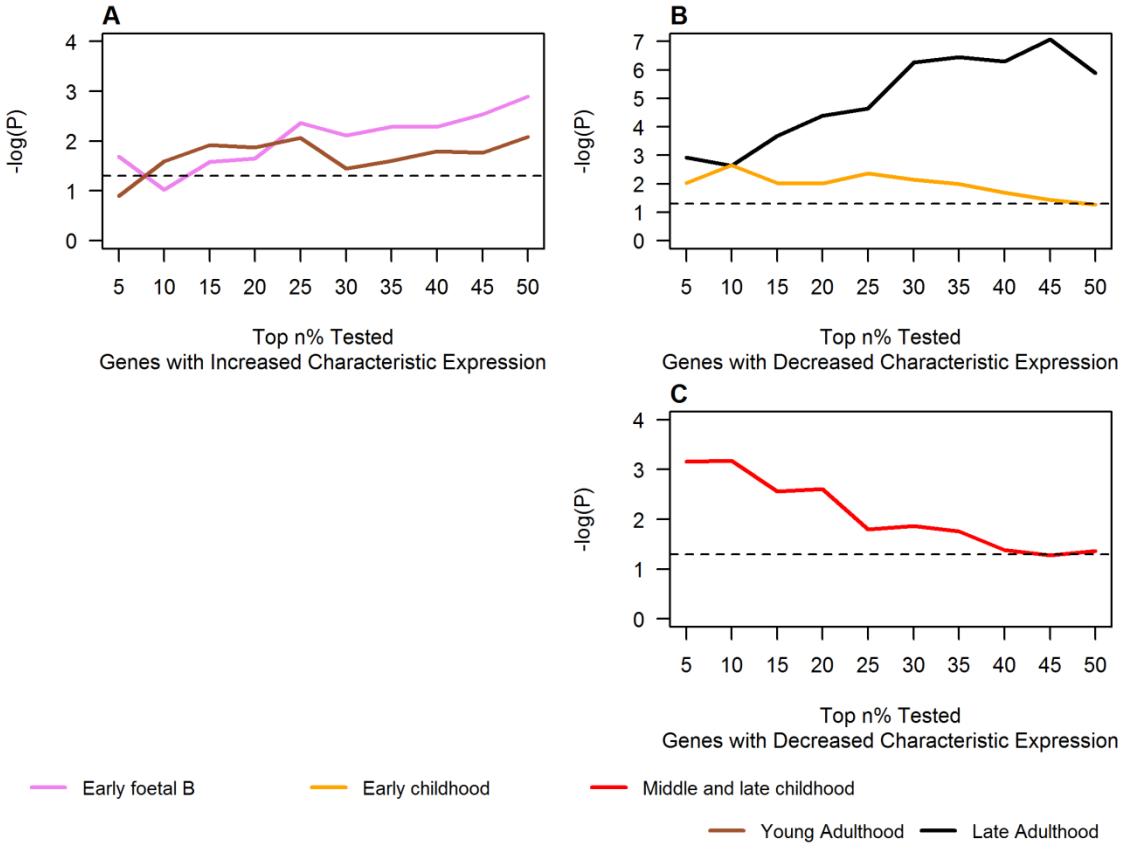

Figure 8.6: Results from Mann-Whitney tests to verify significant regression models between development stage characteristic scores calculated in the Kang microarray dataset and Brown's p values.
Genes were ranked by absolute characteristic scores and the top n% split into positive and negative subsets. The subset consistent with the direction of the significant regression model in Figure 3.2 was tested in a one-sided Mann-Whitney test against the bottom 50%, panel A test positive subset; panels B & C tested negative subset. Panels A & B tested SCZ Brown's p values; panel C tested BPD Brown's p values. Black dashed line is p = 0.05.

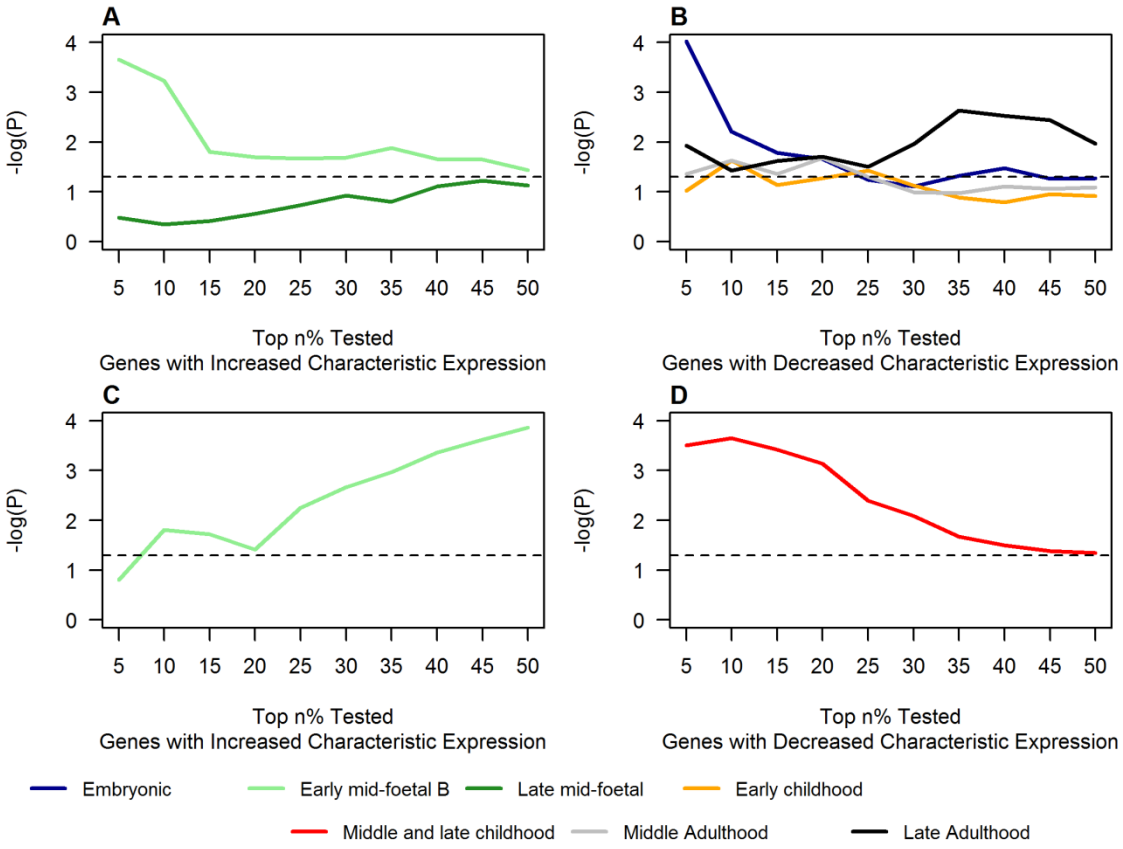

Figure 8.7: Results from Mann-Whitney tests to verify significant regression models between development stage characteristic scores calculated in the Kang microarray dataset and Simes' p values.
Genes were ranked by absolute characteristic scores and the top n% split into positive and negative subsets. The subset consistent with the direction of the significant regression model in Figure 3.2 was tested in a one-sided Mann-Whitney test against the bottom 50%, panels A & C tested positive subsets; panels B & D tested negative subsets. Panels A & B tested SCZ Simes' p values; panels C & D tested BPD Simes' p values. Black dashed line is p = 0.05.

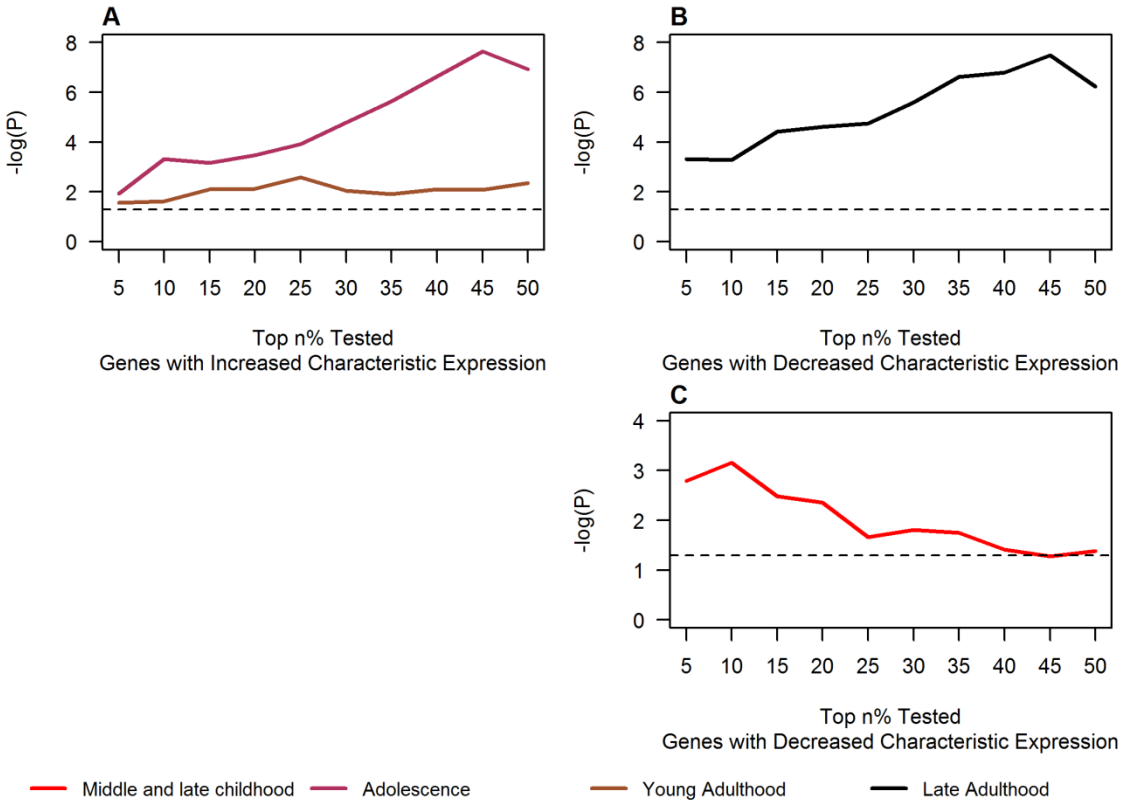

Figure 8.8: Results from Mann-Whitney tests to verify significant regression models between development stage characteristic scores calculated in the Kang microarray dataset and Brown's p values, excluding MHC genes.
Genes were ranked by absolute characteristic scores and the top n% split into positive and negative subsets. The subset consistent with the direction of the significant regression model in Figure 3.3 was tested in a one-sided Mann-Whitney test against the bottom 50%, panel A tested positive subset; panels B & C tested negative subset. Panels A & B tested SCZ Brown's p values; panel C tested BPD Brown's p values. Black dashed line is p = 0.05.

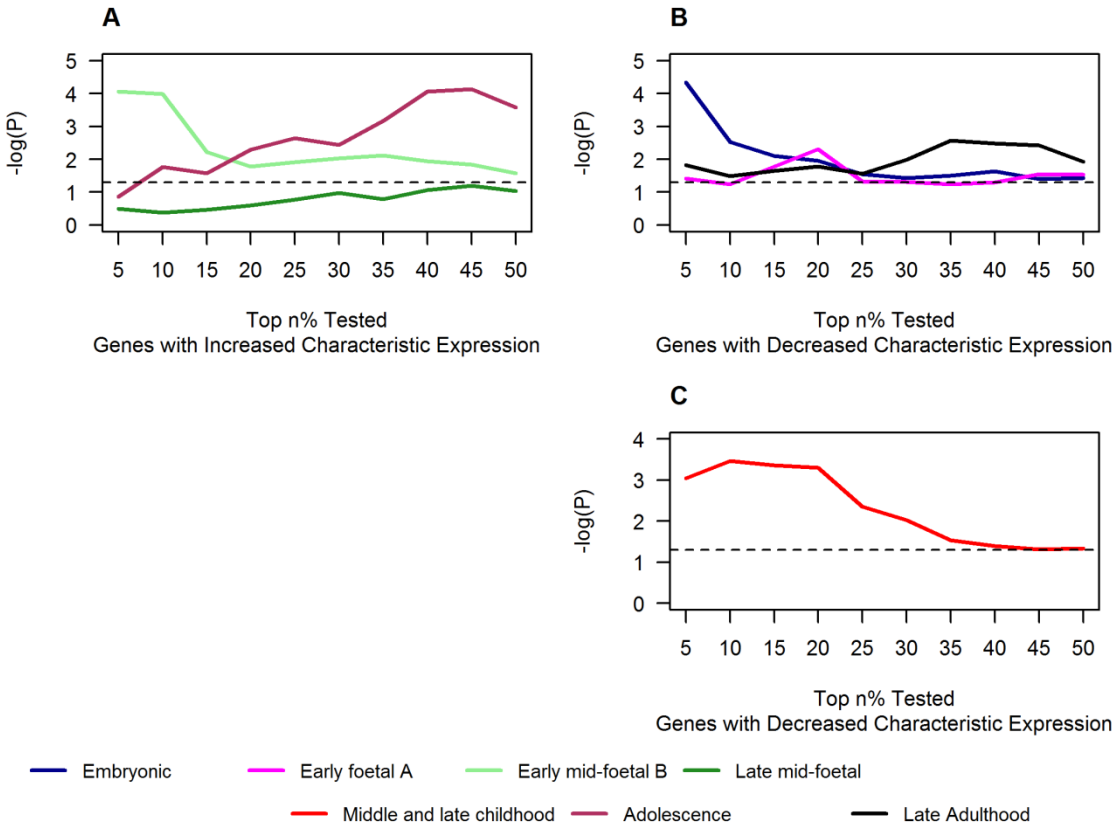

Figure 8.9: Results from Mann-Whitney tests to verify significant regression models between development stage characteristic scores calculated in the Kang microarray dataset and Simes' p values, excluding MHC genes.
Genes were ranked by absolute characteristic scores and the top n% split into positive and negative subsets. The subset consistent with the direction of the significant regression model in Figure 3.3 was tested in a one-sided Mann-Whitney test against the bottom 50%, panel A tested positive subset; panels B & C tested negative subset. Panels A & B tested SCZ Simes' p values; panel C tested BPD Simes' p values. Black dashed line is p = 0.05.

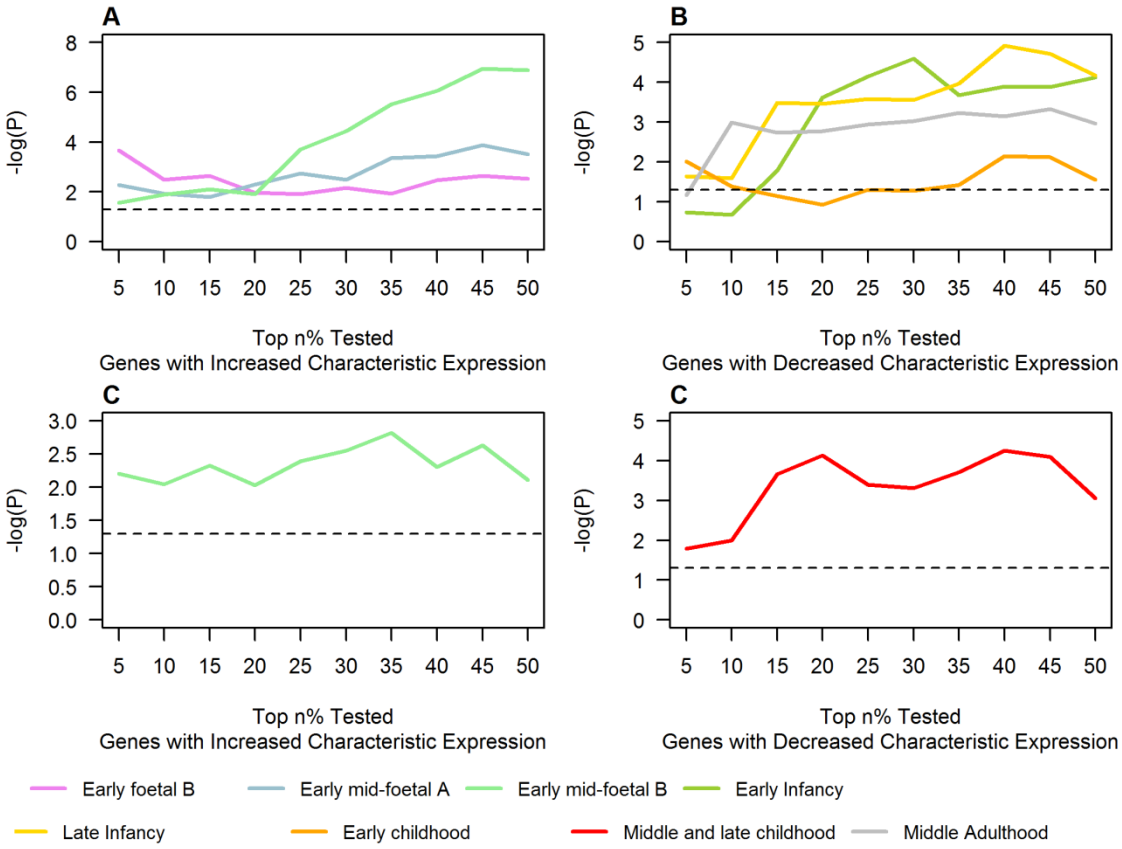

Figure 8.10: Results from Mann-Whitney tests to verify significant regression models between development stage characteristic scores calculated in the Kang microarray dataset without PMI covariate and Brown's p values.
Genes were ranked by absolute characteristic scores and the top n% split into positive and negative subsets. The subset consistent with the direction of the significant regression model in Figure 3.4 was tested in a one-sided Mann-Whitney test against the bottom 50%, panels A & C tested positive subset; panels B & D tested negative subset. Panels A & B tested SCZ Brown's p values; panels C & D tested BPD Brown's p values. Black dashed line is p = 0.05.

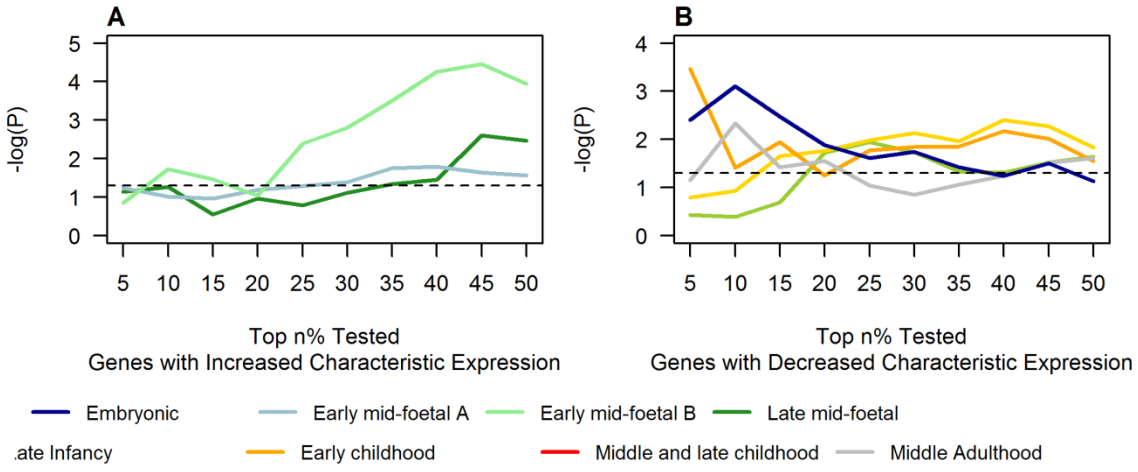

Figure 8.11: Results from Mann-Whitney tests to verify significant regression models between development stage characteristic scores calculated in the Kang microarray dataset without PMI covariate and SCZ Simes' p values.
Genes were ranked by absolute characteristic scores and the top n% split into positive and negative subsets. The subset consistent with the direction of the significant regression model in Figure 3.4 was tested in a one-sided Mann-Whitney test against the bottom 50% for smaller SCZ Simes' p values, panel A tested positive subset; panel B tested negative subset. Black dashed line is p = 0.05.

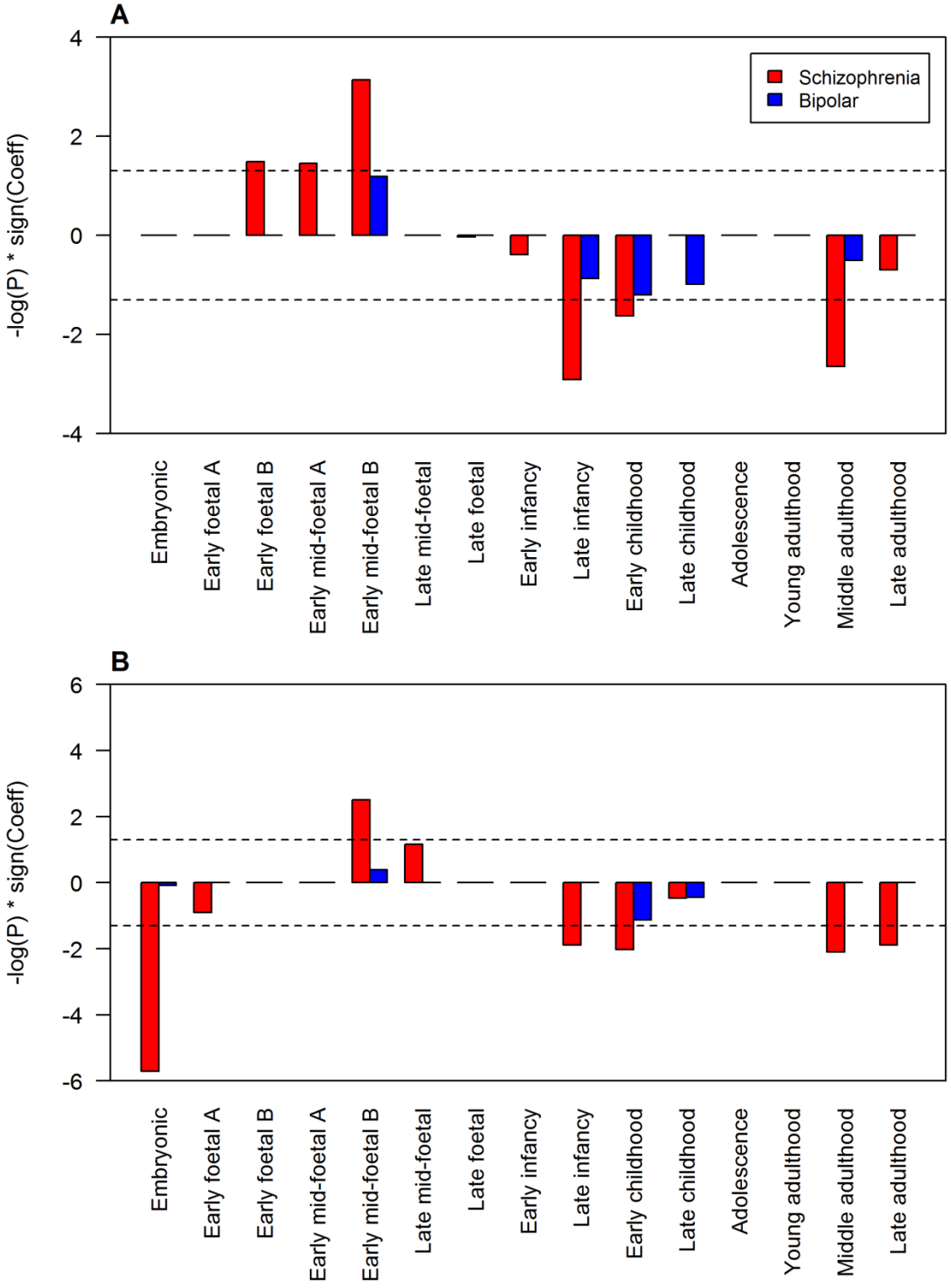

Figure 8.12: Results from linear regression of development stage characteristic scores calculated in the Kang microarray dataset without PMI covariate, excluding MHC genes.
P values were –log10 transformed and multiplied by the sign of the coefficient, therefore bars above the origin indicate positive regression coefficients; bars below the origin indicate negative regression coefficients. Panel A tested Brown's logP; panel B tested Simes' logP. All P values were corrected for 15 development stages using Bonferroni's method. Black dashed line is p = 0.05.

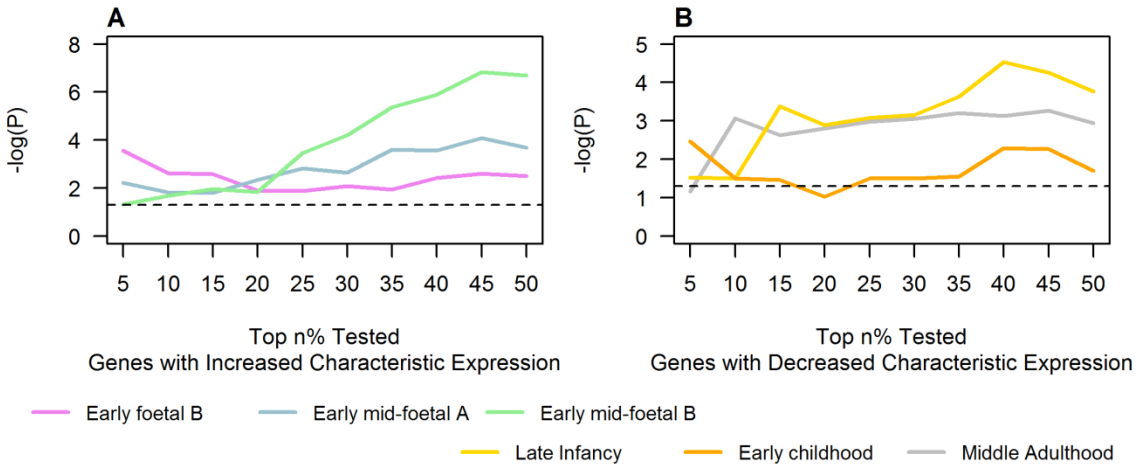

Figure 8.13: Results from Mann-Whitney tests to verify significant regression models between development stage characteristic scores calculated in the Kang microarray dataset without PMI covariate and SCZ Brown's p values, excluding MHC genes.
Genes were ranked by absolute characteristic scores and the top n% split into positive and negative subsets. The subset consistent with the direction of the significant regression model in Figure 8.12 was tested in a one-sided Mann-Whitney test against the bottom 50% for smaller SCZ Brown's p values, panel A tested positive subset; panel B tested negative subset. Black dashed line is p = 0.05.

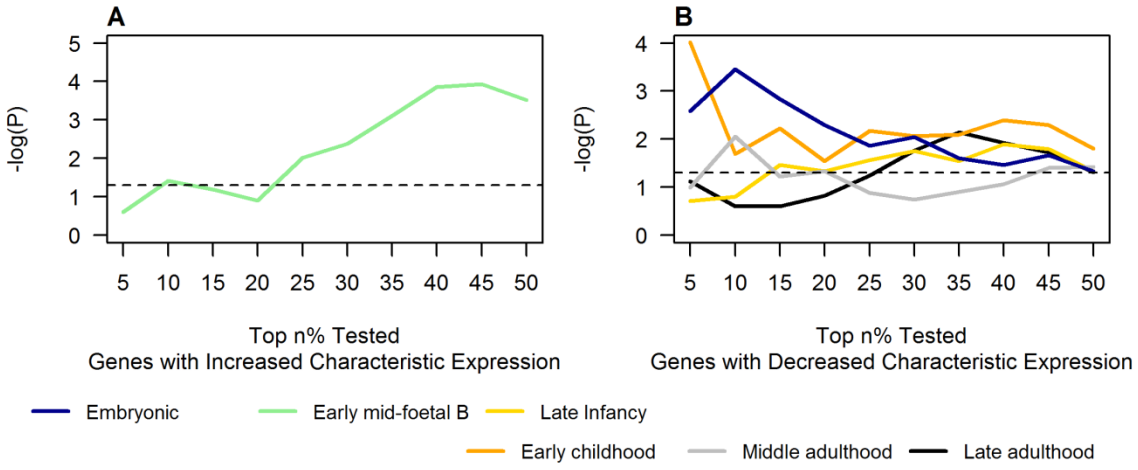

Figure 8.14: Results from Mann-Whitney tests to verify significant regression models between development stage characteristic scores calculated in the Kang microarray dataset without PMI covariate and SCZ Simes' p values, excluding MHC genes.
Genes were ranked by absolute characteristic scores and the top n% split into positive and negative subsets. The subset consistent with the direction of the significant regression model in Figure 8.12 was tested in a one-sided Mann-Whitney test against the bottom 50% for smaller SCZ Brown's p values, panel A tested positive subset; panel B tested negative subset. Black dashed line is p = 0.05.

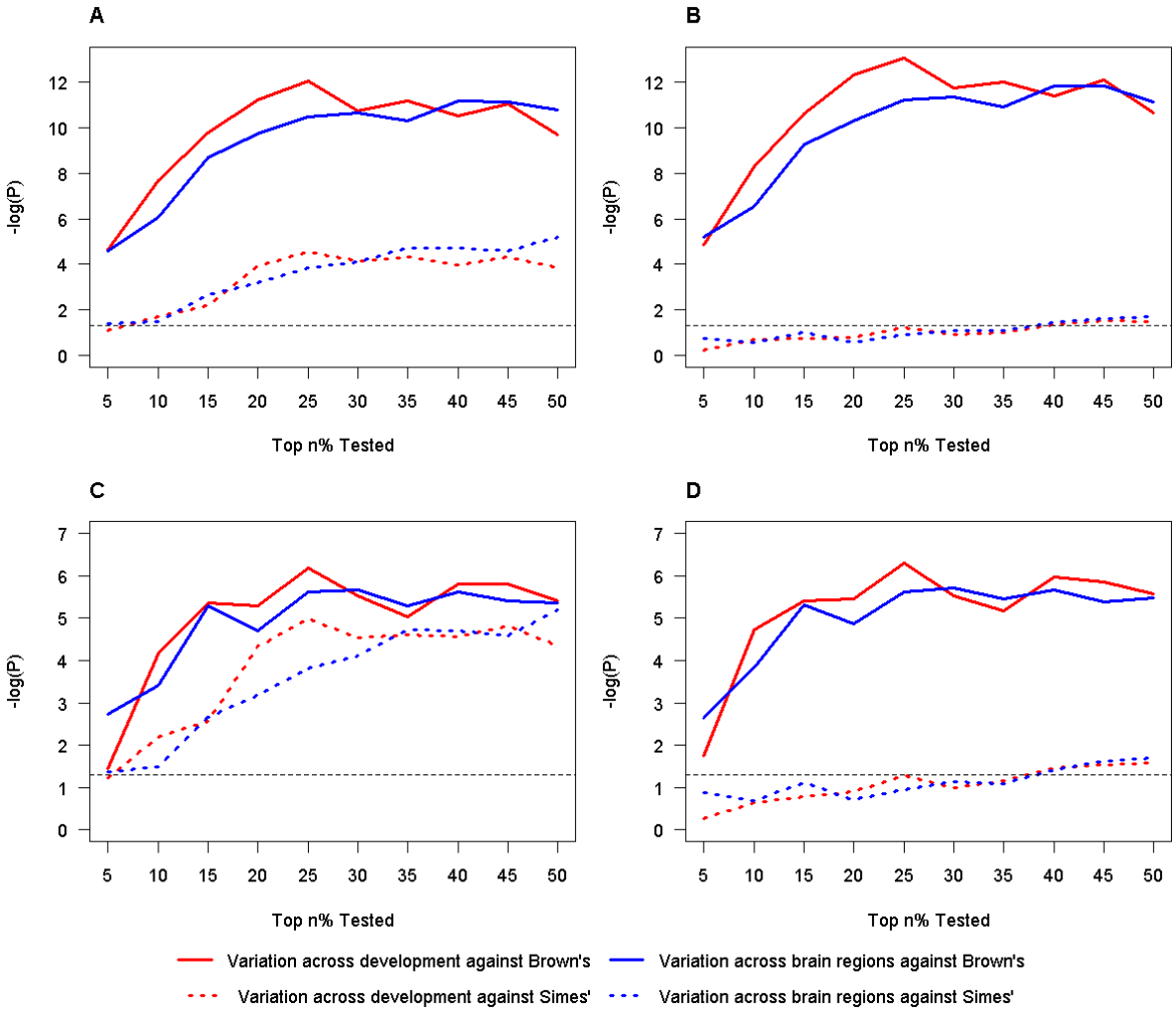

Figure 8.15: Results from Mann-Whitney tests for genes ranked by SCZ risk genes co-expression model p values calculated in the BrainSpan RNA-Seq dataset.
Genes ranked by co-expression model p values and top n% tested against bottom 50%. Panels A & B tested for smaller SCZ p values; panels C & D tested for smaller BPD p values. Panels A & C tested all genes, panels B & D excluded MHC genes. Black dashed line is 0.05.

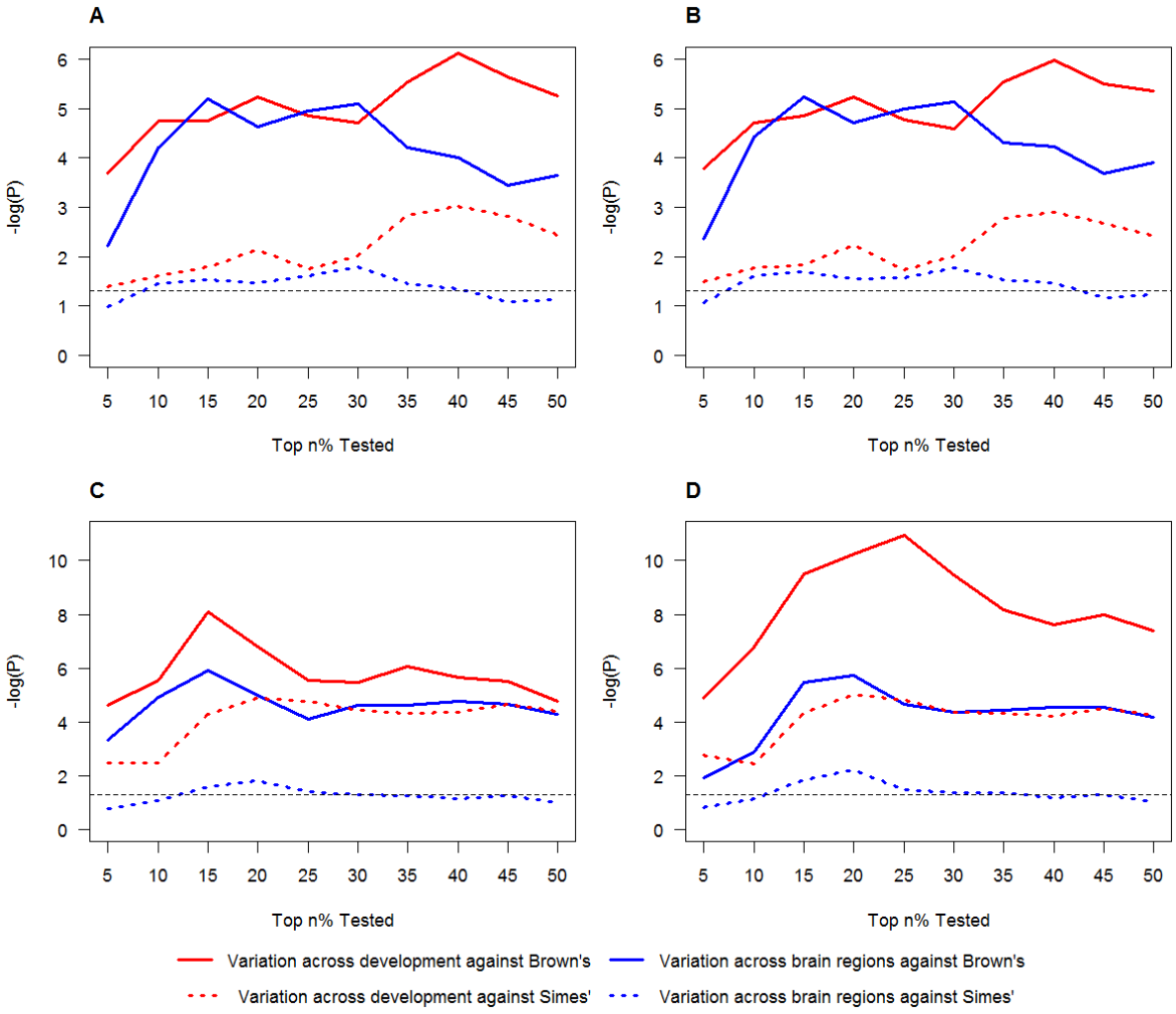

Figure 8.16: Results from Mann-Whitney tests for genes ranked by BPD risk genes co-expression model p values calculated in the BrainSpan RNA-Seq dataset.
Genes ranked by co-expression model p values and top n% tested against bottom 50%. Panels A & B tested for smaller SCZ p values; panels C & D tested for smaller BPD p values. Panels A & C tested all genes; panels B & D excluded MHC genes. Black dashed line is 0.05.

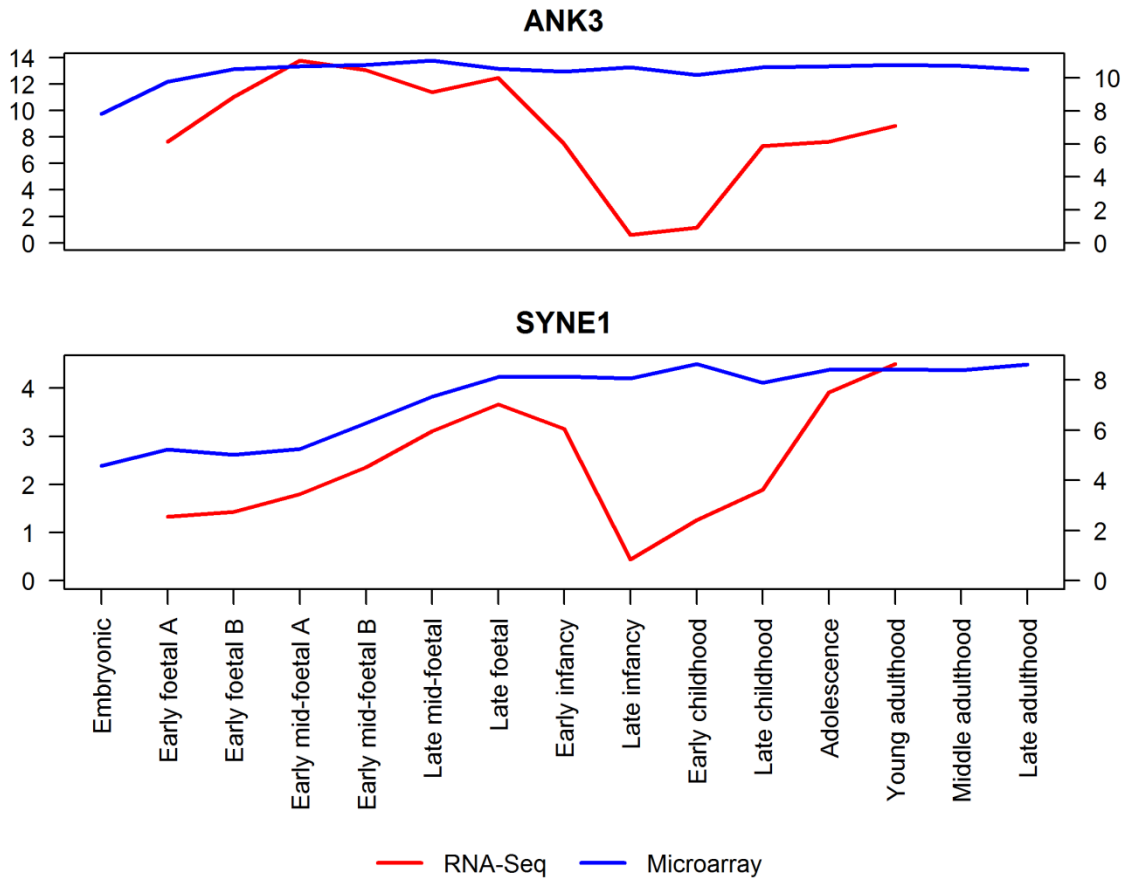

Figure 8.17: BPD risk genes comparing microarray and RNA-Seq expression values.
Risk genes identified from PGC GWAS whose co-expression indexed association in the RNA-Seq dataset and were present in the microarray dataset. Median expression values calculated for each development stage, scale on left for RNA-Seq data, scale on right for microarray data.

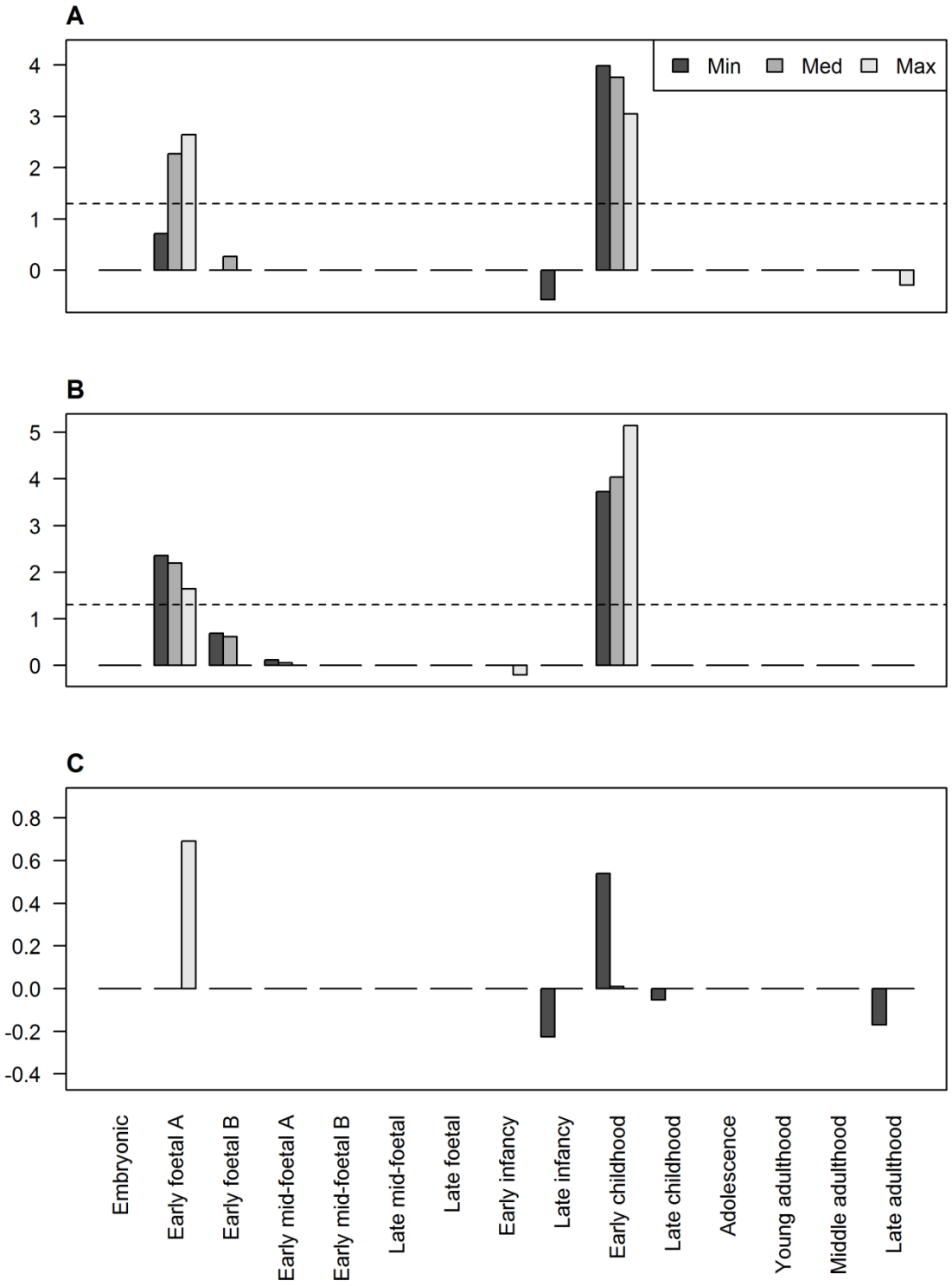

Figure 8.18: Logistic regression results testing CNV singleton status on development stage characteristic scores calculated in the Kang microarray dataset without PMI covariate.
Panel A is all CNVs, panel B deletions, and panel C duplications. P values were corrected for 15 development stages using Bonferroni's method. Black dashed line is 0.05.